ResNet-Conformer for Stereo Sound Event Localization and Distance Estimation in DCASE 2025 task3

Technical Report

*Jehyun Park, Hyeonuk Nam, Yong-Hwa Park**

Korea Advanced Institute of Science and Technology, South Korea,
jsasang1234@gmail.com
{frednam, yhpark}@kaist.ac.kr

ABSTRACT

In the DCASE 2025 Task 3 Track A challenge, we propose a ResNet-Conformer architecture for stereo sound event localization and detection (SELD) with integrated sound distance estimation (SDE). We develop two complementary systems. The first system follows a two-stage training strategy, where the model is initially trained to perform sound event detection (SED) and direction-of-arrival (DOA) estimation, and then fine-tuned to also predict source distance. The second system is based on a dual-branch ensemble that combines a model trained for SED and DOA with another model trained for SED and SDE. Both systems share a common backbone consisting of a ResNet-based convolutional encoder followed by an 8-layer Conformer stack, with separate output branches for SED (sigmoid), DOA (tanh), and SDE (ReLU). To enhance robustness, we apply audio channel swapping (ACS) and FilterAugment as data augmentation techniques. Evaluation on the DCASE 2025 Task 3 development set demonstrates that the proposed ensemble system improves overall SELD performance.

Index Terms— sound event localization and detection (SELD) , conformer, ensemble, audio channel swapping(ACS)

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is the task of detecting sound events and simultaneously estimating their spatial positions. It has gained significant attention in recent years for applications such as surveillance, autonomous vehicles, and human-computer interaction. The DCASE 2025 Task 3 [1] extends this task by requiring systems not only to identify sound events and estimate their direction-of-arrival (DOA) but also to predict the distance of each sound source using stereo recordings [2]. This makes the task inherently more complex, as models must infer 3D spatial information from only two audio channels.

To address these challenges, we propose a ResNet-Conformer-based SELD system designed specifically for the audio-only Track A of DCASE 2025 Task 3. Our system is composed of a ResNet-based convolutional encoder for extracting robust local time-frequency features, followed by an 8-layer Conformer block that captures both global and local temporal dependencies. In this study, instead of using the widely adopted multi-accdoa or accdoa [3,4] approaches, we design the network with three separate output branches dedicated to sound event detection (SED), direction-of-arrival (DOA) estimation, and source distance estimation (SDE). Each branch uses a task-specific activation function sigmoid for SED, tanh for DOA, and ReLU for SDE allowing the model to optimize each objective independently and more effectively.

To improve performance and robustness, we design two complementary systems. The first system adopts a two-stage training strategy: the model is first pretrained on the SED and DOA tasks, and then fine-tuned with the additional SDE objective. This gradual approach facilitates stable convergence and prevents negative transfer between tasks. The second system is an ensemble model that combines the outputs of two independently trained networks: one focused on SED and DOA, and the other focused on SED and SDE. This ensemble strategy allows the model to leverage specialized learning for localization and distance tasks while sharing detection knowledge.

In addition, we apply two data augmentation techniques to enhance generalization: Audio Channel Swapping (ACS), which increases spatial variation by mirroring stereo channels, and FilterAugment which simulates frequency domain distortion by randomly masking or modifying frequency bands. Together with the architectural design and training strategies, our proposed system demonstrates strong performance on the development set of the DCASE 2025 Task 3 challenge.

2. PROPOSED METHOD

2.1. Features

Our SELD system uses stereo log-mel spectrograms as the input representation. Each audio segment is converted into a 2-channel time–frequency map with the shape (2, 251, 64), corresponding to approximately a few seconds of audio divided into 251 frames and 64 mel frequency bands. This configuration provides a good balance between spectral resolution and computational efficiency and is commonly adopted in DCASE SELD baselines. Stereo (two-channel) features preserve important spatial information. Differences in intensity and timing between the left and right

channels—such as interaural time and level differences—serve as useful cues for the model to estimate the direction of sound sources.

2.2. Data Augmentation

In this study, we train a stereo-based sound event localization and detection (SELD) model with integrated source distance estimation (SDE) using the DCASE2025 Task 3 Stereo SELD Dataset, which is derived from the STARSS23 dataset [5,6]. STARSS23 consists of multichannel audio in FOA (First-Order Ambisonics) format and 360-degree video recordings, capturing diverse acoustic scenes in various indoor environments. It provides detailed temporal and spatial annotations for prominent sound events. For this challenge, the original STARSS23 data were converted into stereo audio and perspective-view video to simulate typical consumer recording conditions, while preserving the original direction-of-arrival (DOA) and distance (SDE) annotations. With its realistic and varied acoustic scenes, this dataset is well-suited for training and evaluating spatial audio models.

To enhance data diversity and improve the generalization capability of our model, we employ two complementary on-the-fly data augmentation strategies during training: Audio Channel Swapping (ACS) [7] and FilterAugment [8]. These augmentations are designed to introduce additional spatial and spectral variability into the training set, helping the model become more robust in both localization and distance estimation tasks.

the combination of these two augmentation methods—ACS for spatial variability and spectral masking for frequency robustness—significantly improves the model's ability to generalize across environments. It reduces overfitting and enhances performance on SELD task.

2.3. Network Architecture

The core of our model is a ResNet-Conformer-based neural network that maps input spectrograms to multiple task-specific outputs [9]. First, a ResNet-based convolutional encoder processes the 2-channel spectrogram to extract low-level time–frequency features. These features are then passed through a linear projection layer (either a fully connected layer or a 1×1 convolution) to unify the feature dimensions and project them into an embedding space suitable for sequence modeling. Next, we employ an 8-layer Conformer block as the sequence modeling backbone, which effectively captures both global context and local patterns.

At the output stage, the network branches into three task-specific heads corresponding to SED, DOA, and SDE. Each head is trained with a loss function appropriate to its task. This modular output architecture enables independent optimization of each objective and provides the flexibility to adjust or tune the loss weights for each task as needed.

2.4. Network Training

Training a model to simultaneously perform SELD, DOA, and SDE is a challenging task due to the differing nature of each subtask, and the added complexity introduced by the distance estimation (SDE) component.

Table 1: Performance on the official test set of the DCASE2025 Task 3 stereo SELD dataset.

Model	F20°/1 (%)	DOAE (°)	RDE
Baseline	22.8%	24.5°	0.41
Pretrained Model	35.3%	15.5°	0.30
Ensemble Model	**36.3%**	**14.5°**	**0.28**

To maximize performance across all subtasks, we adopt two complementary training strategies: a two-stage multi-task training approach and an ensemble of task-specialized models. These strategies are designed to leverage the benefits of multi-task learning while mitigating the optimization difficulties that arise when learning all three objectives jointly.

The first training strategy adopts a two-stage learning approach to gradually introduce the SDE task. In the initial stage, the model is trained only on the core SELD tasks: sound event detection (SED) and direction-of-arrival estimation (DOA). By excluding distance estimation at this stage, the model can focus on learning robust features for detecting events and estimating their directions without the added complexity of distance prediction. Once the training converges, the second stage begins by attaching an additional output head for SDE to the pretrained model. The model is then fine-tuned jointly on all three tasks—SED, DOA, and SDE—using a lower learning rate to preserve the representations learned in the first stage. This progressive training strategy enables the model to first master relatively simpler tasks before moving on to the more complex task of distance estimation, leading to improved training stability and better SDE performance without degrading the accuracy of the other tasks.

The second strategy enhances performance by ensembling two models [9], each specialized for a different task combination. One model is trained specifically for SED and DOA, while the other focuses on SED and SDE. Both models share the same ResNet-Conformer architecture but are optimized independently according to their task objectives. During inference, the outputs of the two models are combined: SED predictions are merged, DOA estimates are taken from the SED+DOA model, and SDE values are taken from the SED+SDE model. This ensemble setup allows each model to concentrate on the subtasks it handles best, improving overall prediction accuracy. While the use of two models slightly increases inference cost, evaluation on the development set showed clear performance gains, including fewer detection errors, more accurate localization, and more reliable distance estimation. Therefore, the ensemble of task-specialized models provides a practical and effective solution for improving performance.

3. EXPERIMENTAL RESULTS

Table 1 summarizes the evaluation set performance of three models on the DCASE2025 Task 3 stereo SELD test dataset. We compare the official baseline model [10], a two-stage trained model, and an ensemble model that combines the outputs of separate SED+DOA and SED+SDE models. The reported metrics include the location-dependent F1 score with a 20° angular tolerance (F20°/1), Direction of Arrival Error (DOAE) in degrees, and Relative Distance Error (RDE). Higher F20°/1 and lower DOAE and RDE indicate better overall performance.

The ensemble model achieved the best overall performance on the evaluation set. It recorded an F20°/1 score of 36.3%, showing a slight improvement over the single pre-trained model. With a DOA error of 14.5° and an RDE of 0.28, it achieved the lowest values among the three systems, demonstrating the highest localization accuracy and distance estimation precision. Compared to the pre-trained model, the ensemble showed a gain of approximately +1.0 percentage point in F20°/1, a 1° reduction in DOA error, and a slight decrease in RDE. This suggests that combining the SED+DOA and SED+SDE models provides complementary benefits, ultimately resulting in the most accurate SELD system.

In summary, both advanced approaches significantly outperform the baseline in terms of F20°/1 score and DOA/distance errors, demonstrating the effectiveness of pretraining and ensembling strategies for improving spatial sound event detection.

4. CONCLUSION

In this paper, we proposed a ResNet-Conformer-based SELD system designed for the DCASE2025 Task 3 stereo audio track with integrated sound distance estimation (SDE). To address the challenges of simultaneous sound event detection, localization, and distance estimation from stereo input, we explored two complementary approaches: a two-stage training strategy that gradually introduces the SDE task after SED and DOA pretraining, and an ensemble model that combines SED+DOA and SED+SDE outputs. We also employed data augmentation techniques, including audio channel swapping and FilterAugment, to enhance model robustness.

Experimental results on the test set demonstrate that both proposed methods significantly outperform the baseline across all metrics. In particular, the ensemble model achieved the best overall performance, indicating the effectiveness of task specialization and model combination. These findings highlight the benefit of structured training and ensembling in complex SELD tasks involving 3D spatial reasoning from stereo input.

5. REFERENCES

[1] http://dcase.community/workshop2025/.

[2] Daniel A. Krause, Archontis Politis, and Annamaria Mesaros. *Sound event detection and localization with distance estimation.* In European Signal Processing Conference (EUSIPCO), 286–290. 2024.

[3] 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.

[4] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki 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). Singapore, Singapore, May 2022.

[5] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. *STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events.* In Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), 125–129. Nancy, France, November 2022.

[6] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A. Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. *STARSS23: an audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events.* In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, 72931–72957. Curran Associates, Inc., 2023.

[7] 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.

[8] Nam, Hyeonuk, Seong-Hu Kim, and Yong-Hwa Park. "Filteraugment: An acoustic environmental data augmentation method." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[9] Wang, Qing, et al. "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge." *DCASE2024 Challenge, Tech. Rep.* (2024).

[10] David Diaz-Guerra, Archontis Politis, Parthasaarathy Sudarsanam, Kazuki Shimada, Daniel A. Krause, Kengo Uchida, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Takashi Shibuya, Yuki Mitsufuji, and Tuomas Virtanen. *Baseline models and evaluation of sound event localization and detection with distance estimation in dcase2024 challenge.* In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024), 41–45. Tokyo, Japan, October 2024.