ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890145

Multichannel-to-Multichannel Target Sound Extraction Using Direction and Timestamp Clues

Dayun Choi
School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
cdy3773@kaist.ac.kr

Jung-Woo Choi
School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
jwoo@kaist.ac.kr

***Abstract*—We propose a multichannel-to-multichannel target sound extraction (M2M-TSE) framework for separating multichannel target signals from a multichannel mixture of sound sources. Target sound extraction (TSE) isolates a specific target signal using user-provided clues, typically focusing on single-channel extraction with class labels or temporal activation maps. However, to preserve and utilize spatial information in multichannel audio signals, it is essential to extract multichannel signals of a target sound source. Moreover, the clue for extraction can also include spatial or temporal cues like direction-of-arrival (DoA) or timestamps of source activation. To address these challenges, we present an M2M framework that extracts a multichannel sound signal based on spatio-temporal clues.**

We demonstrate that our transformer-based architecture can successively accomplish the M2M-TSE task for multichannel signals synthesized from audio signals of diverse classes in different room environments. Furthermore, we show that the multichannel extraction task introduces sufficient inductive bias in the DNN, allowing it to directly handle DoA clues without utilizing hand-crafted spatial features.

***Index Terms*—target sound extraction, multichannel extraction, directional clue, timestamps, complex spectral mapping.**

I. Introduction

Humans naturally focus on a specific sound in complex auditory environments with multiple sound sources. This ability allows us to attend to a target sound using clues like its time-frequency pattern or direction [1]. Target sound extraction (TSE) aims to mimic this by extracting the desired sound source using various types of clues. Common clues include class-labels [2], [3], a signal resembling the target [4], images or videos [5], timestamps marking target occurrences [6], [7], text descriptions [8], [9], directions or regions indicating the locations of targets [10], and combinations of these clues [11]–[14].

However, these methods primarily focus on extracting a single-channel target signal. Multichannel mixtures, often recorded using microphone arrays, include spatial characteristics of a sound field. In applications like 3-D audio and virtual reality (VR) audio, interchannel relationships provide important spatial cues for rendering realistic sound. Similarly, in acoustic surveillance systems, interchannel time delays or phase differences are key to determining the direction or location of a target sound source. To fully exploit the spatial information in multichannel recordings, TSE should extract the multichannel source signal as if the microphone array had recorded the target sound alone.

There have been many DNN models designed to capture the spatial features and interchannel relations from multichannel sound sources, especially in speech separation or enhancement [15]–[21] and direction-of-arrival (DoA) estimation [22], [23]. More recently, directional speech extraction [10], [24] has also been proposed to extract speech from a multichannel mixture using its DoA clue. To incorporate DoA clue, its interchannel correlation features are extracted from a complex spectrogram of the mixture or individual one-hot embedding for each channel is integrated into spectral features. Nevertheless, these models utilize interchannel information to extract a single-channel clean sound source or to determine the DoA of the sound source by removing reverberations and noises.

Examples of separating a multichannel source signal can be found in binaural or multichannel speech enhancement and TSE approaches [25]–[29]. These approaches have demonstrated the potential of multichannel-to-multichannel (M2M) extraction. However, the binaural speech enhancement model [26] focuses exclusively on the speech target, limiting its application to a specific source type. Binaural TSE [28] overcomes this limitation and also utilizes spatial losses to minimize the degradation in binaural cues. Despite these advances, the primary clue for extraction remains the class-label related only to the time-frequency (TF) characteristics. In complex mixtures of signals from the same class, i.e., such as multiple instruments or speech sources, the direction of sound or timestamps of activation becomes compelling clues for extraction.

To this end, we propose an M2M-TSE framework capable

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337945) and STEAM research grant (No. RS-2024-00337945) funded by the Ministry of Science and ICT of Korea government (MSIT), the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE).

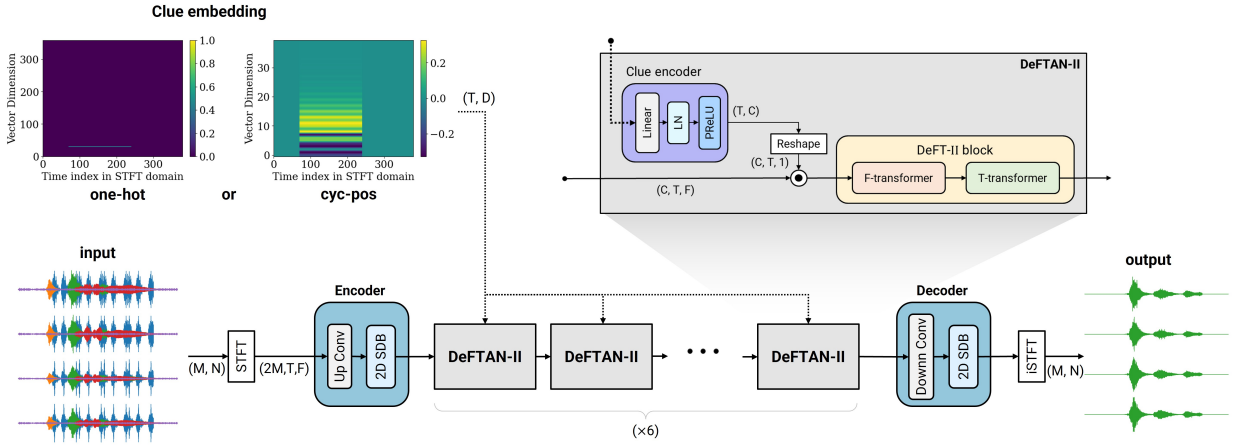

Fig. 1. Model architecture for DoA-based multichannel-to-multichannel target sound extraction (M2M-TSE).

of extracting multichannel sounds from a complex, reverberant mixture using directional clues and timestamp information. We utilize dense frequency-time attentive network II (DeFTAN-II) [21] architecture as a backbone network, which delivers high performance with low time and space complexity in multichannel speech enhancement tasks. The key contribution of this work is the modified model designed to incorporate directional and timestamp clues for extracting multichannel sound sources of various types, rather than a single-channel speech. We demonstrate that the M2M extraction task encourages the model to capture and aggregate spatial features effectively, enabling efficient extraction from simple directional and temporal embeddings without requiring additional spatial input features such as intensity vectors or interchannel correlation features.

II. Proposed Methods

A. Problem Statement

Our research focuses on extracting an M-channel reverberant sound signal $\mathbf{X}_i \in \mathbb{R}^{M\times N}$ of temporal length N, from an input mixture $\mathbf{Y} \in \mathbb{R}^{M\times N}$ of I source signals captured from a microphone array in a reverberant room. The input mixture can be described as

$$\mathbf{Y} = \sum_{i=1}^{I} \mathbf{X}_i + \mathbf{V} = \sum_{i=1}^{I} \mathbf{s}_i * \mathbf{R}_i + \mathbf{V}, \tag{1}$$

where $\mathbf{s}_i \in \mathbb{R}^N$ and $\mathbf{R}_i \in \mathbb{R}^{M\times N}$ are the i-th dry source signal and its M-channel room impulse response (RIR), respectively. Here, $\mathbf{V} \in \mathbb{R}^{M\times N}$ is the multichannel measurement noise, and $*$ indicates the temporal convolution. When the target sound is the g-th source $\mathbf{X}_g$ ($g \in \{1, \cdots, I\}$), the multichannel signal $\hat{\mathbf{X}}_g \in \mathbb{R}^{M\times N}$ extracted from a TSE model with model parameters θ can be written as

$$\hat{\mathbf{X}}_g = \mathrm{TSE}(\mathbf{Y}, \mathbf{C}_g; \theta), \tag{2}$$

where $\mathbf{C}_g$ is a clue embedding derived from DoA and temporal activity (timestamps) of the target source. In this work, we consider the clue embedding $\mathbf{C}_g$ given by either a one-hot or a cyclic positional vector, as described in section II-C.

B. Model Architecture

The backbone of the proposed network is the DeFTAN-II [21] architecture. DeFTAN-II is a transformer-based architecture that performs complex spectral mapping to extract a single-channel clean speech with suppressed noise and reverberation. Our TSE objective is similar, but the main difference is that a multichannel target signal should be extracted and a clue for identifying the target signal should be injected into the network. To achieve these, we introduce the following modifications to the backbone network.

The overview of the proposed architecture is depicted in Fig. 1 without the batch dimension. First, a multichannel input waveform is converted into a complex spectrogram of dimensions $2M \times T \times F$ by short-time Fourier transform (STFT), where T and F denote the number of time and frequency bins, respectively. This spectrogram is then transformed to a tensor with increased channel dimension (C) using a 2-dimensional split dense block (2D SDB) encoder. 2D SDB is a modified version of DenseNet [30] introduced for extracting spatial features and learning local spectral-temporal relations. Since the channel dimension includes encoded spatial features and local time-frequency (TF) information, we combine the clue embedding with the channel dimension of the encoded tensor to extract the target sound of a specific direction.

In a series of DeFTAN-II blocks, the F- and T-transformers analyze the relationships in spectral and temporal sequences. We repeatedly embed the clue during this stage, to gradually align the features developed by DeFTAN-II blocks with those of the target sound. Finally, the aligned features are decoded to a multichannel waveform through a decoder reducing the channel dimension from C to $2M$ as in the original STFT, and the inverse STFT (iSTFT) operation.

C. Spatio-temporal Clues

The clue utilized in this work is the direction and timestamps of a target source. For simplicity, we consider only

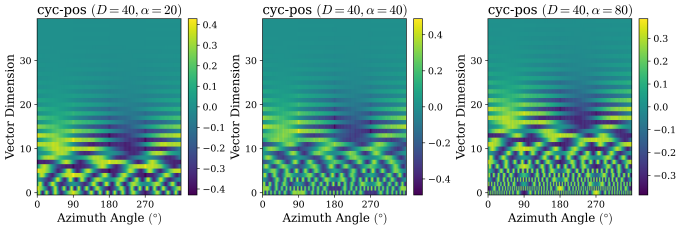

Fig. 2. Clue embeddings given by cyclic positional (cyc-pos) vectors for different scaling factors (α).

azimuth angles as the direction clue and encode them in two different ways: one-hot or cyclic positional encoding. For the one-hot encoding, the clue embedding vector $\mathbf{1}_{\text{one-hot}}(\phi) \in \mathbb{R}^{360}$ defined for direction $\phi \in [0, 360)$ with 1° resolution has a value of one only in the index corresponding to the direction ϕ, with all other values being zero. That is, for the index $j \in [0, 360)$,

$$\mathbf{1}_{\text{one-hot}}(\phi, j) = \begin{cases} 1 & \text{where } j = \phi, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

While the one-hot vector contains unique information for each direction, it cannot represent periodicity correctly, showing the abrupt transition from 359° to 0°. To address this problem, we employ the cyclic positional (cyc-pos) encoding [31]. The cyc-pos vector $\mathbf{PE}_{\text{cyc-pos}}(\phi) \in \mathbb{R}^D$ for embedding dimension D can be represented as:

$$\begin{aligned} \mathbf{PE}_{\text{cyc-pos}}(\phi, 2j) &= \sin(\sin(\phi) \cdot \frac{\alpha}{10000^{2j/D}}), \\ \mathbf{PE}_{\text{cyc-pos}}(\phi, 2j+1) &= \sin(\cos(\phi) \cdot \frac{\alpha}{10000^{2j/D}}), \end{aligned} \tag{4}$$

where $j \in [0, \frac{D}{2})$, and α is the scaling factor controlling the angular range utilized for the positional encoding. In Fig. 2, examples of cyclic positional encoding are presented across various α. Both D and α are hyper-parameters determined empirically, and in this work, $D = 40$ and $\alpha = 20$ were selected from the parameter study. The generated positional embedding is normalized by its L2 norm for each direction.

To further reduce ambiguity in signal extraction, the encoded positional embedding vector is combined with the timestamp clue. The timestamp indicates the occurrences of a target signal, so the positional embedding is broadcasted along the time dimension, such that rows of the final embedding matrix $\mathbf{C}_g \in \mathbb{R}^{T \times D}$ are nonzero only when the target source is active. One example of the embedding matrix is shown in the top left corner of Fig. 1. This embedding is encoded by linear layers, followed by layer normalization (LN) [32] and the parametric rectified linear unit (PReLU) [33] activation, and is then multiplied element-wise with the output of the encoder and DeFTAN-II blocks except for the final block, across the channel and the time dimensions.

III. Experiment and Analysis

A. Datasets

Target signals of training and test datasets were collected from the FSD Kaggle 2018 [34] dataset, a set of sound sources with 41 classes. Following a setup similar to that used for generating reverberant speech in [35], a 4-channel circular microphone array with a radius of 10 cm was positioned in cuboid rooms of which width, depth, and height dimensions were randomly sampled from the uniform distribution within the ranges [5, 10] m, [5, 10] m, and [3, 4] m, respectively. The distance between the center of the microphone array and all sources was also randomly varied within the range of [0.75, 2.5] m, and the minimum angle between two sources relative to the array was set to 20°. The RIRs were generated using the image source method implemented in the `pyroomacoustics`[1] [36] library, and the reverberation time (RT60) of each room was varied between [0.2, 1.3] s. All sound sources convolved with RIRs were mixed using the `scaper`[2] [37] library depending on their timestamp, duration, and magnitude in dB. In addition, noise signals obtained from the 1st, 3rd, 5th, and 7th microphone noises in the REVERB challenge [38] dataset were added. All input mixtures were 6-second-long samples, sampled at 8 kHz for fast computation. The numbers of mixtures constituting the training, validation, and test datasets were 12.5K, 5K, and 2.5K, respectively.

B. Implementation Details

All experiments were conducted in PyTorch framework using automatic mixed precision (AMP) training on a GeForce RTX 4090. The training parameters included a batch size of 4, the Adam optimizer with an initial learning rate of 0.0005, multiplied with 0.1 when the scale-invariant signal-to-noise ratio (SI-SNR) [39] of the validation dataset did not increase after 5 consecutive epochs, and gradient norm clipping was set to 0.5 for 100 epochs. The model parameters were the same as a base model in [21] except for the number of output channels of the decoder ($2 \rightarrow 2M$). The loss function for training was the phase-constrained magnitude (PCM) loss [40], which compares the magnitudes of the real and imaginary components of the ground truth and predicted spectrograms for both target sound and noise. The loss was calculated for each channel, averaged across all time-frequency bins, and then averaged over all channels to produce a final value.

C. Analysis of Results

We evaluated the effectiveness of our method by measuring the improvement in SNR (SNRi) and SI-SNR (SI-SNRi) compared to the input mixture. These metrics assess the models' ability to suppress other sounds and noise while preserving the target signal. To assess how well the interchannel relations between microphone pairs were maintained, we also calculated mean-absolute-error (MAE) of interchannel metrics (Spatial Errors), such as ΔILD, ΔIPD, ΔITD, and ΔITD-GCC[3], utilized in previous studies [25], [28]. The comparative results are shown in Table I.

[1]https://github.com/LCAV/pyroomacoustics

[2]https://github.com/justinsalamon/scaper

[3]Average of the absolute differences between the model prediction and ground truth in interchannel level, phase, time differences using simple cross-correlation, and ITD using generalized cross-correlation phase transform (GCC-PHAT) calculated on all microphone pairs, respectively.

TABLE I
PERFORMANCE COMPARISONS ACROSS TSE MODELS AND CLUES.

Model	Type of Clue	Parameters		SNR Metrics ↑		Spatial Errors ↓			
		D	α	SNRi (dB)	SI-SNRi (dB)	ΔILD (dB)	ΔIPD (rad)	ΔITD (μs)	ΔITD-GCC (μs)
Waveformer [25]	TS only	40	-	9.98	6.65	0.95	0.90	239.46	234.23
	one-hot + TS	360	-	11.35	7.90	0.81	0.89	175.68	185.13
	cyc-pos + TS	40	20	11.90	8.73	0.73	0.87	161.33	190.48
Proposed	TS only	40	-	16.85	14.28	0.42	**0.77**	105.91	122.76
	one-hot + TS	360	-	12.23	8.45	0.91	<u>0.84</u>	164.93	154.81
	cyc-pos + TS	40	80	14.81	12.96	0.79	0.91	96.78	108.32
			40	<u>17.22</u>	<u>15.85</u>	<u>0.37</u>	0.88	<u>80.97</u>	**103.77**
			20	**17.78**	**16.51**	**0.32**	0.87	**77.37**	<u>106.63</u>

Boldface and underlined numbers indicate the best and the second-best results for each metric, respectively. TS denotes the timestamp clue.

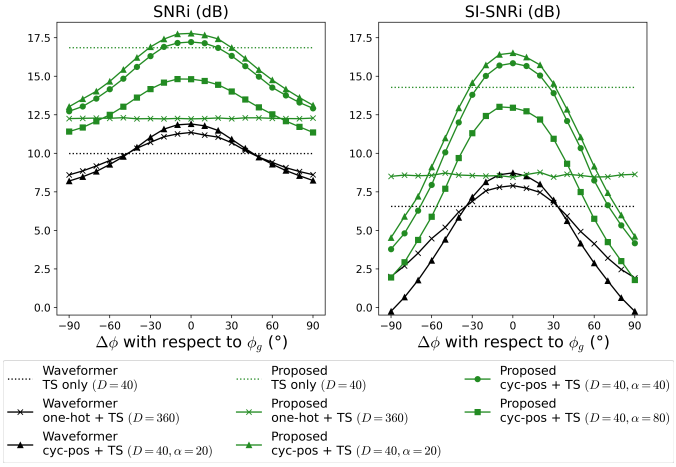

Fig. 3. SNRi and SI-SNRi change with respect to target azimuth angle.

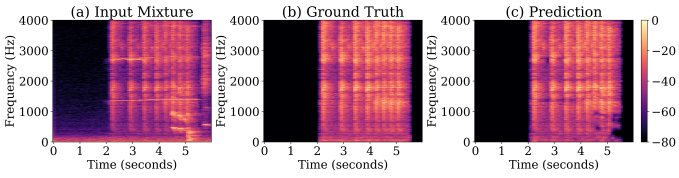

Fig. 4. Spectrograms of the first microphone channel for (a) input mixture, (b) ground truth, and (c) prediction result of our model.

As a baseline model for comparison, we modified the binaural extraction model[4] [25] based on Waveformer [3] into a 4-channel model. Compared to the baseline, our model performs significantly better, yielding higher SNRs and lower spatial errors. This result emphasizes the importance of encoding spatio-temporal information using a complex spectrogram instead of a waveform directly for extracting a multichannel sound. Additionally, the extraction performance is noticeably better when using a cyc-pos vector rather than a one-hot-encoded vector. Since the cyc-pos vector has smooth variation across azimuthal angles and has periodicity unlike the one-hot vector, it better integrates into spatially encoded features. However, with large values of α especially greater than the embedding dimension, performance is degraded because the embedding is no longer smooth and changes rapidly in a short cycle even for closely located directions. Meanwhile, most models employ similar ΔIPD values, encountering difficulties in analyzing interchannel phase differences directly, rather than the level or time difference of arrival.

To evaluate the sensitivity to incorrect directional clues, we measured extraction performance across gradually changing directional clues with a resolution of 10° from the true target direction. Results presented in Fig. 3 indicate that cyc-pos vectors maintain robust extraction performance near the true target direction, showing less than 1 dB decrease in SNRi for azimuth angle difference of $\pm 20^\circ$. In contrast, the one-hot vector embedding showed no variation with changing azimuth angles, indicating that the spatial information was not utilized by the model. Thus, using cyc-pos vectors is more advantageous, as it not only reduces memory usage by lowering the embedding dimension but also maintains periodicity.

One example of M2M-TSE is presented in Fig. 4, for the cyc-pos vector with parameters $D = 40$ and $\alpha = 20$. As depicted in the bottom right part of the input mixture and prediction in Fig. 4, distortion occurs due to the presence of louder competing sounds, which indicates that our model struggles to separate the target in presence of strong interferences. While the proposed approach works well in many scenarios, addressing this issue will be part of our future work, such as enhancing separation in more complex mixtures. Demo is available at https://choishio.github.io/demo_M2M-TSE/.

IV. CONCLUSION

We introduced an M2M-TSE framework for extracting multichannel sound from diverse sound mixtures using target direction and timestamp clues. The direction clue, embedded in form of cyclic positional encoding, was directly integrated with multichannel features of modified DeFTAN-II blocks to enable multichannel sound extraction. The proposed model demonstrated superior performance across multiple channels, outperforming the state-of-the-art extraction model using the same types of clues. This approach to extracting multichannel signals paves the way for separating and editing multichannel audio recordings by preserving spatial information of individual sources.

[4]https://github.com/vb000/SemanticHearing

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[2] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, Shanghai, China, 2020, ISCA, pp. 1441–1445.

[3] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Real-time target sound extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes, Greece, 2023, IEEE, pp. 1–5.

[4] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 31, pp. 121–136, 2022.

[5] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, K. Kashino, et al., "ConceptBeam: Concept driven target speech extraction," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 4252–4260.

[6] H. Wang, D. Yang, C. Weng, J. Yu, and Y. Zou, "Improving target sound extraction with timestamp information," in *Proc. Interspeech*, Incheon, Korea, 2022, ISCA, pp. 1526–1530.

[7] D. Kim, M. S. Baek, Y. Kim, and J. H. Chang, "Improving target sound extraction with timestamp knowledge distillation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seoul, Korea, 2024, IEEE, pp. 1396–1400.

[8] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, W. Wang, et al., "Separate anything you describe," *arXiv preprint arXiv:2308.05037*, 2023, [Online]. Available: https://arxiv.org/abs/2308.05037.

[9] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, W. Wang, et al., "Separate what you describe: Language-queried audio source separation," in *Proc. Interspeech*, Incheon, Korea, 2022, ISCA, pp. 1801–1805.

[10] R. Gu and Y. Luo, "ReZero: Region-customizable sound extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 2576–2589, 2024.

[11] C. Li, Y. Qian, Z. Chen, D. Wang, T. Yoshioka, M. Zeng, et al., "Target sound extraction with variable cross-modality clues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes, Greece, 2023, IEEE, pp. 1–5.

[12] H. Ma, Z. Peng, M. Shao, J. Liu, X. Li, and X. Wu, "CLAPSep: Leveraging contrastive pre-trained models for multi-modal query-conditioned target sound extraction," *arXiv preprint arXiv:2402.17455*, 2024, [Online]. Available: https://arxiv.org/abs/2402.17455.

[13] E. Tzinis, G. Wichern, A. Subramanian, P. Smaragdis, and J. L. Roux, "Heterogeneous target speech separation," in *Proc. Interspeech*, Incheon, Korea, 2022, ISCA, pp. 1796–1800.

[14] E. Tzinis, G. Wichern, P. Smaragdis, and J. L. Roux, "Optimal condition training for target source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes, Greece, 2023, IEEE, pp. 1–5.

[15] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Virtual, 2020, IEEE, pp. 6389–6393.

[16] R. Gu, S. X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Virtual, 2020, IEEE, pp. 7319–7323.

[17] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 542–553, 2023.

[18] D. Lee and J. W. Choi, "DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Process. Letters (SPL)*, vol. 30, pp. 155–159, 2023.

[19] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, IEEE, pp. 1–5.

[20] 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. (TASLP)*, vol. 31, pp. 3221–3236, 2023.

[21] D. Lee and J. W. Choi, "DeFTAN-II: Efficient multichannel speech enhancement with subgroup processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 32, pp. 4850–4866, 2024.

[22] M. A. Chung, C. W. Lin, and H. C. Chou, "Combined multi-sensor based angle clipping algorithm and multi-channel noise removal method for multi-channel sound localization," *IEEE Sens. J.*, vol. 24, no. 1, pp. 700–709, 2023.

[23] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. Wang, "Leveraging sound localization to improve continuous speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seoul, Korea, 2024, IEEE, pp. 621–625.

[24] A. Pandey, S. Lee, J. Azcarreta, D. Wong, and B. Xu, "All neural low-latency directional speech extraction," in *Proc. Interspeech*, Kos, Greece, 2024, ISCA, pp. 4328–4332.

[25] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Proc. 36th Annual ACM Symp. User Interface Software Tech. (UIST)*, San Francisco, CA, USA, 2023, pp. 1–15.

[26] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural speech enhancement using deep complex convolutional transformer networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seoul, Korea, 2024, IEEE, pp. 681–685.

[27] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural multi-channel wiener filtering for hearing aids: preserving interaural time and level differences," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006, IEEE, vol. 5, pp. V–V.

[28] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, and S. Araki, "Interaural time difference loss for binaural target sound extraction," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Aalborg, Denmark, 2024, IEEE, pp. 210–214.

[29] S. Baligar and S. Newsam, "McRTSE: Multi-channel reverberant target sound extraction," in *Proc. European Signal Processing Conference (EUSIPCO)*, Lyon, France, 2024, IEEE, pp. 6–10.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, 2017, pp. 4700–4708.

[31] H. Lee, C. Homeyer, R. Herzog, J. Rexilius, and C. Rother, "Spatio-temporal outdoor lighting aggregation on image sequences using transformer networks," *Int. J. Comp. Vis. (IJCV)*, vol. 131, no. 4, pp. 1060–1072, 2023.

[32] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016, [Online]. Available: https://arxiv.org/abs/1607.06450.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1026–1034.

[34] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, X. Serra, et al., "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, Surrey, UK, 2018, pp. 69–73.

[35] Z. Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Virtual, 2020, IEEE, pp. 486–490.

[36] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Alberta, Canada, 2018, IEEE, pp. 351–355.

[37] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2017, IEEE, pp. 344–348.

[38] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Maas, et al., "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2013, IEEE, pp. 1–4.

[39] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, 2019, IEEE, pp. 626–630.

[40] A. Pandey and D. L. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, vol. 29, pp. 1270–1279, 2021.