# GISP@HEU'S SUBMISSION TO THE DCASE 2025 CHALLENGE: STEREO SELD TASK

Technical Report

*Congyi Fan*[1†], *Shitong Fan*[1†], *Feiyang Xiao*[1], *Wenbo Wang*[2], *Xinyi Che*[3], *Qiaoxi Zhu*[4], *and Jian Guan*[1*]

[1]Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3]Sichuan University
[4]University of Technology Sydney, Ultimo, Australia

## ABSTRACT

This technical report presents our submission to Task 3 of the DCASE 2025 Challenge. To enhance the model's generalization ability, we adopt the official synthetic data generation pipeline to expand the training set. In addition, SpecAugment is applied for data augmentation to improve event recognition performance. To address the challenges of ambiguous localization and long-range temporal dependencies inherent in stereo SELD, we use the Mamba architecture, which effectively captures both local and global temporal dynamics, thereby improving overall system performance.

***Index Terms—*** Sound event localization and detection, audio-visual fusion, Mamba, SpecAugment

## 1. INTRODUCTION

Sound Event Localization and Detection (SELD) [1, 2, 3, 4, 5] involves identifying target-class sound events, tracking their temporal activity, and estimating their spatial positions. These spatial cues are critical for machine perception tasks such as scene understanding, source tracking, and intelligent environment interaction.

Unlike previous challenges that emphasized four-channel formats such as first-order Ambisonics and microphone array recordings, the DCASE 2025 Challenge shifts focus to SELD using stereo audio [6, 7, 8]. Due to inherent ambiguities in stereo audio—particularly in distinguishing front-back and top-bottom directions—the task limits direction-of-arrival (DOA) estimation to azimuth angles along the left-right axis, and also includes distance estimation. Task 3 of the challenge comprises two tracks: Track A uses audio-only input for SELD, while Track B combines audio and visual inputs with a limited field of view, requiring models to also determine whether events are onscreen or offscreen.

This technical report presents our submission systems for Task 3. Our systems enhance the detection of audio signals or audio-visual information by a hybrid framework based on Transformer [9] and Mamba [10] to fuse audio-visual bimodal features and model long sequence features for the fused features to improve the overall model performance. During the implementation, we adopt a simple network structure rather than a complex redundant one, use simple feature extraction and achieve competitive performance.

†These authors contributed equally to this work.
*Corresponding author.

## 2. OUR SYSTEMS

### 2.1. Track-A: Audio-only inference

**A-System1:** For audio-only input, spatial and semantic cues must be inferred from the temporal and directional patterns in stereo audio. We adopt the Mamba [10] architecture to model long-range dependencies, which is crucial for accurate DOA, distance, and on-screen/offscreen classification under ambiguous, overlapping conditions. With its linear-time complexity and selective state space design, Mamba enables precise tracking of sound sources and enhances the spatial-temporal resolution of audio-only SELD systems.

**A-System2:** Building on A-System1, we incorporate a synthetic data augmentation strategy [11, 6] into the training process, resulting in our submission A-System2.

**A-System3:** Building on A-System1, we apply the SpecAugment strategy [12] to augment the audio data, resulting in our submission A-System3.

### 2.2. Track-B: Audio-visual inference

**B-System1:** Based on B-Baseline (i.e., our replicated version of the baseline[13, 14, 15]), we do not modify the model architecture of the baseline. Specifically, we follow the synthetic data pipeline of the baseline to generate 30,000 synthetic samples. These samples are used to augment the original dataset in order to enhance the generalization capability of the model.

**B-System2:** Building on B-System1, we apply the SpecAugment strategy [12] to augment the audio data, resulting in our submission B-System2.

**B-System3:** Building on B-System1, we adopt a hybrid framework based on Transformer and Mamba. First, we extract audio and visual features and fuse them through cross attention mechanism similar to B-System1; then, the fused features are passed through the Mamba architecture to model long-range dependencies in audio-visual sequences for stereo SELD. Given the limited field-of-view and angular ambiguity of stereo input, capturing extended temporal context is crucial. Mamba enables efficient and expressive sequence modeling through selective state space dynamics, allowing our system to better detect event boundaries, estimate DOA, and classify onscreen/offscreen events, thereby improving the overall spatial-temporal accuracy and robustness of SELD.

## 3. EXPERIMENTS

### 3.1. Dataset

We conducted our experiments on the development dataset [13, 15, 16] of the DCASE 2025 Challenge Task 3. To expand the available training data, we follow the official data synthesis pipeline provided by the challenge organizers to generate 30,000 additional training samples [11, 6].

### 3.2. Experimental Setup

The proposed methods are evaluated on a single NVIDIA 3090 GPU, using the same learning rate and learning rate update strategy as the baseline configuration. For performance comparison, we reproduce the baselines for Track A and Track B, and evaluate them on the development dataset provided by Task 3 of the DCASE 2025 Challenge. The results are referred to as A-Baseline and B-Baseline in Table 1 and Table 2, respectively.

### 3.3. Evaluation Metric

According to the evaluation metrics specified for Task 3 of the DCASE 2025 official competition[1], we adopt the F1-score as the primary metric, which jointly considers detection accuracy, azimuth estimation, distance estimation, and, for audio-visual input, onscreen presence estimation. In addition, we report the macro-averaged Direction of Arrival Error (DOAE), Relative Distance Error (RDE), and Onscreen Accuracy (OSA) to provide a comprehensive evaluation of spatial and visual localization performance.

### 3.4. Results

Table 1 presents the performance of the audio-only systems on the development dataset. Among them, A-System3 achieves the highest $F_{20^\circ}$ score (i.e., 25.28), demonstrating the effectiveness of SpecAugment and Mamba in enhancing localization accuracy.A-System1 yields a slightly lower $F_{20^\circ}$ score (22.94) than the A-Baseline (23.67), but achieves the same RDE (0.32), indicating stable performance in distance estimation. In contrast, A-System2, which uses synthetic data, performs almost worst across all metrics, suggesting that synthetic data may introduce distribution mismatch. These results highlight the benefits of SpecAugment and the importance of careful data selection in Mamba-based systems.

Table 1: Experimental results of the audio-only systems on the development dataset.

| System | $\mathbf{F_{20^\circ}}\uparrow$ | DOAE↓ | RDE↓ |
|---|---|---|---|
| A-Baseline | 23.67 | **21.7**° | **0.32** |
| A-System1 | 22.94 | 23.5° | **0.32** |
| A-System2 | 21.95 | 28.2° | 0.44 |
| A-System3 | **25.28** | 23.0° | 0.45 |

Table 2 summarizes the performance of the audio-visual systems on the development dataset. B-System2 achieves the best $F_{20^\circ\text{/on-off}}$ score (19.63) and the lowest DOAE (22.3°), demonstrating strong performance in both angular and direction estimation. B-System3, which replaces the baseline encoder with Mamba, obtains the best OSA score (0.81) and a competitive RDE (0.37), suggesting better capability in detecting whether the target is on screen. In contrast, B-System1 and B-System2 share similar OSA scores (0.80), but B-System1 yields the worst RDE (0.48) and DOAE (25.8°). These results indicate that combining SpecAugment with synthetic data improves localization accuracy, while Mamba-based models show potential for enhancing screen-aware detection.

Table 2: Experimental results of the audio-visual systems on the development dataset.

| System | $\mathbf{F_{20^\circ/on-off}}\uparrow$ | DOAE↓ | RDE↓ | OSA↑ |
|---|---|---|---|---|
| B-Baseline | 14.17 | 23.7° | **0.33** | 0.76 |
| B-System1 | 16.65 | 25.8° | 0.48 | 0.80 |
| B-System2 | **19.63** | **22.3**° | 0.46 | 0.80 |
| B-System3 | 17.93 | 25.8° | 0.37 | **0.81** |

## 4. CONCLUSION

This technical report presents our systems developed for Task 3 of the DCASE 2025 challenge. The experimental results demonstrate the effectiveness of our proposed systems for the stereo SELD task, which provides performance improvement over the baselines, indicating its potential for enhancing stereo SELD.

## 5. REFERENCES


[1] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.

[2] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, " STARSS23: An Audio-Visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events," in *Proc. of Neural Information Processing Systems (NeurIPS)*, 2023.

[3] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021.

[4] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[5] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

[6] A. S. Roman, A. Chang, G. Meza, and I. R. Roman, "Generating diverse audio-visual 360º soundscapes for sound event


[1] https://dcase.community/challenge2025/

localization and detection," *arXiv preprint arXiv:2504.02988*, 2025.

[7] 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," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[8] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)* , 2024.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Neural Information Processing Systems (NeurIPS)*, 2017.

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

[11] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial Scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[13] D. Diaz-Guerra, A. Politis, P. Sudarsanam, K. Shimada, D. A. Krause, K. Uchida, Y. Koyama, N. Takahashi, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE2024 challenge," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2024.

[14] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2024.

[15] 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," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[16] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.