ADL-MVDR: ALL DEEP LEARNING MVDR BEAMFORMER FOR TARGET SPEECH SEPARATION

Zhuohuang Zhang[1,2⋆], *Yong Xu*[2], *Meng Yu*[2], *Shi-Xiong Zhang*[2], *Lianwu Chen*[2], *Dong Yu*[2]

[1] Indiana University, Bloomington, USA [2] Tencent AI Lab

ABSTRACT

Speech separation algorithms are often used to separate the target speech from other interfering sources. However, purely neural network based speech separation systems often cause nonlinear distortion that is harmful for ASR systems. The conventional mask-based minimum variance distortionless response (MVDR) beamformer can be used to minimize the distortion, but comes with high level of residual noise. Furthermore, the matrix inversion and eigenvalue decomposition processes involved in the conventional MVDR solution are not stable when jointly trained with neural networks. In this paper, we propose a novel all deep learning MVDR framework, where the matrix inversion and eigenvalue decomposition are replaced by two recurrent neural networks (RNNs), to resolve both issues at the same time. The proposed method can greatly reduce the residual noise while keeping the target speech undistorted by leveraging on the RNN-predicted frame-wise beamforming weights. The system is evaluated on a Mandarin audio-visual corpus and compared against several state-of-the-art (SOTA) speech separation systems. Experimental results demonstrate the superiority of the proposed method across several objective metrics and ASR accuracy.

Index Terms— Speech separation, speech enhancement, MVDR, RNN-MVDR, deep learning

1. INTRODUCTION

Environmental noises and adverse room acoustics can greatly affect the quality of the speech signal and therefore degrade the effectiveness of many speech communication systems (e.g., digital hearing-aid devices [1], and automatic speech recognition (ASR) systems [2, 3]). Speech enhancement and speech separation algorithms are thus proposed to alleviate this problem. With the renaissance of neural networks, better objective performance can be achieved using deep learning methods [4, 5, 6, 7]. However, it often results in greater amount of nonlinear distortion on the separated target speech [8, 9, 10], which harms the performance of ASR systems.

The minimum variance distortionless response (MVDR) filters [11] aim to reduce the noise while keeping the target speech undistorted. More recently, MVDR systems with neural network (NN) based time-frequency (T-F) mask estimator can help greatly reduce the word error rate (WER) of ASR systems with less amount of distortion [12, 13, 14], yet they still suffer from residual noise problems since chunk- or utterance-level beamforming weights [15, 16, 14, 8] are not optimal for noise reduction. Some frame-level MVDR weights estimation methods have been proposed, in [17], the authors estimate the covariance matrix in a recursive way. Nevertheless, the calculated frame-wise weights are not stable when jointly trained with NNs. Previous studies have indicated that it is feasible for a recurrent neural network (RNN) to learn the matrix inversion efficiently [18, 19] and that RNNs can better stabilize the process of matrix inversion and principal component analysis (PCA) when jointly trained with NNs.

There are three main contributions in this work, firstly, we propose a novel all deep learning MVDR framework (denoted as ADL-MVDR) where the ADL-MVDR can be jointly trained stably with the front-end filter estimator for frame-level beamforming weights estimation. Secondly, we propose to use RNNs to learn the matrix inversion and PCA from the noise and target speech covariance matrices, instead of utilizing the traditional mathematical approach. Thirdly, instead of using the classical per T-F bin mask, we propose a complex ratio filtering method (denoted as cRF) to further stabilize joint training process and estimate the covariance matrices of target speech and noise more accurately. The RNN components of ADL-MVDR system help to recursively estimate the statistical variables (i.e., inverse of the noise covariance matrix and PCA of the steering vector) in an adaptive way. Meanwhile, a Conv-TasNet variant [9, 10] is adopted as the front-end filter estimator to calculate the frame-level covariance matrices.

The proposed cRF based ADL-MVDR system achieves the best performance in many objective metrics as well as the ASR accuracy. To the best of our knowledge, this is the first pioneering study that applies RNNs to derive the MVDR solution by replacing the matrix inversion and PCA. Note that Xiao et al. [20] once proposed a directly NN-learned beamforming weights method which was not successful due to lack of using noise information, whereas our approach still follows the mask-based MVDR framework and explicitly utilizes the

This work was done while Z. Zhang was a research intern at Tencent AI Lab, Bellevue, USA. ⋆ zhuozhan@iu.edu

noise and speech covariance matrices with RNNs.

The rest of the paper is organized as follows: Section 2 introduces the conventional mask-based MVDR beamformer and Section 3 describes the proposed ADL-MVDR beamformer. We present the dataset and experimental setup in Section 4. Results are reported in Section 5. Finally, we draw conclusions in Section 6.

2. SIGNAL MODEL FOR MVDR BEAMFORMER

This section describes the conventional mask-based MVDR beamformer, the proposed ADL-MVDR beamformer will be introduced in the next section. Consider a noisy speech mixture $\mathbf{y} = [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_M}]^{\mathbf{T}}$ recorded with an M-size microphone array. Let $\mathbf{s}$ represent the clean speech and let $\mathbf{n}$ denote the interfering noise with M channels, then we have

$$\mathbf{Y}(t,f) = \mathbf{S}(t,f) + \mathbf{N}(t,f) \tag{1}$$

where (t, f) indicates the time and frequency indices of the acoustic signals in the T-F domain, and $\mathbf{Y}, \mathbf{S}, \mathbf{N}$ denote the corresponding variables in T-F domain. The separated speech $\hat{\mathbf{s}}_{\mathbf{MVDR}}(t,f)$ can be obtained as

$$\hat{\mathbf{s}}_{\mathbf{MVDR}}(t,f) = \mathbf{h}(f)^{\mathrm{H}}\mathbf{Y}(t,f) \tag{2}$$

where $\mathbf{h(f)} \in \mathbb{C}^M$ represents the MVDR weights at frequency index f and $^{\mathrm{H}}$ stands for the Hermitian operator. The goal of the MVDR beamformer is to minimize the power of the noise while keeping the target speech undistorted, which can be formulated as

$$\mathbf{h_{MVDR}} = \arg\min_{\mathbf{h}} \mathbf{h}^{\mathrm{H}}\mathbf{\Phi_{NN}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^{\mathrm{H}}\boldsymbol{v} = \mathbf{1} \tag{3}$$

here $\mathbf{\Phi_{NN}}$ stands for the covariance matrix of the noise power density spectrum (PSD) and $\boldsymbol{v}(f) \in \mathbb{C}^M$ denotes the steering vector of the target speech. Different solutions can be used to derive the MVDR beamforming weights. In our study, we mainly focus the MVDR solution that is based on the steering vector [17, 21], which can be derived by applying principal component analysis (PCA) on the speech covariance matrix.

$$\mathbf{h}(f) = \frac{\mathbf{\Phi}^{-1}_{\mathbf{NN}}(f)\boldsymbol{v}(f)}{\boldsymbol{v}(f)^{H}\mathbf{\Phi}^{-1}_{\mathbf{NN}}(f)\boldsymbol{v}(f)}, \quad \mathbf{h}(f) \in \mathbb{C}^M \tag{4}$$

here $\mathbf{u} \in \mathbb{C}^M$ is the one-hot vector selecting the reference microphone channel. Note that the matrix inversion and PCA in Eq (4) are not stable especially when jointly trained with neural networks.

A complex ratio mask [5] (denoted as cRM) can be used to estimate the target speech accurately with less amount of phase distortion, which benefits human listeners [5, 22]. In this case, the estimated speech $\hat{\mathbf{s}}_{\mathbf{cRM}}$ and covariance matrix of the speech PSD $\mathbf{\Phi_{SS}}$ can be computed as

$$\begin{aligned} \hat{\mathbf{S}}_{\mathbf{cRM}}(t,f) &= \mathrm{cRM_S}(t,f) * \mathbf{Y}(t,f) \\ \mathbf{\Phi_{SS}}(f) &= \frac{\sum_{t=1}^{T} \hat{\mathbf{S}}_{\mathbf{cRM}}(t,f)\hat{\mathbf{S}}^{\mathrm{H}}_{\mathbf{cRM}}(t,f)}{\sum_{t=1}^{T} \mathrm{cRM}^{\mathrm{H}}_{\mathrm{S}}(t,f)\mathrm{cRM_S}(t,f)} \end{aligned} \tag{5}$$

where $*$ denotes the complex multiplier and $\mathrm{cRM_S}$ represents the estimated cRM for speech target. The noise covariance matrix $\mathbf{\Phi_{NN}}$ can be obtained in a similar way. However, the covariance matrix $\mathbf{\Phi}$ derived here is on the utterance level which is not optimal for each frame, resulting in high level of residual noise.

3. PROPOSED RNN-DERIVED MVDR

In this work, we implement two gated recurrent unit (GRU) [23] based networks (denoted as GRU-Nets) to replace the matrix inversion and PCA in Eq. (4) for frame-level beamforming weights estimation. One advantage of using RNNs is that it utilizes the weighted information from all previous frames and does not need any heuristic updating factors between consecutive frames as needed in recursive approaches [17, 24].

3.1. cRF for covariance matrix estimation

To better utilize the nearby T-F information and stabilize the estimated statistical variables (namely, $\mathbf{\Phi_{SS}}(t,f)$ and $\mathbf{\Phi_{NN}}(t,f)$), we introduce a complex ratio filtering (denoted as cRF) method to estimate the speech and noise components. For each T-F bin, the cRF is applied to its $K \times L$ nearby bins where K and L represent the number of nearby frequency and time bins

$$\begin{aligned} \hat{\mathbf{S}}_{\mathbf{cRF}}(t,f) &= \sum_{t=1}^{L}\sum_{f=1}^{K} \mathrm{cRF}(t,f) * \mathbf{Y}(t,f) \\ \mathbf{\Phi_{SS}}(t,f) &= \frac{\hat{\mathbf{S}}_{\mathbf{cRF}}(t,f)\hat{\mathbf{S}}^{\mathrm{H}}_{\mathbf{cRF}}(t,f)}{\mathrm{cRM}^{\mathrm{H}}_{\mathrm{S}}(t,f)\mathrm{cRM_S}(t,f)} \end{aligned} \tag{6}$$

here $\hat{\mathbf{S}}_{\mathbf{cRF}}$ indicates the estimated speech using the complex ratio filter. The cRF is equivalent to $K \times L$ number of cRMs that each applies to the corresponding shifted version (i.e., along time and frequency axes) of the noisy spectrogram. The frame-level speech covariance matrix is then computed where the center mask of the cRF (i.e., $\mathrm{cRM_S}(t,f)$) is used for normalization. Note that we do not sum over the time dimension of $\mathbf{\Phi_{SS}}$ in order to preserve the frame-level temporal information. The frame-level noise covariance matrix $\mathbf{\Phi_{NN}}(t,f)$ can be obtained in a similar way.

3.2. RNNs for matrix inversion and PCA in MVDR

Here we propose to estimate the steering vector and the inverse of noise covariance matrix with two GRU-Nets. The GRU-Nets can better utilize temporal information from previous frames for statistical terms estimation than conventional frame-wise approaches that are based on heuristic updating factors [17, 24]. Additionally, replacing the matrix inversion with a GRU-Net resolves the instability issue during joint training with NNs. These MVDR coefficients can be obtained

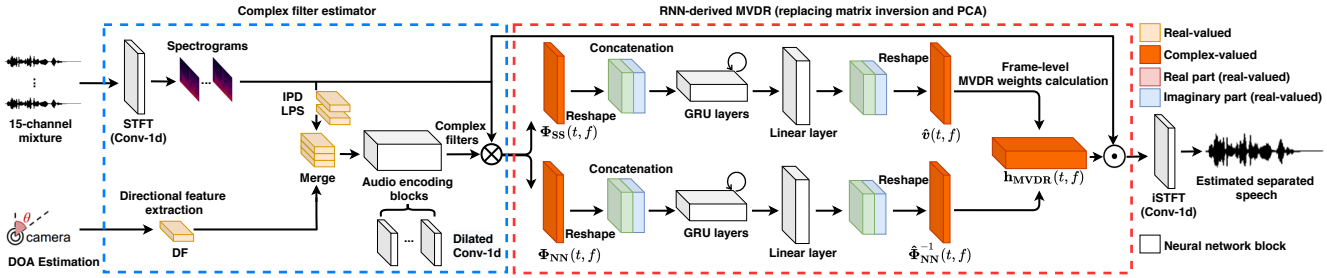

Fig. 1. Network structure of proposed ADL-MVDR beamformer. $\otimes$ and $\odot$ indicate the operations expressed in Eq. (6) and (9), respectively. The complex filter estimation (i.e., cRF) and RNN-derived MVDR (i.e., estimations of $\boldsymbol{v}(t,f)$ and $\mathbf{\Phi}_{\mathbf{NN}}^{-1}(t,f)$) blocks are highlighted in the blue and red dashed boxes, respectively. The real and imaginary parts are reshaped and concatenated before fed into the GRU networks, and then reshaped again as inputs for MVDR weights calculation. The estimated frame-level MVDR weights are then applied to the multi-channel speech.

via the GRU-Nets as

$$\begin{aligned}\hat{\boldsymbol{v}}(t,f) &= \mathbf{GRU\text{-}Net}_{\boldsymbol{v}}(\mathbf{\Phi}_{\mathbf{SS}}(t,f)) \\ \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f) &= \mathbf{GRU\text{-}Net}_{\mathbf{NN}}(\mathbf{\Phi}_{\mathbf{NN}}(t,f))\end{aligned} \tag{7}$$

where the real and imaginary parts of the complex-valued covariance matrix $\mathbf{\Phi}$ are concatenated together as input to the GRU-Net. Here we assume that the explicitly calculated speech and noise covariance matrices are important for RNNs to learn the spatial filtering, which is different from the directly NN-learned beamforming weights in [20]. Leveraging on the temporal structure of RNNs, the model recursively accumulates and updates the covariance matrix for each frame. As shown in Fig. 1, the output of each GRU-Net is fed into a linear layer to obtain the final real and imaginary parts of the complex-valued covariance matrix or steering vector. Then, we compute the frame-level ADL-MVDR weights as

$$\mathbf{h}(t,f) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}{\hat{\boldsymbol{v}}(t,f)^{\mathrm{H}}\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}, \quad \mathbf{h}(t,f) \in \mathbb{C}^{M} \tag{8}$$

Where $\mathbf{h}(t,f)$ is frame-wise and different from the utterance-level weights of conventional mask-based MVDR defined in Eq. (4). Finally, the ADL-MVDR enhanced speech is obtained

$$\hat{\mathbf{S}}_{\mathbf{ADL\text{-}MVDR}}(t,f) = \mathbf{h}(t,f)^{\mathrm{H}}\mathbf{Y}(t,f) \tag{9}$$

4. DATASET AND EXPERIMENTAL SETUP

4.1. System overview and dataset

The proposed system is evaluated based on our previously reported multi-modal multi-channel target speech separation platform [10, 25]. As shown in Fig. 1, we extract the log-power interaural phase difference (IPD) and spectra (LPS) features from the 15-channel microphone recorded mixture that is synchronized with the 180° camera [10]. The direction of arrival (DOA) is roughly estimated using the location of the target speaker's face in the whole camera view [10], then the location guided directional feature (DF) [26] is extracted. The DF feature is then merged with the IPD and LPS features before fed into the audio encoding blocks [10, 25]. We use our previously reported Mandarin audio-visual dataset [10, 25] as the speech corpus. Different from our previous works [10, 25], the lip movement feature is not fed into the model in this study as we focus on the beamforming. The corpus contains 205500 audio clips (roughly 200 hours) with sampling rate set to 16 kHz. The simulated multi-channel audio data contains sources from different speakers (either target or interfering sources). The audios are further mixed with random cuts of noises and different reverberation conditions are applied [10].

4.2. Experimental setup

A 512 frequency size STFT is applied along with 32 ms Hann window and 16 ms step size to extract the audio features. The size of the cRF is empirically set to 3×3 (i.e., center T-F bin with its nearby 8 T-F bins). In the training stage, the audio chunk and batch size are set to 4 s and 12, respectively. Adam optimizer [27] is adopted with the initial learning rate set to $1e^{-3}$. The objective is to maximize the time-domain scale-invariant source-to-noise ratio (Si-SNR) [9]. All models are trained for 60 epochs.

In terms of the GRU-Nets, the $\hat{\boldsymbol{v}}(t,f)$ estimation network consists of two layers of GRU followed by another layer of fully connected (FC) neurons. The hidden size is set to 500 and 250 for the 2-layer GRU with tanh activation function, linear activation function is used for the FC layer with a hidden size of 30. As for $\hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1}(t,f)$ estimation, the corresponding GRU-Net features a similar structure, where each GRU layer contains 500 units with a 450-size FC layer.

Four baseline systems are considered, including a purely NN-based (i.e., a Conv-TasNet variant [9]) cRM system [8], a purely NN-based cRF system, a conventional MVDR-based cRM system [8] and another MVDR-based cRF system (denoted as NN with cRM, NN with cRF, MVDR with cRM, and MVDR with cRF, respectively). We further include two multi-tap (i.e., $[t, t-1]$) MVDR systems that are proposed in our previous work [8], trained with cRM and cRF (denoted as Multi-tap MVDR with cRM/cRF).

Table 1. Experimental results for different speech separation systems across objective evaluation metrics.

Systems/Metrics	PESQ $\in [-0.5, 4.5]$								Si-SNR (dB)	SDR (dB)	WER (%)
	0-15°	15-45°	45-90°	90-180°	1spk	2spk	3spk	Avg.	Avg.	Avg.	Avg.
Reverberant clean (reference)	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	∞	∞	8.26
Noisy Mixture (interfering speech + noise)	1.88	1.88	1.98	2.03	3.55	2.02	1.77	2.16	3.39	3.50	55.14
NN with cRM	2.72	2.92	3.09	3.07	3.96	3.02	2.74	3.07	12.23	12.73	22.49
NN with cRF (3×3)	2.75	2.95	3.12	3.09	3.98	3.06	2.76	3.10	12.50	13.01	22.07
MVDR with cRM [8]	2.55	2.76	2.96	2.84	3.73	2.88	2.56	2.90	10.62	12.04	16.85
MVDR with cRF (3×3)	2.55	2.77	2.96	2.89	3.82	2.90	2.55	2.92	11.31	12.58	15.91
Multi-tap MVDR with cRM (2-tap) [8]	2.70	2.96	3.18	3.09	3.80	3.07	2.74	3.08	12.56	14.11	13.67
Multi-tap MVDR with cRF (2-tap, 3×3)	2.67	2.95	3.15	3.10	3.92	3.06	2.72	3.08	12.66	14.04	13.52
Proposed ADL-MVDR with cRF (3×3)	**3.04**	**3.30**	**3.48**	**3.48**	**4.17**	**3.41**	**3.07**	**3.42**	**14.80**	**15.45**	**12.73**

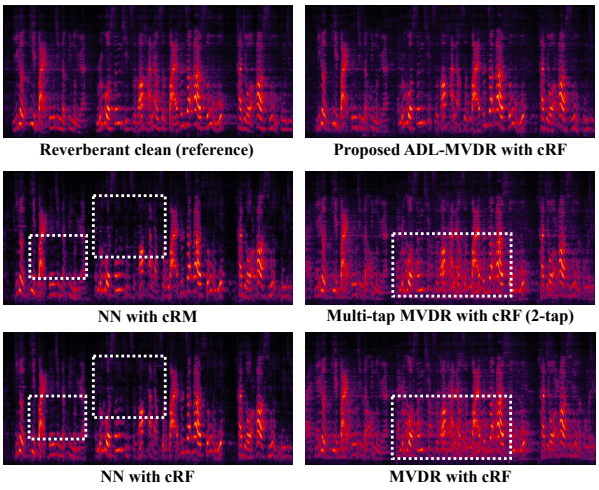

Fig. 2. Sample spectrograms of some evaluated systems

5. RESULTS AND DISCUSSIONS

The systems' performance [1] is evaluated by several objective metrics, including PESQ, Si-SNR [9], and SDR [28]. A Tencent commercial mandarin speech recognition API [29] is used for measuring the WER in this study. The PESQ scores are further presented in specific conditions (i.e., the angle between the target speaker and the closest interfering source, and the number of speakers). Average scores for other metrics are presented. Note that the systems are only trained on speech separation and denoising, without dereverberation.

ADL-MVDR vs. NN: the proposed ADL-MVDR system achieves significantly better results across all metrics and ASR accuracy than purely NN-based systems. Our proposed ADL-MVDR system achieves around 42% improvement on WER (i.e., 12.73% vs. 22.07%) when compared to NN with cRF. Significant improvements aross objective metrics are also observed (i.e., PESQ: 3.42 vs. 3.10, Si-SNR: 14.80 dB vs. 12.50 dB, SDR: 15.45 dB vs. 13.01 dB). For purely NN-based systems, although they perform reasonably well across objective metrics, they perform poorly in ASR system due to large amount of distortion (also highlighted in Fig. 2).

ADL-MVDR vs. MVDR: our proposed ADL-MVDR system achieves about 17% PESQ improvement over the baseline MVDR system with cRF (i.e., 3.42 vs. 2.92). In terms of ASR accuracy, the proposed ADL-MVDR system outperforms MVDR with cRF by a large margin (i.e., 12.73% vs. 15.91%). Considering that the commercial ASR system is already robust to some mild noise, the differences on WER become smaller for multi-tap MVDR systems, yet large gaps can be found in all other metrics (e.g., 0.34 absolute improvement on average PESQ). Although conventional MVDR systems can alleviate the distortion issue, there still remains a lot of residual noise. This can be observed in the objective scores and is also highlighted in Fig. 2. Again, our proposed ADL-MVDR system resolves this issue (i.e., Si-SNR: 14.80 dB vs. 12.66 dB and SDR: 15.45 dB vs. 14.04 dB) while also keeping the target speech undistorted. Specifically, under extreme conditions where interfering sources are very close to each other (e.g., angles between 0-15°), our proposed ADL-MVDR system improves the speech quality by nearly 62% (i.e., PESQ: 3.04 vs. 1.88). The experimental results presented here verify our claims that the proposed ADL-MVDR system not only ensures the distortionless of the target speech (i.e., lowest WER) but also eliminates the residual noise (i.e., highest scores across all objective metrics).

cRM vs. cRF: the NN with cRF achieves better performance in all metrics (e.g., Si-SNR: 12.50 dB vs. 12.23 dB) and ASR accuracy (i.e., 22.07% vs. 22.49%) than NN with cRM. Slight improvements can be found on conventional MVDR systems due to utterance-level weights. The cRF is more important for ADL-MVDR system since ADL-MVDR is recursively getting frame-level weights from the estimated covariance matrices. It indicates that the benefits of introducing T-F filtering include further reducing the residual noise while not distorting the speech.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel all deep learning MVDR method to recursively learn the spatio-temporal filtering for multi-channel target speech separation. The proposed system outperforms prior arts across several objective metrics and ASR accuracy. The future of our proposed ADL-MVDR framework is promising and it could be generalized to many other speech separation systems. We will further verify this idea on single-channel speech separation and move on to dereverberation tasks.

[1] Demos (including real-world recording evaluations) are available at https://zzhang68.github.io/adlmvdr/

7. REFERENCES

[1] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *JASA*, vol. 125, no. 1, pp. 360–371, 2009.

[2] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Interspeech*, 2014.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *LVA/ICA*. Springer, 2015, pp. 91–99.

[4] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[5] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE TASLP*, vol. 24, no. 3, pp. 483–492, 2015.

[6] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018, pp. 696–700.

[7] Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li, "On loss functions and recurrency training for gan-based speech enhancement systems," *arXiv preprint arXiv:2007.14974*, 2020.

[8] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," *arXiv preprint arXiv:2005.03889*, 2020.

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE J-STSP*, 2020.

[11] Barry D. Van V. and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

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

[13] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Interspeech*, 2016, pp. 1981–1985.

[14] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *ICASSP*, 2019, pp. 6745–6749.

[15] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *ICASSP*, 2017, pp. 3246–3250.

[16] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *ICASSP*, 2018, pp. 6697–6701.

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

[18] J. Wang, "A recurrent neural network for real-time matrix inversion," *Applied Mathematics and Computation*, vol. 55, no. 1, pp. 89–100, 1993.

[19] Y. Zhang and S. S. Ge, "Design and analysis of a general recurrent neural network model for time-varying matrix inversion," *IEEE Trans. on Neural Networks*, vol. 16, no. 6, pp. 1477–1490, 2005.

[20] X. Xiao, C. Xu, and et al., "A study of learning based beamforming methods for speech recognition," in *CHiME 2016 workshop*, 2016.

[21] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," in *ICASSP*, 2018, pp. 5734–5738.

[22] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," *arXiv preprint arXiv:2007.14986*, 2020.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] M. Tammen, D. Fischer, and S. Doclo, "Dnn-based multi-frame mvdr filtering for single-microphone speech enhancement," *arXiv preprint arXiv:1905.08492*, 2019.

[25] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J-STSP*, 2020.

[26] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE SLT*, 2018, pp. 558–565.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] "Tencent ASR," https://ai.qq.com/product/aaiasr.shtml.