DySiME: Dynamic Single-Source Multichannel Enhancement Using Time-Varying Directional Cues

Hao Liang*, Yichen Yang*†, Xiao Zhang*, Shoji Makino*, Jingdong Chen†
*: Waseda University, Japan
E-mail: lianghao@akane.waseda.jp, zhang_x07@toki.waseda.jp, s.makino@waseda.jp
†: Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
E-mail: yang_yichen@mail.nwpu.edu.cn, jingdongchen@ieee.org

Abstract—**Single-source multichannel speech enhancement involves extracting speech from a desired, target speaker in noisy, reverberant environments while maintaining spatial fidelity across multichannel outputs. A recent solution to this problem, multichannel-to-multichannel target sound extraction (M2M-TSE), utilizes directional and temporal cues to perform end-to-end complex spectrogram mapping in reverberant environments with static sources. However, its effectiveness is limited by heavy reliance on prior knowledge and handcrafted cyclic positional embeddings, reducing its practicality in real-world applications. To overcome these limitations in dynamic source scenarios where the target speaker is moving, we propose DySiME, dynamic single-source multichannel enhancement, an end-to-end framework tailored for moving sources. DySiME integrates a direction-of-arrival (DOA) estimation module based on full-band and narrow-band fusion for sound source localization (FN-SSL) to continuously track the target source direction. Furthermore, a learnable positional-information adapter incorporates intermediate features from the DOA estimator into the enhancement backbone, enabling the model to utilize time-varying spatial cues for more effective speech enhancement. This design reduces reliance on prior DOA knowledge during inference and enables robust, real-time enhancement of moving sources. We evaluate the system using a simulated 4-channel circular microphone array, and the results show that DySiME consistently outperforms the baseline in both speech quality and spatial accuracy.**

I. Introduction

Multichannel speech enhancement involves extracting and improving a target speech signal from noisy multichannel observations captured by an array of microphones [1]–[3]. Compared to traditional single-channel techniques, multichannel methods leverage spatial information inherent in the multichannel input signals, leading to significantly improved enhancement performance. Consequently, they are widely adopted in applications such as smart speakers, binaural rendering [4], and immersive audio reproduction [5].

Traditional multichannel speech enhancement methods, including beamforming [6], [7] and geometrically constrained source extraction [8]–[11], typically rely on spatial cues like direction-of-arrival (DOA) information to extract the target source signal. These spatial cues are often easily obtainable and effective for enhancing performance. However, the output of such methods is usually a single-channel signal, which results in the loss of spatial information. With the advancement of deep neural networks (DNNs), DNN-based source extraction systems have emerged, extracting the target source using time-frequency masks [12]–[14] or neural beamformers [15]–[17]. While these approaches often outperform traditional methods in terms of signal-to-noise ratio (SNR) and signal quality, they still fail to preserve the spatial characteristics of the source, such as inter-channel time delays and level differences. Since spatial cues are essential for downstream applications like 3D audio, binaural rendering, and virtual reality, this work focuses on extracting the desired target source signal while preserving its spatial properties. To achieve this, we propose a neural network-based multi-input multi-output (MIMO) enhancement framework.

Indeed, recent efforts have begun addressing the challenge of multichannel-to-multichannel target sound extraction (M2M-TSE). Notably, the work in [18], based on the dense frequency-time attentive network II (DeFTAN-II) [19], represents a significant step forward. This approach encodes the ground-truth static target source direction and timestamp using one-hot or cyclic positional (cyc-pos) embeddings [20], which are then integrated into the transformer backbone to guide feature extraction. The method has shown promising performance in scenarios involving static sources, benefiting from directional and temporal cues. However, M2M-TSE depends on handcrafted positional embeddings and requires a fixed DOA prior for static sources. When extended to moving sources, this would correspond to a time-varying DOA trajectory, which is difficult to obtain in real-world conditions.

To eliminate the need for prior DOA information during inference, accurate time-varying DOA estimation becomes essential for tracking moving sources. Among existing approaches [21], [22], the full-band and narrow-band fusion for sound source localization (FN-SSL) model [23] has demonstrated superior performance. This model employs two bidirectional long short-term memory (BLSTM) layers: the full-band layer captures spatial correlations across all frequencies, while the narrow-band layer models temporal dynamics within individual frequency bands. Together, these layers enable the model to robustly and accurately estimate direct-path inter-

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

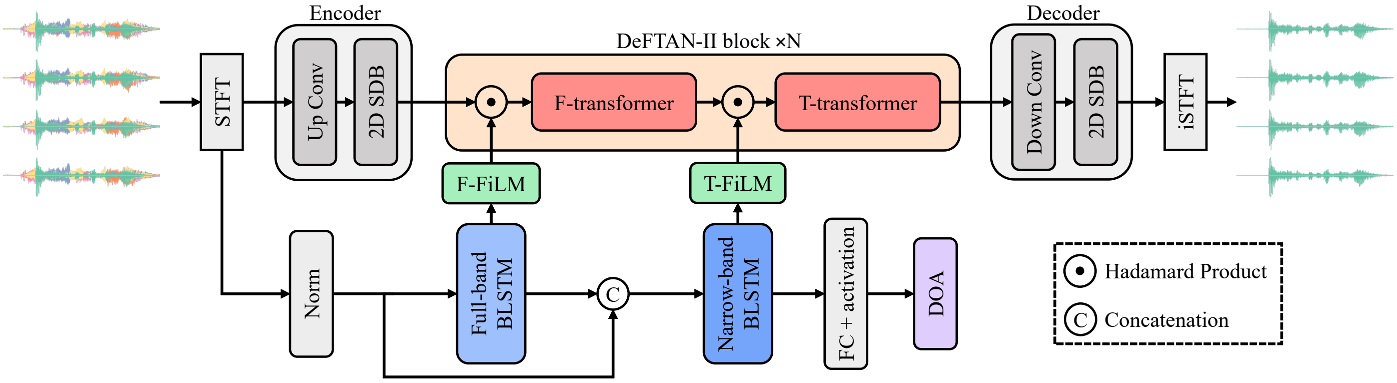

Fig. 1. Structure of the proposed DySiME.

channel phase differences (DP-IPD), even under challenging acoustic conditions, thereby allowing reliable localization of a moving target source.

In this paper, we introduce DySiME, a novel end-to-end dynamic single-source multichannel enhancement framework designed to eliminate reliance on handcrafted positional embeddings and inference-time DOA priors. DySiME integrates a DOA estimator, based on FN-SSL, directly into the M2M-TSE system to estimate frame-level DOAs of the moving target source. These estimated directional cues are then used to guide and enhance target speech extraction. To achieve effective integration, we incorporate a learnable Feature-wise Linear Modulation (FiLM) [24] adapter that fuses DOA features with the enhancement backbone. This enables DySiME to autonomously select the most relevant spatial cues for the enhancement process. Simulation results show that DySiME delivers substantial improvements in both speech quality and spatial consistency, even in challenging acoustic environments involving reverberation, interference, and background noise.

II. Proposed Method

Although the conventional M2M-TSE approach leverages handcrafted directional embeddings to guide static source enhancement, its reliance on such priors, which are unavailable in real-world dynamic scenarios, introduces a critical limitation. To address this, we propose the DySiME framework, an end-to-end system that eliminates the need for directional priors during inference. It seamlessly integrates a BLSTM-based DOA estimator with a transformer-based multichannel enhancement backbone. The system first extracts intermediate DOA features from both full-band and narrow-band BLSTM modules and then employs FiLM to inject these features into the time transformer and frequency transformer blocks, dynamically steering the enhancement process.

A. Model Architecture

The proposed DySiME framework is built upon the DeFTAN-II transformer architecture [19], which is shown in Fig. 1. It takes as input multichannel short-time Fourier transform (STFT) features $\{\mathbf{X}_t\}$, where $\mathbf{X}_t$ is the feature tensor at time frame t of shape $\mathbf{X}_t \in \mathbb{R}^{2M\times T\times F}$ with M, T, and F denoting the numbers of microphones, time frames, and frequency bins, respectively, and the factor of 2 accounting for the real and imaginary components of the complex STFT. These input features are first up-convolved to a hidden channel dimension C, and then passed through a 2D split dense block (SDB), which performs local sub-group feature extraction and time–frequency inter-channel aggregation. The resulting representation is subsequently processed by a stack of DeFTAN-II transformer blocks.

Each block in the framework alternates between a frequency transformer (F-transformer) and a time transformer (T-transformer). The F-transformer is designed to capture dependencies among frequency bins within a single time frame, learning the relationships between spectral components. In contrast, the T-transformer models temporal dependencies across consecutive time frames, allowing the network to track how signal features change over time. Following the approach of M2M-TSE [18], our method integrates positional information from the DOA estimator into these transformer blocks to guide and improve the enhancement process.

Once all DeFTAN-II blocks have refined the feature representations, a down-convolution layer followed by a 2D SDB compresses the hidden channels back into the real and imaginary components. These enhanced spectrograms are then converted into multichannel time-domain waveforms via the inverse STFT. In our implementation, both the input and output consist of four channels.

B. DOA Estimator

Our DOA estimator follows the approach of FN-SSL [23], utilizing both a full-band BLSTM and a narrow-band BLSTM to estimate the direct-path inter-channel phase difference (DP-IPD). The input consists of a 4-channel time-domain waveform $\{x_m[n]\}_{m=0}^{3}$, where each channel is transformed into the STFT domain to obtain $X_m(t,k)$, indexed by time frame t

and frequency bin k. The direct-path component $A_m^d(t,k,\theta)$ is extracted for each channel. The DP-IPD between channel i and j is then computed as $\text{DP-IPD}_{ij}(t,k,\theta) = \angle A_i^d(t,k,\theta) - \angle A_j^d(t,k,\theta)$, where $\angle A_m^d(t,k,\theta)$ represents the phase angle of the direct-path component at microphone m.

Phase-difference values are generated by extracting the DP-IPD across all unique microphone pairs at each frequency, capturing the time-delay-of-arrival (TDOA) information associated with the direct sound path. Unlike raw phase differences, DP-IPD effectively suppresses the influence of late reverberation and noise, offering a more reliable cue that closely correlates with the true source direction. The DOA estimator is trained to predict these DP-IPD values from noisy, reverberant multichannel inputs. Once estimated, the resulting DP-IPD patterns can be used to infer the azimuth of the moving target source on a per-frame basis.

The full-band BLSTM layer processes feature vectors across all K frequency bins at a single time frame t as

$$\mathbf{H}^f(t) = \left[\mathbf{h}^f(t,1) \;\; \mathbf{h}^f(t,2) \;\; \cdots \;\; \mathbf{h}^f(t,K)\right] \in \mathbb{R}^{D\times K}, \quad (1)$$

where $\mathbf{h}^f(t,k)$ is a D-dimensional vector representing the hidden representation for frequency bin k. This layer updates hidden states sequentially along the frequency axis, enabling the network to capture inter-frequency dependencies within a frame. Since all frequency components share the same underlying TDOA, their DP-IPD values are inherently correlated, making frequency-wise modeling critical.

Conversely, the narrow-band BLSTM layer processes features over all T time frames at a single frequency bin k as

$$\mathbf{H}^n(k) = [\mathbf{h}^n(1,k) \;\; \mathbf{h}^n(2,k) \;\; \cdots \;\; \mathbf{h}^n(T,k)] \in \mathbb{R}^{D\times T}, \quad (2)$$

where $\mathbf{h}^n(t,k)$ is a D-dimensional vector, presenting the output from the full-band layer at frequency k. This layer focuses on modeling the temporal evolution of the direct-path features at each frequency. Capturing these frame-wise dynamics is crucial for robust localization, particularly under reverberant environments where DP-IPD patterns may vary over time.

C. FiLM Adapter

To incorporate time-varying DOA cues into the enhancement backbone, we build on a recent study in which DOA information is projected through two different linear mappings to initialize hidden states in both frequency and time long short-term memory (LSTM) [25]. This approach guides the network to focus on spatial features in the frequency and time domains, respectively. Inspired by this, we leverage DOA features extracted from the full-band BLSTM to guide the F-transformer, while those from the narrow-band BLSTM are utilized to guide the T-transformer, achieving a frequency-to-frequency and time-to-time correspondence. Specifically, this enables the F-transformer to leverage time-frame-level spatial cues during cross-frequency attention and concurrently allows the T-transformer to utilize frequency-band-level spatial cues during cross-time attention.

Building on the established frequency-to-frequency and time-to-time correspondence, we employ a lightweight FiLM [24] mechanism as the core modulator for integrating DOA features. Inspired by the clue encoders used in M2M-TSE [18] and Directional Recurrent Network (DRN) [26], the FiLM adapter follows a Conv2D–LayerNorm–PReLU structure. Within each DeFTAN-II block, this adapter is placed between the BLSTM outputs and the inputs to the transformer blocks, enabling effective conditioning of the enhancement network with spatial information.

To generate modulation parameters for the two transformer branches, the adapter first extracts intermediate DOA features from both the full-band and narrow-band BLSTM layers. These features are reshaped into $\mathbf{S}^F$ and $\mathbf{S}^T$ tensors of shape $D \times T \times F$ for modulation. Focusing on the F-transformer branch, the FiLM adapter applies a 1×1 convolution to $\mathbf{S}^F$, projecting the channel dimension from D to $2C$ and producing $\mathbf{H}^F \in \mathbb{R}^{2C\times T\times F}$, where C is the transformer's input dimension. After 4D layer normalization and a PReLU activation, $\mathbf{H}^F$ is split along the channel axis into scaling parameters $\boldsymbol{\gamma}^F \in \mathbb{R}^{C\times T\times F}$ and bias parameters $\boldsymbol{\beta}^F \in \mathbb{R}^{C\times T\times F}$. The T-transformer undergoes the same process on $\mathbf{S}^T$ to produce $\boldsymbol{\gamma}^T$ and $\boldsymbol{\beta}^T$. Given the original inputs $\mathbf{x}^F, \mathbf{x}^T \in \mathbb{R}^{C\times T\times F}$ for the F-transformer and T-transformer respectively, the FiLM-modulated outputs are computed as

$$\widetilde{\mathbf{x}}^F = \boldsymbol{\gamma}^F \odot \mathbf{x}^F + \boldsymbol{\beta}^F, \quad (3)$$

$$\widetilde{\mathbf{x}}^T = \boldsymbol{\gamma}^T \odot \mathbf{x}^T + \boldsymbol{\beta}^T, \quad (4)$$

where $\odot$ denotes element-wise multiplication.

Building upon the loss functions used in M2M-TSE and FN-SSL, we formulate our training objective function as

$$\mathcal{L} = \mathcal{L}_{\text{PCM}} + \lambda \mathcal{L}_{\text{DOA}}, \quad (5)$$

where the phase-constrained magnitude loss $\mathcal{L}_{\text{PCM}}$ penalizes errors in both magnitude and phase to ensure accurate waveform reconstruction [27], and the DOA loss $\mathcal{L}_{\text{DOA}}$ is a mean absolute error loss used for accurate source direction estimation [23].

The proposed framework integrates intermediate features from the DOA estimator into the transformer backbone via FiLM modulation, enabling bidirectional gradient flow between the two components. By establishing both frequency-to-frequency and time-to-time correspondences, the system effectively guides speech enhancement across spectral and temporal dimensions within a unified, end-to-end trainable architecture. As a result, the DySiME model can continuously infer the time-varying direction of a moving target without relying on handcrafted priors, autonomously identifying the most informative positional cues.

TABLE I
Impact of DOA Information on Enhancement Performance and Spatial Consistency.

Methods	**Prior**	**Encode**	**Speech Quality ↑**				**Spatial Errors ↓**		
			PESQ	STOI	SDR (dB)	SI-SNRi (dB)	ΔILD (dB)	ΔIPD (rad)	ΔITD (μs)
M2M-TSE*	✓	cyc-pos	**2.30**	0.86	14.34	18.65	0.0223	**0.6904**	2.1302
M2M-TSE	✗	–	2.06	0.84	13.58	17.85	0.0331	0.9474	2.1667
DySiME (ours)	✗	cyc-pos	2.22	0.86	14.23	18.51	0.0317	0.7787	2.0677
DySiME (ours)	✗	FiLM	2.27	**0.86**	**14.46**	**18.72**	**0.0219**	0.8532	**1.7760**

* Oracle ground-truth DOA was provided.

III. Experiments

This section presents our experimental setup and results. We evaluate the proposed DySiME on a simulated moving-source dataset, measuring its performance in terms of speech quality, spatial consistency, and through focused ablation studies.

A. Datasets

In the simulated dataset, clean speech samples were randomly selected from the train, dev, and test sets of the LibriSpeech corpus [28], serving as both moving target sources and stationary interferers. Room impulse responses (RIRs) were generated using FRAM-RIR [29], with the reverberation time T_{60} (RT60) uniformly sampled between 0.2 and 1.3 seconds. A circular array with four microphones and a 3-cm radius was used, and its center was randomly placed within a 1-m radius circle centered on the room's horizontal midpoint. Room dimensions (width, depth, and height) were uniformly sampled from the ranges $[5, 10]$ m, $[5, 10]$ m, and $[3, 4]$ m, respectively.

The moving trajectory of the target source was generated according to [30], and the trajectory maintained a fixed height. Their start and end points lay on an annulus of radius $[1.5, 2.5]$ m centered on the microphone array, ensuring a path length of at least 2 m and a minimum distance of 0.5 m from the array center. Four stationary interference sources were placed off the target path, with an angular separation of at least 30° between the target's start point, end point, and each interferer. Each interference source was positioned at least 2 m from the array center and 0.1 m away from the nearest wall. The overall signal-to-interference ratio (SIR) between the moving source and four interference sources was uniformly sampled within $[-5, 5]$ dB. Following a setup similar to M2M-TSE [18], diffuse noise signals obtained from the 1st, 3rd, 5th, and 7th microphone noises in the RealMAN [31] dataset were added at a signal-to-noise ratio (SNR) uniformly sampled within $[-5, 15]$ dB. All mixtures were six-second clips sampled at 16 kHz, comprising 16K training, 2K validation, and 2K test examples.

B. Implementation Details

The input to the DOA estimator was formed by concatenating the real and imaginary parts of the STFT coefficients. Although the array consists of four microphones (channels 0–3), only three microphone pairs (0–1, 0–2, and 0–3) were used to compute DP-IPD, reducing computational load and GPU memory usage. To enable full complex spectrogram reconstruction [18], [19], the decoder's output channels were expanded from 2 (real + imaginary) to $2M$, where M is the number of microphone channels. Accordingly, the transformer backbone used four DeFTAN-II blocks. The input and output of our system were both set to four channels.

Training was conducted for 100 epochs using the Adam optimizer, starting with a learning rate of $5e^{-4}$. The learning rate was reduced by a factor of 0.1 if the validation scale-invariant signal-to-noise ratio improvement (SI-SNRi) [32] did not improve for five consecutive epochs. A batch size of 4 was used, with automatic mixed precision (AMP) and gradient norm clipping at 0.5 to balance convergence efficiency and memory usage.

C. Evaluation Metrics

The proposed DySiME was evaluated using three complementary metric groups. DOA estimation accuracy was measured by the frame-level mean absolute error (MAE) of predicted azimuths, along with localization accuracy within 5° and 10° thresholds, calculated over non-silent frames to assess the model's precision in tracking source direction. Speech quality was evaluated using perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), source-to-distortion ratio (SDR), and scale-invariant signal-to-noise ratio improvement (SI-SNRi), capturing the model's effectiveness in enhancing target speech while suppressing noise and interference. Spatial consistency was assessed via the mean absolute error of interchannel cues, including interaural level difference (ΔILD), interaural phase difference (ΔIPD), and interaural time difference (ΔITD), following [33], to evaluate how well the model preserved spatial characteristics in its multichannel output.

D. Ablation Studies

Table I presents the speech quality and spatial consistency metrics under various DOA configurations. To analyze the impact of DOA priors and different integration strategies, four experimental setups were compared. 1) Baseline (M2M-TSE) with oracle DOA clues: uses ground-truth DOA information during inference and serves as an upper-bound reference. 2) Baseline without DOA information: excludes any spatial guidance, representing the lower-bound performance of the backbone. 3) Proposed DySiME with cyc-pos encoding: uses predicted DOA values transformed into cyclic positional embeddings, similar to the baseline's encoding strategy. 4) Proposed DySiME with FiLM adapter: uses intermediate features from the DOA estimator to modulate transformer inputs via learnable FiLM layers, coupling DOA estimation and enhancement in a fully end-to-end framework. Both setups 3) and 4) operate without access to DOA priors during inference.

In the absence of any directional input, the backbone model achieved a PESQ of 2.06, STOI of 0.84, SDR of 13.58 dB, and SI-SNRi of 17.85 dB. The inclusion of oracle DOA information notably improved the performance of the M2M-TSE baseline, resulting in a PESQ of 2.30, STOI of 0.86, SDR of 14.34 dB, and SI-SNRi of 18.65 dB. It also exhibited improved spatial accuracy, with spatial errors reduced to ΔILD= 0.0223 dB, ΔIPD= 0.6904 rad, and ΔITD= 2.1302 μs.

The proposed DySiME with cyc-pos encoding performed slightly below this oracle-enhanced baseline, as expected. However, DySiME with FiLM modulation outperformed all configurations, including the oracle-guided M2M-TSE, in both enhancement and spatial consistency. It achieved a PESQ of 2.27, STOI of 0.86, SDR of 14.46 dB, and SI-SNRi of 18.72 dB. In terms of spatial preservation, it also delivered the lowest errors, with ΔILD= 0.0219 dB and ΔITD= 1.7760 μs, outperforming even the oracle baseline.

E. Results Analysis

The observed improvements in speech quality and spatial accuracy can be attributed to the end-to-end learning of spatial cues. By integrating the DOA estimator with the transformer backbone through a learnable FiLM adapter, the model enables bidirectional gradient flow between the two components. This allows the network to autonomously identify and leverage the most informative spatial features for enhancement. As a result, the system maintains consistent amplitude and timing across channels, leading to more effective noise suppression and improved spatial fidelity of the enhanced multichannel signals, surpassing even configurations with perfect DOA priors.

Nevertheless, DySiME with FiLM exhibited higher ΔIPD error. This drawback may stem from the application of learnable scaling (γ) and shifting (β) at each time-frequency bin, which introduces block-wise, discontinuous magnitude modulation. While this mechanism effectively boosts high-frequency components and compensates low-frequency delays, it may disrupt the smoothness of phase transitions in the complex domain. The resulting phase discontinuities between adjacent frequencies degrade phase consistency in the mid-to-low range, contributing to the elevated ΔIPD.

TABLE II
ACCURACY OF DOA IN THE ENHANCED SIGNALS PRODUCED BY DYSIME.

Encode	acc (5°) [%]	acc (10°) [%]	MAE [°]
cyc-pos	93.27	98.26	1.95
FiLM	93.42	98.18	1.97

The proposed DySiME with cyc-pos encoding showed minimal improvement in ΔILD but notably outperformed the oracle in ΔITD. This behavior is attributed to the inter-frame jitter introduced by imperfect DOA estimates. To avoid high-frequency artifacts resulting from noisy direction cues, the model adopts a conservative gain strategy, limiting its ability to reduce ΔILD. However, these low-frequency fluctuations act as a form of online data augmentation for temporal alignment, encouraging the network to develop a sharper, more dynamic peak-locking mechanism. As a result, DySiME with cyc-pos achieved better ΔITD than the oracle M2M-TSE, which was trained using smoothly varying ground-truth DOAs.

As shown in Table II, both the cyc-pos encoding and the FiLM adapter yielded nearly identical DOA regression accuracy within the DySiME framework. This indicates that the observed gains in speech quality and spatial consistency stem from FiLM's modulation capability, allowing the enhancement backbone to better utilize spatial cues, rather than any improvement in the accuracy of the upstream DOA estimator.

In conclusion, DySiME with FiLM not only matched but in several aspects exceeded the oracle-guided upper bound in terms of speech quality and spatial consistency, with only a minor compromise in phase accuracy. These results underscore the effectiveness of the end-to-end integration of directional cues for enhanced multichannel speech enhancement.

IV. Conclusions

In this work, we presented DySiME, a dynamic single-source multichannel speech enhancement framework that removes the need for prior DOA information during inference. By integrating intermediate features from a DOA estimator into a transformer-based enhancement backbone through a learnable FiLM adapter, DySiME enables end-to-end joint optimization of localization and enhancement. This design provides a robust and adaptive solution for enhancing speech from a moving target in noisy, reverberant environments. Notably, DySiME achieved performance that exceeded even oracle-guided baselines in both speech quality and spatial consistency.

References

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Berlin: Springer, 2005.

[2] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin: Springer, 2008.

[3] J. Benesty, I. Cohen, and J. Chen, Eds., *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE, 2018.

[4] J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki, "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *J. Acoust. Soc. Am.*, vol. 151, pp. 2624–2635, Apr. 2022.

[5] X. Sun, "Immersive audio capture, transport, and rendering: A review," *APSIPA Trans. Signal Inf. Process.*, vol. 10, pp. 1–24, Sep. 2021.

[6] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley-IEEE, 2002.

[7] J. Benesty, J. Chen, and H. Yiteng, *Microphone Array Signal Processing*. Berlin: Springer, 2008.

[8] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, Jun. 2020.

[9] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 846–850.

[10] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[11] Y. Yang, X. Wang, A. Brendel, W. Zhang, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. EUSIPCO*, 2023, pp. 41–45.

[12] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[13] J. Fan, R. Gu, Y. Luo, and C. Pang, "A unified geometry-aware source localization and separation framework for AD-HOC microphone array," in *Proc. IEEE ICASSP*, 2024, pp. 725–729.

[14] Y. Yang, H. Li, X. Wang, W. Zhang, S. Makino, and J. Chen, "Stereophonic music source separation with spatially-informed bridging band-split network," in *Proc. IEEE ICASSP*, 2024, pp. 3779–3783.

[15] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, 2016, pp. 196–200.

[16] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. IEEE ICASSP*, 2021, pp. 6089–6093.

[17] Y. Yang, N. Pan, W. Zhang, C. Pan, J. Benesty, and J. Chen, "Interference-controlled maximum noise reduction beamformer based on deep-learned interference manifold," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4676–4690, Oct. 2024.

[18] D. Choi and J.-W. Choi, "Multichannel-to-multichannel target sound extraction using direction and timestamp clues," in *Proc. IEEE ICASSP*, 2025, pp. 979–984.

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

[20] H. Lee, C. Homeyer, R. Herzog, J. Rexilius, and C. Rother, "Spatio-temporal outdoor lighting aggregation on image sequences using transformer networks," *Int. J. Comput. Vis.*, vol. 131, pp. 1060–1072, Dec. 2022.

[21] B. Yang, H. Liu, and X. Li, "SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization," in *Proc. IEEE ICASSP*, 2022, pp. 721–725.

[22] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *Proc. IEEE ICASSP*, 2022, pp. 716–720.

[23] Y. Wang, B. Yang, and X. Li, "FN-SSL: Full-band and narrow-band fusion for sound source localization," in *Proc. Interspeech*, 2023, pp. 3779–3783.

[24] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018, pp. 3942–3951.

[25] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

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

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

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[29] Y. Luo and R. Gu, "Fast random approximation of multi-channel room impulse response," in *Proc. IEEE ICASSP*, 2024, pp. 449–454.

[30] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 300–311, Nov. 2020.

[31] B. Yang, C. Quan, Y. Wang, *et al.*, "RealMAN: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization," in *Proc. NeurIPS DB&B Track*, 2024, pp. 105 997–106 019.

[32] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, May. 2019.

[33] 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. IWAENC*, 2024, pp. 210–214.