REAL-TIME BINAURAL SPEECH SEPARATION WITH PRESERVED SPATIAL CUES

Cong Han, Yi Luo, Nima Mesgarani

Department of Electrical Engineering, Columbia University, New York, NY

ABSTRACT

Deep learning speech separation algorithms have achieved great success in improving the quality and intelligibility of separated speech from mixed audio. Most previous methods focused on generating a single-channel output for each of the target speakers, hence discarding the spatial cues needed for the localization of sound sources in space. However, preserving the spatial information is important in many applications that aim to accurately render the acoustic scene such as in hearing aids and augmented reality (AR). Here, we propose a speech separation algorithm that preserves the interaural cues of separated sound sources and can be implemented with low latency and high fidelity, therefore enabling a real-time modification of the acoustic scene. Based on the time-domain audio separation network (TasNet), a single-channel time-domain speech separation system that can be implemented in real-time, we propose a multi-input-multi-output (MIMO) end-to-end extension of TasNet that takes binaural mixed audio as input and simultaneously separates target speakers in both channels. Experimental results show that the proposed end-to-end MIMO system is able to significantly improve the separation performance and keep the perceived location of the modified sources intact in various acoustic scenes.

Index Terms— Binaural speech separation, interaural cues, deep learning, real-time

1. INTRODUCTION

In real-world multi-talker acoustic environments, humans can easily separate speech and accurately perceive the location of each speaker due to the binaural acoustic features such as interaural time differences (ITDs) and interaural level differences (ILDs). Speech processing methods aimed to modify the acoustic scene are therefore required to not only separate sound sources, but do so in a way to preserve the spatial cues needed for accurate localization of sounds.

However, most of the binaural speech separation systems [1–3] are multi-input-single-output (MISO), and hence lose the interaural cues at the output level which are important for humans to perform sound lateralization and localization [4,5]. To achieve binaural speech separation as well as interaural cues preservation, the multi-input-multi-output (MIMO) setting is necessary, and currently, such setting can be divided into three main categories.

The first category of methods add another stage for binaural sound rendering, such as head related transfer function (HRTF) hypotheses, after a MISO system [6]. This method decouples speech separation and spatial cues preservation, however, it requires robust speaker localization algorithms and a priori knowledge about the HRTF of the listener [7]. Thus, it not only requires additional efforts but limits the system to be listener-dependent.

The second category calculates a real-valued spectro-temporal mask and then applies the same mask to both left and right microphone channels [6, 8–12]. Because both sides obtain the same zerophase gain, the original interaural cues are preserved. However, the separation performance may be limited because of the singlechannel mask estimation and the constraint due to the same gain assumption.

In the third category, complex-valued filters are applied to all available microphone signals simultaneously to generate binaural outputs with an additional constraint on interaural cues preservation. One approach is to use two beamformers at the same time to generate left and right outputs respectively, such as generalized sidelobe canceller (GSC) [13] and binaural minimum variance distortionless response (MVDR) beamformer [14]. Another approach is multichannel wiener filter (MWF) [15] that is equivalent to the combination of spatial filtering and spectral post-filtering. There has been a method that exploits the deep neural network to estimate complex ideal ratio masks (cIRM) for both left and right channels [16]. Since these multi-channel methods aim at estimating the desired separated sources in each channel, the spatial information could be naturally preserved.

One common issue for the systems mentioned above is that the system latency can be perceivable by humans, and the delayed playback of the separated speakers might affect the localization of the signals due to the precedence effect [17]. To decrease the system latency while maintaining the separation quality, a natural way is to use time-domain separation methods with smaller windows. Recent deep learning-based time-domain separation systems have proven their effectiveness in achieving high separation quality and decreasing the system latency [18–21], however, all those systems are still MISO and their ability to perform binaural speech separation and interaural cues preservation is not fully addressed.

In this paper, we look into multiple methods for formulating such systems into MIMO systems and investigate their capability of high-quality separation and interaural cue preservation. Based on the time-domain audio separation network (TasNet) [20], we propose a MIMO TasNet that takes binaural mixture signals as input and simultaneously separates speech in both channels, then the separated signals can be directly rendered to the listener without postprocessing. The MIMO TasNet exploits a parallel encoder to extract cross-channel information for mask estimation and uses mask-andsum method to perform spatial and spectral filtering for better separation performance. We compare it with other variants of TasNet in the tasks. Experiment results show that MIMO TasNet can perform listener-independent speech separation across a wide range of speaker angles and preserve both ITD and ILD features with significantly higher quality than the single-channel baseline. Moreover, the minimum system latency of the systems can be less than 5 ms, showing the potentials for the actual deployment of such systems into real-world hearable devices.

The rest of the paper is organized as follows. We introduce the problem definition of binaural separation with preserved spatial cues and the MIMO variants of TasNet in Section 2, describe the experiment settings in Section 3, discuss the results in Section 4, and concludes the paper in Section 5.

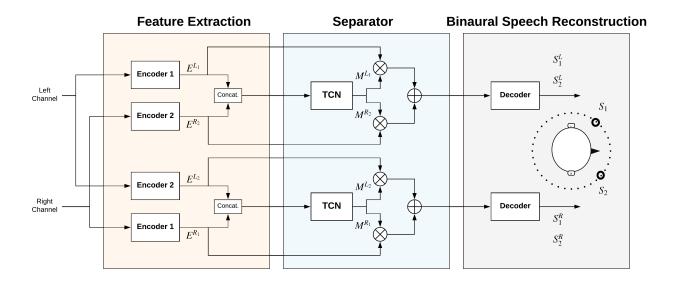


Fig. 1. The architecture of the proposed binaural speech separation network. Two encoders are shared by the mixture signals from both channels, and the encoder outputs for each channel are concatenated together and passed to a mask estimation network. Then, spectral-temporal and spatial filtering are performed by applying the masks to the corresponding encoder outputs and sum them up on both left and right paths. Finally, binaural separated speech are reconstructed by a linear decoder.

2. MIMO TASNET FOR BINAURAL SPEECH SEPARATION

2.1. Problem definition

The problem of binaural speech separation is formulated as the separation of C sources $\mathbf{s}_i^{l,r}(t) \in \mathbb{R}^{1 \times T}$, $i = 1, \ldots, C$ from the binaural mixtures $\mathbf{x}^l(t), \mathbf{x}^r(t) \in \mathbb{R}^{1 \times T}$, where the superscripts l and r denote the left and right channels, respectively. For preserving the interaural cues in the outputs, we consider the case where every single source signal is transformed by a set of head-related impulse response (HRIR) filters for a specific listener:

$$\begin{cases} \mathbf{s}_{i}^{l} = \hat{\mathbf{s}}_{i} \circledast \mathbf{h}_{i}^{l} \\ \mathbf{s}_{i}^{r} = \hat{\mathbf{s}}_{i} \circledast \mathbf{h}_{i}^{r} \end{cases} \quad i = 1, \dots, C$$

$$(1)$$

where $\hat{\mathbf{s}}_i \in \mathbb{R}^{1 \times T'}$ is the monaural signal of source i, $\mathbf{h}_i^l, \mathbf{h}_i^r \in \mathbb{R}^{1 \times (T-T'+1)}$ are the pair of HRIR filters corresponding to the source i, and \circledast represents the convolution operation. Using the HRIR-transformed signals as the separation targets forces the model to preserve interaural cues introduced by the HRIR filters, and the outputs can be directly rendered to the listener.

2.2. MIMO TasNet

2.2.1. TasNet overview

TasNet has been shown to achieve superior separation performance in single-channel mixtures [20]. TasNet contains three modules: a linear encoder first transforms the mixture waveform into a twodimensional representation similar to spectrograms; a separator estimates C multiplicative functions similar to time-frequency masks based on the 2-D representation; and a linear decoder transforms the C target source representations back to waveforms.

Various approaches have been proposed to extend TasNet into

the multi-channel framework [22, 23]. A standard pipeline is to incorporate cross-channel features into the single-channel model, where spatial features such as interaural phase difference (IPD) is concatenated with the mixture encoder output on a selected reference microphone for mask estimation [22]. In various scenarios, such configuration have led to a significantly better separation performance than the signal-channel TasNet.

2.2.2. Design of MIMO TasNet

The proposed MIMO TasNet uses a parallel encoder for spectrotemporal and spatial features extraction and a mask-and-sum mechanism for source separation. A primary encoder is always applied to the channel to be separated, and a secondary encoder is applied to the other channel to jointly extract cross-channel features. In other words, the sequential order of the encoders determines which channel (left of right) the separated outputs belong to. The outputs of the two encoders are concatenated and passed to the separator, and 2C multiplicative functions are estimated for the C target speakers. C multiplicative functions are applied to the *primary* encoder output while the other C multiplicative functions are applied to the secondary encoder output, and the two multiplied results are then summed to create representations for C separated sources. We denote it as the mask-and-sum mechanism to distinguish it from the other methods where only C multiplicative functions were estimated from the separation module and applied to only the reference channel. Similar to TasNet, a linear decoder transforms the C target source representations back to waveforms. Figure 1 shows the flowchart of the system design.

Note that a parallel encoder design for multi-channel TasNet has been discussed in a previous literature [22]. For a N-channel input, N encoders were applied to each of them and the encoder outputs were summed to create a single representation. The multiplicative function was also estimated on the single representation, which resulted in a MISO system design. We can easily find that it is a special case of MIMO TasNet where the two multiplicative functions for the two encoders are equal. Although in [22] an on par performance with respect to the feature concatenation method was reported for the parallel encoder design, in Section 4 we will show that MIMO TasNet is able to significantly surpass feature concatenation TasNet in various configurations in both separation performance and spatial cue preservation accuracy.

2.2.3. Training objective

Scale-invariant signal-to-distortion ratio (SI-SDR) [24] is used as both the evaluation metric and training objective for many recent end-to-end separation systems. SI-SDR between a signal $\mathbf{x} \in \mathbb{R}^{1 \times T}$ and its estimate $\hat{\mathbf{x}} \in \mathbb{R}^{1 \times T}$ is defined as:

$$\text{SI-SDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{||\alpha \mathbf{x}||_2^2}{||\hat{\mathbf{x}} - \alpha \mathbf{x}||_2^2} \right)$$
(2)

where $\alpha = \hat{\mathbf{x}} \mathbf{x}^{\top} / \mathbf{x} \mathbf{x}^{\top}$ corresponds to the rescaling factor. Although SI-SDR is able to implicitly incorporate the ITD information, the scale-invariance property of SI-SDR makes it insensitive to power rescaling of the estimated signal, which may fail in preserving the ILD between the outputs. Hence instead of using SI-SDR as the training objective, we use the plain signal-to-noise ratio (SNR) defined as:

$$\operatorname{SNR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{||\mathbf{x}||_2^2}{||\hat{\mathbf{x}} - \mathbf{x}||_2^2} \right)$$
(3)

3. EXPERIMENTAL SETTINGS

3.1. Dataset

We generated an anechoic speech dataset from the WSJ0-2mix dataset [25]. 30 hours of training data, 10 hours of validation data and 5 hours of test data were generated with the same configuration as the single-channel WSJ0-2mix data, while the clean speech are convolved with randomly sampled HRIR filters from the CIPIC HRTF Database [7]. The CIPIC HRTF Database contains real-recorded HRIR filters for 20 subjects across 25 different interaural-polar azimuths from -80° to 80° and 50 different interaural-polar elevations from -90° to 270° . We randomly sampled two speaker locations in the database for spatial rendering. We used 27 subjects for training and validation sets and 9 unseen subjects for test set, ensuring that the model is evaluated in a listener-independent way. All mixtures were downsampled to 8k Hz.

Anechoic WSJ0-3mix dataset with spatial cues was generated by using the the same method as above. To generate noisy WSJ0-2mix dataset, we added to the training set the noise from one out of eight environmental noises (washing room, kitchen, sport field, city park, office, meeting room) chosen from DEMAND dataset [26] with SNR between -15 and 2.5 dB. The noise in test set is from another eight scenarios (subway station, restaurant, public square, traffic intersection, subway, private car). To generate echoic WSJ0-2mix dataset, HRIR filters are obtained from the BRIR Sim Set¹, which is simulated with different reverberation time (T60). We use rooms with T60 0.1s, 0.2s, 0.4s, 0.5s, 0.7s, 0.8s, 1.0s for training and 0.3s, 0.6s, 0.9s for testing.

3.2. Evaluation metrics

We evaluate the model with both the separation quality and the ability to preserve interaural cues. SNR improvement (SNRi) is used as the signal quality metric instead of SI-SDR improvement according to our discussion in Section 2.2.3. ITD and ILD errors between the separated and target clean signals are used as the metric for the accuracy of preserving interaural cues, which are defined as:

$$\Delta_{ITD} = \left| \text{ITD}(\mathbf{s}^{\text{l}}, \mathbf{s}^{\text{r}}) - \text{ITD}(\bar{\mathbf{s}}^{\text{l}}, \bar{\mathbf{s}}^{\text{r}}) \right|$$
(4)

$$\Delta_{ILD} = \left| 10 \log_{10} \frac{||\mathbf{s}^l||_2^2}{||\mathbf{s}^r||_2^2} - 10 \log_{10} \frac{||\mathbf{s}^l||_2^2}{||\mathbf{s}^r||_2^2} \right|$$
(5)

where $\bar{\mathbf{s}}^l, \bar{\mathbf{s}}^r \in \mathbb{R}^{1 \times T}$ are the separated signals in left and right channels, respectively, $\mathbf{s}^l, \mathbf{s}^r \in \mathbb{R}^{1 \times T}$ are the corresponding target signals, and $|| \cdot ||$ denotes the L_2 -norm of the signal. We use generalized cross-correlation phase transform (GCC-PHAT) algorithm [27] to compute time difference of arrival (TDOA) of \mathbf{s}^l and \mathbf{s}^r as ITD($\mathbf{s}^l, \mathbf{s}^r$). The tool is available online ².

3.3. Network architectures

The configurations of the MIMO TasNet variants are based on the causal setting of the single-channel TasNet [20]. We use 64 filters in the linear encoder and decoder with 2 ms filter length (i.e. 16 samples at 8k Hz). In the causal temporal convolutional network (TCN), there are 4 repeated stacks and each one includes 8 1-D convolutional blocks. The number of parameters in all models are aligned to 1.67M for a fair comparison.

For baseline models, we adopt the following configurations:

- 1. *Single-channel TasNet*: the single-channel model is applied to each channel independently.
- 2. *Feature concatenation TasNet*: cross-channel features are concatenated to the encoder output in the same way as [22]. We use sin(IPD), cos(IPD) and ILD as spatial features, where the IPD and ILD are defined as

$$IPD(\mathbf{X}, \mathbf{Y}) = \angle \mathbf{X} - \angle \mathbf{Y}$$
(6)

$$ILD(\mathbf{X}, \mathbf{Y}) = 10 \log_{10} \left(|\mathbf{X}| \oslash |\mathbf{Y}| \right)$$
(7)

where \mathbf{X}, \mathbf{Y} are the spectrograms of the two channel mixtures, \oslash means element-wise division. The window length of STFT for calculating spectrograms is 256 samples.

3. *Parallel encoder TasNet*: the same configuration as in [22] which is also discussed in Section 2.2.2.

4. RESULTS AND DISCUSSIONS

Table 2 compares different MIMO TasNet variants at various speaker locations on anechoice spatialized WSJ0-2mix. The single-channel baseline is able to achieve the smallest ILD error across all models when the speaker angle is very small, which indicates that the interaural features in this scenario are not helpful in preserving the absolute energy of the separated speech. For all other speaker locations, both the ILD error and separation quality for the singlechannel model are significantly worse than all the MIMO variants. For TasNet concatenated with sin(IPD), cos(IPD) and ILD features, we can observe significant signal quality improvement and ITD/ILD error reduction across all angle ranges, and better performance is

¹http://iosr.uk/software/index.php

²https://www.mathworks.com/help/phased/ref/gccphat.html

	SNRi / Δ_{ITD} / Δ_{ILD}								
Method	Angle								
	<15°	15-45°	45-90°	$>90^{\circ}$	Average				
TasNet	10.0 / 6.0 / 0.29	10.0 / 5.8 / 0.39	10.3 / 4.8 / 0.56	10.7 / 6.4 / 0.59	10.2 / 5.8 / 0.46				
+ILD	11.1 / 3.1 / 0.31	13.4 / 1.9 / 0.12	14.4 / 1.3 / 0.16	14.8 / 1.9 / 0.17	13.4 / 2.0 / 0.19				
+sin(IPD), cos(IPD)	11.7 / 2.4 / 0.34	14.1 / 1.7 / 0.14	14.7 / 1.4 / 0.20	15.3 / 2.0 / 0.20	13.9 / 1.9 / 0.22				
+sin(IPD), cos(IPD), ILD	11.8 / 2.4 / 0.33	14.5 / 1.6 / 0.11	15.3 / 1.2 / 0.16	15.8 / 1.8 / 0.18	14.4 / 1.8 / 0.20				
+parallel encoder	10.6 / 3.0 / 0.47	15.1 / 1.5 / 0.11	16.8 / 1.2 / 0.11	17.7 / 2.0 / 0.12	15.0 / 2.0 / 0.20				
+parallel encoder, mask∑	10.7 / 2.8 / 0.47	15.6 / 1.3 / 0.13	17.7 / 1.1 / 0.09	18.3 / 1.8 / 0.09	15.6 / 1.8 / 0.19				

Table 1. SNR improvement (dB), ITD error (μ s), and ILD error (dB) for different variants of TasNet on anechoic spatialized WSJ0-2mix. The averaged performance on different ranges of speaker angles is reported.

Table 2. Evaluation of TasNet with parallel encoder on several adverse conditions: three-speaker separation, two-speaker separation with environmental noise, and with room reverberance.

Method	$SNRi / \Delta_{ITD} / \Delta_{ILD}$								
	3 speaker	2 speaker with noise (SNR)			2 speaker with reverberance (RT60)				
		12.5 dB	5 dB	-2.5 dB	0.3s	0.6s	0.9s		
TasNet	9.1 / 6.3 / 0.74	9.8/3.4/0.31	10.9 / 3.7 / 0.31	13.8 / 5.2 / 0.57	7.2 / 10.8 / 0.46	6.2 / 45.1 / 0.47	5.7 / 44.5 / 0.50		
+parallel encoder	11.3 / 12.3 / 0.84	13.7 / 2.3 / 0.16	15.0 / 2.5 / 0.18	17.8 / 3.0 / 0.23	9.2 / 6.5 / 0.20	7.7 / 33.2 / 0.25	6.9 / 17.7 / 0.30		
+parallel encoder, mask∑	12.1 / 5.7 / 0.45	14.3 / 2.2 / 0.14	15.3 / 2.3 / 0.15	18.2 / 2.8 / 0.21	9.4 / 5.9 / 0.23	7.8 / 30.0 / 0.21	7.1 / 15.6 / 0.25		

achieved with larger speaker angle. This confirms the previous observations regarding the effectiveness of cross-channel features in end-to-end frameworks [22]. The parallel encoder method has on par performance in preserving ITD/ILD with feature concatenation method, but achieves better separation performance except when the speaker angle is small (less than 15°). The significant improvement for signal quality (SNRi) indicates that the parallel encoders are able to implicitly extract more effective cross-channel features than cross-domain features IPD/ILD for multi-channel speech separation. The further improvement from mask-and-sum mechanism indicates the effectiveness of combining spatial filtering and spectral filtering to separate sources. The correlation (Pearsons r) between SNRi and Δ_{ITD} and between SNRi and Δ_{ILD} are -0.77 and -0.85, respectively (p <0.0001 for both), which means higher separated signal quality helps in preserving ITD/ILD better.

To further examine our proposed MIMO TasNet in more adverse environments, we tested the separation accuracy in three speaker mixtures, noisy speech separation and speech separation with room reverberation. Note that in the evaluation of these three cases, top $5\% \Delta_{ITD}$ and Δ_{ILD} were dropped before averaging to prevent the errors incurred by outliers.

When testing the model on noisy WSJ0-2mix dataset, we set the noise power range at three levels. As shown in Table 2, additive noise contaminates both speech quality and ITD/ILD preservation, but the overall performance compared to the clean condition is still superior and MIMO TasNet with parallel encoder and mask-and-sum achieves the best performance in all metrics across all noise levels, which proves the MIMO TasNet is more robust to the noise.

We observe that three speaker separation is more challenging than noisy speech separation. Both ITD and ILD preservation downgrade significantly than the two-speaker case. That's because the model had failed to separate some of speech with very small power compared to the other two speakers or speech with very similar spatial features to others, and the failure of separation leads to the failure of ITD/ILD preservation.

Finally, we evaluated the model on the echoic spatialized WSJ0-

2mix dataset. The target is the reverberant clean signal. Not surprisingly, convolutive room reverberation is a more challenging condition than additive environmental noises both in terms of signal quality improvement and preserving spatial cues as the sparseness properties of the speech is affected by room reverberation. The smearing casued by reverberation means that the mixture at each instance includes components of the same and different speakers, which makes the mask prediction and TDOA estimation more difficult. As a result, SNRi and Δ_{ITD} are more easily affected by the reverberation. Nonetheless, averaged 9.4 dB SNR improvement, 5.9 μ s ITD error and 0.23 dB ILD error shows that the performance of MIMO TasNet is still helpful in the moderate reverberant environment.

5. CONCLUSION

In this paper, we investigated the problem of real-time binaural speech separation with interaural cues preservation. We proposed a multi-input-multi-output (MIMO) TasNet that uses a parallel encoder and mask-and-sum mechanism to improve performance. Experimental results show that the MIMO TasNet is able to achieve significantly better separation performance and has the ability to preserve interaural time difference (ITD) and interaural level difference (ILD) features in the separated outputs compared to the other existing variants of TasNet. Future works include adapting to environmental noise and room reverberation and incorporate extra microphones for obtaining more cross-channel information, which can pave the way to real-world speech separation solutions for acoustic scene modification.

6. ACKNOWLEDGEMENT

This work was funded by a grant from the National Institute of Health, NIDCD, DC014279; a National Science Foundation CA-REER Award.

7. REFERENCES

- X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.
- [2] Q. Liu, Y. Xu, P. J. Jackson, W. Wang, and P. Coleman, "Iterative deep neural networks for speaker-independent binaural blind speech separation," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 541–545.
- [3] P. Dadvar and M. Geravanchizadeh, "Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target," *Speech Communication*, vol. 108, pp. 41–52, 2019.
- [4] M. Sams, M. Hämäläinen, R. Hari, and L. McEvoy, "Human auditory cortical mechanisms of sound lateralization: I. interaural time differences within sound," *Hearing research*, vol. 67, no. 1-2, pp. 89–97, 1993.
- [5] T. C. Yin, "Neural mechanisms of encoding binaural localization cues in the auditory brainstem," in *Integrative functions in the mammalian auditory pathway*. Springer, 2002, pp. 99– 159.
- [6] 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). IEEE, 2018, pp. 516–520.
- [7] 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). IEEE, 2001, pp. 99– 102.
- [8] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, p. 063297, 2006.
- [9] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures," in 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2007, pp. 315–318.
- [10] K. Reindl, Y. Zheng, and W. Kellermann, "Analysis of two generic wiener filtering concepts for binaural speech enhancement in hearing aids," in 2010 18th European Signal Processing Conference. IEEE, 2010, pp. 989–993.
- [11] J. I. Marin-Hurtado, D. N. Parikh, and D. V. Anderson, "Perceptually inspired noise-reduction method for binaural hearing aids," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1372–1382, 2011.
- [12] M. Azarpour and G. Enzner, "Binaural noise reduction via cuepreserving mmse filter and adaptive-blocking-based noise psd estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 1–17, 2017.
- [13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

- [14] 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, 2015.
- [15] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proc. Int. Workshop Acoust. Echo Noise Control* (*IWAENC*), 2006, pp. 1–4.
- [16] X. Sun, R. Xia, J. Li, and Y. Yan, "A deep learning based binaural speech enhancement approach with spatial cues preservation," in *ICASSP 2019-2019 IEEE International Conference* on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5766–5770.
- [17] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Engineering Society*, vol. 20, no. 2, pp. 146–159, 1972.
- [18] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv* preprint arXiv:1806.03185, 2018.
- [19] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-toend source separation with adaptive front-ends," in 2018 52nd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2018, pp. 684–688.
- [20] 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, 2019.
- [21] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: Endto-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *International Conference* on Multimedia Modeling. Springer, 2020, pp. 653–665.
- [22] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," arXiv preprint arXiv:1905.06286, 2019.
- [23] Y. Luo, E. Ceolini, C. Han, S.-C. Liu, and N. Mesgarani, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019.
- [24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdrhalf-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 626–630.
- [25] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.* IEEE, 2016, pp. 31–35.
- [26] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013.
- [27] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.