



# End-to-End DOA-Guided Speech Extraction in Noisy Multi-Talker Scenarios

Kangqi Jing<sup>1†</sup>, Wenbin Zhang<sup>2†</sup>, Yu Gao<sup>2\*</sup>

<sup>1</sup>School of Information Science and Engineering, Southeast University, China

<sup>2</sup>AI Research Center, Midea Group (Shanghai) Co.,Ltd., China

jing kangqi@seu.edu.cn, {zhangwb87, gaoyu11}@midea.com

## Abstract

Target Speaker Extraction (TSE) plays a critical role in enhancing speech signals in noisy and multi-speaker environments. This paper presents an end-to-end TSE model that incorporates Direction of Arrival (DOA) and beamwidth embeddings to extract speech from a specified spatial region centered around the DOA. Our approach efficiently captures spatial and temporal features, enabling robust performance in highly complex scenarios with multiple simultaneous speakers. Experimental results demonstrate that the proposed model not only significantly enhances the target speech within the defined beamwidth but also effectively suppresses interference from other directions, producing a clear and isolated target voice. Furthermore, the model achieves remarkable improvements in downstream Automatic Speech Recognition (ASR) tasks, making it particularly suitable for real-world applications.

**Index Terms:** directional speech extraction, multichannel, embedding

## 1. Introduction

In complex auditory environments with multiple sound sources, humans are able to selectively focus on a specific sound, a phenomenon commonly known as the “cocktail party effect”. This ability allows us to attend to a target sound using various cues, such as its time-frequency pattern or the direction of arrival (DOA) [1]. Target Sound Extraction (TSE) aims to replicate this selective attention mechanism by isolating and enhancing a desired sound source, typically human speech, from a noisy and reverberant background. This task is critical for applications in fields such as conference systems and hearing aids [2, 3]. The primary challenges in TSE include background noise, speech interference, and reverberation, all of which can severely degrade the quality and intelligibility of the target speech [4]. Among these, speech interference, especially from competing speakers, poses the greatest difficulty. Speech recognition systems, in particular, struggle to distinguish the target speech from the interference, resulting in a significant drop in performance.

Recent advancements in Deep Neural Networks (DNNs) have made significant strides in overcoming the challenges of target sound extraction. These developments have proven especially effective in applications such as speech separation [5, 6], enhancement [7], and DOA estimation [8]. In multichannel audio setups, DOA provides a crucial cue for locating the target speaker. This technique, referred to as Directional Speech Extraction (DSE), aims to isolate speech from a fixed [9, 10] or dynamically adjustable [11, 12] spatial region. Recently, DSE

approaches [13, 14] have been further refined to leverage DOA information from multichannel mixtures, where the DOA cue is used to guide the extraction process.

However, most existing models for TSE rely heavily on accurate DOA information to achieve optimal performance [15]. When the DOA estimates are inaccurate or ambiguous, these models often fail to clearly define the spatial region for sound extraction, resulting in unpredictable outputs that may include both target speech and interfering noise. This limitation significantly degrades their effectiveness in real-world scenarios where precise DOA information is not always available. The CDUNet model [4] proposing a directional TSE approach that integrates beamforming and DNNs for two-microphone and two-speaker setups. This model introduces the enhancement width as an input parameter to control the spatial region for speech enhancement. While innovative, CDUNet lacks generalization to more complex, multi-speaker environments.

In this work, we propose an end-to-end TSE model that leverages DOA embedding to extract speech within a selected beamwidth centered around the DOA. Our model introduces advanced voice focusing and zooming capabilities, enabling robust performance in scenarios involving multiple speakers. By efficiently extracting target speech signals from different directions, our approach significantly improves the performance of downstream ASR tasks. This makes it particularly suitable for real-world applications such as conference systems and smart assistants, where accurate and reliable speech extraction in highly overlapping multi-speaker environments is critical. Code and audio examples are available at <https://github.com/jingkangqi/DSENet>.

## 2. Method

### 2.1. Problem formulation

Let  $\mathbf{X} \in \mathbb{R}^{M \times L}$  be the multichannel mixture recorded by an  $M$ -microphone array, where  $L$  denotes the number of time samples. The signal at the  $m$ -th microphone can be expressed as:

$$\mathbf{X}_m(t) = \sum_{i=1}^N s_i(t) * \mathbf{H}_i^m(t) + \mathbf{N}_m(t), \quad (1)$$

where  $s_i(t) \in \mathbb{R}^L$  represents the clean speech signal of the  $i$ -th speaker,  $\mathbf{H}_i^m(t)$  denotes the room impulse response (RIR) from the  $i$ -th speaker to the  $m$ -th microphone, and  $\mathbf{N}_m(t)$  represents the background noise at the  $m$ -th microphone.

Given a target DOA  $\theta_{\text{target}}$  and a beamwidth  $\theta_{\text{beam}}$ , our model aims to extract speech signals  $s_c$  exists in  $\theta_{\text{beam}} = [\theta_{\text{target}} - \theta_{\text{width}}, \theta_{\text{target}} + \theta_{\text{width}}]$ , while suppressing interference from other directions outside the beam. The selected beamwidth  $\theta_{\text{beam}}$  may contain more than one speaker, in which case the model will

† Equal contribution.

\* Corresponding author.

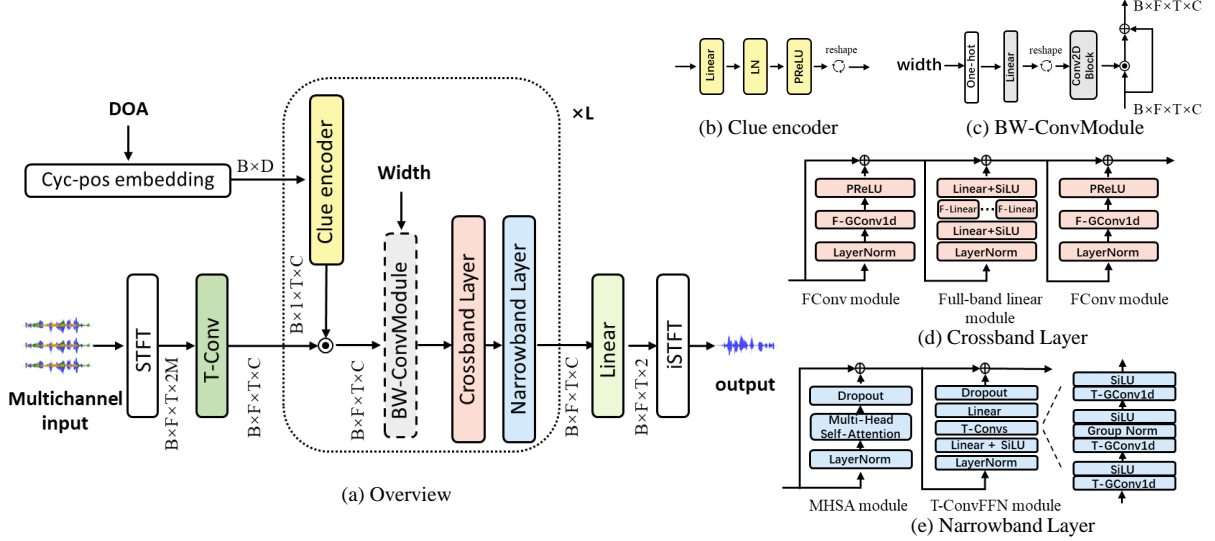


Figure 1: (a) The system overview of the proposed DSE model. The detailed structure of (b) Clue encoder, (c) BW-ConvModule, (d) Crossband Layer and (e) Narrowband Layer.

retain and output all voices within this range. If no speech is present in the range, the model outputs a signal that is either close to silence or silent.

## 2.2. Model Architecture

The overview of the proposed architecture is depicted in Figure 1. First, a multichannel input is converted into a  $2M \times T \times F$  complex spectrogram by short-time Fourier transform (STFT), where  $M$  denotes the number of microphone,  $T$  and  $F$  denote the number of time and frequency bins, respectively. The spectrogram is then processed by a convolutional input layer, expanding the channel dimension to  $C$  while encoding spatial features and local time-frequency information. The feature sequence is then processed through  $L$  stacked layers. In each layer, the clue encoder refines the DOA-related features for spatial selectivity and target extraction. The BW-ConvModule limits the beamwidth around the DOA, dynamically restricting the spatial region to improve focus on target speech. The Crossband blocks capture spectral dependencies by modeling frequency-wise correlations, while the Narrowband blocks focus on temporal patterns within each frequency band. The final layer's output is projected to target STFT coefficients via linear transformation, with iSTFT reconstructing the enhanced speech signal  $\hat{s}_c$ .

### 2.2.1. DOA-based clue embedding and beamwidth control

While the one-hot vector provides a unique representation for each direction, it fails to capture the inherent periodicity, resulting in an abrupt transition from  $359^\circ$  to  $0^\circ$ . Moreover, the one-hot vector requires a high-dimensional encoding (360 dimensions), which is computationally inefficient. To address these issues, we adopt cyclic positional (cyc-pos) encoding [16], which effectively preserves the continuity of direction while reducing the dimensionality. The cyc-pos vector  $\mathbf{PE}_{\text{cyc-pos}}(\phi) \in \mathbb{R}^D$  for an embedding dimension of  $D$  is defined as:

$$\begin{aligned} \mathbf{PE}_{\text{cyc-pos}}(\phi, 2j) &= \sin\left(\sin(\phi) \cdot \frac{\alpha}{10000^{2j/D}}\right) \\ \mathbf{PE}_{\text{cyc-pos}}(\phi, 2j+1) &= \sin\left(\cos(\phi) \cdot \frac{\alpha}{10000^{2j/D}}\right) \end{aligned} \quad (2)$$

where  $j \in [0, \frac{D}{2})$ , and  $\alpha$  is a scaling factor.

The DOA embedding is then broadcast along the time dimension and processed through the clue encoder, as illustrated in Figure 1(b). The clue encoder comprises a linear layer, followed by layer normalization (LN) [17] and parametric rectified linear unit (PReLU) [18]. The encoded DOA embedding is applied via element-wise multiplication with the output of the T-Conv and Narrowband layers excluding the final layer.

The BW-ConvModule shown in Figure 1(c), dynamically adjusts the beamwidth by processing a width input through a one-hot encoding, a Linear layer, and a  $1 \times 1$  Conv2D to generate a mask. This mask filters out noise outside the desired beamwidth, while residual connections preserve essential information for robust speech extraction.

### 2.2.2. Crossband and Narrowband layers

The Crossband and Narrowband layers have the same structure as that described in [19], designed to learn complex spatial information, as shown in Figure 1(d)(e).

The Crossband layer comprises two frequency-convolutional modules and a full-band linear module, processing each time frame independently. The F-GConv1d module utilizes grouped frequency convolutions to model local spectral dependencies. The full-band linear module first compresses channels ( $C \rightarrow C'$ ), followed by a set of frequency-wise linear layers. To improve parameter efficiency, the same F-Linear networks are shared across all instances of the Crossband layer. Finally, a linear layer restores the channel dimension to  $C$ .

The Narrowband layer captures temporal dependencies by processing each frequency independently. The multihead self-attention (MHSA) [20] module computes spatial similarities within each frequency, facilitating the separation of speech components originating from different directions. The time-convolutional feedforward network (T-ConvFFN) enhances temporal modeling via a sequence of operations: a linear layer expands the hidden dimension from  $C$  to  $C''$ , followed by grouped 1-D convolutions along the time axis with group nor-

malization [21], before a final linear layer restores the original dimensionality  $C$ .

The Crossband and Narrowband layers are interleaved to enhance the model’s ability to differentiate and extract target signals while effectively noise, reverberation, and interference.

### 2.3. Loss Function

The proposed model is trained using a combination of a spectral magnitude loss and a scale-invariant signal-to-distortion ratio (SI-SDR) [22] loss. The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{Mag}} + \lambda \mathcal{L}_{\text{SI-SDR}}, \quad (3)$$

where  $\lambda$  is a weighting factor. The spectral magnitude loss, similar to [6], enforces consistency in the TF domain and is defined as:

$$\mathcal{L}_{\text{Mag}} = \frac{\| |\text{STFT}(\hat{s}_c)| - |\text{STFT}(s_c)| \|_1}{\| |\text{STFT}(s_c)| \|_1}. \quad (4)$$

To improve the perceptual quality of the extracted speech, we incorporate the SI-SDR loss:

$$\mathcal{L}_{\text{SI-SDR}} = - \sum_{c=1}^C 10 \log_{10} \left( \frac{\|s_c\|_2^2}{\|\hat{s}_c - \alpha_c s_c\|_2^2} \right), \quad (5)$$

where  $\alpha_c$  is the optimal scaling factor given by

$$\alpha_c = \frac{s_c^T \hat{s}_c}{s_c^T s_c}. \quad (6)$$

## 3. Experiment

### 3.1. Dataset

For this study, the speech data is sourced from the LibriSpeech corpus [23], while the background noise is taken from the DEMAND dataset [24]. The room acoustics, including reverberation and microphone array configuration, are simulated using the `pyroomacoustics`<sup>1</sup> [25] package. The room’s width and depth are randomly sampled from the range [6, 9] m, with a fixed height of 3 m. The reverberation time (RT60) of the room is randomly varied within the range [0.3, 0.5] s. A 3-channel circular microphone array with a radius of 30 mm is placed at the center of each room, at a height of 1 m from the floor.

Each audio mixture contains speech from six simultaneous speakers combined with an independent noise source. All sources are randomly placed in the room, at least 0.3 m away from the walls. The Root-Mean-Square (RMS) level of each mixed signal is randomly sampled from [-20, -15] dB. All audio mixtures are 4 second-long, with a sampling rate of 16 kHz. The training, validation, and test sets contain 14.4k, 3.6k, and 2k mixtures, respectively.

### 3.2. Implementation Details

The network consists of  $L = 8$  blocks, with hidden units configured as  $C = 192$ ,  $C' = 8$ , and  $C'' = 192$ . The kernel sizes of TConv1d, T-GConv1d, and F-GConv1d are set to 5, 5, and 3, respectively, with all group numbers set to 8.

The input STFT uses a window length of 16 ms (256 samples) and a hop length of 8 ms (128 samples), employing a Hanning window and 129 FFT frequency bins. Training batch size is 4, using the Adam [26] optimizer with an initial learning rate of 0.001, which decays exponentially by a factor of 0.99 per epoch. The training is conducted in two stages:

<sup>1</sup><https://github.com/LCAV/pyroomacoustics>

- **Stage 1: Target Extraction (100 epochs)** Using DOA inputs from active speech sources, with the BW-ConvModule **disabled**. The SI-SDR loss weight is set to  $\lambda = 0.5$ .
- **Stage 2: Beam Adaptation (50 epochs)** The beamwidth input to BW-ConvModule is selected from one of  $15^\circ$ ,  $30^\circ$ , or  $45^\circ$ . For each selection, 90% DOA inputs are active ( $\theta_{\text{beam}}$  containing speech sources), while the remaining 10% are inactive ( $\theta_{\text{beam}}$  without speech sources). In inactive cases, the model generates a small, fixed signal instead of complete silence or noise. Silent periods contribute no meaningful information to SI-SDR calculation, and may cause training instability or large loss fluctuations. Here, we use a 20Hz reference tone (-60 dB RMS). The SI-SDR loss weight is decayed to  $\lambda = 0.05$ .

## 4. Results and Analysis

### 4.1. Ablation Study and Comparison results

To evaluate our proposed method’s effectiveness, we compute three metrics: SDR improvement (SDRi) [27], SI-SDR improvement (SI-SDRi), and perceptual evaluation of speech quality (PESQ) [28]. We also analyzed the parameter count (Para.) to assess computational efficiency. For evaluation fairness and simplicity, we use the model trained in Stage 1, where the BW-ConvModule is disabled, and set the input DOA such that only one speaker falls within the beam width.

Table 1: Performance comparison of DSE Models and clues

Model	Emb.	D	$\alpha$	Para.	Evaluation Metrics $\uparrow$		
					SI-SDRi	SDRi	PESQ
Noisy					-17.30	-9.56	1.09
MVDR	-	-	-	-	-1.62	-1.27	1.09
JNF	one-hot	360	-	1.34M	10.22	8.04	1.10
Proposed	one-hot	360	-	1.89M	16.21	12.39	1.31
	cyc-pos	40	10	1.40M	16.60	12.31	1.29
			40		<b>18.29</b>	<b>13.99</b>	<b>1.40</b>
					14.82	11.19	1.25

We conducted an ablation study to evaluate the impact of different DOA embedding types. The results in Table 1 demonstrate that cyclic positional encoding vectors outperform one-hot encoding. Specifically, the one-hot encoding achieves an SI-SDRi of 16.21 dB with 1.89M parameters, while the best-performing cyc-pos configuration attains an SI-SDRi of 18.29 dB with only 1.40M parameters. The smooth and periodic nature of cyc-pos vectors across angles facilitates better integration into spatial feature representations, while simultaneously reducing the parameter count. However, when the scaling factor  $\alpha$  is set too large, performance degrades due to the loss of smoothness and the introduction of rapid variations.

For comparison, we selected the minimum variance distortionless response (MVDR) filter [29] and the Joint Spatial and TempoSpectral Non-linear Filter (JNF) [30] as baseline models. JNF is a deep neural network (DNN)-based spatially selective filter (SSF) that leverages a recurrent neural network (RNN) layer initialized with the target direction to spatially steer and extract the speaker of interest. Due to the high complexity of the scenario involving up to six highly overlapping speakers, the MVDR method, which serves as a conventional benchmark, fails to effectively handle such an extremely challenging environment, resulting in degraded performance, which highlights the difficulty of the task. JNF, which can be attributed to a better use of the spatial information achieve advanced results in Table I. However, with cyc-pos vectors configured at  $D = 40$

and  $\alpha = 20$ , our model significantly outperforms JNF, achieving higher SNRi, SI-SNRi, and PESQ scores while maintaining a comparable parameter count.

## 4.2. Gain-pattern Analysis

To facilitate the understanding of the complex spatial patterns incorporates reverberation effects, the methodology proposed in [31] introduces a simplified 1D representation of the array responses. In this section, we further explore the spatial properties of our method by plotting the gain pattern.

The gain pattern is empirically generated by moving a speaker along a circular trajectory around the fixed microphone array, evaluating the model's response to the reverberant signal as a function of DOA. Let the input and output signals be denoted by  $\mathbf{x}(t)$  and  $y(t)$ , respectively. The power of the dry speech component of  $\mathbf{x}(t)$  in reference channel is defined as  $P_{in} = \mathbb{E}[|x_{dry}(t)|^2]$ , where  $\mathbb{E}[\cdot]$  denotes the expected value. The power of the output signal is  $P_{out} = \mathbb{E}[|y(t)|^2]$ . The gain in a given direction is computed as:

$$Gain \text{ (dB)} = 10 \log_{10} \left( \frac{P_{out}}{P_{in}} \right) \quad (7)$$

We compute the gain for each direction and plot the resulting gain map in Figure 2. The model significantly enhances the target speech signal within the selected beamwidth centered around the input DOA, with the gain in the target region being substantially higher while the gain from other directions remains below  $-20$  dB, effectively suppressing interference sources. The results demonstrate the model's capacity to simultaneously utilize both DOA and beamwidth inputs. Notably, a spillover of approximately  $5^\circ$ - $10^\circ$  is observed, which is more pronounced in wider beamwidths (e.g.,  $45^\circ$ ) compared to narrower ones (e.g.,  $15^\circ$ ). This spillover is an acceptable trade-off given the model's ability to maintain effective speech extraction within the target region. The beamwidth explicitly defines the

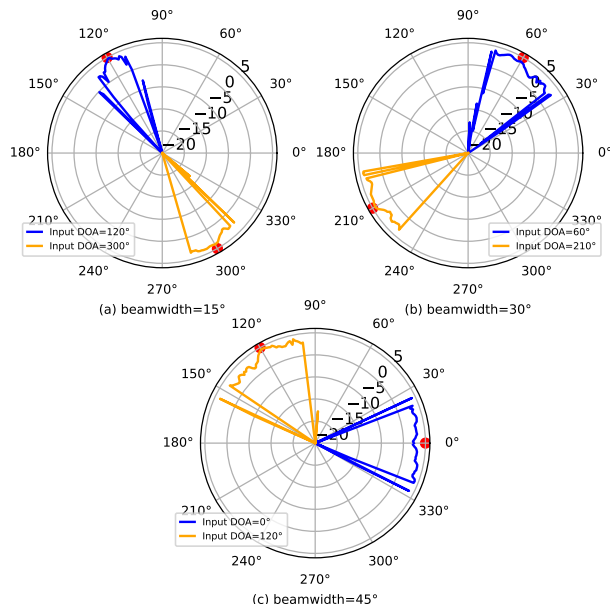


Figure 2: Gain-pattern analysis for different input beamwidths: (a)  $15^\circ$ , (b)  $30^\circ$ , and (c)  $45^\circ$ . Two random DOA inputs are tested for each beamwidth, with the input DOA indicated by a red dot.

speech extraction range and offers a tolerance for DOA estimation errors, making it more aligned with practical requirements.

## 4.3. Effective Speaker Extraction Capability

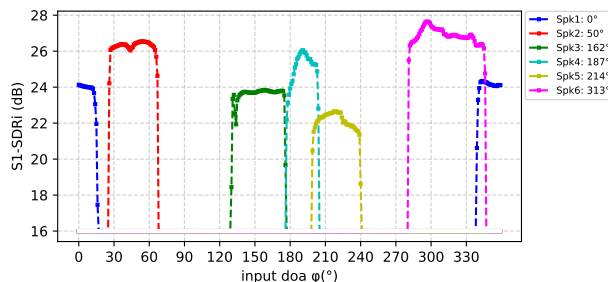


Figure 3: SI-SDRi for six speakers under various input DOAs with a fixed beamwidth of  $30^\circ$ . The DOAs of the six speakers (Spk1-Spk6) are:  $0^\circ$ ,  $50^\circ$ ,  $162^\circ$ ,  $187^\circ$ ,  $214^\circ$ , and  $313^\circ$ .

This work primarily evaluates the model's ability to extract clear speech signals corresponding to different speakers under varying input DOAs. To this end, we calculate the SI-SDRi for each speaker based on the model's output, with the beamwidth input fixed at  $30^\circ$ . As shown in Figure 3, when a speaker's DOA falls within the beamwidth, the corresponding SI-SDRi is substantially improved. As the input DOA varies, the model adaptively extracts speech from different speakers while effectively suppressing interference from other directions.

## 4.4. Downstream ASR Performance

For evaluating ASR performance, we additionally generated two datasets with different numbers of speakers, each comprising 1,000 samples. During testing, the input DOA was adjusted to correspond to the range of active sound sources.

Table 2: WER (%) results on different dataset

Mixed	Noisy	JNF	Proposed
2spk	82.10	38.06	<b>10.52</b>
3spk	96.52	57.04	<b>17.31</b>

Table 2 shows the Word Error Rate (WER) results. Our model achieves significantly lower WER compared to the JNF method, particularly as the number of speakers increases. This performance advantage underscores the model's potential for real-world applications involving multiple speakers, such as smart home devices, where it can accurately capture user commands amidst background noise.

## 5. Conclusion

In this work, we propose a novel end-to-end target speaker extraction model that leverages DOA and beamwidth as soft constraints to form an adaptive neural beam, dynamically focusing on the target speech even in highly complex multi-speaker environments. By integrating DOA and beamwidth embeddings, our approach efficiently captures spatial and temporal features, enabling robust performance in scenarios with significant speaker overlap and background noise. In future work, we will explore the deployment of this model on edge devices to further enhance its practicality in real-world scenarios.

## 6. References

- [1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.
- [2] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.
- [3] K. Patterson, K. Wilson, S. Wisdom, and J. R. Hershey, "Distance-based sound separation," in *Interspeech 2022*, 2022, pp. 901–905.
- [4] W. Wen, Q. Zhou, Y. Xi, H. Li, Z. Gong, and K. Yu, "Neural directed speech enhancement with dual microphone array in high noise scenario," *arXiv preprint arXiv:2412.18141*, 2024.
- [5] S. Wang, X. Kong, X. Peng, H. Movassagh, V. Prakash, and Y. Lu, "Dasformer: Deep alternating spectrogram transformer for multi/single-channel speech separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [6] 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 Transactions on Audio, Speech, and Language Processing*, 2023.
- [7] D. Lee and J.-W. Choi, "Deft-an: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.
- [8] Z. Li, S. He, and X. Zhang, "Robust target speaker direction of arrival estimation," *arXiv preprint arXiv:2412.18913*, 2024.
- [9] A. Kovalyov, K. Patel, and I. Panahi, "Dsenet: Directional signal extraction network for hearing improvement on edge devices," *IEEE Access*, vol. 11, pp. 4350–4358, 2023.
- [10] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2022.
- [11] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," in *Proc. Interspeech 2020*, 2020, pp. 56–60.
- [12] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2023.
- [13] R. Gu and Y. Luo, "Rezero: Region-customizable sound extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [14] A. Pandey, S. Lee, J. Azcarreta, D. Wong, and B. Xu, "All neural low-latency directional speech extraction," in *Proc. Interspeech 2024*, 2024, pp. 4328–4332.
- [15] C. Rascon, "Direction of arrival correction through speech quality feedback," *Digital Signal Processing*, vol. 158, p. 104960, 2025.
- [16] H. Lee, C. Homeyer, R. Herzog, J. Rexilius, and C. Rother, "Spatio-temporal outdoor lighting aggregation on image sequences using transformer networks," *International Journal of Computer Vision*, vol. 131, no. 4, pp. 1060–1072, 2023.
- [17] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv e-prints*, pp. arXiv–1607, 2016.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [19] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.
- [20] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
- [21] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [22] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.
- [25] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.
- [26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.
- [27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [29] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.
- [30] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [31] O. Shmaryahu and S. Gannot, "On the importance of acoustic reflections in beamforming," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.