ON MULTIPLE-INPUT/BINAURAL-OUTPUT ANTIPHASIC SPEAKER SIGNAL EXTRACTION

Xianrui Wang[1], *Ningning Pan*[1], *Jacob Benesty*[2], *and Jingdong Chen*[1]

[1]CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

ABSTRACT

This paper studies the problem of target speaker signal exaction and antiphasic rendering with an array of microphones in the scenarios where there are two active speakers. Based on the important findings achieved in the psychoacoustic field as well as our recent works on single-channel speech enhancement, we present a rendering based approach in which a temporal convolutional network (TCN) is trained to take the multiple signals observed by the microphone array as its inputs and generate two output (binaural) signals. The TCN is trained in such a way that, when binaural output signals are listened by the listener with headsets, the speech signal from the desired speaker is perceived on one side of and close to the listener's head, while the competing speech signal is perceived on the opposite side and also away from the listener's head. Benefited from rendering and the signal-to-interference ratio (SIR) improvement, this antiphasic binaural presentation enables the listener to better focus on the target speaker's signal while ignoring the impact of the competing speech. The modified rhyme tests (MRTs) are performed to validate the superiority of the proposed method.

Index Terms— Antiphasic presentation, speaker extraction, modified rhyme test, multiple-input/binaural-output.

1. INTRODUCTION

In real-world applications, a speech signal of interest is inevitably contaminated by interference, reverberation, and noise, which impair speech quality and/or speech intelligibility [1], thereby degrading the performance of systems such as speech coding, speech communication, and automatic speech recognition (ASR). Therefore, source extraction, which aims at extracting the target signal from its corrupted observations, has attracted a significant amount of attention [2, 3] and many different methods have been developed over the last few decades such as beamforming [4–8], statistical techniques [9–14], and deep learning based approaches [15–19]. Although they have demonstrated promising performance, most existing methods still suffer from a number of limitations. Not only do they degrade significantly in performance when the signal-to-interference ratio (SIR) is low, but they generate only a monaural estimate of the target speech signal, which obviously do not take advantage of the human binaural perception system.

Research in psychoacoustics has shown that speech intelligibility can be significantly improved by using binaural presentation in comparison with the monaural presentation [21–23]. There are three kinds of binaural presentations depending on the perceptual regions of the target signal and the interference, i.e., homophasic, heterophasic, and antiphasic [24, 25]. The antiphasic presentation in which the target and interference signals are perceived at the opposite directions has the highest speech intelligibility, which is followed by the heterophasic presentation. The homophasic case where both the target and interference signals are perceived at the middle of the listener's head correpsonds to the lowest intelligibility [24, 25].

Motivated by the findings in the psychoacoustic field, we developed recently an antiphasic rendering approach to single-channel speech enhancement [23]. This work is an extension of the work in [23] and we present a multiple-input/binaural-output (MIBO) antiphasic speaker signal extraction method, where a temporal convolutional network (TCN) is adopted to achieve antiphasic presentation. Thus, the proposed approach is referred to as TCN-MIBO. The TCN-MIBO network renders the target signal to one side to the listener's head and the interference signal to the other side, and meanwhile the target signal is rendered close to the listener's head while the interference is rendered away from the listener's head to improve SIR. Benefiting from the antiphasic presentation and SIR improvement, the listener is able to better focus on the target signal, leading to intelligibility improvement. Experiments are carried out to demonstrate the superiority of the proposed method.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We consider the acoustic scenario where two competing speakers coexist. The signals observed by a microphone array can be expressed as

$$y_m(t) = h_{\text{tar},m}(t) * s_{\text{tar}}(t) + h_{\text{int},m}(t) * s_{\text{int}}(t), \tag{1}$$

where t is the time index, $m \in \{1, 2, \ldots, M\}$ is the microphone index, $s_{\text{tar}}(t)$ and $s_{\text{int}}(t)$ denote, respectively, the target and interference signals, $h_{\text{tar},m}(t)$ and $h_{\text{int},m}(t)$ are the room impulse responses (RIRs) from the target source and interference to the mth microphone, respectively, and $*$ denotes the linear convolution operation. Note that we neglect the noise term in the signal model to better illustrate the principle. The additive noise can be treated in a similar way to [23].

Unlike most existing target speaker signal extraction methods, which generate only a monaural estimate of the target signal, i.e., $\hat{s}_{\text{tar}}(t)$, the presented method generates two (binaural) output signals: $x_{\text{L}}(t)$ and $x_{\text{R}}(t)$, for the left and right ears, respectively. The training target of the network are, therefore, $x_{\text{L}}(t)$ and $x_{\text{R}}(t)$, which are constructed during the training process as follows:

$$x_{\text{L}}(t) = h_{\text{tar,L}}(t) * s_{\text{tar}}(t) + h_{\text{int,L}}(t) * s_{\text{int}}(t), \tag{2}$$

$$x_{\text{R}}(t) = h_{\text{tar,R}}(t) * s_{\text{tar}}(t) + h_{\text{int,R}}(t) * s_{\text{int}}(t), \tag{3}$$

where $h_{\text{tar,L}}(t)$ and $h_{\text{tar,R}}(t)$ are binaural RIRs (BRIRs) from the designed location of the rendered target speech signal to the left and right ears, and $h_{\text{int,L}}(t)$ and $h_{\text{int,R}}(t)$ are the BRIRs from the designed location of the rendered interference signal to the left and

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102200 and in part by the Key Program of National Science Foundation of China (NSFC) under Grant No. 61831019 and 62192713.

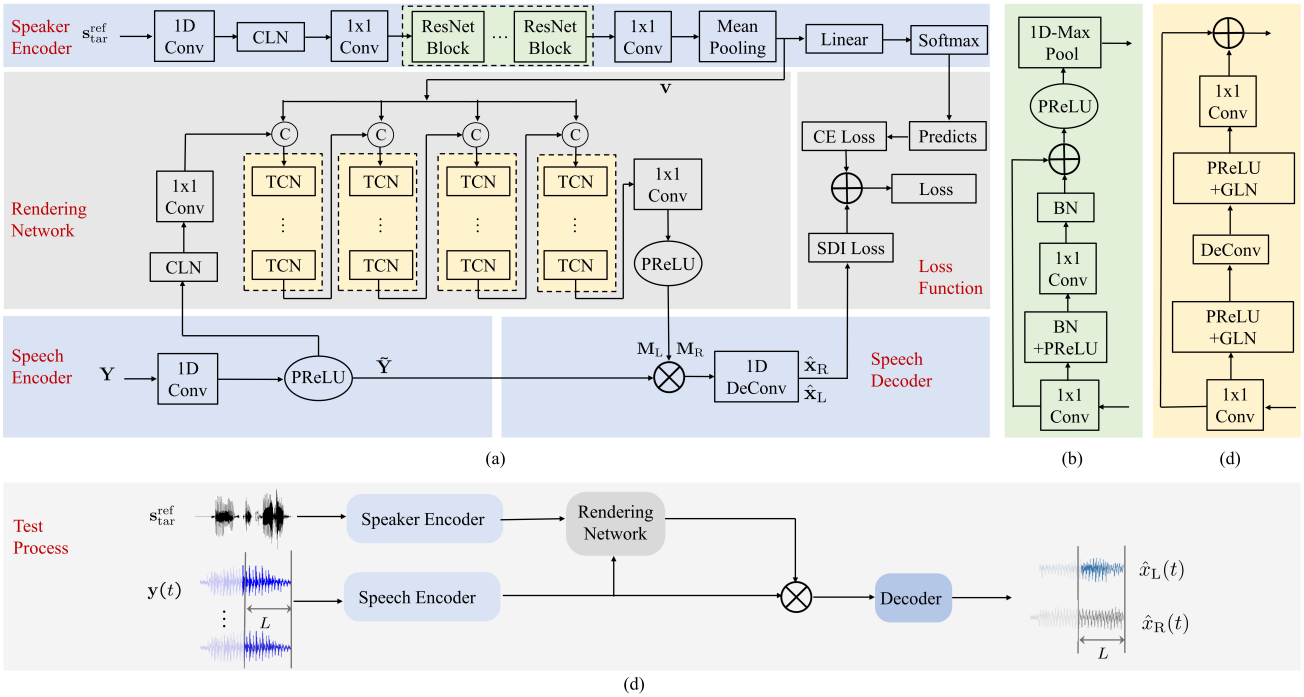

Fig. 1. (a) Diagram of the proposed rendering network."CLN" stands for channel-wise layer normalization. Symbol "C" represents concatenation. There are four TCN stacks in the render part, each consists of 8 TCN blocks. The first TCN block of each TCN stack takes the speaker embedding as an additional input. The ResNet stack consists of three ResNet blocks. CE and SDI denote the cross entropy and signal to distortion ratio, respectively. Symbols "x" and "+" denote element-wise multiplication and addition. (b) Structure of the ResNet block, where "BN" is short for batch normalization. (c) Structure of the TCN block, where "GLN" is global layer normalization and "DeConv" means dilated depth-wise convolution. (d) Example of test process, where L is the kernel size of the *1D-Conv* in speech encoder.

right ears, respectively. The perceived directions/zone of the target speech and interference signals in the binaural outputs are controlled by the corresponding BRIRs, which are selected from a BRIR database [26]. In this paper, the target signal will be rendered to the left-hand side to the listener's head while the interference will be rendered to the right-hand side (note that they can also be rendered the other way around). To distinguish the target speaker and the interference signals, a reference signal $s_{\text{tar}}^{\text{ref}}(t)$, which is a registered speech signal from the target speaker, is needed to supervise the rendering.

3. TCN-MIBO NETWORK

The structure of the TCN-MIBO rendering network is shown in Fig. 1(a). Similar to the state-of-the-art extraction network Spex+ [18], the proposed network consists of four parts: a speaker encoder, which is used to extract the speaker embedding from the reference signal; a speech encoder, which transforms the multichannel observation signals into its latent representation; a rendering network, which generates binaural representations in the latent space; and a speech decoder, which transforms the latent binaural representations back to the time domain signals.

3.1. Speaker Encoder

For the speaker encoder, we adopt the same structure as that in Spex+ [17, 18]. A one dimensional convolution layer (*1D-Conv*) first transforms the reference signal with length T_0 (the length of the longest reference signal within the batch during training and can be any value in test process), i.e., $\mathbf{s}_{\text{tar}}^{\text{ref}} = \left[s_{\text{tar}}^{\text{ref}}(1) \quad \cdots \quad s_{\text{tar}}^{\text{ref}}(t) \quad \cdots \quad s_{\text{tar}}^{\text{ref}}(T_0)\right]^T$, into its latent representation. The number of filters of this layer is 256, the kernel size L is 20, and the stride is 10. After a channel-wise layer normalization (CLN), three residual blocks *ResNet*, as shown in Fig. 1(b), and the last *1D-Conv* with a mean pooling operation, the speaker embedding $\mathbf{v} \in \mathbb{R}^{d_0 \times 1}$ is extracted. We set the dimension of the speaker embedding, i.e., d_0, to 256.

The *ResNet* block consists of two 1×1 convolution layer (*1x1-Conv*, with the kernel size and the stride being 1), each followed by a batch normalization (BN) and a parametric rectified linear unit (PReLU). A skip connection is used to add the input to the output of the second BN. All parameters are set to the same as those in Spex+ [18].

3.2. Speech Encoder

The input of the encoder is the multichannel observation signals with length of T_1 (set to 4 seconds during training and can be any value in test process), $\mathbf{Y} = [\mathbf{y}(1) \quad \cdots \quad \mathbf{y}(t) \quad \cdots \quad \mathbf{y}(T_1)] \in \mathbb{R}^{M \times T_1}$, where $\mathbf{y}(t) = [y_1(t) \quad \cdots \quad y_m(t) \quad \cdots \quad y_M(t)]^T$ is the microphone array observation signal vector at time index t. The speech encoder is a *1D-Conv*, followed by a PReLU. The kernel size and the stride of the *1D-Conv* are set to 20 and 10, respectively. The number of input channels is equal to the number of microphones, i.e., M, and the output dimension d_1 is set to 256. Thus, the output of the encoder is a latent representation of the observation signals, $\tilde{\mathbf{Y}} \in \mathbb{R}^{d_1 \times T_1'}$, where T_1' is the length of the sequence after convolution.

3.3. Rendering Network

The rendering network begins with a *1x1-Conv*, whose output dimension is set to be the same as the input dimension d_1. The rendering network consists of 4 TCN stacks. There are 8 TCN blocks in each stack, which is shown in Fig. 1(c). The configurations of TCN blocks are the same as those in Spex+ [18]. The rendering

network takes the speaker embedding $\mathbf{v}$ together with the latent mixture representation $\tilde{\mathbf{Y}}$ as input to generate binaural masks $\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\mathrm{L}} \\ \mathbf{M}_{\mathrm{R}} \end{bmatrix} \in \mathbb{R}^{2d_1 \times T_1'}$, which can be regarded as the left-ear and right-ear transfer functions in the latent space. The binaural representations in the latent space are calculated as $\mathbf{X}_{\mathrm{L}} = \mathbf{M}_{\mathrm{L}} \otimes \tilde{\mathbf{Y}} \in \mathbb{R}^{d_1 \times T_1'}$ and $\mathbf{X}_{\mathrm{R}} = \mathbf{M}_{\mathrm{R}} \otimes \tilde{\mathbf{Y}} \in \mathbb{R}^{d_1 \times T_1'}$, where $\otimes$ denotes the element-wise multiplication.

3.4. Decoder

The decoder transforms the latent binaural representations, $\mathbf{X}_{\mathrm{L}}$ and $\mathbf{X}_{\mathrm{R}}$, back to the time domain, thereby reconstructing the waveforms $\hat{\mathbf{x}}_L = [\hat{x}_{\mathrm{L}}(1) \quad \cdots \quad \hat{x}_{\mathrm{L}}(t) \quad \cdots \quad \hat{x}_{\mathrm{L}}(T_1)]^T$ and $\hat{\mathbf{x}}_R = [\hat{x}_{\mathrm{R}}(1) \quad \cdots \quad \hat{x}_{\mathrm{R}}(t) \quad \cdots \quad \hat{x}_{\mathrm{R}}(T_1)]^T$ using a deconvolution operation.

3.5. Training Objective

Jointly training the speaker classification and the rendering network is a multi-task learning problem. For speaker classification, we adopt the cross entropy (CE) [27] as the loss function, which is defined as

$$\mathcal{J}_{\mathrm{CE}} = -\sum_{n_{\mathrm{s}}=1}^{N_{\mathrm{s}}} P_{n_{\mathrm{s}}} \log \hat{P}_{n_{\mathrm{s}}}, \tag{4}$$

where $P_{n_{\mathrm{s}}} = \{0 \text{ or } 1\}$ is the ground-truth label of the n_{s}th speaker, N_{s} is the number of speakers in the training set, and $\hat{P}_{n_{\mathrm{s}}}$ is the predicted probability.

For the rendering network, we choose the signal-to-distortion index (SDI) [28] as the cost function, which is defined as

$$\mathcal{J}_{\mathrm{SDI},i} = 10\log_{10}\left\{\frac{E\left[x_i(t) - \hat{x}_i(t)\right]^2}{E\left[x_i^2(t)\right]}\right\}, \; i = \mathrm{L,R}, \tag{5}$$

where $\mathcal{J}_{\mathrm{SDI},L}$ and $\mathcal{J}_{\mathrm{SDI},R}$ denote, respectively, the SDIs of signals estimated for the left and right ears. We use $\mathcal{J}_{\mathrm{SDI}}$, the average of $\mathcal{J}_{\mathrm{SDI,L}}$ and $\mathcal{J}_{\mathrm{SDI,R}}$, as the extraction loss function. $\mathcal{J}_{\mathrm{SDI}}$ and $\mathcal{J}_{\mathrm{CE}}$ are combined to optimize the proposed framework. So, the overall cost function is

$$\mathcal{J} = \mathcal{J}_{\mathrm{SDI}} + \lambda \mathcal{J}_{\mathrm{CE}}, \tag{6}$$

where λ is a parameter, which is set empirically to 10 in this work.

4. SIMULATIONS

4.1. Setup

4.1.1. Training Data

The training and development sets are constructed as follows. The target speech, reference and interference signals are taken from the Wall Street Journal (WSJ0) database [29] with the same configuration of the WSJ0-2mix-extr database [17]. All signals are resampled to 8 kHz. We align the length of the target signal and the interference in a "max" way, i.e, the shorter one is padded with zeros at the end to have the same length as the longer one. Then, a room of size $L_x \times L_y \times L_z$ is considered, where the length L_x, the width L_y , and the height L_z are generated with a uniform distribution, respectively, in the range of $[8\,\mathrm{m}, 10\,\mathrm{m}]$, $[6\,\mathrm{m}, 8\,\mathrm{m}]$, and $[3\,\mathrm{m}, 4\,\mathrm{m}]$. For ease of exposition, we use the 3-dimensional (3D) Cartesian coordinate system to specify the positions and the one corner on the floor is used as the origin of the coordinate system. A uniform linear microphone array consisting of 6 sensors with an element spacing of 5 cm was placed horizontally with the array center at $\left(\frac{L_x}{2}\,\mathrm{m}, 1\,\mathrm{m}, 1.5\,\mathrm{m}\right)$. The positions of both the target speech source and the interference source are randomly generated where the $x-$, $y-$, and $z-$coordinates are all generated with a uniform distribution with ranges being, respectively, $[1\,\mathrm{m}, (L_x - 1)\,\mathrm{m}]$, $[1\,\mathrm{m}, (L_y - 1)\,\mathrm{m}]$, and $[1\,\mathrm{m}, 2\,\mathrm{m}]$. The RIRs are generated with the python package gpuRIR [30], which is based on the image model [31]. The reverberation time T_{60} is set to the range of $[180\,\mathrm{ms}, 200\,\mathrm{ms}]$. The microphone array observation signals are then obtained by convolving the clean speech signal and the interference signal with the corresponding RIRs as in (1). The SIR is controlled to be in the range of $[-5\,\mathrm{dB}, 5\,\mathrm{dB}]$. The training targets, i.e., the binaural signals, were generated by convolving the clean speech signal and the interference signals with the desired BRIRs selected from [26] according to (2) and (3).

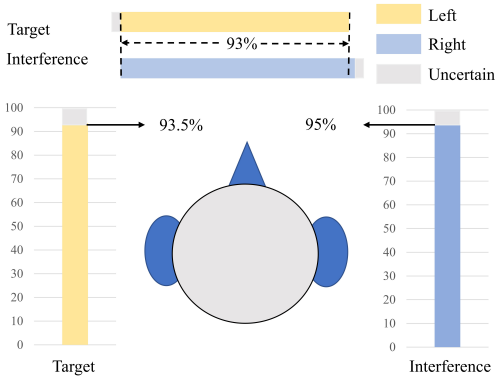

Fig. 2. Percentage of direction perception results.

4.1.2. Training Configuration

For the baseline, we considered a network, which has a similar network structure as TCN-MIBO. The only difference is that the output of the baseline is a monaural estimation of the target speech signal. Note that the baseline can be regarded as an extension of Spex+ [18] from single-channel input to multichannel input. In the rest of the paper, the baseline is denoted as TCN-multiple input/single output (TCN-MISO).

To train TCN-MISO and TCN-MIBO, adaptive moment estimation (Adam) [32] is used as the optimizer. The initial learning rate is set to 10^{-3}, and it is reduced to half if the training loss on the development set does not decrease within 3 consecutive epochs. The training process stops if the loss on the development set does not decrease for successive 20 epochs.

4.2. Experiments

In all experiments, RIRs are generated in the same way as in the training set.

4.2.1. Direction Perception

To evaluate the rendering ability of TCN-MIBO, we generated 40 mixtures with WSJ0-2mix-extr test set, which are not seen in the training set. The SIR is set to 0 dB. The target speech signal is rendered to the left-hand side with 1 m away from the middle of the listen's head, while the interference signal is 1 m away on the right-hand side. Five normal hearing listeners were asked to listen to the binaural outputs and judge the directions of the target speech and the interference signals. Only when their judgements on the directions of both the target speech and the interference sources are correct, the result is deemed to be correct. The direction perception results are shown in Fig. 2. One can see that 93% of the listeners' choices are correct for both signals, which validates the rendering ability of

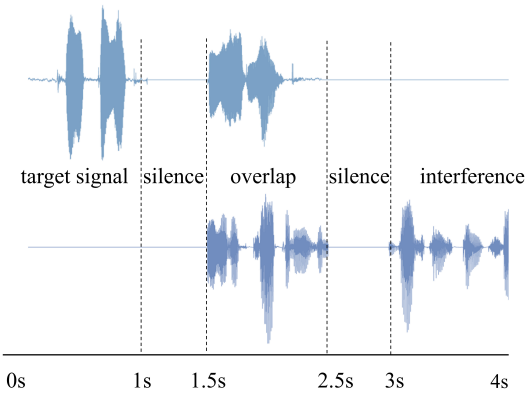

Fig. 3. Illustration of the signal construction.

Table 1. The output SIR of the estimated binaural signals by TCN-MIBO and the ground truth SIR (dB).

	1 m	2 m	4 m
Ground truth	-1.06	5.01	10.8
TCN-MIBO	-1.00	5.12	11.0

TCN-MIBO. Analyzing the signals marked as "uncertain" by most of the listeners, we found that the two active speakers in the mixed signals are difficult to distinguish, which causes the failure of the speaker encoder. This problem could be solved by including more speakers in the training set, which is however beyond the scope of this paper.

4.2.2. *Impact of Rendering Distance*

In this set of experiments, we evaluate the impact of the rendering distance on the performance of TCN-MIBO by measuring the output SIRs in the binaural presentation. 20 observation signals were constructed in a way as illustrated in Fig. 3. Both the target speech and interference signals are selected from the WSJ0-2mix-extr test set and they were arranged to be 4-second long. The 4-second target speech signal is a concatenation of the following 4 segments: a 1-second segment of speech signal, 0.5-second segment of silence, another 1-second segment of speech signal, and 1.5-second segment of silence, while the interference speech signal is concatenated in the opposite order. The SIR of non-overlapped parts was set to 0 dB. The multichannel observation signals are then obtained according to (1).

The output SIR of the estimated binaural signals is computed as

$$\text{biSIR} = 10\log_{10}\left\{\frac{E\left[\tilde{x}_{\text{L}}^2(t)\right]}{E\left[\tilde{x}_{\text{R}}^2(t)\right]}\right\}, \tag{7}$$

where $\tilde{x}_{\text{L}}(t)$ is the first 1-second segment of the signal estimated for the left ear and $\tilde{x}_{\text{R}}(t)$ is the last 1-second segment of the signal estimated for the right ear.

Table 1 lists the biSIR achieved by TCN-MIBO rendering network and the ground truth SIR, which is computed with (7) by constructing the ground truth signals for the left and right ears as in (2). As seen, the biSIR of the estimated binaural signals by TCN-MIBO is close to the ground truth, which validates its capability in adjusting distance. By rendering the interference further away, the SIR is significantly improved.

4.2.3. *Listening Tests*

Finally, we evaluate the intelligibility improvement achieved by TCN-MIBO through listening tests. We adopted the modified

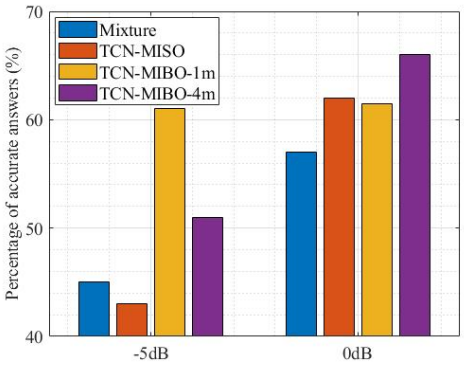

Fig. 4. Percentages of correct answers. Simulation conditions: $T_{60} \in [180\text{ ms}, 200\text{ ms}]$, SIR=$-5$ dB, 0 dB.

rhyme test (MRT), which is a standardized listening test for speech intelligibility measurement [33]. In the MRT database, there are 50 sets of rhyming keywords (each consists of 6 keywords), e.g., [rent, tent, dent, bent, sent, went]. Every keyword is presented in a carrier sentence, e.g., "please select the word went." In total, 50×6 carrier sentences were recorded by each of the 4 female and 5 male native English speakers. We arbitrarily selected 48 utterances from four male speakers and four female speakers with 6 utterances from each speaker: one sentence is used as the reference signal and the other five sentences are used as the target signals. We randomly picked up forty sentences from the WSJ0-2mix-extr test set [17] and used them as interference signals. The input SIRs were set to -5 dB and 0 dB. We considered two configuration for TCN-MIBO: the target signal was rendered to 1 m away from the middle of the listen's head on the left-hand side, while the interference was rendered to 1 m and 4 m away but on the right-hand side. We refer them as TCN-MIBO-1m and TCN-MIBO-4m, respectively. Five normal hearing listeners were asked to select the keyword they perceive.

The percentage of correct answers of the mixed signals and signals estimated by TCN-MISO, TCN-MIBO-1m, and TCN-MIBO-4m are depicted in Fig. 4. One can see that when the input SIR $=$ 0 dB, all methods can improve speech intelligibility and TCN-MIBO-4m achieves highest speech intelligibility, which is benefited from rendering as well as SIR improvement. Meanwhile, TCN-MIBO-1m, which only benefited from rendering, achieves the similar number of correct answers as compared to TCN-MISO. In the case where input SIR $=-5$ dB, TCN-MISO fails in improving speech intelligibility due to target speech distortion and TCN-MIBO-1m significantly outperforms TCN-MISO and TCN-MIBO-4m, indicating that rendering is more helpful in improving intelligibility particularly when SIR is low.

5. CONCLUSIONS

This paper presented a deep learning based multiple-input/binaural-output antiphasic rendering method, in which a TCN was trained to take the multichannel signals observed at a microphone array as its inputs and generate binaural output signals. The TCN was trained in such a way that, when binaural output signals are listened by the listener with headsets, the speech signal from the target speaker is perceived on one side of and close to the listener's head, while the competing speech signal is perceived on the opposite side and also away from the listener's head. Thanks to the antiphasic rendering and the SIR improvement, the deep learning based antiphasic binaural presentation enables the listener to better focus on the target speaker's signal while ignoring the impact of the competing speech. The MRT results validated the feasibility and superiority of the proposed method.

6. REFERENCES

[1] J. Benesty, M. M. Sondhi, Y. Huang, *et al.*, *Springer handbook of speech processing*. Springer, 2008.

[2] S. Makino, *Audio source separation*. Springer, 2018.

[3] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[4] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Hoboken, NJ: Wiley, 2018.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Processing*. Berlin: Springer-Verlag, 2008.

[6] M. Branstein and D. B. Ward, eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Berlin: Springer, 2001.

[7] W. Dong, W. Fanggang, L. Xiran, Y. Pu, and J. Dajie, "Orthogonal time frequency space modulation in multiple-antenna systems," *ZTE Communications*, vol. 19, no. 4, pp. 71–77, 2021.

[8] S. Sun and G. Wen, "Optimal design of wireless power transmission systems using antenna arrays," *ZTE Communications*, vol. 20, no. 2, 2022.

[9] F. Nesta and Z. Koldovskỳ, "Supervised independent vector analysis through pilot dependent components," in *Proc. IEEE ICASSP*, 2017, pp. 536–540.

[10] Z. Koldovskỳ and P. Tichavskỳ, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, 2018.

[11] R. Scheibler and N. Ono, "Fast independent vector extraction by iterative SINR maximization," in *Proc. IEEE ICASSP*, 2020, pp. 601–605.

[12] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[13] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. Signal Process.*, vol. 69, pp. 3252–3267, 2021.

[14] J. Malek, J. Jansky, Z. Koldovsky, T. Kounovsky, J. Cmejla, and J. Zdansky, "Target speech extraction: Independent vector extraction guided by supervised speaker identification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2295–2309, 2022.

[15] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures.," in *Proc. Interspeech*, 2017, pp. 2655–2659.

[16] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE ICASSP*, 2018, pp. 5554–5558.

[17] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. IEEE ASRU*, 2019, pp. 327–334.

[18] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[19] M. Borsdorf, C. Xu, H. Li, and T. Schultz, "Universal speaker extraction in the presence and absence of target speakers for speech of one and two talkers," in *Proc. Interspeech*, 2021, pp. 1469–1473.

[20] M. Elminshawi, W. Mack, S. Chakrabarty, and E. A. Habets, "New insights on target speaker extraction," *arXiv preprint arXiv:2202.00733*, 2022.

[21] J. Rennies and G. Kidd, "Benefit of binaural listening as revealed by speech intelligibility and listening effort," *J. Acoust. Soc. Am.*, vol. 144, pp. 2147–2159, Oct. 2018.

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

[23] 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, Jul. 2021.

[24] J. Peissiga and B. Kollmeierb, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *J. Acoust. Soc. Am.*, vol. 101, pp. 1660–1670, Sep. 1996.

[25] M. Hawley, R. Litovsky, and J. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, vol. 115, pp. 833–843, Feb. 2004.

[26] B. Bacila and H. Lee, "360° binaural room impulse response (BRIR) database for 6DOF spatial perception research," *J. Audio Eng. Soc.*, Mar. 2019.

[27] J. W. Miller, R. Goodman, and P. Smyth, "On loss functions which minimize to conditional expected values and posterior probabilities," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1404–1408, 1993.

[28] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[29] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proc. DARPA Speech Natural Language Workshop*, pp. 357–362, 1992.

[30] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Appl.*, vol. 80, no. 4, pp. 5653–5671, 2021.

[31] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

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

[33] "Method for measuring the intelligibility of speech over communication systems," ANSI/ASA, Standard, 1989.