ROBUST MVDR BEAMFORMING USING TIME-FREQUENCY MASKS FOR ONLINE/OFFLINE ASR IN NOISE

Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{higuchi.takuya, ito.nobutaka, yoshioka.takuya, nakatani.tomohiko}@lab.ntt.co.jp

ABSTRACT

This paper considers acoustic beamforming for noise robust automatic speech recognition (ASR). A beamformer attenuates background noise by enhancing sound components coming from a direction specified by a steering vector. Hence, accurate steering vector estimation is paramount for successful noise reduction. Recently, a beamforming approach was proposed that employs time-frequency masks. In the speech recognition system we submitted to the CHiME-3 Challenge, we employed a new form of this approach that uses a speech spectral model based on a complex Gaussian mixture model (CGMM) to estimate the time-frequency masks and the steering vector without providing technical details. This paper elaborates on this technique and examines its effectiveness for ASR. Experimental results show that the CGMM-based approach outperforms a recently proposed mask estimator based on a Watson mixture model. In addition, the CGMM-based approach is extended to an online speech enhancement scenario, which allows this technique to be used in an online recognition setup. This online version reduces the CHiME-3 evaluation error rate from 15.60% to 8.47%, which is a comparable improvement to that obtained by batch processing.

Index Terms— Noise robust speech recognition, speech enhancement, beamforming, CHiME-3

1. INTRODUCTION

In this paper, we consider beamforming for automatic speech recognition (ASR) in noisy environments. Since background noise greatly degrades the ASR performance, high quality noise reduction is indispensable for noise robustness. A beamforming approach has been shown to improve the ASR performance in tasks ranging from medium vocabulary distant speech recognition [1] to large vocabulary meeting transcription [2, 3]. A beamformer is often parameterized by a steering vector for a target speaker direction, as with a delay-and-sum beamforming and minimum variance distortionless response (MVDR) beamforming.

While accurate steering vector estimation is the key to effective beamforming, conventional steering vector estimators often rely on possibly inaccurate knowledge, such as an array geometry or a plane wave assumption. For example, the baseline beamformer that was provided for the CHiME-3 challenge [4], a research community challenge program conducted in 2015, first estimates a target speaker direction with the steered response power-phase transform (SRP-PHAT) technique [5]. Then a steering vector is obtained by using the estimated direction of arrival (DOA) and a known microphone array geometry with the assumption of plane wave propagation. Although this beamformer works for simulated data in the CHiME-3 task, it does not improve the recognition performance for real data [4].

To overcome this limitation, we recently proposed a time-frequency mask-based approach to beamforming without any extra knowledge such as the array geometry or the plane wave assumption [6]. The central idea is to leverage the spectral sparsity of speech signals by using time-frequency masks estimated with a complex Gaussian mixture model (CGMM). The masks represent the probabilities of background noise dominating the corresponding time-frequency points. Then the steering vector can be estimated solely from the time-frequency masks and the observed data, which are used for constructing an MVDR beamformer. However, details of the method were not shown in [6] due to the limited space.

In this paper, we provide a detailed description of the CGMM-based beamforming method and undertake an extended investigation of this technique. We compare the CGMM-based beamformer with a conventional DOA-based beamformer and a beamformer that uses a Watson mixture model, which has often been used for time-frequency mask estimation [7]. Furthermore, we extend the CGMM-based approach to online speech enhancement, which enables this beamformer to be used for online recognition. Our online algorithm is derived with recursive updates of the CGMM parameters. Experimental results show that the online version of the CGMM-based beamformer runs in real time (even with a Matlab implementation) and reduces the word error rate (WER) from 15.60% to 8.47% for the CHiME-3 evaluation set, which already surpasses the 2nd best result of the challenge without speaker adaptation and system combination.

The rest of this paper is organized as follows. Section 2 explains the difference between our present method and previous studies. Section 3 provides an overview of our speech enhancement system, which comprises a time-frequency mask estimator, a steering vector estimator, and a beamformer. Section 4 describes the CGMM-based time-frequency mask estimation method used in our mask estimator. Section 5 extends this method to an online speech enhancement scenario. Section 6 shows ASR results obtained using the CHiME-3 corpus, which is followed by a conclusion in Section 7.

2. RELATED WORK

There have been several studies related to robust MVDR beamforming in the literature [8, 9, 10]. These studies aimed at making MVDR robust against steering vector estimation errors and sound reflections rather than improving steering vector estimation accuracy. Since MVDR attempts to null signals coming from any direction other than the look direction specified by the steering vector, the presence of a target speech signal component in the 'nuisance' directions, caused by the steering vector estimation errors and the sound reflections, would end up canceling out the target speech. While the methods proposed in [8, 9, 10] allow the beamformer to alleviate the signal cancellation problem, our work attempts to improve

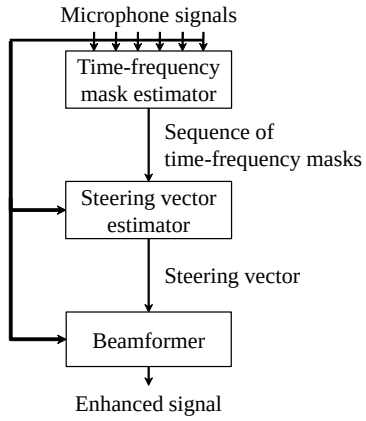

Fig. 1. Schematic diagram of our microphone array system architecture.

the accuracy of steering vector estimation. Although it is possible to combine these previous robust beamforming techniques with the proposed method, our experimental results show that the steering vectors estimated with the proposed method are accurate enough to prevent signal cancellation even with a conventional beamformer.

A mask-based beamforming approach was proposed in [7, 11], but there are two main differences between our CGMM-based method and these proposed methods. One difference is that we use a CGMM for the mask estimation while the previous methods employ a Watson mixture model. The Watson mixture model has fewer parameters than the CGMM, and so it tends to be affected by the fluctuation in the steering vector caused when speakers or recording devices move. The CGMM is parameterized by a full-rank spatial correlation matrix, and so we can deal flexibly with the spatial fluctuation of the steering vector. The superiority of the CGMM to the Watson mixture model is shown experimentally in Section 6. The other difference is that previous methods construct a beamformer without estimating the steering vector. For example, the method in [11] applied a beamformer parameterized by a spatial correlation matrix of a target signal, which is estimated with time-frequency masks. Our preliminary experiments showed that the beamforming proposed in [11] did not perform as well as the MVDR beamformer using the steering vector in the CHiME-3 task.

Regarding mask estimation, the CGMM was used in [12] to perform source separation in reverberant environments. However, the estimated spectral masks were used to perform spectral masking rather than beamforming, which we found to be harmful for ASR [6].

3. OVERVIEW OF OUR MICROPHONE ARRAY SYSTEM

Figure 1 shows a digram of our microphone array system architecture. The system inputs consist of noise-corrupted speech signals that are captured by the microphone array. The system comprises a beamformer, a steering vector estimator, and a time-frequency mask estimator. These three components combine to generate an enhanced speech signal with a beamforming approach.

3.1. Beamforming

The assumed architecture performs MVDR beamforming to enhance a speech signal in the short-time Fourier transform (STFT) domain. Let $y_{f,t,m}$ denote the m-th microphone signal at frequency f and time t. The signals from all M microphones can be represented using vector notation as

$$\mathbf{y}_{f,t} = [y_{f,t,1}, \ldots, y_{f,t,M}]^{\mathrm{T}}, \tag{1}$$

where superscript T denotes non-conjugate transposition. The beamformer applies a linear filter $\mathbf{w}_f$ to the microphone signal vector to produce an enhanced speech signal, $\hat{s}_{f,t}$, as

$$\hat{s}_{f,t} = \mathbf{w}_f^{\mathrm{H}} \mathbf{y}_{f,t}, \tag{2}$$

where superscript H denotes conjugate transposition. The filter $\mathbf{w}_f$ is determined in order to minimize the beamformer output power subject to $\mathbf{w}_f^{\mathrm{H}} \mathbf{r}_f = 1$, where $\mathbf{r}_f$ is the steering vector of the target signal. It should be noted that other types of beamformers such as the multichannel Wiener filter may be a useful alternative to the MVDR beamformer.

3.2. Steering vector estimation

The key to successful noise reduction lies in the accurate estimation of the steering vector. Conventional beamformers often obtain the steering vector by using DOA estimates and the plane wave propagation assumption, which holds only for an ideal anechoic space. Using the DOA estimates could also degrade noise reduction performance as their estimation accuracy deteriorates when SNRs are low.

Our approach does not use such errorful prior knowledge to obtain an accurate estimate of the steering vector. The basic idea is to directly estimate the steering vector using the covariance matrix of a microphone image of a target speech signal. Specifically, we utilize the principal eigenvector of an estimate of the covariance matrix as an estimate of the steering vector. The covariance matrix can be estimated by using time-frequency masks as described below, which allows us to take advantage of recent developments as regards clustering-based speech separation.

Let $\lambda_{f,t}^{(n)}$ denote the time-frequency mask that represents the probability of the time-frequency point (f, t) containing only noise. Then, we can estimate the covariance matrices of noisy speech and noise as

$$\mathcal{R}_f^{(x+n)} = \frac{1}{T} \sum_t \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}, \tag{3}$$

$$\mathcal{R}_f^{(n)} = \frac{1}{\sum_t \lambda_{f,t}^{(n)}} \sum_t \lambda_{f,t}^{(n)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}, \tag{4}$$

respectively. Then, the desired covariance matrix for the target speech signal is obtained by

$$\mathcal{R}_f^{(x)} = \mathcal{R}_f^{(x+n)} - \mathcal{R}_f^{(n)}. \tag{5}$$

An estimate of the steering vector can be obtained by first performing eigenvector decomposition on $\mathcal{R}_f^{(x)}$ and then extracting the eigenvector associated with the maximum eigenvalue.

4. TIME-FREQUENCY MASK ESTIMATION BASED ON COMPLEX GAUSSIAN MIXTURE MODEL

4.1. Observation model based on sparsity assumption in T-F domain

Considering the sparseness of speech in the time-frequency domain [12, 13, 14, 15, 16], we can assume that observed signals are clustered into two categories: one containing the noisy speech signal and

one containing only noise. With this assumption, the observed signal can be described as

$$\mathbf{y}_{f,t} = \mathbf{r}_f^{(\nu)} s_{f,t}^{(\nu)} \quad (\text{where } d_{f,t} = \nu), \tag{6}$$

where $d_{f,t}$ denotes the category index at the time-frequency point (f, t). ν may take $x + n$ or n, where the categories represent noisy speech and noise respectively. $s_{f,t}^{(x+n)}$ denotes a mixed signal of speech and noise at frequency f and time t, while $s_{f,t}^{(n)}$ denotes a noise signal at frequency f and time t. One way to associate the two categories with either noisy speech or noise, is to initialize the parameters of the two categories in different ways, and another way is to utilize some criterion after the clustering (as described in detail in Section 4.3).

4.2. Generative model of observed signal with complex Gaussian mixture model

Based on the above observation model, we design a generative model of the observation and define an objective function for soft mask estimation. First we assume $s_{f,t}^{(\nu)}$ locally follows a complex Gaussian distribution as

$$s_{f,t}^{(\nu)} \sim \mathcal{N}_c(0, \phi_{f,t}^{(\nu)}), \tag{7}$$

where $\phi_{f,t}^{(\nu)}$ corresponds to the variance of the signal at the time-frequency point, and $\mathcal{N}_c(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp - \frac{|x-\mu|^2}{\sigma^2}$. From Eqs. (6) and (7), the multichannel observed signal follows a complex Gaussian distribution

$$\mathbf{y}_{f,t} | d_{f,t} = \nu \sim \mathcal{N}_c(0, \phi_{f,t}^{(\nu)} \mathbf{R}_f^{(\nu)}) \tag{8}$$

conditioned on $d_{f,t}$, where $\mathbf{R}_f^{(\nu)} = \mathbf{r}_f^{(\nu)} \mathbf{r}_f^{(\nu)\mathrm{H}}$. This generative model of the observed signal $\mathbf{y}_{f,t}$ eventually becomes a complex Gaussian mixture model with the indicator $d_{f,t}$. We estimate the parameter $\mathbf{R}_f^{(\nu)}$ as a full-rank unconstrained matrix instead of directly estimating $\mathbf{r}_f^{(\nu)}$, which enables us to deal flexibly with fluctuations in the speaker and microphone positions [17].

4.3. Parameter estimation based on EM algorithm

The CGMM parameters, i.e., $\phi_{f,t}^{(\nu)}$ and $\mathbf{R}_f^{(\nu)}$, are estimated with a Maximum Likelihood (ML) approach. ML estimation can be performed with the Expectation-Maximization (EM) algorithm. The Q function to be maximized in each EM iteration is defined as

$$Q(\Theta) = \sum_{f,t} \sum_{\nu} \lambda_{f,t}^{(\nu)} \log \mathcal{N}_c(\mathbf{y}_{f,t}; 0, \phi_{f,t}^{(\nu)} \mathbf{R}_f^{(\nu)}), \tag{9}$$

where $\lambda_{f,t}^{(\nu)}$ represents the posterior probability of $d_{f,t}$ being ν. This posterior can be computed as

$$\lambda_{f,t}^{(\nu)} \leftarrow \frac{p(\mathbf{y}_{f,t} | d_{f,t} = \nu, \Theta)}{\sum_{\nu} p(\mathbf{y}_{f,t} | d_{f,t} = \nu, \Theta)}, \tag{10}$$

where $p(\mathbf{y}_{f,t} | d_{f,t} = \nu, \Theta) = \mathcal{N}_c(\mathbf{y}_{f,t}; 0, \phi_{f,t}^{(\nu)} \mathbf{R}_f^{(\nu)})$. The parameter values can be updated as follows:

$$\phi_{f,t}^{(\nu)} \leftarrow \frac{1}{M} \mathrm{tr}(\mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}} \mathbf{R}_f^{(\nu)^{-1}}), \tag{11}$$

$$\mathbf{R}_f^{(\nu)} \leftarrow \frac{1}{\sum_t \lambda_{f,t}^{(\nu)}} \sum_t \lambda_{f,t}^{(\nu)} \frac{1}{\phi_{f,t}^{(\nu)}} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}. \tag{12}$$

The time-frequency mask for point (f, t) can be obtained as the value of $\lambda_{f,t}^{(n)}$ after convergence.

After the convergence of the EM algorithm, to associate the two clusters with noise and noisy speech, we compute the entropy among the eigenvalues of the estimated spatial correlation matrix. The spatial correlation matrix with the bigger entropy can be regarded as that of the noise.

5. ONLINE SPEECH ENHANCEMENT WITH CGMM-BASED BEAMFORMING

In this section, we extend the proposed CGMM-based method to enable online speech enhancement. We assume that an observed signal is obtained as a sequence of mini-batches. Here, let $l \in \{1, \ldots, L\}$ be a mini-batch index, and let $\mathcal{B}_l$ denote a set of time frame indices within the l-th mini-batch. For the l-th mini-batch, $\lambda_{f,t}^{(\nu)}$ and $\phi_{f,t}^{(\nu)}$ are estimated with Eqs. (10) and (11), respectively, using the estimates of $\mathbf{R}_f^{(\nu)}$ obtained from the $(l-1)$-th mini-batch, $\mathbf{R}_{f,l-1}^{(\nu)}$. By modifying the update equation given by Eq. (12), the estimate of the spatial correlation matrix at the l-th mini-batch, $\mathbf{R}_{f,l}^{(\nu)}$, is recursively obtained by

$$\begin{aligned} \mathbf{R}_{f,l}^{(\nu)} \leftarrow & \frac{\Lambda_{f,l-1}^{(\nu)}}{\Lambda_{f,l-1}^{(\nu)} + \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)}} \mathbf{R}_{f,l-1}^{(\nu)} \\ & + \frac{1}{\Lambda_{f,l-1}^{(\nu)} + \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)}} \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)} \frac{1}{\phi_{f,t}^{(\nu)}} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}, \end{aligned} \tag{13}$$

where $\Lambda_{f,l}^{(\nu)}$ is the sum of $\lambda_{f,t}^{(\nu)}$ over all the observed time frames, which is also recursively updated by

$$\Lambda_{f,l}^{(\nu)} \leftarrow \Lambda_{f,l-1}^{(\nu)} + \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)}. \tag{14}$$

Using the sequentially estimated soft mask, $\lambda_{f,t}^{(\nu)}$, beamforming can be performed online as follows. First, the covariance matrices for noisy speech and noise are recursively updated by

$$\begin{aligned} \mathcal{R}_{f,l}^{(\nu)} \leftarrow & \frac{\Lambda_{f,l-1}^{(\nu)}}{\Lambda_{f,l-1}^{(\nu)} + \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)}} \mathcal{R}_{f,l-1}^{(\nu)} \\ & + \frac{1}{\Lambda_{f,l-1}^{(\nu)} + \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)}} \sum_{t \in \mathcal{B}_l} \lambda_{f,t}^{(\nu)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}. \end{aligned} \tag{15}$$

The steering vector for the l-th mini-batch is estimated by using the procedure described in Section 3.2. After that, the enhanced signal for $\mathcal{B}_l$ can be obtained with MVDR beamforming. It should be noted that MVDR beamforming can be also performed without updating the CGMM parameter $\mathbf{R}_{f,l}^{(\nu)}$ if we can obtain a reliable initial value.

6. EXPERIMENTAL EVALUATION

We conducted ASR experiments using the CHiME-3 corpus to evaluate the noise reduction performance of the CGMM-based beamformer. The corpus consists of read speech recorded in four different environments with six microphones attached to a tablet device and additional simulated audio data. The sentences were taken from the WSJ0 corpus. The training set comprises 1600 real and 7138 simulated utterances. The training data amounts to about 108 hours when using the audio data from all six microphones for training. The development and evaluation sets consist of 3280 and 2640 utterances, respectively, each containing both simulated data (simu data) and

recorded data (real data). Further details of the corpus can be found in [4].

In our experiments, we performed speaker independent decoding by using a deep convolutional neural network (CNN) acoustic model [18, 19] and a class-based recurrent neural network language model (RNN-LM) [20, 21]. Inputs to the acoustic model comprised 40-dimensional log mel-filter bank channel outputs and their delta and double-delta coefficients. Our CNN was based on the network-in-network concept [22] and consisted of five convolution layers and two max-pooling layers, where all the layers contained 180 feature maps. The last convolution layer was followed by three fully connected layers with 2048 units and a softmax layer. The softmax layer contained 5976 units, i.e., context-dependent HMM states. Our RNN-LM used 10 classes and accommodated 500 units in the hidden recurrent layer. See our CHiME-3 paper [6] for a detailed description of the recognizer.

We investigate the effectiveness of the CGMM-based beamformer with batch- and online-processing setups in Sections 6.1 and 6.2, respectively.

6.1. Batch processing experiments

For the batch-processing setup, we performed beamforming with the configurations shown in Table 1. The initial value of $\mathbf{R}_f^{(x+n)}$ was set at the covariance matrix of an observed signal vector. $\mathbf{R}_f^{(n)}$ was initialized by using an identity matrix. We used two conventional

Table 1. Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame overlap	75%
Window function	Hanning
Number of EM iterations	20
Number of microphones	6

beamformers for comparison: one was the CHiME-3 baseline beamformer, which estimates steering vectors based on an array geometry and a plane wave assumption (see [4] for details); the other was a beamformer that was obtained by replacing a CGMM with a Watson mixture model in the mask-based beamforming scheme.

Table 2 compares the proposed CGMM-based beamformer with its two competitors in terms of WERs. We can see that the CGMM-based method achieved the lowest WERs for both the development and evaluation sets.

Table 2. WERs obtained with the proposed method and its competitors. Following a CHiME-3 challenge regulation, we focused on the results for real data.

systems	dev			eval		
	avg	simu	real	avg	simu	real
not used	8.62	8.24	9.01	12.89	10.17	15.60
conventional	7.10	4.79	9.41	10.79	5.37	16.21
Watson MM	5.71	6.33	5.09	10.60	11.72	9.47
Proposed	4.96	5.09	**4.83**	8.46	8.06	**8.86**

To investigate the impact that the number of microphones has on the noise reduction performance, we performed experiments where we varied the number of microphones from two to five. Table 3 shows the WERs we obtained with different numbers of microphones. Although the use of fewer microphones increased the WER, our proposed beamformer always yielded performance gains for the real data. This means that the CGMM-based beamformer is applicable to the most practically relevant multi-microphone setup that uses two microphones.

Table 3. WERs obtained with different numbers of microphones.

Number of microphones	dev			eval		
	avg	simu	real	avg	simu	real
2	8.82	9.96	7.69	12.05	10.44	13.66
3	6.79	7.16	6.42	9.74	8.09	11.39
4	5.79	6.13	5.45	9.21	7.90	10.51
5	5.43	5.50	5.36	8.67	6.96	10.37

6.2. Online processing experiments

We evaluated the online speech enhancement algorithm described in Section 5. We set the size of the first mini-batch at 500 ms and that of succeeding mini-batches at 250 ms to ensure that the first mini-batch contained the target speech signal. We initialized the spatial correlation matrices by using separate speech and noise signals contained in the CHiME-3 corpus. Specifically, $\mathbf{R}_{f,0}^{(x+n)}$ was obtained from speech signals recorded in a booth while $\mathbf{R}_{f,0}^{(n)}$ was obtained from separate noise signals. This initialization also allowed us to avoid the permutation ambiguity described at the end of Section 4.3 and thus reduce the computational cost. With this setup, the average real-time factor was 0.86 with our Matlab implementation on a 2.6 GHz PC. Therefore we can obtain an enhanced speech signal with a 500 ms delay. Other hyperparameters were set in the same way as for the experiments in Section 6.1. For the online processing, we considered two cases where the CGMM parameters, namely the spatial correlation matrices, were updated or not updated.

Table 4 shows the WERs obtained by batch processing and online processing with/without a CGMM parameter update. Note that we performed batch and online beamforming with the same initial conditions to obtain fair comparisons. Even without the parameter update, our online beamformer yielded performance gains while having an advantage in terms of computational cost. With the parameter update, the performance gains increased and were comparable to those obtained by batch processing. Our online speech enhancement reduced the WERs from 15.60% to 8.47% compared with those obtained without processing shown in Table 2.

Table 4. WERs obtained with online processing.

systems	dev			eval		
	avg	simu	real	avg	simu	real
batch	5.09	5.19	5.00	8.14	7.90	8.37
online w/o updates	6.09	6.66	5.52	10.67	9.74	11.59
online w/ updates	5.27	5.54	5.00	8.20	7.92	8.47

7. CONCLUSION

We described a beamfomer that uses a novel steering vector estimation method based on time-frequency masks. The use of the time-frequency masks allowed us to avoid using inaccurate prior knowledge such as an array geometry and a plane wave propagation assumption and thus provided robust steering vector estimates. The time-frequency masks were estimated by using a spectral model based on a CGMM, which was shown to outperform a recently proposed Watson mixture model. In addition, we extended the CGMM-based beamfoming approach to online speech enhancement. Our experimental results showed that the online processing method reduced the WER from 15.60% to 8.47% in the CHiME-3 task, which is a comparable improvement to that obtained by batch processing.

8. REFERENCES

[1] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, 2015, Article ID 2015:60, doi:10.1186/s13634-015-0245-7.

[2] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant multichannel large vocabulary speech recognition," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2013, pp. 285–290.

[3] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5527–5531.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, accepted.

[5] J. H. DiBiase, H. R. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 157–180. Springer, 2001.

[6] 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. Worksh. Automat. Speech Recognition, Understanding*, 2015, accepted.

[7] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 241–244.

[8] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2677–2684, 1999.

[9] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 994–1008, 2009.

[10] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, 2007.

[11] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[12] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2014.

[13] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[14] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Adv. Neural Inform. Process. Syst.*, vol. 13, pp. 953–960, 2007.

[15] 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. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 33–36.

[16] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

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

[18] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[19] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadrana, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[20] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010, pp. 1045–1048.

[21] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocky, "Empirical evaluation and combination of advanced language modeling techniques," in *Proc. Interspeech*, 2011, pp. 605–608.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint*, 2014, arXiv:1312.4400v3.