A CONFORMER-BASED ENSEMBLE APPROACH FOR SOUND EVENT LOCALIZATION AND DETECTION FOR STEREO DATA

Technical Report

*Arjun Bahuguna**

Universitat Pompeu Fabra
Dept. of Engineering
Barcelona, 08018, Spain
arjunbahuguna251@gmail.com

*Rahul Peter**

Aalto University
Electrical Engineering Dept.
Espoo, 20150, Finland
rahul.peter@gmail.com

ABSTRACT

This report presents our approach to task 3 of the DCASE Challenge 2025[1], which focuses on the localization and detection of stereo sound events (SELD) in regular video content. We propose a three-part ensemble model that operates in the audio domain and outperforms the official baseline. To address class imbalance in the STARSS23[2] dataset, we explore synthetic data generation using SpatialScaper[3] and apply data augmentation techniques such as channel-swapping and time-domain remixing. Our proposed system achieves an F-score of 28%, DOA error of 17.3°, and relative distance error of 0.43 on the development data set. We conclude by suggesting possible future enhancements.

***Index Terms*—** sound event detection and localization, ensemble models, conformer, data augmentation

1. INTRODUCTION

The joint task of Sound Event Localization and Detection (SELD) is critical for machine perception in real-world environments, enabling applications from smart homes to autonomous systems. Traditional SELD systems focused on Sound Event Detection (SED) and Direction-of-Arrival (DOA) estimation [4]. The recent 3D SELD task extends this by including Source Distance Estimation (SDE), providing a more complete spatial understanding of the acoustic scene.

Recent advancements have seen a move from separate modeling of these tasks to joint modeling frameworks. The Activity-Coupled Cartesian DOA (ACCDOA) representation and its multi-track extension (multi-ACCDOA) have been pivotal in unifying SED and DOA prediction into a single output vector, simplifying network design [5]. Architecturally, models have evolved from CNN-RNN structures [6] to more powerful backbones like the Conformer, which effectively captures local and global dependencies in audio feature sequences [7].

For the DCASE 2024 Challenge, a notable strategy involved training separate models for different sub-tasks (e.g., SED-DOA and SED-SDE) and ensembling their predictions, which achieved state-of-the-art results [8]. Our work builds on these insights, proposing an ensemble of specialized Conformer-based models to tackle the SELD task.

*Equal contribution

2. MODEL

Our proposed system is an ensemble of two specialized Conformer-based models. The first model is a multi-ACCDOA system designed to predict class, direction, and distance simultaneously. The second is a task-specific model focused solely on Sound Event Detection (SED) and Direction of Arrival (DOA) estimation. The overall architecture for both models is shown in Figure 1.

2.1. Model Architecture

Both models in our ensemble share a common backbone architecture, which consists of a convolutional front-end followed by a stack of Conformer blocks. This structure is inspired by the successful ResNet-Conformer architecture from the DCASE 2022 Challenge and further adapted based on the NERC-SLIP system for DCASE 2024 [8].

The front-end features three **Convolutional Blocks**. These blocks process the input stereo log-mel spectrograms to extract feature representations. Each block consists of convolutional layers, batch normalization, pooling, and dropout.

The sequential modeling part of the network uses a stack of eight **Conformer Blocks**. Each block integrates multi-head self-attention, depthwise separable convolutions, and feed-forward layers, enabling the model to capture both local and global dependencies in the audio sequence. Each block, operating with an internal dimension (d_{model}) of 128, contributes approximately 220k parameters, with the feed-forward network (≈132k), multi-head self-attention (≈67k), and convolution module (≈21k) being the main components. This results in the Conformer stack having a total of approximately 1.76 million parameters, which forms the majority of the model's complexity.

2.1.1. Multi-ACCDOA Model

This model is designed for the comprehensive 3D SELD task. As shown in Figure 1(a), its output head is composed of fully connected layers that map the Conformer output to the multi-ACCDOA format. This format jointly encodes predictions for up to three simultaneous events (`max_polyphony`: 3) per time frame across all 13 classes.

The associated loss function for this model is the Auxiliary Duplicating Permutation Invariant Training (AD-PIT) loss [5]. This loss function effectively handles multiple overlapping sound events

of the same class by finding the optimal assignment between predicted tracks and ground truth tracks. The core of AD-PIT is to minimize the Mean Squared Error (MSE) on the Activity-Coupled Cartesian DOA (ACCDOA) representation over all possible permutations.

For a given time frame t and class c, with K predicted tracks and J ground truth tracks (where $K = J = 3$ in our case), the loss is calculated. Let $\hat{\mathbf{y}}_{k,c,t}$ be the ACCDOA prediction for track k and $\mathbf{y}_{j,c,t}$ be the ground truth for track j. The AD-PIT loss is formulated as finding the minimum loss over all $K!$ permutations π of the predicted tracks:

$$\mathcal{L}_{\text{ADPIT}} = \min_{\pi \in \mathcal{P}} \sum_{j=1}^{K} \frac{1}{T \cdot C} \sum_{c,t} \left\| \hat{\mathbf{y}}_{\pi(j),c,t} - \mathbf{y}_{j,c,t} \right\|^2 \quad (1)$$

where $\mathcal{P}$ is the set of all permutations of track indices $\{1, ..., K\}$, and the ACCDOA vector $\mathbf{y}$ for a single track is the element-wise product of the sound event activity and its 3D coordinates.

2.1.2. SED-DOA Specific Model

This model is exclusively for the SED and DOA sub-tasks. Unlike the multi-ACCDOA model, it does not predict distance. As illustrated in Figure 1(b), the output head splits the final feature representation into two separate branches for SED and DOA predictions.

The total loss is a weighted sum of the losses from these two branches, a strategy adapted from the NERC-SLIP DCASE 2024 entry [8]. This allows for task-specific tuning, and the loss is formulated as:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{SED}} + \beta \cdot \mathcal{L}_{\text{DOA}} \quad (2)$$

The Sound Event Detection (SED) loss, $\mathcal{L}_{\text{SED}}$, is the Binary Cross-Entropy (BCE) between the predicted activity $\hat{a}_{ct}$ and the ground truth activity a_{ct} for each class c and time frame t:

$$\mathcal{L}_{\text{SED}} = -\frac{1}{CT} \sum_{c,t} [a_{ct} \log \hat{a}_{ct} + (1 - a_{ct}) \log(1 - \hat{a}_{ct})] \quad (3)$$

The Direction of Arrival (DOA) loss, $\mathcal{L}_{\text{DOA}}$, is the Mean Squared Error (MSE) between the predicted DOA vector $\hat{\mathbf{R}}_{ct}$ and the ground truth vector $\mathbf{R}_{ct}$, masked by the ground truth activity a_{ct}:

$$\mathcal{L}_{\text{DOA}} = \frac{1}{CT} \sum_{c,t} \left\| a_{ct} \left(\hat{\mathbf{R}}_{ct} - \mathbf{R}_{ct} \right) \right\|^2. \quad (4)$$

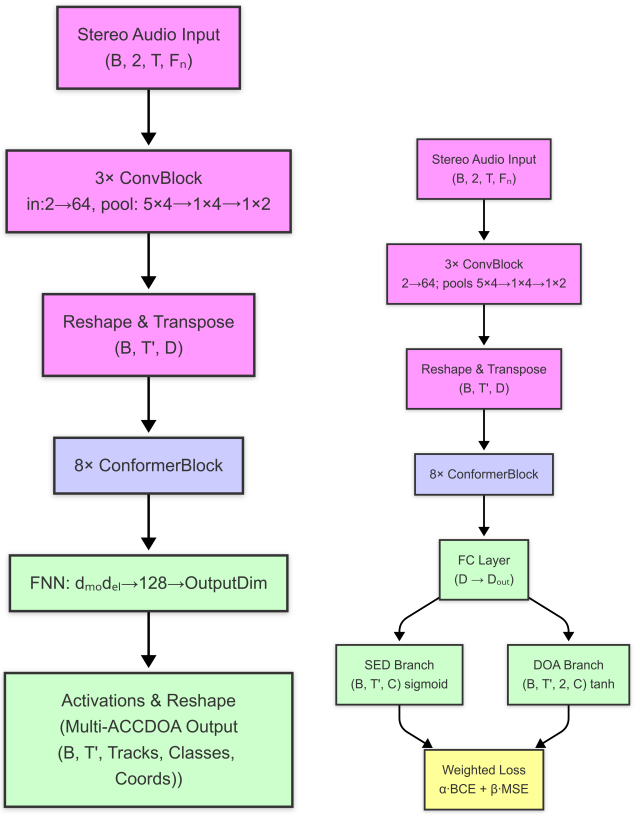
Following the aforementioned work, the weights are set to $\alpha = 0.1$ and $\beta = 1.0$ to prioritize accurate localization [8].

2.2. Ensemble Strategy

Our final system employs a frame-level "winner-takes-all" ensembling strategy, depicted in Figure 2. This method leverages the specialized strengths of each model by consulting a performance lookup table generated from the development-test set [9].

The process is as follows:

1. **Inference**: Both the multi-ACCDOA model (Localization Specialist) and the SED-DOA model (SED Specialist) perform inference on the input audio, generating separate prediction files.
2. **Performance Lookup**: We pre-calculate the F-score and DOA error for each model on a per-class basis using the 'dev-test' split. This creates two lookup tables mapping each class to its expected performance for each model.



(a) Multi-ACCDOA Model (b) SED-DOA Specific Model

Figure 1: Architectures of the two models used in the ensemble. (a) The comprehensive Multi-ACCDOA model. (b) The task-specific SED-DOA model with separate output branches and a weighted loss.

3. **Frame-level Consolidation**: For each frame in the evaluation data, we compare the predictions from both models. We calculate the average F-score of all classes predicted within that frame for each model, using our performance lookup tables.
4. **Winner-Takes-All**: The model with the higher average F-score for a given frame "wins" that frame, and its predictions are written to the final output file. If only one model predicts events in a frame, its predictions are used by default.

This strategy allows the system to select the most reliable prediction at each time step, combining the high F-score of the SED-DOA model with the localization capabilities of the multi-ACCDOA model.

2.3. Data

We train and validate our models on the dataset provided for task 3 by the DCASE Challenge organizers. This dataset has 13 target classes of sound events. As the classes are imbalanced, we use SpatialScaper to generate and augment synthetic data which can balance these classes. We use a quota-system during synthetic data generation, where we first calculate the number of frames assigned to each class and then how many classes need to be added to the

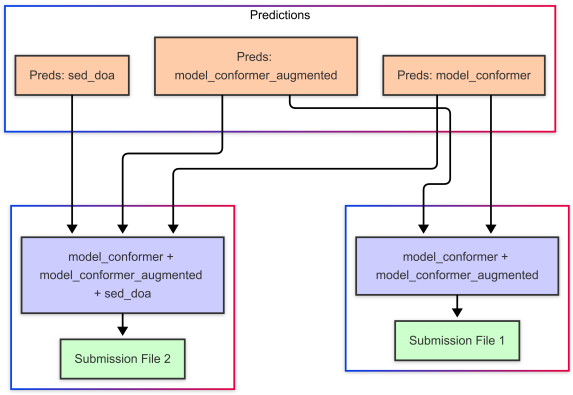


Figure 2: Submission Ensemble Strategy

undersampled classes. The difference is set as the quota per class, the number of frames which SpatialScaper is required to generate.

Apart from quota-based generation, we make some changes to SpatialScaper for our generation process. As SpatialScaper measures rooms in metres, we convert the roomsizes in SpatialScaper to centimetres - this is a done in accordance with a commit by user abreuwallace [10]. We stay close to the distance distribution of the STARSS23 dataset, to avoid localizing sound events at distances that are anamolously larger than STARSS23. We append the sound event datasets used by SpatialScaper, FSD50k[11] and FMA[12], with sound events from the DESED dataset[13]. Finally, we perform data augmentation like channel swapping and time-domain remixing using a pull request by user sivannavis [14].

Finally, all generated data is converted using the dcase2025_seld_generator repository provided by the organizers. We reduce the number of files to be generated from 30k to 20k. Despite our best efforts, we found that using synthetic data did not aid our efforts. We found that generating synthetic data with SpatialScaper did not lead to improved results across classes. An illustrative result of our experiment is shown in Figure 2.3.

Model	F (%)	DOA (°)
Baseline	23.2	22.3
Baseline with synthetic data	3.4	36.0

Table 1: Synthetic data: Performance on dev set.

3. EXPERIMENTS

3.1. Experimental Setup

All models were trained on the official DCASE 2025 Task 3 development dataset in addition to synthetic data for certain rare classes. The development dataset was split into training and testing folds as specified by the challenge organizers.

3.2. Individual Models

We evaluated three individual models to understand their specific strengths and weaknesses:

- **Multi-ACCDOA**: Our primary model for joint SED, DOA, and SDE, trained only on the development data.

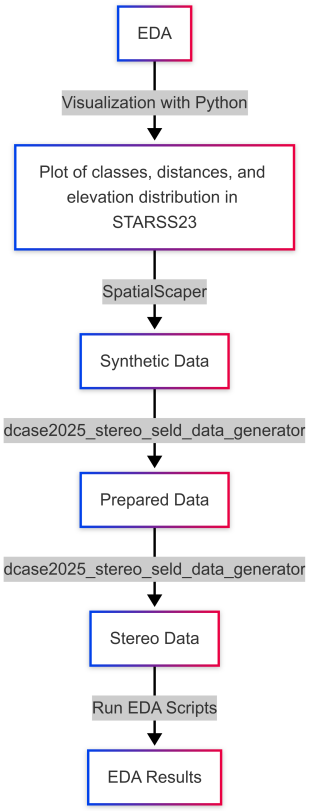


Figure 3: Synthetic Data Generation Strategy

- **SED-DOA**: The specialized model focusing only on SED and DOA, with no distance prediction capabilities.
- **Multi-ACCDOA (Aug.)**: The primary model trained on the development data augmented with our synthetically generated data, which specifically targeted under-represented classes.

4. RESULTS

The performance of our individual models and the final ensemble on the development test set is summarized in Table 4.

Model	F (%)	DOA (°)	RDE
Multi-ACCDOA	24.8	20.6	**0.34**
SED-DOA	**28.4**	20.3	–
Multi-ACCDOA (Aug.)	24.2	23.5	0.36
Ensemble 1 (all)	28.0	**17.3**	0.43
Ensemble 2 (only multi-accdoa)	26.6	20.4	0.36

Table 2: Ensembles: Performance on dev set.

4.1. Analysis of Individual Models

As shown in Table 4, a clear trade-off exists between the models. The **SED-DOA** model achieved the highest F-score (28.4%). However, as it was not trained to predict distance, its distance error is not applicable and its contribution to the final distance prediction is null. Conversely, the **Multi-ACCDOA** model predicted accurate distances, achieving the lowest distance error 0.34 while maintaining the F-score and DOA error.

Our experiments with synthetic data augmentation yielded mixed results. While the **Multi-ACCDOA (Aug.)** model showed improved F-scores for the targeted weak classes (11 and 12), as seen in Table 4.1, its overall performance slightly degraded compared to the model trained without augmentation. This suggests that while targeted augmentation can be beneficial, it may negatively impact performance on more common classes if not balanced carefully.

Class	F-score (No Aug.)	F-score (With Aug.)
Class 11	0.00	0.09
Class 12	0.00	0.11

Table 3: F-score: weak classes with and without data augmentation.

4.2. Ensemble Performance

Our ensemble strategy was designed to combine the strengths of all three models. By using a frame-level "winner-takes-all" approach based on class-wise F-scores, we used the detection capability of the SED-DOA model. For distance, the ensemble always defaulted to the prediction from our distance expert, the Multi-ACCDOA model.

This fusion resulted in a robust final system. The ensemble achieved an F-score of 28.0%, nearly matching our best individual model, while improving the DOA error to a final value of 17.3°. The final distance error of 0.43 cm is a direct result of inheriting predictions from the specialist model, providing a well-balanced system.

Another ensemble was a fusion of the Multi-ACCDOA models (both augmented and unaugmented) which achieved an F-score of 26.6% and a DOAE of 20.4°. The distance error came to around 0.36 which indicated a more well-rounded score while also not taking a massive hit to the rare classes due to augmentation.

5. CONCLUSION

Our report details a Conformer-based ensemble approach for the DCASE 2025 Challenge Task 3, focusing on sound event localization and detection (SELD). Our system utilizes a three-part ensemble model and incorporates synthetic data generation and augmentation to address class imbalance in the STARSS23 dataset.

Future work will explore improving the effectiveness of synthetic data. This includes refining the generation process with SpatialScaper, investigating alternative data augmentation techniques beyond current methods like channel-swapping and time-domain remixing, and developing more nuanced, targeted augmentation strategies that avoid negatively impacting performance on common classes.

Finally, further enhancements to the ensemble strategy are also an important area of exploration in the future. This involves moving beyond the current winner-takes-all approach to investigate more sophisticated methods such as dynamic weighting of models based on real-time performance metrics or confidence scores.

6. REFERENCES

[1] "Dcase 2025 challenge," https://dcase.community/challenge2025/, [Online; accessed 2025-06-14].

[2] A. Politis, K. Shimada, P. Sudarsanam, A. Hakala, S. Takahashi, D. A. Krause, N. Takahashi, S. Adavanne, Y. Koyama, K. Uchida, Y. Mitsufuji, and T. Virtanen, "STARSS23: Sony-TAu Realistic Spatial Soundscapes 2023 (1.0.0)," https://doi.org/10.5281/zenodo.7709052, 2023, data set, Zenodo.

[3] 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," https://arxiv.org/abs/2401.12238, jan 19 2024.

[4] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," https://arxiv.org/abs/2009.02792v2, sep 6 2020, [Online; accessed 2025-06-14].

[5] 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, may 23 2022, pp. 316–320, [Online; accessed 2025-06-14].

[6] T. Komatsu, M. Togami, and T. Takahashi, "Sound event localization and detection using convolutional recurrent neural networks and gated linear units," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, jan 24 2021, [Online; accessed 2025-06-14].

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*. ISCA: ISCA, oct 25 2020, [Online; accessed 2025-06-14].

[8] 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," https://arxiv.org/abs/2501.10755, jan 18 2025.

[9] A. F. Agarap and A. P. Azcarraga, "k-Winners-Take-All Ensemble Neural Network," https://arxiv.org/abs/2401.02092, jan 4 2024, [Online; accessed 2025-06-14].

[10] abreuwallace, ""radius units from m to cm"," https://github.com/marl/SpatialScaper/commit/009ea3393f3c28877884558ffe6690a5eb91aaa8, [Online; accessed 2025-06-14].

[11] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022, [Online; accessed 2025-06-14].

[12] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," https://arxiv.org/abs/1612.01840, dec 6 2016.

[13] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019, [Online; accessed 2025-06-14].

[14] marl, "spatial augmentation on generated soundscapes by sivannavis · Pull Request #47 · marl/SpatialScaper," https://github.com/marl/SpatialScaper/pull/47, [Online; accessed 2025-06-14].