Recurrent Neural Network-based Estimation and Correction of Relative Transfer Function for Preserving Spatial Cues in Speech Separation

Zicheng Feng
Shenzhen Key Laboratory of Robotics Perception and Intelligence, Southern University of Science and Technology
Shenzhen, China
fengzc503@foxmail.com

Yu Tsao
Research Center for Information Technology Innovation Academia Sinica
Taipei, Taiwan
yu.tsao@citi.sinica.edu.tw

Fei Chen
Shenzhen Key Laboratory of Robotics Perception and Intelligence Southern University of Science and Technology
Shenzhen, China
fchen@sustech.edu.cn

***Abstract*—Although deep learning-based algorithms have achieved great success in single-channel and multi-channel speech separation tasks, limited studies have focused on the binaural output and the preservation of spatial cues. Existing methods indirectly preserve spatial cues by enhancing signal-to-noise ratios (SNRs), and the accuracy of spatial cue preservation remains unsatisfactory. A framework has been proposed before to directly restore the spatial cues of the separated speech by applying relative transfer function (RTF) estimation and correction after speech separation. To further improve this framework, a new RTF estimator based on recurrent neural network is proposed in this study, which directly estimates the RTF from the separated speech and the noisy mixture. The upgraded framework was evaluated with spatialized WSJ0-2mix dataset with diffused noise. Experimental results showed that the interaural time difference and interaural level difference errors of the separated speech were significantly reduced after RTF correction, and its SNR was not sacrificed. The new RTF estimator further improved the performance of the system, with about 5 times smaller model than the previous one. As the proposed framework does not rely on any specific type of model structure, it could be incorporated with both multi-channel and single-channel speech separation models.**

Keywords—binaural speech separation, spatial cue preservation, recurrent neural network,

I. INTRODUCTION

It is a common situation for humans to hear multiple sources of sounds simultaneously in real life, and the human's auditory system has the ability to focus on the desired sources and perceive their locations while suppressing undesired sources , inferred from the interaural time differences (ITDs) and interaural level difference (ILDs) of the sounds reaching both ears [1-2]. Studies in binaural hearing have shown that preserving spatial cues (e.g., ITD and ILD) in speech separation not only offers the localization information of target sources, but also improves speech intelligibility [2-3].

For conventional speech separation methods, several solutions have been proposed to preserve spatial cues. One can simply apply an identical real-value mask to both left and right channels of the speech signal [e.g., 4-5], so that the interaural relation remains unchanged, but the separation quality is sacrificed. A more efficient way is to upgrade existing beamformers with additional constraints on spatial cues. A binaural output version of speech-distortion-weighted multi-channel Wiener filter (SDW-MWF) was introduced in [6], which added a penalty term into the cost function to maintain the relative transfer function (RTF) of the target speech. As a frequently-used representation of spatial cues, RTF is defined as the ratio of the acoustic transfer functions related to the source position and two ears [7]. It is suitable for modeling directional sounds [8], and its phase and magnitude correspond to the ITD and ILD of the sound respectively [6]. In [9], the minimum variance distortionless response (MVDR) was extended to binaural output by adding a linear constraint of RTF into MVDR's cost function. Most of these methods added penalty terms or constraints into the beamformer's cost function to minimize the distortion of spatial cues.

In recent years, deep learning-based approaches have dramatically advanced the performance of speech separation systems. A deep neural network (DNN) can be trained to estimate the time-frequency (T-F) mask of the target speech [10], or directly model the mapping function from the noisy mixture to the target speech [11]. Most of the early separation systems work in the short-time Fourier transform (STFT) domain, while more and more current systems replace the fixed STFT with a learned encoder, e.g., the Conv-TasNet [12]. Some studies extended the application of DNN from single-channel scenarios to multi-channel scenarios by introducing inter-channel features [13] or combining DNN with beamformers [14]. However, the preservation of spatial cues for binaural output has been rarely studied. A multiple-input-multiple-output (MIMO) extension of Conv-TasNet was proposed in [15] where all channels were encoded by different encoders and subsequently concatenated as spatial-sensitive features. Compared to the single-channel TasNet, the MIMO-TasNet achieved significantly better SNR performance and reduced the ITD and ILD errors. More recently, a self-attentive gated recurrent neural network, named SAGRNN [16], exploited the self-attention mechanism and dense connectivity to further improve the speech separation performance, and it was extended to a MIMO system as in MIMO-TasNet. Unlike the beamforming-based methods, the preservation of spatial cues was not included in their training objective, since both of them used signal-to-noise ratio (SNR) as their training objective. Improving SNR will certainly reduce the difference of phase and level between the separated

This work was supported by Shenzhen Key Laboratory of Robotics Perception and Intelligence (ZDSYS20200810171800001), the National Natural Science Foundation of China (61971212), and Shenzhen Sustainable Support Program for High-level University (20200925154002001).

speech and the target speech, which is beneficial to spatial cue preservation, but improving SNR and spatial cue preservation are two fundamentally different tasks. It will be more efficient to optimize the system directly on spatial cues rather than only improving SNR [9].

In our previous work, a framework was proposed to preserve spatial cues for deep learning-based speech separation models [17]. It involves two main steps to directly optimize spatial cue preservation, i.e., 1) to estimate the RTF of the target speech, and 2) to correct the distorted RTF of the separated speech based on the estimated RTF. The estimation of RTF is critical for preserving accurate spatial cues, and conventionally it is obtained by solving the eigenvector decomposition of the covariance matrix [18]. It has been proved that replacing eigenvector decomposition with the recurrent neural network (RNN) can achieve comparable RTF estimation results [e.g., 19]. Therefore in this study, an RNN-based RTF estimator is proposed, which can further utilize the temporal information and frequency context. The new RTF estimator can achieve higher accuracy in RTF estimation, with 5.4 times fewer parameters compared to the previous model. The whole system is evaluated in a complex acoustic scenario including two competing speakers and a diffused background noise. The rest of this paper is as follows. Section II describes the system in detail. Section III explains the experimental setup. The experimental results and discussion are presented in Section IV, and Section V concludes this paper.

II. System Description

A. Overview

Given a binaural noisy mixture $\mathbf{y}[n] \in \mathbb{R}^2$ which consists of I speech sources $\mathbf{x}_i[n] \in \mathbb{R}^2$ (i denotes the index of speakers) and background noise $\mathbf{n}[n] \in \mathbb{R}^2$, the target speech is estimated by a speech separation neural network, yielding the separated speech $\hat{\mathbf{x}}_i[n] \in \mathbb{R}^2$. The STFT coefficients of $\mathbf{x}_i[n]$ and $\hat{\mathbf{x}}_i[n]$ are denoted as $\mathbf{X}_i(t,f)$ and $\hat{\mathbf{X}}_i(t,f)$, respectively. The variable $f \in \{1, \dots, F\}$ is the index of each frequency bin out of a total F bins, and $t \in \{1, \dots, T\}$ is the index of each time frame out of a total F frames. According to the definition in [7], the RTF of the target speech (r_i^{in}) and the separated speech (r_i^{out}) are respectively defined as:

$$r_i^{\text{in}}(t,f) = \frac{X_{\text{L},i}(t,f)}{X_{\text{R},i}(t,f)},$$
$$r_i^{\text{out}}(t,f) = \frac{\hat{X}_{\text{L},i}(t,f)}{\hat{X}_{\text{R},i}(t,f)}, \quad (1)$$

where $X_{\text{L},i}(t,f)$ and $X_{\text{R},i}(t,f)$ represent the left and right channels of $\mathbf{X}_i(t,f)$, respectively. $\hat{X}_{\text{L},i}(t,f)$ and $\hat{X}_{\text{R},i}(t,f)$ are defined similarly. When a sound source locates at a fixed position, its RTF will be time-invariant. In this study, the preservation of spatial cues is achieved by reducing the RTF error between the clean speech and the separated speech. The procedure consists of 3 steps: First, the preliminarily separated speech is obtained by performing a common speech separation task. Second, an estimation of RTF for source i, denoted as $\hat{r}_i(f)$, is extracted from the separated speech by an RTF estimator. Finally, the incorrect RTF of the separated speech is modified by the estimated RTF $\hat{r}_i(f)$, yielding the corrected speech $\tilde{\mathbf{X}}_i(t,f)$. Figure 1 illustrates the whole procedure of the framework.

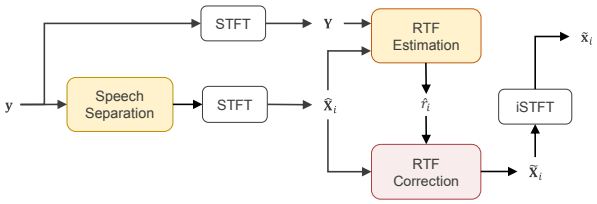

Fig. 1. The flowchart of the proposed spatial cue preservation framework.

B. RTF estimation

In order to alleviate the impact of speech separation error on RTF estimation, a gated recurrent unit (GRU) [20] based estimator is designed to extract the accurate RTF of the target speech. Figure 2 shows the pipeline of the proposed RTF estimator. The inputs of the estimator are the separated speech signals and the noisy mixture in STFT domain, respectively denoted as $\hat{\mathbf{X}}_i(t,f) \in \mathbb{C}^2$ and $\mathbf{Y}(t,f) \in \mathbb{C}^2$. Letting $(\cdot)^{\text{H}}$ denote the conjugate transpose, their covariance matrixes are:

$$\hat{\mathbf{\Phi}}_{\mathbf{XX},i}(t,f) = \hat{\mathbf{X}}_i(t,f)\hat{\mathbf{X}}_i(t,f)^{\text{H}},$$
$$\mathbf{\Phi}_{\mathbf{YY}}(t,f) = \mathbf{Y}(t,f)\mathbf{Y}(t,f)^{\text{H}}. \quad (2)$$

These covariance matrixes are flattened and concatenated into a vector $\mathbf{K}_i(t,f) \in \mathbb{C}^8$. To provide context information of nearby frequency bins above and below the central frequency, the spatial features are further extracted by a 2-dimensional convolutional layer, with a kernel size of C in the frequency dimension, and 1 in the time dimension. The real and imaginary parts of the complex-valued $\mathbf{K}_i(t,f)$ are concatenated together as input to a global layer normalization (gLN) and the 2-D convolutional layer. The dimension of the output channel of the 2-D convolutional layer is set to N, yielding the spatial feature $\bar{\mathbf{U}}_i(t,f) \in \mathbb{R}^N$. The feature vector $\bar{\mathbf{U}}_i(t,f)$ centered at each frequency bin is processed by the same GRU network in parallel, followed by the tanh activation function. The last frame of the GRU network outputs is projected to two complex values, which are $V_{\text{L},i}(f)$ for the left channel and $V_{\text{R},i}(f)$ for the right channel. The final RTF estimation $\hat{r}_i(f)$ is obtained as:

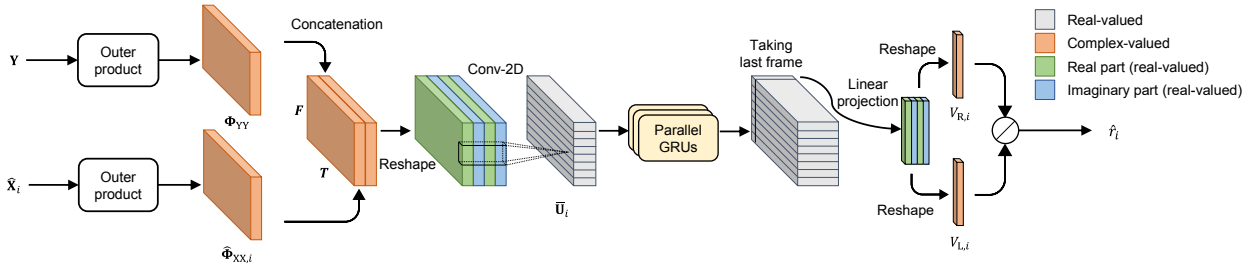

Fig. 2. The architecture of the proposed RTF estimation network. The outer product represents the operation expressed in (2), and ⊘ denotes the element-wise division. T and F denote the size of the tensor, where T is the number of time frames and F is the number of frequency bins.

$$\hat{r}_i(f) = \frac{V_{\mathrm{L},i}(f)}{V_{\mathrm{R},i}(f)}. \quad (3)$$

The estimation result is evaluated by the difference between the RTF of the target speech and the estimated RTF. The final score is calculated by averaging the RTF error of all frequency bins, and weighted by the energy of the target speech, as:

$$\Delta\mathrm{RTF}_i(f) = 10\log_{10}\left[\frac{\sum_f P_{\mathbf{X},i}(f)\frac{|r_i(f)-\hat{r}_i(f)|}{|r_i(f)|}}{\sum_f P_{\mathbf{X},i}(f)}\right],$$

$$P_{\mathbf{X},i}(f) = \frac{1}{T}\sum_t\left(\left\|X_{\mathrm{L},i}(t,f)\right\|_2^2 + \left\|X_{\mathrm{R},i}(t,f)\right\|_2^2\right), \quad (4)$$

where $\Delta\mathrm{RTF}_i(f)$ is the RTF error of source i, $P_{\mathbf{X},i}(f)$ is the energy of $\mathbf{X}_i(t,f)$. The reason for using energy weighting is to reduce the contribution of RTF errors at unimportant frequency bins, since the frequency components of low energy have little effect on the spatial characteristic of the whole speech utterance.

C. RTF correction

After obtaining the estimation of RTF, the distorted RTF of the separated speech has to be changed into the estimated one at each T-F unit, meanwhile the separation quality cannot be sacrificed. It is realized by solving the following optimization problem:

$$\tilde{\mathbf{X}}_i(t,f) = \arg\min_{\tilde{\mathbf{X}}_i(t,f)} \left\|\tilde{\mathbf{X}}_i(t,f) - \hat{\mathbf{X}}_i(t,f)\right\|_2^2,$$

$$\text{s.t.}\frac{\tilde{X}_{\mathrm{L},i}(t,f)}{\tilde{X}_{\mathrm{R},i}(t,f)} = \hat{r}_i(f), \qquad \forall t \in \{1,\dots,T\}. \quad (5)$$

Even though $\tilde{\mathbf{X}}_i(t,f)$ will not be identical to the target speech $\mathbf{X}_i(t,f)$, minimizing $\left\|\tilde{\mathbf{X}}_i(t,f) - \hat{\mathbf{X}}_i(t,f)\right\|_2^2$ can largely ensure that the correction will not introduce too much extra noise.

III. Experiment

A. Dataset

A spatialized and noisy version of the WSJ0-2mix dataset [21] was generated for the training and evaluation of the proposed system. The mono utterances in WSJ0-2mix were convolved by randomly selected head-related impulse response (HRIR) from the ITA database [22], and the location of speech sources was randomly selected from 72 azimuths (with 5° resolution in all directions) and 32 elevations (with 5° resolution from −66° to 90°). Data from 36 subjects were used for training and evaluation, and data from 9 unseen subjects were used for testing. The spatialized speech signals were mixed with randomly selected noises from the DEMAND dataset [23], which contained diffused noises recorded in 18 different scenarios. Each of the mono noise from the DEMAND dataset was rendered by the ITA HRIRs and averaged in all directions, to simulate an isotropic noise field as in [9] and [24]. The noises of 9 scenarios were chosen for training and evaluation, and those of the rest scenarios were used for testing. The noise level relative to the speech mixture was randomly chosen between –10 dB and 10 dB, with the average SNR of the noisy mixtures being –5.57 dB. All audios were downsampled to 8 kHz.

B. Network Configurations

The non-causal MIMO-TasNet was implemented as the speech separation module, with the same configuration reported in [15]. For RTF estimation, the kernel size of the 2-D convolutional layer was set to $(1, 11)$, and the number of output channels (i.e. N) was set to 32. The GRU network included 4 layers of GRU with 64 hidden channels. The analysis window for STFT was a square-root-Hann window, with a frame length of 512 samples, an overlap of 128 samples, and an FFT size of 512 samples. The speech separation module was trained on SNR, and the RTF estimator was trained on ΔRTF. The Adam optimizer [26] was adopted with the initial learning rate set to $1e^{-3}$. These two modules were trained sequentially to avoid the potential problem of balancing multiple tasks in one loss function.

C. Evaluation

Both speech separation quality and the accuracy of pre-served spatial cues were considered in the evaluation. The speech separation quality was evaluated by SNR improvement (ΔSNR), and the preservation of spatial cues was evaluated by the ITD error (ΔITD) and the ILD error (ΔILD) between the estimated speech signals and their clean references. The ITD and ILD were calculated by the same method in [16], which originated from a sound localization algorithm [27]. Specifically, the binaural speech signal was filtered by a gammatone filter-bank with 32 channels, and then segmented into T-F units of 160 samples (i.e. 20 ms). The time delay (measured by cross-correlation) and the level differences were calculated as the T-F unit level ITDs and ILDs, respectively. The ITD of the entire utterance was summarized by plotting the histogram of the ITD for each T-F unit, and taking the center value of the highest bin. Due to the dominant role of ITD in localization at low frequencies, only the T-F units under 1.5 kHz were taken into count. The ILD was summarized through a similar procedure, but separately counted at 3 different filter-banks with center frequencies at roughly 2.07, 3.08, and 3.75 kHz, respectively, because of the frequency-dependence of ILD.

IV. Results and Discussion

Table 1 shows the evaluation results of the proposed system in the noisy condition, in which MIMO-TasNet is chosen as the speech separation module. The performance of MIMO-TasNet itself and binaural MWF are also listed as baselines. The covariance matrixes required by binaural MWF were generated from the separation result of MIMO-TasNet. Three different types of RTF estimators were compared in the same framework, which were: the GRU network proposed in this paper (named as "RNN-EVD"), the previously proposed RTF estimator in [17] (named as "RNN-mask"), and the conventional method based on eigenvector decomposition (named as "EVD"). Combined with the best RTF estimator, the proposed framework significantly outperformed MIMO-TasNet in spatial cue preservation. ΔITD was reduced from 28.95 μs to 17.60 μs, and ΔILDs in three frequency bands were reduced from 0.96, 0.79, and 1.34 dB to 0.36, 0.29, and 0.53 dB, respectively. Besides, although the RTF corrector was not designed to improve the separation quality, ΔSNR still slightly increased by about 0.7 dB after correction, which is consistent with the results in the noise-free condition [17]. These results indicate that the RTF corrector can efficiently restore the spatial cue of the separated speech. The RNN-EVD estimator produced the most accurate RTF, and ΔRTF was 3.62 dB lower than that of the eigenvector decomposition

TABLE I. SEPARATION QUALITY AND ACCURACY OF PRESERVED SPATIAL CUES OF DIFFERENT METHODS ON SPATIALIZED WSJ0-2MIX WITH DIFFUSED NOISE.

Method				ΔRTF (dB)	ΔSNR (dB)	ΔITD (μs)	ΔILD (dB)		
Separation module	RTF estimator	RTF estimator input							
		Mixture	Context				2.07 kHz	3.08 kHz	3.75 kHz
BMWF [6]	EVD	—	—	—	15.61	25.66	0.76	0.68	1.07
MIMO-TasNet [15]	None	—	—	—	20.96	28.95	0.96	0.79	1.34
	EVD	—	—	–10.42	21.44	23.63	0.76	0.68	1.05
	RNN-mask [17]	—	—	–13.06	21.60	20.27	0.51	0.38	0.64
	RNN-EVD	✓	✓	**–14.04**	**21.67**	**17.60**	**0.36**	**0.29**	**0.53**
		✓	×	–13.59	21.64	18.91	0.41	0.31	0.55
		×	✓	–11.19	21.51	22.13	0.63	0.51	0.80
		×	×	–11.04	21.49	23.29	0.64	0.51	0.80
DPTNet [29]	None	—	—	—	15.62	76.95	1.85	1.85	2.34
	EVD	—	—	–7.47	16.52	27.52	3.07	2.94	3.21
	RNN-EVD	✓	✓	**–14.03**	**16.86**	**19.12**	**0.45**	**0.39**	**0.68**

method, which shows the superior accuracy of the proposed RTF estimator. The evaluation results with different noise levels are separately plotted in Fig. 3, where the noise levels relative to speech mixture were set to –12, –6, 0, and 6 dB. It can be found that the proposed system provided the best result constantly.

In addition to the improvement in performance, the RNN-EVD estimator has a much smaller model size with 135K parameters compared to the RNN-mask with 735K parameters. This is because the spatial features are processed at multiple sub-bands rather than full band in RNN-EVD estimator, which makes the RNN model learn common features across multiple sub-bands, and helps to save the model size.

To clarify the importance of each component in the RTF estimator, several variants of the estimator were created and tested, which are also shown in Table 1. The column "Mixture" indicates whether the input of the RTF estimator includes the noisy mixture, and the column "Context" indicates whether the input includes the frequency context. For variants without frequency context, the number of context frequency bins (i.e. C) is set to 0. It is shown that introducing frequency context and extra features from the noisy mixture reduced the ΔRTF by 3 dB, from –11.04 dB to –14.04 dB. The improvement in ΔRTF reveals that the proposed RTF estimator can utilize the complementary features in separated speech and noisy mixture, and the relation of RTF among nearby frequency bins to improve the accuracy of the estimated RTF.

Since the RTF preservation modules in our proposed framework do not rely on any specific speech separation model, they could be flexibly combined with any deep learning-based speech separation model. To illustrate this property, the framework was tested with a single-channel speech separation model called DPTNet [28]. The DPTNet is an end-to-end speech separation model which incorporates the transformer into the dual-path RNN network [29]. The configuration of DPTNet was not modified, except that the frame length of the encoder was set to 4 samples, due to the limitation of hardware resources. The DPTNet was applied to each channel of the binaural input independently, so that the interaural features were not available. The RTF estimator was retrained based on the separation result of DPTNet. As a result, the spatial cues of the separated speech contained larger distortion than the multi-channel models. The lower part of Table 1 shows the evaluation results of DPTNet before and after RTF correction in the noisy condition. Even though the spatial cues were highly distorted, the RNN-EVD estimator still provided a relatively accurate estimation of RTF, with a ΔRTF of –14.03 dB. Consequently, the gap between MIMO-TasNet and DPTNet on spatial accuracy was narrowed after the RTF correction. This indicates that the proposed framework is suitable for both multi-channel and single-channel speech separation models, and the proposed RTF estimator is robust to spatial cue distortion caused by the speech separation models.

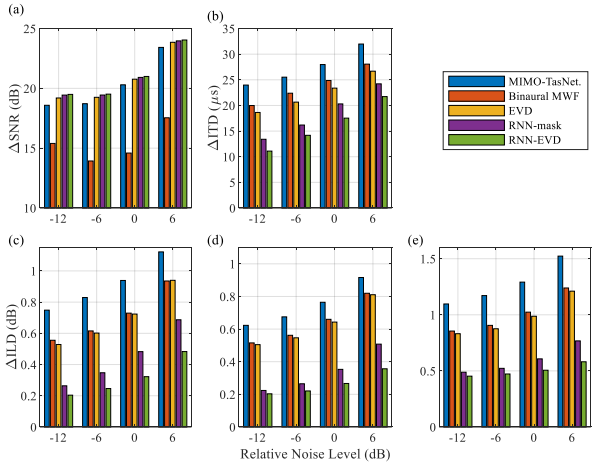

Fig. 3. The evaluation results of different methods with constant relative noise levels. (a) SNR improvement. (b) ITD error. (c)-(e) ILD errors at 2.07, 3.08, and 3.75 kHz, respectively.

V. CONCLUSIONS

In this paper, a new RNN-based RTF estimator is proposed to upgrade the framework of preserving spatial cues for speech separation models. The framework with the new RTF estimator was evaluated in a 2-speaker scenario with diffused noise. The experimental results showed that the framework can further reduce the ITD and ILD errors of the separated speech in the noisy condition, and slightly increase the SNR at the same time. The proposed framework does not rely on any specific type of speech separation model, hence it is suitable for both multi-channel and single-channel speech separation models. Future work could include the real-time solution to spatial cue preservation, and its application to assistive hearing devices (e.g., hearing aids).

REFERENCES

[1] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," The Journal of the Acoustical Society of America, vol. 106, no. 6, pp. 3578–3588, 1999.

[2] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," The Journal of the Acoustical Society of America, vol. 115, no. 2, pp. 833–843, 2004.

[3] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," The Journal of the Acoustical Society of America, vol. 83, no. 4, pp. 1508–1516, Apr. 1988.

[4] M. Zohourian and R. Martin, "Gsc-based binaural speaker separation preserving spatial cues," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 516–520.

[5] M. Azarpour and G. Enzner, "Binaural noise reduction via cue-preserving MMSE filter and adaptive-blocking-based noise PSD estimation," EURASIP Journal on Advances in Signal Processing, vol. 2017, no. 1, p. 49, Jul. 2017.

[6] S. Haykin and K. R. Liu, Handbook on array processing and sensor networks, vol. 63. John Wiley & Sons, 2010.

[7] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural multi-channel wiener filtering for hearing aids: preserving interaural time and level differences," in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, May 2006, vol. 5, p. V–V.

[8] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[9] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[10] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[11] E. Nachmani, Y. Adi, and L. Wolf, "Voice Separation with an Unknown Number of Multiple Speakers," in International Conference on Machine Learning, Nov. 2020, pp. 7164–7175.

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[13] R. Gu et al., "End-to-end multi-channel speech separation," arXiv:1905.06286, May 2019.

[14] 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 ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 6384–6388.

[15] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 6404–6408.

[16] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated rnn for binaural speaker separation with interaural cue preservation," IEEE Signal Processing Letters, vol. 28, pp. 26–30, 2021.

[17] Z. Feng, Y. Tsao, and F. Chen, "Estimation and correction of relative transfer function for binaural speech separation networks to preserve spatial cues," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021, pp. 1239–1244.

[18] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2016, pp. 5210–5214.

[19] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, pp. 6089–6093.

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

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 31–35.

[22] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in Proceedings of Meetings on Acoustics 172ASA, 2016, vol. 29, no. 1, p. 050002.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," Proc. Mtgs. Acoust., vol. 19, no. 1, p. 035081, Jun. 2013.

[24] H. As'ad, M. Bouchard, and H. Kamkar-Parsi, "Beamforming designs robust to propagation model estimation errors for binaural hearing aids," IEEE Access, vol. 7, pp. 114837–114850, 2019.

[25] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-speech: End-to-end multi-channel multi-speaker speech recognition," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Dec. 2019, pp. 237–244.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," ICLR, 2015.

[27] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[28] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," arXiv:2007.13975, Aug. 2020.

[29] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 46–50.