2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Speaker Localization in Classroom Environments Using GCC-PHAT Features and Mamba State Space Models with Ad-hoc Microphone Arrays

Rashed Iqbal*, Christian Ritz*, Jack Yang* and Sarah Howard
* University of Wollongong, NSW, Australia
E-mail: mri510@uowmail.edu.au, critz@uow.edu.au, jiey@uow.edu.au
University of Leeds, UK
E-mail: S.K.Howard@leeds.ac.uk

***Abstract*— Accurately understanding classroom interactions is essential for assessing teaching quality and student engagement. This study presents a novel audio-only source localization (SSL) framework designed for locating speakers within classroom environments using two microphone arrays located at unknown locations in the room. The proposed method integrates Generalized Cross-Correlation with Phase Transform (GCC-PHAT) for time difference of arrival (TDOA) feature extraction and the Mamba State Space Model (SSM) for sequence modelling to estimate speaker locations. A dataset of reverberant speech recordings corresponding to three simulated classroom environments of different size and reverberation levels containing a teacher and multiple students and recorded with two different microphone setups, including First Order Ambisonic (FOA) B-format arrays and 6-channel circular arrays. Experimental results demonstrate that our approach significantly outperforms conventional baselines in both Mean Angular Error (MAE) and Mean Distance Error (MDE).**

I. Introduction

Monitoring speaker activity in classroom environments, such as the location of students plays a vital role can be used to evaluate student engagement[1, 2], identifying participation disparities, understanding pedagogical effectiveness and diarization of speech recordings for further analysis. Sound Source Localisation (SSL) methods applied to microphone array recordings of classroom audio to locate individual speakers (i.e. students and the teacher) is an alternative to traditional approaches that rely on manual observation or analysis of video recordings. The authors have previously developed a low disturbance classroom observation system that uses multiple platforms distributed in a classroom to record classroom activity [3]. This paper proposes an SSL approach for use with this system using microphone arrays attached to each platform.

Finding the 3D position of speakers from multi-microphone recordings commonly rely estimating the Direction of Arrival (DOA) and techniques such as triangulation to estimate the distance from the microphone arrays to the source. Traditional DOA estimation algorithms, such as Multiple Signal Classification (MUSIC) [4], Steered Response Power with Phase Transform (SRP-PHAT) [5] and generalized cross-correlation with Phase Transform (GCC-PHAT) [6], can be effective under anechoic conditions but suffer performance degradation in acoustically challenging environments, such as reverberation and noise encountered in classrooms.

Recent deep learning developments have greatly enhanced DOA accuracy in such environments. These methods leverage spatial audio features like GCC, phase spectrograms, and spatial pseudo-spectra [7, 8] combined with Convolutional neural networks (CNNs) or recurrent networks (RNNs) [8-11]. More recent models like FN-SSL [12] achieve state-of-the-art results by fusing narrowband and full-band information through LSTM layers, effectively modeling time-frequency dependencies for robust localization.

The MAMBA neural SSM [13, 14] has recently been introduced as an alternative to attention-based transformers and RNNs. Mamba excels at capturing long-range dependencies with improved computational efficiency. Applications of Mamba have shown competitive results in speech enhancement, separation, and sequential modeling tasks [15-17], positioning it as a compelling architecture for localization under realistic constraints.

While DOA estimation provides angular localization, three-dimensional (3D) speaker localization estimates the Cartesian coordinates of active sound sources. Various deep learning-based methods have been proposed to address 2D [18, 19] and 3D [20] localization problems. However, the majority of these studies rely on constrained setups involving fixed microphone configurations and locations, thereby limiting their generalization to practical approaches, such as the targeted where recording platforms are arbitrarily distributed in a classroom. In contrast, the classroom observation system,

979-8-3315-7206-8/25/$31.00

which is the target application of this work, contains multiple distributed nodes of arrays and corresponds to the ad-hoc microphone array paradigm. While recent research into speaker localisation using ad-hoc arrays [28] explicit triangulation of DOA estimates from multiple distributed nodes, this paper employs a data-driven approach where Mamba learns spatial relationships implicitly from GCC-PHAT features, reducing reliance on precise geometric calibration compared to traditional triangulation methods.

This paper proposes a 3D speaker localization framework that leverages GCC-PHAT features and models spatio-temporal dependencies using a structured state-space model (SSM) known as Mamba. Unlike prior studies, our system supports ad-hoc multi-microphone configurations and employs soft spatial labels over a discretized 3D grid, enabling probabilistic spatial inference. The approach is validated on a simulated dataset of three classroom recordings with different reverberation levels. Each room uses two microphone arrays and results compare using First Order Ambisonic (FOA) B-format and circular microphone arrays. Results also compare several alternative SSL approaches, including MUSIC, GCC-PHAT, SRP-PHAT and combinations of these with different types of neural networks.

Section II of this paper presents the proposed GCC-PHAT-approach applied to an ad-hoc array formed by two B-format microphones while Section III describes the proposed Mamba-based SSL methodology. Section IV describes the experimental setup using dual circular microphone arrays and B-format recordings in simulated classroom environments. Section V analyzes the comparative results across different SNR conditions and microphone configurations. Section VI concludes with key findings and future directions.

II. GCC PHAT FOR AD-HOC B FORMAT ARRAYS

A. Signal model and application of GCC-PHAT

Assume $s_l[n], s_m[n]$ are discrete-time audio samples from microphones recording a sound source $s(n)$ in a reverberant and noisy environment are given in (1):

$$\begin{aligned} s_l[n] &= \mathrm{s[n]}*h_l[n]+w_l[n] \\ s_m[n] &= \mathrm{s[n]}*h_m[n]+w_m[n] \end{aligned} \quad (1)$$

Where in (1), $h_l[n], h_m[n]$ are the room impulse response between the speech source and each of the microphones and $w_l[n], w_m[n]$ is the additive noise at each microphone, where it is assumed that this noise is approximately uniform across the recording environment i.e. $w_l[n] = w_m[n]$. The location of the recorded sound source is related to the Time Difference of Arrival (TDOA) of this source at each microphone, which is estimated by finding the cross-correlation $r_{l,m}(\tau)$ (2) between the two recorded microphone signals $s_l[n], s_m[n]$, where τ represents the range of different possible time delays and L is the length of the two signals in samples $(-L \le \tau < L)$

$$r_{l,m}(\tau) = \sum_{n=0}^{L-1} s_l[n] s_m[n-\tau] \quad (2)$$

However, the standard GCC performs poorly in reverberant or noisy environments due to its sensitivity to signal spectrum variations and hence the GCC-PHAT function operates in the frequency domain using the PHAT weighting and given by (3), where $S_l(\omega), S_m(\omega)$ are the frequency domain versions of the recorded signals and ω is frequency.

$$R_{l,m-PHAT}(\omega) = \frac{S_l(\omega).S_m^*(\omega)}{|S_l(\omega).S_m(\omega)|} \quad (3)$$

The time domain correlation (4) is obtained by finding the inverse Fourier transform of (3).

$$r_{l,m-PHAT}(\tau) = \mathcal{F}^{-1}[R_{l,m-PHAT}(\omega)] \quad (4)$$

The estimated TDOA corresponds to the value of τ that maximises (4), as given in (5).

$$\hat{\tau} = \max_{\tau} r_{l,m-PHAT}(\tau) \quad (5)$$

The SRP-PHAT algorithm extends the GCC-PHAT approach to arrays of more than two microphones, by finding the GCC-PHAT for all possible pairs of microphones, with the TDOA estimated as the corresponding time delay resulting in the maximum overall correlation.

B. GCC-PHAT for two B-Format Microphone Arrays

B-format audio consists of four channels representing the omnidirectional encoding of the soundfield, $W(n)$, and three encodings in the X, Y and Z directions, $X(n)$, $Y(n)$ and $Z(n)$, respectively, which in theory correspond to (6) to (9), where $s(n)$ is the source signal at time n, ϕ is the azimuth angle and θ is the elevation angle.

$$W(n) = s(n) \quad (6)$$

$$X(n) = s(n)\cos(\phi)\sin(\theta) \quad (7)$$

$$Y(n) = s(n)\sin(\phi)\cos(\theta) \quad (8)$$

$$Z(n) = s(n)\sin(\theta) \quad (9)$$

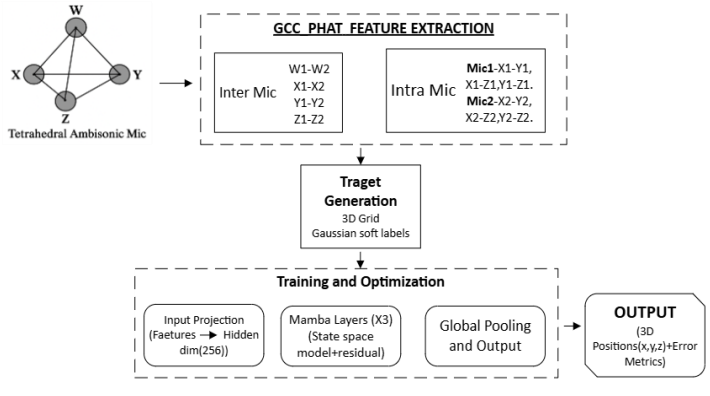

Fig. 1 Process flow diagram of the GCC-PHAT Mamba

Assume there are two B-format microphones located at two different positions in the room and given by $M_1(n) = [W_1(n), X_1(n), Y_1(n), Z_1(n)]$ and $M_2(n) = [W_2(n), X_2(n), Y_2(n), Z_2(n)]$. Rather than finding estimating the TDOA as the time delay that maximises the GCC-PHAT (see (5)), the proposed approach forms GCC-PHAT signals that are used as input features for a neural network. In particular, the proposed approach finds the GCC-PHAT between the same B-format channels of the two B-format microphones, r_{inter}, as well as the GCC-PHAT between microphone channels within the same B-format microphone, r_{intra}, as given by (10) and (11) and using the based GCC-PHAT in (4).

$$r_{inter}(\tau) = \{r_{W_1W_2}(\tau), r_{X_1X_2}(\tau), r_{Y_1Y_2}(\tau), r_{Z_1Z_2}(\tau)\} \quad (10)$$

$$r_{intra}^{(i)}(\tau) = \{r_{X_iY_i}(\tau), r_{X_iZ_i}(\tau), r_{Y_iZ_i}(\tau)\} \quad (11)$$

Where $i = 1,2$ in (11). Denoting the inter and intra GCC-PHAT features as vectors of length corresponding to the range of possible TDOAs, τ, the final time-domain feature vector as given in (12).

$$\mathbf{r} = [\mathbf{r}_{inter}, \mathbf{r}_{intra}^{(1)}, \mathbf{r}_{intra}^{(2)}] \in \mathbb{R}^D \quad (12)$$

Where in (12), $\mathbf{r}_{inter}, \mathbf{r}_{intra}^{(1)}, \mathbf{r}_{intra}^{(2)}$ are the vector representations of (10) and (11). These feature vectors are then used in the neural network approaches in described in this paper as in Fig.1.

III. MAMBA-BASED SSL APPLIED TO FOA GCC PHAT

A. Overview

Mamba is based on structured state space models (SSMs). The widely used H3 architecture underpins many SSM designs, typically consisting of a linear attention-inspired block alternated with a multi-layer perceptron (MLP) block. In this work, we simplify the design by merging these two components into a single, uniform block that is stacked repeatedly, drawing inspiration from the gated attention unit (GAU) proposed by Hua et al. [37], which applied a similar approach to attention. By combining GCC-PHAT features with the Mamba state space model, the proposed method captures both precise inter-channel time differences and their temporal dependencies. This synergy enables robust 3D sound source localization, particularly effective in ad-hoc microphone setups and acoustically challenging environments.

The core equations governing the Mamba model are given by (13) and (14).

$$h_t = A(u).h_{t-1} + B.u \quad (13)$$

$$x_t = C(u).h_t + D.u \quad (14)$$

Here, h_t represents the hidden state at time t, u is the input token at time t, and x_t is the output. The matrices A (·), B (·), C(.) and D (·) represent the state transition, input, output, and feedthrough matrices respectively, [21].

B. Feature Embedding and processing

To prepare GCC-PHAT features for the Mamba model, we first apply input projection where $\mathbf{r}$ represents the concatenated GCC-PHAT features from (12), where W_{proj}, b_{proj} are weights and biases.

$$X^{(0)} = \mathbf{r}.W_{proj} + b_{proj} \quad (15)$$

The projected features $X^{(0)}$ in (15) are then reshaped into a sequence format $(u_1, u_2, \ldots . u_T)$ and served as the input to mamba state space. Mamba block processing in (16), Where l represents the layer index $(l = 1,2, \ldots . , L)$ and $Mamba(X^{(l-1)}$ applies the state space operations from (13) and (14) to each time step in the sequence. The output of the final layer L gives the sequence $(x_1, x_2, \ldots . , x_T)$.

$$X^{(l)} = LayerNorm(Mamba(X^{(l-1)} + X^{(l-1)}) \quad (16)$$

To obtain a fixed-size representation for localization, the output sequence is aggregated (e.g., via mean pooling) in (17), where T is the sequence length:

$$z = \frac{1}{T}\sum_{t=1}^{T} X_i^{(L)} \quad (17)$$

This aggregated z *pooled latent* vector is then passed through a fully connected layer with a ReLU activation to produce a probability distribution over predefined spatial grid locations in (18):

$$O = Dropout\big(ReLU(z.W_1 + b_1)\big)W_2 + b_2 \quad (18)$$

where W and b are learnable parameters. The predicted distribution O is used for soft classification [22].

C. Training and Evaluation

The model is trained using a dataset comprising synchronized multi-microphone recordings and corresponding ground-truth source positions and as illustrated in Fig. 2. The ground-truth positions are converted into soft labels over a predefined spatial grid using a Gaussian kernel centered at the true location [23] and as described by (19).

$$x_{soft}[i] = \frac{\exp\left(-\frac{\|q_i - P_{true}\|^2}{2\sigma^2}\right)}{\sum_j \exp\left(-\frac{\|q_j - P_{true}\|^2}{2\sigma^2}\right)} \quad (19)$$

where q_i denotes the i^{th} spatial grid point, P_{true} is the true source position, and σ controls the spread of the Gaussian.

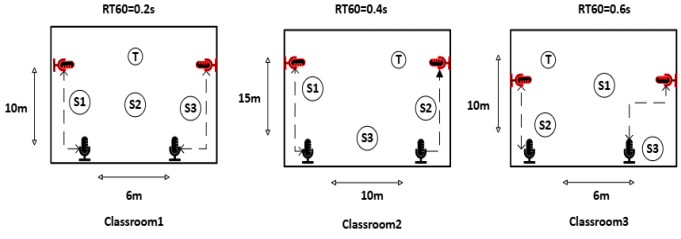

Fig. 2 Classroom settings: T-teacher, S1-student1, S2-student2, S3-Student 3 location.

The model's soft labels $x_{soft}[i]$ are found via (20) using the Kullback-Leibler divergence loss.

$$\mathcal{L} = \sum_i x_{soft}[i] \log \left(\frac{x_{soft}[i]_i}{softmax(O)[i]}\right) \quad (20)$$

During evaluation, the predicted source position is estimated as the expectation over the grid as per (21).

$$\hat{p} = \sum_i softmax(O)[i].q_i \quad (21)$$

IV. Experiment Setup

The experimental evaluation was conducted using pyroomacoustics [24] simulated B-format and circular microphone array recordings across three classroom environments with varying acoustic conditions: Classroom 1 (10×6×3m, RT60=0.2s); Classroom 2 (15×10×4m, RT60=0.4s); and Classroom 3 (10×6×4m, RT60=0.6s) in Fig. 2. Two B-format microphones and two circular microphone arrays were positioned in ad-hoc locations within each room to capture 8-channel B-format audio (W,X,Y,Z per microphone) and 12 channels for the circular microphones (6 channels each, 8cm radius) at 16 kHz sampling frequency. Speech signals from LibriSpeech clean dataset were used to simulate one teacher and three student speakers, with white noise and babble noise added at SNR levels of 10, 20, and 30 dB. The experimental protocol implements a 50/25/25 train/validation/test split ensuring spatial coverage across all regions and utilizes mixed-precision training on CUDA-enabled hardware. Performance on the test database is evaluated using the Mean Angular Error (MAE) and Mean Distance Error (MDE) between estimated and ground truth speaker locations.

The proposed approach was compared with several alternatives. MUSIC [25] applies jointly across both B-format arrays using a combined covariance matrix, utilizing the orthogonality between steering vectors and the noise subspace to estimate the DOA. SRP-PHAT [26] computes phase-weighted cross-correlations between all B-format channel combinations and steers the array response across a predefined 3D spatial grid to identify the direction yielding maximum power response. GCC-PHAT [27] is based on the approaches described in Section II.A and geometric triangulation between two B-format microphone arrays with known positions, computing time delays between corresponding channels and solving TDOA equations to determine 3D source positions. SRP-PHAT-CNN [28] feeds the computed SRP power spectrum as 3D feature maps into convolutional neural networks. GCC-PHAT-CNN [29] processes GCC-PHAT features from B-format channel pairs through one-dimensional convolutional networks. CNN-ULD (Unbiased Label Distribution) [30] applies CNN with continuous probability labels (ULD) to each B-format array separately for quantization-error-free DOA estimation, then triangulates between the two arrays' DOA estimates to obtain precise 3D sound source localization.. GCC-PHAT-MLP [31] utilizes multi-layer perceptrons to process GCC-PHAT features extracted from B-format channels.

V. Results and Analysis

Results from Table1 shows classroom 1, 2, and 3 reveal consistent high performance of the proposed method, achieving mean angular errors of 1.2°, 2.2°, and 1.2° respectively for B-format recordings at 30dB SNR, representing a 5-10×

Table 1. MAE and MDE Results for the Three Classrooms for both Ambisonics(A), Circular (C) microphone arrays (SNR 30dB).

Models	Classroom1				Classroom2				Classroom3			
	MAE^O		MDE(m)		MAE^O		MDE(m)		MAE^O		MDE(m)	
	A	**C**	**A**	**C**	**A**	**C**	**A**	**C**	**A**	**C**	**A**	**C**
MUSIC	25.2	24.7	3.1	3.4	21.2	22.0	2.9	3.7	22.0	23.0	3.9	4.0
SRP-PHAT	21.1	23.0	3.2	3.1	21.9	20.2	2.6	2.8	23.9	21.1	4.1	3.5
SRP-CNN	20.2	23.1	1.6	1.9	20.0	16.5	2.3	1.3	14.2	20.1	1.5	3.4
CNN-ULD	8.1	9.5	0.1	0.1	25.3	27.9	2.1	3.6	24.1	25.3	1.1	1.2
GCC-PHAT	20.1	30.0	3.4	3.8	5.4	23.7	0.8	3.7	6.7	25.2	0.5	2.6
GCC-CNN	14.9	9.8	1.6	1.3	18.9	22.0	2.1	3.2	15.4	11.1	1.4	1.2
GCC-MLP	7.8	5.1	0.9	0.2	9.8	8.0	1.2	2.0	3.8	4.0	0.6	0.8
GCC-PHAT-Mamba	**1.2**	**3.5**	**0.1**	**0.3**	2.2	4.9	0.4	1.3	1.2	6.0	0.3	0.4

improvements over traditional methods and 2-3× improvements over alternative neural approaches. The cross-environment standard deviation of 0.6° for GCC-PHAT-Mamba demonstrates exceptional stability compared to other methods, with CNN-ULD showing high variance ($\sigma = 9.93°$) and traditional GCC-PHAT exhibiting environment-dependent performance ranging from 5.4° to 30.0°. Moreover, for the 10dB and 20dB SNR the results show similar significant improved accuracy for proposed method compared to other baseline method shows in Table 2.

Table 2. GCC-Mamba performance at SNR 10dB and 20dB

Models	SNR	Classroom 1				Classroom 2				Classroom 3			
		MAE°		MDE(m)		MAE°		MDE(m)		MAE°		MDE(m)	
		A	C	A	C	A	C	A	C	A	C	A	C
GCC-CNN	10	15.0	10.2	1.6	1.5	18.9	22.9	2.4	3.5	15.9	12.0	1.6	1.3
	20	14.1	9.9	1.6	1.4	18.9	22.9	2.2	3.2	15.6	11.2	1.4	1.2
GCC-MLP	10	8.2	5.4	0.7	0.4	10.2	8.4	1.3	2.2	3.8	4.0	0.6	0.9
	20	7.2	5.2	0.9	0.2	9.8	8.2	1.2	2.1	3.8	4.0	0.6	0.8
GCC-Mamba	10	**1.9**	3.9	**0.2**	0.5	**2.4**	**5.1**	**0.4**	**1.4**	**1.3**	**6.5**	**0.3**	**0.4**
	20	**1.6**	**3.6**	**0.1**	**0.4**	**2.3**	**5.1**	**0.4**	**1.3**	**1.3**	**6.0**	**0.3**	**0.4**

Results also reveal that B-format arrays consistently outperform circular arrays across all environments, with the proposed method maintaining sub-2.5° angular accuracy in all tested conditions. Circular array performance shows greater sensitivity to acoustic environments, with errors ranging from 3.5° to 6.0° for GCC-PHAT-Mamba, though still significantly outperforming competing approaches. Distance accuracy analysis confirms the method's practical applicability, achieving sub-meter precision in most cases with minimum distance errors of 0.094 m (Classroom 1, Ambisonics) and maximum of 1.298 m (Classroom 2, Circular). The acoustic characterization reveals Classroom 1 as optimal with low reverberation, Classroom 2 as correlation-friendly favoring feature-based approaches, and Classroom 3 providing moderate challenge conditions.

Statistical analysis (paired t-test with multiple comparison correction) demonstrates highly significant improvements of the proposed GCC-PHAT-Mamba method over all alternatives ($p < 0.001$), with consistent superiority maintained across varying SNR conditions from 10dB to 30dB. The method's robustness to noise degradation is particularly noteworthy, showing minimal performance reduction compared to competing neural approaches that exhibit substantial degradation at lower SNR levels. Feature-based neural methods consistently outperform end-to-end approaches, validating the design choice of combining GCC-PHAT feature extraction with Mamba's sequential modeling capabilities. Traditional methods, while showing consistent but moderate performance (20-30° MAE range), demonstrate limited accuracy for practical applications requiring precise localization.

VI. Conclusions

This paper presents a novel GCC-PHAT-Mamba approach for sound source localization applied to classroom audio recordings made by ad-hoc microphone arrays formed from two arbitrary located microphone arrays. Through comprehensive evaluation across three distinct classroom environments with varying SNR conditions (10-30 dB), the proposed method demonstrates exceptional performance, achieving mean angular errors as low as 1.16° for Ambisonics B-format recordings, representing a significant improvement over both traditional methods (MUSIC, SRP-PHAT, GCC-PHAT) and over alternative neural approaches (CNN-ULD, GCC-PHAT-CNN, GCC-PHAT-MLP). Comparative analysis demonstrates that B-format configurations consistently outperform circular microphone arrays, though both benefit significantly from the proposed Mamba-based approach. Future work will investigate the extension to scenarios involving multiple simultaneous speakers.

VII. Acknowledgment

This research in this project was supported by ARC Discovery Projects DP130100481 and DP210101426.

References

[1] R. Iqbal, C. Ritz, J. Yang, and S. Howard, "Few-Shot Audio Classification Model for Detecting Classroom Interactions Using LaSO Features in Prototypical Networks," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024: IEEE, pp. 1-6.

[2] O. Canovas, F. J. Garcia-Clemente, and F. Pardo, "Ai-driven teacher analytics: Informative insights on classroom activities," in *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 2023: IEEE, pp. 1-8.

[3] H. Zhou, F. Jiang, J. Si, L. Xiong, and H. Lu, "Stuart: Individualized classroom observation of students with automatic behavior recognition and tracking," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: IEEE, pp. 1-5.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation,* vol. 34, no. 3, pp. 276-280, 1986.

[5] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University, 2000.

[6] J. Thyssen, A. Pandey, and B. J. Borgström, "A novel Time-Delay-of-Arrival estimation technique for multi-microphone audio processing," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015: IEEE, pp. 21-25.

[7] 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*

Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2626-2637, 2020.

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018: IEEE, pp. 1462-1466.

[9] D. Suvorov, G. Dong, and R. Zhukov, "Deep residual network for sound source localization in the time domain," *arXiv preprint arXiv:1808.06429,* 2018.

[10] B. Yang, H. Liu, and X. Li, "SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: IEEE, pp. 721-725.

[11] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 12, pp. 2444-2453, 2017.

[12] Y. Wang, B. Yang, and X. Li, "FN-SSL: Full-band and narrow-band fusion for sound source localization," *arXiv preprint arXiv:2305.19610,* 2023.

[13] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396,* 2021.

[14] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in Neural Information Processing Systems,* vol. 35, pp. 35971-35983, 2022.

[15] X. Zhang, J. Ma, M. Shahin, B. Ahmed, and J. Epps, "Rethinking mamba in speech processing by self-supervised models," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025: IEEE, pp. 1-5.

[16] X. Zhang *et al.*, "Mamba in speech: Towards an alternative to self-attention," *IEEE Transactions on Audio, Speech and Language Processing,* 2025.

[17] X. Jiang, C. Han, and N. Mesgarani, "Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025: IEEE, pp. 1-5.

[18] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Computer Speech & Language,* vol. 49, pp. 83-106, 2018.

[19] G. Zhang, L. Geng, F. Xie, and C.-D. He, "A dynamic convolution-transformer neural network for multiple sound source localization based on functional beamforming," *Mechanical Systems and Signal Processing,* vol. 211, p. 111272, 2024.

[20] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Transactions on Industrial Electronics,* vol. 65, no. 8, pp. 6403-6413, 2017.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752,* 2023.

[22] Y. Xiao and R. K. Das, "Tf-mamba: A time-frequency network for sound source localization," *arXiv preprint arXiv:2409.05034,* 2024.

[23] S. de Vries and D. Thierens, "Generating the ground truth: Synthetic data for soft label and label noise research," *International Journal of Data Science and Analytics,* pp. 1-13, 2025.

[24] 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)*, 2018: IEEE, pp. 351-355.

[25] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009: IEEE, pp. 2027-2032.

[26] A. Levi and H. F. Silverman, "An alternate approach to adaptive beamforming using srp-phat," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010: IEEE, pp. 2726-2729.

[27] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments," *IEEE Access,* vol. 8, pp. 7373-7382, 2020.

[28] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards domain independence in CNN-based acoustic localization using deep cross correlations," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021: IEEE, pp. 226-230.

[29] J. Zhao and C. Ritz, "Adapting GCC-PHAT to co-prime circular microphone arrays for speech direction of arrival estimation using neural networks," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022: IEEE, pp. 815-819.

[30] S. Liu *et al.*, "Deep learning based stage-wise two-dimensional speaker localization with large ad-hoc microphone arrays," *Speech Communication,* p. 103247, 2025.

[31] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018: IEEE, pp. 74-79.