# A FRAMEWORK FOR SELD USING CONFORMER AND MULTI-ACCDOA STRATEGIES

Technical Report

*Priyanshu Kumar, Amit Kumar, Shwetank Choudhary, Jiban Prakash, Sumit Kumar*

Samsung Research Institute Bangalore
{priyanshu.k, amit.kumar, sj.choudhary, p.jiban, sumit.kr }@samsung.com

## ABSTRACT

This technical report describes our submission system for the task 3A of the DCASE 2023 challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes, which uses only audio data as compared to task 3B which leverages audio-visual input. We build our models based on the official baseline system and improve our models in terms of model architecture and data augmentation. Since recent works in Deep Learning, have experimented with replacing traditional Recurrent Neural Networks with Transformer based architectures, we replace the Gated Recurrent Units layers with Conformer blocks. In order to have more training data, we apply Audio Channel Swapping (ACS) augmentation on the DCASE 2023 official dataset. Thus, our experimentations lead to improved SELD score as compared to the official baseline. The proposed system is evaluated on the dev-test set of Sony-TAu Realistic Spatial Soundscapes 2023 (STARS2023) dataset and obtains an improvement of 14.5% in SELD score as compared to the baseline.

***Index Terms*—** Sound Source Localisation, Conformer, Audio Channel Swapping

## 1. INTRODUCTION

The Task 3 of DCASE 2023 Challenge, Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes comprises of the challenge to locate and predict trajectories of sound sources in space. The official dataset provides multichannel audio input and a sound event localization and detection system must outputs a temporal activation track for each of the target sound classes, along with one or more corresponding spatial trajectories when the track indicates activity.

The official dataset of the challenge, STARSS23: Sony-TAu Realistic Spatial Soundscapes 2023 [1] , builds upon the 2022 version of the data by introducing additional 2hrs 30mins of recordings in the development set, from 5 new rooms distributed in 47 new recording clips.

We particularly focus on Track A, which consists of audio only data for the SELD system. We strengthen the official baseline architecture by replacing Gated Recurrent Units layers and Attention layers with Conformer [2] blocks, which have proven their superiority in Automatic Speech Recognition. In addition, we augment the STARSS23 dataset with Audio Channel Swapping [3], which leverages the arrangement of the microphones during data collection.

The rest of the report is organized as follows. In Section 2, the proposed method is described in detail, including data, model architecture and experimental parameters. Experimental results on development dataset is shown in Section 3. Conclusions are summarized in Section 4.

## 2. PROPOSED METHOD

### 2.1. Data

We train our systems with the FOA (First Order Ambionics) dataset. Log-Mel spectrograms are extracted from the 4 channel data of the raw audio files (4 microphones were used while data collection) at 24 kHz sampling rate. Hop length is set at 0.02 times the sampling rate and number of mel bins is set to 64. In addition, 3 channel Intensity Vectors are extracted and concatenated with the 4 channel spectrogram data, leading to a 7 channel input to the model.

In addition to competition data, we also include synthetic data provided with DCASE 2022 challenge. The synthetic data is generated through convolution of isolated sound samples with real spatial room impulse responses (SRIRs) captured in various spaces of Tampere University using sound samples from FSD50K dataset [4].

We experiment with Audio Channel Swapping data augmentation which leverage the tetrahedral positioning of the microphones. It relies on the simple intuition that a rotation of the sound source in intervals of 90 degrees is equivalent to a permutation of the audio channel data. This augmentation technique helps us to increase the data with a 8x factor. The augmentation is not applied on the synthetic data as it leads to deterioration of performance. We also experiment with SpecAugment augmentations like frequency masking and time masking.

### 2.2. Model Architecture

The official baseline architecture comprises of a convolutional module which extracts features from the spectrogram and intensity vector features. The input of the model consists of 250 frames of data, which is downsized to 50 frames data and passed to the recurrent neural network layers. We replace these recurrent layers with 8 Conformer blocks with input dim of 128, feed forward dim of 512, 8 attention heads and depthwise convolution kernel of size 3.

### 2.3. Training Details

The models are trained in Multi ACCDOA [5] setting so as to handle multiple overlapping sound events. The models are trained for 300 epochs with a learning rate of $5e-4$. For experiments with SpecAugment, we train the model without SpecAugment for the first half of the training and then turn on the augmentation for the second half (we apply time and frequency masking augmentations). We train our models on folds 1, 2 and 3 of the dataset and evaluate

| Model | Experiment Description | F1 | ER | LE | LR | SELD Score |
|---|---|---|---|---|---|---|
| Baseline | DCASE 2023 + Syn Data | 0.61 | 0.31 | 22.41 | 0.51 | 0.48 |
| Conformer | DCASE 2023 + Syn Data | 0.60 | 0.36 | 20.83 | 0.58 | 0.44 |
| Baseline | DCASE 2023 + Syn Data + Aug Data | 0.56 | 0.39 | 18.89 | 0.54 | 0.43 |
| Conformer | DCASE 2023 + Syn Data + Aug Data | 0.56 | 0.39 | 20.30 | 0.63 | **0.41** |
| Conformer | DCASE 2023 + Syn Data + Aug Data + SpecAug | 0.60 | 0.38 | 19.54 | 0.55 | 0.44 |
| Conformer | DCASE 2023 + Syn Data + Aug Data + Progressive SpecAug | 0.56 | 0.39 | 20.30 | 0.63 | **0.41** |

Table 1: Evaluation results of DCASE validation split (fold 4)

on fold 4. Random seed is set to 42 to ensure reproducibility of results.

## 3. EVALUATION RESULTS

We tabulate the results of our experiments in Table 1. We compute the F1 score at 20 degrees, Error Rate (ER), class-aware Localization Error (LE), Localization Recall (LR) and SELD score which is an aggregate of the previous 4 metrics.

We observe that ACS data augmentation with Conformer based model achieves the best performance. Our experiments of using SpecAugment progressively i.e. using it only in the second half of the training, produces the same model as the former model. Despite a spike in the training loss at the beginning of the second half of the training, the time and frequency masking augmentation do not provided any nodel learning signal to the model.

## 4. CONCLUSION

In this report, we present our experimentations for Task3A SELD task in DCASE 2023 challenge. We experiment with data augmentation approaches to expand the official dataset. Architectural enhancements of using a Conformer based architecture provides considerable improvements in the evaluation metrics.

## 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,” *arXiv preprint arXiv:2206.01948*, 2022.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[3] 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.

[4] 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, 2021.

[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, 2022, pp. 316–320.