MULTI-ACCDOA-BASED SELD IN STEREO AUDIO: FEATURE EXTRACTION AND DATA AUGMENTATION STRATEGIES

Technical Report

Bingnan Duan[1] *Yinhuan Dong*[1] *Liuyuan Na*[1]

[1]The University of Edinburgh, School of Engineering, Edinburgh, UK
{bduan, ydong3, lna3}@ed.ac.uk

ABSTRACT

This technical report describes the proposed system submitted to the DCASE2025 Task3: Stereo sound event localization and detection in regular video content (Track A: Audio-only inference). To improve SELD performance, we replace the convolutional blocks in the baseline model with ResNet blocks, extract a 3-channel input feature consisting of log-mel spectrograms and short-term power of autocorrelation (stpACC), and employ two data augmentation techniques: Time Masking and Frame Shuffle. Our system uses the Multi-ACCDOA output representation with an ADPIT loss function to support overlapping sound events. Evaluated on the development dataset, our proposed method achieves significant improvements over the official baseline across F1-score, DOA error, and relative distance error.

Index Terms— Sound Event Localization and Detection, Feature Extraction, Distance Estimation, Stereo Audio

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is a combined task that involves identifying sound events through sound event detection (SED), while simultaneously estimating the spatial characteristics of the sources, including their direction-of-arrival (DOA) and distance [1]. The output of a SELD system supports many machine cognition tasks. These include environment understanding, self-localization, navigation toward hidden targets, sound source tracking, smart-home automation, scene visualization, and acoustic monitoring [1].

Recent advancements in SELD have been driven by two complementary developments: robust model architectures and effective output representations. A commonly adopted baseline is the convolutional recurrent neural network (CRNN) SELDnet model [2], which combines convolutional layers for spatial feature extraction with recurrent layers for temporal modeling. This architecture outputs separate branches for SED and DOA estimation. However, handling overlapping sound events requires additional post-processing to associate detected events with their spatial positions. To address this limitation, the Multi-activity coupled Cartesian DOA (ACCDOA) representation was introduced [3]. Building upon the ACCDOA framework [4], Multi-ACCDOA enables the model to directly output multiple activity-coupled Cartesian DOA vectors per frame, each encoding both the presence and direction of a sound event. As a result, the combination of CRNN and Multi-ACCDOA is used as the baseline system in this year's DCASE SELD challenge [5].

This technical report presents our SELD systems submitted to the audio-only track of the 2025 DCASE Challenge. Specifically, we replaced the ConvBlock in the baseline model, extracted two types of audio features, and applied two data augmentation techniques. Our models were trained solely on the DCASE2025 Task 3 Stereo SELD development dataset. The results demonstrate that our approach achieves significant improvements over the baseline system.

2. PROPOSED METHOD

2.1. Feature Extraction

Selecting the right input features is crucial in designing a SELD system [6]. Experiments on the STARSS24 dataset have shown that combining log-mel spectrograms with intensity vectors (IV) improves both distance estimation accuracy and overall SELD performance [7]. However, unlike previous years where multichannel first-order Ambisonics (FOA) audio was used, the 2025 DCASE Task 3 provides only stereo audio, which does not support the extraction of IV features. To address this limitation, we used stpACC [6] as an additional input feature alongside log-mel spectrograms. Recent empirical results show that stpACC significantly reduces relative distance error (RDE) and enhances SELD performance, validating its effectiveness as a reverberation-based distance cue [6].

To implement the above feature selection, our SELD system uses the following settings for feature extraction. The stereo audio is sampled at 24 kHz. A 1024 point Fast Fourier Transform (FFT) is applied to both log-spectral and stpACC feature extraction using a 40 ms Hanning window and a 20 ms hop length. The resulting 2 channel log-mel spectrogram and 1 channel stpACC feature are then concatenated to form a 3 channel audio feature representation.

2.2. Data Augmentation

After extracting the 3-channel audio features, we apply data augmentation to further enhance model robustness. The official development dataset contains 30,000 audio clips, each 5 seconds long, totaling 41.7 hours. Of these, 16,214 clips are used for training and 13,786 for testing. To improve the generalization ability of our SELD model and reduce overfitting, we employ two complementary data augmentation techniques.

The first technique is Time Masking [8].In this method, contiguous blocks of spectrogram frames—each spanning five 20 ms frames (i.e., 100 ms)—are randomly zeroed out on the two log-mel channels, while leaving the stpACC channel untouched. By aligning each mask with the label grid, we maintain label consistency

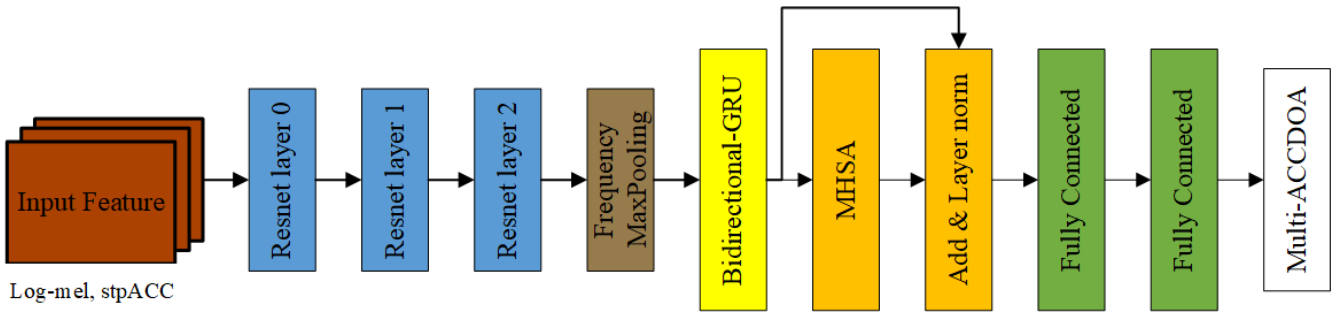

Figure 1: The Architecture of the proposed SELD model.

while encouraging the model to infer sound event activity, direction-of-arrival, and distance from incomplete temporal information.

The second technique, which we propose, is Frame Shuffle. Here, all five log-mel spectrogram frames within each 100 ms block are randomly permuted across the two log-mel channels. This serves as a strong regularizer by introducing new temporal patterns without altering spectral or spatial cues. It effectively increases the diversity of training data without requiring additional recordings.

2.3. Network Architecture

The architecture of our proposed SELD model is illustrated in Figure 1. Our SELD model is based on CRNN-MHSA model [5], which integrates convolutional layers, bidirectional GRUs, and multi-head self-attention (MHSA) layers to capture both local and long-range temporal dependencies. To enhance spatial feature extraction, we replace the baseline ConvBlocks with ResNet blocks [9], allowing deeper feature representation with residual connections.

The output head follows the Multi-ACCDOA format, which directly encodes both the presence and direction of multiple sound events per frame. To support the Multi-ACCDOA output representation, we adopt the Auxiliary Duplicating Permutation Invariant Training (ADPIT) loss function [3], which allows the model to handle multiple overlapping sound events from the same class. In the proposed system, the Multi-ACCDOA vector is extended to include distance estimation, forming the Multi-ACCDDOA representation. Each output track is represented as:

$$\mathbf{y}_{nct} = a_{nct} \cdot \mathbf{R}_{nct}, \quad \text{where} \quad \mathbf{R}_{nct} = [x, y, z, d],$$

where $a_{nct} \in \{0, 1\}$ indicates activity, $\mathbf{R}_{nct} \in \mathbb{R}^4$ denotes a Cartesian vector consisting of a unit DOA vector $[x, y, z]$ and a distance component $d > 0$. The network predicts N such vectors per class per time frame.

The ADPIT loss computes the minimum matching cost over all permutations α of the output tracks:

$$\mathcal{L}_{\text{ADPIT}} = \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} \min_{\alpha \in \text{Perm}[c,t]} \ell_{\text{Multi-ACCDOA}}^{\alpha,c,t},$$

where

$$\ell_{\text{Multi-ACCDOA}}^{\alpha,c,t} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{y}_{\alpha,nct}, \hat{\mathbf{y}}_{\alpha,nct}).$$

Here, $\hat{\mathbf{y}}$ represents the model prediction, and $\mathcal{L}(\cdot)$ is the pointwise loss function. We use mean squared error (MSE) as our primary loss:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2.$$

Following the findings in [5], we also explored mean absolute error (MAE) as an alternative to evaluate the impact on distance estimation. However, we observed that MSE provided a better trade-off across all SELD metrics and was thus chosen as the final loss function for training.

2.4. Training

We train our SELD model using the Adam optimizer with an initial learning rate of 1×10^{-3}. A learning rate scheduler monitors the validation macro F1-score [1] and reduces the rate by a factor of 0.5 if no improvement is observed for 15 consecutive epochs. The model is trained for up to 200 epochs with a batch size of 256. We use only the official DCASE2025 Task 3 stereo SELD development dataset for training and evaluation.

The training loss is computed using the ADPIT framework, combining event activity, DOA, and distance components. We use dropout and early stopping strategies to further mitigate overfitting.

Table 1: Experimental results of the proposed SELD system and the baseline system, evaluated on the development dataset. The metrics of macro F_{20° (%), DOAE, and RDE represent location-dependent F1-score, class-dependent DOA error, and class-dependent relative distance error, respectively.

Model	macro F_{20° (%) ↑	DOAE(°) ↓	RDE(%) ↓
Baseline	22.8	24.5	41
Proposed	**29.0**	**19.3**	**30**

3. RESULTS

We evaluate our proposed SELD system on the development dataset. Table 1 presents the experimental results comparing our proposed SELD system with the baseline system. Our proposed system achieves a macro F_{20° of 29.0%, significantly outperforming the baseline's 22.8%. The DOAE is reduced from 24.5°to 19.3°, demonstrating better spatial localization. Additionally, the RDE is lowered from 41% to 30%, showing more accurate distance estimation. These results confirm the effectiveness of our feature extrac-

tion strategy and data augmentation techniques in enhancing SELD performance across all evaluated metrics.

4. CONCLUSION

In this technical report, we present the proposed system to solve the DCASE 2025 challenge task 3 (audio-only task). By introducing ResNet-based convolutional blocks, combining log-mel and stpACC features, and applying two data augmentation methods, we enhanced the model's ability to detect and localize sound events in stereo audio. Our system leverages the Multi-ACCDOA representation and ADPIT loss function to effectively handle overlapping events. Experimental results on the development dataset show clear improvements over the baseline system in all key metrics, validating the robustness and generalization capability of our approach.

5. REFERENCES

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

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[3] 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 *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[4] 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," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[5] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 286–290.

[6] D. Berghi and P. J. Jackson, "Reverberation-based features for sound event localization and detection with distance estimation," *arXiv preprint arXiv:2504.08644*, 2025.

[7] Y. Dong, Q. Wang, H. Hong, Y. Jiang, and S. Cheng, "An experimental study on joint modeling for sound event localization and detection with source distance estimation," *arXiv preprint arXiv:2501.10755*, 2025.

[8] 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.

[9] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.