STEREO SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION USING DATA-DRIVEN RESNET-CONFORMER ENSEMBLE

Technical Report

Changjiang He[1,2]*, Jian Chen*[1]*, Siyao Cheng*[1,2]*, Jiahua Bao*[1,2]*, jie Liu*[1,2]

[1] Harbin Institute of Technology, Faculty of Computing, China,
[2] State Key Laboratory of Smart Farm Technologies and Systems, China,
cjhe@stu.hit.edu.cn, jianchen_1997@163.com, jhbao@stu.hit.edu.cn, {csy,jieliu}@hit.edu.cn

ABSTRACT

This technical report presents our submitted system for Task 3 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2025 Challenge: Stereo Sound Event Localization and Detection in Regular Video Content (SELD). The DCASE task 3 includes two tracks, and we participate exclusively in the audio-only track. First, we perform data augmentation by employing audio channel swapping (ACS) and data simulation techniques, expanding the dataset to 3.7 times its original size. Subsequently, a single ResNet-Conformer model is used to perform SELD predictions. To further optimize the model and submit multiple model ensemble solutions, we fine-tuned it on the original dataset after training it on the augmented dataset. During model ensemble, we integrate two models—SED-DoA and SED-SDE. Our approach is evaluated on the development test set of the dataset.

Index Terms— Sound event localization and detection, source distance estimation, ResNet-Conformer, model ensemble

1. INTRODUCTION

Given multichannel audio input, a Sound Event Localization and Detection (SELD) system outputs one or more localization estimates for each target sound class whenever such events are detected. SELD plays a crucial role in machine auditory perception, supporting applications such as smart homes [1] and audio-visual scene understanding [2]. Moreover, audio-visual SELD datasets have shown benefits in improving source separation [3] and speech recognition [4]..

The SELD task can be divided into three sub-tasks: Sound Event Detection (SED), Direction of Arrival (DoA), and Source Distance Estimation (SDE). Traditional approaches address these components separately using methods such as Hidden Markov Models (HMM) for SED [5], MUSIC for DoA [6], and DRR for SDE [7]. With the advent of the DCASE challenges and advancements in deep learning, there has been growing interest in using deep neural networks to jointly solve these tasks. From DCASE 2019 [8] to 2024 [9], the SELD challenges utilized four-channel audio data, including first-order Ambisonics (FOA) and microphone array recordings. In contrast, the current challenge adopts stereo audio data to address SELD tasks within common audio and media scenarios. Given the inherent angular ambiguities of stereo audio, particularly in elevation and front-back resolution, this task restricts DoA estimation to azimuthal angles along the horizontal (left-right) plane.

Prior to DCASE 2024, SELD tasks typically included only SED and DoA. [10] simplified the DoA estimation into a multi-direction classification problem and employed a CNN to jointly address both SED and DoA. However, due to the CNN's limited capacity for modeling temporal features, [11, 12] proposed using a CRNN for the SELD task. The CRNN architecture uses a shared backbone to predict both SED and DoA, with the two outputs sharing the CNN and RNN modules except for the final output layers. However, due to limited capacity, CRNNs often struggle to capture complex temporal dependencies, motivating the integration of more expressive architectures such as ResNet-GRU [13], RD3Net [14], ResNet-Conformer [15, 16, 17, 18] , CST-Former [19], and SELD-SSAST[20].

In addition to increasing model capacity, optimizing the output representation is also an effective strategy for improving model performance. Since SELD involves two outputs, balancing the respective losses during training introduces additional challenges. The activity-coupled cartesian direction of arrival (ACCDOA) [21] framework addresses this by integrating the SED and DoA outputs, where the SED output is represented by the magnitude of the DoA vector. The ACCDOA framework elegantly reformulates SELD as a single-task problem by embedding SED confidence into the magnitude of the DoA vector. To address the issue of unrecognized overlapping sound events of the same class in polyphonic scenarios, [22] proposed the Multi-ACCDOA representation. Additionally, [23] introduced the ENV2 approach. With the incorporation of SDE into the SELD task, [24] further proposed a framework that integrates SED and SDE.

Building on these existing works and aligned with the objectives of DCASE2025, we adopt the ResNet-Conformer as our model. We augment the training dataset, increasing its size by a factor of 3.7 through a combination of audio channel swapping (ACS) and synthetic generation methods. Finally, we apply the model ensemble strategy from [18] to integrate the SED-DoA and SED-SDE models.

2. PROPOSED METHOD

2.1. Audio Data Augmentation

The provided dataset consists of 30,000 audio clips, each approximately 5 seconds in length, including 16,214 training samples and 13,786 testing samples, totaling around 41 hours of audio. Compared to the datasets from the previous two years, DCASE2025 features a significant increase in data volume; however, there remains

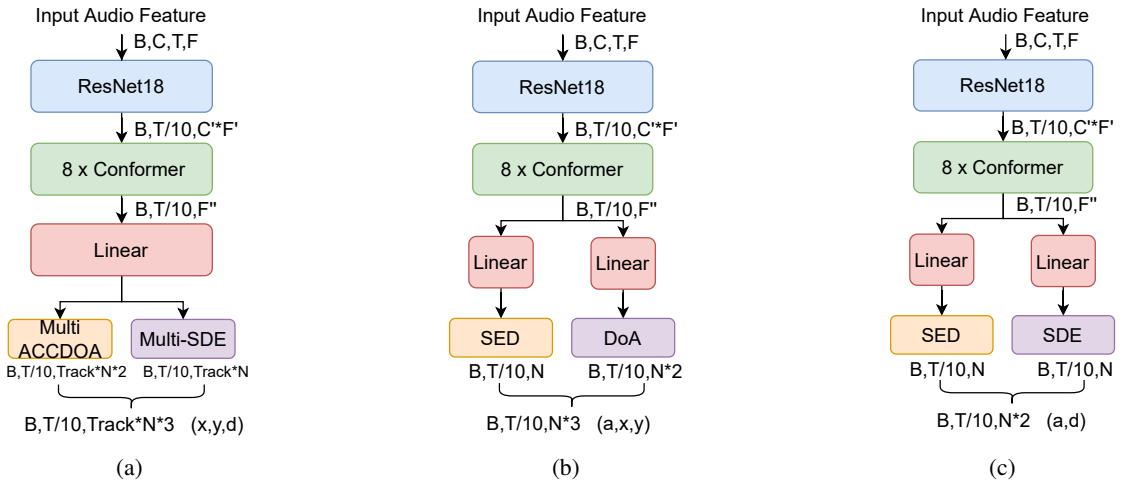

Figure 1: Figures (a-c) illustrate three different audio-only SELD models. The descriptions are as follows: (a) Multi-ACCDOA representation output (Multi-ACCDOA-SDE); (b) SED-DoA representation output; (c) SED-SDE representation output; The dimensions B, C, T, and F denote the input's batch size, feature channel count, feature frames, and Mel bands, respectively. N represents the number of classes, and $Track$ indicates the number of tracks in the Multi-ACCDOA representation. The output represents the output for one class in a single frame, where a indicates whether the sound source is active, (x, y) represents the Cartesian coordinates of the sound source, and d represents the distance of the sound source.

a need for further data augmentation. We apply data augmentation only to the training set, keeping the test set unchanged.

The DCASE2025 dataset is derived from STASS2023 using the Stereo SELD Data Generator. As a first step, we apply the ACS technique to the training set of STASS2023, expanding it to eight times its original size. We then use the Stereo SELD Data Generator to generate 56,214 training samples, which replace the original training set. To further enrich the dataset, we incorporate single-source audio samples from FSD50k and STASS2023 to create a new dataset. Using Spatial Scaper [25], we synthesize 2,400 audio clips, each 60 seconds long. These are subsequently segmented using Stereo SELD Data Generator to produce 40,000 clips of 5 seconds each.

To generate data more closely aligned with the characteristics of STASS2023, we do not use the default initialization parameters when generating datasets with Spatial Scaper. Instead, we adjust the parameters to match the properties of STASS2023 better. In discussing data distribution, our first priority is to ensure that the total number of active event frames within a one-minute interval is consistent with STASS2023. Additionally, we control the maximum polyphony level to align the polyphonic distribution of the generated data with that of STASS2023. The final parameter settings are as follows: Mean number of foreground events in a soundscape = 25, Standard deviation of the number of foreground events = 3, Maximum number of events allowed to overlap at any point in time = 4, Maximum duration of any single sound event in seconds = 10s.

As a result, our final dataset comprises 110,000 audio clips of 5 seconds in length, with 96,214 used for training and 13,786 for testing.

2.2. Features

In previous work, the FOA format was the preferred input representation. However, this year's task adopts a two-channel stereo format. We extract two log-Mel spectrograms sampled at 24 kHz. For the short-time Fourier transform (STFT), a Hann window of 480 points (20 ms) with a hop size of 240 points (10 ms) is used, resulting in a 257-dimensional complex spectrogram. Both the log-Mel spectrogram and intensity vectors are computed using 128-dimensional real-valued vectors.

To enhance the robustness of the submitted results, we incorporate multi-channel features into one of the ensemble systems by constructing stereo-based channels: W+Y and W–Y. First, we sum the original stereo channels to obtain the W channel. Then, using the W+Y, W–Y, and W signals, we compute the intensity vector, following the method described in [20].

2.3. Network Architecture

In this work, we adopt the ResNet-Conformer architecture. As shown in Figure 1(a), the input to the ResNet-Conformer model has the shape (B, 2, 500, 128). Given this input, the ResNet module learns the time-frequency relationships across multiple channels while also capturing inter-channel differences. The output from the ResNet module is reshaped and passed to the Conformer, which models the temporal dependencies and further refines the features. Finally, a fully connected layer maps the features to the output, which is represented in the Multi-ACCDOA format. In the single-model setting, the model uses a single fully connected module for output. Our experiments show that under the hybrid loss, this design achieves superior performance over two separate fully connected layers.

For the model ensemble, we adopt the strategy proposed in Method [18], integrating the SED-DoA and SED-SDE models. The architectures of the SED-DoA and SED-SDE models are illustrated in Figures 1(b) and 1(c), respectively.

Table 1: Add caption

System	Data Aug	Multi	Submit	$F_{20^\circ}(\%)\uparrow$	$DOAE(^\circ)\downarrow$	$RDE(\%)\downarrow$
Baseline	✗	✓	✗	22.8%	24.5	41%
ResNet-Conformer	✗	✓	✗	32.1%	15.7	33.5%
ResNet-Conformer	✓	✓	✓	47.8%	13	30%
SED-DoA	✓	✗	✗	50.4%	13.1	-
SED-SDE	✓	✗	✗	55.9%	-	30.5%
SED-DoA + SED-SDE (1)	✓	✗	✓	50%	13.1	36.6%
SED-DoA + SED-SDE (2)	✓	✗	✓	48.9%	13.1	30.4%
SED-DoA + SED-SDE (3)	✓	✗	✓	51.3%	12.5	33.4%

2.4. Network Training

The maximum number of training epochs is set to 100, with a batch size of 32. We use the Adam optimizer along with a learning rate scheduler. The initial learning rate is set to 0.0001 and is reduced by half if no improvement is observed for 10 consecutive epochs. When fine-tuning on the original dataset, the learning rate is initialized to 0.00001.

3. RESULTS

We evaluate our proposed method using the DCASE2025 Task 3 Stereo SELD dataset. Initially, replacing the baseline model with the ResNet-Conformer under the original dataset scale leads to a noticeable performance improvement. However, due to the limited data size, the ResNet-Conformer only achieves around a 10% improvement in the F_{20°. To further enhance task performance, we apply data augmentation to expand the dataset to 3.7 times its original size. Training the ResNet-Conformer on the augmented dataset results in a 25% improvement in the F_{20° compared to the baseline.

To further boost the final results, we adopt a model ensemble approach combining SED-DoA and SED-SDE models. First, we evaluate the individual performance of the SED-DoA and SED-SDE models. It is evident that excluding one task can significantly improve performance on the remaining task. We submit three ensemble systems: the first is a direct ensemble of the SED-DoA and SED-SDE models; the second further fine-tunes these models on the original dataset; and the third system trains the SED-DoA and SED-SDE models using five-channel features.

4. REFERENCES

[1] S. Krstulović, "Audio event recognition in the smart home," *Computational Analysis of Sound Scenes and Events*, pp. 335–371, 2018.

[2] A. Owens and A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[3] D. Michelsanti, Z. Tan, S. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1368–1396, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3066303

[4] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 7613–7617.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European signal processing conference*. IEEE, 2010, pp. 1267–1271.

[6] Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe, "Multiple source direction of arrival estimations using relative sound pressure based MUSIC," *TASLP*, vol. 29, pp. 253–264, 2021.

[7] H. Chen, T. D. Abhayapala, P. N. Samarasinghe, and W. Zhang, "Direct-to-reverberant energy ratio estimation using a first-order microphone," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 2, pp. 226–237, 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2601222

[8] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 10–14. [Online]. Available: https://dcase.community/workshop2019/proceedings

[9] A. P. D. A. Krause and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv preprint arXiv:2403.11827*, 2024.

[10] hirvonen toni, "classification of spatial audio location and content using convolutional neural networks," *journal of the audio engineering society*, no. 9294, may 2015.

[11] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[12] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "Seld-tcn: Sound event localization & detection via temporal convolutional networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 16–20.

[13] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," DCASE2019 Challenge, Tech. Rep., June 2019.

[14] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The ustc-iflytek system for sound event localization and detection of dcase2020 challenge," DCASE2020 Challenge, Tech. Rep., July 2020.

[15] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of accdoa- and einv2-based systems with d3nets and impulse response simulation for sound event localization and detection," DCASE2021 Challenge, Tech. Rep., November 2021.

[16] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[17] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C.-H. Lee, "The nerc-slip system for sound event localization and detection of dcase2023 challenge," DCASE2023 Challenge, Tech. Rep., June 2023.

[18] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

[19] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8686–8690.

[20] C. He, S. Cheng, J. Bao, and J. Liu, "Adapting single-channel pre-trained transformer models for multi-channel sound event localization and detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.

[21] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 915–919.

[22] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 316–320.

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 885–889. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9413473

[24] H. H. R. W. M. H. S. C. Y. J. M. C. X. F. J. D. Q. Wang, Y. Dong, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

[25] C. He, S. Cheng, J. Bao, and J. Liu, "Adapting single-channel pre-trained transformer models for multi-channel sound event localization and detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.