Speaker Clustering Based on Utterance-oriented Dirichlet Process Mixture Model

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

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

Abstract

This paper provides the analytical solution and algorithm of UO-DPMM based on a non-parametric Bayesian manner, and thus realizes fully Bayesian speaker clustering. We carried out preliminary speaker clustering experiments by using a TIMIT database to compare the proposed method with the conventional Bayesian Information Criterion (BIC) based method, which is an approximate Bayesian approach. The results showed that the proposed method outperformed the conventional one in terms of both computational cost and robustness to changes in tuning parameters.

Index Terms: Non-parametric Bayesian model, Gibbs sampling, utterance-oriented DPMM, speaker clustering

1. Introduction

The great progress made in archiving speech data found on the Web increases the demands on how to find desirable speech data among the archives by using speech attributes (e.g., spoken keywords, speakers) as queries. In this situation, we are faced with the problem that how to automatically provide these attributes with archived speech data, which is growing from day to day. Therefore, the framework of a speaker clustering system should be flexible in order to handle an infinite number of clusters. In such a system, the number of speaker clusters would be adequately increased or decreased as new data are observed.

The hierarchical agglomerative clustering method [1] has been used frequently for speaker clustering. This method uses Bayesian information criterion (BIC) to optimize the number of clusters. However, this method is an approximate Bayesian approach, and has the following problems. 1.) computational cost increases exponentially with an increase in the amount of data because this method needs to compare the BIC scores for all data pairs, and 2.) heuristically given optimal parameters are necessary because the estimated number of clusters depends highly on a penalty parameter of the BIC. There is another method that increases the number of clusters with an increase in the number of data [2], but it cannot give a global solution.

To solve these problems, we propose a fully Bayesian speaker clustering method based on non-parametric Bayesian learning and apply sampling-based methods for estimating a model. Since non-parametric Bayesian learning deals with unobserved data in its formulation, this method is suitable for a framework in which the number of speaker clusters increases or decreases as new (i.e., unobserved) data are observed. Infinite hidden Markov models (IHMM) [3, 4] are an example of clustering based on non-parametric Bayesian modeling. A hidden Markov model with an infinite number of states is trained using the frame-oriented observations in all the utterances. We can then regard the states of an IHMM as the speaker clusters. Although this model can segment speech by a speaker and cluster the segmented data for each speaker simultaneously, the number of speakers is frequently overestimated. In contrast, we propose a novel speaker clustering method called an "utterance-oriented Dirichlet process mixture model (UO-DPMM)," in which a generative unit is not a frame but an utterance. Because the speaker change usually occurs utterance-by-utterance, the UO-DPMM would be a reasonable model to represent the speaker variability in speech data. This paper provides the analytical solution and algorithm of UO-DPMM based on a non-parametric Bayesian manner, and thus realizes fully Bayesian speaker clustering.

The rest of the present paper is organized as follows. In section 2, the algorithm of the UO-DPMM is described in detail. In section 3, a speaker clustering experiment to verify the effectiveness of the proposed method is presented. In section 4, the paper is concluded, and future works are mentioned.

2. Speaker clustering based on an utterance-oriented DPMM

We attempt to assign speaker labels to each speech segment (i.e., utterance) by estimating the number of speakers. First, we define a mixture distribution whose components correspond to each speaker. This means that the problem of speaker clustering is replaced with the problem of estimating the latent variables of the mixture distribution. Let $\mathbf{O}_1$, $\mathbf{O}_2$, $\cdots$,and $\mathbf{O}_U$ be observation sequences, each of which represents an utterance, and z_1, z_2, $\cdots$,and z_U be the corresponding latent variables. The optimal set of these latent variables ($\bar{z_1}$, $\bar{z_2}$, $\cdots$,and $\bar{z_U}$) are estimated by maximizing the posterior distribution as follows:

$$\bar{z_1}, \bar{z_2}, \cdots, \bar{z_U} = \underset{z_1, z_2, \cdots, z_U}{\arg\max}\ P(z_1, z_2, \cdots, z_U | \mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_U) \quad (1)$$

From the viewpoint of computational cost, it is not realistic to evaluate all pairs of z_1, z_2, $\cdots$, and z_U. Therefore, we approximately estimated the pairs of z_1, z_2, $\cdots$, and z_U that maximize $P(z_1, z_2, \cdots, z_U | \mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_U)$ by applying Gibbs sampling. By using this method with Gibbs sampling, we sampled each latent variable z_u from the conditional distribution that is conditioned by the other latent variables ($\mathbf{z}_{\setminus u} = \{z_i\}_{i \neq u}$) and iterated the aforementioned sampling process until certain conditions were met.

The rest of this section is organized as follows. In 2.1, we define an utterance-oriented generative model in which the number of speaker clusters is fixed. We call this model the "finite speaker model (FSM)." In this subsection, we also present a method for assigning speaker labels to each utterance by using collapsed Gibbs sampling [5]. The FSM was extended so that

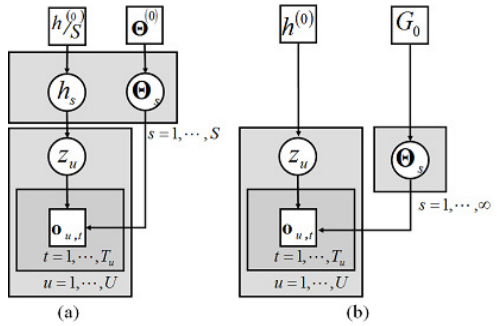

Figure 1: *Graphical models for (a) finite and (b) infinite speaker models.*

it could estimate not only the latent variables but also the number of speakers. We present this extension, named the "infinite speaker model (ISM)." In the ISM, the number of speakers can be estimated optimally from the data by introducing the Dirichlet process to the prior distribution of the FSM (2.2, 2.3, 2.4)D

2.1. Utterance-oriented generative model

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

We assume that a D-dimensional Gaussian distribution for each speaker generates the utterances from the corresponding speaker and that a speaker model is represented by a mixture of these distributions [i.e., a Gaussian mixture model (GMM)]. We then assume that each utterance is generated as an i.i.d. from this GMM and that each feature vector $\mathbf{o}_{u,t}$ is generated as an i.i.d. from a mixture component to which the utterance is assigned. This utterance-oriented generative model is described as follows, where $\{z_u\}_{u=1}^{U}$ denotes the latent variables assigned to the u-th utterance (i.e., these variables represent the indices of speaker clusters), $\mathcal{M}\left(\cdot|\{h_s\}_{s=1}^{S}\right)$ represents a multinomial distribution whose parameter h_s corresponds to the weight of speaker cluster s, S represents the number of speaker clusters, $\mathcal{N}(\cdot|\mathbf{\Theta}_s = \{\boldsymbol{\mu}_s, \mathbf{\Sigma}_s\})$ denotes a Gaussian distribution, $\boldsymbol{\mu}_s$ and $\mathbf{\Sigma}_s$ are respectively the mean vector and covariance matrix, and $h^{(0)}$ and $\mathbf{\Theta}^{(0)}$ are respectively the hyper-parameters of the prior distribution of h_s and $\mathbf{\Theta}_s$:

$$h_s \sim p(\cdot|h^{(0)}/S) \tag{2}$$
$$z_u \sim \mathcal{M}(\cdot|\{h_s\}_{s=1}^{S}) \tag{3}$$
$$\mathbf{o}_{u,t} \sim \mathcal{N}(\cdot|\mathbf{\Theta}_{z_u}) \tag{4}$$
$$\mathbf{\Theta}_s \sim p(\cdot|\mathbf{\Theta}^{(0)}) \tag{5}$$

The graphical model is shown in Fig. 1(a). Note that, in this case, the number of speakers S is fixed (i.e., finite).

In utterance-generative models, the likelihood for the set of observation vectors given the latent variable sequence ($\mathbf{Z} \triangleq \{z_u\}_{u=1}^{U}$) is expressed as follows:

$$p(\mathbf{O}|\mathbf{Z}, \mathbf{h}, \boldsymbol{\mu}, \mathbf{\Sigma}) = \prod_{u=1}^{U} h_{z_u} \prod_{t=1}^{T_u} \mathcal{N}(\mathbf{o}_{u,t}|\boldsymbol{\mu}_{z_u}, \mathbf{\Sigma}_{z_u}) \tag{6}$$

By introducing conjugate prior distributions of h_s, $\boldsymbol{\mu}_s$, and $\mathbf{\Sigma}_s$, we can derive analytically the posterior distributions for the latent variables. We can then use collapsed Gibbs sampling for the latent variables. In [6], $\mathcal{N}(\cdot)$ in Eq. (6) was formulated using GMM.

Table 1: *Algorithm of the proposed method.*

```
Initialize S and {z_u}_{u=1}^U.
repeat
    for u = shuffle (1, ··· , U) do
        Sample z_u according to Eq. (9).
        if z_u = S + 1 then
            Θ_{S+1} ~ G_0 ( Θ | Θ^(0) )
            S ← S + 1
        end if
    end for
until some condition is met
```

2.2. Dirichlet process mixture model (DPMM)

We attempt to extend the FSM described in 2.1 to the ISM. The ISM corresponds to the DPMM [7], which uses the Dirichlet process for the prior distribution of mixture weights in mixture models. In this case, the DPMM was implemented using the Chinese restaurant process (CRP). The reason we used the CRP is that it has the potential to avoid local solutions due to its sampling-based implementation. Furthermore, we can easily apply other sophisticated methods, such as simulated annealing, to the CRP. The graphical model for ISM is shown in Fig. 1(b). Table 1 provides a sample code of the proposed method.

In 2.3, we describe a method that uses the CRP to deal with an infinite number of speakers. In 2.4, we formulate in detail conditional posterior distributions $P(z_u|\mathbf{z}_{\backslash u}, \mathbf{O})$, which are required for Gibbs sampling, where $\mathbf{z}_{\backslash u}$ denotes a set of latent variables other than z_u.

2.3. Chinese restaurant process (CRP)

The CRP is given by taking the limit of S (i.e., the $S \to \infty$) in the following posterior probability in the FSM, where c_k denotes the number of utterances assigned to the k-th speaker cluster

$$\begin{aligned} P(z_u = k|\mathbf{z}_{1:u-1}) &= \int P(z_u = k|\mathbf{h}) p\left(\mathbf{h}|\mathbf{z}_{1:u-1}\right) d\mathbf{h} \\ &= \frac{c_k + h^{(0)}/S}{U - 1 + h^{(0)}}, \end{aligned} \tag{7}$$

In this case, $P(z_u = k|\mathbf{z}_{\backslash u})$ is computed separately for when the u-th utterance is assigned to a cluster having more than one utterance and for when the u-th utterance is assigned to a new cluster having no utterance (i.e., c_k = 0). Furthermore, joint distribution $P(z_1, z_2, \cdots, z_U) = P(z_1)P(z_2|z_1)\cdots P(z_U|\mathbf{z}_{1:U-1})$ is invariant even if z_u is exchanged on any u (i.e., exchangeability) [8]. From the aforementioned discussion, $P(z_u = k|\mathbf{z}_{\backslash u})$ is computed as follows:

$$P(z_u = k|\mathbf{z}_{\backslash u}) = \begin{cases} \frac{c_k}{U-1+h^{(0)}}, & \text{if } k = z_i \text{ for } \exists i \neq u \\ \frac{h^{(0)}}{U-1+h^{(0)}}, & \text{if } k \neq z_i \text{ for } \forall i \neq u \end{cases} \tag{8}$$

This equation means that the u-th utterance is assigned to either the existing k-th cluster having a probability of $c_k/(U-1+h^{(0)})$, or a new cluster having a probability of $h^{(0)}/(U-1+h^{(0)})$. In the latter case, the number of clusters increases. Since the number of speaker clusters increases with an increase in the observed data, the CRP can deal with the infinite number of speakers. In this case, $h^{(0)}$ denotes the concentration parameter, which is the degree for choosing a new cluster.

2.4. Conditional posterior distribution for the ISM

The posterior probability of $z_u = k$, given all utterances $\mathbf{O}$ and the other latent variables $\mathbf{z}_{\backslash u}$, is computed as follows:

$$P(z_u = k|\mathbf{z}_{\backslash u}, \mathbf{O}, \mathbf{\Theta}^{(0)}) \propto P(z_u = k|\mathbf{z}_{\backslash u}) \cdot p(\mathbf{O}_u|\mathbf{O}_{\backslash u}, z_u = k, \mathbf{\Theta}^{(0)}), \quad (9)$$

where $\mathbf{O}_{\backslash u} = \{\mathbf{O}_i : \forall i, i \neq u\}$, $\mathbf{O} = \{\mathbf{O}_u, \mathbf{O}_{\backslash u}\}$. The first term of the right side of Eq. (9) is obtained by the CRP. The second term is obtained as follows:

$$p(\mathbf{O}_u|\mathbf{O}_{\backslash u}, z_u = k, \mathbf{\Theta}^{(0)}) = \begin{cases} \int p(\mathbf{O}_u|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathbf{O}_{\backslash u}^{(k)}, \mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}_k, \\ \qquad \text{if } z_i = k \text{ for } \exists i \neq u \\ \int p(\mathbf{O}_u|\mathbf{\Theta}) G_0(\mathbf{\Theta}|\mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}, \\ \qquad \text{if } z_i \neq k \text{ for } \forall i \neq u \end{cases} \quad (10)$$

where $\mathbf{O}_{\backslash u}^{(k)} = \{\mathbf{O}_i : \forall i, i \neq u, z_i = k\}$. Thus, the conditional posterior probability of $z_u = k$ is computed as follows:

$$P(z_u = k|\mathbf{z}_{\backslash u}, \mathbf{O}, \mathbf{\Theta}^{(0)}) = \begin{cases} \frac{c_k}{U-1+h^{(0)}} \int p(\mathbf{O}_u|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathbf{O}_{\backslash u}^{(k)}, \mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}_k, \\ \qquad \text{if } z_i = k \text{ for } \exists i \neq u \\ \frac{h^{(0)}}{U-1+h^{(0)}} \int p(\mathbf{O}_u|\mathbf{\Theta}) G_0(\mathbf{\Theta}|\mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}, \\ \qquad \text{if } k \neq z_i \text{ for } \forall i \neq u \end{cases} \quad (11)$$

where $G_0(\mathbf{\Theta}|\mathbf{\Theta}^{(0)})$ denotes the base measure for the parameters. Equation (11) can be analytically integrated if the base measure is conjugate. In this case, we use the base measure as follows:

$$G_0(\mathbf{\Theta}|\mathbf{\Theta}^{(0)}) = \begin{cases} \boldsymbol{\mu} \sim \mathcal{N}\left(\cdot|\boldsymbol{\mu}^{(0)}, (\xi^{(0)})^{-1}\mathbf{\Sigma}\right) \\ (\sigma_d)^{-1} \sim \mathcal{G}\left(\cdot|\eta^{(0)}, \sigma_d^{(0)}\right) \end{cases} \quad (12)$$

where $\mathbf{\Sigma}$ denotes a diagonal covariance matrix of a Gaussian distribution whose (d, d)-th element is represented by σ_d and $\mathcal{G}\left(\cdot|\eta^{(0)}, \sigma_d^{(0)}\right)$ denotes a Gamma distribution with the hyperparameters $\eta^{(0)}$ and $\sigma_d^{(0)}$. Equation (11) can then be analytically integrated over $\mathbf{\Theta}$ as follows:

$$\begin{aligned} &\int p(\mathbf{O}_u|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathbf{O}_{\backslash u}^{(k)}, \mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}_k \\ &= \frac{\int p(\mathbf{O}_u, \mathbf{O}_{\backslash u}^{(k)}|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}_k}{\int p(\mathbf{O}_{\backslash u}^{(k)}|\mathbf{\Theta}_k) p(\mathbf{\Theta}_k|\mathbf{\Theta}^{(0)}) \mathrm{d}\mathbf{\Theta}_k} \\ &= \frac{\Psi_k\left(\mathbf{O}_u, \mathbf{O}_{\backslash u}^{(k)}\right)}{\Psi_k\left(\mathbf{O}_{\backslash u}^{(k)}\right)} \end{aligned} \quad (13)$$

where

$$\Psi_k\left(\mathbf{O}_u, \mathbf{O}_{\backslash u}^{(k)}\right) = (2\pi)^{-\frac{n_k D}{2}} \times \frac{(\xi^{(0)})^{\frac{D}{2}} (\Gamma(\frac{\eta^{(0)}}{2}))^{-D}}{(\tilde{\xi}_k)^{\frac{D}{2}} (\Gamma(\frac{\tilde{\eta}_k}{2}))^{-D}} \times \frac{(\prod_d \sigma_d^{(0)})^{\frac{\eta^{(0)}}{2}}}{(\prod_d \tilde{\sigma}_{k,d})^{\frac{\tilde{\eta}_k}{2}}} \quad (14)$$

The parameters of the conditional posterior distribution and the sufficient statistics are obtained as follows:

$$n_k = \sum_{u=1}^{U} T_u \delta_{z_u,k} \quad (15)$$

$$\tilde{\eta}_k = \eta^{(0)} + n_k \quad (16)$$

$$\tilde{\xi}_k = \xi^{(0)} + n_k \quad (17)$$

$$\mathbf{m}_k = \sum_{u=1}^{U} \delta_{z_u,k} \sum_{t=1}^{T_u} \mathbf{o}_{u,t} \quad (18)$$

$$r_{k,d} = \sum_{u=1}^{U} \delta_{z_u,k} \sum_{t=1}^{T_u} (o_{u,t,d})^2 \quad (19)$$

$$\tilde{\boldsymbol{\mu}}_k = \frac{\xi^{(0)}\boldsymbol{\mu}^{(0)} + \mathbf{m}_k}{\tilde{\xi}_k} \quad (20)$$

$$\tilde{\sigma}_{k,d} = \sigma_d^{(0)} + r_{k,d} + \xi^{(0)}(\mu_d^{(0)})^2 - \tilde{\xi}_k(\tilde{\mu}_{k,d})^2 \quad (21)$$

where $\delta_{i,j}$ denotes Kronecker's delta function, which is one if $i = j$ and zero if otherwise. From Eqs. (14) and (15)- (21), we can see that $\Psi_k\left(\mathbf{O}_u, \mathbf{O}_{\backslash u}^{(k)}\right)$ is obtained from the sufficient statistics of data $\{\mathbf{O}_u, \mathbf{O}_{\backslash u}^{(k)}\}$. Similarly, we can obtain $\Psi_k\left(\mathbf{O}_{\backslash u}^{(k)}\right)$ from $\mathbf{O}_{\backslash u}^{(k)}$ by calculating the right side of Eq. (14).

If the lower-right term of Eq. (11) is chosen, then we derive a new cluster from basis G_0, and the number of clusters increases. We can obtain the appropriate number of clusters and latent variables from their posterior distribution by iterating the sampling of latent variables in Eq. (11).

3. Speaker clustering experiments

To evaluate the effectiveness of the proposed method, we carried out speaker clustering experiments. In this paper, we compare the performance of the proposed method with that of the existing BIC based method [1] by using TIMIT database.

3.1. Experimental condition

3.1.1. Speech data

We used two evaluation sets. One set was the "core test set" in a TIMIT database. We call the evaluation using this set "Eval. 1." The other set was the "complete test set" that excludes the core test set in the TIMIT database. We call the evaluation using this set "Eval. 2." Evaluation 1 includes 192 utterances. These utterances were spoken by 24 speakers (16 male and 8 female speakers), and each speaker spoke 8 utterances. Evaluation 2 includes 1152 utterances. These utterances were spoken by 144 speakers (96 male and 48 female speakers) and each speaker spoke 8 utterances. Speech data were sampled at 16 kHz and quantized into 16-bit data.

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.

3.1.2. Measurement

We applied the average cluster purity (ACP), the average speaker purity (ASP), and their geometric mean (K value) to the evaluation criteria in speaker clustering [10]. In this case, the correct speaker label for each utterance was manually annotated. Let S_T be the correct number of speakers, S be the estimated number of speakers, n_{ij} be the estimated number of utterances assigned to speaker cluster i in all utterances of speaker j, n_j be the estimated number of utterances of speaker j, n_i be the estimated number of utterances assigned to speaker cluster i, and U be the number of all utterances. Cluster purity p_i and speaker purity q_j are then calculated as follows:

$$p_i = \sum_{j=0}^{S_T} \frac{n_{ij}^2}{n_i^2}, \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2} \quad (22)$$

The cluster purity is the ratio of the number of utterances derived from the same speaker to the number of utterances as-

Table 2: *Evaluation results. Here, α and $h^{(0)}$ denote a threshold/penalty parameter in the BIC and a logarithmic concentration parameter in the DPMM, respectively.*

evaluation data	method	parameter	#clusters	ACP	ASP	K value
Eval. 1 (#clusters: 24)	BIC	$\alpha = 1.3$	44	0.910	0.686	0.790
	BIC	$\alpha = 1.6$	31	0.864	0.764	0.812
	proposed	$h^{(0)} = -440$	34	0.898	0.751	**0.821**
Eval. 2 (#clusters: 144)	BIC	$\alpha = 1.3$	222	0.603	0.469	0.531
	BIC	$\alpha = 1.6$	117	0.330	0.594	0.442
	proposed	$h^{(0)} = -440$	186	0.577	0.492	**0.533**

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

$$V_{\mathrm{ACP}} = \frac{1}{U}\sum_{i=0}^{S} p_i n_i, \;\; V_{\mathrm{ASP}} = \frac{1}{U}\sum_{j=0}^{S_T} q_j n_j \tag{23}$$

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

$$K = \sqrt{V_{\mathrm{ACP}} \cdot V_{\mathrm{ASP}}} \tag{24}$$

The number of iterations was set to 200. We considered the first 180 iterations as the burn-in period, so the K values obtained from this period were rejected. The average of the K values from the remaining 20 iterations was measured. Furthermore, we carried out the same experiment 100 times but with different seeds for generating random numbers and then measured the average of their K values.

3.1.3. Conditions for speaker clustering

The terms $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ in Eqs. (20) and (21) were computed as the mean and covariance of all data in the database used. The terms $\xi^{(0)}$ and $\eta^{(0)}$ in Eqs. (16), (17), (20), and (21) were determined so that the best performance for Eval. 1 would be given. The initial number of clusters was set to one as in the preliminary experiment. To evaluate the robustness to the changes of data, we used multiple values for penalty parameter α in the existing BIC-based method [1] and concentration parameter $h^{(0)}$ in the proposed method for both Eval. 1 and Eval. 2.

3.2. Experimental results

Table 2 shows the ACPs, ASPs, and K values of the proposed and conventional methods. For the conventional BIC-based method, we show the result for $\alpha = 1.6$, where the best performance was given with Eval. 1, and the result for $\alpha = 1.3$, where the best performance was given with Eval. 2. For the proposed method, we only show the result for $h^{(0)} = -440$ for both Eval. 1 and Eval. 2.

The conventional method performed considerably worse for Eval. 2. The proposed method performed better even when the same hyper-parameter was used for both Eval. 1 and Eval. 2. Therefore, the proposed method is more robust to changes in the tuning parameters than the conventional method.

Now we will discuss computational costs. For Eval. 1, the conventional method took about 3,381 seconds on average in our environment (Intel core i7 Q720 1.60 GHz). In contrast, the proposed method took only 0.77 seconds per iteration and 156 seconds for 200 iterations. This demonstrates the effectiveness of the proposed method. The computational cost in the conventional method became exponentially larger with an increase in the amount of data because this method requires enormous amounts of BIC-value comparisons. For example, we need to evaluate the BIC value $n(n-1)(n-2)/6$ times until the number of clusters is unified at one if the amount of data is n. Thus, the conventional method is difficult to apply to tasks having large amounts of data. In contrast, the proposed method required less computation because it converged comparatively earlier, although sampling-based methods generally require many iterations until the value of the samples converges. This is attributed to the utterance-oriented samplings. The proposed method performed comparatively to the conventional method for Eval. 2 in the linear time of Eval. 1 (about six seconds). Therefore, the proposed method can be applied to large amounts of data.

4. Conclusion

In this paper, a speaker clustering method based on utterance-oriented DPMM was proposed. Speaker clustering experiments showed the effectiveness of the proposed method in terms of the computational cost and the robustness against tuning parameters. In this paper, we assumed a Gaussian distribution to be a speaker distribution. In [6], it was shown that the GMM was valid for the speaker distribution to express an inner utterance variety. Therefore, we will extend each speaker distribution to the GMM. In addition, we will apply the proposed method to speaker diarization tasks [9].

5. References

[1] S. S. Chen *et al.*, "Clustering via the Bayesian information criterion with applications in speech recognition," Proc. ICASSP, vol.2, pp.645–648, May 1998.

[2] J. Geiger *et al.*, "GMM-UBM based open-set online speaker diarization," Proc. Interspeech, pp.2330-2333, Sept. 2010.

[3] F. Valente, "Infinite models for speaker clustering," Proc. ICSLP, Sept. 2006

[4] E. B. Fox *et al.*, "The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states, " MIT LIDS, Cambridge, MA, Tech. Rep. P–2777, Nov. 2007.

[5] J. S. Liu, MonteCarlo Strategies in Scientific Computing, Springer, 2001.

[6] S. Watanabe *et al.*, "Gibbs sampling based multi-scale mixture model for speaker clustering," Proc. ICASSP, May 2011. (to appear)

[7] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," Ann. Statist., vol.1, no.2, pp.209–230, March 1973.

[8] D. Aldous, "Exchangeability and related topics," École dété de probabilités de Saint-Flour, XIII, pp.1–198, 1983.

[9] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans. Audio, Speech & Lang. Process., vol.14, no.5, pp.1557–1565, Sept. 2006.

[10] A. Solomonoff *et al.*, "Clustering speakers by their voices," Proc. ICASSP, vol.2, pp.757–760, May 1998.