# A PROBABILISTIC SPEAKER CLUSTERING FOR DOA-BASED DIARIZATION

*Katsuhiko Ishiguro, Takeshi Yamada, Shoko Araki and Tomohiro Nakatani*

NTT Communication Science Laboratories
Kyoto, 619-0237, Japan
ishiguro@cslab.kecl.ntt.co.jp

## ABSTRACT

We present a probabilistic speaker clustering and diarization model. Speaker diarization determines "who spoke when" from the recorded conversation of unknown number of people. We formulate this problem as the clustering of sequential auditory features generated by an unknown number of latent mixture components (speakers). We employ a probabilistic model which automatically estimates the number of speakers and time-varying speaker proportions. Experiments with synthetic and real sound recordings confirm that the proposed model can successfully infer the number and features of speakers and obtained better speaker diarization results than conventional models.

***Index Terms***— Probabilistic clustering, speaker diarization, direction of arrival, variational Bayes

## 1. INTRODUCTION

Speaker diarization, i.e. estimating "who spoke when", is one of key technologies for meeting recognition from audio recordings (e.g. [1]). Speaker diarization is useful for, among other functions, the auto-annotation of minutes, speech signal enhancements and human-computer interfaces.

A speaker diarization system first estimates the number of speakers and their classification characteristics. After that, the system classifies the recorded signals into speaker-wise speech signal fragments based on the estimated speaker information.

The main topic of the speaker diarization is speaker clustering: clustering of the auditory features into an unknown number of clusters (speakers). We also note that on-line (incremental) clustering is necessary for real-system application [2]. We measure the time difference in sound signal arrival to the microphone array. Given this TDOA (time difference of arrival) feature and the microphone array geometry, we compute DOA (direction of arrival) features to estimate speaker location. TDOA and DOA features have been proven to be useful in speaker clustering and diarization tasks [2, 3]. DOA-based diarization requires that the speakers do not move during the conversation, but it is robust against voice overlapping and this is highly desirable for speaker diarization in meeting situations.

In this paper, we introduce a new probabilistic model called dLDA for speaker clustering and diarization. Our proposed model estimates the number of speakers attending the conversation by clustering DOA features. In addition, the model formulates time-varying speaker proportions as a simple Markov model, depending on the previous time frame. Experiments on synthetic and real-world recording data show that the proposed model is more effective than conventional models.

## 2. SYSTEM OVERVIEW AND RELATED WORKS

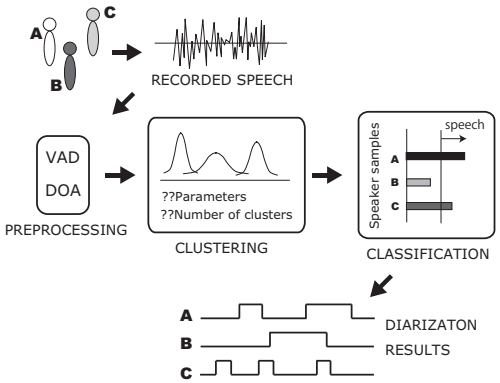


Figure 1: Diarization system overview

Fig. 1 overviews a typical DOA-based speaker diarization system. A sound recording of the conversations 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.

In our model, we combine two feature extractors (preprocessors). The first one, called VAD (voice activity detector) [4], computes the probability of the frame containing (any) voices. If this probability is low, the frame is assumed to be a noise frame, i.e. an environmental noise, and is excluded from further processes. The second and main extractor is DOA (direction of arrival). We utilize the DOA extractor proposed in [5]. This feature extractor estimates i) the direction (angle) of the sound source from a microphone array and ii) 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.

Given the sequence of DOA features (power-orientation features from frames), the clustering processor infers the number and the 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, the clustering is the most difficult and challenging part. We briefly review the clustering techniques used in the previous models and clarify their problems.

In [2], the authors proposed a real-time and on-line diarization system based on DOA features. They employed a simple on-line clustering technique called leader-follower clustering [6]. This algorithm is simple and fast to compute, but has an apparent draw-

back that the clustering result strongly depends on a time independent threshold that is difficult to optimize a priori.

In [3] 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, thus the computational load increases exponentially.

One of the most popular clustering models is Gaussian mixture model (GMM). However a simple GMM, where each Gaussian with parameter $\theta$ corresponds to a speaker, cannot deal with dynamics of time series. It assumes and estimates a time-invariant mixture model $G_0(\theta) = \sum_{k=1}^{K} \beta_k \delta_{\theta_k}(\theta)$ where $\beta_k$ denotes the mixing ratio, $\theta_k$ a parameter of the cluster $k$, and $\delta_{\theta_k}$ a delta function peaked at $\theta_k$, respectively. However, the actual distribution at each time frame is time varying because of "turn-takings" 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 speaker speaks. The distribution at $t$ is apparently different from the time-invariant distribution which generates THREE speakers' observations.

## 3. PROPOSED MODEL

### 3.1. Preprocessing

We assume meeting situations where the speakers remained 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 line. DOA features are $D = 360$ dimensional power vector $\boldsymbol{f}_t = (f_{t,-179}, \ldots, f_{t,180})$. Each $f_{t,d}$ denotes the estimated signal power from the direction (angle) $d$[deg] at a time $t$.

To meet the model description, we convert this DOA feature $\boldsymbol{f}_t$ to a set of discrete samples $x_t = \{x_{ti} \in R^1\}$. From each $d \in \{-179, \ldots, 180\}$, multiple samples $x_{ti} = d$ are reproduced. The number of samples $n_{td}$ is proportional to $f_{t,d}$. Going through this converting step from $d = -179$ to $d = 180$, we have $x_t = \{x_{ti}\}$ containing $n_t = \sum_d n_{td}$ samples. The distribution of $x_{ti}$ reflects the power-orientation distribution at time $t$. Because human voices have large power (large $f_{t,d}$), a sample concentration indicates a location of a speaker. Therefore we can estimate the locations the speakers by clustering samples.

Speaker clustering is understood as the on-line clustering of sample set $\{x_1, \ldots, x_t\}$. We expect $x_{ti}$ are partitioned into clusters. Each cluster corresponds to a speaker, and has latent parameter $\theta$ representing the location of the speaker. Samples $x_{ti}$ are generated from observation function $F(\theta)$. We would like to infer the number of speakers, and their locations, and partition the sound fragments. This is equivalent to finding the optimal number of clusters, their parameters $\theta$, and the assignment of $x_{ti}$ to the clusters.

### 3.2. Dynamic LDA model: idea

In this paper, we adopt a new probabilistic model called dynamic Latent Dirichlet Allocation (dLDA). The dLDA model not only infers the number of clusters (speakers), but also flexibly deals with the time evolution of the parameter (sample) distributions. The unique point of dLDA is that it has a Markov property between the distributions of two consecutive time frames. This property is described as follows:

$$H_t = \sum_{k=1}^{K} \pi_{tk} \delta_{\theta_k} \tag{1}$$

$$w_{t\neq 1} \,|\, a_0, b_0 \sim Beta\,(a_0, b_0)\,, \quad w_1 = 1 \tag{2}$$

$$G_t = (1 - w_t)\, G_{t-1} + w_t H_t = \sum_{k=1}^{K} \beta_{tk} \delta_{\theta_k}. \tag{3}$$

$G_t$ is a mixture which represents who actually produces speeches at the time frame $t$. As in (3), $G_t$ is constructed as a linear interpolation of the previous distribution $G_{t-1}$ and a newly introduced distribution (innovation measure) $H_t$, which is responsible for the change in the sample distribution between $t-1$ and $t$. It is worth noting that $w_t$, an interpolation factor, is a probabilistic variable sampled from Beta distribution (2) and that by regulating $w_t$ the dLDA model can handle an irregular distribution changes, ranging from small to significant changes.

It is beneficial to consider two extreme cases as follows. If we set $w_t = 0$ for all $t$, then all $G_t$ is equivalent to $G_0$ and the dLDA model is reduced to the simple GMM [7]. On the other hand, setting $w_t = 1$ for all $t$ makes all the time frames independent from each other. In this case, we recover the latent Dirichlet allocation (LDA) model [8], which is popular in machine learning community. The dLDA is a natural extension of GMM and LDA.

### 3.3. Dynamic LDA model: the generative model

The generative model of dLDA is shown below.

$$\theta_k \,|\, H \sim H \quad k = 1, \ldots, K \tag{4}$$

$$\boldsymbol{\pi}_t \,|\, \alpha_0 \sim Dirichlet\left(\frac{\alpha_0}{K}, \frac{\alpha_0}{K}, \cdots, \frac{\alpha_0}{K}\right) \tag{5}$$

$$w_{t\neq 1} \,|\, a_0, b_0 \sim Beta\,(a_0, b_0)\,, \quad w_1 = 1 \tag{6}$$

$$v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m) \tag{7}$$

$$d_{ti} \,|\, \boldsymbol{v}_t \sim Multinomial\,(\boldsymbol{v}_t) \tag{8}$$

$$z_{ti} \,|\, d_{ti}, \boldsymbol{\pi}_t \sim Multinomial\left(\boldsymbol{\pi}_{d_{ti}}\right) \tag{9}$$

$$x_{ti} \,|\, z_{ti}, \theta_k \sim F\left(\theta_{z_{ti}}\right) \tag{10}$$

In (4) we sample $K$ parameters for clusters. As discussed, we assume each cluster corresponds to a speaker, and its parameter represents the speaker location. The dLDA puts a Dirichlet prior on the mixing ratio of $K$. 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.

$H_t$ is generated from Dirichlet as in (5), and $G_t$ is represented as the weighted sum of $H_t$. From (3) it is easy to see that

$$G_t = \sum_{l=1}^{t} \left\{ \prod_{m=l+1}^{t} (1 - w_m) \right\} w_l H_l \triangleq \sum_{l=1}^{t} v_{tl} H_l. \tag{11}$$

Rather than considering the temporal mixing ratio $\beta_{tk}$ directly, it is easier to work on $\boldsymbol{v}_t$ for inference. Next, we sample the interpolation factor $w_t$ (6) and compute $v_{tl}$ $(l = 1, \ldots, t)$ as in (7).

As in (11), $G_t$ is a mixture of $H_l$ with the mixing ratio $\boldsymbol{v}_t$. We pick an innovation measure index $d_{ti} = l$, meaning that innovation

measure $H_l$ generates $x_{ti}$ (8). Using $\pi_{d_{ti}}$, we pick a cluster index $z_{ti} = k$ meaning the cluster with parameter $\theta_k$ is responsible for generating $x_{ti}$. In other words, $z_{ti}$ is an index of the speaker that produces the sound heard from the direction of $x_{ti}$ at time $t$ (9). This model can represent voice overlapping situations naturally because each $z_{ti}$ may take a different value independently.

Finally, $x_{ti}$ is sampled from the observation distribution (function) $F$ with picked parameter $\theta_k$. We assume $F$ is a one dimensional Gaussian (14). To maintain conjugacy, we use a Normal-Gamma distribution for the prior of parameter (15).

Please note that dLDA can be extended to dynamic Hierarchical Dirichlet Process (dHDP) [9] as the limit of $K \to \infty$, where the finite Dirichlet prior is replaced by an infinite stick-breaking prior. However, it is difficult to derive VB inference algorithm for dHDP. [10] proposes a VB inference solution for finite approximated version of dHDP, but their model is in fact dLDA.

Another closely related model is recently proposed by Fox and others [11]. The model formulates the infinite number of speakers and time-dependent turn-takings in a form of extended HMM. The main difference from our model is that their model does not allow overlaps of speakers.

### 3.4. On-line VB inference

There are two major approaches to solve the probabilistic model. One is Gibbs sampling that is accurate but slow in terms of convergence, and the other is variational Bayes (VB) that is fast but may be trapped in a locally optimal solution. We prefer the VB inference for on-line speaker diarization system.

In this paper, we develop an on-line and incremental inference by VB for dLDA. VB computes the variational (approximated) posterior $q(\cdot)$ of the hidden variables and parameters ( (16-20) in the appendix). VB iteratively computes i) the posterior parameters of hidden variables ($r_{tik}$ and $s_{til}$) and ii) sufficient statistics.

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

$$\|z_{t,k}\| = \sum_{i=1}^{n_t} r_{tik}. \tag{12}$$

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

$$\hat{\beta}_{tk} = \frac{\|z_{t,k}\|}{\sum_{k=1}^{K} \|z_{t,k}\|}, \quad \hat{\beta}_k = \frac{\sum_t \|z_{t,k}\|}{\sum_{k=1}^{K} \sum_t \|z_{t,k}\|}. \tag{13}$$

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

## 4. EXPERIMENTS

### 4.1. Data Sets

The simulated data has 422 frames with 64 [msec] intervals, excluding no-speech periods based on VAD. We merged five consecutive frames into a larger frame in order to stabilize the distributions of $x_{ti}$ within a time frame. Hereafter index $t$ refers to the time index after the merge, and $T = 82$. This data simulated the conversation by three people with speaker overlaps.

We used four data sets of the real recordings gathered in [2]. Their specifications are shown in Table 1. Each data set is a 300 [sec] long recording, and was sampled at 32 [msec] intervals. We merged 10 consecutive frames into a single frame, yielding four sequences of $T = 938$.

Table 1: Specification of the real recording data sets

| Data | # 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 |

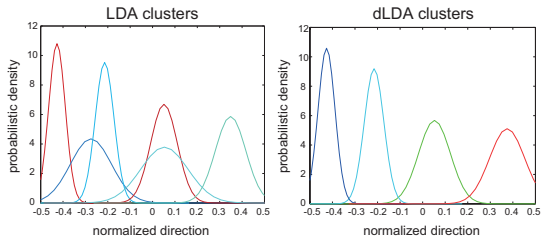


Figure 2: Clustering results on the CP1 real record data set (four speakers). Vertical axis denotes the probability density, and the horizontal axis denotes the normalized angle (location). Left: clustering results by LDA. Right: clustering results by dLDA.

### 4.2. Experiment on speaker clustering

In this experiment, we tested the performance of dLDA with regard to the speaker clustering tasks, and compared it against the GMM and the LDA models. The GMM and LDA are simulated by setting all $w_t$ as $w_t \approx 0$ or $w_t \approx 1$, respectively. After the on-line clustering of the record data up to the last time frame $T$, we examined the resultant number of clusters and their parameters.

We present the part of clustering results in Fig. 2 and Fig. 3. As you can see, dLDA clustering obtained better results than GMM and LDA models. We assume that this difference comes from the ability to model the intermittent changes of speaker (cluster) distributions. The dLDA model could not achieve perfect clustering with the DC and CN datasets: the model produced extra clusters corresponding to noise inputs. However those "noise" clusters had smaller mixing ratios than the those of "speaker" clusters. Therefore we can still improve the clustering of dLDA by carefully studying the threshold ($\hat{\beta}_k > \frac{1}{K}$).

### 4.3. Experiment on speaker diarization precision

Next, we evaluated the classification performance to verify the entire diarization system (Fig. 1). We employ the DER (diarization error ratio) measure [12] for the evaluations. DER is a percentage of the error time length in the total sound recording length.

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

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

$\tau_1, \tau_2$ are predefined thresholds. This simple classification rule is based on the one used in [2], and effectively suppress the "noise" clusters estimated in CN and DC datasets (Fig. 4).

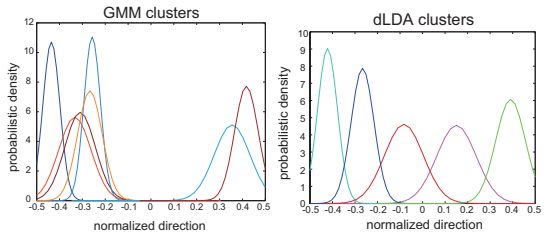


Figure 3: Clustering results on the CN real record data set (three speakers). Vertical axis denotes the probability density, and the horizontal axis denotes the normalized angle (location). Left: clustering results by GMM. Right: clustering results by dLDA.

Table 2: DER measures (%) achieved in diarization experiments

| Method | Simulate | CP1 | CP2 | DC | CN |
|---|---|---|---|---|---|
| [2] | 7.1 | 21.9 | 25.0 | 29.9 | 34.3 |
| dLDA(proposed) | **6.7** | **21.6** | **22.3** | **27.0** | **30.9** |

Table 2 presents the computed DER measures; the DER values reported in [2] are shown for comparison. It is clear that the proposed model basically outperforms the previous method. We note that DERs of the GMM and the LDA models are much worse than those of dLDA (results not shown).

## 5. CONCLUSION

In this paper, we proposed a new probabilistic model for speaker diarization tasks. We employ dynamic LDA (dLDA) model for speaker clustering. The dLDA model automatically infer the number of clusters and data partitioning, and is able to handle time varying cluster distributions. We developed an on-line inference algorithm, and experimentally confirmed the improvements in DER performance yielded by the dLDA model.

## A. MODELS AND VB POSTERIORS

The observation function $F$ and the parameter prior $H$ are:

$$x_i \sim N\left(\cdot; m, \sigma^2\right), \tag{14}$$

$$p\left(m, \sigma^{-2}\right) = NormalGamma\left(\mu_0, \beta_0, c_0, d_0\right). \tag{15}$$

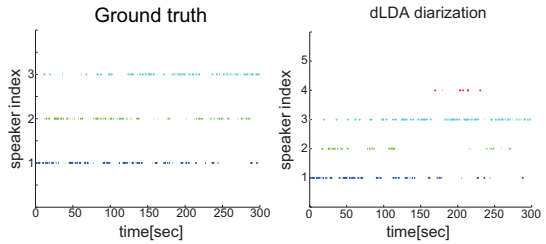


Figure 4: Diarization results on the DC real record data set (best viewed in color). The horizontal axis denotes the actual time. Left: ground truth. Right: diarization results after dLDA clustering.

VB posteriors are defined as follows:

$$q^*\left(\theta_k\right) = NormalGamma\left(\mu_1, \beta_1, c_1, d_1\right), \tag{16}$$

$$q^*\left(\boldsymbol{\pi}_t\right) = Dirichlet(\frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{mi1}\, s_{mit}, \ldots,$$

$$\frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{mik}\, s_{mit}, \ldots, \frac{\alpha_0}{K} + \sum_{m=t}^{T}\sum_{i=1}^{n_m} r_{miK}\, s_{mit},), \tag{17}$$

$$q^*\left(w_t\right) = Beta\left(a_0 + \sum_{i=1}^{n_t} s_{tit}, b_0 + \sum_{i=1}^{n_t}\sum_{m=1}^{t-1} s_{tim}\right), \tag{18}$$

$$q^*\left(d_{ti}\right) = Multinomial\left(s_{ti1}, \ldots, s_{til}, \ldots, s_{tit}\right), \tag{19}$$

$$q^*\left(z_{ti}\right) = Multinomial\left(r_{ti1}, \ldots, r_{tik}, \ldots, r_{tiK}\right). \tag{20}$$

Parameters of the hidden variables' posteriors $(s, r)$ are computed via EM algorithms. Required suffiient statistics are:

$$\mu_1 = \frac{\beta_0\mu_0 + N_k\tilde{x}_k}{\beta_0 + N_k}, \quad \beta_1 = \beta_0 + N_k,$$

$$c_1 = c_0 + \frac{N_k}{2}, \quad d_1 = d_0 + \frac{\tilde{S}}{2} + \frac{\beta_0 N_k}{\beta_0 + N_k}\frac{(\tilde{x}_k - \mu_0)^2}{2},$$

$$N_k = \sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{tik}, \quad \tilde{x}_k = \frac{1}{N_k}\sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{tik}x_{ti}, \quad \tilde{S}_k = \sum_{t=1}^{T}\sum_{i=1}^{n_t} r_{tik}\left(x_{ti} - \tilde{x}_k\right)^2.$$

## B. REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Proc. HSCMA*, 2008, pp. 29–32.

[3] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Proc. MLMI*, 2006, pp. 705–711.

[4] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on adaptive integration of multiple speech feature and signal decision scheme," in *Proc. IEEE ICASSP*, 2008, pp. 4441–4444.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE ICASSP*, vol. 5, 2006, pp. 33–36.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2001.

[7] 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. ICASSP*, 2009, pp. 33–36.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] L. Ren, D. B. Dunson, and L. Carin, "The dynamic hierarchical Dirichlet process," in *Proc. ICML*, 2008, pp. 824–831.

[10] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin, "Dynamic hierarchical Dirichlet process for modeling topics in time-stamped documents," *IEEE Trans. PAMI*, vol. submitted, 2008.

[11] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. ICML*, 2008.

[12] NIST, "Spring 2007 (rt-07) rich transcription meeting recognition evaluation plan," 2007. [Online]. Available: http://www.nist.gov/speech/tests/rt/2007/index.html