MAXIMUM-LIKELIHOOD ONLINE SPEAKER DIARIZATION IN NOISY MEETINGS BASED ON CATEGORICAL MIXTURE MODEL AND PROBABILISTIC SPATIAL DICTIONARY

Nobutaka Ito, Takashi Makino, Shoko Araki, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{ito.nobutaka, araki.shoko, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

In this paper, we propose a maximum-likelihood online diarization method based on a *probabilistic spatial dictionary*. This dictionary consists of the given probability distribution of spatial features for each possible direction of arrival (DOA) of source signals. Recently, we have developed an online, noise-robust diarization method by utilizing this dictionary as spatial prior information. In this method, DOA estimation is first performed frame-wise based on the dictionary, and subsequently diarization is performed. Although the DOA estimation is performed optimally in the maximum-likelihood sense, the diarization is performed suboptimally based on some heuristics. In contrast, the proposed method performs DOA estimation and diarization jointly and optimally in the maximum-likelihood sense. This is realized by introducing a *categorical mixture model (CMM)*, which has source-wise DOA information and diarization information as unknown parameters. We conducted an experiment on a real-world meeting dataset, and confirmed that the proposed method reduced a diarization error rate by absolute 2.7 % compared to the above conventional method.

Index Terms— Microphone array signal processing, speaker diarization, maximum-likelihood method.

1. INTRODUCTION

Automatic speech recognition (ASR) can work quite well when the desired speaker speaks in proximity to the microphones. However, when the speaker speaks at a distance from the microphones, the ASR performance degrades significantly. We are conducting research that aims to realize accurate ASR even in such a setting.

Specifically, this work aims to realize ASR in a meeting situation in which multiple speakers are conversing spontaneously at a distance from the microphones. In such a situation, there are many obstacles to accurate ASR, such as reverberation, speech overlap, and background noise. To deal with these obstacles, diarization and speech enhancement are crucial, and we focus on the former in this paper.

Diarization refers to the estimation of the speech intervals of each speaker, *i.e.*, the estimation of *who spoke when*. Such information is rarely available in real-world meetings, and therefore needs to be estimated from observed signals. Diarization is crucial not only to exclude noise-only segments and give speaker labels in the back-end but also to stabilize the adaptation of the beamformer for each speaker in the front-end [1]. Although diarization has been addressed by many researchers [2–5], it still remains to be an important research topic, especially with distant microphones in *noisy* environments. Furthermore, although it may be relatively easy to improve diarization performance for batch processing, *online* processing ability is compromised in that case.

Recently, we have proposed an *online, noise-robust* diarization method based on a *probabilistic spatial dictionary* [1, 6]. In this method, we prepare a probabilistic spatial dictionary composed of the given probability distribution of spatial features for each *possible DOA*. Here, we assume that each of the *source signals* (including the speech signals and the background noise) arrives from one of predetermined DOA candidates, which we call *possible DOAs*. The dictionary enables online, noise-robust diarization by providing spatial prior information. In this method, DOA estimation is first performed frame-wise based on the dictionary, and subsequently diarization is performed based on it. Although the DOA estimation is performed optimally in the maximum-likelihood sense, diarization is performed suboptimally based on some heuristics.

In this paper, we propose a novel diarization method based on the probabilistic spatial dictionary. Unlike the conventional method, the proposed method performs DOA estimation and diarization jointly and optimally in the maximum-likelihood sense. This is realized by introducing a *categorical mixture model (CMM)*, which has source-wise DOA information and diarization information as unknown parameters.

We follow the following conventions throughout the rest of this paper. Signals are represented in the short-time Fourier transform (STFT) domain with the time and the frequency indices being t and f respectively. T denotes the number of frames, and F the number of frequency bins up to the Nyquist frequency, $(\cdot)^{\mathsf{T}}$ transposition, and ϕ the empty set.

2. CONVENTIONAL METHOD

This section describes the conventional diarization method [1, 6] based on the probabilistic spatial dictionary.

Figure 1 shows the processing flow of the conventional method. The input data are signals observed at M ($\geq$ 2) microphones in the presence of multiple speakers along with background noise and reverberation. These observed signals are represented by an M-dimensional complex vector $\boldsymbol{y}_{tf}$ in the STFT domain, which we refer to an *observation vector*. The output is the diarization result represented by a binary variable $d_t^{(n)}$, which equals 1 if the nth speaker is speaking in the tth frame and 0 otherwise.

2.1. Feature Extraction

A spatial feature vector $\boldsymbol{z}_{tf}$ is extracted at each time-frequency point. An example of such a feature vector is given by

$$\boldsymbol{z}_{tf} = \frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|_2}, \tag{1}$$

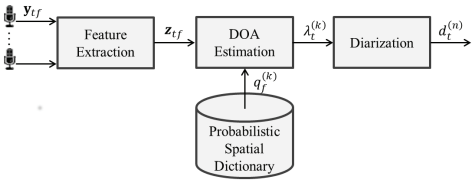

Fig. 1. Processing flow of the conventional diarization method [1, 6].

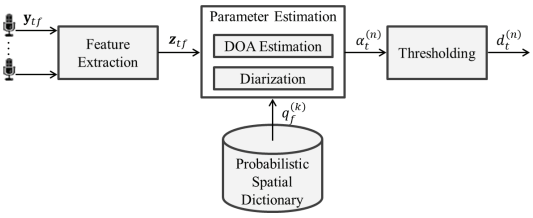

Fig. 2. Processing flow of the proposed diarization method.

where $\|\cdot\|_2$ denotes the Euclidean norm (see [1]).

2.2. Probabilistic Spatial Dictionary

We assume that each of the *source signals* (including the speech signals and the background noise) arrives from one of predetermined DOA candidates, which we call *possible DOAs*. The possible DOAs can be represented by the indices $1, 2, \ldots, K$, where K is assumed to be much larger than the number of speech signals, N. We also regard diffuse noise from all directions as a possible DOA, and represent it by the index 0. This enables us to model diffuse noise explicitly to realize accurate diarization even in the presence of diffuse noise.

In this method, we prepare a probabilistic spatial dictionary composed of the given probability distribution $q_f^{(k)}$ of the feature vector $\boldsymbol{z}_{tf}$ for each possible DOA $k \in \{0, 1, \ldots, K\}$ and each frequency bin $f \in \{1, 2, \ldots, F\}$. For example, for the feature vector in (1), the dictionary can be prepared as follows [1]. The distribution $q_f^{(k)}$ is modeled by a complex Watson distribution [7, 8]

$$q_f^{(k)}(\boldsymbol{z}_{tf}) = \mathcal{W}(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}), \tag{2}$$

where $\boldsymbol{a}_f^{(k)}$ is a centroid parameter and $\kappa_f^{(k)}$ is a concentration parameter. These parameters are prepared as follows. For $k = 1, 2, \ldots, K$, they are pre-trained on training data composed of a real-recorded source signal from each possible DOA. To model diffuse noise, we set $\kappa_f^{(0)}$ at zero and $\boldsymbol{a}_f^{(0)}$ at an arbitrary unit vector, for which (2) reduces to the uniform distribution. See [9] for more details of dictionary preparation.

2.3. Probabilistic Model of Feature Vector Based on Probabilistic Spatial Dictionary

The probability distribution of the feature vector $\boldsymbol{z}_{tf}$ is modeled by a mixture model composed of the distributions $q_f^{(k)}$ in the probabilistic spatial dictionary:

$$p(\boldsymbol{z}_{tf}) = \sum_{k=0}^{K} \lambda_t^{(k)} \underbrace{q_f^{(k)}(\boldsymbol{z}_{tf})}_{\text{dictionary}}. \tag{3}$$

Since the DOAs are unknown, the mixture weight $\lambda_t^{(k)}$ is assumed to be unknown. In the conventional method, no constraint is imposed on $\lambda_t^{(k)}$ except the trivial one:

$$\sum_{k=0}^{K} \lambda_t^{(k)} = 1. \tag{4}$$

We call the mixture weight $\lambda_t^{(k)}$ a *frame-wise DOA probability*. It is the probability of the observed signals in the tth frame arriving from the kth possible DOA. The estimation of $\lambda_t^{(k)}$ corresponds to frame-wise DOA estimation.

An assumption underlying (3) is sparseness: each of the source signals has non-zero power only at a small percentage of time-frequency points. Based on this, the observation vector $\boldsymbol{y}_{tf}$ at each (t, f) is assumed to be composed of one source signal only. This implies that, at each (t, f), the sound arrives from only one DOA, which is represented by a DOA index k_{tf} in the following.

On the above assumptions, the feature vector $\boldsymbol{z}_{tf}$ can be considered to be generated by the following generative process. First, the DOA index k_{tf} is generated from the categorical distribution $P(k_{tf} = k) = \lambda_t^{(k)}$. Then, under the condition $k_{tf} = k$, the feature vector $\boldsymbol{z}_{tf}$ is generated from the conditional distribution $p(\boldsymbol{z}_{tf} \mid k_{tf} = k) = q_f^{(k)}(\boldsymbol{z}_{tf})$. It is straightforward to confirm that this generative process leads to the mixture model (3).

In practice, the background noise may not be sparse so that it has non-zero power at all time-frequency points. In this case, the observation vector $\boldsymbol{y}_{tf}$ is composed of one speech signal plus the background noise if $k_{tf} \geq 1$, and only the background noise if $k_{tf} = 0$. Therefore, (3) may be invalid in an extremely noisy environment, but still (3) remains approximately valid for a moderate noise level. This is because the background noise is negligible as compared to the speech signal present at each (t, f) for a moderate noise level.

2.4. DOA Estimation

The frame-wise DOA probability $\lambda_t^{(k)}$ in (3) is unknown and estimated in the maximum-likelihood sense. Specifically, it is estimated by maximizing the likelihood function $\prod_{t=1}^{T} \prod_{f=1}^{F} p(\boldsymbol{z}_{tf})$ based on an EM algorithm, in which an E step and an M step are alternated until convergence. The E step updates the posterior distribution $\zeta_{tf}^{(k)} \triangleq P(k_{tf} = k \mid \boldsymbol{z}_{tf})$ of the DOA index k_{tf} based on the current estimate of $\lambda_t^{(k)}$ by

$$\zeta_{tf}^{(k)} \leftarrow \frac{\lambda_t^{(k)} q_f^{(k)}(\boldsymbol{z}_{tf})}{\sum_{k'=0}^{K} \lambda_t^{(k')} q_f^{(k')}(\boldsymbol{z}_{tf})}. \tag{5}$$

The M step updates $\lambda_t^{(k)}$ so as to maximize an auxiliary Q function by

$$\lambda_t^{(k)} \leftarrow \frac{1}{F} \sum_{f=1}^{F} \zeta_{tf}^{(k)}. \tag{6}$$

It is theoretically guaranteed that this EM algorithm increases the likelihood function monotonically.

Alternatively, the gradient method can also be employed instead of the EM algorithm [1].

2.5. Diarization

Once the frame-wise DOA probability $\lambda_t^{(k)}$ is obtained as in Section 2.4, diarization can be performed as follows.

First, peak picking of $\lambda_t^{(k)}$ is performed in each frame t. That is, the DOA indices k that locally maximize $\lambda_t^{(k)}$ in the range $1 \leq k \leq K$ (*i.e.*, with the noise DOA index $k = 0$ excluded) are detected. The detected DOA indices k are regarded as the DOAs of the source signals present in the frame t, and collected in a set $R_t \subset \{1, 2, \ldots, K\}$.

Second, diarization is performed based on frame-wise classification of the detected DOA indices in R_t into source classes. The speaker location for each speaker is assumed to be given, which makes this classification straightforward. In each frame, the speaker n is considered to be speaking, if the nth source class is non-empty. The diarization result is stored in the variable $d_t^{(n)}$.

2.6. Discussion

The conventional method first estimates the frame-wise DOA probability $\lambda_t^{(k)}$, and then performs DOA detection and diarization. Although $\lambda_t^{(k)}$ is estimated optimally in the maximum-likelihood sense, the DOA detection and the diarization are performed sub-optimally based on some heuristics. This motivates the proposed method described in the next section.

3. PROPOSED METHOD

In this section, we describe the proposed diarization method. Unlike the conventional method, the proposed method estimates source-wise DOA information $\beta^{(n,k)}$ and diarization information $\alpha_t^{(n)}$ jointly and optimally in the maximum-likelihood sense. This is realized by modeling the frame-wise DOA probability $\lambda_t^{(k)}$ by a categorical mixture model (CMM) parametrized by $\beta^{(n,k)}$ and $\alpha_t^{(n)}$. In Section 4, we show through an experiment that the proposed method outperforms the conventional method in terms of a diarization error rate.

Figure 2 shows the processing flow of the proposed method. The input and the output are the same as in the conventional method in Fig. 1. The feature extraction procedure and the probabilistic spatial dictionary are also the same as in the conventional method.

In the following, we focus on the main differences from the conventional method: the probabilistic model of the feature vector (Section 3.1), the parameter estimation procedure (Section 3.2), and the diarization procedure (Section 3.3). We also describe online implementation of the proposed method (Section 3.4).

3.1. Probabilistic Model of Feature Vector Based on Probabilistic Spatial Dictionary and Categorical Mixture Model

In the conventional method, the frame-wise DOA probability $\lambda_t^{(k)}$ is unconstrained except for (4). In contrast, the proposed method models it by

$$\lambda_t^{(k)} = \sum_{n=0}^{N} \beta^{(n,k)} \alpha_t^{(n)}. \tag{7}$$

Here, $\alpha_t^{(n)}$ and $\beta^{(n,k)}$ are unknown parameters. The index $n \in \{0, 1, \cdots, N\}$ is the source index, which corresponds to a speech signal for $n \geq 1$, and the background noise for $n = 0$. N denotes the number of speakers. The parameter $\alpha_t^{(n)}$ is the presence probability of the nth source signal in the tth frame, which satisfies $\sum_{n=0}^{N} \alpha_t^{(n)} = 1$. The parameter $\alpha_t^{(n)}$ is called a *frame-wise source presence probability (SPP)*, and can be regarded as the diarization information. The parameter $\beta^{(n,k)}$ is the probability of the nth source signal arriving from the kth possible DOA, which we call a *source-wise DOA probability*. It satisfies $\sum_{k=0}^{K} \beta^{(n,k)} = 1$. It can be easily confirmed that $\lambda_t^{(k)}$ in (7) satisfies (4).

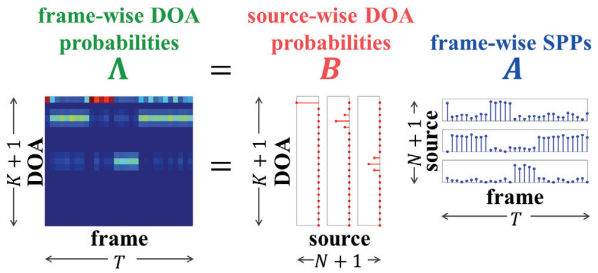

Fig. 3. In the proposed method, the frame-wise DOA probabilities (matrix $\boldsymbol{\Lambda}$) are modeled by the matrix product of the source-wise DOA probabilities (matrix $\boldsymbol{B}$) and the frame-wise SPPs (matrix $\boldsymbol{A}$).

Equation (7) can be rewritten in matrix form as follows (see Fig. 3):

$$\underbrace{\begin{pmatrix} \lambda_1^{(0)} & \cdots & \lambda_T^{(0)} \\ \vdots & \ddots & \vdots \\ \lambda_1^{(K)} & \cdots & \lambda_T^{(K)} \end{pmatrix}}_{\boldsymbol{\Lambda}} = \underbrace{\begin{pmatrix} \beta^{(0,0)} & \cdots & \beta^{(N,0)} \\ \vdots & \ddots & \vdots \\ \beta^{(0,K)} & \cdots & \beta^{(N,K)} \end{pmatrix}}_{\boldsymbol{B}} \underbrace{\begin{pmatrix} \alpha_1^{(0)} & \cdots & \alpha_T^{(0)} \\ \vdots & \ddots & \vdots \\ \alpha_1^{(N)} & \cdots & \alpha_T^{(N)} \end{pmatrix}}_{\boldsymbol{A}}. \tag{8}$$

This is the same model as in the PLSA (probabilistic latent semantic analysis) [10], which is a variant of the NMF (non-negative matrix factorization) [11]. Equation (8) can also be regarded as a categorical mixture model (CMM), which can be seen by rewriting (8) as follows:

$$\begin{pmatrix} \lambda_t^{(0)} \\ \vdots \\ \lambda_t^{(K)} \end{pmatrix} = \sum_{n=0}^{N} \alpha_t^{(n)} \underbrace{\begin{pmatrix} \beta^{(n,0)} \\ \vdots \\ \beta^{(n,K)} \end{pmatrix}}_{\text{categorical distribution}}. \tag{9}$$

By plugging (7) in (3), we obtain the proposed probabilistic model as follows:

$$p(\boldsymbol{z}_{tf}) = \underbrace{\sum_{k=0}^{K} \sum_{n=0}^{N} \beta^{(n,k)} \alpha_t^{(n)}}_{\text{CMM}} \underbrace{q_f^{(k)}(\boldsymbol{z}_{tf})}_{\text{dictionary}}. \tag{10}$$

The generative process behind (10) is as follows. First, a source index n_{tf} indicating the source signal present at (t, f) is generated from the categorical distribution $P(n_{tf} = n) = \alpha_t^{(n)}$. Second, under the condition $n_{tf} = n$, the DOA index k_{tf} is generated from the conditional distribution $P(k_{tf} = k \mid n_{tf} = n) = \beta^{(n,k)}$. Finally, under the condition $k_{tf} = k$, the feature vector $\boldsymbol{z}_{tf}$ is generated from the conditional distribution $p(\boldsymbol{z}_{tf} \mid k_{tf} = k) = q_f^{(k)}(\boldsymbol{z}_{tf})$. The model (10) can be easily derived from this generative model.

As we will describe in Section 3.2, the parameters $\alpha_t^{(n)}$ and $\beta^{(n,k)}$ in (10) are estimated in the maximum-likelihood sense. The estimation of $\alpha_t^{(n)}$ corresponds to diarization, and the estimation

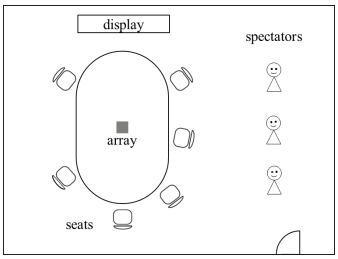

Fig. 4. Experimental setting.

of $\beta^{(n,k)}$ corresponds to DOA estimation. Furthermore, time-frequency masks can also be obtained based on these parameters as follows:

$$P(n_{tf} = n \mid \boldsymbol{z}_{tf}) = \frac{\sum_{k=0}^{K} \alpha_t^{(n)} \beta^{(n,k)} q_f^{(k)}(\boldsymbol{z}_{tf})}{\sum_{n'=0}^{N} \sum_{k=0}^{K} \alpha_t^{(n')} \beta^{(n',k)} q_f^{(k)}(\boldsymbol{z}_{tf})}. \quad (11)$$

These can be utilized for further processing such as mask-based beamforming [1], although here we focus on diarization, which utilizes $\alpha_t^{(n)}$ only.

3.2. Parameter Estimation

The parameters $\alpha_t^{(n)}$ and $\beta^{(n,k)}$ are estimated based on the EM algorithm. This corresponds to joint DOA estimation and diarization. The E step updates the posterior distribution $\gamma_{tf}^{(n,k)} \triangleq P(n_{tf} = n, k_{tf} = k \mid \boldsymbol{z}_{tf})$ based on the current estimates of $\alpha_t^{(n)}$ and $\beta^{(n,k)}$ by

$$\gamma_{tf}^{(n,k)} \leftarrow \frac{\alpha_t^{(n)} \beta^{(n,k)} q_f^{(k)}(\boldsymbol{z}_{tf})}{\sum_{n'=0}^{N} \sum_{k'=0}^{K} \alpha_t^{(n')} \beta^{(n',k')} q_f^{(k')}(\boldsymbol{z}_{tf})}. \quad (12)$$

The M step updates $\alpha_t^{(n)}$ and $\beta^{(n,k)}$ so as to maximize an auxiliary Q function by

$$\alpha_t^{(n)} \leftarrow \frac{1}{F} \sum_{f=1}^{F} \sum_{k=0}^{K} \gamma_{tf}^{(n,k)}, \quad (13)$$

$$\beta^{(n,k)} \leftarrow \frac{\sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_{tf}^{(n,k)}}{\sum_{k'=0}^{K} \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_{tf}^{(n,k')}}. \quad (14)$$

3.3. Thresholding

As already pointed out, the frame-wise SPP $\alpha_t^{(n)}$ can be regarded as diarization information. The diarization result $d_t^{(n)}$ is obtained by applying a predetermined threshold to $\alpha_t^{(n)}$ $(1 \leq n \leq N)$.

3.4. Online Implementation

This section describes online implementation of the above EM algorithm. Equations (12) and (13) along with the feature extraction and the diarization procedures can be performed frame-wise. In contrast, (14) cannot be computed frame-wise, because it involves temporal summation.

To make (14) computable frame-wise, note that (14) can be rewritten as

$$\beta^{(n,k)} \leftarrow \frac{\left\langle \sum_{f=1}^{F} \gamma_{tf}^{(n,k)} \right\rangle}{\sum_{k'=0}^{K} \left\langle \sum_{f=1}^{F} \gamma_{tf}^{(n,k')} \right\rangle}, \quad (15)$$

where $\langle \cdot \rangle$ denotes temporal averaging. We replace the average $\langle \sum_{f=1}^{F} \gamma_{tf}^{(n,k)} \rangle$ in (15) by a moving average $\mu_t^{(n,k)}$:

$$\beta_t^{(n,k)} \leftarrow \frac{\mu_t^{(n,k)}}{\sum_{k'=0}^{K} \mu_t^{(n,k')}}. \quad (16)$$

The moving average $\mu_t^{(n,k)}$ is updated frame-wise by

$$\mu_t^{(n,k)} \leftarrow (1-\delta)\mu_{t-1}^{(n,k)} + \delta \sum_{f=1}^{F} \gamma_{tf}^{(n,k)}. \quad (17)$$

Here, δ is the forgetting factor.

Table 1. Diarization error rate.

session ID	conventional [6]	proposed
20141208_ses03	27.7%	**24.8%**
20141208_ses04	24.8%	**20.9%**
20141208_ses05	17.2%	**13.0%**
20141208_ses06	**18.9%**	19.8%
20141216_ses02	9.3%	**8.8%**
20141216_ses03	12.2%	**9.8%**
20141217_ses02	15.6%	**13.8%**
20141217_ses03	18.9%	**12.3%**
average	18.1%	**15.4%**

4. DIARIZATION EXPERIMENT

We conducted a diarization experiment on a real-world meeting dataset [6]. Figure 4 depicts the recording setting for the dataset. Four to six speakers conversed at a table in a meeting room, which was recorded by an eight-channel microphone array on the table. Background noise was simulated by playing babble noise from ten loudspeakers outside the room. See [6] for the details of the dataset. The proposed method and the conventional method [6] based on the probabilistic spatial dictionary were compared. The feature vector in (1) was used. The dictionary was designed as in the example in Section 2.2.

Table 1 shows the diarization error rate (DER) [12] for each session. In average, the proposed method reduced the DER by 2.7 % compared to the conventional method.

5. CONCLUSIONS

In this paper, we proposed a diarization method based on the probabilistic spatial dictionary and the categorical mixture model. The proposed method estimates source-wise DOA information and diarization information jointly and optimally in the maximum-likelihood sense. In the experiment, the proposed method reduced the DER by absolute 2.7 % compared to the conventional method [6].

Future work includes application of the proposed approach to meeting speech enhancement.

6. REFERENCES

[1] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. ICASSP*, Mar. 2017, pp. 681–685.

[2] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. ICASSP*, Apr. 2003, pp. 572–575.

[3] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Machine Learning for Multimodal Interaction*, S. Renals, S. Bengio, and J. G. Fiscus, Eds. Springer Berlin Heidelberg, 2006, pp. 257–264.

[4] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. ASLP*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[5] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. ASLP*, vol. 20, no. 2, pp. 499–513, Feb. 2012.

[6] M. Fakhry, N. Ito, S. Araki, and T. Nakatani, "Modeling audio directional statistics using a probabilistic spatial dictionary for speaker diarization in real meetings," in *Proc. IWAENC*, Sept. 2016.

[7] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[8] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

[9] N. Ito, S. Araki, and T. Nakatani, "Data-driven and physical model-based designs of probabilistic spatial dictionary for online meeting diarization and adaptive beamforming," in *Proc. EUSIPCO*, Aug. 2017, pp. 1205–1209.

[10] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, July 1999, pp. 289–296.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[12] The 2009 (RT-09) rich transcription meeting recognition evaluation plan. http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html.