Estimation and Correction of Relative Transfer Function for Binaural Speech Separation Networks to Preserve Spatial Cues

Zicheng Feng*, Yu Tsao† and Fei Chen*
*Department of Electrical and Electronic Engineering, Southern University of Science and Technology
†Research Center for Information Technology Innovation, Academia Sinica
E-mail: fchen@sustech.edu.cn

***Abstract*— Deep learning based approaches have achieved great success in mono-channel and multi-channel speech separation, but limited studies have focused on the binaural output, not even to mention the preservation of spatial cues. Existing speech separation networks preserve spatial cues by improving the signal-to-noise ratio (SNR) of the separated speech, regardless of the different requirements between reducing noise and preserving spatial cues. This work proposed a framework to optimize spatial cue preservation for binaural speech separation. It consisted of a relative transfer function (RTF) corrector that modified the distorted RTF of the separated speech into a correct one, and an RTF estimator to extract the correct RTF. A new RTF estimator was designed to obtain an accurate RTF. The framework was evaluated on a binaural version of WSJ0-2mix dataset, which was spatialized by anechoic head-related impulse responses. Experimental results showed that the proposed framework significantly reduced the interaural time difference (ITD) and interaural level difference (ILD) errors of the existing binaural separation networks, but did not notably sacrifice the SNR of the separated speech signals.**

I. INTRODUCTION

Speech communication in real life is frequently contaminated by various types of noises and interferences, and human listeners can focus on a target speech and extract it from undesired sources. Especially when the target source and undesired sources are spatially separated, the human auditory system can easily separate them utilizing spatial cues like interaural time differences (ITDs) and interaural level differences (ILDs) [1-2]. Hence, preserving spatial cues can provide extra advantages in intelligibility improvement for binaural speech separation algorithms [2-4], and offer localization information of sound sources for listeners.

The spatial cue preservation for conventional speech separation or enhancement methods has been well studied [e.g., 5-9]. One simple way is to apply an identical real-value mask to both left and right channels [e.g., 5-6], so that the ITD and ILD of the original speech signals will not be modified, but the separation performance is sacrificed. A more efficient way is to apply beamformers with constraints on spatial cues. The speech-distortion-weighted multi-channel Wiener filter (SDW-MWF) has been extended to binaural output in [7], where a cost function of interaural transfer functions (ITFs) is added to the total cost function as a penalty term to preserve spatial cues. Binaural minimum variance distortionless response (BMVDR) beamformer [8] achieves preservation of spatial cues by introducing a linear constraint of the relative transfer function (RTF) into MVDR's cost function. Relaxed binaural linearly constrained minimum variance (LCMV) beamformer [9] replaces the linear equality constraints in BMVDR with inequality constraints into cost function, which ignores some spatial cues that are inaudible to humans and further improves its noise reduction ability. These methods adopt a similar strategy of preserving spatial cues, i.e., adding penalty terms or constraints related to the distortion of spatial cues into the beamformer's cost function. As a frequently used metric for quantifying spatial cues, ITF is defined as the ratio of the acoustic transfer functions related to the source position and the two ears [10], and it is equivalent to RTF for a binaural setup.

In recent years, the development of deep learning algorithms has dramatically improved the performance of speech separation systems. Given a noisy mixture in the short-time Fourier transform (STFT) domain, the clean speech signal can be separated by applying time-frequency (T-F) masks which are estimated by a deep neural network [11]. Besides, deep learning based speech separation can also work in time domain, where the mixture waveform is directly modeled. A typical example is Conv-TasNet [12], which replaces STFT and inverse STFT (iSTFT) by trainable encoder and decoder for feature extraction, and achieves comparable performance to the T-F domain systems.

Except for the mono-channel approaches, the multi-channel separation systems based on deep learning have been also investigated. Some studies attempted to extend mono-channel separation systems into multi-channel separation systems [e.g., 13-14] by introducing inter-channel features, while others combined neural networks with beamformers [e.g., 15-16]. 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 [17] which exploited parallel encoders to extract inter-channel spatial features, and the ITD and ILD of the unmixed speech signals were preserved. Based on the gated recurrent neural network [18], the later presented MIMO self-attentive gated RNN (SAGRNN) [19] surpasses the MIMO-

TasNet on both separation performance and spatial cue preservation by incorporating self-attention mechanism and dense connectivity. Both of them use signal-to-noise ratio (SNR) as their training objective, since the ITD and ILD errors are already involved in SNR. It cannot be denied that improving SNR will largely benefit spatial cue preservation, but improving SNR and spatial cue preservation are two fundamentally different tasks. Examples can be found in [8], where perfect preservation of spatial cues does not require a perfect speech separation. If one aims to preserve spatial cues, optimizing the system directly on spatial cues rather than SNR could be a more efficient way.

In this study, a framework is proposed for binaural speech separation to further improve the accuracy of the preserved spatial cues. The spatial cues of the separated speech signals are directly controlled by an RTF corrector, and an RTF estimator is designed to ensure the accuracy of the spatial cues. The rest of this paper is organized as follows. Section II describes the speech separation system in details. Section III introduces the experimental setup. The experimental results and discussion are presented in Section IV, and Section V concludes this paper.

II. System Description

A. Problem Definition

Assuming that a time-domain binaural mixture signal consisting of C sources is formulated as $\mathbf{y}[n] = \sum_{i=1}^{C} \mathbf{x}_i[\mathrm{n}]$, the sound propagation from each source to ears is usually modeled by head-related impulse response (HRIR) [20], as:

$$\begin{cases} x_{\mathrm{L},i}[n] = s_i[n] \circledast h_{\mathrm{L},i}[n] \\ x_{\mathrm{R},i}[n] = s_i[n] \circledast h_{\mathrm{R},i}[n] \end{cases} \quad i = 1, \dots, C, \tag{1}$$

where $s_i[n]$ is the monaural signal of source i, $x_{\mathrm{L},i}[n]$ and $x_{\mathrm{R},i}[n]$ represent, respectively, the left and right channels of the received speech signal $\mathbf{x}_i[\mathrm{n}]$, and $h_{\mathrm{L},i}[n]$ and $h_{\mathrm{R},i}[n]$ are the HRIRs of the corresponding sources. The symbol $\circledast$ represents the convolution operator. According to the narrowband approximation [21], the convolution in time domain can be approximate by the multiplication in STFT domain as:

$$\mathbf{X}_i(t,f) = S_i(t,f)\mathbf{A}_i(f)\,, \tag{2}$$

where $\mathbf{X}_i(t,f)$ and $S_i(t,f)$ are the STFTs of $\mathbf{x}_i[n]$ and $s_i[n]$, respectively, and $\mathbf{A}_i(f)$ denotes the head-related transfer function (HRTF). In the remainder of the paper, the variables f and t will be omitted for the sake of brevity. The binaural cues can be extracted through the relative transfer function (RTF) which is defined as the ratio of the left and right channels, as:

$$\mathrm{RTF}_i^{\mathrm{in}} = \frac{X_{\mathrm{L},i}}{X_{\mathrm{R},i}} = \frac{A_{\mathrm{L},i}}{A_{\mathrm{R},i}}. \tag{3}$$

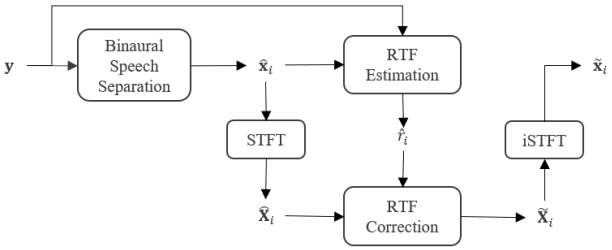

Fig. 1 Flowchart of the proposed spatial cue preservation framework.

RTF is chosen as the representation of spatial cues for two reasons: i) It is suitable for modeling directional sounds [22]; and ii) ITD and ILD can be directly derived from RTF's phase and magnitude, respectively [7]. Thus, the preservation of spatial cues can be achieved by maintaining the RTF of the input sources as:

$$\mathrm{RTF}_i^{\mathrm{out}} = \frac{\hat{X}_{\mathrm{L},i}}{\hat{X}_{\mathrm{R},i}} = \mathrm{RTF}_i^{\mathrm{in}} = \frac{A_{\mathrm{L},i}}{A_{\mathrm{R},i}}, \tag{4}$$

where $\hat{X}_{\mathrm{L},i}$ and $\hat{X}_{\mathrm{R},i}$ represent the left and right channels of estimated speech source $\hat{\mathbf{X}}_i$, respectively.

B. Overview of the Framework

The proposed framework consists of 3 modules: binaural speech separator, the RTF estimator, and the RTF corrector. The speech mixture is initially separated by an arbitrary binaural separation neural network with binaural output. Multiple-input single-output (MISO) separation systems are also alternative, which can estimate the left and right channels separately. The separated speech of source i is denoted as $\hat{\mathbf{x}}_i$, and its STFT form is $\hat{\mathbf{X}}_i$. Then the RTF estimator extracts accurate RTF $\hat{r}_i$ from the separated speech and provides it to the RTF corrector. Finally, the distorted RTF of the separated speech is modified by the RTF corrector in the STFT domain, according to the estimated accurate RTF. The corrected speech is denoted as $\tilde{\mathbf{X}}_i$, and it is transformed into time-domain as the final result $\tilde{\mathbf{x}}_i$. An illustration of the framework is displayed in Fig. 1.

C. RTF Estimation

Given an estimated speech from a speech separation network, a commonly used estimation of RTF is the eigenvector decomposition of covariance matrix [23]:

$$\begin{aligned} \hat{\mathbf{\Phi}}_i &= \frac{1}{T}\sum_t \hat{\mathbf{X}}_i\, \hat{\mathbf{X}}_i^{\mathrm{H}} \\ \hat{\mathbf{A}}_i &= \mathcal{P}\{\hat{\mathbf{\Phi}}_i\} \\ \hat{r}_i &= \frac{\hat{A}_{\mathrm{L},i}}{\hat{A}_{\mathrm{R},i}}, \end{aligned} \tag{5}$$

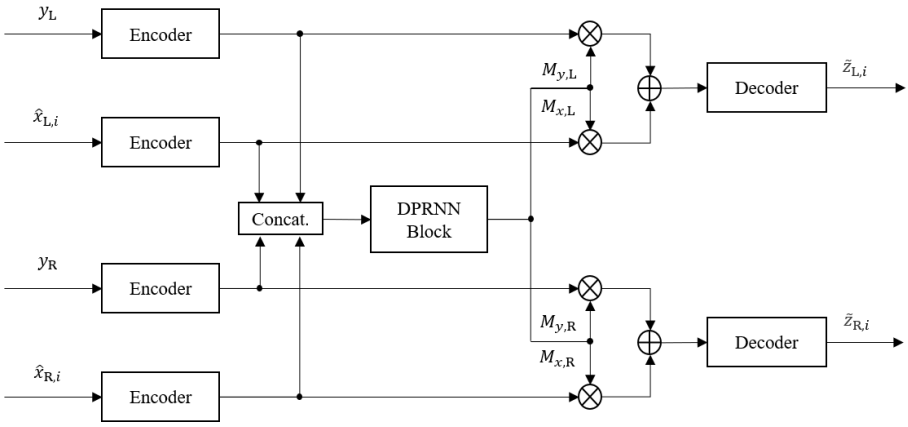

Fig. 2 The architecture of the proposed RTF estimation network.

where $\widehat{\mathbf{\Phi}}_i$ denotes the estimated covariance matrix of signal, T denotes the total number of frames, $\mathcal{P}\{\cdot\}$ extracts the principal eigenvector, and $\hat{r}_i$ denotes the estimated RTF.

Considering that the errors of speech separation and narrowband approximation could affect the estimation accuracy of RTF, here a neural network based estimator is designed to overcome these errors. The architecture of the estimator is illustrated in Fig. 2. Similar to [13] and [17], parallel encoders are used for feature extraction. The separated speech signals $\hat{x}_{\mathrm{L},i}$, $\hat{x}_{\mathrm{R},i}$ and speech mixtures y_{L}, y_{R} are encoded by a shared encoder. The encoded inputs are concatenated and fed to a DPRNN [24] mask estimator. Four masks in total are estimated and applied to encoded inputs. The masked signals that correspond to the same channel are summed and decoded, producing a binaural time-domain signal $\tilde{\mathbf{z}}_i = [\tilde{z}_{\mathrm{L},i} \quad \tilde{z}_{\mathrm{R},i}]^{\mathrm{T}}$. The function of the separator is not to separate different sources, but to separate the RTF-preserved components from the RTF-distorted components. Thus, the output signal will be notably different from the reference speech signal, but with accurate RTFs. Let $\tilde{Z}_{\mathrm{L},i}$ and $\tilde{Z}_{\mathrm{R},i}$ represent the STFTs of $\tilde{z}_{\mathrm{L},i}$ and $\tilde{z}_{\mathrm{R},i}$, respectively, the estimation of RTF $\hat{r}_i$ is summarized by:

$$\hat{r}_i = \frac{\sum_t W_{\tilde{Z},i} \times \frac{\tilde{Z}_{\mathrm{L},i}}{\tilde{Z}_{\mathrm{R},i}}}{\sum_t W_{\tilde{Z},i}}, \qquad W_{\tilde{Z},i} = \left\|\tilde{Z}_{\mathrm{L},i}\right\|_2^2 + \left\|\tilde{Z}_{\mathrm{R},i}\right\|_2^2. \tag{6}$$

The RTF estimation error is used as the training objective, which is defined as the average error across frequencies weighted by speech magnitude:

$$\Delta\mathrm{RTF} = 10\log_{10}\left(\frac{\sum_f \frac{|r_i - \hat{r}_i|}{|r_i|} W_{X,i}}{\sum_f W_{X,i}}\right),$$
$$W_{X,i} = \sum_t \left\|X_{\mathrm{L},i}\right\|_2^2 + \sum_t \left\|X_{\mathrm{R},i}\right\|_2^2. \tag{7}$$

The motivation for using magnitude weights is that: RTF components at different frequencies will not contribute equally to the RTF correction result. Generally, the RTF errors at T-F bins with large magnitude will bring severe performance degradation to the correction results, while the influence will be much smaller at T-F bins with small magnitude. The magnitude weights are expected to guide the network to focus on frequencies that are critical to the RTF correction.

D. RTF Correction

To ensure that the separated speech signals maintain the RTFs of the original unmixed speech sources, the outputs of the separation neural network are modified by solving the following optimization problem:

$$\widetilde{\mathbf{X}}_i = \arg\min_{\widetilde{\mathbf{X}}_i} \left\|\widetilde{\mathbf{X}}_i - \widehat{\mathbf{X}}_i\right\|_2^2 \quad \text{s.t.} \ \frac{\tilde{X}_{\mathrm{L},i}}{\tilde{X}_{\mathrm{R},i}} = \mathrm{RTF}_i^{\mathrm{in}}. \tag{8}$$

However, the actual RTFs are not available in the real situation. $\mathrm{RTF}_i^{\mathrm{in}}$ has to be replaced by the estimated one, i.e., $\hat{r}_i$. The closed-form solution of (8) is easy to obtain as:

$$\widetilde{\mathbf{X}}_i = \mathbf{d}_i\left(\mathbf{d}_i^{\mathrm{H}}\mathbf{d}_i\right)^{-1}\mathbf{d}_i^{\mathrm{H}}\widehat{\mathbf{X}}_i, \qquad \mathbf{d}_i = \begin{bmatrix}\hat{r}_i \\ 1\end{bmatrix}. \tag{9}$$

The purpose of this correction is to enforce the RTF of the input source $\mathbf{X}_i$ and output source $\widetilde{\mathbf{X}}_i$ to be identical, meanwhile the spectra differences between $\widetilde{\mathbf{X}}_i$ and $\widehat{\mathbf{X}}_i$ are minimized to prevent the correction from introducing too much new noise. From the perspective of geometrical explanation, the process of RTF correction is to project the input estimate to the nearest point on the subspace of the RTF-preserved estimates, as illustrated in Fig. 3. If the estimated RTF is accurate, the projected point could be closer to the speech reference. As a result, the separated speech might have a chance to obtain a higher SNR after RTF correction.

TABLE I
SEPARATION PERFORMANCE AND SPATIAL CUE PRESERVATION OF DIFFERENT METHODS

Method	ΔRTF (dB)	ΔSNR (dB)	ΔITD (μs)	ΔILD (dB)		
				2.07 kHz	3.08 kHz	3.75 kHz
MIMO-TasNet	–	21.02	19.64	0.73	0.66	0.97
MIMO SAGRNN	–	**26.88**	14.95	0.53	0.45	0.70
Oracle	–	22.36	7.77	0.24	0.22	0.37
Eig.	-15.22	21.99	14.58	0.52	0.54	0.83
Proposed	**-20.21**	22.31	**8.52**	**0.29**	**0.32**	**0.46**
Masked mix. only	-19.66	22.30	8.72	0.31	**0.32**	0.47
Masked sep. only	-18.42	22.28	9.72	0.32	0.34	0.46

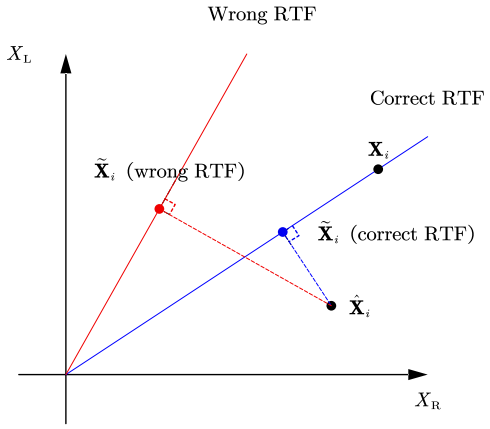

Fig. 3 The geometric illustration of RTF correction with examples of correct RTF and wrong RTF.

III. EXPERIMENT

A. Datasets

A spatialized and anechoic version of WSJ0-2mix dataset [25] was created for training and evaluation of the proposed speech processing system. The mono-channel utterances in WJS0-2mix were convolved by randomly selected HRIR from CIPIC database [26]. The CIPIC HRTF database contains HRIR of 45 subjects, covering 25 azimuths (from $-80°$ to $-80°$) and 50 elevations (from $-90°$ to $-270°$). Data from 36 subjects were used for training and evaluation, and data from 9 unseen subjects were used for testing. All audios were downsampled to 8 kHz.

B. Network Configurations

The non-causal MIMO-TasNet was implemented as the binaural speech separation module, with the same configuration reported in [17]. For the RTF estimation, both the proposed network and eigenvector decomposition were evaluated. Linear encoder and decoder with 2 ms filter length were used in the proposed RTF estimator, and the number of filters was 64. The DPRNN block was implemented with 128 bottleneck channels and 128 hidden channels. The analysis window for STFT was a square-root-Hann window, with frame length 512 samples, overlap 128 samples, and an FFT size of 512 samples.

C. Evaluation

The accuracy of preserving spatial cues was evaluated in the same way as [19], which applied a binaural sound localization algorithm [27] to compute the ITD and ILD of binaural speech signals. The ITD for each T-F unit was firstly plotted as a histogram, and then one ITD value was summarized for the whole utterance, by taking the center value of the highest bin. Only the frequency bands below 1.5 kHz were taken into count when evaluating ITD, due to its dominant role in localization at low frequencies. Following a similar procedure to evaluate ITD, ILD is separately evaluated at 3 different filter-banks with their center frequencies at roughly 2.07, 3.08, and 3.75 kHz, which was due to the frequency-dependence of ILD. The ITD error (ΔITD) and ILD error (ΔILD) were respectively calculated as the difference of the summarized ITD and ILD values between the estimated speech signals and clean references. The separation performance of the system was evaluated by SNR improvement (ΔSNR).

IV. RESULTS AND DISCUSSION

The evaluation results are presented in Table I. The performance of MIMO SAGRNN in the noise-free condition is also presented as a baseline. The RTF correction was evaluated with providing three different RTFs: i) an oracle RTF which was calculated from clean speech, named as "oracle"; ii) an estimated one from eigenvector decomposition, named as "eig."; and iii) an estimated one from proposed RTF estimator, named as "proposed". The speech signals separated by MIMO-TasNet have already preserved some of the spatial cues, but the ΔITD and ΔILD were further reduced after RTF correction. The proposed corrector and estimator reduced the ΔITD of MIMO-TasNet from 19.64 μs to 8.52 μs, and the ΔILDs in three frequency bands from 0.73, 0.66 and 0.97 dB to 0.29, 0.32 and 0.46 dB, respectively. The accuracy of spatial cue preservation was even better than MIMO SAGRNN, with ΔITD of 14.95 μs and ΔILD of 0.53, 0.45, and 0.70 dB. These results indicate the effectiveness of the RTF corrector in preserving spatial cues. The proposed RTF estimation neural network reduced ΔRTF by 5 dB over eigenvector decomposition, and its ΔITD and ΔILD were quite close to those of the oracle RTF, which shows the superior accuracy of the proposed RTF estimator.

Although the RTF corrector was not designed to reduce separation errors, ΔSNR still slightly increased by about 1.3

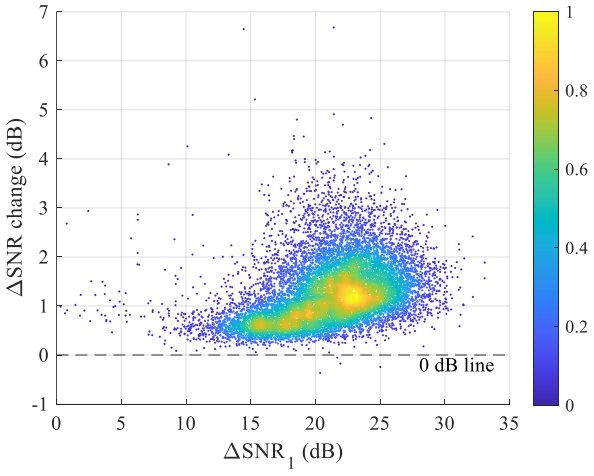

Fig. 4 Scatter plot for the ΔSNR before RTF correction and the corresponding ΔSNR change after correction. Color indicates density.

dB after correction. This confirms the deduction in Section II. Denoting the ΔSNR of the speech signal before correction as ΔSNR_1, and the ΔSNR of speech signal after correction as ΔSNR_2, the ΔSNR changes (i.e., $\Delta\text{SNR}_2 - \Delta\text{SNR}_1$) caused by RTF correction are presented in Fig. 4. The scatter plot shows that the ΔSNR changes are consistently larger than 0 dB, with almost no exception.

Notably, the proposed system achieved better ITD and ILD preservation than MIMO SAGRNN at a disadvantage of SNR. It is another example proving that good preservation of spatial cues does not entirely rely on speech separation performance.

Two variants of the proposed RTF estimator were also evaluated and presented in in Table I. One variant only used masked speech mixture to generate RTF estimation, named as "masked mix. only", and the other only used masked separated speech to generate RTF estimation, named as "masked sep. only". The input features still contained both mixture and separated speech as the vanilla estimator. Both of the two variants outperformed the MIMO-TasNet baseline and obtained better RTF estimation than eigenvector decomposition, with ΔRTF of -19.66 dB and -18.42 dB, respectively. This indicates that DPRNN can separate spatial cue components from either speech mixture or separated speech. Comparing the two variants, the "masked mixture only" condition yielded a better performance than the "masked separated only" condition, which means that the masked speech mixture contributes more to the estimation performance in the masking procedure. This deviation might result from the noise-free setup of the experiment design, in which case the unmodified spatial information can be extracted from segments of mixtures that are dominant by a single source. If the mixture is corrupted by an ambient noise, the separated speech might contribute more to the RTF estimation.

V. Conclusions

In this paper, the problem of spatial cue preservation was investigated in the context of deep learning based speech separation approaches. A framework of preserving spatial cues was proposed, which used a neural estimator to estimate RTFs, and a linear corrector to recover the distorted RTFs of the separated speech signals. Experimental results showed that the performance of spatial cue preservation of the existing binaural separation approaches could be further improved by the proposed framework. The ITD and ILD errors were significantly reduced after RTF correction, without any significant loss of SNR. The framework does not depend on any specific binaural separation network, so it is suitable to be adopted as an addon of speech separation system for spatial cue preservation. Future works could include adapting this framework to noisy and reverberate environments and developing a real-time implementation.

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] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jun. 2006.

[5] 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.

[6] 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.

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

[8] 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.

[9] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed Binaural LCMV Beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[10] 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.

[11] 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.

[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 preprint arXiv:1905.06286*, 2019.

[14] R. Gu *et al.*, "Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7319–7323.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep Learning Based Speech Beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5389–5393.

[16] 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.

[17] 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.

[18] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 7164–7175.

[19] 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.

[20] B. Xie, *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

[21] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, Oct. 2001, pp. 99–102.

[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.