A Single-Input/Binaural-Output Perceptual Rendering Based Speech Separation Method in Noisy Environments

Tianqin Zheng[†], Hanchen Pei[*], Ningning Pan[‡], Jilu Jin[†], Gongping Huang[*], Jingdong Chen[†], and Jacob Benesty[§]
[*]School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China
[†]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[‡]RIIS, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China
[§]INRS-EMT, University of Quebec, Montreal, Canada

***Abstract*—In this paper, we address the challenge of single-channel speech separation in noisy environments, where two active speakers and background noise are present in the observed signal. We propose using a dual path recursive neural network (DPRNN) to estimate the desired binaural signals from the single-channel noisy input. When the estimated binaural signal is played through headsets, listeners perceive the two speakers as originating from opposite directions, with the background noise coming from a separate direction. Additionally, the background noise is perceived to be further away from the two speakers, resulting in an improved signal-to-noise ratio (SNR). Research in psychoacoustics indicates that spatial unmasking in the perceptual domain enhances speech intelligibility in complex auditory scenes. This hypothesis is supported by both subjective and objective evaluations, including a significant 26% improvement in modified rhyme test (MRT) scores reported in this paper.**

***Index Terms*—Source separation, binaural hearing, speech enhancement, speech intelligibility.**

I. Introduction

Single-channel speech separation (SCSS), which focuses on isolating speech signals from two competing speakers using a single microphone observation, for enhancing both speech quality and intelligibility, consistently poses a challenge within the realms of source separation and speech enhancement [1], [2], [3]. Numerous approaches have been explored to address the SCSS problem, such as deep clustering methods [4], time-domain end-to-end methods [5], [6], and frequency-domain mask-based techniques [7]. With the development of more well-designed and compounded neural networks, the performance of speech separation is reported to be further improved [8], [9], [10]. However, the existence of acoustic background noise makes the SCSS problem more complicated and degrades the performance of existing methods by a large margin [11], [12]. Recently, many robust SCSS approaches have been introduced to mitigate the impact of additive noise. For example, [13] demonstrates the relative inseparability of noise and introduces a training objective inspired by SI-SDR. This approach exploits the inseparability of noise to implicitly segment the signal and reduce noise separation errors, thereby facilitating the training of more efficient separation systems using noisy oracle sources. Similarly, [14] proposes a novel network that unifies speech enhancement and separation using gradient modulation to improve noise robustness. Despite these advancements, these methods still encounter issues such as speech distortion and residual noise.

Despite these efforts and advances in solving the problem of SCSS in noisy environments, the majority of existing methods focus on producing a monaural output for each target speaker, which ignores the benefits of the human binaural auditory perception. Research on psychoacoustics has shown that (normal) human hearing is intrinsically capable of localizing sound sources and suppressing unwanted noise, known as the cocktail party effect, suggesting that properly spatializing sound sources in the perceptual domain helps in speech separation. More specifically, three typical scenarios of binaural presentations were investigated [15], [16], i.e., homophasic, heterophasic, and antiphasic. Among these methods, antiphasic presentation, where speech and noise are perceived from opposite directions, was demonstrated to offer the highest speech intelligibility [15], [17], [18], [19]. Based on the aforementioned findings, a single-input/binaural-output (SIBO) antiphasic noise reduction method [20] and a multiple-input/binaural-output (MIBO) antiphasic target speaker extraction method [21] were proposed, where significant improvement in speech intelligibility was reported compared to the corresponding monaural methods, especially in environments with low signal-to-noise ratios (SNRs) and low signal-to-interference ratios (SIRs).

In this paper, we adopt a dual-path recurrent neural network (DPRNN) structure [8], which includes an encoder, a rendering network, and a decoder, to produce a binaural output from a single-channel noisy observation. When the binaural signal is played back to listeners through headsets, it creates the perception of two competing speakers from opposite directions (e.g., the right and left sides of the head), while the background noise appears to be coming from behind. Additionally, the background noise is perceived to be further away from the two speakers, leading to an improved SNR. Simulations and experiments are conducted to evaluate the performance of the

978-616-590-477-3

proposed method. Both objective and subjective test results indicate that this method effectively improves speech quality and intelligibility.

II. Signal Model and Problem Formulation

We consider the scenario where a single microphone captures a mixture of two competing speech signals along with additional background noise. The observed signal can be formulated as

$$y(n) = s_1(n) + s_2(n) + v(n), \tag{1}$$

where $s_1(n)$ and $s_2(n)$ are the clean speech signals of the two speakers, $v(n)$ denotes the additive noise, and n is the discrete time index, which will be omitted hereinafter for readability.

This work aims to produce a binaural signal (for the left and right ears) from y utilizing a deep neural network (DNN). When the estimated binaural signal is presented to listeners through headsets, the two competing speakers and background noise will be perceived from three distinct directions/zones in the perceptual domain. Thus, the training target of the network can be formulated as

$$y_{\mathrm{L}} = \sum_{i=1}^{2} h_{\mathrm{L}}^{i} * s_i + h_{\mathrm{L}}^{v} * v, \tag{2}$$

$$y_{\mathrm{R}} = \sum_{i=1}^{2} h_{\mathrm{R}}^{i} * s_i + h_{\mathrm{R}}^{v} * v, \tag{3}$$

where $*$ denotes the linear convolution operation, h_{L}^{i}, $i = 1, 2$, and h_{L}^{v} denote the binaural room impulse responses (BRIRs) from the desired rendering locations of s_i, $i = 1, 2$, and v to the left ear, respectively, and h_{R}^{i}, $i = 1, 2$, and h_{R}^{v} denote the BRIRs from the desired locations to the right ear correspondingly. The BRIRs control the perceived locations of each source. In this work, we aim to render s_1 and s_2 to opposite directions, i.e., to the left and right side of the head, respectively, (or vice versa), while rendering v to the backside of the head.

III. Method Derivation

A. Model Architecture

As shown in Fig. 1(a), we adopt the DPRNN [8] as the backbone network to estimate the binaural signal in an end-to-end manner. The network consists of three modules: an encoder, a rendering network, and a decoder.

1) Encoder: The encoder is a 1-dimensional (1D) convolution layer with kernel size 16 and stride 8, followed by a rectified linear unit (ReLU) activation function. The input noisy speech sequence is then mapped into a latent feature space with dimension $C = 256$, thus giving a representation $\mathbf{A} \in \mathbb{R}^{C \times L}$, where L is the sequence length after the convolution operation.

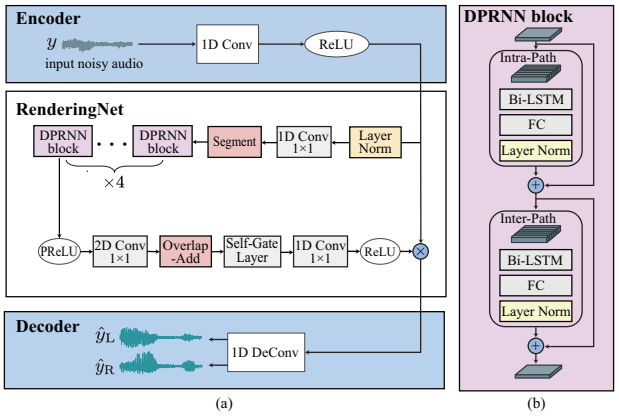

Fig. 1. Model architecture: (a) flowchart of the rendering based DPRNN-SIBO and (b) details of the DPRNN block.

2) Rendering Network: The rendering network is the essential module, which begins with a layer normalization operation to stabilize the training process. Then, it is followed by a special 1D convolution layer with a kernel size and stride of 1, denoted as a 1×1 convolution, which serves as a bottleneck to further fuse the information. The processed feature is then segmented into short frames of length $K = 250$ with a frame shift of $P = 125$. Thus, the feature dimension is changed to $\mathbb{R}^{C \times K \times T}$, where T is the number of frames.

The core part of the rendering network consists of 4 repetitions of a DPRNN block, illustrated in Fig. 1(b). *Each DPRNN block contains an intra-path and an inter-path module. Both modules include a bidirectional LSTM (Bi-LSTM) layer, a fully connected (FC) layer and the layer normalization operation. In addition, skip connections are adopted for both modules.* The 4 repeats of DPRNN block fully capture and exploit both local and long-term dependencies of the feature in and across short time frames. A PReLU function with one parameter follows and introduces a nonlinear transform. Then, a 2D convolution layer (kernel size 1) is employed to expand the feature dimension from C to $2 \times C$ to produce stacked two gain vectors related to the left and right ear outputs.

After the overlap-add operation, each gain vector is transformed back to size $\mathbb{R}^{C \times L}$, which is then passed through a self-gate module (composed of a hyperbolic tangent and a sigmoid activation function), a 1×1 1D convolution layer, and a ReLU function. Finally, the rendering network yields the binaural transfer functions, which are then elementwise multiplied in with the noisy representation $\mathbf{A}$ to produce estimates of the binaural signal in latent space.

3) Decoder: The decoder is a 1D deconvolution layer to reconstruct binaural estimates, i.e., $\hat{y}_{\mathrm{L}}$ and $\hat{y}_{\mathrm{R}}$, from the latent-space binaural estimates, which is a reverse process of the encoder.

B. Rendering Permutation Invariant Training Objective

The loss function is formulated using the scale-invariant SNR (SI-SNR) [22]. Specifically, for the left and right ear channels, the loss function is defined as follows:

$$\mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}}\right) = 10\log_{10}\frac{E\left(|y_{\mathrm{L}} - \hat{y}_{\mathrm{L}}|^2\right)}{E\left(|y_{\mathrm{L}}|^2\right)}, \tag{4}$$

$$\mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}}\right) = 10\log_{10}\frac{E\left(|y_{\mathrm{R}} - \hat{y}_{\mathrm{R}}|^2\right)}{E\left(|y_{\mathrm{R}}|^2\right)}, \tag{5}$$

where $E(\cdot)$ denotes mathematical expectation. Considering the binaural output, the loss function should be

$$\mathcal{L}\left(\hat{y}_{\mathrm{L}},\ \hat{y}_{\mathrm{R}}\right) = \mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}}\right) + \mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}}\right). \tag{6}$$

Note that rendering s_1 and s_2 to the left and right sides, respectively, or vice versa, are both acceptable for training. Thus, there are two possible permutations. To handle this, we exploit a direction permutation invariant loss during training:

$$\mathcal{L} = \min\left\{\mathcal{L}(\hat{y}_{\mathrm{L}_1},\ \hat{y}_{\mathrm{R}_1}),\ \mathcal{L}(\hat{y}_{\mathrm{L}_2},\ \hat{y}_{\mathrm{R}_2})\right\}, \tag{7}$$

where $\mathcal{L}(\hat{y}_{\mathrm{L}_1},\ \hat{y}_{\mathrm{R}_1}) = \mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}_1}\right) + \mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}_1}\right)$ denotes the loss when rendering s_1 and s_2 to left and right sides, respectively. In contrast, $\mathcal{L}(\hat{y}_{\mathrm{L}_2},\ \hat{y}_{\mathrm{R}_2})$ denotes the permutation case, i.e., the right and left side, respectively.

IV. Simulations

A. Training Setup

1) Dataset: We employ the WHAM! dataset [11] for both training and development purposes, which is constructed by incorporating noise from the WHAM! noise dataset [11] into the standard Wall Street Journal (WSJ0) two-speaker mixture dataset [4]. During the process of blending the two speech signals, the longer signal is adjusted to the length of the shorter one by truncating it in a 'minimum' mode, with a sampling rate set at 8 kHz. Our objective is to spatially position the two speech signals to the left and right sides, each at a distance of 1 meter from the listener's head. Concurrently, the additional noise is spatially rendered to the rear of the head, at varying distances of 1 meter, 2 meters, and 4 meters.

The location of each source is controlled by the corresponding BRIRs taken from the dataset [23]. The training targets are then generated by (2) and (3).

2) Training Configuration: For comparison, we exploit the original DPRNN (referred to as DPRNN-base in the following) with the same network structure as the baseline model. Unlike the proposed model (named as DPRNN-SIBO hereinafter), DPRNN-base achieves source separation by directly estimating the waveform of s_1 and s_2.

To train the models, we use the adaptive moment (Adam) estimation optimizer [24]. The initial learning rate is set to 10^{-3}. Given the case when training loss does not decrease in 3 consecutive epochs, the learning rate would accordingly half its value.

B. Simulation Results

1) Directional Perception Validation: To assess the effectiveness of the DPRNN-SIBO's rendering outcomes, we have synthesized 20 distinct test mixtures derived from the WHAM! dataset. It is important to note that these test signals were not present in either the training or development datasets. Additionally, we have established an input SNR of 0 dB, which is calculated as the ratio of the power of the mixed speech signals to that of the accompanying noise.

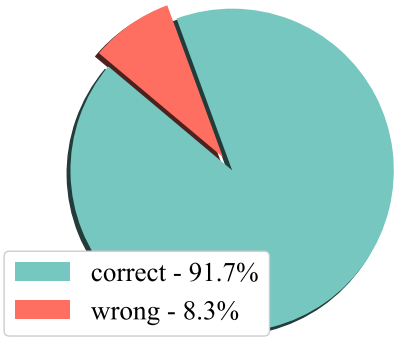

Fig. 2. Percentage of correct answers in the direction perception test. Correct answers include either "speaker 1 from left, speaker 2 from right, noise from back" or "speaker 2 from left, speaker 1 from right, noise from back." Any other choices are considered incorrect.

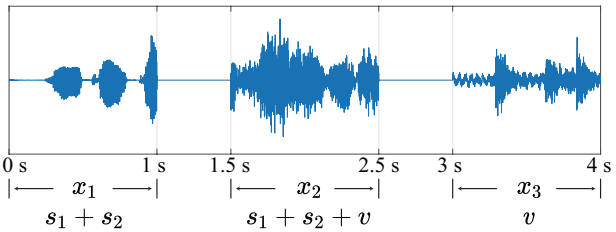

Fig. 3. Schematic diagram of observed signal construction, composed of five segments: x_1, x_2, and x_3 are the mixed signal, observed signal, and noise segment, respectively, with two silent segments of 0.5 seconds in the middle.

We invited five participants without hearing loss to listen to the binaural outputs of DPRNN-SIBO and listeners were required to select the direction of two speakers and noise. They are instructed to choose one from the following options for each source: 1) left, 2) right, and 3) back. Since we train DPRNN-SIBO with direction permutation invariant loss, both choices, i.e., "speaker 1 from left, speaker 2 from right, noise from back" and "speaker 2 from left, speaker 1 from right, noise from back" are considered to be correct. The findings are depicted in Fig. 2. It is evident that the direction of speakers and noise were properly chosen by most participants, leading to a 91.7% accuracy rate, which proves that DPRNN-SIBO can render sources to desired directions.

2) Impact of Rendering Distance: To evaluate the impact of rendering distance of the additive noise, we constructed 20 sets of observation signals as illustrated in Fig. 3. Both speech signals and noise are the same as those used in Section IV-B1. As shown in Fig. 3, the observed signal is 4-second long, consisting of five segments: a 1-second segment of 2-mixed speech signal, denoted as $x_1 = s_1 + s_2$, 0.5-second of silence, another 1-second of 2-mixed speech and noise, denoted as $x_2 = s_1 + s_2 + v$, another 0.5-second of silence, and 1-second

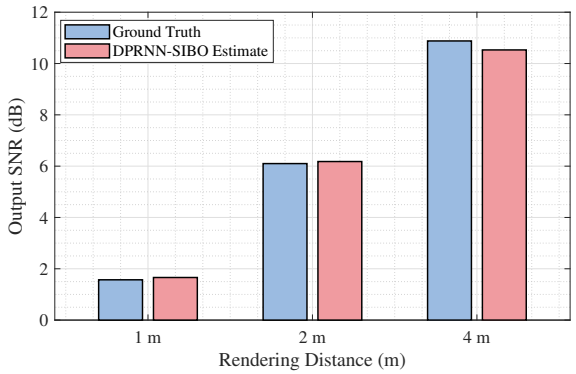

Fig. 4. Output SNR of the binaural signals estimated by DPRNN-SIBO and the ground truth SNR (dB).

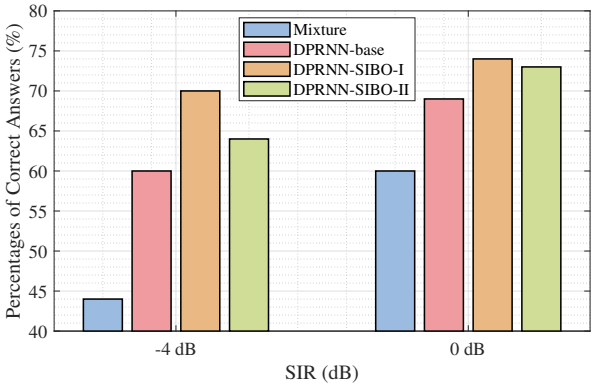

Fig. 5. Percentages of correct answers for input mixture, estimation of DPRNN-base, DPRNN-SIBO-I, and DPRNN-SIBO-II in the MRT test.

of additive noise, denoted as $x_3 = v$. Note that segment x_1, x_3 is utilized to calculate SNR and segment x_2 is used to make the observed signal similar to that in the training set. The input SNR is set to be 0 dB.

With the proposed DPRNN-SIBO, we can obtain the left ear and right ear outputs for x_1 and x_3, which are denoted as $\{\hat{x}_{\mathrm{L},1}, \hat{x}_{\mathrm{R},1}\}$, and $\{\hat{x}_{\mathrm{L},3}, \hat{x}_{\mathrm{R},3}\}$, respectively. Then the output SNRs of left and right ears are defined as follows:

$$\mathrm{SNR_L} = 10\log_{10}\frac{E\left(|\hat{x}_{\mathrm{L},1}|^2\right)}{E\left(|\hat{x}_{\mathrm{L},3}|^2\right)}, \tag{8}$$

$$\mathrm{SNR_R} = 10\log_{10}\frac{E\left(|\hat{x}_{\mathrm{R},1}|^2\right)}{E\left(|\hat{x}_{\mathrm{R},3}|^2\right)}. \tag{9}$$

Here, we define the binaural SNR (biSNR) as follows:

$$\mathrm{biSNR} = \frac{1}{2}(\mathrm{SNR_L} + \mathrm{SNR_R}). \tag{10}$$

Figure 4 lists the biSNRs achieved by DPRNN-SIBO and the ground-truth biSNRs, which are computed using (10) with the ground-truth signals formulated by (2) and (3). It can be shown that the biSNR of the output of DPRNN-SIBO closely approximates the ground-truth biSNR, proving that DPRNN-SIBO can render the additive noise to the desired distance.

3) MRT Tests: To evaluate the intelligibility improvement achieved by DPRNN-SIBO, modified rhyme tests (MRTs) [25] were conducted. We selected 48 carrier utterances from MRT database in the same way as our previous work [20], [21]. Then, the MRT speech signals were mixed with speech signals randomly selected from WSJ0 test dataset and additive noise from WHAM! noise dataset. The input SIRs were set to -4 dB and 0 dB and SNRs were randomly selected from -6 dB to 3 dB. DPRNN-SIBO and DPRNN-base are adopted to render or separate the speech. We consider two configurations of DPRNN-SIBO, denoted as DPRNN-SIBO-I and DPRNN-SIBO-II, both of which were designed to render the 2 speech signals to 1 m away from the listener to the left and right direction, while rendering the additive noise to 1 m and 4 m away from the listener to the backward direction, respectively. To evaluate speech intelligibility, participants were required to choose the word they hear in the MRT carrier sentence "please select the word –" from all the testing signals. More correct answers provided by the participants indicate higher speech intelligibility achieved by the tested method.

Figure 5 shows the percentage of correct answers of MRT from DPRNN-base and the proposed two methods, DPRNN-SIBO-I and DPRNN-SIBO-II. We can see that all tested methods are capable of increasing speech intelligibility by a large margin as compared to the MRT results of the mixed signal. Among them, DPRNN-base increases the accuracy by 9% and 16% in the 0 dB and -4 dB experiments while DPRNN-SIBO-I increases the accuracy by 14% and 26%, respectively. Also, both configurations of the proposed method outperform DPRNN-base in both input SIR scenarios. Specifically, DPRNN-SIBO-I achieves 10% increase than DPRNN-base in -4 dB, and 5% increase in 0 dB. Moreover, we note that DPRNN-SIBO-I consistently performs better than DPRNN-SIBO-II. This can be attributed to the fact that DPRNN-SIBO-II pushes the noise source to a distance of 4 m for more noise suppression, which results in more distortion of the speech signal. This also explains the lower percentage of correct answers for DPRNN-base.

V. Conclusions

This paper presented a deep learning-based single-input/binaural-output rendering method to improve speech intelligibility in complex acoustic environments, where speech signals from two competing speakers and background noise are captured by a single microphone. A rendering network based on DPRNN was trained to take the single-channel mixture as input and generate a binaural signal. When delivered through headphones, this binaural signal allows listeners to perceive each speaker from opposite directions while placing the noise further away in the back direction. The proposed rendering method effectively reduces interference from competing speakers and improves speech intelligibility, especially in low-SIR conditions. MRT results show a notable 26% improvement over the mixed speech signal, demonstrating the effectiveness of the proposed method.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer Science & Business Media, 2006.

[2] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, Oct. 2018.

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

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

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, Aug. 2019.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[8] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, pp. 46–50, IEEE, 2020.

[9] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[10] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE ICASSP*, pp. 21–25, IEEE, 2021.

[11] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[12] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. IEEE ICASSP*, pp. 1–5, IEEE, 2023.

[13] M. Maciejewski, J. Shi, S. Watanabe, and S. Khudanpur, "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," in *Proc. IEEE ICASSP*, pp. 5774–5778, 2021.

[14] Y. Hu, C. Chen, H. Zou, X. Zhong, and E. S. Chng, "Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation," in *Proc. IEEE ICASSP*, pp. 1–5, 2023.

[15] J. Licklider, "The influence of interaural phase relations upon the masking of speech by white noise," *J. Acoust. Soc. Am.*, vol. 20, pp. 150–159, March. 1948.

[16] L. A. Jeffress, "A place theory of sound localization.," *J. Comp. Physiol. Psychol.*, vol. 41, p. 35, Feb. 1948.

[17] Y. Wang, J. Chen, J. Benesty, J. Jin, and G. Huang, "Binaural heterophasic superdirective beamforming," *Sensors*, vol. 21, no. 1, p. 74, 2020.

[18] J. Jin, J. Chen, J. Benesty, Y. Wang, and G. Huang, "Heterophasic binaural differential beamforming for speech intelligibility improvement," *IEEE Trans. Veh. Technol.*, vol. 69, pp. 13497–13509, Nov. 2020.

[19] J. Benesty, G. Huang, J. Chen, and N. Pan, *Microphone Arrays*, vol. 22. Berlin, Germany: Springer-Verlag, 2023.

[20] N. Pan, Y. Wang, J. Chen, and J. Benesty, "A single-input/binaural-output antiphasic speech enhancement method for speech intelligibility improvement," *IEEE Signal Process. Lett.*, vol. 28, pp. 1445–1449, 2021.

[21] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple-input/binaural-output antiphasic speaker signal extraction," in *Proc. IEEE ICASSP*, pp. 1–5, IEEE, 2023.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE ICASSP*, pp. 626–630, IEEE, 2019.

[23] B. I. Bacila and H. Lee, "360 binaural room impulse response (BRIR) database for 6dof spatial perception research," in *Audio Engineering Society Convention 146*, Audio Engineering Society, 2019.

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

[25] S. ANSI, "3.2, methods for measuring the intelligibility of speech over communication systems," *American National Standards Institute*, pp. 3–2, 1989.