Comparative Study on DNN-based Minimum Variance Beamforming Robust to Small Movements of Sound Sources

Kohei Saijo*, Kazuhiro Katagiri†, Masaru Fujieda†, Tetsunori Kobayashi* and Tetsuji Ogawa*
* Waseda University, Communications and Computer Engineering Dept. , Tokyo, JAPAN
† OKI Electric Industry Corporation, Tokyo, JAPAN

***Abstract*—This paper discusses a deep neural network (DNN)-based minimum variance (MV) beamformer suitable for the case where the target sound source moves slightly in front of the microphones. In practical applications of speech enhancement, such as a guidance terminal installed in a train station, the target sound source can be assumed to be located approximately in front of the microphones, although it may move slightly. Speech enhancement techniques used under such conditions can be classified into two types: one is to enhance the sound source while adaptively estimating its location, and the other is to enhance the area in front of the microphone array. The former requires localization of the target source but has a high degree of freedom of the beamformer, which can lead to high noise suppression performance, while the latter does not require the source localization but has a low degree of freedom of the beamformer. Speech enhancement experiments conducted to compare the performance of these approaches demonstrated that the MV beamformer based on adaptive sound source localization can provide more accurate enhancement than that based on area enhancement even when the sound source is moving.**

I. Introduction

Speech enhancement technologies play an important role in situations where speech interfaces are used in noisy environments. Since the position of the speaker's mouth (i.e., the target source location) changes from time to time with facial and body movements, it is necessary to provide a speech enhancement system that is robust to movements of the target source. On the other hand, there are many applications where the target source can be assumed to be located in front of the microphones [1], such as guidance terminals in train stations and online conference systems. The speech enhancement technologies used in such applications can be roughly classified into two approaches: one is the localization and enhancement, which aims to enhance the moving target source while tracking it, and the other is area enhancement, which aims to direct a wide beam in front of the microphones for enhancing the area including the target source. It should be noted that there is a trade-off between these two approaches in terms of the robustness to sound source deviation and the noise suppression accuracy. Specifically, the former can achieve accurate speech enhancement if source localization is accurate, but the localization error may adversely affect the speech enhancement performance. The latter does not require exact source localization, but the accuracy of speech enhancement is limited because it enhances not a sound source but an area including the source.

In recent years, deep neural networks (DNNs) have attracted much attention for their high capability in source modeling [2]–[4]. Many attempts have been made for integrating beamformers [5] with DNNs. In particular, minimum variance distortionless response (MVDR) beamforming [6], which aims to minimize noise under the constraint that makes speech arriving from the direction of the target source distortionless, has been successfully integrated with DNN-based time-frequency (TF) masking [7]–[11]. The MVDR beamformer can be computed using the spatial covariance matrices (SCMs) for the target speech and the interfering noise [12]. By updating these SCMs sequentially [13], [14], it is possible to track and enhance the target source even if it moves slightly. Thus, this beamformer (referred to as L-MVDR) is regarded as the localization and enhancement approach.

In MVDR, the distortionless constraint is usually given to a single target source direction, but by extending it to multiple directions, it is possible to direct a wider beam to enhance a certain area around the target source [15]. This multiple-constraint MVDR beamformer (referred to as MC-MVDR) is regarded as the area enhancement approach. Note that the MVDR beamformer has a trade-off between the strength of the distortionless constraints and the degrees of freedom of the beamformer (i.e., its noise suppression performance). Therefore, allowing distortion in the target speech (i.e., relaxation of the constraints) may contribute to the improvement of the noise suppression performance [16], [17]. Considering this trade-off, a new area enhancement approach, the relaxed multiple constraint minimum variance (RMC-MV) beamformer, is proposed to improve the denoising performance by relaxing the constraints.

To the best of our knowledge, no multi-constrained MV beamformer has been trained jointly with DNNs. So when focusing on DNN-based MV beamformers, the superiority of the localization and enhancement approach versus the area enhancement approach is not clarified in the situation where the target source moves slightly in the frontal direction of the microphones. The present study therefore attempts to identify the design of MV beamformers that are robust against the small movement of the target source by comparing the performance of DNN-based MV beamformers estimated with these two approaches. Such experimental comparisons would

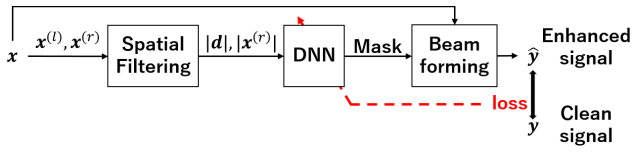

Fig. 1. Schematic diagram of beamforming with DNN-based mask estimation. $\boldsymbol{x}^{(i)}$ denotes microphone observations and $\boldsymbol{d}$ denotes noise-dominated signal. DNN-based mask estimator is trained to minimize noise in beamformer output.

provide useful knowledge for achieving robust and accurate speech enhancement in real environments.

II. Overview of DNN-based Beamforming

This section describes an overview of DNN-based MV beamforming, whose schematic diagram is illustrated in Fig. 1. For simplicity, the frequency indices are omitted in the following equations.

A. Spatial filtering for extracting noise-dominated signal

Let $\boldsymbol{x} \in \mathbb{C}^{F\times T\times M}$ be the output of the short-time Fourier transform (STFT) of the noisy speech signal observed at an M-channel linear microphone array, where F and T denote the number of frequencies and that of frames, respectively. Now, it is assumed that the target source is located in front of the microphones (ideally, on the perpendicular bisector of the two central microphones). The disturbance sounds arriving from the directions other than the front direction (i.e., noise-dominated signal), $\boldsymbol{d} \in \mathbb{C}^{F\times T}$, can be obtained by directing the null to the front direction [1] as:

$$\boldsymbol{d} = \boldsymbol{x}^{(r)} - \boldsymbol{x}^{(l)}, \tag{1}$$

where $\boldsymbol{x}^{(l)}$ and $\boldsymbol{x}^{(r)} \in \mathbb{C}^{F\times T}$ denote the observed signals at the two central microphones.

B. DNN-based mask estimation

DNN is used to estimate a TF mask for extracting the target source. The amplitude spectrum of the noisy speech $|\boldsymbol{x}^{(r)}|$ and that of the noise-dominated signal $|\boldsymbol{d}|$ are input to the mask estimator $\mathcal{D}$ as:

$$\boldsymbol{\mathcal{M}}^{(s)} = \mathcal{D}(|\boldsymbol{x}^{(r)}|, |\boldsymbol{d}|), \tag{2}$$

where $\mathcal{D}(\cdot)$ denotes a nonlinear transformation represented by DNN (the details are explained in Sect. IV-D), and $\boldsymbol{\mathcal{M}}^{(s)} \in \mathbb{R}^{F\times T}$ denotes the TF mask for the target source. The TF mask for the disturbance $\boldsymbol{\mathcal{M}}^{(n)}$ is given by subtracting each component of $\boldsymbol{\mathcal{M}}^{(s)}$ from one.

C. Beamforming

Using TF masks estimated by DNNs, the SCM for the target source and that for the interfering noise, $\boldsymbol{R}^{(s)}$ and $\boldsymbol{R}^{(n)} \in \mathbb{C}^{F\times M\times M}$, are calculated respectively as follows [18]:

$$\boldsymbol{R}^{(s)} = \frac{1}{\sum_t \boldsymbol{\mathcal{M}}_t^{(s)}} \sum_t \boldsymbol{\mathcal{M}}_t^{(s)} \boldsymbol{x}_t \boldsymbol{x}_t^H, \tag{3}$$

$$\boldsymbol{R}^{(n)} = \frac{1}{\sum_t \boldsymbol{\mathcal{M}}_t^{(n)}} \sum_t \boldsymbol{\mathcal{M}}_t^{(n)} \boldsymbol{x}_t \boldsymbol{x}_t^H, \tag{4}$$

where t denotes the time index and H denotes the Hermitian transpose of a matrix or a vector. The filter coefficients of beamformers $\boldsymbol{w} \in \mathbb{C}^{F\times M}$ are computed using the SCMs and applied to the noisy speech to obtain the enhanced signal $\hat{\boldsymbol{y}} \in \mathbb{C}^{F\times T}$ as:

$$\hat{\boldsymbol{y}} = \boldsymbol{w}^H \boldsymbol{x}. \tag{5}$$

The details of estimating beamformers are described in Sect. III.

D. Joint training of deep neural network

A DNN for mask estimation is trained to minimize the squared error between the clean speech signal $\boldsymbol{y}$ and the beamformer output $\hat{\boldsymbol{y}}$ as:

$$\mathcal{L} = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2. \tag{6}$$

It should be noted that the DNN is trained to estimate a mask that is suitable for minimizing the noise in the beamformer output, rather than reducing the noise in the time-frequency masking output.

III. Minimum Variance Beamformers Robust Against Small Movement of Sound Sources

To design MV beamformers that are robust to the movement of the sound sources, this study examines the localization and enhancement approach, which aims to enhance the target source while tracking it, and the area enhancement approach, which aims to enhance the area including the target source. This section describes the MV beamformers to be compared. All beamformers are computed using the SCMs estimated by DNN.

A. MVDR beamformer

The general MVDR beamformer is chosen as the localization and enhancement approach. The MVDR beamformer aims to minimize noise under the constraint that the sound source arriving from the desired direction is not distorted as:

$$\boldsymbol{w}_{\mathrm{MVDR}} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^H \boldsymbol{R}^{(n)} \boldsymbol{w} \quad s.t. \quad \boldsymbol{w}^H \boldsymbol{a} = 1, \tag{7}$$

where $\boldsymbol{a} \in \mathbb{C}^{F\times M}$ denotes the steering vector of the target source. Solving (7) gives the filter coefficients:

$$\boldsymbol{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{R}^{(n)-1}\boldsymbol{a}}{\boldsymbol{a}^H \boldsymbol{R}^{(n)-1}\boldsymbol{a}}. \tag{8}$$

The MVDR beamformer can also be formulated in a form that does not use a steering vector [12] as:

$$\boldsymbol{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{R}^{(n)-1}\boldsymbol{R}^{(s)}}{tr(\boldsymbol{R}^{(n)-1}\boldsymbol{R}^{(s)})}\boldsymbol{u}, \tag{9}$$

where $tr(\cdot)$ denotes the trace of a matrix and $\boldsymbol{u} \in \mathbb{C}^{F\times M\times 1}$ denotes the one-hot vector to choose one output channel. The MVDR beamformer represented in (9) simultaneously localizes and enhances the target source by using the SCM estimated for the target. In this paper, the beamformer in (9) is referred to as the localization-MVDR (L-MVDR) to distinguish from the beamformer in (8).

B. Multiple-constraint MVDR beamformer (MC-MVDR)

By applying the distortionless constraint to multiple directions around the target source, a wider beam can be constructed to enhance the area including the target source. The Multiple-constraint MVDR (MC-MVDR) beamformer, which has multiple distortionless constraints [15], is formulated as:

$$\boldsymbol{w}_{\mathrm{MCMVDR}} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^H \boldsymbol{R}^{(n)} \boldsymbol{w} \quad s.t. \quad \boldsymbol{w}^H \boldsymbol{A} = \boldsymbol{f}, \tag{10}$$

where $A \in \mathbb{C}^{F \times M \times N_c}$ is a matrix whose elements are the steering vectors in multiple directions; N_c denotes the number of constraints; and $\boldsymbol{f} \in \mathbb{R}^{F \times N_c}$ denotes a matrix whose elements are all ones. Solving (10) gives the following solution:

$$\boldsymbol{w}_{\mathrm{MCMVDR}} = \boldsymbol{R}^{(n)-1} \boldsymbol{A} (\boldsymbol{A}^H \boldsymbol{R}^{(n)-1} \boldsymbol{A})^{-1} \boldsymbol{f}. \tag{11}$$

The MC-MVDR beamformer can enhance the target source robustly to source movement within the given constraints. On the other hand, the distortionless constraints in multiple directions decrease the degree of freedom of the filter, leading to a degradation in the noise suppression performance.

C. Relaxed multiple-constraint MV beamformer (RMC-MV)

In general, there is a trade-off between the strength of constraints and the degrees of freedom of a beamformer (i.e., its noise suppression performance). Thus, relaxing the constraints (i.e., allowing for distortion of the target source) of MC-MVDR can improve the denoising accuracy [16], [17]. Based on this idea, a new beamformer, the relaxed multiple-constraint minimum variance (RMC-MV) beamformer is presented in the form:

$$\begin{aligned} \boldsymbol{w}_{\mathrm{RMCMV}} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^H \boldsymbol{R}^{(n)} \boldsymbol{w} \\ s.t. \quad |\boldsymbol{w}^H \boldsymbol{A} - \boldsymbol{f}|^2 \leq E, \end{aligned} \tag{12}$$

where E is the distortion tolerance of the target source. Solving (12) gives the following filter coefficients as:

$$\boldsymbol{w}_{\mathrm{RMCMV}} = (\boldsymbol{R}^{(n)} + \lambda \boldsymbol{A} \boldsymbol{A}^H)^{-1} \lambda \boldsymbol{A} \boldsymbol{f}, \tag{13}$$

where λ is a hyper-parameter to adjust the aforementioned trade-off. The smaller the value of λ is, the better the noise suppression performance becomes, but the distortion of the target source becomes larger due to the relaxation of the constraint. In contrast, increasing the value of λ can reduce the distortion of the target source but deteriorate the noise suppression performance.

IV. Source Enhancement Experiments

To identify an appropriate MV beamformer for the case where the sound sources move slightly but occasionally, experimental comparisons were conducted using the L-MVDR, MC-MVDR, and RMC-MV beamformers. The first one is considered as the localization and enhancement approach and the other two as the area enhancement approach.

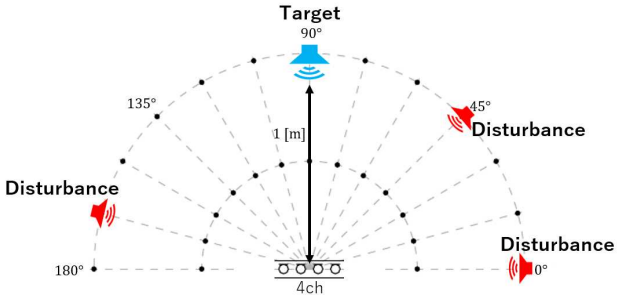

Fig. 2. Acoustic field used in experiments. One target source was placed in front of four-channel microphones and one to three disturbance sources were placed in directions other than front.

A. Speech materials

The simulated acoustic field is illustrated in Fig. 2. The target speech, one to three interfering voices, and ambient noise were observed at the microphones. The room size was $7 \times 5 \times 3$ m, the number of microphones was 4, and the microphone spacing was 3 cm. The sound absorption rate was 0.25, the number of reflections was 20, and the reverberation time RT_{60} was 0.31 s. Pyroomacoustics [19] was used to obtain room impulse responses.

Microphone observations were simulated by convolving the room impulse response with the speech signal (dry source) and then superimposing the ambient noise. The spoken utterances for the target and disturbance sources were selected from the TIMIT corpus [20] and different from each other. From the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) [21], five types of non-stationary ambient noise (NRIVER, NPARK, DLIVING, OOFFICE, and OMEETING) were chosen. The sampling rate was 16 kHz, the frame size and frame shift for STFT were 1024 and 256, respectively. The number of frequency bins was 513. The number of training data, validation data, and testing data were 2000, 320, and 448, respectively.

During training, the target speech source was placed in one of three directions, 80°, 90°, and 100°, and one to three disturbance speech sources were randomly placed in the following eight directions, 0°, 15°, 30°, 45°, 135°, 150°, 165°, and 180°. During testing, the following two types of data were evaluated:

- *matched* data: the target sound source is placed in the position used for training, and
- *mismatched* data: the target source may deviate from the position used for training.

The *mismatched* data were used to examine the robustness of the MV beamformers to the small movement of the sound source. To produce the *mismatched* data, the simulation of the sound field was conducted assuming a situation where a speaker continuously speaks, changing his/her position for each utterance. In this case, the difference in the source direction between utterances was assumed to be zero to two degrees. Such pseudo movement was applied to both the target and disturbance speech sources, but not to the ambient noise. The target source was initially placed in the 80° to 100° direction, and varied within a range of 80° to 100°. The

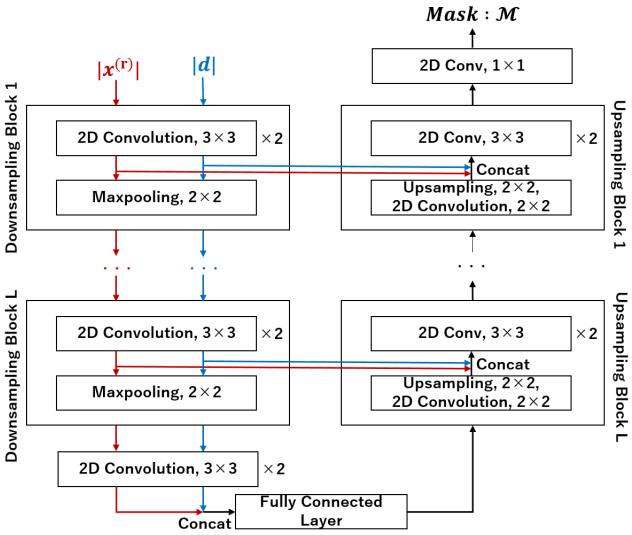

Fig. 3. DNN architecture for mask estimation.

disturbance sources were initially placed in the same directions as training data, and varied within a range of 0° to 45° or 135° to 180°.

B. Setups for updating spatial covariance matrices

Assuming that sound sources move, it is necessary to localize them at certain intervals. In this experiment, the SCMs were updated every 0.51 seconds. The formula to update the SCMs was the same as the one applied in [13]. Let $\boldsymbol{\mathcal{M}}'_{l,t}$ be an estimate of the mask for block B_l, where l is the block index, and $\boldsymbol{R}'_l$ be the SCM for block B_l that is computed using $\boldsymbol{\mathcal{M}}'_{l,t}$. The update formula of the SCM in block B_l is represented in a recursive form as:

$$\boldsymbol{R}_l \leftarrow \frac{\boldsymbol{\mathcal{M}}_{l-1} \cdot \boldsymbol{R}_{l-1}}{\boldsymbol{\mathcal{M}}_{l-1} + \sum_{t \in B_l} \boldsymbol{\mathcal{M}}'_{l,t}} + \frac{\sum_{t \in B_l} \boldsymbol{\mathcal{M}}'_{l,t} \cdot \boldsymbol{R}'_l}{\boldsymbol{\mathcal{M}}_{l-1} + \sum_{t \in B_l} \boldsymbol{\mathcal{M}}'_{l,t}}, \tag{14}$$

where $\boldsymbol{\mathcal{M}}_l$ denotes the sum of the masks in all time frames until block B_l, which is obtained as:

$$\boldsymbol{\mathcal{M}}_l \leftarrow \boldsymbol{\mathcal{M}}_{l-1} + \sum_{t \in B_l} \boldsymbol{\mathcal{M}}'_{l,t}. \tag{15}$$

C. Setups for minimum variance beamformers

The number of constraints should be determined to balance the robustness against the source movement and noise suppression performance. As a result of the preliminary experiment, the constraints for the MC-MVDR and RMC-MV beamformers were applied in two directions, 80° and 100° when the target source was in the 90° direction. More constraints at 80°, 90°, and 100° may improve the robustness against the source movement, but actually deteriorated the noise suppression performance. For the RMC-MV beamformer, the hyper-parameter λ was empirically determined to 1.0×10^6.

TABLE I
SPEECH ENHANCEMENT PERFORMANCE FOR L-MVDR, MC-MVDR, AND RMC-MV BEAMFORMERS. NUMBERS REPRESENT SDR AND STOI FOR OBSERVED SIGNAL, AND SDR IMPROVEMENT AND STOI IMPROVEMENT FOR THREE BEAMFORMERS.

	80, 90, 100° *matched*		80 to 100° *mismatched*	
Beamformer	SDRi [dB]	STOIi	SDRi [dB]	STOIi
Observation	-1.72	0.751	-1.77	0.739
L-MVDR	7.94	0.139	**7.86**	**0.145**
MC-MVDR	4.84	0.127	4.89	0.134
RMC-MV	7.50	0.125	7.48	0.132

D. Neural network architecture

The architecture of the DNN used is shown in Fig. 3. It has a U-Net-like architecture [22], which has been used not only in segmentation but in speech enhancement and separation [23], [24]. The amplitude spectrum of the observed signal $|\boldsymbol{x}^{(r)}|$ and that of the noise-dominated signal $|\boldsymbol{d}|$ were downsampled through four downsampling blocks and then input to two two-dimensional convolution layers. The outputs of the convolution layers were concatenated and input to a fully-connected layer. This process is expected to be equivalent to the subtraction of noise components from the observed signal of a microphone in the latent space. In these encoder parts, the network parameters were shared for the observed signal and the noise-dominated signals. Then, the output of the fully-connected layer was upsampled through four upsampling blocks, in which the downsampled features are concatenated as shown in Fig. 3. The convolution with kernel size 1×1 was performed to adjust the number of channels, and finllay, the TF mask was obtained.

The Adam optimizer [25] was utilized during training. The learning rate was 5.0×10^{-3}. The mini-batch size was 16, and the number of epochs was set to 100.

E. Experimental result

The signal-to-distortion ratio improvement (SDRi) [26] and short-time objective intelligibility improvement (STOIi) [27] were used for evaluation criteria. Table I lists the average values of these measures over the evaluation data.

First, MC-MVDR was compared with RMC-MV to investigate the effect of relaxing the constraints. The RMC-MV beamformer yielded significant improvements in SDR over the MC-MVDR beamformer. It indicates that relaxation of the constraints contributed to the improvement of the noise suppression performance. In contrast, since RMC-MV allows for distortion of the target speech, no improvement was obtained for STOI, which measures objective sound quality.

Next, the robustness of the MV beamformers against source movement was investigated by comparing the performance for *matched* and *mismatched* data. For the MC-MVDR and RMC-MV beamformers, the performance difference for *matched* and *mismatched* data was not significant. This result suggests the effectiveness of the multiple constraints introduced to improve the robustness of these beamformers to small source deviations. The L-MVDR beamformer, which adaptively localizes the target source, also showed no significant difference in the performance for both data. This result suggests that DNNs are

capable enough to enhance a slightly displaced source while tracking it.

Finally, L-MVDR and RMC-MV were compared using *mismatched* data to investigate the superiority of the localization and enhancement approach and the area enhancement approach for the MV beamformer design. The results showed that the L-MVDR beamformer outperformed the RMC-MV beamformer on both SDR and STOI. The results suggest that the high representational power of DNNs enables accurate source tracking, and that speech enhancement based on such source tracking can work well even when the source location changes from time to time.

V. Conclusion

The present paper explored how to design MV beamformers suitable for situations where the target source moves in front of the microphone. The MV beamformers were designed using two approaches: one is to enhance the target source while localizing it, and the other is to enhance the area including the target source. For the latter area enhancement, this paper presented a formulation to relax the constraint of multiple-constraint MV beamformers and a method to train multiple-constrained MV beamformers jointly with DNNs. The experimental comparisons demonstrated the superiority of the approach that localizes and enhances the target source in both SDR and STOI. The robustness of the MV beamformers for the case when the source location changes in a shorter period needs to be investigated in the future.

References

[1] K. Katagiri, T. Yamaguchi, T. Yazu, and M. Nonaka, "Multiple beamforming area sound enhancement (mubase) and stereophonic area sound reproduction (sasr) system," in *SIGGRAPH Asia 2015 Emerging Technologies*, 2015, p. Article No.17.

[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[3] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE ICASSP*, 2017, pp. 246–250.

[4] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–255.

[5] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[6] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in *Proc. of the IEEE*, vol. 57, no.8, 1969, pp. 1408–1418.

[7] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016, pp. 1981–1985.

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

[9] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming," in *Proc. IEEE ICASSP*, 2017, pp. 286–290.

[10] T. Ochiai, S. Watanabe, and T. Hori, "Multichannel end-to-end speech recognition," in *Proc. PMLR*, 2017, pp. 2632–2641.

[11] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-tasnet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. IEEE ICASSP*, 2020, pp. 6384–6388.

[12] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 18, pp. 260–276, Feb. 2010.

[13] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," in *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, 2017, pp. 780–793.

[14] T. Higuchi, K. Kinosita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *Proc. IEEE ICASSP*, 2018, pp. 531–535.

[15] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. on Signal Process.*, vol. 53, pp. 1684–1696, May. 2005.

[16] Y.Kaneda, "Directivity characteristics of adaptive microphone-array for noise reduction (amnor)," in *J. Acoust. Soc. Jpn. (E)*, 1991, pp. 179–187.

[17] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction in hearing aids," in *IWAENC2003*, 2003.

[18] 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. on Audio, Speech, and Language Process.*, vol. 21, pp. 1913–1928, Sept. 2013.

[19] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE ICASSP*, 2018, pp. 351–355.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[21] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, pp. 3591–3596, 2013.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *arXiv:1505.04597*, 2015.

[23] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *arXiv:1806.03185*, 2018.

[24] X. Hao, X. Su, Z. Wang, H. Zhang, and Batushiren, "Unetgan: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," in *Proc. Interspeech 2019*, 2019, pp. 1786–1790.

[25] D. P. Kingmaa and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980*, 2014.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, pp. 1462–1469, July. 2006.

[27] R. H. C. H. Taal, R. C. Hendriks and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, pp. 2125–2136, Sept. 2011.