INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Binaural Speech Separation of Moving Speakers With Preserved Spatial Cues

Cong Han, Yi Luo, Nima Mesgarani

Department of Electrical Engineering, Columbia University, USA
ch3212@columbia.edu, yl3364@columbia.edu, nima@ee.columbia.edu

Abstract

Binaural speech separation algorithms designed for augmented hearing technologies need to both improve the signal-to-noise ratio of individual speakers and preserve their perceived location in space. The majority of binaural speech separation methods assume nonmoving speakers. As a result, their application to real-world scenarios with freely moving speakers requires block-wise adaptation which relies on short-term contextual information and limits their performance. In this study, we propose an alternative approach for utterance-level source separation with moving speakers and in reverberant conditions. Our model makes use of spectral and spatial features of speakers in a larger context compared to the block-wise adaption methods. The model can implicitly track speakers within the utterance without the need for explicit tracking modules. Experimental results on simulated moving multitalker speech show that the proposed method can significantly outperform block-wise adaptation methods in both separation performance and preserving the interaural cues across multiple conditions, which makes it suitable for real-world augmented hearing applications.

Index Terms: Binaural speech separation, moving speakers, interaural cues, real-time

1. Introduction

Speaker-independent speech separation is crucial for improving speech perception in multitalker listening environments [1]. The developments in hardware technologies have made it possible to build binaural hearing devices with microphones placed on both left and right ears that can communicate wirelessly. These technological advances have enabled binaural speech separation algorithms [2–4] with simultaneous access to both microphone signals. A crucial factor in such algorithms is preserving the spatial cues such as interaural level difference (ILD) and the interaural time difference (ITD) of all directional sources to enable a listener to perceive the correct location of sources in space [5, 6]. Various techniques have been developed for source separation with preserved spatial cues [7–11]. These algorithms can enable immersive audio for applications such as augmented reality (AR) and brain-controlled hearing devices that can detect and amplify an attended speaker to improve the intelligibility and perception of the relative location of speakers in space [12, 13]. Conventionally, most speech separation methods assume the sources are nonmoving which limits their application in real-world scenarios. A commonly used solution is block-wise adaptation of these methods, meaning to split the signal into short blocks in which the sources within each block are assumed stationary since spatial features vary smoothly and slowly [14–16]. Choosing an appropriate block size is not trivial: on one hand, the block size must be small enough to satisfy stationarity assumption; on the other hand, the block size should be long enough for successful separation. For example, independent component analysis (ICA) methods require block size that is large enough for the independence assumption to hold within the block. Besides, block-wise methods require a tracking module to resolve the source permutation problem across consecutive blocks [17, 18].

An alternative solution is to divide the moving source separation problem into speech source localization, tracking, and separation problems, and to tackle them separately [19, 20] or jointly [21]. In [22], an approximate Bayesian tracker that employs the Markovian property of the speaker motion model is used to associate time-frequency bins to each source, based on which spatial filters are estimated to separate moving sources. However, these approaches depend on the high fidelity of source localization and/or tracking. In real scenarios, speech turns, speaker appearance and disappearance, and room reverberation [23] can complicate the source tracking considerably. Moreover, the tracking methods mainly utilize the past spatial information but do not take advantage of long-term spectral information. Long-term spectral information has proven beneficial for source separation [24, 25]. Instead, a method that is able to use both spectral-temporal and spatial-temporal information in a larger context is desired for resolving these challenges.

In this paper, we describe a deep learning approach that performs utterance-level source separation of moving speakers. The approach does not require localization and tracking modules and crucially, is able to preserve the spatial cues in the outputs which enables the correct localization of the separated moving source. Our framework that uses a binaural separation module and a binaural post enhancement module. The binaural speech separation module takes binaural mixed signals as input and simultaneously separates speech in both channels; then the left and right channel speech of each speaker are concatenated and further enhanced by the binaural post enhancement module; the output of the binaural post enhancement module is the separated stereo sound rendered to the listener. The modules employ the TasNet framework [24] that can achieve latency as low as 2 ms which is important for deployment in hearing devices. We created a dataset to simulate moving speakers in reverberant rooms. Experimental results show that utterance-level separation significantly outperforms the block-wise adaptation methods both in terms of signal quality and spatial cue preservation.

The rest of the paper is organized as follows. We introduce the separation framework in Section 2, present the experiment configurations in Section 3, analyze the experiment results in Section 4, and conclude the paper in Section 5.

2. Method

Figure 1 shows the overall flowchart of the proposed framework. In the first step, the binaural speech separation module simultaneously separates the speaker in each channel of the mixed input. In the second step, the binaural post enhancement module enhances each speaker individually. We first describe these two modules and then discuss a speaker localizer for evaluating the preservation of the interaural cues in the stereo output.

http://dx.doi.org/10.21437/Interspeech.2021-1372

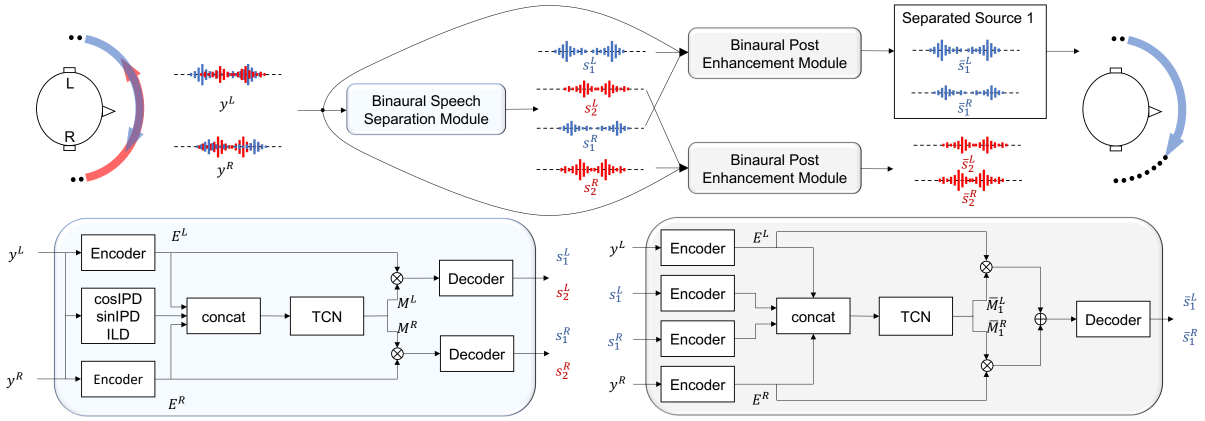

Figure 1: *(top) The architecture of the proposed system. Two moving speakers, denoted in blue and red are being separated. The bottom left and bottom right figures illustrate the details of binaural speech separation and binauarl post enhancement modules.*

2.1. Binaural speech separation module

TasNet has shown superior separation performance on various conditions, and TasNet can be implemented with causal configuration with low latency which is needed for real-time applications [10, 24]. We used multi-input-multi-output (MIMO) TasNet for binaural speech separation. MIMO TasNet contains three steps: (1) spectral and spatial feature extraction, (2) estimation of multiplicative functions, which is similar to 2-D time-frequency masks, and (3) speech reconstruction.

Two linear encoders transform the left- and right-channel of the mixed signals $\mathbf{y}^L, \mathbf{y}^R \in \mathbb{R}^T$ into 2-D representations $\mathbf{E}^L, \mathbf{E}^R \in \mathbb{R}^{N \times H}$, respectively, where N is the number of encoder basis and H is the number of time frames. To enhance the extraction of the spatial features, we explicitly added interaural phase difference (IPD) and interaural level difference (ILD) as additional features. Specifically, we calculate $\text{cosIPD}, \text{sinIPD}, \text{ILD} \in \mathbb{R}^{F \times H}$,

$$\text{cosIPD} = \cos(\angle \mathbf{Y}^L - \angle \mathbf{Y}^R) \tag{1}$$

$$\text{sinIPD} = \sin(\angle \mathbf{Y}^L - \angle \mathbf{Y}^R) \tag{2}$$

$$\text{ILD} = 10 \log_{10} \left(|\mathbf{Y}^L| \oslash |\mathbf{Y}^R| \right) \tag{3}$$

where $\mathbf{Y}^L, \mathbf{Y}^R \in \mathbb{R}^{F \times H}$ are the spectrograms of $\mathbf{y}^L, \mathbf{y}^R$, respectively, F is the number of frequency bins, and $\oslash$ is element-wise division operation. The hop size for calculating $\mathbf{Y}^L, \mathbf{Y}^R$ is the same as that for $\mathbf{E}^L, \mathbf{E}^R$ to guarantee they have the same number of time frames H, although the window length in the encoder is typically much shorter than that in the STFT. Finally, these cross-domain features are concatenated into $\mathbf{E}^M = [\mathbf{E}^L, \mathbf{E}^R, \text{cosIPD}, \text{sinIPD}, \text{ILD}] \in \mathbb{R}^{(2N+3F) \times H}$ as the spectro-temporal and spatial-temporal features.

Subsequently, $\mathbf{E}^M$ is fed into a series of temporal convolutional network (TCN) blocks for estimating multiplicative function $\mathbf{M}^L, \mathbf{M}^R \in \mathbb{R}^{C \times N \times H}$, where C is the number of speakers. We apply $\mathbf{M}^L$ and $\mathbf{M}^R$ to $\mathbf{E}^L$ and $\mathbf{E}^R$, respectively, and use a linear decoder to transforms the multiplied representations back to the waveforms, $\{\mathbf{s}_i^L\}_{i=1}^C$ and $\{\mathbf{s}_i^R\}_{i=1}^C$. Due to the permutation problem [26], the order of the estimated speakers in each channel cannot be pre-determined. However, we constrain the speaker order in two channels to be the same, which is important to pair the left- and right- channel signals of the individual speaker in a real-time system.

2.2. Binaural post enhancement module

Post enhancement processing stages have proven effective in further improving the signal quality [27]. Each stereo sound, $\mathbf{s}_i^L$ and $\mathbf{s}_i^R$, from the separation module, combined with the mixed signals ($\mathbf{y}^L$, $\mathbf{y}^R$), is sent to a multi-input-single-output (MISO) TasNet for post enhancement. Similar to the speech separation module, we concatenated all the encoder outputs and passed them to the TCN blocks for estimating multiplicative functions $\mathbf{M}_i^L, \mathbf{M}_i^R \in \mathbb{R}^{2 \times N \times H}$,

$$\mathbf{s}_i^L = \text{decoder}(\mathbf{E}^L \cdot \mathbf{M}_i^L[0,:,:] + \mathbf{E}^R \odot \mathbf{M}_i^L[0,:,:]) \tag{4}$$

$$\mathbf{s}_i^R = \text{decoder}(\mathbf{E}^L \cdot \mathbf{M}_i^L[1,:,:] + \mathbf{E}^R \odot \mathbf{M}_i^L[1,:,:]) \tag{5}$$

where, $\odot$ denotes element-wise multiplication. Different from the speech separation module that only applies multiplicative functions, which is equivalent to spectral filtering, speech enhancement module performs multiplication and sum, which is equivalent to both spectral and spatial filtering. This is similar to multichannel wiener filtering [9]. We denote it as as the mask-and-sum mechanism [10].

Since the input stereo sound, $\mathbf{s}_i^L$, $\mathbf{s}_i^R$, contains both spectral and spatial information of the speaker i, the enhancement module essentially performs informed speaker extraction without the need for permutation invariant training.

2.3. Speaker localizer

The speaker localizer adopts the similar architecture as the speech enhancement module but performs classification of the direction of arrival (DOA). We discretize the DOA angles into K classes. The speaker localizer takes only stereo sound, $\mathbf{s}_i^L$, $\mathbf{s}_i^R$, as input, concatenates two encoders' outputs, and passes them to the TCN blocks to estimate a single-class classification matrix $\mathbf{V}_i \in (0,1)^{K \times H}$, where "single-class" means that in each time frame, there is exactly one class labeled with 1 and all the other classes are labeled with 0.

We splits $\mathbf{V}_i$ into B small chunks $\{\mathbf{V}_i^b\}_{b=1}^B \in \mathbb{R}^{K \times Q}$, where Q is the number of time frames in each chunk and $B = \frac{H}{Q}$. In each chunk, we count the frequency of each class labeled with 1, and regard the most frequent class as the estimated DOA for that chunk.

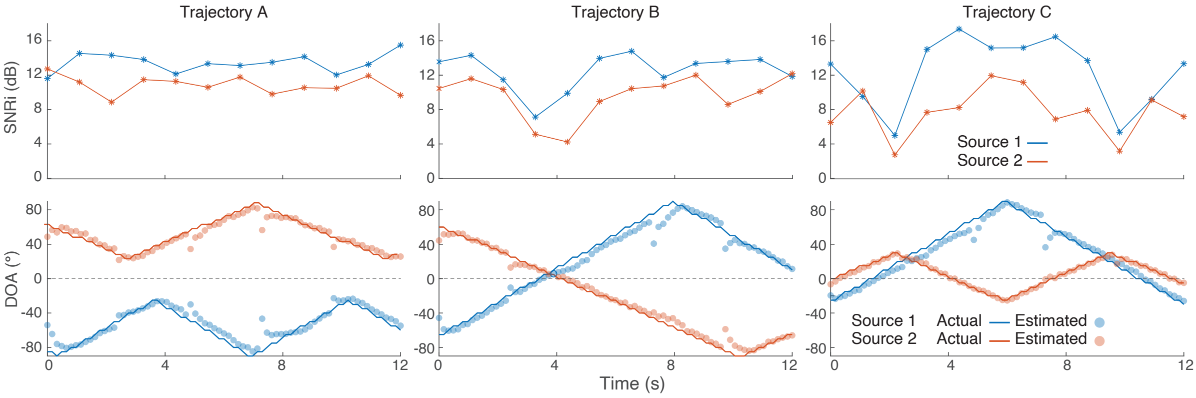

Figure 2: *SNR improvement (top) and DOA estimation (bottom) over time for two moving speakers on three examples of trajectories in the reverberant room with RT60 0.2s. The result for each trajectory is averaged over 100 instances of speech separation.*

2.4. Training objective

We use the signal-to-noise ratio (SNR) as the training objective for the speech separation and enhancement modules. SNR is sensitive to both time shift and power scale of the estimated waveform, so it's able to force the ITD and IPD to be preserved in the estimated waveform [10]. SNR is defined as:

$$\text{SNR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{||\mathbf{x}||_2^2}{||\hat{\mathbf{x}} - \mathbf{x}||_2^2} \right) \tag{6}$$

where $\hat{\mathbf{x}}$ and $\mathbf{x}$ is the estimated and reference signal, respectively. In the speech separation module, we used utterance-level permutation invariant training [28].

$$\mathcal{L} = \min_{\pi \in P} \sum_{c=1}^{C} \text{SNR}(\hat{\mathbf{x}}_c^L - \mathbf{x}_{\pi(c)}^L) + \text{SNR}(\hat{\mathbf{x}}_c^R - \mathbf{x}_{\pi(c)}^R) \tag{7}$$

where P is the set of all C! permutations. The same permutation π for left- and right-channel signals guarantees the speaker order is consistent in both channels.

3. Experimental settings

3.1. Dataset

3.1.1. *Binaural room impulse responses and speech data*

We used two types of binaural room impulse responses (BRIRs); one obtained from simulated rooms and other which was measured in real rooms, [1]. There were 11 simulated rooms with reverberation time (RT60) varying from 0 to 1 s with 0.1 s increments. In this study, we only used 8 rooms with RT60 from 0 to 0.7 s. There were 4 real rooms with RT60 0.32 s, 0.47 s, 0.68 s, 0.89 s, respectively. The impulse responses were calculated with the sound source located on the frontal azimuthal plane between -90° and 90° with 5° increments at a distance of 1.5 m to the receiver. Two speakers were randomly selected from the 100-hour Librispeech dataset [29]. Both speech data and BRIRs were sampled at 16 kHz.

3.1.2. *Moving source simulation*

Given a monaural speech $\mathbf{s}$ and a set of BRIRs $\{h_j^L\}_{j=1}^N, \{h_j^R\}_{j=1}^N \in \mathbb{R}^{L_h}$, where h_j^L and h_j^R are the BRIR filters of length L_h from the location j to the left ear and right ear, respectively, and N is the number of locations (37 in this study), the moving binaural source was simulated as:

$$\mathbf{s}^L[n] = \sum_{j=0}^{N} \sum_{k=0}^{L_h} \mathbb{I}_j(n) \cdot h_j^L[k] \cdot s[n-k] \tag{8}$$

$$\mathbf{s}^R[n] = \sum_{j=0}^{N} \sum_{k=0}^{L_h} \mathbb{I}_j(n) \cdot h_j^R[k] \cdot s[n-k] \tag{9}$$

where $\mathbb{I}_j(n)$ is an indicator function which is 1 when $\mathbf{s}$ is at location j at time step n, and is 0 otherwise. This method simulates the stereo sound with time-varying spatial cues.

3.1.3. *Training, development, and test sets*

For the training, development, and test set, we respectively created 40000, 10000, and 6000 utterances of length 2.4-second using only the simulated BRIRs. For each utterence, we randomly sampled a set of BRIRs and two speech samples. The speech signals were rescaled to a random relative SNR between 0 and 5 dB. The moving velocity of each speaker was randomly chosen between 8 and $15^\circ/s$ and moving direction was randomly chosen between clockwise and counter-clockwise. In addition, to compare the proposed method in different rooms with different velocity, we chose 3 simulated rooms with RT60 0.3 s, 0.5 s, 0.7 s; 3 real rooms with RT60 0.32 s, 0.47 s, 0.68 s; and 3 velocity ranges $5-10^\circ/s$, $10-15^\circ/s$, $15-20^\circ/s$, and generated 1000 utterances on each condition for testing only.

3.2. Evaluation metrics

We evaluate the models by measuring the separation quality and the preservation of spatial cues. We use SNR as the signal quality metric. In this study, ITD and ILD errors are not suitable for utterance-level evaluation of spatial cues preservation because the moving sources are at different locations in an utterance. Therefore, we trained a speaker localizer for moving source localization on reverberant clean signals. We set the chunk size as 80 ms, so the localizer predicts the DOA every 80 ms. Since the DOA classes are ordered, even a wrong classification can correspond to a close DOA estimation, e.x., 5° angular error. As a result, We report absolute DOA errors as the metric for the accuracy of preserving spatial cues.

[1] http://iosr.uk/software/index.php

3.3. Network architectures

Our MIMO TasNet was based on the causal configuration of TasNet [24]. We used 64 filters in the linear encoder and decoder with 4 ms filter length (i.e. 64 samples at 16 k Hz). We used 5 repeated stacks each having seven 1-D convolutional blocks in the TCN module. The effective receptive filed of he model is approximately 2.5 s. We set the STFT window size to 32 ms and the window shift to 2 ms when calculating cosIPD, sinIPD, and ILD in Section 2.1.

4. Results and discussions

Table 1: *Experimental results of moving source separation in reverberant rooms with various configurations of TasNet. SNR (dB) and DOA error (°).*

Method	Context size (s)	SNR (dB)	DOA error (°)
Unprocessed	-	0	-
SIMO TasNet	2.4	5.1	20.9
MIMO TasNet+block-wise	0.1	5.7	16.3
	0.2	6.0	15.4
	0.3	6.2	14.1
MIMO TasNet	2.4	**8.4**	**8.3**
-sinIPD, cosIPD, ILD		7.3	11.0
MISO TasNet	2.4	**9.4**	**6.1**
-mask&sum		8.8	7.3
Reverberant clean	-		0.5

Table 1 compares different methods for moving source separation. The single-input-multi-output (SIMO) TasNet uses only spectral-temporal information for separation, yielding an average of 5.1 dB SNR improvement. The block-wise adaptation of MIMO TasNet with oracle block tracking achieves better performance than the SIMO TasNet even though it relies on a much shorter context window. This observation suggests the importance of spatial information in source separation. As the duration increases, both SNR and DOA estimation becomes better. When the MIMO TasNet performs utterance-level separation, it achieved 8.4 dB SNR improvement, outperforming block-wise adaptation by a large margin. The huge improvement confirms the effectiveness of the longer context for moving source separation, as the model could take advantage of longer spectral-temporal and spatial-temporal information. We also notice that the performance of MIMO TasNet drops greatly when the frequency-domain features, i.e., cosIPD, sinIPD, and ILD, are removed. It's likely because the frequency-domain features provide more stable spatial features than those extracted by the parallel encoders in the reverberant environments as STFT uses a longer window size than the linear encoders. In the binaural enhancement stage, the MISO TasNet further improves the SNR and reduces the DOA error. The performance gain from the mask-and-sum mechanism shows the effectiveness of combining spatial filtering and spectral filtering to separate sources. Our results show that better signal quality (higher SNR) always leads to better preservation of spatial cues (lower DOA error), which is consistent with the observations in [10,11] that conducted source separation on nonmoving sources.

Figure 2 reports SNR improvement and DOA estimation on three example trajectories. In example A, two speakers move in left and right planes, and the proposed method achieves good performance constantly because two speakers always have distinct spatial information. In B and C trajectories, the SNR improvement becomes less as the two speakers move closer to each other, and improves when they move apart. Interestingly, we notice that the DOA estimation is less affected by speakers' co-location than the SNR. A possible explanation is that the speaker localizer uses the velocity and directional information to compensate for decreased signal quality.

Table 2: *Experimental results of the proposed system with different moving velocity in different rooms. SNR (dB) and DOA error (°) are reported.*

Room condition (RT60)		Velocity of motion (°/s)		
		5-10	10-15	15-20
Simulated room	a (0.3 s)	8.8/5.8	8.7/5.6	8.9/6.2
	b (0.5 s)	7.6/7.7	7.6/7.3	7.5/8.3
	c (0.7 s)	6.8/9.2	6.7/9.3	6.7/10.0
Real room	A (0.32 s)	7.5/9.6	7.5/9.9	7.5/9.3
	B (0.47 s)	6.6/15.2	6.5/15.0	6.5/15.2
	C (0.67 s)	6.5/11.9	6.5/12.3	6.3/12.3

Table 2 compares the proposed method on speakers with different moving velocities in different rooms. The model trained in the simulated rooms can generalize to the real rooms well. In both simulated and real rooms, reverberation substantially deteriorates the performance of the model in terms of the signal quality and accuracy of the preserved spatial cues. This degradation is due to the temporal smearing of the mixed signal which combines the components of the same and different speakers over time and makes the mask estimation more challenging. We notice that velocity has little impact on SNR and , DOA estimation only becomes slightly worse in the highest velocity range. It shows the robustness of the system across various conditions.

5. Conclusion

In this paper, we investigated the problem of binaural speech separation of moving speakers with interaural cues preservation. We proposed a two-stage framework that performs utterance-level binaural speech separation and enhancement sequentially. The proposed method fully utilizes long-term spectral and spatial information and implicitly tracks the speakers within the utterance without the need for the external speaker tracking module. Experimental results show that the proposed model is able to achieve significantly better separation performance and preserves spatial cues more accurately compared to the conventional block-wise adaptation method even with oracle block tracking. The model trained with simulated room imluse responses could successfully generalize real rooms. Future works include alleviating the performance degradation when moving sources are co-located and incorporating extra microphones to deal with room reverberation.

6. Acknowledgements

This work was funded by a grant from the National Institute of Health, NIDCD, DC014279; and a grant from Marie-Josée and Henry R. Kravis.

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. Alinaghi, W. Wang, and P. J. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2013, pp. 684–688.

[3] 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.

[4] 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.

[5] A. 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, 1988.

[6] ——, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3132–3139, 1992.

[7] 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.

[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, 2015.

[9] 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 *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2006, pp. 1–4.

[10] C. Han, Y. Luo, and N. Mesgarani, "Online deep attractor network for real-time single-channel speech separation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 361–365.

[11] 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*, 2020.

[12] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science advances*, vol. 5, no. 5, p. eaav6134, 2019.

[13] P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Joint representation of spatial and phonetic features in the human core auditory cortex," *Cell reports*, vol. 24, no. 8, pp. 2051–2062, 2018.

[14] A. Koutvas, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2000, pp. II1133–II1136.

[15] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation for moving speech signals," in *Speech Enhancement*. Springer, 2005, pp. 353–369.

[16] J. Zhang and P.-C. Ching, "Blind separation of moving speech sources using short-time lod based ica method," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2007, pp. III–957.

[17] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–469.

[18] ——, "Blind source separation for moving speech signals using blockwise ica and residual crosstalk subtraction," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 87, no. 8, pp. 1941–1948, 2004.

[19] N. Chong, S. Nordholm, B. T. Vo, and I. Murray, "Tracking and separation of multiple moving speech sources via cardinality balanced multi-target multi bernoulli (cbmember) filter and time frequency masking," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS).* IEEE, 2016, pp. 88–93.

[20] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel nmf and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[21] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based ondoa-hmm," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 3191–3195.

[22] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[23] X. Li, L. Girin, R. Horaud, S. Gannot, X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1997–2012, 2017.

[24] 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.

[25] C. Li, Z. Chen, Y. Luo, C. Han, T. Zhou, K. Kinoshita, M. Delcroix, S. Watanabe, and Y. Qian, "Dual-path modeling for long recording speech separation in meetings," *arXiv preprint arXiv:2102.11634*, 2021.

[26] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2017, pp. 241–245.

[27] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speaker separation," *arXiv preprint arXiv:2010.01703*, 2020.

[28] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[29] G. Hu, "100 nonspeech sounds,"," *Online: http://www. cse. ohio-state. edu/pnl/corpus/HuCorpus. html*, 2006.