ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446934

LEVERAGING SOUND LOCALIZATION TO IMPROVE CONTINUOUS SPEAKER SEPARATION

Hassan Taherian[1], *Ashutosh Pandey*[3], *Daniel Wong*[3], *Buye Xu*[3], and *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
[3]Meta Reality Labs, USA

ABSTRACT

Continuous speaker separation aims to separate overlapping speakers in real-world environments like meetings, but it often falls short in isolating speech segments of a single speaker. This leads to split signals that adversely affect downstream applications such as automatic speech recognition and speaker diarization. Existing solutions like speaker counting have limitations. This paper presents a novel multi-channel approach for continuous speaker separation based on multi-input multi-output (MIMO) complex spectral mapping. This MIMO approach enables robust speaker localization by preserving inter-channel phase relations. Speaker localization as a byproduct of the MIMO separation model is then used to identify single-talker frames and reduce speaker splitting. We demonstrate that this approach achieves superior frame-level sound localization. Systematic experiments on the LibriCSS dataset further show that the proposed approach outperforms other methods, advancing state-of-the-art speaker separation performance.

Index Terms— MIMO complex spectral mapping, continuous speaker separation, robust speaker localization.

1. INTRODUCTION

In the presence of speech overlap, the performance of automatic speech recognition (ASR) systems degrades drastically as they are tailored for single-talker speech. To tackle this, various approaches have been proposed. Some utilize end-to-end multi-talker ASR for direct transcription of overlapped speech [1, 2], while others integrate speaker separation with diarization [3, 4] or employ continuous speaker separation (CSS) [5, 6, 7]. The latter, in particular, is designed for processing long audio recordings and handling overlapping speech with a variable number of speakers. CSS divides an audio stream into shorter, partially overlapped segments, typically around 2-3 seconds long, which contains a maximum of two speakers. This partitioning facilitates a two-talker separation model for each individual segment. The separation model processes each segment independently and produces two estimated signals for each segment. When dealing with segments without overlapped speech, the model focuses on speech enhancement. In these cases, the enhanced signal is mapped to one of the outputs, while the other output generates a zero signal. Finally, the adjacent segments are stitched together to ensure that any single-talker utterance spanning two segments can be integrated into the same output stream.

This research was supported in part by an National Science Foundation grant (ECCS-2125074), a research contract from Meta Reality Labs, the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

However, when processing single-talker segments, a separation model sometimes fails to isolate the speaker in one stream and maintain silence in the other stream. This results in speech mistakenly split into two streams and creates residual speech signals in the stream that is supposed to produce no speech. These split signals harm downstream speech applications, such as ASR or speaker diarization, as the residual signals are processed as if they originated from a valid talker. To mitigate the speaker splitting issue, Wang et al. utilized a supervised speaker counting (SC) network to detect single-talker segments in order to suppress residual speech [8, 9].

When multiple microphones are available, spatial information could be leveraged for speaking counting or addressing the speaker splitting issue. Standard localization techniques like generalized cross-correlation with phase transform (GCC-PHAT) [10] can be applied to determine the number of speakers. However, the accuracy of these techniques suffers in noisy-reverberant conditions due to spurious or broad peaks in the cross-correlation dimension. In a recent study, we introduced multi-input multi-output (MIMO) complex spectral mapping to estimate the target signal at all microphones simultaneously, which achieved strong separation performance [11]. MIMO complex spectral mapping retains inter-channel phase relations, which can be utilized for accurate direction of arrival (DOA) estimation.

Assuming speakers are still within each segment, we propose to use speaker localization as the byproduct of MIMO separation to detect the frames originating from the same speaker in order to reduce speaker splitting. This approach is based on the observation that multiple speakers cannot occupy the same location at the same time. We show that localization using MIMO complex spectral mapping produces superior frame-level localization results. Moreover, experiments on the LibriCSS corpus [5] demonstrate that the proposed approach yields better separation performance than other competitive methods, achieving a new state-of-the-art on this open dataset.

The paper is structured as follows: Section 2 covers related works. The proposed algorithm is detailed in Section 3. Experimental setup and results are presented in Sections 4 and 5, respectively. We conclude in Section 6.

2. RELATED WORKS

Most works on deep learning-based speaker localization predominantly focus on single-talker scenarios (See [12] for a recent review). Typically, these methods leverage deep neural networks (DNNs) to directly estimate the DOA. For instance, [13] employs a DNN to process phase components from the Short-Time Fourier Transform (STFT) across all microphones, subsequently generating posterior probabilities for each DOA class.

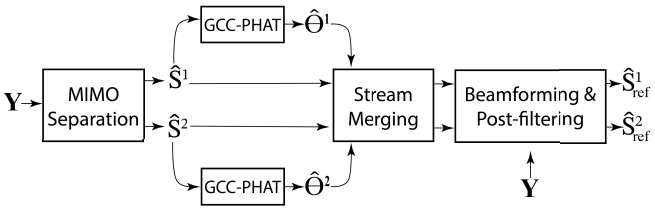

Fig. 1: Schematic diagram of MIMO-BF-MISO system and proposed stream merging method.

For multi-speaker localization, a common assumption is that the number of speakers is known. [14] employs a DNN to estimate the DOAs of two speakers. Instead of utilizing a multi-label classification approach, they estimate speaker-specific DOAs and their posterior probabilities through a source-splitting mechanism. For situations where the number of speakers is unknown, solutions range from employing dedicated DNNs for speaker counting [15] to training systems that combine speaker counting and localization [16].

Another approach involves separating and enhancing noisy mixture signals and subsequently using traditional signal processing techniques for DOA estimation. This indirect approach sidesteps the need for additional DNN-based DOA estimation, reducing computational overhead. They also isolate each speaker in the mixture into its own output stream, thereby facilitating the application of single-talker localization [17, 18]. Wang et al. [18] utilizes a DNN to estimate monaural time-frequency (T-F) masks for speech enhancement. These estimated masks help identify T-F units dominated by the target speaker and selectively use them for localization. Such T-F units contain cleaner phase information, enabling more accurate localization. This information serves as a weighting mechanism alongside the phase from noisy-reverberant multi-channel input signals in a GCC-PHAT algorithm to estimate the DOA. An extension to this approach is to use multi-input single-output (MISO) complex spectral mapping to directly estimate both the magnitude and phase components of the target speaker at a reference microphone [8, 19, 20]. While this method offers better estimated phase quality, it comes with higher computational costs. This is because the enhanced phase from all microphones is needed for localization, requiring the MISO model to be applied as many times as the number of microphones. Additionally, inter-channel phase relations are not guaranteed to be preserved in the enhanced signals.

In contrast, MIMO complex spectral mapping estimates the enhanced phase from all microphones at once while retaining inter-channel phase relations. This enables the use of enhanced phase information from all microphones for localization with GCC-PHAT. In this context, we demonstrate that MIMO complex spectral mapping achieves superior localization performance compared to both MISO complex spectral mapping and masking-based GCC-PHAT approaches.

3. PROPOSED ALGORITHMS

3.1. MIMO Complex Spectral Mapping

We employ MIMO complex spectral mapping to estimate both real and imaginary (RI) components of the target signal across M microphones using a multi-channel noisy mixture. Given a M-channel mixture signal $\boldsymbol{Y} = [Y_1, \ldots, Y_M]$ in STFT domain, the MIMO separation model estimates the direct-path complex spectrograms

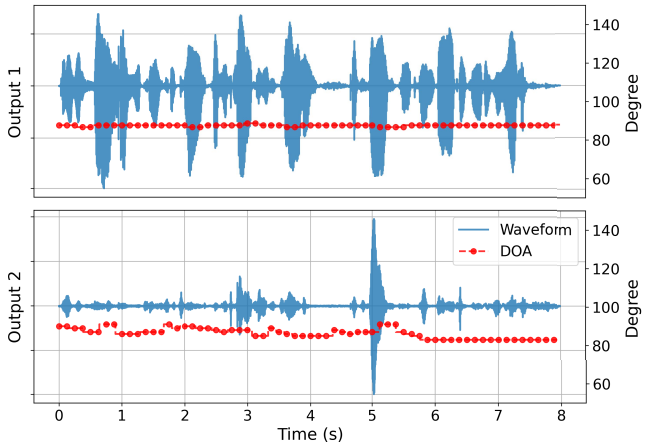

Fig. 2: DOA estimates and separated waveforms for a single-talker segment from LibriCSS. The blue lines represent the audio waveforms, while the red dots represent frame-wise DOA estimates. The model fails to isolate the speaker in one stream and maintain silence in the other.

$\hat{S}_q^n, q \in [1, \ldots, M]$ of target speaker n for each microphone. Our approach employs TF-GridNet [21], which processes speech spectrograms in a grid-like fashion. In adapting the architecture to accommodate the microphone count, we extend the output layer, introducing a slight increase in computational overhead. We train the model with an ℓ_1 norm loss on the RI components of the estimated $\hat{\boldsymbol{S}} = [\hat{S}_1, \ldots, \hat{S}_M]$ and target speech $\boldsymbol{S} = [S_1, \ldots, S_M]$, with an additional magnitude loss term averaged across all microphones:

$$\mathcal{L}(\hat{\boldsymbol{S}}, \boldsymbol{S}) = \frac{1}{M} \sum_{m=1}^{M} \Big\| \Re(\hat{S}_m) - \Re(S_m) \Big\|_1 + \Big\| \Im(\hat{S}_m) - \Im(S_m) \Big\|_1 + \Big\| |\hat{S}_m| - |S_m| \Big\|_1 \tag{1}$$

where $\Re(.)$ and $\Im(.)$ extract real and imaginary parts, $|\,.\,|$ computes magnitude and $\|\,.\,\|_1$ computes ℓ_1 norm. For speaker separation, we use the location-based training [22] criterion to tackle the permutation ambiguity problem [23].

For enhanced performance, we combine the MIMO separation model with a multi-channel Wiener filter (MCWF) beamformer and MISO speech enhancement [24, 21]. In each separation output, we concatenate the beamformed signal with the RI components of the mixture and estimated signals from all microphones. This combined input is then fed into the enhancement model. The enhancement model is trained using the aforementioned loss function. We denote this model as MIMO-BF-MISO.

3.2. Speaker Localization and Stream Merging

We perform speaker localization with GCC-PHAT [12, 17, 18] to attenuate speech residuals produced by the separation model. Fig. 1 illustrates the proposed stream merging method for MIMO-BF-MISO. Instead of relying on the mixture signal, we leverage the multi-channel estimated signals generated by the separation model for speaker localization. MIMO complex spectral mapping yields separated signals with cleaner phase information compared to the mixture signal, all while preserving vital inter-channel cues. For

Algorithm 1 Stream Merge Algorithm

```
candidate_frames ← []
for each frame t do
    Find DOA for stream 1, θ̂¹(t)
    Find DOA for stream 2, θ̂²(t)
    if |θ̂¹(t) − θ̂²(t)| < 5° then
        Append candidate_frames with t
    end if
end for
for each continuous interval in candidate_frames with length ≥ 3 frames do
    Merge two streams
    Attenuate the weaker stream
end for
```

a given microphone pair (p, q), the GCC-PHAT coefficient at time frame t and frequency bin f is defined as:

$$\text{GCC}^{j}_{p,q}(t, f, \theta) = \cos\left(\angle\hat{S}^{j}_{p}(t, f) - \angle\hat{S}^{j}_{q}(t, f) - 2\pi\frac{f}{N}f_s\tau_{p,q}(\theta)\right) \quad (2)$$

where N represents the number of discrete-time Fourier transform frequencies, f_s is the sampling rate, and $\tau_{p,q}(\theta)$ signifies the time delay between microphones p and q for direction θ. The symbol j denotes the output stream index. Compared to mixture signal, the inter-channel phase difference between estimated signals, $\angle\hat{S}^{j}_{p}(t, f) - \angle\hat{S}^{j}_{q}(t, f)$ is more robust to reverberation and noise. The subsequent step involves summing the GCC-PHAT coefficients across all microphone pairs and frequency units. The direction θ that results in the highest summation is deemed the estimated direction:

$$\hat{\theta}^{j}(t) = \underset{\theta}{\text{argmax}} \sum_{(p,q),f} |\hat{S}^{j}_{p}(t, f)||S^{j}_{q}(t, f)|\text{GCC}^{j}_{p,q}(t, f, \theta). \quad (3)$$

In Eq. (3), we include the magnitude of the estimated signals as a weighting term to emphasize speech-dominant T-F units. Once we acquire DOA estimates for both stream at each time frame, we identify frames where two streams likely originate from the same talker. Fig. 2 illustrates an example of a single-talker segment processed by the separation model where the segment is split into two output streams with similar frame-wise DOA estimates. We merge continuous frame intervals with minor DOA differences and attenuate the weaker stream by multiplying it with a small constant. Algorithm 1 outlines the stream merging process for each segment.

4. EXPERIMENTAL SETUP

We evaluate the proposed method for conversational speech recognition tasks using the LibriCSS corpus. This dataset comprises 10 hours of meeting-style speech recordings that include some overlapping speech. Originally sourced from the LibriSpeech development set, these recordings were played back through loudspeakers to simulate real-world room acoustics. The audio was captured using a circular microphone array, consisting of six microphones arranged in a circle with a 4.25 cm radius and one additional microphone in the center. The LibriCSS corpus is divided into six sessions, each featuring varying degrees of speech overlap: 0S (no overlap with pauses of 0.1-0.5 seconds between utterances), 0L (no overlap with pauses of 2.9-3.0 seconds), and 10%, 20%, 30%, and 40% overlaps. All recordings were made at a 16 kHz sampling rate.

Table 1: Comparison of localization accuracy for MIMO, MISO, and masking-based GCC-PHAT on SMS-WSJ.

Localization Method	Frame-level	Utterance-level
MIMO GCC-PHAT	95%	100%
MISO GCC-PHAT	85%	99%
Masked-based GCC-PHAT	71%	100%

We utilized the standard ASR system that comes with the LibriCSS dataset. The corpus offers two evaluation setups: one focused on individual utterances and another on continuous speech. In the utterance-wise evaluation, the precise starting and ending points of each spoken segment are given. The ASR system then evaluates each separated audio signal on its own, choosing the one with the lowest word error rate (WER) as the best. In the continuous speech setup, the boundaries between utterances are not defined, and each recording contains 8-10 utterances. Unlike the utterance-wise setup, the decoded results from both separated audio are used to calculate the final WER. For single-talker segments, this continuous approach demands that one audio stream exclusively contains the speaker's voice, while the other must be completely silent. Otherwise, any residual signal may introduce insertion errors into the evaluation.

To create training and validation data, we used the approach outlined in [8] using simulated room impulse responses (RIRs) [25, 26]. We generated 192K two-speaker audio mixtures with varying overlap ratios from the LibriSpeech dataset. These mixtures were processed using 7-channel microphone array RIRs, matching the LibriCSS recording setup. The RIRs were created in virtual rooms with varying random dimensions, where the microphone array was centrally located. Speaker positions were randomly selected from 360 possible angles, and the reverberation time ranged from 0.2 to 0.6 seconds.

For TF-GridNet, we employed 4 layers with a kernel size of $I = 4$, a stride of $J = 1$, and embedding dimensions set at $D = 48$, along with BLSTM hidden units of $H = 192$. We sequentially trained the separation and enhancement networks using a learning rate of 0.001. The frame length and shift for training were set at 32 ms and 8 ms, respectively. In the case of MISO models, the first microphone served as the reference ($q = 1$), while for MIMO models, we used the estimated complex spectrograms from this same microphone for ASR evaluation. We applied sample variance normalization to the multi-channel input. For more accurate and stable results in stream merging with localization, we used a larger frame length of 256 ms and a frame shift of 128 ms.

The LibriCSS dataset lacks ground-truth speaker positions, so to assess the localization accuracy of our MIMO separation model, we use the SMS-WSJ dataset [28] instead. This dataset includes two-speaker mixtures in reverberant conditions and has an 8 kHz sampling rate. We train the MIMO separation model using the SMS-WSJ array configuration, which features a circular array of six microphones evenly spaced on a 10 cm radius circle. For localization evaluation, we opt for smaller frame lengths and shifts—specifically, 20 ms and 10 ms, respectively. Localization accuracy is determined by the percentage of frames where the estimated DOA is within 5 degrees of the actual direction. This accuracy is averaged across

Table 2: WER (%) results of comparison systems for utterance-wise and continuous evaluation on LibriCSS. 'SC' and 'LOC' denote the speaker counting and localization methods used for stream merging, respectively.

Models	Stream Merging	Utterance-wise						Continuous					
		0S	0L	10%	20%	30%	40%	0S	0L	10%	20%	30%	40%
Unprocessed	–	11.8	11.7	18.8	27.2	35.6	43.3	15.4	11.5	21.7	27.0	34.3	40.5
BLSTM [5]	–	8.3	8.4	11.6	16.0	18.4	21.6	11.9	9.7	13.4	15.1	19.7	22.0
Conformer [27]	–	7.2	7.5	9.6	11.3	13.7	15.1	11.0	8.7	12.6	13.5	17.6	19.6
MISO-BF-MISO (UNet) [8]	SC	5.8	5.8	5.9	6.5	7.7	8.3	7.7	7.5	7.4	8.4	9.7	11.3
MIMO-BF-MISO (UNet) [11]	–	6.3	6.1	6.0	6.8	7.4	8.5	7.4	7.5	7.2	7.4	8.8	9.6
MISO-BF-MISO (TF-GridNet)	–	6.1	6.3	5.9	6.1	6.7	7.8	8.0	8.4	7.4	7.1	9.0	9.3
MISO-BF-MISO (TF-GridNet)	SC	5.7	5.8	5.6	5.9	7.1	8.0	7.0	**6.8**	**6.7**	**6.9**	8.5	9.5
MISO-BF-MIMO (TF-GridNet)	–	6.1	6.3	5.9	6.1	6.9	8.0	8.0	8.2	7.7	7.1	8.8	9.7
MISO-BF-MIMO (TF-GridNet)	SC	5.6	6.0	5.7	6.0	7.3	8.2	7.2	6.8	6.8	7.1	8.8	9.7
MISO-BF-MIMO (TF-GridNet)	LOC	5.4	**5.6**	5.6	6.1	6.9	8.3	7.1	7.0	6.8	6.9	8.7	9.3
MIMO (TF-GridNet)	–	7.5	7.4	7.3	8.3	9.6	10.3	9.2	12.2	9.9	10.1	11.9	12.2
MIMO (TF-GridNet)	LOC	5.8	6.4	6.7	7.9	9.5	10.3	8.4	9.1	9.4	10	11.3	12.0
MIMO-BF-MISO (TF-GridNet)	LOC	**5.3**	5.7	**5.5**	**5.8**	**6.8**	**7.1**	**6.8**	**6.8**	**6.7**	**6.9**	**8.4**	**9.0**

both speakers in the mixture, and we apply WebRTC[1] voice activity detector to exclude non-speech frames when calculating accuracy.

5. EVALUATION RESULTS

We begin by assessing the frame-level and utterance-level performance of the MIMO complex spectral mapping for speaker localization. For utterance-level localization, we aggregate the GCC coefficients across all frames to determine the DOA for each speaker. When utilizing a MISO model to derive these coefficients, we perform separation multiple times by rotating the microphone order. This trick only works for uniform circular arrays; for non-circular arrays, a dedicated MISO model must be trained for each microphone. For the masking-based GCC-PHAT, we followed [18] and applied an ideal ratio mask [29] as a weighting term to the GCC coefficients, using the multi-channel mixture signals in Eq. (2). As Table 1 shows, the MIMO separation model achieves an impressive 95% frame-level accuracy, outperforming both the MISO separation model and masking-based GCC-PHAT. This superior performance is attributed to the MIMO models' ability to simultaneously estimate the target signal at all microphones, thereby preserving inter-channel phase relations for more accurate DOA estimation. Not surprisingly, all three methods perform similarly well in utterance-level localization, as summing over all frames significantly sharpens the peak in the accumulated GCC coefficients.

In Table 2, we evaluate the separation performance of our proposed method for both utterance-wise and continuous ASR evaluations using the LibriCSS dataset. For a comprehensive comparison, we report other competitive separation systems evaluated with the default ASR of LibriCSS. The systems proposed in [5] and [27] leverage BLSTM and conformer architectures, respectively, for real-valued mask estimation. The MISO-BF-MISO system from [8] employs MISO complex spectral mapping via a UNet model. It incorporates an additional beamformer and MISO enhancement model for post-filtering. Moreover, a dedicated SC network is utilized specifically for stream merging. This SC network performs frame-wise three-class classification—counting 0, 1, or 2 speakers—and is trained using cross-entropy. Segments identified as single-talker are then merged, and the weaker stream is suppressed by multiplying it with a small constant. In contrast, the system in [11] mirrors MISO-BF-MISO but utilizes MIMO complex spectral mapping for separation and excludes the SC network. To ensure a fair comparison, we also trained a MISO-BF-MISO system using TF-GridNet. For this TF-GridNet-based system, we observed that applying SC leads to significant improvements in lower overlap ratio conditions, which mainly consist of single-talker segments, albeit with slight degradation under higher overlap conditions.

We also explored the efficacy of our proposed stream merging method by training a MISO-BF-MIMO model that incorporates an enhancement model with MIMO complex spectral mapping. We found that both the localization method and the SC method perform similarly. This suggests that comparable performance can be achieved without the need to train a separate SC network, simply by using the localization method. Additionally, for MIMO separation models that do not include post-filtering, our proposed stream merging method enhances performance to a large extent. Overall, we achieved state-of-the-art results with a TF-GridNet based MIMO-BF-MISO model that uses stream merging based on the localization method.

6. CONCLUSIONS

We have proposed a novel approach to improve the performance of continuous speaker separation. We utilize MIMO complex spectral mapping to estimate the target signal at all microphones simultaneously, which retains inter-channel phase relations for accurate DOA estimation. We then leverage the retained phase information from all microphones for speaker localization with GCC-PHAT. We have demonstrated that MIMO complex spectral mapping yields excellent localization performance. We have further shown that our proposed stream merging method improves continuous speaker separation performance, advancing state-of-the-art performance on the LibriCSS dataset.

[1] Available at: https://github.com/wiseman/py-webrtcvad

7. REFERENCES

[1] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming Multi-Talker ASR with Token-Level Serialized Output Training," in *Proc. Interspeech*, 2022, pp. 3774–3778.

[2] I. Sklyar, A. Piunova, X. Zheng, and Y. Liu, "Multi-turn RNN-T for streaming recognition of multi-party speech," in *Proc. ICASSP*, 2022, pp. 8402–8406.

[3] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *IEEE Spoken Language Technology Workshop*, 2021, pp. 897–904.

[4] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. L. Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *arXiv:2303.03849*, 2023.

[5] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*, 2020, pp. 7284–7288.

[6] T. v. Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Segment-less continuous speech separation of meetings: Training and evaluation criteria," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 576–589, 2023.

[7] Y. Zhang, Z. Chen, J. Wu, T. Yoshioka, P. Wang, Z. Meng, and J. Li, "Continuous speech separation with recurrent selective attention network," in *Proc. ICASSP*, 2022, pp. 6017–6021.

[8] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2001–2014, 2021.

[9] Z.-Q. Wang and D. L. Wang, "Count and separate: Incorporating speaker counting for continuous speaker separation," in *Proc. ICASSP*, 2021, pp. 11–15.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Process.*, vol. 24, pp. 320–327, 1976.

[11] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. L. Wang, "Multi-input multi-output complex spectral mapping for speaker separation," in *Proc. Interspeech*, 2023, pp. 1070–1074.

[12] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, pp. 107–151, 2022.

[13] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. WASPAA*, 2017, pp. 136–140.

[14] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, p. 101360, 2022.

[15] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "High-resolution speaker counting in reverberant rooms using CRNN with ambisonics features," in *Proc. EUSIPCO*, 2021, pp. 71–75.

[16] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2626–2637, 2020.

[17] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. ICASSP*, 2017, pp. 6125–6129.

[18] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, pp. 178–188, 2019.

[19] K. Tan, Z.-Q. Wang, and D. L. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 605–621, 2022.

[20] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated RNN for binaural speaker separation with interaural cue preservation," *IEEE Signal Processing Letters*, vol. 28, pp. 26–30, 2021.

[21] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[22] H. Taherian and D. L. Wang, "Multi-resolution location-based training for multi-channel continuous speech separation," in *Proc. ICASSP*, 2023, pp. 1–5.

[23] 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 Trans. Audio Speech Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[24] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, pp. 692–730, 2017.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[27] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. ICASSP*, 2021, pp. 5749–5753.

[28] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv:1910.13934*, 2019.

[29] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, pp. 1849–1858, 2014.