PROBABILISTIC SPATIAL DICTIONARY BASED ONLINE ADAPTIVE BEAMFORMING FOR MEETING RECOGNITION IN NOISY AND REVERBERANT ENVIRONMENTS

Nobutaka Ito, Shoko Araki, Marc Delcroix, Tomohiro Nakatani

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

ABSTRACT

Here we propose online adaptive beamforming for automatic speech recognition (ASR) in meetings in noisy, reverberant environments. The proposed method is based on recently developed mask-based beamforming, in which accurate mask estimation and diarization are paramount. Real-world experiments have shown that mask-based beamforming enables accurate ASR in meetings in small noise and reverberation with a signal-to-noise ratio (SNR) of 15–25 dB and a reverberation time (RT) of 120–350 ms. In this paper, we deal with a more adverse condition: meetings in large noise and reverberation with an SNR of 3–15 dB and an RT of 500 ms. To this end, we exploit a *probabilistic spatial dictionary*, a dictionary that consists of a pre-trained probability distribution of source location features for each potential speaker location. This dictionary enables us to perform mask estimation and diarization for beamforming accurately, even in the above adverse condition. The proposed method reduced the word error rate (WER) on real meeting data by 54.8 % relative to our previous beamforming method.

Index Terms— Automatic speech recognition, maximum likelihood estimation, microphone arrays, speech enhancement.

1. INTRODUCTION

When a desired speech is recorded by a distant microphone, the recording contains not only the desired speech but also interferences and background noise, which degrade the ASR performance. To resolve this problem, speech enhancement has been studied extensively, which aims to estimate the desired speech from the recording.

A promising speech enhancement technique as a front-end of ASR is a minimum variance distortionless response (MVDR) beamformer, which can reduce interferences and background noise without distorting the target speech. The key to the effectiveness of the MVDR beamformer is accurate estimation of a steering vector, which represents the acoustic transfer characteristics from a sound source to the microphone array. Conventionally, the steering vector is estimated based on the assumption of planewave propagation and a known array geometry. However, these assumptions are often violated in the real world, and the estimated steering vector is inaccurate. This degrades the performance of the MVDR beamformer and thus that of ASR.

To resolve this issue, we have recently proposed mask-based MVDR beamforming [1]. In this approach, the steering vector is estimated based on masks, which indicate the dominant source signal (or background noise) at each time-frequency point. The approach does not make the above unrealistic assumptions, and therefore it is more robust in the real world than the conventional MVDR beamformer. The mask-based MVDR beamforming has been applied to denoising in the NTT CHiME-3 system successfully [1]. It has also been applied to meeting ASR, but an environment with small noise and reverberation was considered with an SNR of 15–25 dB and an RT of 120–350 ms [2].

In this paper, we extend the mask-based MVDR beamforming so that it can deal with meetings in noisy, reverberant environments. For example, we consider an environment with an SNR of 3–15 dB and an RT of 500 ms in the experiment. To deal with such an adverse condition, we exploit a *probabilistic spatial dictionary* as prior knowledge about source location features. The probabilistic spatial dictionary consists of a pre-trained probability distribution of the source location features for each potential speaker location. This pre-trained dictionary reduces the number of unknown parameters, enabling accurate mask estimation and diarization from short observation and in adverse environments. Moreover, the pre-trained dictionary takes into account the acoustic transfer characteristics of the environment properly, which also makes the proposed method robust against reverberation.

The proposed method is suitable for online processing because of the following features. First, the reduced number of unknown parameters results in reduced computational costs. Second, the distributions in the dictionary are labeled with the index of the potential speaker location consistently in all frequency bins, which prevents the permutation problem.

The rest of this paper is organized as follows. Section 2 describes background. Section 3 elaborates on probabilistic spatial dictionary for mask estimation and diarization. Section 4 describes the proposed method. Section 5 presents a meeting ASR experiment on real meeting data. Section 6 concludes the paper.

2. BACKGROUND

2.1. Mask-based Microphone Array Signal Processing

We employ mask-based approach to microphone array signal processing. This approach has been applied to various tasks, such as source separation [2, 7–14], denoising [1, 2, 15], source localization [11, 16], and source counting [10, 17]. The approach has been employed in the best system in many evaluation campaign [1, 18, 19], which demonstrates its effectiveness and robustness in the real world. The approach employs source location features, such as time and level differences between microphones.

2.2. Feature Vector: Directional Statistics [8, 13]

Here we employ *directional statistics* [8, 12–15, 17, 20] as the source location features, which are known to be effective and robust in the real world [15], instead of the better-known time and level differences between microphones. In this paper, the directional statistics refer to a normalized vector $\boldsymbol{z}_{tf}$ in (1), where $\boldsymbol{y}_{tf}$ denotes the vector

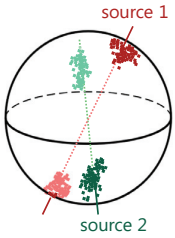

Fig. 1. Illustration of directional statistics for two sources. Here, $\mathbb{C}^M$ has been simplified to $\mathbb{R}^3$ for illustration.

composed of the observed signals at M microphones in the short-time Fourier transform (STFT) domain as in (2).

$$\boldsymbol{z}_{tf} \triangleq \frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|_2}. \tag{1}$$

$$\boldsymbol{y}_{tf} \triangleq \begin{bmatrix} y_{tf}^{(1)} & y_{tf}^{(2)} & \dots & y_{tf}^{(M)} \end{bmatrix}^{\mathsf{T}}. \tag{2}$$

As illustrated in Fig. 1, the directional statistics (1) lie on the unit hypersphere centered at the origin, and form cluster for each source signal. In the above, t denotes the frame index; f the frequency bin index; $^{\mathsf{T}}$ transposition;

$$\|\boldsymbol{y}_{tf}\|_2 \triangleq \sqrt{\sum_{m=1}^{M} |y_{tf}^{(m)}|^2}. \tag{3}$$

3. PROBABILISTIC SPATIAL DICTIONARY FOR MASK ESTIMATION AND DIARIZATION

3.1. Basic Idea of Probabilistic Spatial Dictionary

A real meeting involves reverberation and background noise, as well as speaker overlap and turn-taking, which make mask estimation and diarization challenging. To deal with such a challenging task, here we exploit prior knowledge about the feature vector (1). Specifically, we exploit a *probabilistic spatial dictionary* [21], a dictionary that consists of a pre-trained probability distribution of the feature vector for each potential speaker location. For example, when the speakers are seated around a table as in Fig. 2, we can choose the potential speaker locations as indicated by the dots. We pre-train the distributions in the dictionary on training data composed of single-channel reverberant speech signals for each potential speaker location. The single-channel reverberant speech signals can be obtained, *e.g.,* by recording a speech signal played by a loudspeaker at each potential speaker location, or by convolving a dry source signal and measured room impulse responses for each potential speaker location.

3.2. Modeling Observed Signals using Probabilistic Spatial Dictionary

Here we describe our modeling of the feature vector $\boldsymbol{z}_{tf}$ in (1) based on the probabilistic spatial dictionary.

Based on speech sparseness [7], we assume that, at each time-frequency point, $\boldsymbol{z}_{tf}$ is dominated by a speech signal from one of the potential speaker locations, or by the background noise. Let us define a speaker location indicator g_{tf}, which takes $k \in \{1, \dots, K\}$ if $\boldsymbol{z}_{tf}$ is dominated by a speech signal from the kth potential speaker location, and takes 0 if $\boldsymbol{z}_{tf}$ is dominated by the background noise.

We model $\boldsymbol{z}_{tf}$ by a complex Watson mixture model (cWMM) in (4), where the mixture components correspond to the cases $g_{tf} = 0, \dots, K$.

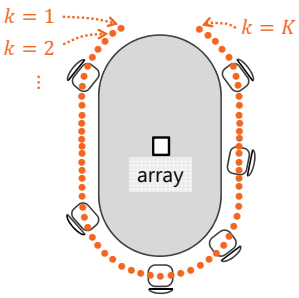

Fig. 2. Discrete index k indicating a potential speaker location.

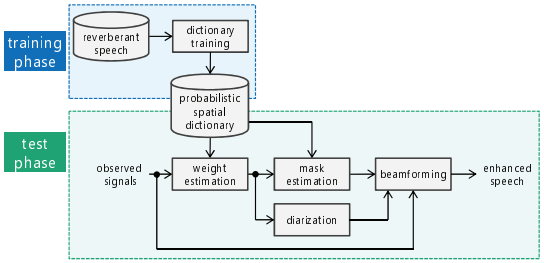

Fig. 3. Processing flow of the proposed method.

$$p(\boldsymbol{z}_{tf}) = \sum_{k=0}^{K} \alpha_t^{(k)} \mathcal{W}\big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\big). \tag{4}$$

In (4), $\mathcal{W}$ denotes a complex Watson distribution defined by (5); $\boldsymbol{a}_f^{(k)}$ a centroid satisfying $\|\boldsymbol{a}_f^{(k)}\|_2 = 1$; $\kappa_f^{(k)}$ a concentration parameter; $\alpha_t^{(k)}$ a mixture weight satisfying $\sum_{k=0}^{K} \alpha_t^{(k)} = 1$.

$$\mathcal{W}(\boldsymbol{z}; \boldsymbol{a}, \kappa) \triangleq \frac{(M-1)!}{2\pi^M \mathcal{K}(1, M, \kappa)} \exp\big(\kappa |\boldsymbol{a}^{\mathsf{H}} \boldsymbol{z}|^2\big). \tag{5}$$

In (5), $\mathcal{K}$ denotes the confluent hypergeometric function of the first kind, and $^{\mathsf{H}}$ Hermitian transposition.

Of the model parameters in (4), $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ are pre-trained from the single-speaker reverberant speech signals, and stored as the probabilistic spatial dictionary. On the other hand, $\alpha_t^{(k)}$ is estimated from a meeting recording, which can be regarded as speaker presence probability at the kth potential speaker location in the tth frame. Once $\alpha_t^{(k)}$ has been estimated, the posterior probability $\gamma_{tf}^{(k)} \triangleq P(g_{tf} = k | \boldsymbol{z}_{tf})$ of $g_{tf} = k$ can be estimated by

$$\gamma_{tf}^{(k)} = \frac{\alpha_t^{(k)} \mathcal{W}\big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\big)}{\sum_{l=0}^{K} \alpha_t^{(l)} \mathcal{W}\big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(l)}, \kappa_f^{(l)}\big)}. \tag{6}$$

The masks are obtained based on $\gamma_{tf}^{(k)}$ as detailed later. Diarization can be performed based on peak picking of $\alpha_t^{(k)}, 1 \le k \le K$.

4. PROPOSED METHOD

This section describes the processing of the proposed method, which consists of training and test phases (Fig. 3).

4.1. Training Phase

In the training phase, $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ of the cWMM (4) are pre-trained from the single-speaker *reverberant speech* (*dictionary training*), and stored as the *probabilistic spatial dictionary*.

4.1.1. Dictionary Training

In dictionary training, the feature vector (1) is first extracted from the single-speaker reverberant speech for each potential speaker location $k = 1, \ldots, K$, which is denoted by $\boldsymbol{z}_{tf}^{(k)}$. $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ are then estimated for $k = 1, \ldots, K$ by the maximization of the following likelihood function:

$$\prod_{t=1}^{T} \prod_{f=1}^{F} \mathcal{W}\big(\boldsymbol{z}_{tf}^{(k)}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\big). \tag{7}$$

Here, T denotes the number of frames, and F the number of frequency bins up to the Nyquist frequency. The resulting update rules are similar to those in [22] and omitted here.

Under the assumption that the background noise is isotropic, $\kappa_f^{(0)}$ is fixed to 0, in which case the Watson distribution (5) reduces to the uniform distribution on the unit hypersphere. Since the distribution is independent of $\boldsymbol{a}_f^{(0)}$ in this case, $\boldsymbol{a}_f^{(0)}$ can be set to an arbitrary unit vector.

$\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ for $k = 0, \ldots, K$ are stored as the probabilistic spatial dictionary.

4.2. Test Phase

In the test phase, beamforming is performed based on the probabilistic spatial dictionary. To estimate the steering vector for MVDR beamforming accurately from a mixture of multiple speech signals and background noise, we propose to exploit masks and diarization. The masks allow for estimation of the steering vector for each speaker from the mixture. Diarization indicates which speakers are active in each frame [3–6], and prevents the beamformer from degrading gradually while the corresponding speaker remains silent.

4.2.1. Weight Estimation

In *weight estimation*, the feature vector $\boldsymbol{z}_{tf}$ in (1) is first extracted from the observed signals. $\alpha_t^{(k)}$ is then estimated by the maximization of the following likelihood function:

$$\prod_{t=1}^{T} \prod_{f=1}^{F} \sum_{k=0}^{K} \alpha_t^{(k)} \mathcal{W}\big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\big). \tag{8}$$

The maximization can be performed by a gradient ascent-based algorithm that iterates the following update rules alternately:

$$\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t + \lambda \sum_{f=1}^{F} \frac{1}{\boldsymbol{\alpha}_t^{\mathsf{T}} \boldsymbol{w}_{tf}} \boldsymbol{w}_{tf}, \tag{9}$$

$$\boldsymbol{\alpha}_t \leftarrow \frac{1}{\mathbf{1}^{\mathsf{T}} \boldsymbol{\alpha}_t} \boldsymbol{\alpha}_t. \tag{10}$$

Here, $\lambda > 0$ denotes a learning rate; $\boldsymbol{\alpha}_t \triangleq \begin{bmatrix} \alpha_t^{(0)} & \ldots & \alpha_t^{(K)} \end{bmatrix}^{\mathsf{T}}$; $\boldsymbol{w}_{tf} \triangleq \begin{bmatrix} w_{tf}^{(0)} & \ldots & w_{tf}^{(K)} \end{bmatrix}^{\mathsf{T}}$; $w_{tf}^{(k)} \triangleq \mathcal{W}\big(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\big)$; $\mathbf{1} \triangleq \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^{\mathsf{T}}$. $\boldsymbol{\alpha}_t$ can be initialized by, *e.g.*, $\boldsymbol{\alpha}_t \leftarrow \frac{1}{K+1}\mathbf{1}$.

4.2.2. Diarization

In *diarization*, a binary variable $d_t^{(n)}$ indicating the diarization result is estimated based on $\alpha_t^{(k)}$, where $n = 1, \ldots, N$ denote the indexes of the seats, and N the number of seats. $d_t^{(n)}$ takes 1 if there is voice activity from the nth seat in the tth frame, and 0 otherwise.

First, the peaks of $\alpha_t^{(k)}, 1 \leq k \leq K$, are picked to detect source activity. Specifically, the set $I_t (\subset \{1, \ldots, K\})$ that consists of all values of $k \in \{1, \ldots, K\}$ at which $\alpha_t^{(k)}$ has a peak is estimated.

Then, the K potential speaker locations are grouped into N classes corresponding to the N seats based on tracking of the peaks. Specifically, disjoint sets $J_t^{(n)}, n = 1, \ldots, N$, satisfying $\cup_{n=1}^{N} J_t^{(n)} = \{1, \ldots, K\}$ are estimated. $J_t^{(n)}$ is the set of all values of k that correspond to the nth seat. The initial value $J_0^{(n)}$ of $J_t^{(n)}$ is computed based on seat locations, which are assumed to be roughly known.

Finally, $d_t^{(n)}$ is computed by

$$d_t^{(n)} \leftarrow \begin{cases} 1, I_t \cap J_t^{(n)} \neq \emptyset, \\ 0, I_t \cap J_t^{(n)} = \emptyset. \end{cases} \tag{11}$$

4.2.3. Mask Estimation

In *mask estimation*, $\gamma_{tf}^{(k)}$ is first estimated for $k = 0, \ldots, K$ by (6), and the masks $\mathcal{M}_{tf}^{(n)}, n = 0, \ldots, N$, are then estimated by

$$\mathcal{M}_{tf}^{(n)} \leftarrow \begin{cases} \sum_{k \in J_t^{(n)}} \gamma_{tf}^{(k)}, & \text{if } n \in \{1, \ldots, N\}, \\ \gamma_{tf}^{(0)}, & \text{if } n = 0. \end{cases} \tag{12}$$

Here, $\mathcal{M}_{tf}^{(n)}, n \in \{1, \ldots, N\}$, denotes the mask for the nth seat, and $\mathcal{M}_{tf}^{(0)}$ that for the background noise.

4.2.4. Beamforming

In *beamforming*, MVDR beamforming is performed based on $d_t^{(n)}$ and $\mathcal{M}_{tf}^{(n)}$. $d_t^{(n)}$ and $\mathcal{M}_{tf}^{(n)}$, which enable accurate estimation of the steering vector. The steering vector represents the acoustic transfer characteristics from a sound source to the microphones, and is the key to the MVDR beamforming.

In each frame, the covariance matrix of the observed signals, $\boldsymbol{\Phi}_{tf}^{\boldsymbol{y}} \triangleq E\big[\boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}}\big]$, is updated by

$$\boldsymbol{\Phi}_{tf}^{\boldsymbol{y}} \leftarrow (1 - \beta) \boldsymbol{\Phi}_{t-1,f}^{\boldsymbol{y}} + \beta \boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}}, \tag{13}$$

where $\beta \in [0, 1]$ denotes a forgetting factor. A covariance matrix corresponding to the background noise, $\boldsymbol{\Phi}_{tf}^{(0)}$, is updated by

$$\boldsymbol{\Phi}_{tf}^{(0)} \leftarrow (1 - \beta) \boldsymbol{\Phi}_{t-1,f}^{(0)} + \beta \mathcal{M}_{tf}^{(0)} \boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}}. \tag{14}$$

A covariance matrix corresponding to the speech signal from the nth seat, $\boldsymbol{\Phi}_{tf}^{(n)}, n \in \{1, \ldots, N\}$, is updated by

$$\boldsymbol{\Phi}_{tf}^{(n)} \leftarrow \begin{cases} (1 - \beta) \boldsymbol{\Phi}_{t-1,f}^{(n)} + \beta \mathcal{M}_{tf}^{(n)} \boldsymbol{y}_{tf} \boldsymbol{y}_{tf}^{\mathsf{H}}, & \text{if } d_t^{(n)} = 1, \\ \boldsymbol{\Phi}_{t-1,f}^{(n)}, & \text{if } d_t^{(n)} = 0. \end{cases} \tag{15}$$

(15) implies that $\boldsymbol{\Phi}_{tf}^{(n)}, n \in \{1, \ldots, N\}$, is updated only if $d_t^{(n)} = 1$. Such update control is crucial in meetings to prevent the beamformer from degrading gradually while the corresponding speaker remains silent.

Then, an eigenvector $\boldsymbol{u}_{tf}^{(n)}$ corresponding to the largest eigenvalue of the generalized eigenvalue problem

$$\boldsymbol{\Phi}_{tf}^{(n)} \boldsymbol{u} = \mu \boldsymbol{\Phi}_{tf}^{(0)} \boldsymbol{u} \tag{16}$$

Table 1. Meeting data

	office-exh.	office	sound-proof
reverberation time	500 ms	350 ms	120 ms
SNR	3–15 dB	15–20 dB	20–25 dB
# of speakers	4–6	4	4
training data	40 sessions	14 sessions	30 sessions
development data	8 sessions	4 sessions	4 sessions
evaluation data	8 sessions	8 sessions	8 sessions
# of microphones	8	8	8

Table 2. ASR performance in terms of WER (%).

	microphone	AM	enhancement	WER
(a)	headset	headset	off	15.9
(b)	headset	multi-condition	off	18.8
(c)	table	headset	off	93.5
(d)	table	multi-condition	off	40.5
(e)	table	multi-condition	on [23]	53.3
(f)	table	multi-condition	on (masking)	57.5
(g)	table	multi-condition	on (MVDR)	**24.1**

is computed for $n = 1, \ldots, N$. The steering vector $\boldsymbol{h}_{tf}^{(n)}$ for the nth seat is estimated by

$$\boldsymbol{h}_{tf}^{(n)} \leftarrow \boldsymbol{\Phi}_{tf}^{(0)} \boldsymbol{u}_{tf}^{(n)}, \tag{17}$$

$$\boldsymbol{h}_{tf}^{(n)} \leftarrow \frac{\boldsymbol{h}_{tf}^{(n)}}{h_{tf}^{(1,n)}}, \tag{18}$$

where $h_{tf}^{(1,n)}$ denotes the first entry of $\boldsymbol{h}_{tf}^{(n)}$. Finally, the speech signal from the nth seat, $s_{tf}^{(n)}$, is estimated by using the MVDR beamformer as

$$s_{tf}^{(n)} \leftarrow \frac{\boldsymbol{h}_{tf}^{(n)\mathsf{H}} (\boldsymbol{\Phi}_{tf}^{\boldsymbol{y}})^{-1} \boldsymbol{y}_{tf}}{\boldsymbol{h}_{tf}^{(n)\mathsf{H}} (\boldsymbol{\Phi}_{tf}^{\boldsymbol{y}})^{-1} \boldsymbol{h}_{tf}^{(n)}}. \tag{19}$$

The steering vector estimation here enables noise-robust estimation owing to the generalized eigenvalue problem. In [1], we have also proposed steering vector estimation based on subtraction of covariance matrices. A main advantage of the method here over that in [1] is that the former works well even if the noise component in $\boldsymbol{\Phi}_{tf}^{(n)}$ and that in $\boldsymbol{\Phi}_{tf}^{(0)}$ have different scales, which is the case in the diarization-based approach here, because $\boldsymbol{\Phi}_{tf}^{(n)}$ and $\boldsymbol{\Phi}_{tf}^{(0)}$ are updated in different frames.

5. EXPERIMENTAL EVALUATION

5.1. Dataset

As in Table 1, we employed meeting datasets recorded in the following three different environments:

- office-exh.: an office room next to an exhibition hall;
- office: a quiet office room;
- sound-proof: a sound-proof room.

Only the first dataset was used for evaluation, while the latter two were used only to train the ASR system. The recordings in the first dataset may contain babble noise from audience standing behind meeting participants and from loudspeakers in the exhibition hall, depending on sessions. Also, the door of the office room is either open or close, depending on sessions.

We also employed a dataset composed of single-speaker reverberant speech signals to pre-train the probabilistic spatial dictionary. This dataset was generated by convolving a dry speech signal with room impulse responses measured in office-exh. for each of $K = 22$ potential speaker locations around a table (see Fig. 2).

5.2. Back-end System

We employed a DNN-HMM acoustic model (AM) [24] with seven hidden layers with 2048 units each. The input to the DNN was a 1320-dimensional feature vector that consisted of 40 log-mel filterbank coefficients and their delta and acceleration with five left and five right context frames. The output was 4100 HMM states. The DNN was first trained on the clean training set of Corpus of Spontaneous Japanese (CSJ) [25], then retrained on the headset recordings in the training set of the three meeting datasets in Table 1 (headset AM in Table 2), and finally retrained on all channels of the array recordings in the training set of the office-exh. dataset (multi-condition AM in Table 2). In the training of the multi-condition AM, we used HMM state alignments obtained by using the headset recordings.

We employed a Kneser-Ney smoothed word trigram language model [26], trained on the CSJ, the training set of the meeting data, and topic-related WWW data. The mixture weights were determined based on the minimization of the perplexity on the development set of the meeting data.

We used manual annotation for voice activity detection for ASR. ASR was performed in an utterance batch instead of an online manner, while the meeting speech enhancement was performed in an online manner.

5.3. Results

Table 2 shows the meeting recognition result in terms of the word error rate (WER). (a) and (b) show the results for headset recordings as the performance upper bound, and (c)–(g) those for the array recordings. (c) and (d) are results without enhancement, where (c) corresponds to the headset AM and (d) to the multi-condition AM. (e), (f), and (g) are results with enhancement, where (e) corresponds to our previous beamforming [23], (f) to masking with masks estimated based on the probabilistic spatial dictionary, and (g) to the proposed method.

Although the WER for (d) was much lower than that for (c), it was still as high as 40.5 %. Our previous beamforming in (e) and masking in (f) gave even high WERs than (d) without enhancement. The proposed method (*i.e.*, MVDR beamforming with masks estimated based on the probablistic spatial dictionary) in (g) gave the WER of 24.1 %, reducing the WER by $54.8\,\%$ relative to our previous beamforming in (e). The real-time factor was 0.6.

6. CONCLUSION

We proposed an online beamforming method for meeting recognition in noisy and reverberant environments based on the probabilistic spatial dictionary.

The proposed method as it is requires the single-speaker reverberant speech signals for the same room as the meeting takes place. Otherwise, the performance of meeting enhancement may degrade due to environment mismatch between training and test data. We plan to extend the proposed method so that it can adapt to a new environment.

7. REFERENCES

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*, Dec. 2015, pp. 436–443.

[2] S. Araki, M. Okada T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. ICASSP*, Mar. 2016, pp. 385–389.

[3] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998.

[4] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macías-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. ICASSP NIST Meeting Recognition Workshop*, 2004.

[5] 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.

[6] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *Proc. ASRU*, Dec. 2007, pp. 238–247.

[7] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. SP*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[9] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, Oct. 2007, pp. 147–150.

[10] 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*, Apr. 2009, pp. 33–36.

[11] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[12] 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.

[13] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[14] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. ICASSP*, Mar. 2012, pp. 253–256.

[15] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. ASLP*, vol. 21, no. 12, pp. 2516–2531, Dec. 2013.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, vol. 63, no. 3, pp. 265–275, June 2011.

[17] L. Drude, A. Chinaev, D.H. Tran Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *Proc. ICASSP*, May 2014, pp. 6834–6838.

[18] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. CHiME 2011 Workshop on Machine Listening in Multisource Environments*, Sept. 2011, pp. 12–17.

[19] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Workshop*, May 2014.

[20] L. Drude, F. Jacob, and R. Haeb-Umbach, "DOA-estimation based on a complex Watson kernel method," in *Proc. EUSIPCO*, Sept. 2015.

[21] 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.

[22] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238–3242.

[23] 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.

[24] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *Proc. Interspeech*, 2013, pp. 2992–2996.

[25] S. Furui, K. Maezawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proc. ISCA ASR*, 2000.

[26] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.