SPATIAL CORRELATION MODEL BASED OBSERVATION VECTOR CLUSTERING AND MVDR BEAMFORMING FOR MEETING RECOGNITION

Shoko Araki[†] *Masahiro Okada*[†,‡] *Takuya Higuchi*[†] *Atsunori Ogawa*[†] *Tomohiro Nakatani*[†]

† NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
‡ Graduate School of Design, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka-shi, Fukuoka, 815-8540, Japan

ABSTRACT

This paper addresses a minimum variance distortionless response (MVDR) beamforming based speech enhancement approach for meeting speech recognition. In a meeting situation, speaker overlaps and noise signals are not negligible. To handle these issues, we employ MVDR beamforming, where accurate estimation of the steering vector is paramount. We recently found that steering vector estimation by clustering the time-frequency components of microphone observation vectors performs well as regards real-world noise reduction. The clustering is performed by taking a cue from the spatial correlation matrix of each speaker, which is realized by modeling the time-frequency components of the observation vectors with a complex Gaussian mixture model (CGMM). Experimental results with real recordings show that the proposed MVDR scheme outperforms conventional null-beamformer based speech enhancement in a meeting situation.

Index Terms— Minimum variance distortionless response (MVDR), speech enhancement, meeting speech recognition, diarization, complex Gaussian mixture model (CGMM)

1. INTRODUCTION

In recent years, the use of voice-operable smart-phones and tablets has become widespread, and their usefulness has been widely recognized. When a user speaks carefully into a terminal, that is, a microphone(s), his/her voice is usually accurately recognized, and the device works as expected. On the other hand, there is a growing need for voice interfaces that can work when a user speaks at a certain distance from the microphones. One typical scenario is a group conversation in a meeting, where we may want to avoid the use of headset microphones and employ microphones on the table. In such a meeting scenario, the interference from acoustic noise and reverberation is not negligible. Moreover, in an informal relaxed conversation, the utterances of speakers sometimes overlap. To achieve high speech recognition accuracy in such a situation, we must take account of the reverberation, utterance overlaps, and noise. That is, speech enhancement plays an important role for meeting recognition. This paper describes a beamforming based speech enhancement approach for meeting recognition.

Meeting recognition has long been studied [1–8], and one major example of speech enhancement for meetings is Wiener filter based single channel noise reduction followed by delay-and-sum beamforming (e.g., [9]). We have also developed a prototype meeting recognizer [10] for a small party meeting conversation, where we employed a microphone array at the center of the table. In the system, we employed multi-channel dereverberation and null-steering based beamforming for speech enhancement. The system worked well with a GMM-HMM based speech recognizer at that time. However, we recently realized that the null-steering based beamforming used in our old prototype cannot work well with a state-of-the-art DNN-HMM based speech recognition system. This motivated us to develop more sophisticated speech enhancement techniques.

As a state-of-the-art speech enhancement technique, this paper employs a new minimum variance distortionless response (MVDR) beamforming technique, and investigates its performance in meeting scenarios. For MVDR beamforming, accurate and blind estimation of the steering vectors is essential. With conventional MVDR beamforming, the steering vectors are given by an (estimated) speaker direction and the microphone array geometry, which is not always given/accurate especially in a meeting situation. Moreover, speaker positions tend to slowly change during a meeting. We recently found that steering vector estimation by clustering the time-frequency components of the observation vectors performs well as regards noise reduction of real-world recordings [11, 12]. The clustering is performed by taking a cue from the spatial correlation matrix of each speaker location [13–15], which is realized by modeling the time-frequency components of the microphone observation vectors with a complex GMM (CGMM). This paper describes the application of our above-mentioned MVDR beamforming scheme to real recordings of multi-speaker meeting conversations. The proposed method has an ability to estimate steering vectors blindly, without relying on an (estimated) speaker direction and the microphone array geometry. Full-batch and block-batch modes of the proposed approach will also be presented.

The rest of this paper is organized as follows: Section 2 describes the speech enhancement task in a meeting, and Sec. 3 explains the proposed approach. Section 4 reports the experimental results, and Sec. 5 concludes the paper.

2. PROBLEM FORMULATION

Let $s_k(t,f)$ be the short time Fourier transform (STFT) coefficient of a speech source of speaker k, and $\mathbf{h}_k(f) = [h_{1,k}, \cdots, h_{M,N}]^T$ be its steering vector, where t and f are the time and frequency indices, respectively. Then, the observation vector $\mathbf{y}(t,f) = [y_1(t,f), \cdots, y_M(t,f)]^T$ at M microphones becomes

$$\mathbf{y}(t,f) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(t,f) + \mathbf{n}(t,f), \tag{1}$$

where $\mathbf{n}(t,f)$ is noise observed at the microphones.

In this paper, we assume that $M \geq N$. We also assume that the speakers are seated during a meeting, and therefore the steering

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

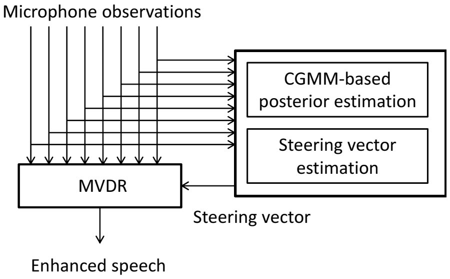

Fig. 1. Algorithm block diagram

vector $\mathbf{h}_k(f)$ does not change greatly. The objective of this paper is to estimate the speech source $\hat{s}_k(t,f)$ of each speaker k from the observations $\mathbf{y}(t,f)$.

3. ALGORITHM

3.1. Basic scheme

Figure 1 shows the fundamental scheme of our algorithm. The main part is the MVDR beamformer, where its steering vectors are estimated with CGMM-based clustering.

The MVDR beamforming coefficients are calculated by

$$\mathbf{w}_k(f) = \frac{\mathbf{R}_{\mathbf{yy}}^{-1}(f)\hat{\mathbf{h}}_k(f)}{\hat{\mathbf{h}}_k^H(f)\mathbf{R}_{\mathbf{yy}}^{-1}(f)\hat{\mathbf{h}}_k(f)}, \tag{2}$$

where $\hat{\mathbf{h}}_k$ is the estimated steering vector of speaker k, and $\cdot^H$ denotes the conjugate transpose of a vector. By using these coefficients, the enhanced speech estimate is obtained by

$$\hat{s}_k(t,f) = \mathbf{w}_k^H(f)\mathbf{y}(t,f). \tag{3}$$

In (2), $\mathbf{R}_{\mathbf{yy}}(f)$ is easily calculated with the observation vectors: $\mathbf{R}_{\mathbf{yy}}(f) = \sum_t \mathbf{y}(t,f)\mathbf{y}^H(t,f)$. On the other hand, accurate estimation of the steering vector $\mathbf{h}_k$ is essential, but it is a challenging problem in a meeting situation.

The following subsections explain the approach for estimating steering vectors.

3.1.1. *Steering vector estimation*

We estimate the steering vector $\mathbf{h}_k(f)$ of each speaker k by computing the principal eigenvector of the correlation matrix $\mathbf{R}_k(f)$ of observations when only speaker k speaks. Assuming the independence of speech and noise, and the sparseness of each speaker, i.e., only one speaker utterance is dominant at each time-frequency slot, such a correlation matrix can be estimated as

$$\mathbf{R}_k(f) = \mathbf{R}_{k+n}(f) - \mathbf{R}_n(f) \tag{4}$$

where $\mathbf{R}_{k+n}(f)$ and $\mathbf{R}_n(f)$ are the correlation matrices of noisy speech of speaker k and noise, respectively. They can be estimated by

$$\mathbf{R}_{k+n}(f) = \frac{1}{\sum_{t=1}^T M_k(t,f)}\sum_{t=1}^T M_k(t,f)\mathbf{y}(t,f)\mathbf{y}^H(t,f) \tag{5}$$

$$\mathbf{R}_n(f) = \frac{1}{\sum_{t=1}^T M_n(t,f)}\sum_{t=1}^T M_n(t,f)\mathbf{y}(t,f)\mathbf{y}^H(t,f) \tag{6}$$

where $M_k(t,f)$ and $M_n(t,f)$ denote the posterior probability of speaker k and noise existence in each time-frequency slot, respectively.

In the next subsection, we explain how we can estimate the posterior probabilities for obtaining correlation matrices $\mathbf{R}_k(f)$ in (4).

3.1.2. *Spatial correlation model based posterior estimation*

We assume that the source $s_k(t,f)$ and noise $\mathbf{n}(t,f)$ follow a Gaussian distribution of zero mean and a variance $|s_k(t,f)|^2 = \phi_{tfk}$:

$$p(s_k(t,f);\phi_{tfk}) = \mathcal{N}(0,\phi_{tfk}). \tag{7}$$

Then, the observation vector follows a complex Gaussian mixture model (CGMM) [15]:

$$p(\mathbf{y}(t,f);\theta) = \sum_{k=1}^{N+1}\alpha_{fk}p(\mathbf{y}(t,f)|C(t,f)=k;\theta) \tag{8}$$

$$p(\mathbf{y}(t,f)|C(t,f)=k;\theta) = \mathcal{N}_c(0,\phi_{tfk}\mathbf{B}_{fk}), \tag{9}$$

where α_{fk} is a mixture weight ($\sum_k^{N+1}\alpha_{fk} = 1$), and $\mathbf{B}_{fk} = \hat{\mathbf{h}}_k(f)\hat{\mathbf{h}}_k^H(f)$ is the spatial correlation matrix of source k. $C(t,f) = k(k=1,\cdots,N)$ corresponds to the source classes, and $C(t,f) = N+1$ corresponds to a noise class.

The log likelihood function is defined as

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_t\sum_f \log p(\mathbf{y}(t,f);\theta) \\ &= \sum_t\sum_f \log\sum_k \alpha_{fk}\mathcal{N}_c(0,\phi_{tfk}\mathbf{B}_{fk}),\end{aligned}$$

where $\theta = \{\theta_k\} = \{\{\phi_{tfk},\mathbf{B}_{fk},\alpha_{fk}\}\}$ is a parameter set.

We can maximize the log likelihood function by using the Expectation-Maximization (EM) algorithm. The Q function is given as

$$Q = \sum_t\sum_f\sum_{k=1}^{N+1} p(C(t,f)=k|\mathbf{y}(t,f),\theta)\log\alpha_{fk}\mathcal{N}_c(0,\phi_{tfk}\mathbf{B}_{fk}) \tag{10}$$

Hereafter, we denote the posterior probability with $M_k(t,f) = p(C(t,f)=k|\mathbf{y}(t,f),\theta)$. This posterior probability is required by the steering vector estimation in Sec. 3.1.1. The posterior for noise $M_n(t,f)$ in Sec. 3.1.1 is given by $M_{N+1}(t,f)$.

The Q function is maximized by iterating the following E- and M-steps, and the posterior is obtained in the E-step.

E-step: We calculate the posterior:

$$M_k(t,f) = p(C(t,f)=k|\mathbf{y}(t,f),\theta) \tag{11}$$

$$= \frac{\alpha_{fk}p(\mathbf{y}(t,f)|\theta_k)}{\sum_k\alpha_{fk}p(\mathbf{y}(t,f)|\theta_k)} \tag{12}$$

M-step: From the Q function, the update rules of the parameters θ are obtained as:

$$\phi_{tfk} = \frac{1}{M}\mathbf{y}^H(t,f)\mathbf{B}_{fk}^{-1}\mathbf{y}(t,f) \tag{13}$$

$$\mathbf{B}_{fk} = \frac{\sum_t^T\frac{M_k(t,f)}{\phi_{tfk}}\mathbf{y}(t,f)\mathbf{y}^H(t,f)}{\sum_t^T M_k(t,f)} \tag{14}$$

$$\alpha_{fk} = \frac{1}{T}\sum_t^T M_k(t,f) \tag{15}$$

3.2. Posterior estimation with pre-trained spatial correlation matrix

This relates to a natural situation where the seating configuration in a meeting room is fixed. In such a case, it would be useful to employ a pre-trained spatial correlation matrix $\mathbf{B}_{fk}^{trained}$ for a robust posterior calculation. The pre-trained matrix $\mathbf{B}_{fk}^{trained}$ can be calculated by applying E- and M-steps to a set of training data.

We may fix the spatial correlation matrix (14) to $\mathbf{B}_{fk}^{trained}$ without updating it. As another option, we can also adapt the spatial correlation matrix to the current meeting data as follows

$$\mathbf{B}_{fk} = \eta \frac{\sum_t \frac{M_k(t,f)}{\phi_{tfk}} \mathbf{y}(t,f)\mathbf{y}^H(t,f)}{\sum_t M_k(t,f)} + (1-\eta)\mathbf{B}_{fk}^{trained}, \quad (16)$$

where η is an adaptation parameter.

3.3. Block batch implementation

In the previous section, we implicitly assume that we can use the whole recording of a meeting. That is, we discussed the full batch version of our MVDR beamforming, which can be utilized for an off-line mode. On the other hand, block batch or online algorithms are also useful when real-time processing is required. Here, we discuss a block batch implementation.

In a block batch mode, MVDR beamformer coefficients are updated every F frames ($F = 100$ in the next section). More concretely, we calculate the correlation matrix (4) and the steering vectors $\mathbf{h}_k(f)$ to update MVDR beamformer coefficients (2) every F frames. Even in the block batch mode, parameters for estimating the posterior probability (12) should be calculated at all the time-frequency slots to calculate the correlation matrix (4). Therefore, the parameter updates (12), (13), (14) and (15) in the E- and M-steps are performed in every time-frequency slot. In this paper we update the parameters only once (i.e., one iteration) at each time frame.

It is worth mentioning that, in the M-step, the spatial correlation matrix at time T can be estimated in an online manner:

$$\begin{aligned}\mathbf{B}_{fk,T} \quad = \quad & \frac{\sum_t^{T-1} M_k(t,f)}{\sum_t^{T-1} M_k(t,f) + M_k(T,f)} \mathbf{B}_{fk,T-1} \\ & + \frac{\frac{M_k(T,f)}{\phi_{tfk}} \mathbf{y}(T,f)\mathbf{y}^H(T,f)}{\sum_t^{T-1} M_k(t,f) + M_k(T,f)}\end{aligned}$$

4. EXPERIMENTS

4.1. Experimental setups

We conducted experiments to evaluate our proposed approach.

We recorded several sessions of spontaneous Japanese meeting conversation with an 8-element microphone array at the center of the table (see Fig. 2). The recordings were made in two different rooms: an office and a sound-proof room. The reverberation times and SNR conditions are summarized in Table 1. As references, we also recorded the meetings with headset microphones.

In the meetings, four participants were seated around a table (Fig. 2) and freely discussed a given topic, which was selected from 28 themes. The participants did not change their seats during the session. Each session lasted approximately 15-25 minutes. The recordings are divided into training, development and evaluation sets as shown in Table 1.

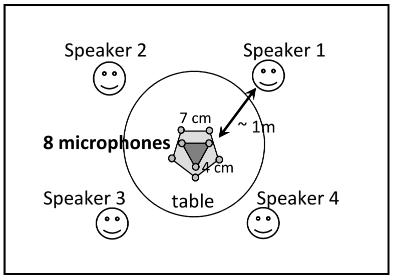

Fig. 2. Recording room

Table 1. Recording room conditions and data setups

	office	sound-proof
Reverberation time	350 msec.	120 msec.
SNR	15-20 dB	20-25 dB
Evaluation set	8 sessions	8 sessions
Training set for calculating $\mathbf{B}_{fk}^{trained}$ (subset of training set for ASR)	10 sessions	12 sessions
Training set for ASR	14 sessions	30 sessions
Development set for ASR	4 sessions	4 sessions

4.1.1. Algorithm setups

We first dereverberate the microphone observation vectors [16], and then employ the proposed MVDR beamforming.

We investigated four setups for the proposed algorithm:

(Ex1) Full batch mode: The MVDR coefficients were calculated by using the whole recording of each meeting.

(Ex2) Block batch MVDR: The spatial correlation matrix $\mathbf{B}_{fk}$ was estimated for each meeting in full batch mode, on the other hand, the MVDR coefficients were updated every 100 frames.

(Ex3) Block batch with pre-trained spatial correlation matrix: By using the pre-trained spatial correlation matrix $\mathbf{B}_{fk}^{trained}$, which was trained with training data (see Table 1), the MVDR coefficients were updated every 100 frames ($\eta = 0$ in (16)).

(Ex4) Block batch with spatial correlation matrix adaptation: The pre-trained spatial correlation matrix $\mathbf{B}_{fk}^{trained}$ was employed and the spatial correlation matrix was adapted to the meeting recordings ($\eta = 0.1$ in (16)). The MVDR coefficients were updated every 100 frames.

We compare the proposed method with our conventional beamformer [10]:

$$\mathbf{W}(f) \quad = \quad \mathbf{H}^{-}(f),$$

where the kth row of $\mathbf{W}(f)$ is the beamformer coefficient $\mathbf{w}_k(f)$, the (m,n)th component of $\mathbf{H}(f)$ is $h_{mn}(f) = E\left[\frac{y_m(t,f)}{y_1(t,f)}\right]_{t\in\{C(t)=k\}}$, $^{-}$ denotes a pseudo inverse, and $C(t) = k$ is estimated by clustering the direction of arrival estimates. The beamformer coefficients were estimated in a block batch mode, i.e., every 100 frames.

4.1.2. ASR evaluation setups

We evaluate the speech enhancement performance in terms of the word error rate (WER) of the evaluation set. For the speech recognition, we utilized our DNN-based automatic speech recognition (ASR) back-end system [17], where we employed 40 log mel filterbank coefficients with their delta and acceleration, and 5 left and 5 right context windows as the DNN input. The DNN structure had

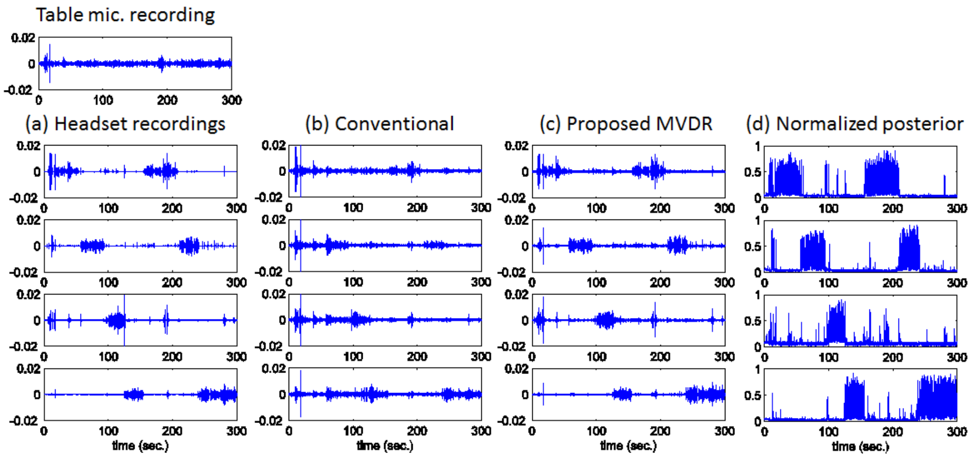

Fig. 3. Example waveforms of (a) office recordings, and processed speech signals with (b) conventional method and (c) proposed method (Ex1). (d) shows example normalized posteriors corresponding to (c), which was calculated by $\sum_f M_k(t,f)/N_F$ in (17).

Table 2. Meeting recognition results (WER [%]) for two window sizes with a half shift. ”sound-p.” denotes ”sound-proof”,

	window size = 64 msec.			window size = 32 msec.		
	office	sound-p.	ave.	office	sound-p.	ave.
(1) headset mic.	17.6	24.6	21.3	17.6	24.6	21.3
(2) table mic.	92.6	65.5	79.1	92.6	65.5	79.1
(3) conventional	67.8	48.6	58.2	-	-	-
(3’) conventional	56.6	58.3	57.5	-	-	-
(4) proposed (Ex1)	49.9	42.0	46.0	45.9	39.8	42.9
(5) proposed (Ex2)	48.6	42.4	45.5	47.3	42.3	44.8
(6) proposed (Ex3)	52.5	44.1	48.3	47.9	40.5	44.2
(7) proposed (Ex4)	52.0	43.9	48.0	47.2	40.4	43.8

7 hidden layers (2048 units each) and 4100 output HMM states. We trained an acoustic model with headset recordings from the training dataset (see Table 1). Note that we did not retrain the DNN with the enhanced speech. As the voice activity detection for ASR, we used manual annotation.

We used a Kneser-Ney smoothed word trigram language model [18], which was trained with transcripts of Japanese lecture speech data from the Corpus of Spontaneous Japanese (CSJ) [19] and the training set of the meeting recordings, in addition to the topic-related WWW data. These three text sets were mixed with weights that minimize the perplexity for the meeting development set.

4.2. Experimental results and discussion

Table 2 summarizes the speech recognition results in terms of WER. In Table 2, (1) and (2) show the WERs of headset and table microphone observations. With conventional beamforming (3), the WER is still poor. We also show conventional beamforming with our old GMM-based ASR system [10] as (3’). We can see that the WER of (3) is worse than that of (3’), i.e., our conventional beamforming does not perform well with the DNN-based ASR system. This motivated us to utilize a new beamforming technique.

The results from (4) to (7) show the WERs with our proposed beamforming approach. We confirmed that the proposed beamforming technique outperforms conventional beamforming. The block batch approaches (5), (6) and (7) performed quite well: their performance did not become very poor compared with the full-batch mode (4). Moreover, it can be seen that the use of the pre-trained spatial correlation matrix works well ((6) and (7)), and the matrix adaptation (7) improves the recognition performance.

Table 3. Diarization performance (%)

	DER	MST	FST	MST
conventional [10]	33.9	28.7	3.6	1.6
proposed (Ex1)	15.9	11.9	3.2	0.8
proposed (Ex4)	18.3	15.1	2.4	0.8

Figure 3 shows the example waveforms of (b) conventional and (c) proposed beamforming.

By using the estimated posterior $M_k(t,f)$, we can also perform speaker diarization, that is, we can estimate ”who speaks when”. One way to perform diarization is

$$\mathrm{dia}_k(t) = \begin{cases} 1 & \text{if } \sum_f M_k(t,f)/N_F > \text{threshold} \\ 0 & \text{otherwise} \end{cases} \qquad (17)$$

where N_F is the number of frequency bins. An example of this normalized posterior is shown in Fig. 3 (d). The diarization performance with threshold = 0.2 is summarized in Table 3, where the diarization error rate (DER), missed speaker time (MST), false-alarm speech time (FST) and speaker error time (SET) [20] were evaluated. We can see that posterior based diarization outperforms our previous approach [10].

5. CONCLUSION

This paper proposed an MVDR beamforming scheme for multi-speaker meeting conversation. For accurate steering vector estimation, we utilize a CGMM-based clustering of the observation vectors. We confirmed that the proposed MVDR approach performs well even in a block batch mode, and outperforms conventional null-beamformer based speech enhancement in a real meeting situation. We also confirmed that we can perform speaker diarization with the posterior probability, which is estimated by CGMM-based clustering. Future work includes the online extension of the proposed MVDR beamforming, DNN-AM retraining with enhanced speech for ASR, and an attempt to deal with more dynamic meeting situations.

6. REFERENCES

[1] A. Waibel, M. Bett, and M. Finke, "Meeting browser: tracking and summarizing meetings," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 281–286.

[2] 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'04 Meeting Recognition Workshop*, 2004.

[3] P. Wellner, M. Flynn, and M. Guillemot, "Browsing recorded meetings with Ferret," in *Proc. ICMI-MLMI*, 2004, pp. 12–21.

[4] F. Asano, K. Yamamoto, J. Ogata, M. Yamada, and M. Nakamura, "Detection and separation of speech events in meeting recordings using a microphone array," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, 2007, Article ID 27616, doi:10.1155/2007/27616.

[5] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings The AMI and AMIDA projects," in *Proc. ASRU'07*, 2007, pp. 238–247.

[6] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system," in *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, 2008.

[7] "AMI: Augmented Multi-party Interaction," Available online: http://www.amiproject.org/ami-scientific-portal.

[8] "Rich Transcription Evaluation Project," Available online: http://www.itl.nist.gov/iad/mig/tests/rt/.

[9] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proc. Workshops CLEAR 2007 and RT 2007*, 2008, pp. 509–519.

[10] 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. Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 499–513, 2012.

[11] 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 of ASRU2015*, 2015.

[12] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc of ICASSP2016*, 2016, (to appear).

[13] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[14] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1369–1380, 2013.

[15] N. Ito, S. Araki, T. Nakatani, and T. Yoshioka, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc of IWANEC2014*, 2014, pp. 269–273.

[16] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[17] 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. of Interspeech2013*, 2013, pp. 2992 –2996.

[18] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. of ICASSP'95*, 1995, pp. 181–184.

[19] S. Furui, K. Maezawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proc of ISCA ASR*, 2000, pp. 244–248.

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