ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Fully Bayesian speaker clustering based on hierarchically structured utterance-oriented Dirichlet process mixture model

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

[1]Waseda University, Tokyo, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

Abstract

We have proposed a novel speaker clustering method based on a hierarchically structured utterance-oriented Dirichlet process mixture model. In the proposed method, the number of speakers can be determined from the given data using a nonparametric Bayesian manner and intra-speaker variability is successfully handled by multi-scale mixture modeling. Experimental result showed that the proposed method is computationally-efficient and effective in speaker clustering. The proposed method significantly improve the accuracy of speaker clustering systems as compared with the conventional method, particularly for the case in which the number of utterances varied from speaker to speaker.

Index Terms Speaker clustering, nonparametric Bayesian model, Gibbs sampling, utterance-oriented Dirichlet process mixture model.

1. Introduction

Agglomerative hierarchical clustering based on a Bayesian information criterion (AHC-BIC) [1] is one of the most well-known strategies for speaker clustering when the number of speakers is unknown. This method, however, has several problems as follows: 1) The deterministic procedure in model estimation causes degradation in clustering accuracy for the case in which a large number of utterances are given. 2) Each speaker model was represented by a single Gaussian distribution. This limitation degrades the accuracy of the speaker clustering systems with respect to the speakers whose utterances have various expressions (speaking styles, contents of speech, etc.).

To solve the former problem, we proposed a fully Bayesian model called "utterance-oriented Dirichlet process mixture model (UO-DPMM) [2]." In the UO-DPMM, we modeled the entire speaker space as a DPMM and estimated its structure with a sampling-based procedure. The UO-DPMM can thus estimate the number of speakers and assign utterances to speakers, efficiently avoiding a local optima in a fully Bayesian manner. In the UO-DPMM, since the speaker distribution is modeled as a single Gaussian distribution, there remains weakness in modeling the speakers with large intra-speaker variability.

In the present paper, in order to solve the problems in AHC-BIC simultaneously, we extend the UO-DPMM such that each speaker distribution could be represented by a Gaussian mixture model (GMM). The proposed model can be interpreted as a nonparametric Bayesian version of the model called the "multi-scale mixture model" [3, 4]. In the proposed method, strict sampling cannot be achieved because the base measure for the parameter space is no longer defined in conjugate. In the present paper, we also provide an efficient computation for estimating the proposed nonparametric Bayesian model.

[1]Shinji Watanabe is now with Mitsubishi Electric Research Laboratories (MERL).

2. Finite speaker utterance-oriented generative model (FSM)

In this section, we define a finite speaker utterance-oriented generative model, in which the number of speaker clusters is fixed (**2.1** and **2.2**). Then, we demonstrate that speaker labels can be optimally assigned to utterances by stochastically estimating the structure of the proposed model (**2.3**).

2.1. Utterance-oriented generative model

Let $\mathbf{o}_{ut} \in \mathcal{R}^D$ be a D-dimensional observation vector at the t-th frame-wise observation 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 $\mathbf{O} \triangleq \{\mathbf{O}_u\}_{u=1}^{U}$ 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 a 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, we introduce two types of latent variables: $\mathbf{Z} = \{z_u\}_{u=1}^{U}$ 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; and $\mathbf{V} = \{v_{ut}\}_{u,t=1}^{U,T_u}$ represents the frame-level latent variables, each of which identifies the intra-speaker GMM component to which the t-th frame-wise observation in the u-th utterance is assigned. 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 $\{\{h_j\},\{w_{ij}\},\{\boldsymbol{\mu}_{ij}\},\{\mathbf{\Sigma}_{ij}\}\}$; h_j, the weight for the entire speaker MoGMM component; and 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 a Bayesian approach, the conjugate prior distributions of the parameters are introduced as follows:

$$\begin{aligned} P(h_i) &= \mathcal{D}(\mathbf{h}^0), \quad p(\boldsymbol{\mu}_{ij}) = \mathcal{N}(\boldsymbol{\mu}_j^0, \xi^0 \mathbf{\Sigma}_{ij}) \\ P(w_{ij}) &= \mathcal{D}(\mathbf{w}_j^0), \quad p(\sigma_{ij,d}) = \mathcal{G}(\eta^0, \sigma_{j,d}^0), \end{aligned} \quad (2)$$

where $\mathcal{D}\left(\mathbf{h}^0\right)$ denotes the Dirichlet distribution with a hyper parameter $\mathbf{h}^0 = \left\{h^0/S, \cdots, h^0/S\right\}$ and $\mathcal{G}\left(\eta^0, \sigma_{j,d}^0\right)$ denotes the

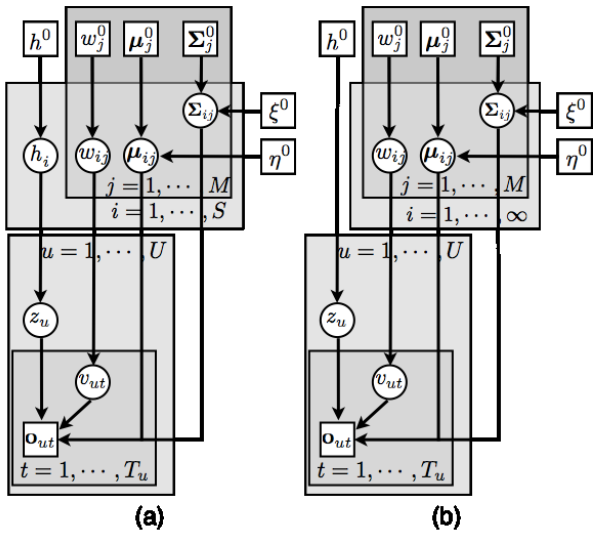

Figure 1: *Graphical representations of the (a) finite and (b) infinite speaker utterance-oriented generative models.*

Gamma distribution with hyper parameters η^0 and $\sigma^0_{j,d}$. Figure 1 (a) shows a graphical representation of the finite speaker utterance-oriented generative model.

2.2. Marginalized likelihood for the complete data

We derive the marginalized likelihood for the complete data. When the complete data, i.e., the pairs of observations and latent variables, are given, all assignments of utterances to speaker clusters as well as those of frame-wise observations to Gaussian components in each speaker's GMM are determined. Then, the sufficient statistics of this model are described as follows:

$$\begin{cases} n_i &= \sum_u \delta(z_u, i), \\ n_{ij} &= \sum_{u,t} \delta(z_u, i)\delta(v_{ut}, j), \\ \mathbf{m}_{ij} &= \sum_{u,t} \delta(z_u, i)\delta(v_{ut}, j) \cdot \mathbf{o}_{ut}, \\ r_{ij,d} &= \sum_{u,t} \delta(z_u, i)\delta(v_{ut}, j) \cdot (o_{ut,d})^2, \end{cases} \tag{3}$$

where $\delta(a, b)$ denotes the Kronecker's delta which is 1 if $a = b$ and 0 otherwise. n_i denotes the number of utterances that are assigned to the i-th component of the entire speaker MoGMM; n_{ij}, the number of frame-wise observations that are assigned to the j-th component of the intra-speaker GMM of the i-th component of the MoGMM; $\mathbf{m}_{ij}$ and $r_{ij,dd}$, the first and the second order sufficient statistics, respectively. On the basis of Eqs. 1 and 3, the likelihood for the complete data can be described as follows:

$$\begin{aligned} &p(\mathbf{O}, \mathbf{V}, \mathbf{Z}|\mathbf{\Theta}) \\ &= \prod_i (h_i)^{n_i} \prod_j (w_{ij})^{n_{ij}} \\ &\quad \prod_{u,t} \delta(z_u, i)\delta(v_{ut}, j)\mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_{ij}, \mathbf{\Sigma}_{ij}). \end{aligned} \tag{4}$$

From Eqs. 2 and 4, the marginalized likelihood for the complete data, is derived as follows:

$$\begin{aligned} &p(\mathbf{O}, \mathbf{V}, \mathbf{Z}|\mathbf{\Theta}^0) \\ &= \int p(\mathbf{Z}|\mathbf{h})p(\mathbf{h}|\mathbf{h}^0)\mathrm{d}\mathbf{h} \cdot \int p(\mathbf{V}|\mathbf{Z}, \mathbf{w})p(\mathbf{w}|\mathbf{w}^0)\mathrm{d}\mathbf{w} \\ &\quad \cdot \int p(\mathbf{O}|\mathbf{V}, \mathbf{Z}, \mathbf{\Theta})p(\mathbf{\Theta}|\mathbf{\Theta}^0)d\mathbf{\Theta} \\ &= p(\mathbf{Z}|\mathbf{h}^0)p(\mathbf{V}|\mathbf{Z}, \mathbf{w}^0)p(\mathbf{O}|\mathbf{V}, \mathbf{Z}, \mathbf{\Theta}^0), \end{aligned} \tag{5}$$

where $p(\mathbf{Z}|\mathbf{h}^0)$, $p(\mathbf{V}|\mathbf{Z}, \mathbf{w}^0)$ and $p(\mathbf{O}|\mathbf{V}, \mathbf{Z}, \mathbf{\Theta}^0)$ in Eq. 5, are described as follows:

$$p(\mathbf{Z}|\mathbf{h}^0) = \frac{\Gamma(h^0)\prod_i \Gamma(\tilde{h}_i)}{\Gamma(h^0/S)^S \Gamma(\sum_i \tilde{h}_i)}, \tag{6}$$

$$p(\mathbf{V}|\mathbf{Z}, \mathbf{w}^0) = \prod_i \frac{\Gamma(\sum_j w_j^0)\prod_j \Gamma(\tilde{w}_{ij})}{\prod_j \Gamma(w_j^0)\Gamma(\sum_j \tilde{w}_{ij})}, \tag{7}$$

$$\begin{aligned} &p(\mathbf{O}|\mathbf{V}, \mathbf{Z}, \mathbf{\Theta}^0) \\ &= \prod_{i,j} (2\pi)^{-\frac{n_{ij}D}{2}} \frac{(\xi^0)^{\frac{D}{2}} \left(\Gamma\left(\frac{\eta_j^0}{2}\right)\right)^{-D} \left(\prod_d \sigma_{j,dd}^0\right)^{\frac{\eta_j^0}{2}}}{(\tilde{\xi}_{ij})^{\frac{D}{2}} \left(\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right)\right)^{-D} \left(\prod_d \tilde{\sigma}_{ij,dd}\right)^{\frac{\tilde{\eta}_{ij}}{2}}}. \end{aligned} \tag{8}$$

The parameters $\tilde{h}_i, \tilde{\eta}_{i,dd}, \tilde{\xi}_{i,dd}, \tilde{\boldsymbol{\mu}}_i$, ,and $\tilde{\sigma}_{i,dd}$ in Eqs. 6, 7 and 8 denote the hyper-parameters of the posterior distribution for $\mathbf{\Theta}$, which are described as follows:

$$\begin{cases} \tilde{h}_i &= \frac{h^0}{S} + n_i, \\ \tilde{w}_{ij} &= w_j^0 + n_{ij}, \\ \tilde{\xi}_{ij} &= \xi^0 + n_{ij}, \\ \tilde{\eta}_{ij} &= \eta^0 + n_{ij}, \\ \tilde{\boldsymbol{\mu}}_{ij} &= \frac{\xi^0 \boldsymbol{\mu}_j^0 + \mathbf{m}_{ij}}{\tilde{\xi}_{ij}}, \\ \tilde{\sigma}_{ij,d} &= \sigma_{j,d}^0 + r_{ij,d} + \xi^0(\mu_{j,d}^0)^2 + \tilde{\xi}_{ij}(\tilde{\mu}_{ij,d})^2, \end{cases} \tag{9}$$

where we used Eq. 3 to obtain Eq. 9.

2.3. Model estimation based on sampling approach

The speaker clustering problem attempts to determine assignments of every utterances to speaker clusters; hence, this problem reduces to the estimation of the optimal utterance-level latent variables $\mathbf{Z}$. The optimal estimates of $\mathbf{Z}$ are obtained such that the marginalized posterior distribution $P(\mathbf{Z}|\mathbf{O}) = \sum_{\mathbf{V}} \int P(\mathbf{Z}, \mathbf{V}|\mathbf{O})\mathrm{d}\mathbf{\Theta}$ would be maximized. However, the strict evaluation of the marginalized posterior distribution for all the possible combinations of the utterance-level latent variables is computationally infeasible. In this study, therefore, we use the Gibbs sampling [5] to obtain the utterance-level latent variables $\mathbf{Z}$ directly from their marginalized posterior distribution $P(\mathbf{Z}|\mathbf{O})$. In each step of the 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., $\mathbf{Z}_{\backslash u} = \{z_{u'}|u' \neq u\}$). This means that each utterance-level latent variable is individually sampled from its conditional posterior distribution as follows:

$$\begin{aligned} &p(z_u = i'|\mathbf{O}, \mathbf{Z}_{\backslash u}) \\ &\propto P(z_u = i'|\mathbf{Z}_{\backslash u}) \sum_{\mathbf{V}} p(\mathbf{O}_u, \mathbf{V}|\mathbf{O}_{\backslash u}, \mathbf{Z}_{\backslash u}, z_u = i') \\ &= \frac{P(\mathbf{Z}_{\backslash u}, z_u = i')}{P(\mathbf{Z}_{\backslash u})} \sum_{\mathbf{V}} \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{Z}_{\backslash u}, z_u = i')}{p(\mathbf{O}_{\backslash u}, \mathbf{V}_{\backslash u}|\mathbf{Z}_{\backslash u})}. \end{aligned} \tag{10}$$

Here, we omitted the hyper-parameters for the prior distributions, $\{\mathbf{h}^0, \mathbf{\Theta}^0\}$, in Eq. 10 in order to keep the notation uncluttered. Using the property of Gamma function that $\Gamma(n+1)/\Gamma(n) = 1$ and the fact that the difference between $\tilde{h}_i$ given $\mathbf{Z}$ and $\tilde{h}_i$ given $\mathbf{Z}_{\backslash u}$ is 1, we can the first term on the right side of Eq. 10 as follows:

$$\frac{P(\mathbf{Z}_{\backslash u}, z_u = i')}{P(\mathbf{Z}_{\backslash u})} = \frac{h^0/S + n_{i'}}{U - 1 + h^0}. \tag{11}$$

Algorithm 1 Algorithm of the proposed method.

Initialize S, $\{z_u, v_{ut} : u = 1, \cdots, U, t = 1, \cdots, T_u\}$.
repeat
  for all u such that $1 \leq u \leq U$ **do**
    for all t such that $1 \leq t \leq T_u$ **do**
      Sample v_{ut} from Eq. 12
    end for
  end for
  for all u such that $1 \leq u \leq U$ **do**
    Sample z_u from Eq. 16
    if $z_u = S + 1$ **then**
      $\boldsymbol{\Theta}_{S+1} \sim G_0\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0\right)$
      $S \leftarrow S + 1$
    end if
  end for
until some condition is met

The evaluation of the second term on the right side of Eq. 10 needs the summation with respect to the frame-level latent variables $\mathbf{V}$. This summation is usually infeasible and thus is approximated using the sampled values that are obtained from the posterior of the frame-level latent variables as $\mathbf{V}^* \sim P(\mathbf{V}|\mathbf{O}, \mathbf{Z}_{\backslash u}, z_u = i')$ with the Gibbs sampler of $\mathbf{V}$ as

$$
\begin{aligned}
&P(v_{ut}^* = j'|\mathbf{O}, \mathbf{V}_{\backslash t}, \mathbf{Z}_{\backslash u}, z_u = i) \\
&\quad \propto \frac{P(v_{ut} = j', \mathbf{V}_{\backslash t})}{P(\mathbf{V}_{\backslash t})} \cdot \frac{p(\mathbf{O}|\mathbf{V}, v_{ut} = j', \mathbf{Z}_{u\backslash t}, z_u = i')}{p(\mathbf{O}_{u\backslash t}|\mathbf{V}_{u\backslash t}, \mathbf{Z}_{u\backslash t})} \\
&\quad = \frac{\tilde{w}_{i'j'}}{T_u - 1 + w_{j'}^0} \cdot \exp\left(g_{i'j'}(\tilde{\boldsymbol{\Theta}}_{i',j'}) - g_{i'j'}(\tilde{\boldsymbol{\Theta}}_{i',j'\backslash t})\right), \quad (12)
\end{aligned}
$$

where $g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j})$ is defined as follows:

$$
\begin{aligned}
\ln g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j}) \;=\; & \log\Gamma(\tilde{w}_{ij}) + D\log\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right) \\
& - \frac{D}{2}\log\tilde{\xi}_{ij} - \frac{\tilde{\eta}_{ij}}{2}\sum\nolimits_d \log\tilde{\sigma}_{ij,d}. \quad (13)
\end{aligned}
$$

3. Infinite speaker utterance-oriented generative model (ISM)

We attempt to extend the FSM described in the previous section to the infinite speaker utterance-oriented generative model (ISM), which can estimate the number of speakers.

The ISM is derived by taking the limit of S (i.e., $S \to \infty$) in Eq. 11. In this case, we separately compute Eq. 11 for the case in which the u-th utterance is assigned to an existing cluster with one or more than one utterances (i.e., $n_{i'} > 0$) and the case in which the u-th utterance is assigned to a new cluster with no utterances (i.e., $n_{i'} = 0$). Taking the limit $S \to \infty$, the number of utterances U satisfies $U \ll S$. In this case, most of the clusters are empty because there are at most U speaker clusters to which at least one utterance is assigned. Therefore, we lump the empty clusters together and consider the limit $S \to \infty$. Eq. 11, then, can be described as follows:

$$
\begin{aligned}
&\frac{P(\mathbf{Z}_{\backslash u}, z_u = i')}{P(\mathbf{Z}_{\backslash u})} \\
&\quad = \begin{cases} \frac{n_{i'}}{U-1+h^0}, & \text{if } z_k = i' \text{ for } \exists k \neq u, \\ \frac{h^0}{U-1+h^0}, & \text{if } z_k \neq i' \text{ for } \forall k \neq u. \end{cases} \quad (14)
\end{aligned}
$$

In this case, the right side of Eq. 10 can be also separately described as follows:

$$
\begin{aligned}
&\sum_{\mathbf{V}} \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{Z}_{\backslash u}, z_u = i')}{p(\mathbf{O}_{\backslash u}, \mathbf{V}_{\backslash u}|\mathbf{Z}_{\backslash u})} \\
&\quad = \sum_{\mathbf{V}} \frac{p(\mathbf{O}|\mathbf{V}, \mathbf{Z}_{\backslash u}, z_u = i')p(\mathbf{V}|\mathbf{O}, \mathbf{Z}_{\backslash u}, z_u = i')}{p(\mathbf{O}_{\backslash u}|\mathbf{V}_{\backslash u}, \mathbf{Z}_{\backslash u})p(\mathbf{V}_{\backslash u}|\mathbf{O}_{\backslash u}, \mathbf{Z}_{\backslash u})} \\
&\quad \approx \frac{p(\mathbf{O}|\mathbf{V}^*, \mathbf{Z}_{\backslash u}, z_u = i')p(\mathbf{V}^*|\mathbf{O}, \mathbf{Z}_{\backslash u}, z_u = i')}{p(\mathbf{O}_{\backslash u}|\mathbf{V}^*_{\backslash u}, \mathbf{Z}_{\backslash u})p(\mathbf{V}^*_{\backslash u}|\mathbf{O}_{\backslash u}, \mathbf{Z}_{\backslash u})} \\
&\quad \propto \begin{cases} \exp\Big(\log\frac{\Gamma(\sum_j \tilde{w}_{i'\backslash u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} + \\ \qquad \sum_j \left(g_{ij}(\tilde{\boldsymbol{\Theta}}_{i',j}) - g_{ij}(\tilde{\boldsymbol{\Theta}}_{i'\backslash u,j})\right)\Big), \\ \qquad \text{if } z_k = i' \text{ for } \exists k \neq u, \\ \int \sum_{\mathbf{V}} p(\mathbf{O}_u, \mathbf{V}|\boldsymbol{\Theta}) G_0(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0)\mathrm{d}\boldsymbol{\Theta}, \\ \qquad \text{if } z_k \neq i' \text{ for } \forall k \neq u, \end{cases} \quad (15)
\end{aligned}
$$

where $\mathbf{V}^*$ are the sampled values of the frame-level latent variables that are obtained from Eq. 12. The lower term on the right side of Eq. 15 describes the conditional likelihood for the u-th utterance to be assigned to a new empty cluster. This term is evaluated as the marginalized likelihood over base measure on measurable space $G_0(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0)$. Although we should marginalize the frame-level latent variables $\mathbf{V}$ besides the model parameters, $\boldsymbol{\Theta}$, this marginalization is infeasible because this requires significantly high computational cost that exponentially grows with an increase in the number of data. Therefore, we attempt to approximately use current values of the frame-level latent variables instead of strictly marginalizing these variables.

On the basis of Eqs. 14 and 15, we can finally describe, the posterior probability of the utterance-level latent variables in the case of the ISM as follows:

$$
\begin{aligned}
&p(z_u = i'|\mathbf{O}, \mathbf{Z}_{\backslash u}) \\
&\quad = \begin{cases} \frac{n_{i'}}{U-1+h^0} \cdot \exp\Big(\log\frac{\Gamma(\sum_j \tilde{w}_{i'\backslash u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} + \\ \qquad \sum_j \left(g_{ij}(\tilde{\boldsymbol{\Theta}}_{i',j}) - g_{ij}(\tilde{\boldsymbol{\Theta}}_{i'\backslash u,j})\right)\Big), \\ \qquad \text{if } z_k = i' \text{ for } \exists k \neq u, \\ \frac{h^0}{U-1+h^0} \int \sum_{\mathbf{V}} p(\mathbf{O}_u, \mathbf{V}|\boldsymbol{\Theta}) G_0(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0)\mathrm{d}\boldsymbol{\Theta}, \\ \qquad \text{if } z_k \neq i' \text{ for } \forall k \neq u. \end{cases} \quad (16)
\end{aligned}
$$

If the lower term on the right side of Eq. 16 is chosen, then we derive a new cluster from the basis G_0 and consequently, the number of clusters increases. Eq. 16 indicates that the ISM corresponds to an application of Dirichlet process mixture model (DPMM) [6] implemented by Chinese restaurant process (CRP) [7] to the FSM.

The graphical representation of ISM is shown in Fig. 1(b). Algorithm 1 provides a sample code of the proposed method.

4. Speaker clustering experiments

We compared the speaker clustering performance of the proposed method with that of the conventional AHC-BIC [1] and UO-DPMM [2] using the TIMIT and the corpus of spontaneous Japanese (CSJ) databases.

4.1. Experimental condition

4.1.1. Speech data

We used five evaluation sets that were obtained from the TIMIT and the CSJ databases. We used two evaluation sets in TIMIT. One set, T-1, was the "core test set," including 192 utterances spoken by 24 speakers. The other set, T-2, was the "complete test set," that did not include the core test set in the TIMIT database. T-2 includes 1,152 utterances spoken by 144 speakers. The remaining three evaluation sets were obtained from

Table 1: Details of test set. # of spkr., # of utt.,# of samp., and total dur. denote the number of speakers, that of utterances, that of frame-wise observations, and the total duration, respectively.

	T-1	T-2	C-1	C-2	C-3
# of spkr.	24	144	10	20	30
# of utt.	192	1,152	786	4,642	1,983
(# of samp.)	(5.8 K)	(353 K)	(515 K)	(3.1 M)	(1.3 M)
total dur.	9.7 [m]	59 [m]	1.4 [h]	8.6 [h]	3.6 [h]

CSJ as follows: all lectures were divided into utterance units on the basis of the silence segments in their transcriptions; these segments were longer than 500 ms. Moreover, 10, 20, and 30 speakers were randomly selected and all of their utterances were selected. We called these sets C-1, C-2, and C-3, respectively. Each utterance has a duration of 5-10 s. Table 1 lists the number of speakers and utterances. The speech data were sampled at 16 kHz and quantized into 16-bit data.

We used 12-dimensional mel-frequency cepstral coefficients (MFCCs) as the feature parameters. The frame length and the frame shift were 25 ms and 10 ms, respectively.

4.1.2. 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 [8]. The number of iterations was set to 50 in the proposed method. We considered the first 49 iterations as the burn-in period; hence, the K values obtained from this period were rejected. Furthermore, we performed the same experiment 50 times with different seeds for generating random numbers and then measured the average of their K values.

4.1.3. Conditions for speaker clustering

The hyper-parameters in Eq. (2) were set as follows: $\eta^0 = 1$ and $\xi^0 = 1$; $\left\{\mathbf{w}, \boldsymbol{\mu}_j^0, \boldsymbol{\Sigma}_j^0\right\}_{j=1}^M$ were set to the weight, mean and covariance matrix in the GMM trained with 28,171 sentences that are not included in the TIMIT and CSJ databases. The remaining parameters, the parameter h^0 in the proposed method and the penalty parameter α in the AHC-BIC method, were critical for estimation of the number of speakers. Hence, in order to evaluate the robustness to changes in data, we separately decided these values for T-* and C-* as follows: we tune the parameter for one dataset and evaluated the remaining datasets using these values. We performed this evaluation for all sets and calculated their average results. The initial number of clusters was set to one in the preliminary experiment. The number of mixture in each speaker GMM was set to 2.

4.2. Experimental results

Table 2 shows the average of the estimated number of clusters (#cl.), ACPs, ASPs and K values for the three speaker clustering methods. The proposed method outperformed both the AHC-BIC and the UO-DPMM for all datasets. In particular, the proposed method significantly improved accuracy of both the AHC-BIC and UO-DPMM for use in the T-2 dataset. This result could be attributed to the relatively large variations in each speaker's utterances in the T-2 dataset. This result shows that the proposed method could represent these variations by using a GMM for modeling each speaker, as compared with the conventional methods in which each speaker was represented by a single Gaussian distribution. The proposed method also achieved remarkably higher accuracy than that of the AHC-BIC-based method for use in the relatively large-scale C-2 dataset. This result showed that the proposed method was effective in the case of a large-scale data, while the AHC-BIC caused significant degradation in clustering accuracy for these data.

Table 2: Speaker clustering accuracy for the proposed method and conventional methods such as the UO-DPMM-based method [2] and AHC-BIC-based method [1].

Eval.	Method	#cl.	ACP	ASP	K
T-1	proposed	38.4	0.92	0.73	**0.82**
	UO-DPMM	32.4	0.84	0.72	0.78
(spkr: 24)	AHC-BIC	34.0	0.85	0.71	0.78
T-2	proposed	169.4	0.71	0.66	**0.68**
	UO-DPMM	145.0	0.53	0.55	0.54
(spkr: 144)	AHC-BIC	174.0	0.54	0.49	0.52
C-1	proposed	10.2	0.75	0.90	**0.82**
	UO-DPMM	10.6	0.68	0.88	0.77
(spkr: 10)	AHC-BIC	18.0	0.98	0.53	0.72
C-2	proposed	36.3	0.88	0.63	**0.74**
	UO-DPMM	18.2	0.69	0.74	0.71
(spkr: 20)	AHC-BIC	15.0	0.08	0.75	0.24
C-3	proposed	41.9	0.85	0.73	**0.79**
	UO-DPMM	31.7	0.76	0.69	0.72
(spkr: 30)	AHC-BIC	54.0	0.91	0.49	0.67

Finally, we will discuss the computational cost. In the experiment for use in the C-2 dataset (i.e., 20 speakers and 4,642 utterances), the proposed approach took only approximately 53.0 s in average for one epoch of iterative calculation. Therefore, the proposed method is substantially practical for speaker clustering tasks.

5. Conclusion

In this paper, we proposed a novel speaker clustering method based on hierarchically structured utterance-oriented DPMM. The speaker clustering experiments showed that the proposed method outperformed the conventional methods for all datasets. In particular, the proposed method helped achieve significantly higher accuracy than the conventional methods, for use in datasets that included large intra-speaker variability. In this study, the number of components in each speaker's GMM was fixed. Therefore, we are going to extend the proposed model such that the number of the Gaussian components could be determined from the given data.

6. References

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

[2] N. Tawara *et al.*, "Speaker Clustering Based on Utterance-oriented Dirichlet Process Mixture Model," Proc. Interspeech2011, pp2905-2908, Aug., 2011.

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

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

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

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

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

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