INTEGRATING DNN-BASED AND SPATIAL CLUSTERING-BASED MASK ESTIMATION FOR ROBUST MVDR BEAMFORMING

Tomohiro Nakatani Nobutaka Ito Takuya Higuchi Shoko Araki Keisuke Kinoshita

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237 Japan

ABSTRACT

Recently, time-frequency mask-based beamforming has been extensively studied as the frontend of deep neural network (DNN) based automatic speech recognition (ASR) in noisy environments. Two mask estimation approaches have been separately developed for this beamforming method, namely the the DNN-based approach, which exploits the time-frequency features of the signal, and the spatial clustering-based approach, which exploits the spatial features of the signal. This paper proposes a new method that integrates the two approaches in a probabilistic way to further improve mask estimation by exploiting the advantages of both approaches. Experiments using the real data of the CHiME-3 multichannel noisy speech corpus show that the proposed method almost always outperforms the conventional approaches in terms of word error rate (WER) improvement.

Index Terms: Beamforming, automatic speech recognition, time-frequency mask, deep neural network, spatial clustering

1. INTRODUCTION

When we capture our speech using distant microphones in our daily lives, various types of ambient noise are mixed with the captured signals, and severely degrade the ASR performance. To solve this problem, beamforming is being extensively studied as the noise reduction frontend for ASR. Delay-and-sum beamforming, minimum variance distortionless response (MVDR) beamforming, and maximum signal-to-noise ratio (MaxSNR) beamforming are often employed [1, 2, 3], and have been shown to improve the ASR performance in tasks ranging from medium vocabulary distant speech recognition [4] to large vocabulary meeting transcription [5, 6].

An accurate estimation of captured signals' spatial acoustic characteristics, such as the spatial covariance matrices of the speech and the noise, is crucial if we are to make beamforming work effectively. For this purpose, researchers have recently proposed time-frequency mask-based beamforming approaches [7, 8, 9, 10, 11]. The central idea is to leverage the spectral sparsity of speech signals by using time-frequency masks that represent the probability of speech (or noise) dominating the corresponding time-frequency points [12, 13, 14]. Then, the spatial covariance matrices of the speech and the noise are estimated solely from the time-frequency masks and the captured signal, and used for constructing beamformers. A feature of this approach is that it can estimate the beamformers accurately without relying on any assumptions regarding the microphone array geometry or the acoustic conditions of the room (e.g., a plane wave condition), from short observation of the order of a few seconds. This is particularly advantageous for many ASR scenarios using distant microphones.

The two main techniques proposed for mask estimation are the DNN-based approach and the spatial clustering-based approach. With the DNN-based approach [10, 11, 15], a DNN is trained in advance on training data so that it can estimate masks from the time-frequency features of a single channel noisy speech signal. Then, the test data from multiple microphones are used to estimate multiple sets of masks, which are then merged into a single set. With the spatial clustering based-approach [7, 9, 16, 17, 18], on the other hand, the masks are estimated from the test data in an unsupervised learning manner. Assuming that the spatial features of speech and noise have different distributions, this approach finds these two distributions based on the clustering of the spatial features, and the masks are estimated as the posteriors of each cluster at the corresponding time-frequency points.

Although both approaches have been shown effective, they have different advantages and disadvantages. In particular, the DNN-based approach inevitably degrades when there is a mismatch between the training and test conditions, which is often the case in real acoustic environments. In contrast, the spatial clustering-based approach is insensitive to such a mismatch thanks to its unsupervised learning scheme. So, to make the DNN-based approach more robust against mismatches, this paper proposes a new mask estimation method that integrates, in a probabilistic way, DNN-based and spatial clustering based mask estimation. With the proposed method, initial masks are estimated based solely on the DNN-based approach. Then, the masks are adapted to the test conditions via the clustering of the spatial features based on the expectation-maximization (EM) algorithm, where the initial masks are utilized as the time-frequency dependent mixture weight of each cluster. The resultant posteriors of each cluster are determined as the integrated masks. Finally, an MVDR beamformer is formed based on the estimated masks and applied to the noisy speech. Experiments using the real data of the CHiME-3 multichannel noisy speech corpus [20] show that the proposed method almost always outperforms conventional DNN-based and spatial clustering-based approaches.

In the remainder of this paper, after reviewing existing mask-based MVDR beamforming in Section 2, the proposed method is presented in Section 3. Section 4 summarizes related work. Sections 5 and 6, respectively, provide experimental results and concluding remarks.

2. REVIEW OF MASK-BASED MVDR BEAMFORMING

Figure 1 shows the processing flow of the mask-based MVDR beamforming method [9] used in this paper. It receives a set of noisy speech signals captured by multiple microphones and generates a single enhanced speech signal. The method employs a time-frequency mask estimator, a steering vector estimator, and an

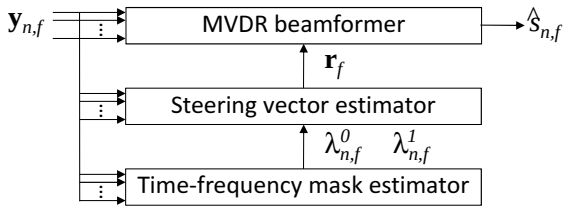

Fig. 1. Processing flow of mask-based MVDR beamforming.

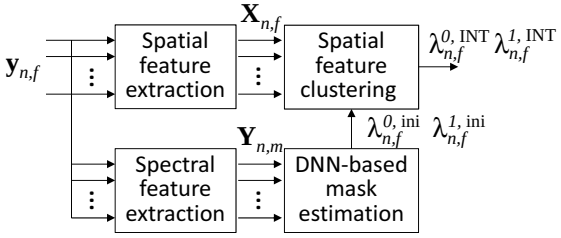

Fig. 2. Processing flow of proposed time-frequency mask estimator.

MVDR beamformer.

2.1. MVDR Beamformer

The method performs MVDR beamforming in the short-time Fourier transform (STFT) domain. Let $y_{n,f,m}$ be a signal captured by the m-th microphone ($M \geq m \geq 1$) at time n ($N \geq n \geq 1$) and frequency f ($F \geq f \geq 1$). The signals from all the microphones are represented using vector notation as

$$\mathbf{y}_{n,f} = [y_{n,f,1}, \cdots, y_{n,f,M}]^T$$

where superscript T denotes non-conjugate transposition.

The beamformer applies a linear filter $\mathbf{w}_f$ to the captured signal $\mathbf{y}_{n,f}$ to produce an enhanced speech signal, $\hat{s}_{n,f}$, as

$$\hat{s}_{n,f} = \mathbf{w}_f^H \mathbf{y}_{n,f}$$

where superscript H denotes conjugate transposition. The filter $\mathbf{w}_f$ is an MVDR beamformer [1] when it is determined so that it minimizes the power of the beamformer output subject to $\mathbf{w}_f^H \mathbf{r}_f = 1$, where $\mathbf{r}_f$ is the estimated steering vector of the speech signal, or the look direction of the beamformer. It should be noted that other types of beamformers such as the MaxSNR beamformer [1, 10, 21, 22] are useful alternatives to mask-based beamforming.

2.2. Mask-based steering vector estimation

The key to successful noise reduction with MVDR beamforming is the accurate estimation of the steering vector. Conventional MVDR beamformers often obtain the steering vector assuming that the spatial features of the captured signal obey the plane wave propagation condition. However, the condition holds only in an ideal anechoic far-field space, and thus the accuracy of the steering vector estimation deteriorates severely in real acoustic environments.

In contrast, the mask-based approach estimates the steering vector in a data-driven manner and does not rely on such an erroneous assumption. Let $\lambda^0_{n,f}$ be an estimated mask for noise, which represents the probability of the corresponding time-frequency point being dominated by the noise, we can estimate the respective spatial covariance matrices of noisy speech and noise as

$$\mathcal{R}_f^{(s+v)} = \frac{1}{N} \sum_n \mathbf{y}_{n,f} \mathbf{y}_{n,f}^H \quad (1)$$

$$\mathcal{R}_f^{(v)} = \frac{1}{\sum_n \lambda^0_{n,f}} \sum_n \lambda^0_{n,f} \mathbf{y}_{n,f} \mathbf{y}_{n,f}^H \quad (2)$$

Then, the spatial covariance matrix of the speech is obtained by

$$\mathcal{R}_f^{(s)} = \mathcal{R}_f^{(s+v)} - \mathcal{R}_f^{(v)} \quad (3)$$

Finally, the steering vector is estimated as the principal eigenvector of the estimated spatial covariance matrix, $\mathcal{R}_f^{(s)}$.

With the above method, the accuracy of the beamforming clearly depends largely on that of the mask estimation.

3. PROPOSED MASK ESTIMATION METHOD

Figure 1 shows the processing flow of our proposed mask estimation method. First, it extracts two different features from the input signal, i.e., an F-dimensional spectral feature $\mathbf{Y}_{n,m}$ at each time n and microphone m, and an M-dimensional spatial feature $\mathbf{X}_{n,f}$ at each time-frequency point (n, f). They are respectively defined as

$$\mathbf{Y}_{n,m} = [Y_{n,1,m}, \cdots, Y_{n,F,m}]^T \quad (4)$$

$$Y_{n,f,m} = \log |y_{n,f,m}| \quad (5)$$

$$\mathbf{X}_{n,f} = \frac{\mathbf{y}_{n,f}}{||\mathbf{y}_{n,f}||} \quad (6)$$

where $||\cdot||$ denotes the Euclidean norm. Next, in the processing flow, the DNN-based mask estimation receives the time sequence of the spectral features for all microphones, and estimates the initial masks at all time-frequency points (n, f). Then, the spatial feature clustering clusters the spatial features into speech and noise clusters taking the initial masks into account, and finally estimates the integrated masks as the posteriors of the clusters.

The proposed method employs the likelihood function below as the optimization criterion for the integration.

$$\mathcal{L}(\theta^{\text{SC}}) = p(\mathcal{X}, \mathcal{Y}; \theta^{\text{SC}}, \theta^{\text{DNN}}) \quad (7)$$

where $\mathcal{X}$ and $\mathcal{Y}$, respectively, are sets of features containing all the spatial and spectral features that are available for the integration, and θ^{DNN} and θ^{SC}, respectively, are the model parameter sets of the DNN and the model for spatial clustering. In this paper, we assume that θ^{DNN} is learned in advance using training data, and θ^{SC} is estimated using the test data. By maximizing the above likelihood, θ^{SC} is optimized taking both spatial and spectral features into consideration.

Let $d_{n,f}$ be a binary random variable that represents the dominant source index at a time-frequency point (n, f), which takes 1 when speech dominates the point and takes 0 otherwise. Then, with proper assumptions on conditional independence over time-frequency points that are commonly used in the conventional approaches [23], and disregarding constant terms, the likelihood function can be rewritten as

$$\mathcal{L}(\theta^{\text{SC}}) = \prod_f \mathcal{L}_f(\theta^{\text{SC}}) \quad (8)$$

$$\mathcal{L}_f(\theta^{\text{SC}}) = \prod_n \sum_{d=0}^{1} \lambda_{n,f}^{d,\text{ini}} p(\mathbf{X}_{n,f} | d_{n,f} = d; \theta^{\text{SC}}) \quad (9)$$

$$\lambda_{n,f}^{d,\text{ini}} = p(d_{n,f} = d | \mathcal{Y}; \theta^{\text{DNN}}) \quad (10)$$

where eq. (8) decomposes the likelihood function into frequency-wise functions, $\mathcal{L}_f(\theta^{\text{SC}})$. In eqs. (9) and (10), the values $\lambda_{n,f}^{d,\text{ini}}$ for $d = 0$ and 1 correspond to the conditional probability of the time-frequency point being dominated, respectively, by noise and speech given the spectral features, and are referred to as the initial masks for noise and speech. Note that $\lambda_{n,f}^{0,\text{ini}} + \lambda_{n,f}^{1,\text{ini}} = 1$. This paper assumes that these values are estimated by the DNN in the proposed method. $p(\mathbf{X}_{\mathbf{n,f}}|d_{n,f} = d; \theta^{\text{SC}})$, on the other hand, is the conditional probability density function (pdf) of the spatial features given the dominant source index, and it corresponds to the pdf of the spatial features of speech for $d = 1$ and that of noise for $d = 0$ according to the sparsity assumption.

When we look closely at eq. (9), we see that it is in the form of a mixture distribution model of the spatial features at frequency f, where d is a hidden variable, $p(\mathbf{X}_{n,f}|d_{n,f} = d; \theta^{\text{SC}})$ is a component distribution, and $\lambda_{n,f}^{d,\text{ini}}$ is the time-frequency dependent mixture weight. With this interpretation and employing appropriate component distributions, the likelihood function can be efficiently maximized to a stationary point using the EM algorithm as shown in Section 3.3. Finally, with the optimized model parameter set θ^{SC}, the integrated speech masks $\lambda_{n,f}^{d,\text{INT}}$ for $d = 0$ and 1 are obtained as the posteriors of the dominant source index.

In the following subsections, we describe the specification of each processing block in more detail.

3.1. DNN-based mask estimation

Recently, mask estimation based on DNNs has been extensively studied. Most techniques are developed so that DNNs receive the spectral features and output the estimate of the masks, and therefore can be used for the initial mask estimation of the proposed method. This paper employs the bidirectional long short-term memory (BLSTM) based neural network proposed in [10].

The BLSTM network used in [10] is composed of a BLSTM layer followed by three feed-forward neural network layers. To train the network, the input is the spectral features defined in eq. (4), and the desired output is a concatenation of two types of ideal binary masks, one for speech and the other for noise. The ideal binary masks for speech (or those for noise) take 1 when the time-frequency points are dominated by the speech (or noise) and take 0 otherwise. Note that, to obtain the ideal binary masks, only simulated data can be used as the training data, from which we can extract the microphone images of speech and noise separately. Then, the output of the trained BLSTM network can be interpreted as the conditional probability of the speech and the noise dominating the time-frequency points given the spectral features, namely $p(d_{n,f} = 1|\mathcal{Y}; \theta^{\text{DNN}})$ and $p(d_{n,f} = 0|\mathcal{Y}; \theta^{\text{DNN}})$. We use them as the respective conditional probabilities in the proposed method although they do not necessarily sum to 1.

3.2. Spatial clustering based-mask estimation

With the spatial features in eq. (6), several useful mixture distribution models, including a complex Watson mixture model [7, 17] and a complex angular central Gaussian mixture model (cACGMM) [18], have been proposed for spatial clustering-based mask estimation. This paper employs cACGMM because it can yield the same mask estimates as a complex Gaussian mixture model (cGMM), which have been shown in [9] to be very useful for mask-based beamforming. Note that cGMM is a model for multichannel complex Fourier spectra, and thus the conditional independence introduced in eqs. (8) and (9) is not appropriate when we use cGMM.

A component distribution of cACGMM is modeled by a complex angular central Gaussian (cACG) distribution [19], which is defined on a hypersphere as

$$\mathcal{A}(\mathbf{z}; B) = \frac{(M-1)!}{2\pi^M \det B} \frac{1}{(\mathbf{z}^H B \mathbf{z})^M} \tag{11}$$

where $\mathbf{z}$ is an M-dimensional complex random variable vector on the hypersphere, which corresponds to a spatial feature in the proposed method, and B is an $M \times M$ positive definite Hermitian matrix, which is a model parameter of the distribution. Unlike a complex Watson distribution, which can only model the mode and concentration of the distribution, the cACG distribution can also model the shape and rotation of the distribution by B, thus it can better model the distribution of the spatial features. With cACGMM, two component distributions with different model parameters at each frequency f, namely B_f^0 and B_f^1, are introduced that correspond to noise and speech, respectively.

The model parameters of the cACGMM are estimated from the captured signal in an unsupervised manner based on the EM algorithm, and the masks are estimated based on the optimized parameters. See the concrete estimation steps in [18]. In the following, this paper describes the estimation steps when they are combined with the DNN based initial mask estimation.

3.3. Processing steps of integrated mask estimation

With the proposed method, after estimating the initial masks with the DNN, the model parameters of the component distributions in eq. (8), or B_f^0 and B_f^1 of cACG distributions, can be estimated based on the EM algorithm using the initial masks as the time-frequency dependent mixture weights of the cACGMM. The overall processing steps including the integrated mask estimation can be summarized as follows.

1. Extract spectral and spatial features, $\mathbf{Y}_{n,m}$ and $\mathbf{X}_{n,f}$, as in eqs. (4) and (6).
2. Calculate the initial masks, $\lambda_{n,f}^{d,\text{ini}}$, for $d = 0$ and 1 using the BLSTM network trained in advance on training data.
3. Set the initial values of the integrated masks, $\lambda_{n,f}^{d,\text{INT}}$, for $d = 0$ and 1 as
$$\lambda_{n,f}^{d,\text{INT}} = \lambda_{n,f}^{d,\text{ini}} \tag{12}$$
4. Iterate the following until convergence is obtained.
   (a) Update B_f^d for $d = 0$ and 1 as (M-step)
$$B_f^d = M \frac{\sum_n \lambda_{n,f}^{d,\text{INT}} \frac{\mathbf{X}_{n,f}\mathbf{X}_{n,f}^H}{\mathbf{X}_{n,f}^H (B_f^d)^{-1} \mathbf{X}_{n,f}}}{\sum_n \lambda_{n,f}^{d,\text{INT}}} \tag{13}$$
   (b) Update $\lambda_{n,f}^{d,\text{INT}}$ for $d = 0$ and 1 as (E-step)
$$\lambda_{n,f}^{d,\text{INT}} = \frac{\lambda_{n,f}^{d,\text{ini}} \mathcal{A}(\mathbf{X}_{n,f}; B_f^d)}{\sum_{d'} \lambda_{n,f}^{d',\text{ini}} \mathcal{A}(\mathbf{X}_{n,f}; B_f^{d'})} \tag{14}$$

4. RELATED WORK

DNN-based mask estimation approaches have been proposed that use both time-frequency features and spatial features as the input of the DNN [24, 25]. As the spatial features, [24] employs level difference, time difference, and cross-correlation over different channels, while [25] employs masks estimated using spatial clustering. The

Table 1. WERs (%) obtained using different mask estimators (Mask) and different beamformers (BF) for the CHiME-3 development set (dt_05) and evaluation set (et_05). Bold fonts indicate the best score for each condition.

			simu					real					
Mask	BF	Set	Ave	BUS	CAF	PED	STR	Ave	BUS	CAF	PED	STR	Ave (all)
BLSTM	GEV	dt_05	5.23	4.66	7.09	4.53	4.65	4.86	5.68	4.50	4.60	4.66	5.05
BLSTM	MVDR	dt_05	**4.47**	**4.04**	**5.62**	**3.86**	**4.38**	4.50	**5.53**	4.28	4.03	4.17	4.49
cACGMM	MVDR	dt_05	4.99	4.99	6.03	4.56	4.40	4.54	5.92	**3.75**	4.04	4.47	4.77
Proposed	MVDR	dt_05	4.53	4.20	5.72	3.95	4.26	**4.40**	5.56	3.94	**4.01**	**4.08**	**4.46**
BLSTM	GEV	et_05	6.57	5.53	6.89	7.02	6.85	7.28	**8.24**	7.28	6.61	7.00	6.93
BLSTM	MVDR	et_05	**5.27**	4.52	**5.19**	**5.66**	**5.70**	7.15	9.05	6.95	6.11	6.48	6.21
cACGMM	MVDR	et_05	6.94	4.99	6.50	7.28	8.97	7.28	9.93	6.63	**5.74**	6.84	7.11
Proposed	MVDR	et_05	5.37	**4.31**	5.29	5.70	6.16	**6.71**	8.38	**6.61**	5.81	**6.03**	**6.04**

masks estimated with these methods are used for nonlinear noise reduction, but have not been tested in combination with mask-based beamforming to the best of our knowledge.

Another approach has been proposed that integrates the spatial and spectral features for beamforming, where DNNs are used to model the spectral features and combined with the multichannel Gaussian model to exploit the spatial information [26]. The model parameters are estimated by iterative optimization and used to derive a time-varying multichannel Wiener filter. The approach is developed for improving the SNR of the signal, but has not yet been well evaluated in terms of ASR improvement.

5. EXPERIMENTS

The proposed approach is evaluated in terms of the ASR performance achieved on the CHiME-3 Speech Separation and Recognition Challenge corpus [20]. The CHiME-3 corpus was created by using a 6-channel microphone array attached to a tablet device. The recordings were obtained in four different noisy environments, i.e., public transport (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR), and they feature several male and female speakers, uttering Wall Street Journal (WSJ) [27] sentences. The corpus is divided into 3 individual subsets, namely a training set, containing 8738 noisy utterances, a development set (dt_05), containing 3280 noisy utterances, and an evaluation set (et_05), containing 2640 noisy utterances. Each subset consists of real and simulated data, abbreviated as real and simu in the following, where the latter has been generated by artificially mixing the clean WSJ utterances with the environmental noise recordings.

5.1. Experimental Setup

We compared three mask estimation methods, namely a DNN-based method using BLSTM (BLSTM), a spatial clustering-based method using cACGMM (cACGMM), and a method integrating them both (Proposed). The same MVDR beamformer (see section 3) [9] and the same ASR backend (see the next paragraph) were used for comparison. For BLSTM, we used a software tool used in [10] and provided at [28]. BLSTM was trained on simu in the training data and used for testing. In this paper, we also compared two beamformers in combination with BLSTM, i.e., the MVDR beamformer and the generalized eigenvalue decomposition (GEV) beamformer used in [10, 21]. For cACGMM and the proposed method, the model parameters of the cACGMM were estimated from each utterance in the test data and used for the mask estimation, except that the time-frequency dependent mixture weights were estimated by BLSTM for the proposed method.

The speech recognizer that we used for the evaluation was based on a multi-condition convolutional neural network (CNN) acoustic model and a recurrent neural network (RNN) language model [29] in addition to a trigram language model. A detailed description of the system is the same as that of the 1-pass SI system in [30].

5.2. Results

Table 1 summarizes the WERs obtained in the experiments. While BLSTM with MVDR outperformed the others under all simu conditions except for BUS in et_05, the proposed method outperformed the others with MVDR under all real conditions except for BUS in dt_05 and PED in et_05. Note that BLSTM is trained on simu in the training set, and the mismatch between the training and the simulated test conditions is very small. Thus, the results obtained with simu suggest that BLSTM is very effective when such mismatch is very small. In contrast, the results with real suggest that BLSTM becomes less effective as the mismatch increases, which is often the case in real acoustic environments. And the integration with cACGMM can offer good mitigation for the influence of the mismatch and achieve the best performance. It should also be noted that the MVDR beamformer used in this paper outperformed the GEV beamformer in combination with BLSTM under all conditions except for real BUS in et_05.

6. CONCLUDING REMARKS

This paper proposed a new time-frequency mask estimation method that can improve the performance of ASR in noisy environments when it is used as the ASR frontend in combination with mask-based beamforming. The proposed method is constructed by integrating BLSTM-based mask estimation with cACGMM based spatial clustering. By exploiting the discrimination capability of BLSTM and the unsupervised learning scheme of the spatial clustering, the proposed method achieves mask estimation that is highly accurate and adaptive to the test conditions. Experiments using the CHiME-3 multichannel noisy speech corpus showed the effectiveness of the proposed method for real data, which will include a certain mismatch between the training and test conditions, compared with the conventional approaches in terms of word error rate (WER) improvement.

7. REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, 5 (2), pp. 4–24, April, 1988.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. ASLP*, 15 (7), pp. 2011–2022, 2007.

[3] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. ASLP*, 18 (2), pp. 260–276, 2007.

[4] 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.*, Article ID 2015:60, doi:10.1186/s13634-015-0245-7, 2015.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant multichannel large vocabulary speech recognition," in *Proc. IEEE ASRU-2013*, pp. 285–290, 2013.

[6] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proc. IEEE ICASSP-2014*, pp. 5527–5531, 2014.

[7] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE ICASSP-2010*, 2010.

[8] 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. ASLP*, 21 (9), pp. 1913–1928, 2013.

[9] T. Higuchi, N. Ito, T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE ICASSP-2016*, pp. 5210–5214, 2016.

[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP-2016*, pp. 196–200, 2016.

[11] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," *Proc. Interspeech-2016*, 2016.

[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE trans. Signal Processing*, 52 (7), pp. 1830–1847, 2004.

[13] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197, Springer, 2005.

[14] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, doi:10.1016/j.sigpro.2007.02.003, 2007

[15] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. ASLP*, 21 (7), pp. 1381–1390, July 2013.

[16] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound: Sources in reverberant environments," in *Proc. the 2006 Conference on Advances in Neural Information Processing Systems*, MIT Press, pp. 953–960, 2007.

[17] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, 19 (3), pp. 516–527, 2011.

[18] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. EUSIPCO-2016*, 2016.

[19] J. T. Kent, "Data analysis for shapes and images," *Journal of Statistical Planning and Inference*, 57 (2), pp. 181–193, Feb. 1997.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: dataset, task and baselines," in *Proc. IEEE ASRU-2015*, 2015.

[21] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. ASLP*, 15 (5), pp. 1529–1539, 2007.

[22] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in Proc. IEEE ICASSP-2007, pp. 41–44, 2007.

[23] 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*, 21 (12), pp. 2516–2531, 2013.

[24] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE Trans. ASLP*, 22 (12), pp. 2112–2121, 2014.

[25] H. Meutzner, S. Araki, M. Fujimoto, and T. Nakatani, "A generative-discriminative hybrid approach to multi-channel noise reduction for robust automatic speech recognition," in *Proc. IEEE ICASSP-2016*, pp. 5740–5744, 2016.

[26] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, 24 (10), 2016.

[27] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *HLT-91 Proc. the Workshop on Speech and Natural Language*, pp. 357–362, 1992.

[28] https://github.com/fgnt/nn-gev

[29] 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, 2011.

[30] 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. IEEE ASRU-2015*, December 2015.