

Enhancing Stereo Sound Event Detection with BiMamba and Pretrained PSELDnet

Wenmiao Gao*, Han Yin†

* Department of Electrical and Photonics Engineering, Denmark Technical University, Lyngby, Denmark

† School of Electrical Engineering, KAIST, Daejeon, Republic of Korea

E-mail: wenmiaogao@163.com, hanyin@kaist.ac.kr

Abstract—Pre-training methods have greatly improved the performance of sound event localization and detection (SELD). However, existing Transformer-based models still face high computational cost. To solve this problem, we present a stereo SELD system using a pre-trained PSELDnet and a bidirectional Mamba sequence model. Specifically, we replace the Conformer module with a BiMamba module. We also use asymmetric convolutions to better capture the time and frequency relationships in the audio signal. Test results on the DCASE2025 Task 3 development dataset show that our method performs better than both the baseline and the original PSELDnet with a Conformer decoder. In addition, the proposed model costs fewer computing resources than the baselines. These results show that the BiMamba architecture is effective for solving key challenges in SELD tasks. The source code is publicly accessible at https://github.com/alexandergwm/DCASE2025_TASK3_Stereo_PSELD_Mamba.

I. INTRODUCTION

Sound event localization and detection (SELD) aims to identify sound events from a set of known classes, follow their changes over time, and estimate their positions in space when they occur [1]. This task is important in many real-world applications. For example, it is used in robotic hearing systems, human-computer interaction, and immersive audio experiences. These technologies rely on accurate sound understanding to interact better with the environment.

The SELD task plays a key role in audio understanding by detecting sound events and locating their sources. To support this field, Detection and Classification of Acoustic Scenes and Events (DCASE) has featured SELD as a regular task since 2019, increasing its difficulty each year to reflect real conditions. Early datasets used clean, simulated audio. In 2022, real recordings were added. In 2023, 360-degree video was introduced, creating an audio-visual track. In 2024, distance estimation was included. The 2025 version uses stereo audio and perspective video to better match media content. These changes highlight SELD’s rising importance and complexity.

Deep learning has brought major progress to sound event detection (SED) [2]–[7] and direction of arrival (DOA) estimation [8], [9], clearly outperforming earlier signal processing and machine learning techniques. A key milestone came with SELDnet by Adavanne et al. [1], which was the first end-to-end model to detect and localize sound events at the same time. However, this method could not handle cases where the same type of sound occurred at different locations. To solve this, the EINV2 model [10] introduced permutation-invariant

training with a track-wise output format. It used two separate branches for SED and DOA, plus an extra shared branch. While effective, this design added training complexity and increased computation due to the need to balance different loss functions. To simplify the task, the class-wise ACCDOA format [11] was proposed. It combined detection and localization into a single vector for each sound class, making SELD a single-target problem. Later, the ADPIT method was developed to better handle overlapping sounds of the same class. The MultiACCDOA output format [12], which extends this approach, is now widely used as the standard baseline.

Despite better output formats from MultiACCDOA, SELD still faces challenges, especially with training data. While pre-training and techniques like Audio Channel Swapping (ACS) [13] help, it remains difficult to gather large, spatially rich datasets. This issue is worse with stereo data from FOA, as generating realistic audio-video pairs is resource-intensive. Limited data often leads to overfitting and weak generalization. Architecture-wise, models like PSELDnet [14] rely on powerful pre-trained backbones [15]–[17] but still use heavy transformer-based architectures or large CNN models. These high-parameter models increase computational complexity significantly. Also, many systems treat spatial and temporal features separately, limiting full scene understanding.

To address this, we consider Mamba [18], [19], a new architecture combining CNNs’ local modeling with Transformers’ long-range ability using a state-space approach. Mamba scales linearly with input length ($O(n)$), making it efficient for long audio sequences. Its ability to capture both short and long-term patterns fits SELD’s needs well. Recent studies confirm Mamba’s value. In speech separation and enhancement [20]–[24], BiMamba has replaced BiLSTM and Multi-Head Self-Attention, achieving strong results with less complexity. Mu et al. [24] also replaced the Conformer module with BiMamba based on the EINV2 framework, surpassing EINV2’s performance in multi-task output SELD systems. These outcomes show that BiMamba can improve performance while remaining computationally efficient.

Based on these strengths, we propose a new approach that integrates BiMamba into the PSELDnet framework. Specifically, we replace the Conformer decoder with a BiMamba block to better model spatiotemporal patterns in stereo input and reduce computing cost. We also introduce asymmetric convolutions to decouple time and frequency features, further

reducing computational load. This design supports the needs of the DCASE2025 task, where generating large-scale spatial audio-visual data is not feasible, and stereo input is required. By fine-tuning the pre-trained PSELDnet models with this improved structure, we aim to boost SELD performance while keeping the system efficient.

II. PROPOSED METHOD

Training SELD systems on large datasets and fine-tuning pre-trained models on specific downstream datasets has shown excellent transfer and generalization [14]. This approach often uses the FSD50K dataset [25] to collect clean, single-source sound clips, which are then convolved with room impulse responses generated by the mirror source method [26], before being converted to FOA format [27].

However, for tasks involving conventional videos with stereo audio, creating audio-video datasets is technically challenging and requires significant resources. This makes it difficult to train models directly on large paired datasets. Therefore, effectively using pre-trained models is essential for achieving strong performance. To solve this problem and make stereo inputs compatible with pre-trained FOA-based models, we apply ACN/SN3D conversion techniques, which transforms stereo signals—simulating Mid/Side technology—into pseudo-FOA format, enabling smooth adaptation to existing pre-trained models. Specifically, the binaural signals $L(n)$ and $R(n)$ are transformed into FOA components according to Equation (1):

$$\begin{aligned} W(n) &= \frac{L(n) + R(n)}{2} \\ Y(n) &= \frac{L(n) - R(n)}{2} \\ X(n) &= 0 \\ Z(n) &= 0 \end{aligned} \quad (1)$$

where $W(n)$ represents the omnidirectional component, $Y(n)$ represents the left-right directional component, $X(n)$ represents the front-back directional component, and $Z(n)$ represents the up-down directional component. Since stereo signals inherently lack precise spatial information, $X(n)$ and $Z(n)$ are set to zero. This four-channel pseudo-FOA representation is then processed to form the concatenation of log-mel spectrograms and intensity vectors as input.

A. Theory of Mamba

The Mamba architecture, developed from the Structured State Space Sequence (S4) model [21], [28], combines the strengths of CNNs and RNNs in a unified framework. It uses CNNs' ability for parallel computation during training, while leveraging RNN-like temporal modeling during inference. This design helps the model efficiently capture both local and long-term dependencies in audio signals. Additionally, Mamba introduces a state-selection mechanism, which allows the network to focus on or ignore specific parts of the input sequence. This selective attention is essential for accurately identifying overlapping sound events, a common challenge in real-world SELD tasks.

Specifically, inspired by continuous linear time-invariant systems in signal processing and control systems, it transforms the input sequence $x(t) \in \mathbb{R}$ to the output sequence $y(t) \in \mathbb{R}$ using higher dimensional hidden states $h(t) \in \mathbb{R}^{N \times 1}$, which can be written as follows:

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}^T\mathbf{h}'(t) + \mathbf{D}x(t) \end{aligned} \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{N \times 1}$, and \mathbf{D} represent the state transition matrix, the input projection matrix, the output projection matrix, and the skip connection matrix respectively. In practice, to handle discrete sequences, discretization of the SSM is necessary. Using the Zero Order Holding method, we introduce a time step Δ to sample the continuous matrices \mathbf{A} and \mathbf{B} , obtaining discrete representations $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, as follows:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \bar{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp \Delta\mathbf{A} - \mathbf{I}) \cdot \Delta\mathbf{B} \end{aligned} \quad (3)$$

and the discretized structured SSM are as follows:

$$\begin{aligned} \mathbf{h}_k &= \bar{\mathbf{A}}\mathbf{h}_{k-1} + \bar{\mathbf{B}}x_k \\ y_k &= \mathbf{C}^T\mathbf{h}_k \end{aligned} \quad (4)$$

The discretized parameters in Mamba vary over time through a selective state space modeling (SSM) approach, which works similarly to gating mechanisms in RNNs. This allows the model to decide at each time step whether to focus on or ignore certain input features, improving its ability to process information effectively. Building on this idea, Mamba2 introduces the SSD module [19], which further decomposes state evolution into two processes: intra-chunk parallel computation and inter-chunk recurrence. This design enables linear computational complexity while maintaining high accuracy when modeling long audio sequences, making it well-suited for demanding SELD tasks.

B. Overall Pipeline

Our proposed model builds on the CNN14-Conformer backbone from PSELDnet, using the pre-trained CNN14 encoder to extract strong feature representations. The model produces predictions in the Multi-ACCDOA (MACCDOA) format. To enhance temporal and spatial modeling, we replace the original Conformer blocks with BiMamba blocks, which are further improved by asymmetric convolutional layers.

As shown in Figure 1, the overall architecture includes the pre-trained CNN14 block as the encoder and BiMamba blocks as the decoder. To resolve the temporal resolution mismatch caused by the multiple pooling layers in CNN14, we add a Temporal Module before the fully connected (FC) layers. This module ensures the model's output aligns with the temporal resolution required by the labels.

The Temporal Module has two main components: Time Interpolation and Frame Aggregation. Time Interpolation restores temporal resolution lost through CNN14's pooling layers, while Frame Aggregation adjusts the frame alignment to

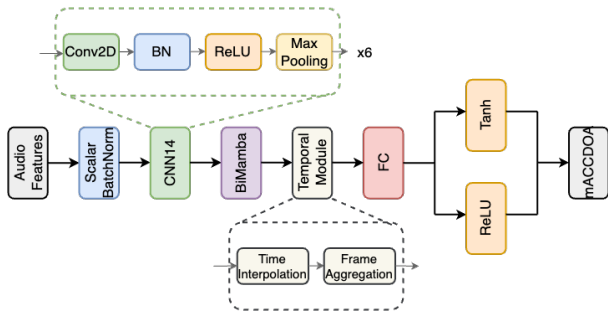


Fig. 1. The overall architecture of the proposed system.

match the target label resolution. Finally, the model applies activation functions suited to each output type: Tanh is used for Cartesian coordinate outputs, and ReLU is applied to distance predictions. The final output, in Multi-ACCDOA format, provides comprehensive spatial and temporal localization of sound events. Specifically, the CNN14 encoder backbone consists of six VGG-style convolutional blocks. Each block includes 3×3 convolution layers, followed by batch normalization, max-pooling, and ReLU activation functions. This hierarchical structure allows the model to progressively extract detailed spectral-temporal features from the input audio. As the network goes deeper, each block captures more abstract and complex acoustic patterns, enabling the model to build strong feature representations for sound event localization and detection.

For the decoder, our BiMamba module extends the original Mamba architecture by introducing a bidirectional processing mechanism. This design allows the model to capture both forward and backward temporal dependencies, overcoming the causal limitations found in traditional Mamba models. By processing information in both directions, the decoder gains a more complete understanding of the temporal context within audio sequences.

A key innovation in our approach is the use of asymmetric convolution, which separately processes the temporal and frequency dimensions. This design decouples feature extraction into two specialized pathways: one focused on dynamic patterns over time, and the other on acoustic features across frequencies. Figure 2 shows the detailed structure of the BiMamba block with asymmetric convolution, where gray blocks indicate Mamba layer components. This decoupled approach offers several advantages. First, replacing traditional 2D convolutions with 1D convolutions reduces computational complexity while maintaining strong feature extraction capabilities. Second, the dual-pathway design captures complementary aspects of the input, enabling richer and more robust feature representations. Finally, independent processing of temporal and spectral dimensions allows for targeted optimization in each domain, improving the model's ability to detect dynamic events and identify their acoustic characteristics accurately.

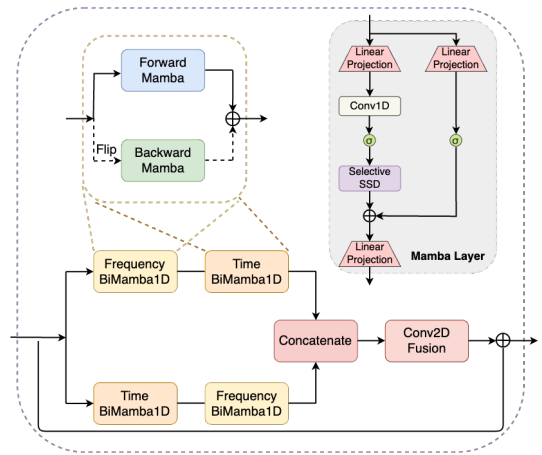


Fig. 2. The detailed structure of the BiMamba2DAC Block.

III. EXPERIMENTS

A. Benchmark Dataset

The proposed method is evaluated on the public benchmark dataset **DCASE2025 Task3 Stereo SELD Dataset** [29], which is derived from the original **STARSS23** [30]. This dataset consists of 5-second audio segments sampled from the source data, where the original FOA audio recordings are converted to stereo audio based on fixed viewpoints. The dataset encompasses 13 distinct sound event categories for comprehensive evaluation of sound event localization and detection performance.

B. Comparable Models

For comparative evaluation, we benchmark our proposed method against several models. The comparison includes the original **CNN14-Conformer** model [15] as the foundational baseline, the **HTS-AT** model [17] based on pure Transformer architecture, the **CNN14-ConBiMamba** [23] and **CNN14-ConBiMambaAC** models, which replace the Multi-Head Self-Attention (MHSA) components in Conformer blocks with BiMamba and BiMambaAC mechanisms, respectively. The **CRNN** model [1] is also regarded as an additional reference.

C. Model Configurations

Audio signals are resampled to 24 kHz and processed using 64 mel-scale filters. STFT computation uses a 20ms hop length (480 samples) and 40ms window length (960 samples). The model employs 2 BiMamba2DAC blocks with recommended parameters: $d_{state} = 64$, $d_{conv} = 4$, and $expand = 2$.

For training, all models use the Adam optimizer with ReduceLROnPlateau scheduling, monitoring validation F1-score and reducing the learning rate by half after 5 stagnant epochs. Models are trained for 120 epochs with full parameter fine-tuning. Training hyperparameters include a learning rate of $1e-4$ and weight decay between $1e-4$ and $5e-6$, with batch sizes ranging from 32 to 256 depending on model size.

Evaluation follows the DCASE2025 Task 3 protocol, selecting the checkpoint with the highest validation location-dependent F1-score. Performance metrics include location-dependent F1-score at a 20° threshold (F_{20°), DOA Error (DOAE), and Relative Distance Error (RDE). Model complexity is assessed by parameter count and MACs computed on 5-second audio segments sampled at 24 kHz.

IV. RESULTS

We evaluate our proposed BiMambaAC against several baseline models under two training scenarios: training from scratch and fine-tuning from pretrained PSELDnet weights.

A. Training from Scratch

Table I compares different model architectures trained from scratch on the development set. Conformer achieves the highest F_{20° score of 32.7% with excellent directional accuracy (14.5° DOAE) but requires 210M parameters and suffers from poor distance estimation (47% RDE). Our BiMambaAC delivers competitive performance with 32.1% F_{20° and comparable directional accuracy (14.8° DOAE) using only 76M parameters while maintaining significantly better distance localization (36% RDE). HTS-AT shows 25.2% F_{20° with the most efficient parameter usage (28M), demonstrating a favorable accuracy-efficiency trade-off. The lightweight CRNN baseline establishes a lower bound at 22.8% F_{20° .

TABLE I
COMPARISON OF DIFFERENT MODEL ARCHITECTURES ON THE DEVELOPMENT SET TRAINED FROM SCRATCH

Model	Params	MACs	$F_{20^\circ} \uparrow$	DOAE \downarrow	RDE \downarrow
Conformer	210M	4.69G	32.7%	14.5°	47%
HTS-AT	28M	2.88G	25.2%	18.0°	40%
CRNN	0.7M	57M	22.8%	24.5°	41%
BiMambaAC	76M	4.63G	32.1%	14.8°	36%

B. Fine-tuning from Pretrained PSELDnet

Table II shows that pretrained PSELDnet initialization significantly improves performance across all models. Our BiMambaAC achieves the best overall performance with 39.6% F_{20° and 15.8° DOAE using only 76M parameters, outperforming BiMamba (+3.4% F_{20°) while reducing parameters by 57% and computation by 39%. The ablation study in Table III validates the importance of each architectural component. Removing bidirectional processing causes the largest performance drop (-8.6% F_{20°), while eliminating asymmetric convolution reduces accuracy by 6.4%. Although increasing d_{state} from 64 to 128 slightly improves performance (+0.3% F_{20°), it requires 6GB additional GPU memory during training, making the current configuration more practical for computational efficiency.

The ConBiMamba variants show mixed results. ConBiMamba exhibits poor distance localization (53% RDE) despite substantial computational overhead (338M parameters), while

ConBiMambaAC improves efficiency but underperforms BiMambaAC, suggesting potential incompatibility between additional convolutional layers and our asymmetric convolution design.

TABLE II
COMPARISON OF DIFFERENT MODEL ARCHITECTURES ON THE DEVELOPMENT SET FINE-TUNED FROM PRETRAINED PSELDNET

Model	Params	MACs	$F_{20^\circ} \uparrow$	DOAE \downarrow	RDE \downarrow
Conformer	210M	4.69G	38.2%	15.9°	33%
HTS-AT	28M	2.88G	35.1%	16.5°	30%
CRNN	0.7M	57M	22.8%	24.5°	41%
BiMamba	178M	7.57G	36.2%	16.6°	33%
ConBiMamba	338M	7.98G	36.2%	16.9°	53%
ConBiMambaAC	236M	5.02G	33.7%	16.3°	39%
BiMambaAC	76M	4.63G	39.6%	15.8°	33%

TABLE III
ABLATION STUDY OF BIMAMBAAC COMPONENTS

Architectural Variant	$F_{20^\circ} \uparrow$	DOAE \downarrow	RDE \downarrow
Proposed	39.6%	15.8°	33%
w/o bidirectional Mamba	31.0%	17.1°	37%
w/o Asymtric convolution	33.2%	17.6°	39%
Increased States ($d_{state} = 128$)	39.9%	15.5°	35%

Our BiMambaAC method achieved 9th place in DCASE 2025 Task 3 using only single-task training with official dataset and simple channel swapping augmentation. This demonstrates the effectiveness of pretrained model fine-tuning and asymmetric convolution design. As shown in Table IV, our approach outperforms ensemble-based methods while maintaining a simpler single-model architecture compared to higher-ranked methods that rely on external datasets and multiple augmentation strategies.

TABLE IV
COMPARISON OF SELD CHALLENGE SUBMISSIONS BASED ON OFFICIAL EVALUATION DATASET

Team	$F_{20^\circ} \uparrow$	DOAE \downarrow	RDE \downarrow	Aug kinds	Ext. Data	Mode
Wan_XJU [31]	35.4%	18.6°	34.9%	1	1	Single
Zhao_MITC-MG [32]	34.0%	16.8°	36.6%	5	2	Single
Proposed	31.0%	17.4°	40.1%	1	—	Single
Park_KAIST [33]	30.3%	14.6°	32.4%	2	—	Ensemble
Bahuguna_UPF [34]	28.8%	21.2°	100%	—	2	Ensemble

V. CONCLUSIONS

This work presents a stereo SELD approach that combines PSELDnet pretraining with BiMamba sequence modeling. Results show that pre-trained models, especially HTS-AT, significantly improve F1-score and DOAE over the CRNN baseline. BiMamba-based models further enhance performance, confirming the strength of state space modeling for audio sequences. The BiMambaAC model, which integrates asymmetric convolution, achieves the best performance across all metrics with fewer parameters than larger models like ConBiMamba. These findings highlight the benefits of combining BiMamba and asymmetric convolution for accurate and efficient stereo SELD, offering a strong direction for future research.

REFERENCES

- [1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *arXiv*, Jun. 2018. eprint: 1807.00129.
- [2] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *J. AUDIO SPEECH MUSIC PROC.*, vol. 2013, no. 1, pp. 1–13, Dec. 2013, ISSN: 1687-4722.
- [3] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European Signal Processing Conference*, 2010, pp. 1267–1271.
- [4] Y. Xiao, H. Yin, J. Bai, and R. K. Das, "FMSG-JLESS submission for DCASE 2024 task4 on sound event detection with heterogeneous training dataset and potentially missing labels," DCASE 2024 Challenge, Tech. Rep., 2024.
- [5] Y. Xiao and R. K. Das, "WildDESED: An LLM-powered dataset for wild domestic environment sound event detection system," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024, pp. 196–200.
- [6] Y. Xiao and R. K. Das, "UCIL: An Unsupervised Class Incremental Learning Approach for Sound Event Detection," *arXiv:2407.03657*, 2024.
- [7] Y. Xiao, H. Yin, J. Bai, and R. K. Das, "Mixstyle based domain generalization for sound event detection with heterogeneous training data," *arXiv preprint arXiv:2407.03654*, 2024.
- [8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
- [9] Y. Xiao and R. K. Das, *Where's that voice coming? continual learning for sound source localization*, 2025. arXiv: 2407.03661 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2407.03661>.
- [10] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 06–11.
- [11] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization and Detection," *arXiv*, Oct. 2020. eprint: 2010.15306.
- [12] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training," *arXiv*, Oct. 2021. eprint: 2110.07124.
- [13] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A Four-Stage Data Augmentation Approach to ResNet-Former Based Acoustic Modeling for Sound Event Localization and Detection," *arXiv*, Jan. 2021. eprint: 2101.02919.
- [14] J. Hu, Y. Cao, M. Wu, *et al.*, "PSELDNets: Pre-trained Neural Networks on Large-scale Synthetic Datasets for Sound Event Localization and Detection," *arXiv*, Nov. 2024. eprint: 2411.06399.
- [15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *arXiv*, Dec. 2019. eprint: 1912.10211.
- [16] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," *arXiv*, Oct. 2021. eprint: 2110.05069.
- [17] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," *arXiv*, Feb. 2022. eprint: 2202.00874.
- [18] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv*, Dec. 2023. eprint: 2312.00752.
- [19] T. Dao and A. Gu, "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality," *arXiv*, May 2024. eprint: 2405.21060.
- [20] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: State-space model is all you need in speech separation," *arXiv*, Apr. 2024. eprint: 2404.02063.
- [21] Y. Xiao and R. K. Das, *Tf-mamba: A time-frequency network for sound source localization*, 2025. arXiv: 2409.05034 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2409.05034>.
- [22] Y. Xiao and R. K. Das, "XLSR-Mamba: a dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, vol. 32, pp. 1276–1280, 2025.
- [23] X. Zhang, Q. Zhang, H. Liu, *et al.*, "Mamba in Speech: Towards an Alternative to Self-Attention," *arXiv*, May 2024. eprint: 2405.12609.
- [24] D. Mu, Z. Zhang, H. Yue, Z. Wang, J. Tang, and J. Yin, "SELD-Mamba: Selective State-Space Model for Sound Event Localization and Detection with Source Distance Estimation," *arXiv*, Aug. 2024. eprint: 2408.05057.
- [25] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K," *Zenodo*, 2025.
- [26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979, ISSN: 0001-4966.
- [27] B. Rafaely, *Fundamentals of Spherical Array Processing*. Cham, Switzerland: Springer International Publishing, 2019, ISBN: 978-3-319-99561-8. [Online]. Available: <https://link.springer.com/book/10.1007/978-3-319-99561-8>.

- [28] A. Gu, K. Goel, and C. Ré, “Efficiently Modeling Long Sequences with Structured State Spaces,” *arXiv*, Oct. 2021. eprint: 2111.00396.
- [29] K. Shimada, A. Politis, I. R. Roman, *et al.*, “DCASE2025 Task3 Stereo SELD Dataset,” *Zenodo*, Jun. 2025.
- [30] A. Politis, K. Shimada, P. Sudarsanam, *et al.*, “STARSS23: Sony-TAU Realistic Spatial Soundscapes 2023,” *Zenodo*, Mar. 2023.
- [31] Q. Wan, Y. Hu, J. Liu, *et al.*, “A multi-level feature extraction network for sound event localization and detection in dcase 2025 task 3,” DCASE2025 Challenge, Tech. Rep., Jun. 2025.
- [32] T. Zhao, Z. Han, and M. Liu, “Enhancing stereo sound event localization and detection through pre-trained audio representations and hybrid architectures,” DCASE2025 Challenge, Tech. Rep., Jun. 2025.
- [33] J. Park, H. Nam, and Y.-H. Park, “Resnet-conformer for stereo sound event localization and distance estimation in dcase 2025 task3,” DCASE2025 Challenge, Tech. Rep., Jun. 2025.
- [34] A. Bahuguna and R. Peter, “A conformer-based ensemble approach for sound event localization and detection for stereo data,” DCASE2025 Challenge, Tech. Rep., Jun. 2025.