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

A sampling-based speaker clustering using utterance-oriented Dirichlet process mixture model and its evaluation on large scale data

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

TR2015-153 October 2015

Abstract

An infinite mixture model is applied to model-based speaker clustering with sampling-based optimization to make it possible to estimate the number of speakers. For this purpose, a framework of nonparametric Bayesian modeling is implemented with the Markov chain Monte Carlo (MCMC) and incorporated in the utterance-oriented speaker model. The proposed model is called the utterance-oriented Dirichlet process mixture model (UO-DPMM). The present paper demonstrates that UO-DPMM is successfully applied on large-scale data and outperforms the conventional hierarchical agglomerative clustering, especially for large amounts of utterances.

APSIPA Transactions on Signal and Information Processing

A sampling-based speaker clustering using utterance-oriented Dirichlet process mixture model and its evaluation on large scale data.

Naohiro Tawara *, Tetsuji Ogawa *, Shinji Watanabe †, Atsushi Nakamura ‡ and Tetsunori Kobayashi *,
*Waseda University, †Mitsubishi Electric Research Laboratories, MA, US ‡Nagoya City University.

***Abstract*—An infinite mixture model is applied to model-based speaker clustering with sampling-based optimization to make it possible to estimate the number of speakers. For this purpose, a framework of nonparametric Bayesian modeling is implemented with the Markov chain Monte Carlo (MCMC) and incorporated in the utterance-oriented speaker model. The proposed model is called the utterance-oriented Dirichlet process mixture model (UO-DPMM). The present paper demonstrates that UO-DPMM is successfully applied on large-scale data and outperforms the conventional hierarchical agglomerative clustering, especially for large amounts of utterances.**

I. Introduction

Speaker clustering is the challenge of grouping the utterances spoken by the same speaker into a cluster. Hierarchical agglomerative clustering (HAC) is one of the best-known strategies for speaker clustering when the number of speakers needs to be estimated. In this framework, the utterances are clustered by progressively merging the most similar pair of clusters on the basis of criteria such as the Bayesian information criterion [1]. This method, however, can diminish clustering accuracy if an inappropriate pair of clusters is merged. This is considered a local solution problem caused by the lack of a procedure for dividing the merged cluster. This problem becomes more serious when the number of speakers is large due to increasing improper merging of the clusters.

An alternative approach to agglomerative clustering is partitional clustering, which directly divides the utterances into homogeneous k clusters. Partitional clustering can yield the advantage of avoiding of the local optimum problem caused at the merging steps in the agglomerative clustering framework. The model-based methods such as k-means clustering and Gaussian mixture model (GMM) are popular in partitional clustering and adopt the generative model in which the utterances spoken by a speaker are expected to be generated from a distribution expressing the speaker. In this approach, the speaker clustering is reduced to estimation of this generative model. The model-based clustering, however, can suffer from the over-learning problem especially when the amount of data is limited, and also be trapped into a local optimum solution when deterministic algorithms are used for estimation.

Sampling-based optimization such as Markov chain Monte Carlo (MCMC) has been shown to effectively address the problems in the model-based approach. We therefore proposed the utterance-oriented speaker mixture model [2], [3] and the MCMC-based sampling techniques to estimate this model [4]. This model is demonstrated to be accurate and efficient in speaker clustering but needs a technique to estimate the number of speakers.

We attempt to develop a model-based technique able to estimate the number of speakers by employing a nonparametric Bayesian framework [5]. Here, we derive the utterance-oriented speaker mixture model for infinite speakers by simply taking the limit of the formula of the finite speaker mixture model as the number of speakers approaches infinity. We call this model the utterance-oriented Dirichlet process mixture model (UO-DPMM). We preliminarily confirmed that UO-DPMM performed well in limited conditions where the number of utterances is small and balanced for each speaker, e.g., only eight utterances per speaker and a total of 1,192 utterances spoken by 192 speakers [6]. The present study therefore demonstrates that UO-DPMM can cope with practically large-scale data including a total of 15,435 utterances (i.e., over ten times the size of the data we used previously [6]) in a realistic computational time.

The remainder of the present paper is organized as follows. In section II, we define the utterance-oriented mixture model for finite speakers, in which the number of speakers is fixed. In section III, we extend the finite speaker mixture model described in section II to the nonparametric Bayesian model, namely UO-DPMM. We also describe the model estimation algorithm of UO-DPMM in detail. In section IV, the speaker clustering experiment used to verify the effectiveness of the proposed method is presented. In section V, we clarify the difference between UO-DPMM and the conventional nonparametric Bayesian method. In section VI, the paper is concluded, and future works are suggested.

II. Utterance-oriented mixture model for finite speakers

In this section, we define an utterance-oriented mixture model to represent all speakers. In the present study, we focus on using a Gaussian distribution to represent each speaker's cluster. Applying a single Gaussian distribution limits the flexibility of the model and actually a Gaussian mixture model (GMM) has been used in the existing approaches [7], [8], [2], [3]. This simple model, however, can be easily extend in a non-parametric Bayesian manner in order to handle infinite speakers (i.e., the optimal number of speakers is automatically determined). The aim of the present study thus is to investigate

the potential of the utterance-oriented mixture model in a non-parametric Bayesian manner.

First, we derive the utterance-oriented mixture model when the number of speaker clusters is fixed. Here, we describe how to estimate the utterance-oriented speaker model for the finite speakers and how to assign speaker labels to each utterance by using this model.

A. Utterance-oriented mixture model

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 the T_u observation vectors, and $\mathbf{O} \stackrel{\triangle}{=} \{\mathbf{O}_u\}_{u=1}^{U}$ 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 the variability for all speakers is modeled 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 $\mathbf{o}_{ut}$ is generated as an i.i.d. from a mixture component to which the utterance is assigned. We call this model "utterance-oriented mixture model." $\mathbf{Z} \stackrel{\triangle}{=} \{z_u\}_{u=1}^{U}$, represents the indices of speaker clusters. In this utterance-oriented mixture model, the likelihood for the set of observation vectors given the sequence of the latent variables is expressed as follows [1]:

$$p(\mathbf{O}|\mathbf{Z},\mathbf{\Theta}) = \prod_{u=1}^{U}\prod_{i=1}^{S}\prod_{t=1}^{T_u}\mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_i,\mathbf{\Sigma}_i)^{\delta(z_u,i)}, \quad (1)$$

$$P(\mathbf{Z}|\mathbf{h}) = \prod_{u=1}^{U}\prod_{i=1}^{S} h_i^{\delta(z_u,i)}, \quad (2)$$

where $\delta(a,b)$ denotes the Kronecker delta, which is 1 if $a=b$ and 0 otherwise. $\mathbf{h} = \{h_i\}_{i=1}^{S}$ and $\mathbf{\Theta} = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}_{i=1}^{S}$ denote the set of weights, mean vector, and covariance matrix for S speaker clusters, respectively. $\mathbf{\Sigma}_i$ is a diagonal covariance matrix whose (d,d)-th element is represented by $\sigma_{i,dd}$.

Since z_u denotes the index of a speaker cluster to which he u-th utterance is assigned, the speaker clustering problem is reduced to the estimation of the optimal values of the latent variables $\mathbf{Z}$. In other words, we can obtain the optimal assignment of utterances to speaker clusters by estimating $\mathbf{Z}$ which maximizes the likelihood function defined in Eqs. 1 and 2. This can be easily obtained by introducing expectation maximization (EM) algorithm [9].

B. Fully Bayesian approach for utterance-oriented mixture model

The maximum likelihood-based approach described in the previous subsection often suffers from an over-learning problem, especially when the amount of data is limited [10]. In order to solve this problem, we introduce a fully Bayesian approach to our utterance-oriented mixture model.

To derive the Bayesian representation, we introduce the following conjugate prior distributions of the model parameters $\mathbf{\Theta}$:

$$p(\mathbf{\Theta},\mathbf{h}) = \begin{cases} \{h_i\}_{i=1}^{S} & \sim \mathcal{D}(\mathbf{h}^0) \\ \{\boldsymbol{\mu}_i,\mathbf{\Sigma}_i\}_{i=1}^{S} & \sim \prod_d \mathcal{N}\mathcal{G}(\mu_d^0,\xi^0,\eta^0,\sigma_{dd}^0),\forall i \end{cases} \quad (3)$$

where $\mathcal{D}\left(\mathbf{h}^0\right)$ denotes the Dirichlet distribution with a hyper parameter $\mathbf{h}^0 = \{h_0/S,\cdots,h_0/S\}$ and $\mathcal{N}\mathcal{G}\left(\mu_d^0,\xi^0,\eta^0,\sigma_{dd}^0\right)$ denotes the Gaussian-Gamma distribution with hyper parameters μ_d^0, ξ^0, η^0, and σ_{dd}^0. Note that these hyper-parameters do not depend on each cluster [1]. The graphical model for this model is shown in Fig. 1 (a). By using these prior distributions, we can derive the joint distribution for the complete data case.

1) Marginalized likelihood for the complete data case: For the complete data case, the posterior probabilities of the latent variables, $P(\mathbf{Z}|\mathbf{O})$, return 0 or 1 because all assignments of utterances to speaker clusters are available. Then, the sufficient statistics of this model can be described as follows:

$$\begin{cases} n_i^{\mathrm{utt}} & = \sum_u \delta(z_u,i) \\ n_i^{\mathrm{frm}} & = \sum_u \delta(z_u,i) T_u \\ \mathbf{m}_i & = \sum_u \delta(z_u,i) \sum_t \mathbf{o}_{ut} \\ r_{i,dd} & = \sum_u \delta(z_u,i) \sum_t (o_{ut,d})^2 \end{cases} \quad (4)$$

where n_i^{utt} and n_i^{frm} are the number of utterances and that of frames assigned to the i-th cluster, respectively; $\mathbf{m}_i$ and $r_{i,dd}$ are the first- and second-order sufficient statistics, respectively. By using Eqs. 2 and 4, the likelihood for the complete data case can be expressed as follows:

$$p(\mathbf{O},\mathbf{Z}|\mathbf{\Theta},\mathbf{h}) = \prod_i (h_i)^{n_i^{\mathrm{utt}}} \prod_{u,t} \mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_i,\mathbf{\Sigma}_i)^{\delta(z_u,i)}. \quad (5)$$

Here, recalling that the speaker clustering problem aims to estimate the optimal assignment of utterances to speaker clusters, we can see that the parameter $\mathbf{\Theta}$ need not be estimated. We can therefore marginalize this parameter out from the joint distribution described in Eq. 5. This marginalization allows us to optimize the model on the latent variable space. By restricting the search space of the latent variables, we can obtain a model estimation algorithm that is robust against the local optima problem.

From Eqs. 3 and 5, the marginalized likelihood for the complete data case, integrated by using the parameter $\mathbf{\Theta}$, can be factorized to the following two integrals:

$$\begin{aligned} p(\mathbf{O},\mathbf{Z}) &= \int p(\mathbf{O},\mathbf{Z}|\mathbf{\Theta},\mathbf{h}) \cdot p(\mathbf{\Theta},\mathbf{h}) \mathrm{d}\mathbf{\Theta}\mathrm{d}\mathbf{h} \\ &= \int P(\mathbf{Z}|\mathbf{h}) p(\mathbf{h}) \mathrm{d}\mathbf{h} \cdot \int p(\mathbf{O}|\mathbf{Z},\mathbf{\Theta}) p(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta}. \end{aligned} \quad (6)$$

The first term on the right-hand side of Eq. 6 is described as follows:

$$\int P(\mathbf{Z}|\mathbf{h}) p(\mathbf{h}) \mathrm{d}\mathbf{h} = C(\mathbf{h}^0) \frac{\prod_i \Gamma(\tilde{h}_i)}{\Gamma(\sum_i \tilde{h}_i)}, \quad (7)$$

where $C(\mathbf{h}^0)$ denotes the normalization term that is independent of n_i^{utt}. The second term on the right-hand side of Eq. 6

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

[1] The detailed definition of Dirichlet and Gaussian-Gamma distributions is described in Appendix A.

is described as follows:

$$\begin{aligned}
&\int p(\mathbf{O}|\mathbf{Z},\mathbf{\Theta})p(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \\
&= \prod_i (2\pi)^{-\frac{n_i^{\mathrm{frm}} D}{2}} \frac{(\xi^0)^{\frac{D}{2}} \left(\Gamma\left(\frac{\eta_i^0}{2}\right)\right)^{-D} \left(\prod_d \sigma_{i,dd}^0\right)^{\frac{\eta_i^0}{2}}}{(\tilde{\xi}_i)^{\frac{D}{2}} \left(\Gamma\left(\frac{\tilde{\eta}_i}{2}\right)\right)^{-D} \left(\prod_d \tilde{\sigma}_{i,dd}\right)^{\frac{\tilde{\eta}_i}{2}}} \\
&= \prod_i \frac{Z(\tilde{\xi}_i, \tilde{\eta}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{\Sigma}}_i)}{Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0)} (2\pi)^{-\frac{n_i D}{2}},
\end{aligned} \tag{8}$$

where $\tilde{\mathbf{\Theta}} \triangleq \left\{\tilde{h}_i, \tilde{\eta}_{i,dd}, \tilde{\xi}_{i,dd}, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_{i,dd}\right\}$ denotes the hyper-parameter of the posterior distribution for $\mathbf{\Theta}$, which is described as follows:

$$\begin{cases}
\tilde{h}_i & = h_i^0 + n_i^{\mathrm{utt}} \\
\tilde{\xi}_i & = \xi^0 + n_i^{\mathrm{frm}} \\
\tilde{\eta}_i & = \eta^0 + n_i^{\mathrm{frm}} \\
\tilde{\boldsymbol{\mu}}_i & = \frac{\xi^0 \boldsymbol{\mu}_i^0 + \mathbf{m}_i}{\tilde{\xi}_i} \\
\tilde{\sigma}_{i,dd} & = \sigma_{i,dd}^0 + r_{i,dd} + \xi^0 (\mu_{i,d}^0)^2 - \tilde{\xi}_i (\tilde{\mu}_{i,d})^2
\end{cases} \tag{9}$$

where we have used Eq. 4.

2) MCMC-based posterior estimation: We again emphasize that the speaker clustering problem is reduced to the estimation of the latent variables $\mathbf{Z}$, which maximize the posterior distribution $P(\mathbf{Z}|\mathbf{O})$. We can then derive the posterior distribution for the latent variables as $p(\mathbf{Z}|\mathbf{O}) \propto p(\mathbf{O}, \mathbf{Z})$. The evaluation of all combinations of these latent variables in $p(\mathbf{Z}|\mathbf{O})$, however, is obviously infeasible if the number of utterances (i.e. the number of latent variables) is large. Instead, we use collapsed Gibbs sampling [11] to obtain the optimal value of $\mathbf{Z}$ directly from its posterior distribution $P(\mathbf{Z}|\mathbf{O})$.

In each step of the collapsed Gibbs sampling process, 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., $\mathbf{Z}_{\backslash u}^* = \{z_{u'} | u' \neq u\}$). In this case, the latent variables are sampled from the conditional posterior distribution as follows:

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

Note that the hyper-parameters of prior distributions, $\left\{\mathbf{h}^0, \mathbf{\Theta}^0\right\}$, are omitted in Eq. 10. From Eq. 7, the first term on the right-hand side of Eq. 10 can be described as follows:

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

From Eq. 8, the second term on the right-hand side of Eq. 10 is described as follows:

$$\frac{p(\mathbf{O}|\mathbf{Z}_{\backslash u}^*, z_u = i')}{p(\mathbf{O}_{\backslash u}|\mathbf{Z}_{\backslash u}^*)} \propto \exp\left(g_{i'}(\tilde{\mathbf{\Theta}}_{i'}) - g_{i'}(\tilde{\mathbf{\Theta}}_{i'\backslash u})\right), \tag{12}$$

where

$$\begin{aligned}
&g_{i'}(\tilde{\mathbf{\Theta}}_{i'}) \\
&\quad \triangleq \ln p(\mathbf{O}|\mathbf{Z}_{\backslash u}^*, z_u = i') \\
&\quad = D \log \Gamma\left(\frac{\tilde{\eta}_{i'}}{2}\right) - \frac{D}{2} \log \tilde{\xi}_{i'} - \frac{\tilde{\eta}_{i'}}{2} \sum\nolimits_d \log \tilde{\sigma}_{i',dd}
\end{aligned} \tag{13}$$

$$\begin{aligned}
&g_{i'}(\tilde{\mathbf{\Theta}}_{i'\backslash u}) \\
&\quad \triangleq \ln p(\mathbf{O}_{\backslash u}|\mathbf{Z}_{\backslash u}^*) \\
&\quad = D \log \Gamma\left(\frac{\tilde{\eta}_{i'\backslash u}}{2}\right) - \frac{D}{2} \log \tilde{\xi}_{i'\backslash u} - \frac{\tilde{\eta}_{i'\backslash u}}{2} \sum\nolimits_d \log \tilde{\sigma}_{i'\backslash u,dd}.
\end{aligned} \tag{14}$$

$\tilde{\mathbf{\Theta}}_{i'\backslash u}$ in Eq. 14 denotes the hyper-parameter of the posterior distribution for $\mathbf{\Theta}$ after removing u-th utterance, which is described as follows:

$$\tilde{\mathbf{\Theta}}_{i'\backslash u} \triangleq \begin{cases}
\tilde{\xi}_{i'\backslash u} & = \tilde{\xi}_i - T_u \\
\tilde{\eta}_{i'\backslash u} & = \tilde{\eta}_i - T_u \\
\tilde{\boldsymbol{\mu}}_{i'\backslash u} & = \frac{\tilde{\xi}_{i'} \tilde{\boldsymbol{\mu}}_{i'} - \sum_t \boldsymbol{o}_{ut}}{\tilde{\xi}_{i'\backslash u}} \\
\tilde{\sigma}_{i'\backslash u,dd} & = \sigma_{i',dd}^0 + r_{i',dd} - \sum_t (o_{ut,d})^2 \\
& \quad + \xi^0 (\mu_{i',d}^0)^2 - \tilde{\xi}_{i'\backslash u} (\tilde{\mu}_{i'\backslash u,d})^2
\end{cases} \tag{15}$$

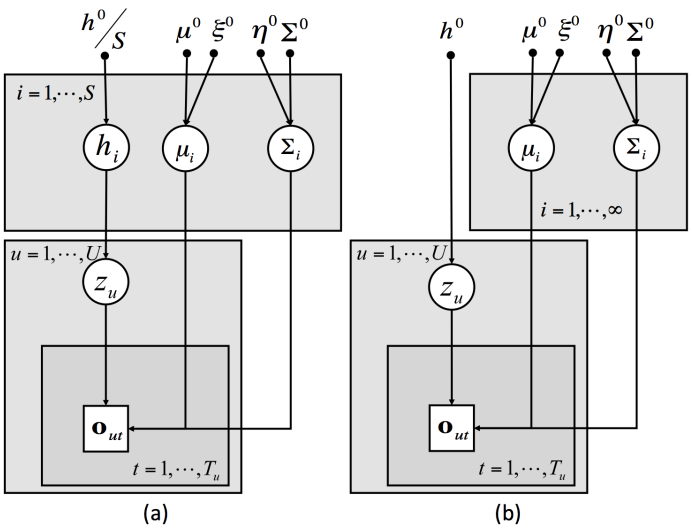
The optimal values of $\mathbf{Z}$ (i.e., the optimal assignments of utterances to clusters) can be obtained from its posterior distribution $P(\mathbf{Z}|\mathbf{O})$ by iterating to sample z_u from its conditional posterior distribution in Eq. 10 until convergence.

III. Utterance-oriented mixture model for infinite speakers

In this section, we attempt to extend the utterance-oriented mixture model for finite speakers in order to deal with infinite speakers. For this purpose, we introduce Dirichlet process as the prior distribution of mixture weights. The derived model (i.e., the UO-DPMM) is a type of Dirichlet process mixture model (DPMM) [5], but it differs from the original DPMM in that the generative unit is not a frame but rather an utterance. In the present study, UO-DPMM was built by using Chinese restaurant process (CRP) [12], which can avoid local solutions because of its sampling-based implementation. Furthermore, we can easily integrate CRP with other sophisticated methods, such as simulated annealing. The graphical model of the utterance-oriented mixture model for infinite speakers is shown in Fig. 1 (b). Table 1 provides a pseudo code of this method.

CRP is found by taking the limit of S (i.e., $S \to \infty$) in Eq. 10. Note that there are at most $U(< S)$ speaker clusters to which at least one utterance is assigned. In the case of S being infinite, most clusters should be empty. In this case, we can separately compute Eq. 11 for the case where the u-th utterance is assigned to a cluster with more than one utterance (i.e., $n_{i'} > 0$) and the case where the u-th utterance is assigned to a new cluster with no utterance (i.e., $n_{i'} = 0$).

$$\frac{P(\mathbf{Z}_{\backslash u}^*, z_u = i')}{P(\mathbf{Z}_{\backslash u}^*)} = \begin{cases}
\frac{\frac{h^0}{S} + n_{i'}}{U - 1 + h^0}, & \text{if } i' = z_k \text{ for } \exists k \neq u \\
\frac{\frac{h^0}{S}}{U - 1 + h^0}, & \text{if } i' \neq z_k \text{ for } \forall k \neq u
\end{cases} \tag{16}$$



h^0 — Hyper-parameter of concentration parameter
$\boldsymbol{\mu}^0, \xi^0$ — Hyper-parameter of prior for the mean vector
$\boldsymbol{\Sigma}^0, \eta^0$ — Hyper-parameter of prior for the covariant matrix
h_i — Weight of the i-th cluster
$\boldsymbol{\mu}_i$ — Mean vector of the i-th cluster
$\boldsymbol{\Sigma}_I$ — Covariant matrix of the i-th cluster
z_u — Latent variable of the u-th utterance
$\mathbf{o}_{ut}$ — Frame-wise observation of the t-th frame in the u-th utterance
u — Utterance index
t — Frame index
i — Cluster index
U — Number of utterances
T_u — Number of frames in u-th utterance
S — Number of clusters in FSM

Fig. 1. Graphical models of utterance-oriented mixture models for (a) finite and (b) infinite speakers.

By taking the limit of $S \to \infty$, the number of utterances U satisfies $U \ll S$ and thus we can assume that there are S empty clusters. Therefore, by combining the empty clusters, Eq. 16 is described as follows:

$$\frac{P(\mathbf{Z}^*_{\backslash u}, z_u = i')}{P(\mathbf{Z}^*_{\backslash u})} = \begin{cases} \frac{\frac{h^0}{S}+n_{i'}}{U-1+h^0}, & \text{if } i' = z_k \text{ for } \exists k \neq u \\ S \cdot \frac{\frac{h^0}{S}}{U-1+h^0}, & \text{if } i' \neq z_k \text{ for } \forall k \neq u \end{cases} \tag{17}$$

Taking the limit of $S \to \infty$ in Eq. 17 allows us to derive the following equation:

$$\frac{P(\mathbf{Z}^*_{\backslash u}, z_u = i')}{P(\mathbf{Z}^*_{\backslash u})} = \begin{cases} \frac{n_{i'}}{U-1+h^0}, & \text{if } i' = z_k \text{ for } \exists k \neq u \\ \frac{h^0}{U-1+h^0}, & \text{if } i' \neq z_k \text{ for } \forall k \neq u \end{cases} \tag{18}$$

From Eq. 8, we can also separately compute the second term on the right-hand side of Eq. 10 as follows:

$$\frac{p(\mathbf{O}, \mathbf{Z}^*_{\backslash u}, z_u = i')}{p(\mathbf{O}_{\backslash u}, \mathbf{Z}^*_{\backslash u})} = \begin{cases} \exp\left(g_{i'}(\tilde{\boldsymbol{\Theta}}_{i'}) - g_{i'}(\tilde{\boldsymbol{\Theta}}_{i'\backslash u})\right), \\ \qquad \text{if } z_k = i' \text{ for } \exists k \neq u \\ \exp\left(g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}}) - g_{\text{new}}(\boldsymbol{\Theta}_0)\right) \\ \qquad \text{if } z_k \neq i' \text{ for } \forall k \neq u \end{cases} \tag{19}$$

where $g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}})$ and $g_{\text{new}}(\boldsymbol{\Theta}_0)$ denote the logarithmic likelihood for $\mathbf{O_u}$ to the new cluster, and the prior likelihood of the parameter itself, respectively.

We can evaluate both $g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}})$ and $g_{\text{new}}(\boldsymbol{\Theta}_0)$ by using Eq. 13, noting that only the u-th utterance is assigned to the new cluster for $g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}})$ and no ones are assigned to the new cluster for $g_{\text{new}}(\boldsymbol{\Theta}_0)$. That is, we can respectively evaluate $g_{\text{new}}(\boldsymbol{\Theta}_0)$ and $g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}})$ by substituting $\tilde{\boldsymbol{\Theta}}_{i'}$ in Eq. 13 to $\boldsymbol{\Theta}_0$ and $\tilde{\boldsymbol{\Theta}}_{\text{new}}$, which is described as follows:

$$\tilde{\boldsymbol{\Theta}}_{\text{new}} \triangleq \begin{cases} \tilde{\xi}_{\text{new}} & = \xi_0 + T_u \\ \tilde{\eta}_{\text{new}} & = \eta_0 + T_u \\ \tilde{\boldsymbol{\mu}}_{\text{new}} & = \frac{\boldsymbol{\mu}_0 + \sum_t \mathbf{o}_{ut}}{\tilde{\xi}_{\text{new}}} \\ \tilde{\sigma}_{new,dd} & = \sigma^0_{dd} + \sum_t (o_{ut,d})^2 \\ & \quad + \xi^0 (\mu^0_d)^2 - \tilde{\xi}_{\text{new}} (\tilde{\mu}_{\text{new},d})^2 \end{cases} \tag{20}$$

From Eqs. 18 and 19, the posterior probability of the latent variables can be finally described as follows:

$$p(z_u = i' | \mathbf{O}, \mathbf{Z}_{\backslash u}) \propto \begin{cases} \frac{n_{i'}}{U-1+h^0} \cdot \exp\left(g_i(\tilde{\boldsymbol{\Theta}}_{i'}) - g_i(\tilde{\boldsymbol{\Theta}}_{i'\backslash u})\right), \\ \qquad \text{if } z_k = i' \text{ for } \exists k \neq u \\ \frac{h^0}{U-1+h^0} \cdot \exp\left(g_{\text{new}}(\tilde{\boldsymbol{\Theta}}_{\text{new}}) - g_{\text{new}}(\boldsymbol{\Theta}_0)\right) \\ \qquad \text{if } z_k \neq i' \text{ for } \forall k \neq u \end{cases} \tag{21}$$

We iteratively reassign each utterance to one of the existing clusters or the new cluster in proportion to Eq. 21 until the value of the samples converges. As shown in Eq. 21, the hyper-parameter h^0 determines how frequently each utterance is reassigned to the new cluster. The estimated number of speaker clusters, therefore, depends on the value of h^0. In the next section, we demonstrate that this parameter can be tuned by using a development set.

IV. Speaker Clustering Experiments

We carried out the speaker clustering experiments by using the TIMIT [13] and Corpus of Spontaneous Japanese (CSJ) [14] databases. We compared UO-DPMM described in Section III with existing hierarchical agglomerative clustering based on the Bayesian information criterion (HAC-BIC) [1] in speaker clustering with estimation of the number of speakers.

HAC-BIC is similar to UO-DPMM in terms of the model structure, i.e., both methods assume that each speaker can be represented by a single Gaussian and estimate the number of clusters using model complexity. Here, the aim of the

Algorithm 1 Speaker clustering by using UO-DPMM. Threshold is 30 for TIMIT and 50 for CSJ.

Initialize S and $\{z_u\}_{u=1}^{U}$.
repeat
  for all u = shuffle $(1, \cdots, U)$ **do**
    Sample z_u according to Eq. 21.
    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 number of iterations exceeds threshold

present study is to verify if the model-based speaker clustering approach can be extended so as to estimate the number of speakers by incorporating the nonparametric Bayesian techniques in the utterance-oriented speaker mixture model. We therefore are determined to focus on comparing UO-DPMM and HAC-BIC to make this experiment comparable.

A. Speech data

We performed the speaker clustering experiments by using six evaluation sets obtained from the TIMIT and CSJ databases. We used two evaluation sets in TIMIT. T-1 was the "core test set," which included 192 utterances spoken by 24 speakers. T-2 was the "complete test set," which excluded the core test set in the TIMIT database and included 1,152 utterances spoken by 144 speakers. T-1 and T-2 are balanced data, in which each speaker spoke the same number of utterances. The remaining four evaluation sets were obtained from lectures in CSJ as follows. First, all lectures were divided into utterance units based on the segments of silence in their transcriptions that were longer than 500 ms; 5 and 10 speakers were then randomly selected and their 100 utterances were selected for C-1 and C-2. Each utterance was between 2 and 10 s long. Next, we selected another 5 and 10 speakers and all their utterances for C-3 and C-4. C-3 and C-4 are "unbalanced" and large-scale data (they include approximately 4 and 6 million samples, respectively). Table I lists the number of speakers and utterances in the evaluation sets used. Speech data were sampled at 16 kHz and quantized into 16-bit data.

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

B. Measurement

We used the average cluster purity (ACP), the average speaker purity (ASP), and their geometric mean value (K) for the evaluation criteria in speaker clustering [15]. The correct speaker labels for utterances were manually annotated. Let S_T be the correct number of speakers, S the estimated number of speakers, n_{ij} the estimated number of utterances assigned to speaker cluster i in all utterances of speaker j, n_j the estimated number of utterances of speaker j, n_i the estimated number of utterances assigned to speaker cluster i, and U the total number of utterances. Cluster purity p_i, speaker purity q_j, and the K value are then calculated as follows:

$$\begin{aligned} p_i &= \sum_{j=0}^{S_T} \frac{n_{ij}^2}{n_i^2}, \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2} \\ K &= \sqrt{\frac{\sum_i p_i \cdot \sum_j q_j}{S_T S}} \end{aligned} \tag{22}$$

We additionally calculated the speaker diarization error rate (DER) [16] in the experiments for CSJ. The DER is the ratio of incorrectly attributed speech time, which is calculated as follows

$$DER = \frac{U_{\text{fa}} + U_{\text{error}}}{U_{\text{ref}}} \tag{23}$$

where U_{fa} denotes the total length of utterances not aligned with the speaker labels in the case where $S_T > S$ (i.e. false alarm utterances), namely the speech time of utterances assigned to improper speakers in the case that the estimated number of speakers is larger than the true number of speakers. U_{error} denotes the total length of utterances aligned with the wrong speaker labels and U_{ref} denotes the total length of all utterances in a test set. The clustering result and speaker labels concurred in order to minimize DER.

The number of iterations was set to 50 for TIMIT and 30 for CSJ. We considered the first 49 and 29 iterations for TIMIT and CSJ as the burn-in periods, respectively, leading the K values obtained from these periods to be rejected. The K value from the remaining one iteration was then measured. We carried out the same experiment 50 times but using different seeds to generate random numbers and then measured the average of their K values.

C. Experimental setup

The hyper-parameters in Eq. 3 were set as follows: $h^0 = 1$, $\xi^0 = 1$, and $\eta^0 = 1$. $\boldsymbol{\mu}_i^0$ and $\boldsymbol{\Sigma}_i^0$ were computed as the mean and covariance of all data used in the database. In this experiment, we first estimated the optimal number of clusters as well as the optimal assignments of utterances to clusters. Next, we carried out the speaker clustering experiments using the TIMIT and CSJ databases. We then cross-validated for each pair of {T-1, T-2}, {C-1, C-2}, and {C-3, C-4} to decide the penalty parameter in the BIC-based method and the hyper-parameter h^0 in UO-DPMM.

D. Experimental results

Table II lists the speaker clustering results for TIMIT. These results show that UO-DPMM outperformed BIC-HAC in terms of estimating the number of speakers for both T-1 and T-2. UO-DPMM also outperformed BIC-HAC in terms of the K value for T-2. Table III shows the speaker clustering results for CSJ. UO-DPMM outperformed BIC-HAC for all evaluation sets. Specifically, BIC-HAC performed considerably worse for C-3 and C-4. These results indicate that UO-DPMM can be robustly estimated for the unbalanced and large-scale data while BIC-HAC significantly diminishes the clustering accuracy for these data.

TABLE I
DETAILS OF TEST SET. # SPEAKERS, # UTTERANCES, # SAMPLES, AND TOTAL DURATION DENOTE THE NUMBER OF SPEAKERS, NUMBER OF UTTERANCES, NUMBER OF FRAME-WISE OBSERVATIONS, AND TOTAL DURATION.

	T-1	T-2	C-1	C-2	C-3	C-4
# speakers	24	144	5	10	5	10
# utterances	192	1,152	500	1,000	9,333	15,435
(# samples)	(5.8 K)	(353 K)	(209 K)	(404 K)	(4.0 M)	(6.4 M)
total duration	9.7 [min.]	59.0 [min.]	35.0 [min.]	1.1 [hour]	11.1 [hour]	17.6 [hour]

TABLE II
SPEAKER CLUSTERING RESULTS FOR TIMIT. $\#$CL. DENOTES THE NUMBER OF CLUSTERS ESTIMATED.

Eval.	Method	#cl.	ACP	ASP	K
T-1	UO-DPMM	32.4	0.84	0.72	**0.78**
(spkr:24, utt:192)	HAC-BIC	34.0	0.85	0.71	**0.78**
T-2	UO-DPMM	145.0	0.53	0.55	**0.54**
(spkr:144, utt:1,152)	HAC-BIC	174.0	0.54	0.49	0.52

TABLE III
SPEAKER CLUSTERING RESULTS FOR CSJ. $\#$CL. DENOTES THE NUMBER OF CLUSTERS ESTIMATED.

Eval.	Method	#cl	ACP	ASP	K	DER [%]
C-1	UO-DPMM	9.15	0.96	0.78	**0.87**	**0.13**
(spkr:5, utt:500)	HAC-BIC	9.50	0.85	0.72	0.78	0.25
C-2	UO-DPMM	10.4	0.87	0.84	**0.81**	**0.20**
(spkr:10, utt:1,000)	HAC-BIC	16.5	0.73	0.68	0.70	0.36
C-3	UO-DPMM	10.9	0.91	0.70	**0.80**	**0.23**
(spkr:5, utt:9,333)	HAC-BIC	2.00	0.21	0.55	0.34	0.72
C-4	UO-DPMM	13.7	0.73	0.68	**0.71**	**0.28**
(spkr:10, utt:15,435)	HAC-BIC	4.00	0.12	0.29	0.19	0.83

Next, we discuss the convergence of the sampling procedure in UO-DPMM. For that purpose, experiments were conducted with the same dataset but different seeds of a pseudo-random number generator. Figure 2 shows the K values obtained from UO-DPMM. The eight lines in each figure show the respective results from the eight trials using the different seeds. This figure shows that all samples from all trials converge to the unique distributions. This result indicates that the proposed method is robust against the local optima problem depending on the initial states.

Finally, we discuss computational costs. In the experiment for C-4, UO-DPMM took 11.8 seconds per iteration and 588 seconds for 50 iterations on average when Intel Xeon 3.00 GHz was used. UO-DPMM required comparatively less computation time because of its fast convergence, although sampling-based methods generally require many iterations until the value of the samples converges. Figure 2 shows that all samples converge to the unique distributions within 30 iterations for all datasets. The advantage of UO-DPMM is yielded by using utterance-oriented sampling. The general Gibbs sampler induces the slow convergence speed due to its sampling procedure in which only one sample is reassigned in each iteration. In contrast, the utterance-oriented sampling simultaneously reassigns a set of frames in each iteration.

V. DISCUSSION

We employed nonparametric Bayesian techniques to make it possible to estimate the number of speakers in the model-based speaker clustering system. A recently proposed sticky hierarchical Dirichlet process hidden Markov model (HDP-HMM) [17] is another approach to incorporate a nonparametric Bayesian manner in model-based speaker clustering. Here, we discuss the difference between UO-DPMM and HDP-HMM.

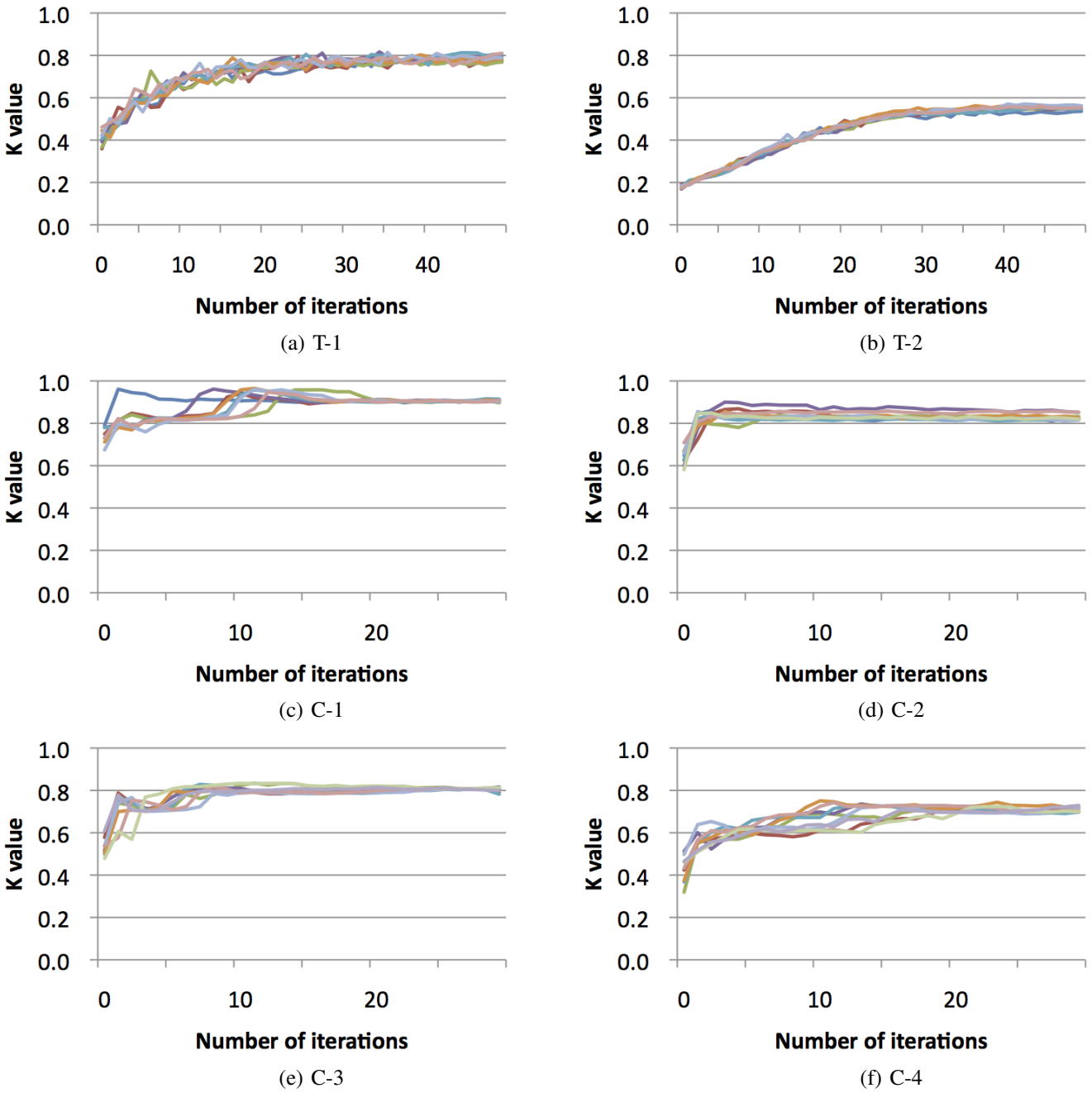
The most obvious difference is a generative unit. The unit is an utterance in UO-DPMM but a frame in HDP-HMM. This difference affects the definition of latent variables and the inference method of those variables. In UO-DPMM, the latent variable is defined for each utterance, which is composed of a set of frames, and sampled from the posterior distribution conditioned on the other utterances. In HDP-HMM, on the other hand, the latent variable is defined for each frame and sampled from the posterior distribution conditioned on the other frames. UO-DPMM, therefore, converges much faster than HDP-HMM when the boundaries of speech are given. In fact, HDP-HMM needs over 10,000 iterations of Gibbs sampling and is hard to apply on the large-scale data that we deal with in the present study.

In this paper, we introduced MCMC-based approach to estimate the model structure of UO-DPMM. Frame-wise observation approach for DPMM is addressed in previous researches [7], [18]. In these methods, however, MCMC-based approach is not applicable because sampling of frame-wise hidden variables requires impractically heavy computational cost. In order to avoid this computation, these methods introduce the deterministic approach based on stick-breaking process [7] and variational Bayesian method [18]. These methods however often suffer from local solutions and over-learning problems. The proposed UO-DPMM, on the other hand, realizes MCMC-based approach by introducing the utterance-oriented assumption.

VI. CONCLUSION AND FUTURE WORK

A nonparametric Bayesian speaker modeling based on UO-DPMM was proposed to make it possible to estimate the number of speakers in model-based speaker clustering. The experimental comparison demonstrated that the proposed method was successfully applied to speaker clustering on practically large-scale data and outperformed the existing hierarchical agglomerative clustering method.

The present study assumed that each speaker is distributed in accordance with a Gaussian. The speaker distribution can be represented by a GMM instead of a single Gaussian, and each utterance can be assumed to be generated from a mixture of these GMMs (MoGMMs). GMM-based speaker distributions have been applied to the HAC-based speaker clustering, i.e., HAC-GMM. We have also already developed utterance-oriented speaker modeling with MoGMMs for the



(a) T-1 (b) T-2

(c) C-1 (d) C-2

(e) C-3 (f) C-4

Fig. 2. K values obtained from proposed method for (a) T-1, (b) T-2, (c) C-1, (d) C-2, (e) C-3 and (f) C-4. Eight lines in each figure show results of eight trials using different seeds.

finite speakers [2], [3]. In future, we aim to derive an effective Gibbs sampling algorithm to incorporate the GMM-based speaker distributions in UO-DPMM and compare it with HAC-GMM.

Acknowledgment

This work was supported in part by Grants for Excellent Graduate Schools, MEXT, Japan.

References

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

[2] S. Watanabe, D. Mochihashi, T. Hori, and A. Nakamura, "Gibbs sampling based multi-scale mixture model for speaker clustering.," ICASSP, pp.4524–4527, 2011.

[3] N. Tawara, T. Ogawa, S. Watanabe, and T. Kobayashi, "Fully bayesian inference of multi-mixture gaussian model and its evaluation using speaker clustering," ICASSP, pp.5253–5256, 2012.

[4] N. Tawara, T. Ogawa, S. Watanabe, A. Nakamura, and T. Kobayashi, "Blocked Gibbs sampling based multi-scale mixture model for speaker clustering on noisy data," MLSP, 2013.

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

[6] N. Tawara, S. Watanabe, T. Ogawa, and T. Kobayashi, "Speaker clustering based on utterance-oriented dirichlet process mixture model.," INTERSPEECH, pp.2905–2908, 2011.

[7] F. Valente, "Infinite models for speaker clustering," International Con-

ference on Spoken Language Processing, 2006.
[8] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," IEEE Automatic Speech Recognition Understanding Workshop, 2003.
[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, vol.39, no.1, pp.1–38, 1977.
[10] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," IEEE Trans. Speech Audio Process., vol.2, pp.291–298, 1994.
[11] J.S. Liu, Monte Carlo strategies in scientific computing, Springer, Jan. 2008.
[12] D. Aldous, "Exchangeability and related topics," École dété de probabilités de Saint-Flour, vol.XIII–1983, pp.1–198, 1985.
[13] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.
[14] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," Proc. LREC2000, pp.947–952, 2000.
[15] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," ICASSP, pp.757–760, 1998.
[16] J.G. Fiscus, J. Ajot, and J.S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," CLEAR, pp.373–389, 2007.
[17] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky, "A sticky HDP-HMM with application to speaker diarization," Annals of Applied Statistics, vol.5, no.2A, pp.1020–1056, 2011.
[18] A.H.H.N. Torbati, J. Picone, and M. Sobel, "Applications of Dirichlet process mixtures to speaker adaptation," ICASSP, pp.4321–4324, 2012.

APPENDIX

This appendix derives the joint posterior distribution $p(\mathbf{O}, \mathbf{Z})$ described in Eq. 5 along with the Dirichlet and Gaussian-Gamma conjugate priors.

Priors:

The Dirichlet distribution is written as

$$\begin{aligned} P(\mathbf{h}) &= \mathcal{D}(\mathbf{h}|\mathbf{h}^0) \\ &= \frac{\Gamma(h^0)}{S \cdot \Gamma(h^0)} \prod_i h_i^{\frac{h_i^0}{S}-1} \end{aligned} \tag{24}$$

The Gaussian-Gamma distributions for the parameter of the i-th cluster are written as

$$\begin{aligned} & p(\mathbf{\Theta}_i) \\ &= p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i|\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0) \\ &= \prod_i \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{\mu}^0, (\xi^0)^{-1}\mathbf{\Sigma}_i) \prod_d \mathcal{G}(\sigma_{dd}|\eta^0, \sigma_{dd}^0) \\ &= \prod_i \prod_d \frac{\xi^0}{(2\pi)^{1/2}(\sigma_{i,dd})^{1/2}} \exp\left\{-\frac{\xi^0(\mu_{i,d}-\mu_d^0)^2}{2\sigma_{i,dd}}\right\} \\ &\quad \cdot \frac{1}{\Gamma(\eta^0)}(\sigma_{dd}^0)^{\frac{\eta^0}{2}} \sigma_{i,dd}^{-\eta^0+1} \exp\left(-\frac{\sigma_{dd}^0}{2\sigma_{i,dd}}\right) \\ &= \prod_i \frac{(\xi^0)^{\frac{D}{2}} \left(\prod_d \sigma_{dd}^0\right)^{\frac{\eta^0}{2}}}{(2\pi)^{D/2}\Gamma(\eta^0)^{\frac{D}{2}}} \left(\prod_d \sigma_{i,dd}\right)^{-\eta^0+\frac{1}{2}} \\ &\quad \cdot \exp\left\{-\sum_d \frac{1}{2\sigma_{i,dd}}\left(\xi^0(\mu_{i,d}-\mu_d^0)^2 + \sigma_{dd}^0\right)\right\} \\ &= \prod_i \frac{1}{Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0)} \left(\prod_d \sigma_{i,dd}\right)^{-\eta^0+\frac{1}{2}} \\ &\quad \cdot \exp\left\{-\sum_d \frac{1}{2\sigma_{dd}}\left(\xi^0(\mu_{i,d}-\mu_d^0)^2 + \sigma_{dd}^0\right)\right\} \end{aligned} \tag{25}$$

where,

$$Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0) = \frac{(2\pi)^{\frac{D}{2}}\Gamma(\eta^0)^D}{(\xi^0)^{\frac{D}{2}}\left(\prod_d \sigma_{dd}^0\right)^{\frac{\eta^0}{2}}} \tag{26}$$

Joint distribution:

We derive the joint distribution for $\{\mathbf{O}, \mathbf{\Theta}\}$ by conditioning on the latent variable $\mathbf{Z}$ as follows:

$$\begin{aligned} & p(\mathbf{O}, \mathbf{\Theta}|\mathbf{Z}) \\ &= p(\mathbf{O}|\mathbf{Z}, \mathbf{\Theta})p(\mathbf{\Theta}) \\ &= \prod_{u,t} p(\mathbf{o}_{ut}|\boldsymbol{\mu}_{z_u}, \mathbf{\Sigma}_{z_u}) \prod_i p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i|\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0) \\ &= \prod_{u,t} \left\{ \frac{(2\pi)^{\frac{D}{2}}}{Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0)} \left(\prod_d \sigma_{z_t,dd}\right)^{-\frac{1}{2}} \right\} \\ &\quad \cdot \exp\left[-\sum_u \sum_t \sum_d \left\{\frac{(o_{ut,d}-\mu_{z_t,d})^2}{2\sigma_{z_t,dd}}\right\}\right] \\ &\quad \cdot \exp\left[-\sum_i \sum_d \frac{1}{2\sigma_{i,dd}}\left\{\xi^0(\mu_{i,d}-\mu_d^0)^2 + \sigma_{dd}^0\right\}\right] \\ &= \prod_i \left\{ \frac{(2\pi)^{\frac{n_i^{frm} D}{2}}}{Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0)} \left(\prod_d \sigma_{i,dd}\right)^{-\frac{n_i^{frm}}{2}} \right\} \\ &\quad \cdot \exp\left[-\sum_i \sum_d \frac{1}{2\sigma_{i,dd}}\left\{\tilde{\xi}_i\left(\mu_{i,d}-\tilde{\mu}_{i,d}\right)^2 + \tilde{\sigma}_{i,dd}\right\}\right] \\ &= \prod_i \frac{(2\pi)^{\frac{n_i^{frm} D}{2}} Z(\tilde{\xi}_i, \tilde{\eta}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{\Sigma}}_i)}{Z(\xi^0, \eta^0, \boldsymbol{\mu}^0, \mathbf{\Sigma}^0)} \\ &\quad \cdot \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_i, \tilde{\xi}_i^{-1}\mathbf{\Sigma}_i\right) \prod_d \mathcal{G}\left(\tilde{\eta}_i, \tilde{\sigma}_{i,dd}\right) \end{aligned} \tag{27}$$

Marginalized distribution:

We derive the likelihood for the complete data case, $p(\mathbf{O}, \mathbf{Z})$, by marginalizing the joint distribution $p(\mathbf{O}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{h}) = p(\mathbf{O}, \mathbf{\Theta}|\mathbf{Z})p(\mathbf{Z}, \mathbf{h})$ with respect to the hyper-parameters $\{\mathbf{h}, \mathbf{\Theta}\}$.

First, we derive the likelihood $P(\mathbf{Z})$ by marginalizing $p(\mathbf{Z}, \mathbf{h})$ with respect to $\{\mathbf{w}_i, \mathbf{h}, \mathbf{\Theta}_i\}_i$. Assuming the independence of the utterance-level latent variables z_u, this can be analytically derived as follows:

$$\begin{aligned} P(\mathbf{Z}) &= \int p(\mathbf{h}) \prod_{u=1}^{U} P(z_u|\mathbf{h}) \mathrm{d}\mathbf{h} \\ &= \frac{\Gamma(\sum_i h_i^0)}{\prod_i \Gamma(h_i^0)} \int \prod_{i=1}^{S} h_i^{\sum_u \delta(z_u,i)+h_i^0-1} \mathrm{d}h_i \\ &= \frac{\Gamma(\sum_i h_i^0)}{\prod_i \Gamma(h_i^0)} \frac{\prod_i \Gamma(\tilde{h}_i)}{\Gamma(\sum_i \tilde{h}_i)} \end{aligned} \tag{28}$$

Finally, we derive the likelihood $p(\mathbf{\Theta})$ by marginalizing $p(\mathbf{O}, \mathbf{\Theta})$ with respect to the model parameter $\mathbf{\Theta}$. By using

Eq. 27, this can be analytically derived as follows:

$$
\begin{aligned}
&p(\mathbf{O}|\mathbf{Z}) \\
&\quad = \int p(\mathbf{O}|\mathbf{Z},\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \\
&\quad = \int p(\mathbf{O},\boldsymbol{\Theta}|\mathbf{Z})\mathrm{d}\boldsymbol{\Theta} \\
&\quad = \prod_i (2\pi)^{\frac{n_i^{frm}D}{2}} \frac{Z(\tilde{\xi}_i,\tilde{\eta}_i,\tilde{\boldsymbol{\mu}}_i,\tilde{\boldsymbol{\Sigma}}_i)}{Z(\xi^0,\eta^0,\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0)} \\
&\qquad \cdot \int \mathcal{N}\left(\boldsymbol{\mu}_i,\tilde{\xi}_i^{-1}\boldsymbol{\Sigma}_i\right)\mathrm{d}\boldsymbol{\mu}_i \prod_d \int \mathcal{G}\left(\tilde{\eta}_i,\tilde{\sigma}_{i,dd}\right)\mathrm{d}\sigma_{i,dd} \\
&\quad = \prod_i (2\pi)^{\frac{n_i^{frm}D}{2}} \frac{Z(\tilde{\xi}_i,\tilde{\eta}_i,\tilde{\boldsymbol{\mu}}_i,\tilde{\boldsymbol{\Sigma}}_i)}{Z(\xi^0,\eta^0,\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0)} \\
&\quad = \prod_i (2\pi)^{-\frac{n_i^{frm}D}{2}} \frac{(\xi^0)^{\frac{D}{2}}\left(\Gamma\left(\frac{\eta^0}{2}\right)\right)^{-D}\left(\prod_d \sigma_{dd}^0\right)^{\frac{\eta^0}{2}}}{(\tilde{\xi}_i)^{\frac{D}{2}}\left(\Gamma\left(\frac{\tilde{\eta}_i}{2}\right)\right)^{-D}\left(\prod_d \tilde{\sigma}_{i,dd}\right)^{\frac{\tilde{\eta}_i}{2}}}
\end{aligned}
\tag{29}
$$

By using Eqs. 28 and 29, the marginalized distribution for the complete data case can be finally described as follows:

$$
\begin{aligned}
p(\mathbf{O},\mathbf{Z}) \quad &= p(\mathbf{O}|\mathbf{Z})P(\mathbf{Z}) \\
&= \frac{\Gamma(\sum_i h^0)}{\prod_i \Gamma(h^0)} \frac{\prod_i \Gamma(\tilde{h}_i)}{\Gamma(\sum_i \tilde{h}_i)} \prod_i (2\pi)^{-\frac{n_i^{frm}D}{2}} \\
&\quad \cdot \left\{ \frac{(\xi^0)^{\frac{D}{2}}\left(\Gamma\left(\frac{\eta^0}{2}\right)\right)^{-D}\left(\prod_d \sigma_{dd}^0\right)^{\frac{\eta^0}{2}}}{(\tilde{\xi}_i)^{\frac{D}{2}}\left(\Gamma\left(\frac{\tilde{\eta}_i}{2}\right)\right)^{-D}\left(\prod_d \tilde{\sigma}_{i,dd}\right)^{\frac{\tilde{\eta}_i}{2}}} \right\}
\end{aligned}
\tag{30}
$$