Probabilistic Speaker Diarization With Bag-of-Words Representations of Speaker Angle Information

Katsuhiko Ishiguro, *Member, IEEE*, Takeshi Yamada, *Member, IEEE*, Shoko Araki, *Member, IEEE*, Tomohiro Nakatani, *Senior Member, IEEE*, and Hiroshi Sawada, *Senior Member, IEEE*

***Abstract*—Speaker diarization determines "who spoke when" from the recorded conversations of an unknown number of people. In general, we have no *a priori* information about the number, the locations, or even the characteristics of the speakers. Additionally, speakers' speech utterances vary dynamically because of turn-taking during the conversations. These conditions make the speaker-clustering task extremely difficult. The problem becomes even harder if online (incremental) processing is required. In this paper, we formulate the speaker-clustering problem as the clustering of the sequential audio features generated by an unknown number of latent mixture components (speakers). We employ a probabilistic model that assumes time-sensitive speaker mixtures at every time frame, which, surprisingly, suits the diarization scenario. We combine the time-varying probabilistic model with direction of arrival (DOA) information calculated from a microphone array in a bag-of-words (BoW)-style feature representation. The proposed system effectively estimates the number and locations of the speakers in an online manner based on the standard Bayes inference scheme. Experiments confirm that the proposed model can successfully infer the number and features of speakers and yield better or comparable speaker diarization results compared with conventional methods in several datasets.**

***Index Terms*—Bag-of-words (BOW), clustering, direction of arrival (DOA), latent Dirichlet allocation (LDA), speaker diarization, microphone arrays, variational Bayes inference.**

I. Introduction

THE automatic indexing of multi-party conversations such as group meetings has been studied intensively because it can allow their rapid retrieval from archives, automatic minutes taking, and automatic speech enhancements in conversations. Many multi-party conversations have been recorded in relation to, for example, Augmented Multi-party Interaction with Distant Access (AMIDA) [1], Computers in the Human Interaction Loop (CHIL) [2], and the NIST Rich Transcription Meeting Recognition Project [3]. Also, a number of methodologies for this problem have been proposed [4]–[6]. One essential function for indexing such recorded data is speaker diarization, i.e., estimating "who spoke when" from audio recordings [7]–[9]. A speaker diarization system estimates the speaker segment boundaries by classifying speech features, such as acoustic and/or location features.

A key issue of speaker diarization is speaker clustering: clustering the features into an unknown number of clusters (speakers). As features, acoustic and location information are widely employed. The former is the mel-frequency cepstral coefficient (MFCC) features, which were used in [10], for example. MFCC features are so strong that we are able to discern the speakers very precisely if we can capture the sole clean utterances of each speaker in a stable manner. Friedland and Vinyals [11] proposed a very fast diarization system. In this paper, MFCC features were effectively combined with a Gaussian mixture model (GMM) model. However, MFCC features tend to be degraded by environmental noise and the overlap in different speakers' utterances. This is inconvenient for the speaker diarization task, where we expect the system to recognize complex dialogs with many overlaps in various environments. The latter, those adopted by this paper, are related to the location information of the speakers, including time difference of arrival (TDOA) or direction of arrival (DOA), estimated via microphone arrays. DOA-based diarization, which requires that the speakers do not move during the conversation, is robust against voice overlap, and this is highly desirable for speaker diarization in meeting situations. Some papers [8], [12] successfully combined MFCC and TDOA features for diarization. However, the use of solely TDOA or DOA features also has been proven to be useful in speaker clustering and diarization tasks [13], [14]. In this paper, we determine the relative time difference of sound signal arrival between microphone pairs of a microphone array following [13].

In most cases, we have no *a priori* information about the number, the locations and the characteristics of the speakers. We also note that online (incremental) clustering is necessary for some applications. For example, in the teleconference system we need instant speaker diarization for better speech enhancements of the speaker. Instant speaker diarization is also required for human–computer interface situations such as robots serving many persons. In both cases, participants in the conversation may vary every moment, and the overlapping of utterances makes the problem even more difficult. The diarization system needs to solve these problems on-site; thus, online clustering is an important issue. It is evident that we

Manuscript received June 30, 2010; revised January 05, 2011; accepted April 21, 2011. Date of current version January 13, 2012. An earlier version of this work was presented at the 2009 Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2009) as a workshop paper. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Friedland.

K. Ishiguro, S. Araki, T. Nakatani, and H. Sawada are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: ishiguro.katsuhiko@lab.ntt.co.jp; araki.shoko@lab.ntt.co.jp; nakatani.tomohiro@lab.ntt.co.jp; sawada.hiroshi@lab.ntt.co.jp).

T. Yamada is with NTT Science and Core Technology Laboratory Group, NTT Corporation, Kanagawa 243-0198, Japan (e-mail: yamada.tak@lab.ntt.co.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2151858

cannot employ simple mixture models such as the GMM, whose mixing ratio is fixed for all time frames, because the "mixture" of speakers' speech fragments is time sensitive due to turn-taking during the conversations.

In this paper, we adopted the new probabilistic model called dynamic Latent Dirichlet Allocation (dLDA), which was proposed in [15]. Employing dLDA enables us to incorporate the dynamic properties of the conversation, especially turn-taking, into the model to represent time-frame-sensitive audio feature distribution. The model formulates time-varying speaker mixtures as a simple Markov model that depends on the previous time frame. Thanks to this time-varying nature of the dLDA model, the proposed model can effectively estimate the number of speakers participating in the conversation by clustering location information. We also incorporate a bag-of-words (BoW)-style feature representation for audio signal processing, which is attracting a lot of attention in other research fields such as natural language processing and computer vision. To utilize a BoW representation in DOA-based speaker diarization, we propose replacing the standard Dirichlet-Multinomial model [15], [16] with a Gaussian mixture model as in our previous research [17]. We validate the performance of the dLDA model with quantitative and qualitative experiments, compared with conventional speaker diarization techniques.

In Section II, we give an overview of the proposed system and related works in the literature. In Section III, we use the probabilistic representation to describe the speaker-clustering problem and compare several mixture models. In Section IV, we introduce the dLDA model for speaker clustering. Section V describes the Bayesian inference algorithm. Section VI presents the experiments and the results, which confirm the superiority of the proposed model. Section VII concludes this paper.

II. System Overview and Related Works

Fig. 1 gives an overview of a typical DOA-based speaker diarization system. A sound recording of a conversation held by an unknown number of speakers is given to the system as a sequence of time frames. We then extract features for speaker clustering from the sequence of frames.

Our model is a combination of two feature extractors (preprocessors). The first, called voice activity detector (VAD) [18], computes the probability of a frame containing any voice. If this probability is low, the frame is assumed to be a noise frame, i.e., an environmental noise, and is excluded from further processing. The second, and main extractor, is DOA. We utilize the DOA extractor proposed in [19]. This feature extractor estimates 1) the direction (angle) of the sound source from a microphone array and 2) the power of the sound heard from that direction. We expect that many high-power vocal signals will be emitted from the locations of the speakers (see Appendix for details).

Given the sequence of DOA features (power-orientation features from frames), the clustering processor infers the number and locations of the speakers. In the last step, the classification processor determines the utterance status of each speaker at each time based on the clustering results.

As discussed earlier, clustering is the most challenging part. We briefly review the clustering techniques used in previous models and clarify their problems. In [13], the authors proposed a real-time (online) diarization system based on DOA features. They employed a simple online clustering technique called leader-follower clustering [20]. This algorithm is simple and fast, but has an apparent drawback in that the clustering result strongly depends on a time-independent threshold that is difficult to optimize *a priori*. In [14], Bayesian information criteria (BIC) is used to find the optimal number of clusters. The algorithm finds the optimal number by comparing the BIC scores before and after merging any two clusters. It requires the computation of BIC improvements for every possible merge, so the computation load increases exponentially.

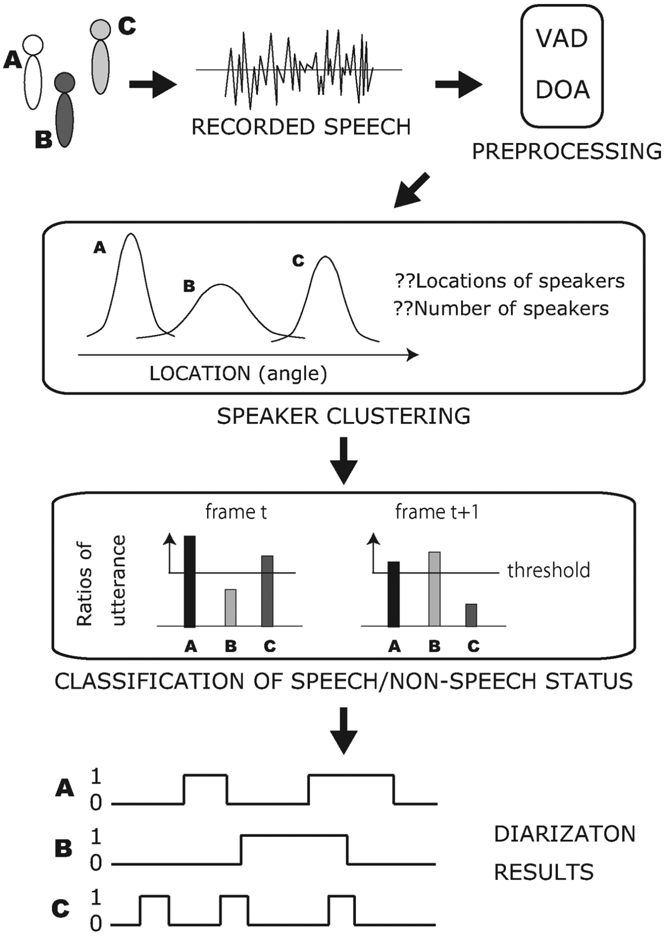

Fig. 1. Diarization system overview.

In this paper, we seek yet another interpretation of the speaker clustering problem, which naturally captures the nature of meeting conversations, based on speaker location information.

III. Naive Modeling by Mixture Models

In this paper, we study three mixture-based models: GMM, LDA, and dLDA. First, we examine GMM and LDA, which are less complicated than the dLDA model. We introduce the probabilistic representations of these two, and study the weak points of these models. In the next section, we introduce dLDA to overcome the problems.

A. *Bag-of-Words Representations*

Throughout this paper, we adopt the BoW feature representation. BoW approaches first appeared in natural language processing and information retrieval research. BoW is a simple representation of documents, each of which consists of many words, that can handle high (thousands) cardinality orders. We consider a dataset that is a collection of documents. BoW representation describes a document by a histogram of words: aggregate of the counts of word appearance regardless of the order of words in the sentences. Identifying the word histogram (or a distribution), we can easily analyze the contents of a document, and this is useful for document handling such as search and abstraction. Though this representation is very simple, BoW is known as a powerful technique for document analysis. These days, other research fields such as computer vision (e.g., [21]–[25]), and very recently audio and acoustics processing [26], employ BoW combined with the topic model, which we will discuss later in this section.

In this paper, we work on the BoW representation for speaker clustering and diarization based on speaker location information. A conversation recording is a collection of time frame-wise observations. Every time frame contains a set of information about multiple speakers' locations. We interpret this data structure as a variant of BoW representation. A series of consecutive time frames (possibly one time frame) is considered as a document. Every document consists of a set of location information; therefore, speaker location information serves as a word in documents. We analyze the location word histogram and try to reveal "who spoke" in the given document. Repeating this analysis over all documents, we can identify "who spoke when" from a conversation recording.

BoW is, in essence, discrete feature representation: counts of word tokens. Therefore, we first quantize the original features into discrete tokens, which serve as "words." If we quantize a visual feature such as SIFT [27], the quantized features are called "visual words." Similarly, we consider "angle words" in this paper.

The audio feature used in this paper is DOA. We assume meeting situations in which the speakers remain seated around a table, i.e., do not move during the meeting. We put an array of three microphones on the table and represent the locations of the speakers by angles from a reference direction. DOA features are $D = 360$ dimensional power vectors:

$$\boldsymbol{f}_t = (f_{t,-179}, \dots, f_{t,180}). \tag{1}$$

Each $f_{t,d}$ denotes the estimated signal power from direction (angle) $d(\mathrm{deg})$ at time t. For details, please see the Appendix.

In accordance with the BoW representation, we convert DOA feature $\boldsymbol{f}_t$ into a set of discrete words $x_t = \{x_{t,i} \in R^1\}$ as in Fig. 2. From each $d \in \{-179, \dots, 180\}$, multiple samples $x_{t,i} = d/360$ are reproduced. The division by $360(deg)$ is for normalization purposes. The number of words, $n_{t,d}$, is proportional to $f_{t,d}$. Going through this conversion from $d = -179$ to $d = 180$, we have

$$x_t = \{x_{t,i}\}, \quad i = 1, 2, \dots, n_t, \tag{2}$$

$$n_t = \sum_d n_{t,d}, n_{t,d} \propto f_{t,d}. \tag{3}$$

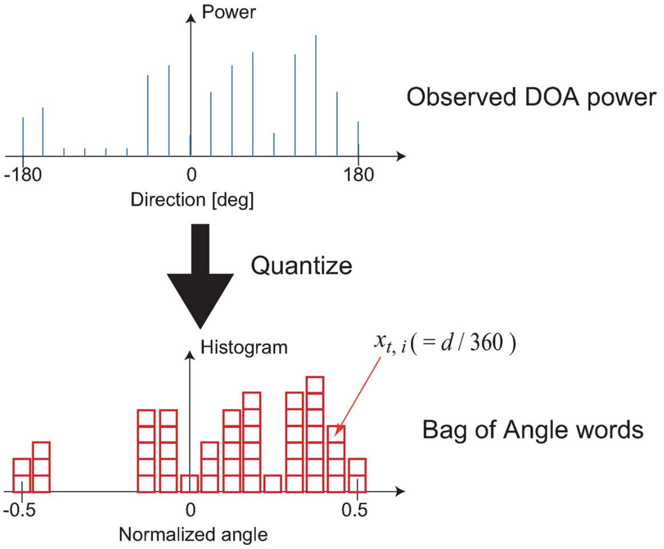

Fig. 2. Bag-of-Angle words representation of observations. We quantize the observed DOA powers into discrete "word" samples $x_{t,i}$.

Now we have angle "words." Next, we define a document. In this paper, we merge some consecutive time frames as one document (frame document). Each document has its own word histogram, which summarizes all angle word occurrences in the time frames within the document. The merging process serves to smooth document-word distributions (the histogram) among documents and to aid computational efficiency.

Finally, each frame document is represented by "Bag-of-Angle words." Hereafter, we slightly abuse notations. An index t refers to the time index of the frame documents. We have n_t "angle words" for the frame document t. Each word (a sample) $x_{t,i}$ denotes the angle of the speech fragment. The distribution (or, histogram) of $x_{t,i}$ reflects the power-oriented distribution at time t. Because human voices have large power (large $f_{t,d}$), a sample concentration indicates the location of a speaker. Therefore, we can estimate the locations of the speakers by clustering these samples, namely angle words.

B. *Gaussian Mixture Model*

Standard mixture models are among the most popular statistical models for clustering, especially the GMM. Araki *et al.* [28] studied the use of GMM for a batch (offline) speaker clustering scenario, which is not our goal. We explain how GMM can be used in a fully probabilistic manner for online speaker clustering for BoW representation.

In this paper, speaker clustering is understood as the online clustering of observed sample set (BoW) $X_{1:t} = \{x_1, \dots, x_t\}$. We hope to partition $x_{t,i}$ into clusters. Each cluster corresponds to a speaker, and latent parameter θ represents the location of the speaker. Samples $x_{t,i}$ are generated from observation function $F(\theta)$. We would like to infer the number of speakers, and their locations. This is equivalent to finding the optimal number of clusters and their parameters θ, and assigning $x_{t,i}$ to each cluster.

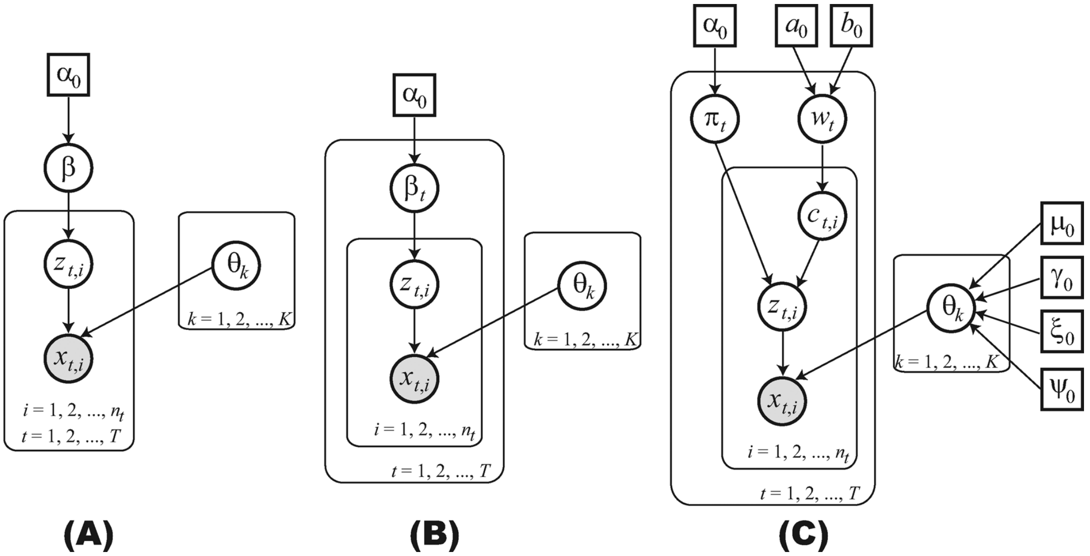

Fig. 3. Graphic models of the three mixture models. Circle nodes are random variables and square nodes are the constants. Shaded nodes denote the observed values, and rounded rectangular "plates" are the repetitions. Arcs indicate the probabilistic dependencies between variables. (A) GMM, (B) LDA, and (C) dLDA.

It may be that this process can be tackled by an ordinary GMM. Usually, BoW indicates that the observation distribution is a multinomial distribution [16]. This is because words (natural language words) are discrete symbols. However, each angle-word $x_{t,i}$ is a continuous real value. Therefore, in this paper we adopt the Gaussian prior as the distribution of $x_{t,i}$, which makes the model a GMM with BoW (angle words). Each cluster corresponds to a Gaussian parameterized by θ. We assume K mixtures of Gaussians, and formulate the distribution of $x_{t,i}$ as follows:

$$p(x_{t,i}) = \sum_{k=1}^{K} \beta_k \mathrm{N}(x; \theta_k) \tag{4}$$

where $\theta_k = (\mu_k, \sigma_k)$ is the parameter of the kth cluster. μ_k is the mean, and σ_k is the variance of the kth Gaussian. β_k is the mixing ratio of $\sum_k \beta_k = 1$.

For simplicity and clarity of later discussions, we describe GMM in another but equivalent way. We introduce a hidden assignment variable $z_{t,i} = k \in \{1, 2, \ldots, K\}$. $z_{t,i}$ is a mixture assignment variable for observed word $x_{t,i}$; it indicates the cluster that generated $x_{t,i}$. We assume $z_{t,i}$ is a K-dimensional binary vector, whose elements equal zero except for the element of 1. Without confusion, we write $z_{t,i} = k$, meaning the kth element of the vector $z_{t,i,k} = 1$; the others are zero.

$z_{t,i}$ is a random variable that distributes proportional to the mixing ratio vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$.

Also assuming the prior distribution on $\boldsymbol{\beta}$, we have the following fully probabilistic GMM model for speaker clustering:

$$p(\boldsymbol{\beta}) = \text{Dirichlet}(\boldsymbol{\alpha}) \tag{5}$$

$$p(z_{t,i}|\boldsymbol{\beta}) = \text{Multinomial}(\boldsymbol{\beta}) \tag{6}$$

$$p(x_{t,i}|z_{t,i}, \{\theta_k\}) = \mathrm{N}\left(\theta_{z_{t,i}}\right). \tag{7}$$

$\boldsymbol{\alpha}$ is a K-dimensional hyperparameter for the Dirichlet distribution. The graphical model of the above GMM is shown in Fig. 3(a). In the figures, circle nodes represent the random variables, and square nodes indicate constants. The shaded nodes are the observed data (angle words), and rounded rectangular "plates" denote repetition with respect to the indices below.

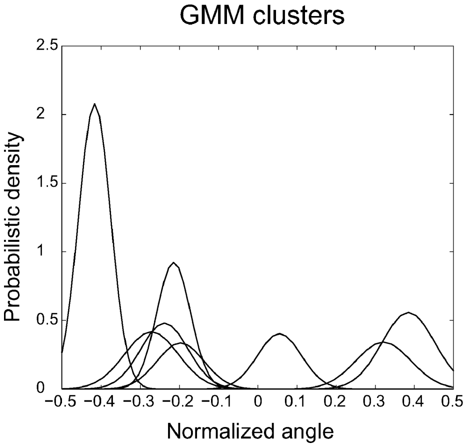

Fig. 4. Result of speaker cluster estimation by GMM model (ground truth: four speakers). Each cluster (Gaussian) is assumed to be a speaker. The horizontal axis is the angle value (normalized from −0.5 to 0.5).

This simple GMM set up is easy to understand, and seems well suited to the speaker clustering problem: running this model includes inferring assignments $z_{t,i}$, which indicate the clustering of the observations into K mixtures. Given these clustering assignments, we can easily infer the parameter of the kth speaker cluster (mixture) parameter θ_k and its mixing ratio β_k.

Now we present one part of the results of the experiments for better understanding. We build the above model, and estimate all unknown variables ($z_{t,i}$, β_k, and θ_k) by Bayesian inference (discussed in the "Inference" section). Inference is performed in an online manner: we iterate the inference process every time the bag-of-angle words of x_t, a new frame document, is received. Fig. 4 shows the final results of the GMM estimation for four-speaker conversation data. As you can see, GMM yields many more clusters than there are speakers (four); it fails to estimate the number of speakers (clusters).

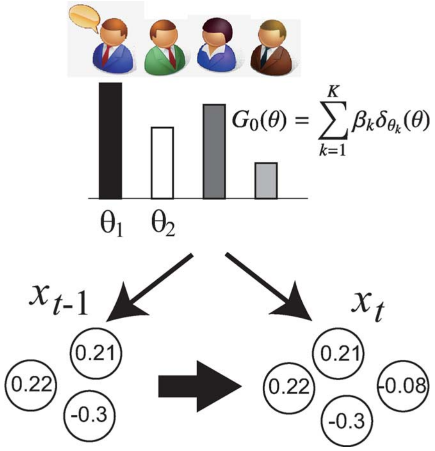
$G_0(\theta) = \sum_{k=1}^{K} \beta_k \delta_{\theta_k}(\theta)$
θ1
θ2
$x_{t\text{-}1}$
x_t
0.21
0.22
-0.3
0.21
0.22
-0.08
-0.3

Fig. 5. Illustration of GMM model for speaker clustering in diarization tasks. Time-invariant mixture G_0 is assumed to represent time-varying observation x_t, which reflects turn-taking.

This result is disappointing but understandable because a simple GMM cannot deal with the dynamics of conversations. Choosing GMM for speaker clustering is equivalent to assuming a time-invariant mixture model in the parameter space:

$$G_0(\theta) = \sum_{k=1}^{K} \beta_k \delta_{\theta_k}(\theta) \tag{8}$$

where δ_{θ_k} is a delta function that peaks at θ_k. Since each parameter corresponds to a speaker, GMM assumes a time-invariant speaker mixture for any conversation. We can see this in Fig. 3(a) as we have only one $\boldsymbol{\beta}$ node, located outside of the repetition plate. However, the actual distribution at each time is time-varying because of "turn-taking," i.e., the number of speakers who produce speech at time t is not constant. Suppose we have a conversation of three speakers, and at time t only one person speaks. The distribution at t differs from the time-invariant distribution, which suits three speakers' observations. We illustrate this problem in Fig. 5.

Hereafter, we use the term "turn-taking" with a more general meaning. By turn-taking, we mean the changes of speaker proportions along the timeline caused by a speaker's beginning and end of one utterance. These changes will be induced not only by speaker changes (strict meaning of turn-taking), but also by an intermittent monologue that contains several switches between silent and speech periods.

Next, we extend a simple GMM to a more powerful probabilistic model, namely LDA.

C. Latent Dirichlet Allocation

We can instead employ the Latent Dirichlet Allocation (LDA) model [16], which is a famous BoW representation model originally designed for text data. In the LDA model, we have a set of K latent mixture components called "topics." Each topic is characterized by its own word distribution. A document is represented as a mixture of topics, and each observed word in the document is generated from one of the topics.

This LDA scheme is readily applicable to our speaker clustering problem. Topic components correspond to speakers represented as Gaussians with location parameters θ, time intervals t to documents, and samples $x_{t,i}$ to words. Analogous to the GMM model [(5)–(7)], we can formalize a speaker clustering model with LDA as follows:

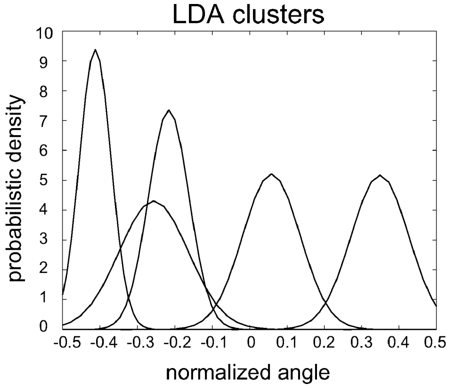

Fig. 6. Result of speaker cluster estimation by LDA model (ground truth: four speakers). Each cluster (Gaussian) is assumed as a speaker. The horizontal axis is the angle value (normalized from -0.5 to 0.5).

$$p(\boldsymbol{\beta}_t) = \text{Dirichlet}(\boldsymbol{\alpha}) \tag{9}$$

$$p(z_{t,i}|\boldsymbol{\beta}_t) = \text{Multinomial}(\boldsymbol{\beta}_t) \tag{10}$$

$$p(x_{t,i}|z_{t,i}, \{\theta_k\}) = \text{N}\left(\theta_{z_{t,i}}\right). \tag{11}$$

The graphical model of LDA is shown in Fig. 3(b). The main difference from a simple GMM is that the mixing ratios of topics (clusters) are time-variant. The mixing ratio vector at the tth frame document is denoted as $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{t,K})$ in (9) ($\sum_k \beta_{t,k} = 1$). $\boldsymbol{\beta}_t$ is located inside of the t-plate in Fig. 3(b). This indicates that $\boldsymbol{\beta}_t$ is variable over time; therefore, we can freely represent time-varying speaker mixtures due to turn-taking. Equation (12) represents the way LDA models the parameter mixture behind the conversation observation:

$$G_t(\theta) = \sum_k \beta_{t,k} \delta_{\theta_k}(\theta). \tag{12}$$

In the above equation, the mixture is defined by time-varying mixing ratio $\beta_{t,k}$. This means that LDA allows the parameter mixture to change over time, and this feature enables LDA to treat turn-taking during conversations in a more reasonable and natural way.

As in the discussion of GMM, we present the results of an experiment on LDA for elucidation. The same four-speaker dataset was tested by LDA, and the result is presented in Fig. 6. As can be seen, LDA performs better than GMM in terms of the estimated number of clusters $=$ speakers. However, it is also obvious that modeling with LDA still has a problem. This is due to the assumption of G_t independence in (12). This can be observed from (9) where every $\boldsymbol{\beta}_t$ is sampled independent and identically distributed (i.i.d.) from the common Dirichlet. This does not suit the speaker clustering task, since the speaker topic mixtures at t and at $t+1$ are indeed correlated and similar to each other. We illustrate the idea and the problem of LDA in Fig. 7.

Originated in natural language processing, LDA usually assumes multinomial distribution for observed words. Please note that we again adopt a Gaussian distribution for angle-words for the same reason as in the GMM case.

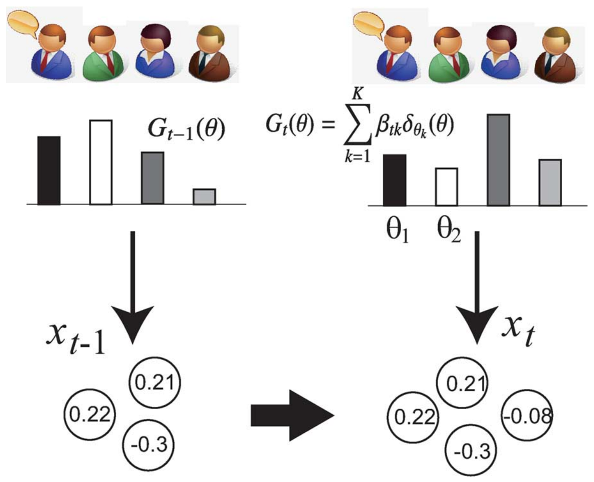

Fig. 7. Illustration of LDA model for speaker clustering in diarization tasks. Time-varying mixtures G_ts are introduced, but they are completely independent of each other, even between successive frame documents.

IV. Dynamic LDA Model

A. Idea of Dynamic LDA

From the above discussions, we conclude that we need a new mixture model that can represent both of the following points:

1) time-varying speaker mixtures;
2) correlation between successive frame documents.

For that purpose, we adopt a new probabilistic model called dynamic Latent Dirichlet Allocation (dLDA). dLDA was first introduced in [15] to capture the time evolution of the data properties of documents, but it is applicable to many time-series data in other domains. We found that the idea and formulation of dLDA is surprisingly well-suited to our speaker clustering problem in diarization tasks. This article is, to the best of our knowledge, the first attempt to utilize dLDA in speaker diarization.

Its idea is straightforward: dLDA sets a Markov property between the distributions of two consecutive time steps. This property is described as follows:

$$H_t(\theta) = \sum_{k=1}^{K} \pi_{t,k}\delta_{\theta_k}(\theta) \tag{13}$$

$$p(w_t) = \text{Beta}(a_0, b_0) \;\; \text{for } t > 1, \quad w_1 = 1 \tag{14}$$

$$G_t(\theta) = (1 - w_t)\, G_{t-1}(\theta) + w_t H_t(\theta) = \sum_{k=1}^{K} \beta_{t,k}\delta_{\theta_k}\theta \tag{15}$$

$$= \sum_{l=1}^{t} \left\{ \prod_{m=l+1}^{t} (1 - w_m) \right\} w_l H_l(\theta) = \sum_{l=1}^{t} v_{t,l} H_l(\theta). \tag{16}$$

As in the case of LDA [(12)], $G_t(\theta)$ is a mixture of speaker parameters that reflect the actual utterances at time t. The key point of (15) is that $G_t(\theta)$ is constructed as a linear interpolation of the previous distribution $G_{t-1}(\theta)$ and a newly introduced "innovation distribution" $H_t(\theta)$, which is responsible for the change in the sample distribution between $t-1$ and t. $H_t(\theta)$ is parameterized by the time-varying mixing vector $\boldsymbol{\pi}_t = (\pi_{t,1}, \ldots, \pi_{t,K})$, whose sum of elements equals one. In this way, we can hold both requirements for the mixture model to solve speaker clustering. This idea is illustrated in Fig. 8.

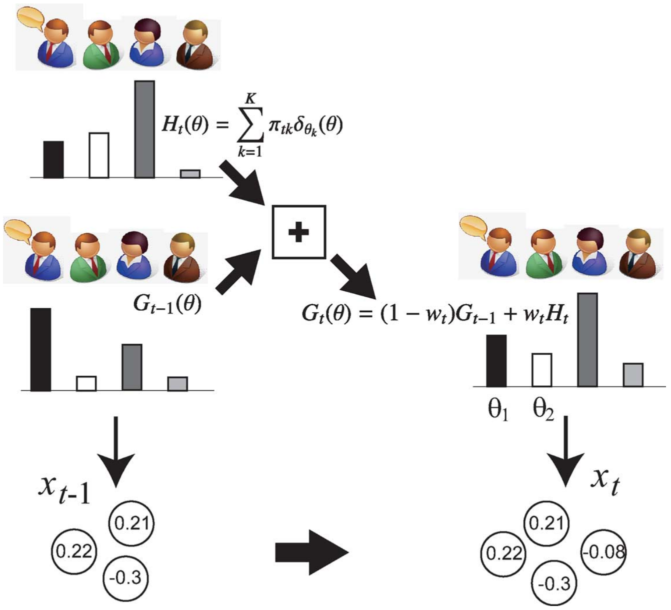
$H_t(\theta) = \sum_{k=1}^{K} \pi_{tk}\delta_{\theta_k}(\theta)$
+
$G_{t-1}(\theta)$
$G_t(\theta) = (1 - w_t)G_{t-1} + w_t H_t$
θ_1 θ_2
$x_{t\text{-}1}$
x_t
0.21
0.22
-0.3
0.21
0.22
-0.08
-0.3

Fig. 8. Illustration of dLDA model for speaker clustering in diarization tasks. Markov dependency is assumed for G_t, enabling smooth speaker mixture changes between frame documents to deal with turn-taking.

Also, (16) gives us another interpretation of $G_t(\theta)$: the current speaker mixture in the parameter space is in fact a mixture of independent innovation distributions. This interpretation is helpful in solving the model.

It is worth noting that w_t, an interpolation factor, is a probabilistic random variable sampled from Beta distribution (14) and that regulating w_t allows the dLDA model to handle irregular distribution changes, ranging from small to significant changes. This indicates that dLDA is extremely flexible in representing the time-varying correlation between successive time steps.

It is beneficial to consider two extreme cases as follows. If we set $w_t = 0$ for all t, then all $G_t(\theta)$ is equivalent to $G_0(\theta)$ and the dLDA model is reduced to the simple GMM [(8)]. On the other hand, setting $w_t = 1$ for all t makes all time steps independent of each other. In this case, we recover the latent Dirichlet allocation (LDA) model [(12)]. Therefore, dLDA naturally generalizes GMM and LDA.

B. Generative Model of Dynamic LDA

The generative model of dLDA is shown as follows:

$$p(\boldsymbol{\pi}_t) = \text{Dirichlet}\left(\frac{\alpha_0}{K}, \frac{\alpha_0}{K}, \ldots, \frac{\alpha_0}{K}\right) \tag{17}$$

$$p(w_t) = \text{Beta}(a_0, b_0) \;\; \text{for } t > 1, \; w_1 = 1 \tag{18}$$

$$v_{t,l} = w_l \prod_{m=l+1}^{t} (1 - w_m) \tag{19}$$

$$p(c_{t,i} \mid \boldsymbol{v}_t) = \text{Multinomial}(\boldsymbol{v}_t) \tag{20}$$

$$p(z_{t,i} \mid \{c_{t,i}\}, \boldsymbol{\pi}_t) = \text{Multinomial}\left(\boldsymbol{\pi}_{c_{t,i}}\right) \tag{21}$$

$$p(\theta_k) = \text{NormalGamma}(\mu_0, \gamma_0, \xi_0, \psi_0). \tag{22}$$

$$p(x_{t,i} \mid z_{t,i}, \{\theta_k\}) = \text{N}\left(\theta_{z_{t,i}}\right). \tag{23}$$

The graphical model of the above dLDA is shown in Fig. 3(c).

```
VB inference of dLDA model
$t = 1$;
while $x_t$ exists
    input $x_t$;
    load $x_1 \sim x_t$;
    load variables to be estimated up to $t$;
    for $j = 1$; $j < N_{VB}$; $++j$
        Phase I computation using Eqs. (36-39);
        Phase II computation using Eqs. (40-43);
    end
    compute the variational posterior using Eqs. (24-28);
    output the estimated results;
    $t = t + 1$;
end
```

Fig. 9. Pseudo-algorithm of a simple online VB inference of dLDA. N_{VB} is a pre-defined constant, number of iterations.

In (17), the probabilistic distribution of mixing ratio vector $\boldsymbol{\pi}_t$ for H_t, not $\boldsymbol{\beta}_t$ for G_t, is defined as a Dirichlet distribution. α_0 is a hyperparameter for the Dirichlet. It is easier to work with $\boldsymbol{v}_t$ for inference than to consider the temporal mixing ratio $\beta_{t,k}$ directly since H_t are conditionally independent of each other.

Next, the interpolation factor w_t in (18) is defined as we have discussed. a_0, b_0 are the hyperparameters for w_t. Given w_t, we compute $v_{t,l}$ ($l = 1, \ldots, t$) as in (19) to move from the past mixtures' information to that of the current mixture.

To generate the observed angle word $x_{t,i}$, first we pick an index, $c_{t,i} = l$, meaning that the lth innovation distribution H_l generates the angle word $x_{t,i}$ based on (20). Using $\pi_{c_{t,i}}$, we choose a cluster index $z_{t,i} = k$, which means that the cluster with parameter θ_k is responsible for generating $x_{t,i}$. In other words, $z_{t,i}$ is an index of the speaker that produces the sound heard from the direction of $x_{t,i}$ at time t [(21)]. This model can represent voice overlap situations naturally because $z_{t,i}$ may take a different value for each i independently. In (22) we sample K parameters for clusters from an appropriate prior distribution. Finally, $x_{t,i}$ is sampled from the normal observation distribution with picked parameter θ_k. Note that we employ a normal distribution instead of the multinomial distribution used in the original dLDA [15] because of the angle-word features.

As discussed, we assume each cluster corresponds to a speaker, and its parameter represents the speaker location. Thus, the choice of parameter prior $p(\theta)$ is important in obtaining good results. In this paper, we assume a one-dimensional Gaussian distribution for the actual angle words. Its parameter θ_k consists of mean m_k and precision (inverse of variance) σ^{-2}. Therefore, priors for these parameters are chosen from a Normal–Gamma distribution [29] to maintain conjugacy. $\mu_0, \gamma_0, \xi_0, \psi_0$ are the hyperparameters. If we assume a multi-dimensional Gaussian for the observation distribution, then the parameter prior is chosen to be Normal–InverseWishart [29]. Because of conjugacy, the dLDA inference is fast and reliable as discussed in the next section.

C. Some Notes and Related Works

Since the dLDA model is fully formulated in a probabilistic generative model, it can not only infer the number of clusters (speakers), but also flexibly estimate the time evolution of the parameter (sample) distributions from an observed sequence. These elements are subject to the Dirichlet prior in (17). If we set K large enough, we will have an appropriate number of "effective" clusters that have a large mixing ratio, and the mixing ratios of the other clusters will become negligibly small, i.e., the number of "effective" clusters (speakers) and their mixing ratios are automatically estimated. Values of the parameters θ_k are also fully probabilistically modeled, and are also optimized automatically in a Bayesian manner.

There are two works closely related to dLDA. Both are strongly formulated in a probabilistic (Bayesian) manner; thus, it may be fruitful to discuss the difference between these models.

dLDA can be extended to a dynamic Hierarchical Dirichlet Process (dHDP) [30] by taking the limit of $K \rightarrow \infty$, where the finite Dirichlet prior is replaced by an infinite stick-breaking prior. In their paper, dHDP is used for segmentation of a music piece. However, it is difficult to derive a VB inference algorithm for dHDP, which is preferable for (future) real-time systems.

Another closely related model was recently proposed by Fox *et al.* [10]. The model formulates an infinite number of speakers and time-dependent turn-taking in a form of extended HMM. In the aspect of model construction, their model differs from dLDA in the assumption of dynamics in sequences. Since the HMM allows discrete state jumps, their model suits time-series sequence with many abrupt structural changes. The dLDA model, on the other hand, prefers smoother changes able to be captured by linear interpolation [(15)] with the dependencies controlled by w_t [(14)].

One main difference in terms of practicality is that their model does not allow overlapping of speakers, while dLDA does. This is because the hidden states of extended HMM correspond to "speaker A speaks," "speaker B speaks," and so forth. On the other hand, our dLDA model estimates the mixture of K clusters at each time frame. Thus, dLDA is able to represent overlapping of speakers in a more natural way than Fox's model.

V. Inference

A. Inference via Variational Bayes

There are two major approaches to solving the probabilistic model. Gibbs sampling is accurate but slow in terms of convergence, while variational Bayes (VB) [29], [31] is fast but may be trapped in locally optimal solutions. We prefer VB inference for online speaker diarization.

In this paper, we develop an online and incremental inference by VB for dLDA. VB estimates the variational (approximated) posterior $q(\cdot)$ of the hidden variables and parameters. Since a VB solution for a Bayes model is just an approximation, it may be trapped by local optima. However, it provides fast iterative computation, which is important when applying a complex probabilistic model to real applications. The performance of VB is better than that of point estimation methods such as maximum-likelihood (ML) and maximum *a posteriori* (MAP) solutions in many cases. In signal/sound processing research, [31], [32] are pioneer works on the use of VB.

We first describe the derivation of the VB solution for the dLDA model [15]. VB posteriors $q(\cdot)$ are approximated posteriors in the sense that all posteriors are independent of other variables since the true posteriors are not. Because of the conjugacy incorporated in the generative model (17)–(23), we expect the following forms of the variational posteriors:

$$\begin{aligned} q^*(\boldsymbol{\pi}_t) = \text{Dirichlet}\Bigg(&\frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{m,i,1}s_{m,i,t},\\ &\ldots, \frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{m,i,k}s_{m,i,t},\\ &\ldots, \frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{m,i,k}s_{m,i,t},\Bigg) \end{aligned} \tag{24}$$

$$q^*(w_t) = \text{Beta}\left(a_0 + \sum_{i=1}^{n_t} s_{t,i,t}, b_0 + \sum_{i=1}^{n_t}\sum_{m=1}^{t-1} s_{t,i,m}\right) \tag{25}$$

$$q^*(c_{t,i}) = \text{Multinomial}\left(s_{t,i,1}, \ldots, s_{t,i,l}, \ldots, s_{t,i,t}\right) \tag{26}$$

$$q^*(z_{t,i}) = \text{Multinomial}(r_{t,i,1}, \ldots, r_{t,i,k}, \ldots, r_{t,i,K}) \tag{27}$$

$$q^*(\theta_k) = \text{NormalGamma}\left(\mu_1, \gamma_1, \xi_1, \psi_1\right) \tag{28}$$

where

$$\mu_1 = \frac{\gamma_0\mu_0 + N_k\tilde{x}_k}{\gamma_0 + N_k} \tag{29}$$

$$\gamma_1 = \gamma_0 + N_k \tag{30}$$

$$\xi_1 = \xi_0 + \frac{N_k}{2} \tag{31}$$

$$\psi_1 = \psi_0 + \frac{\tilde{S}}{2} + \frac{\gamma_0 N_k}{\gamma_0 + N_k}\frac{(\tilde{x}_k - \mu_0)^2}{2} \tag{32}$$

$$N_k = \sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{t,i,k} \tag{33}$$

$$\tilde{x}_k = \frac{1}{N_k}\sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{t,i,k}x_{t,i} \tag{34}$$

$$\tilde{S}_k = \sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{t,i,k}\left(x_{t,i} - \tilde{x}_k\right)^2. \tag{35}$$

It is easy to derive these variational posteriors if you follow the textbook [29]. The objective of the inference is to estimate all of the parameters $s_{t,i,l}, r_{t,i,k}$ to represent the posteriors above.

1) Phase I—Computing Variational Expectations: The inference process consists of two phases that are iteratively computed as in the standard EM algorithm. The first phase computes expectations of several hidden variables over the variational posteriors.

Let us denote $\mathbb{E}_x[f]$ as the expectation of function f over the variational posterior of variable x. Then, in phase I, we compute the following expectations:

$$\begin{aligned} \mathbb{E}_{w_t}[\log w_t] = \psi&\left(a_0 + \sum_{i=1}^{n_t} s_{t,i,t}\right)\\ &-\psi\left(a_0 + b_0 + \sum_{i=1}^{n_t}\sum_{m=1}^{t} s_{t,i,m}\right) \end{aligned} \tag{36}$$

where $\mathbb{E}_{w_1}[\log w_1] = 0$:

$$\begin{aligned} \mathbb{E}_{w_t}[\log(1-w_t)] = \psi&\left(b_0 + \sum_{i=1}^{n_t}\sum_{m=1}^{t-1} s_{t,i,m}\right)\\ &-\psi\left(a_0 + b_0 + \sum_{i=1}^{n_t}\sum_{m=1}^{t} s_{t,i,m}\right) \end{aligned} \tag{37}$$

$$\begin{aligned} \mathbb{E}_{\pi_{t,k}}[\log \pi_{t,k}] = \psi&\left(\frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{m,i,k}s_{m,i,t}\right)\\ &-\psi\left(\alpha_0 + \sum_{k=1}^{K}\sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{m,i,k}s_{m,i,t}\right) \end{aligned} \tag{38}$$

$$\begin{aligned} \mathbb{E}_{\theta_k}[\log p\left(x_{t,i}\,|\,\theta_k\right)] = &-\frac{1}{2}\log(2\pi) + \frac{1}{2}\psi(\xi_1) - \frac{1}{2}\log(\psi_1)\\ &-\frac{1}{2}\gamma_1^{-1} - \frac{1}{2}\frac{\xi_1}{\psi_1}\left(x_{t,i} - \mu_1\right)^2. \end{aligned} \tag{39}$$

Note ψ indicates the digamma function.

2) Phase II—Computing the Posterior Parameters: In phase II, we compute $s_{t,i,l}$s and $r_{t,i,k}$s to estimate the parameters of variational posteriors in (24)–(28).

It is obvious that these quantities indicate the average probabilities of 1) $s_{t,i,l}$: picking the innovation measure H_l and 2) $r_{t,i,k}$: picking cluster k to sample $x_{t,i}$ since these parameters appear as they are in (26) and (27), respectively.

We can compute these quantities by the equations shown as follows:

$$\begin{aligned} \sigma_{t,i,l} = &\mathbb{E}_{w_l}[\log w_l] + \sum_{m=l+1}^{t}\mathbb{E}_{w_m}[\log(1-w_m)]\\ &+\sum_{k=1}^{K} r_{t,i,k}\mathbb{E}_{\pi_{l,k}}[\log \pi_{l,k}] \end{aligned} \tag{40}$$

$$s_{t,i,l} = \frac{\exp(\sigma_{t,i,l})}{\sum_{m=1}^{t}\exp(\sigma_{t,i,m})} \tag{41}$$

$$\rho_{t,i,k} = \mathbb{E}_{\theta_k}[\log p\left(x_{t,i}|\theta_k\right)] + \sum_{l=1}^{t} s_{t,i,l}\mathbb{E}_{\pi_{l,k}}[\log \pi_{l,k}] \tag{42}$$

$$r_{t,i,k} = \frac{\exp(\rho_{t,i,k})}{\sum_{j=1}^{K}\exp(\rho_{t,i,j})}. \tag{43}$$

B. Estimating the Mixing Ratios and the Number of Speakers

We describe how to estimate the mixing ratios and the number of speakers. First, we estimate the expected number of samples assigned to cluster k (a speaker) at time t as

$$||z_{t,k}|| = \sum_{i=1}^{n_t} r_{t,i,k}. \tag{44}$$

Second, we use $||z_{t,k}||$ to determine the posterior estimates of temporal mixing ratios $\hat{\beta}_{tk}$ and global mixing ratio $\hat{\beta}_k$ as follows:

$$\hat{\beta}_{tk} = \frac{||z_{t,k}||}{\sum_{k=1}^{K}||z_{t,k}||}, \quad \hat{\beta}_k = \frac{\sum_t ||z_{t,k}||}{\sum_{k=1}^{K}\sum_t ||z_{t,k}||}. \tag{45}$$

We count the number of "effective clusters" (speakers) as those that have mixing ratio $\hat{\beta}_k$ larger than chance level $1/K$. In many cases, the mixing ratios of other minor clusters are negligible.

C. Techniques for Fast Online VB Processing

Inherently, the computational cost of the inference process grows with the time step, since the number of variables to be estimated and the samples to be stored keep increasing. Therefore, we developed several techniques to accelerate online dLDA inference. Estimation precision can be well balanced against computation costs.

1) Limiting Re-Estimated Variables: One of the most commonly used techniques for adaptive estimation is to re-estimate (update) only the variables of the most recent time steps. We expect that the hidden variables or parameters of older time steps have converged to stable values because they have been repeatedly updated by VB.

Thus, we minimize computation cost by limiting the variables to be re-estimated. Equations (36)–(43) are updated only for the variables of the most recent $W_1(\ll t)$ steps. This reduces the computational cost to W_1/T, where T is the total number of time steps.

2) Limiting Data Used in Inferences: The above trick reduces the number of variables that need to be re-estimated. However, inference of each variable requires all currently available observations and estimation results. Thus, the computational cost basically increases with the time step.

We studied the dLDA model and found that this cost could be reduced based on the model characteristics. As discussed, G_t has a time-dependency and the strength of the dependency is controlled by w_t [(18)]. w_t will be close to 1 when the sample distributions change drastically (e.g., speaker turn-taking).

Assuming that the most recent turn-taking is observed at $t = m$, $w_m \approx 1$, and it is easy to see that $v_{t,l} \approx 0, l < m$ in (19). This indicates that the posterior probability of $c_{til} = 1$, namely s_{til}, will be close to zero for $l < m$. In turn, the corresponding computations for VB inference can be eliminated.

Therefore, we can forget those quantities, which greatly reduces the computation cost of each update equation. We retain only the information (estimation results and observations) of the $W_2(\ll t)$ most recent time steps, where t is the time index of the oldest variables to be updated.

3) Limiting Cluster Updates: We expect that the mixing ratio concentrates on fewer clusters K_{eff}. However, all K clusters and their parameters must be maintained during the inference process, including "negligible" $K - K_{\text{eff}}$ clusters.

Thus, we introduce another trick for computational cost saving. At each VB iteration, we update the kth cluster, which has the mixing ratio of $\hat{\beta}_k$, with the probability of $p_{\text{update}}(k)$. We can define $p_{\text{update}}(k)$ in several ways. Our experiments use the following definition:

$$p_{\text{update}}(k) = K \times \min\left(\frac{1}{K}, \hat{\beta}_k\right). \tag{46}$$

This equation indicates that "effective" clusters that have larger mixing ratios than chance levels $1/K$ are updated at each time step, while the other clusters are re-computed according to the mixing ratio.

TABLE I
SPECIFICATIONS OF REAL RECORDED DATA SETS

Data ID	# Speaker	Overlap [%]	# Turn taking	# Utterance
CP1	4	18.6	149	185
CP2	4	13.0	183	218
DC	3	10.8	126	172
CN	3	34.8	243	278

To summarize, we only update the estimation of variational posteriors for the most recent W_1 steps, and retain the estimation results and observations of the most recent W_2 steps. Re-computations of cluster parameters and likelihoods are further confined probabilistically. These techniques drastically reduce the computation cost of dLDA.

D. Inference of Hyperparameters

Thus far, we have discussed how to infer the hidden variables in a Bayesian manner. In previous discussions, hyperparameters such as α_0, a_0 were assumed to be constant. In several cases, however, the initial values of these hyperparameters will affect the estimation of hidden variables. This problem disturbs VB inference because VB is a deterministic iterative procedure that stops at local optima.

To overcome this problem, we infer the hyperparameters in the inference process simultaneously. Since VB is fully Bayesian, we would like to estimate the hyperparameters in a Bayesian manner as well, instead of resorting to heuristics or offline model selection techniques.

One of the most common ways is to put vague priors on hyperparameters and estimate their posteriors by VB as well. Another approach [33] is to reparameterize and sample each hyperparameter in terms of $a \in (0, 1)$, instead of precise posterior evaluations of the original hyperparameter. For example, if hyperparameter α_0 is assumed to be Gamma-distributed, we convert α_0 to $a = \alpha_0/(1 + \alpha_0)$. Sampling a can be achieved from a uniform grid on $(0, 1)$. We estimate (unnormalized) posterior probability densities at several a values and choose one to update the hyperparameter.

In this paper, we select the latter approach because it is more convenient to implement since it is applicable to non-conjugate cases and prevents the VB solutions from being trapped at local optima due to the jumps provided by the hyperparameter sampling.

VI. EXPERIMENTS

A. Data Sets

We collected four data sets of the real recordings gathered in [13]. Their specifications are shown in Table I. We note that the recordings contain much turn-taking and voice-overlapping periods; thus, diarization of these four data sets are challenging. Each data set took the form of a 300-(s) recording that was sampled at 16 (kHz). The frame length was 64 (ms) and the frame shift was 32 (ms).

We also employed the benchmark datasets provided by the AMI project [34]. We used some samples from the IDIAP subset ("IS" meetings). The datasets are used in many studies, such as [35], [36]. Each recording involves four participants engaged in a scenario-based meeting with different durations. Approximately 18% of speech in the recordings overlaps.

We computed angle-word representations, estimated speaker clusters, and inferred who spoke when for all speakers.

B. Quantitative Evaluation: Speaker Diarization Precision

First, we show quantitative performance comparisons. We evaluated the classification performance to verify the entire diarization system (Fig. 1). We employ the diarization error rate (DER) measure [3] for the evaluations. DER is a percentage of the error time length against the total sound-recording length. The error time consists of the following three errors.

1) MST (missed speaker time): the length of time intervals in which the system detected no utterances while a speaker(s) actually spoke.
2) FAT (false alarm speaker time): the length of time intervals in which the system detected speaker utterances while no speaker actually spoke.
3) SET (speaker error time): the length of time intervals over which the system correctly detected the speaker utterances, but failed to associate the utterance events with the correct speaker.

Given these three errors, DER is computed as follows:

$$\mathrm{DER}(\%) = 100 \times \frac{\mathrm{MST} + \mathrm{FAT} + \mathrm{SET}}{\mathrm{TOTAL}} \tag{47}$$

where TOTAL denotes the length of the total speech. The evaluation criteria also follow those provided by NIST [3], that is, the speech segments were split with non-speech periods of more than 300 (ms) in length, and the allowed tolerance for the difference between the system outputs and the correct labels was 250 (ms). The correct speech onset and offset labels for the recorded data were generated manually. All DER values are computed using a script provided by NIST.

We test a simple classification rule based on the posterior of sample assignments $z_{t,i}$ to estimate each speaker's diarization activity (speaks or not). The rule is

$$k \text{ speaks at } t \qquad \text{if} \quad \|z_{t,k}\| > \tau_1 \quad \& \quad \hat{\beta}_{tk} > \tau_2, \\ k \text{ does not speak at } t \qquad \text{otherwise}$$

where τ_1, τ_2 are predefined thresholds. This simple classification rule is based on the one used in [13].

We compare the performance of our proposed model with two previously studied techniques. The first is [13], which is a non-probabilistic online speaker clustering model with the same DOA features. We also evaluated the diarization performance with a method proposed by ICSI [8], which is a state-of-the-art model. The method is based on agglomerative clustering using ΔBIC, where each cluster is modeled with one GMM per feature stream and embedded into an ergodic HMM, and both a speech feature (MFCC) and a location feature (TDOA) are utilized. Please note that [8] is a batch (offline) algorithm; thus, we do not expect the proposed online (incremental) dLDA model to outperform [8] in the DER scores.

TABLE II
DER Measures (%) Achieved in Diarization Experiments. For the First Four Data, "CP2" is Used as a Development Set. For the IS Dataset, "IS1008d" is Used as a Development Set. Development Sets are Marked by Asterisks. Parenthetic DER Values of [8] on the First Four Data Sets Were Computed by Our Implementation: They Might be Worse Than the Original Implementation. We Excerpt the DER Values of [8] on the IS Dataset From the Online Appendix of [36]

Dataset	[13]	[8]	GMM	LDA	dLDA(proposed)
CP1	21.9	(37.1)	55.8	32.7	**21.7**
*CP2	25.0	(35.8)	32.8	24.5	**19.7**
DC	**29.9**	(47.0)	60.6	48.0	31.0
CN	34.3	(56.4)	57.3	48.5	**34.1**
IS1000a	41.9	46.26	35.2	76.9	**32.2**
IS1001a	31.7	30.58	26.7	33.8	**23.7**
IS1001c	32.2	**12.07**	68.2	40.7	27.2
IS1006d	64.3	**54.56**	67.4	69.9	69.7
IS1008a	13.1	**5.13**	77.8	65.3	62.7
IS1008b	19.6	**16.47**	57.8	55.9	23.1
IS1008c	22.6	**12.09**	30.1	30.8	20.4
*IS1008d	15.8	20.83	21.9	32.1	**13.6**

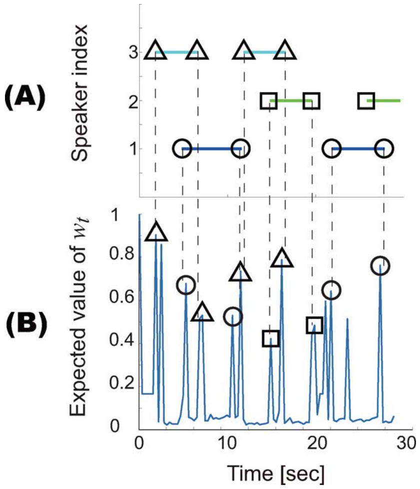

Fig. 10. Turn-taking detection by the trajectory of estimated w_t values in simulation data (best viewed in color). Circles, rectangles, and triangles denote the turn-taking. (A): Ground truth of speaker utterances. (B): Variational expectation of w_t.

For each data collection, we choose one conversation recording as a development set to tune the parameters of [13] and the proposed models. For the proposed models, we tuned parameters related to preprocessing and postprocessing, i.e., parameters for the BoW feature extractions (length of merged consecutive frames for a document, noise threshold for weak power responses, strength of VAD, and map from f_t to n_t) and parameters for the classification (τ_1 and τ_2). Please recall that the model parameters (hidden variables and hyperparameters) of GMM, LDA and dLDA models are automatically estimated by Variational Bayes updates.

Table II summarizes the computed DER measures. Along with [8], [13], we present the DER values of GMM and LDA models. GMM and LDA models are simulated by setting all w_t as $w_t \approx 0$ or $w_t \approx 1$, respectively. As expected, we confirm that DERs of the GMM and the LDA models are much worse than those of dLDA. This is reasonable since the assumptions made in GMM and LDA models are not realistic, ignoring the natural properties of conversations. Compared to previous research, the dLDA model presents comparable or better performance against the non-Bayesian online diarization system of

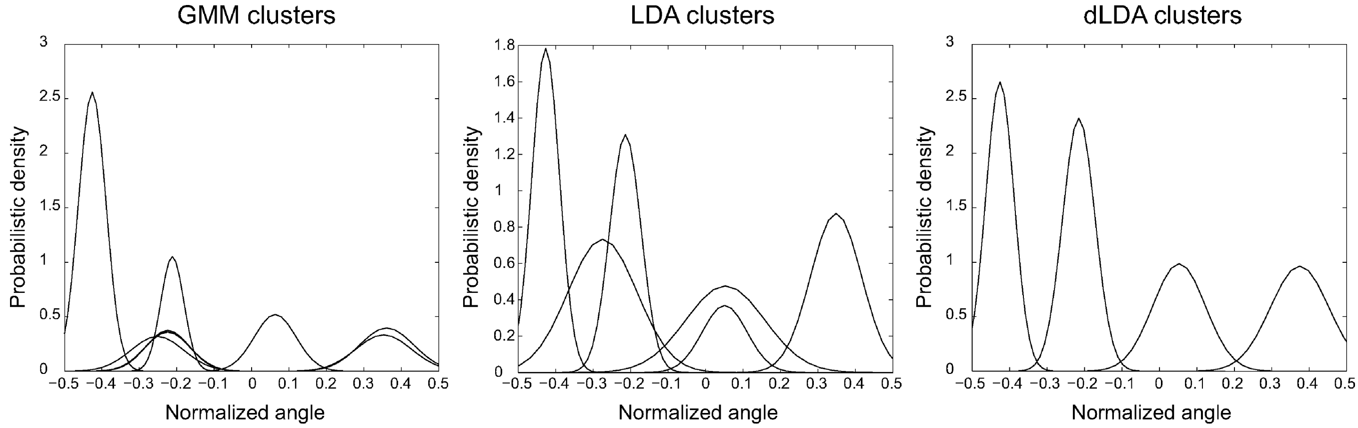

Fig. 11. Clustering results on CP1 real record data set (four speakers). Vertical axis denotes the probability density, and horizontal axis denotes the normalized angle (location). Left: clustering results by GMM. Center: clustering results by LDA. Right: clustering results by dLDA.

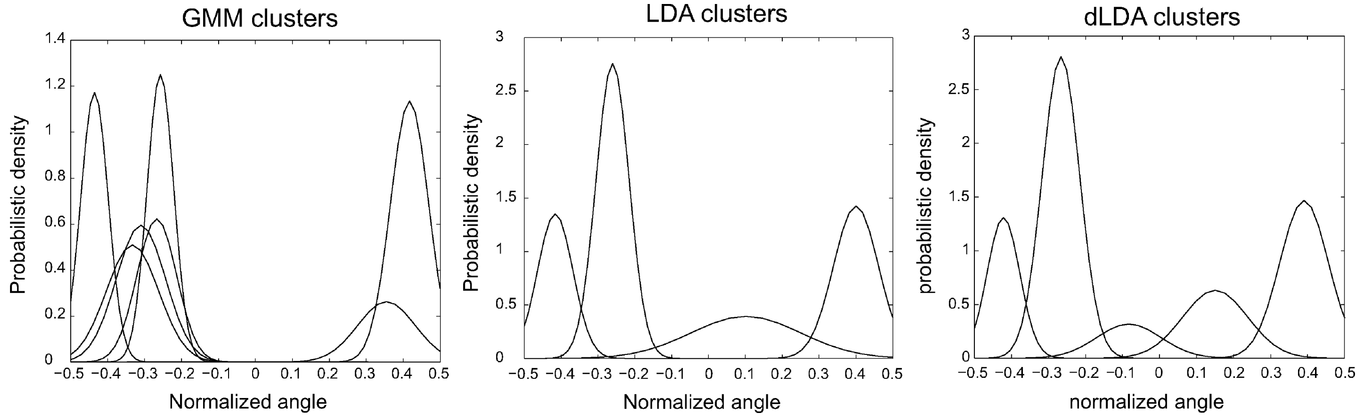

Fig. 12. Clustering results on CN real record data set (three speakers). Vertical axis denotes the probability density, and horizontal axis denotes the normalized angle (location). Left: clustering results by GMM. Center: clustering results by LDA. Right: clustering results by dLDA.

[13] in many recordings. Concerning the comparisons with [8], the DER values of [8] in Table II for the first four data were computed by our implementation. Thus, these values could be worse than the true performance of [8]. For the IS meeting dataset, DER values were taken from the online appendix of [36]. As expected, [8], which is one of the best offline diarization methods, exhibits the best DER values in many IS meeting recordings. However, the proposed dLDA model also marks comparable or better results for a few IS recordings. Considering that dLDA is tested as an online (incremental) model in the experiments, this result is quite promising.

These results provide more understanding about the dLDA model. The main difference between the first four datasets and IS corpus recordings is the duration of each speaker's turn. In IS recordings, each turn of utterance has a longer time duration. On the other hand, the first four datasets are characterized with much shorter durations of turns, and frequent turn-taking. The generative process of turn-taking in dLDA [(14)–(16)] indicates that dLDA is able to estimate and track the changes of speaker distributions for every frame document thanks to w_t. At the same time, the dLDA may overfit if the changes of speaker distributions are not so frequent and drastic, as in the case of IS recordings (with long durations of speech turns). We also note that locations of speakers are not strictly fixed during the recordings for some IS recordings. The dLDA model does not assume the changes of the speaker locations; therefore, we think this model mismatch is one reason for severely degraded performance of dLDA for some recordings (e.g., IS1008a).

C. Qualitative Evaluations

Next, we examine several qualitative results to elucidate the reasons for the good DER scores of dLDA.

1) Turn-Taking Detection by Interpolation Variable: One distinct feature of dLDA is the notion of interpolation factor w_t in (14) and (15). Larger w_t indicates that turn-taking is likely at time t.

Fig. 10 presents estimated w_t for the simulated recording data. The simulated data has 467 frames with 64 (ms) frame shift intervals. Out of 467 frames, 45 frames were excluded in advance as no-speech periods based on VAD. We merged five consecutive frames into a document. The data simulated the conversation of three people with speaker overlapping, convolving English speech signals and measured impulse responses in a room [37].

Colored plots in the upper panel, (A), are the actual speaker utterance periods. Circles, rectangles and triangles denote turn-taking. The lower panel, (B), shows the expectation of variational posterior of w_t. The trajectory is characterized by a few spikes. These spikes indicate large values of w_t, which imply the existence of turn-taking. In fact, we find that most of the (actual) turn-taking matches these spikes in w_t. This result shows that dLDA is able to model and represent turn-taking in the recorded conversation automatically.

2) Speaker Clustering Results: Finally, we tested the performance of dLDA with regard to the speaker clustering tasks, in a comparison against the GMM and the LDA models. After the online clustering of the recorded data up to the last time T, we examined the resultant number of clusters and their parameters.

We present a part of the clustering results from the real recorded data sets in Fig. 11 (CP1), and Fig. 12 (CN). Illustrated normal distributions are "effective" clusters in terms of estimated $\hat{\beta}_k$ in (45).

dLDA clustering yielded better results than either GMM or LDA. We assume that this superiority comes from the ability to model the intermittent changes of speaker (cluster) distributions. The dLDA model could not achieve perfect clustering on the DC and CN data sets: the model produced extra clusters due to noise inputs. However, those "noise" clusters had smaller mixing ratios than the "speaker" clusters. Therefore, we can further improve the clustering of dLDA by carefully refining threshold ($\hat{\beta}_k > 1/K$). Since speaker clustering is performed so precisely, it is no surprise that dLDA achieved good DER values.

Based on these results, we conclude that the dLDA model is able to infer the speaker clusters from speakers' location information. The quantitative results show that the dLDA model especially suits the conversations with frequent turn-taking and shorter speech turn durations.

VII. Conclusion

In this paper, we introduced a new probabilistic model for speaker diarization tasks. We adopted the dynamic LDA (dLDA) model for speaker clustering. The dLDA model automatically infers the number of clusters and data partitioning, and is able to handle time-varying cluster distributions. We proposed a fully probabilistic model for speaker diarization in combination with a bag-of-angle word representation of DOA features. We developed an online inference algorithm based on variational Bayes, and experimentally confirmed that dLDA is able to infer the number of speakers successfully and improves DER performance of online diarization. We also found that the online dLDA recorded good DER values compared to the state-of-the-art offline diarization system.

Some previous studies concerning online diarization proposed using additional cues. For example, [36], [38] proposed a multi-modal (visual + audio) setup for online diarization. Including other features such as MFCC and visual cues is very interesting. We think augmenting other feature models is a promising direction for improving the proposed method.

Appendix

This Appendix explains our methods of calculating the DOA feature and the power vector $\boldsymbol{f}_{t,d}$.

Let $y_j(f,t)$ be a signal observed by microphone j $(j = 1,\ldots,M)$. To calculate the DOA of a sound source active at time t, we first calculate the time difference of arrival (TDOA) between microphone pair $j - j'$ at each time–frequency slot:

$$q'_{jj'}(f,t) = \frac{1}{2\pi f}\arg\left[y_j(f,t)y^*_{j'}(f,t)\right] \tag{48}$$

where * denotes the complex conjugate, and j' is the index of the reference microphone that is arbitrarily selected from one of M microphones. We then calculate the direction of arrival (DOA) information at each time–frequency slot using the TDOA information $\boldsymbol{q}'(f,t)$, which consists of the $q'_{jj'}(f,t)$ of all microphone pairs:

$$\boldsymbol{q}(f,t) = c\boldsymbol{D}^{+}\boldsymbol{q}'(f,t) \tag{49}$$

where c is the propagation velocity of the signals, and $^+$ denotes the Moore–Penrose pseudo-inverse. $\boldsymbol{D}$ is the microphone coordinate information $\boldsymbol{D} = [\boldsymbol{d}_1 - \boldsymbol{d}_{j'},\ldots,\boldsymbol{d}_M - \boldsymbol{d}_{j'}]^T$, where a 3-D vector $\boldsymbol{d}_j$ represents the location of microphone j. When the source azimuth is θ and the elevation is ϕ, the DOA vector $\boldsymbol{q}$ can be written as

$$\boldsymbol{q} = [\cos\theta\cos\phi, \sin\theta\cos\phi, \sin\phi]^T.$$

In this paper, we employ only azimuth θ for simplicity as the DOA information.

The signal power $\boldsymbol{f}_{t,d}$ from direction d at time t is calculated by adding the power of the time–frequency component:

$$\boldsymbol{f}_{t,d} = \sum_{d-0.5^\circ \le \theta(f,t) < d+0.5^\circ} |y(f,t)|^2. \tag{50}$$

Acknowledgment

The authors would like to thank D. Zoeke from the University of Erlangen-Nuremberg for helping to implement the system in [8] and calculate the diarization performance, Dr. T. Matsubayashi for computational support, and anonymous reviewers for insightful comments and advice.

References

[1] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Workshop Multimodal Interaction Related Mach. Learn. Algorithms (MLMI)*, 2006, vol. 3869, Lecture Notes in Compuer Science.

[2] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Lang. Resources Eval.*, vol. 41, no. 3-4, pp. 389–407, 2007.

[3] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Proc. Int. Eval. Workshops CLEAR 2007 and RT 2007: Multimodal Technol. for Perception of Humans*, 2008, vol. 4625, Lecture Notes in Computer Science, pp. 373–389.

[4] A. Waibel, M. Bett, F. Metze, K. Rise, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, pp. 597–600.

[5] R. Cutler, Y. Rui, A. Gupta, J. J. Cadiz, I. Tashev, L.-W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *Proc. ACM Int. Conf. Multimedia*, 2002, pp. 503–512.

[6] Z. Yu, M. Ozeki, Y. Fujii, and Y. Nakamura, "Towards smart meeting: Enabling technologies and a real-world application," in *Proc. ACM Int. Conf. Multimodal Interface*, 2007, pp. 86–93.

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

[8] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization systems," in *Proc. Int. Eval. Workshops CLEAR 2007 and RT 2007: Multimodal Technol. for Perception of Humans*, 2008, vol. 4625, Lecture Notes in Computer Science, pp. 509–519.

[9] J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos, "The IBM RT07 evaluation systems for speaker diarization on lecture meetings," in *Proc. Int. Eval. Workshops CLEAR 2007 and RT 2007: Multimodal Technol. for Perception of Humans*, 2008, vol. 4625, Lecture Notes in Computer Science, pp. 497–508.

[10] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, Jul. 2008.

[11] G. Friedland and O. Vinyals, "Live speaker identification in conversations," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008.

[12] X. Anguera, C. Wooters, J. M. Pardo, and J. Hernando, "Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 4, pp. 241–244.

[13] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2008, pp. 29–32.

[14] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Proc. 3rd Joint Workshop Multimodal Interaction and Rel. Mach. Learn. Algorithms*, Washington, DC, 2006, pp. 257–264.

[15] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin, "Hierarchical Bayesian modeling of topics in time-stamped documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 996–1011, Jun. 2010.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[17] K. Ishiguro, T. Yamada, S. Araki, and T. Nakatani, "A probabilistic speaker clustering for DOA-based diarization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA 2009)*, 2009, pp. 241–244.

[18] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on adaptive integration of multiple speech features and a signal decision scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Mar. 2008, pp. 4441–4444.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, May 2006, vol. 5, pp. 33–36.

[20] R. O. Duda, P. E. Hart, and D. G. Stork*, Pattern Classification*. New York: Wiley, 2001.

[21] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition (CVPR)*, Anchorage, AK, Jun. 2008, pp. 1–8.

[22] E. Bart, I. Porteous, P. Perona, and M. Welling, "Unsupervised learning of visual taxonomies," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition (CVPR)*, Anchorage, AK, Jun. 2008, pp. 1–8.

[23] T. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 1165–1172.

[24] X. Zhang, Z. Li, L. Zhan, W.-Y. Ma, and H.-Y. Shum, "Efficient indexing for large scale visual search," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 1103–1110.

[25] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 1389–1396.

[26] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2009, pp. 37–40.

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[28] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. 34th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 33–36.

[29] C. M. Bishop*, Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[30] L. Ren, D. B. Dunson, and L. Carin, "The dynamic hierarchical Dirichlet process," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, Jul. 2008, pp. 824–831.

[31] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 209–215.

[32] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 365–381, Jul. 2004.

[33] P. D. Hoff, "Subset clustering of binary sequences, with an application to genomic abnormality data," *Biometrics*, vol. 61, no. 4, pp. 1027–1036, 2005.

[34] "AMI Project," in *AMI Meeting Corpus* [Online]. Available: http://corpus.amiproject.org

[35] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 4353–4356.

[36] G. Friedland, C. Yeo, and H. Hung, "Dialocalization: Acoustic speaker diarization and visual localization as joint optimization problem," *ACM Trans. Multimedia Comput., Commun., Applicat. (TOMCCAP)*, vol. 6, no. 4, pp. 27:1–27:18, 2010.

[37] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 1, pp. 41–44.

[38] A. Noulas and B. J. A. Krose, "On-line multi-modal speaker diarization," in *Proc. Int. Conf. Multimodal Interfaces (ICMI)*, 2007, pp. 350–357.

Katsuhiko Ishiguro (M'09) received the B.Eng. and M.Inf. degrees from the University of Tokyo, Tokyo, Japan, in 2000 and 2004, respectively, and the Ph.D. degree from the University of Tsukuba, Ibaraki, Japan, in 2010.

He has been a Researcher at the NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan, since 2006. His research interests include multimedia data modeling with Bayesian approaches, probabilistic models for data mining, and time series analysis.

Dr. Ishiguro is a member of the IEICE and IPSJ.

Takeshi Yamada (M'08) received the B.S. degree in mathematics from the University of Tokyo, Tokyo, Japan, in 1988 and the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2003.

He is currently a Senior Research Scientist, Supervisor, and Leader of the Emergent Learning and Systems Research Group, NTT Communication Science Laboratories, Kyoto, Japan. His research interests include data mining, statistical machine learning, graph visualization, metaheuristics, and combinatorial optimization.

Dr. Yamada is a member of the IEICE.

Shoko Araki (M'01) received the B.E. and M.E. degrees from the University of Tokyo, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree from Hokkaido University, Sapporo, Japan in 2007.

She is with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation (BSS) applied to speech signals, meeting diarization, and auditory scene analysis.

Dr. Araki is a member of the organizing committee of the ICA 2003, the finance chair of IWAENC 2003, the co-chair of a special session on undetermined sparse audio source separation in EUSIPCO 2006, the registration chair of WASPAA 2007, and the evaluation co-chair of SiSEC2008, 2010 and 2011. She received the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from the Institute of Electronics, Information, and Communication Engineers (IEICE) in 2006, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2008. She is a member of the IEICE and ASJ.

Tomohiro Nakatani (M'03–SM'06) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 2002, respectively.

He is a Senior Research Scientist at NTT Communication Science Labs, NTT Corporation, Kyoto. Since he joined NTT Corporation in 1991, he has been investigating speech enhancement technologies for developing intelligent human–machine interfaces. From 2005 to 2006, he was a Visiting Scholar at the Georgia Institute of Technology. Since 2007, he has been a Visiting Associate Professor at Nagoya University.

Dr. Nakatani was honored to receive the 1997 JSAI Conference Best Paper Award, and the 2005 IEICE Paper Award. He is a member of IEEE Signal Processing Audio and Electroacoustics Technical Committee and IEEE CAS Blind Signal Processing Technical Committee, an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and a Technical Program Chair of IEEE WASPAA-2007. He is a Technical Program Committee Chair of the IEEE Kansai Section. He is a member of IEICE and ASJ.

Hiroshi Sawada (M'02–SM'04) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993, and 2001, respectively.

He joined NTT Corporation, Kyoto, in 1993. He is now the Group Leader of the Learning and Intelligent Systems Research Group, NTT Communication Science Laboratories, Kyoto. His research interests include statistical signal processing, machine learning, latent variable model, graph-based data structure, and computer architecture.

Dr. Sawada served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2006 to 2009. He is a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE SP Society. He received the Ninth TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 1994, the Best Paper Award of the IEEE Circuit and System Society in 2000, and the MLSP Data Analysis Competition Award in 2007. Dr. Sawada is a member of IEICE and ASJ.