JLESS SUBMISSION TO DCASE2024 TASK3: Conformer with Data Augmentation for Sound Event Localization and Detection with Source Distance Estimation

Technical Report

Wenqiang Sun[1], *Dongzhe Zhang*[1,2], *Jisheng Bai*[1,2], *Jianfeng Chen*[1,2]

[1] Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
[2] LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{sunwenqiang, dongzhezhang2022, baijs}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

This technical report, we describe our proposed system for DCASE2024 task3: Sound Event Localization and Detection(SELD) with Source Distance Estimation in Real Spatial Sound Scenes. At first, we review the famous deep learning methods in SELD. To augment our dataset, we employ channel rotation techniques. In addition to existing features, we introduce a novel feature: the sine value of the inter-channel phase difference. Finally, we validate the effectiveness of our approach on the Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23) dataset and the results demonstrate that our method outperforms the baseline across multiple metrics.

Index Terms— sound event localization and detection, Data augmentation, model ensemble, real spatial scenes

1. INTRODUCTION

The objective of the Sound Event Localization and Detection (SELD) task is to detect occurrences of sound events from specific target classes, track their temporal activity, and estimate their directions-of-arrival (DOA) or positions. Given multichannel audio input, a SELD system outputs localization estimates of one or more events for each target sound class whenever such events are detected [1]. This process results in a spatial temporal characterization of the acoustic scene, which can be applied to a wide range of machine cognition tasks. These tasks include environmental inference, self-localization, navigation with visually occluded targets, tracking specific types of sound sources, smart-home applications, scene visualization systems, and acoustic monitoring, among others [2,3].

The SELD system was first introduced in the DCASE2019 Task 3, focusing on single static sound sources. In this task, multichannel audio files were synthesized by combining mono audio files with impulse responses in real rooms, allowing manual control over factors such as signal-to-noise ratio (SNR), event occurrence, and arrival direction. However, subsequent SELD challenges introduced several new complexities, including new impulse responses, moving sources, polyphonic events, and overlapping events of the same class [4,5,6,7]. This year, the challenge introduces distance estimation of the detected events [8], significantly increasing the task's difficulty. Evaluation metrics have also been updated to account for this additional dimension.

In this report, we propose a SELDnet-based neural network with data augmentation for SELD. The entire framework of our SELD system is built upon two main components: SELDnet and multi-track ACCDDOA [9,10]. SELDnet is a neural network architecture that integrates spatial information with spectrogram representations to accurately classify and localize sound events. The multi-track ACCDDOA algorithm addresses same-class overlapping sound events and extracts precise localization information for each event. To prevent the model from overfitting on the synthesized data, we employ a strategy of training the model on a combination of real and synthesized data, followed by fine-tuning on real recordings.

2. PROPOSED METHOD

In this section, we first introduce the input features of the proposed SELD system. Then we introduce the data augmentation, network architecture and training procedures.

2.1. Features

Our network's input features consist of signals from four channels of first-order ambisonics (FOA). We emphasize using FOA signals because they do not contain spatial aliasing within the range of 9 kHz. Additionally, the FOA format was preferred as it outperformed the MIC format in the baseline system. The FOA features comprise ten channels, including four log-mel spectrograms and three intensity vectors. Furthermore, we introduce a new feature set [11]: the sine values of the phase differences of the Short-Time Fourier Transform (STFT) after passing through a mel filter bank. This new feature set, which includes three channels, is processed through the mel filter bank to ensure the dimensions are consistent with the other feature sets.

2.2. Data augmentation

Since the dataset provided by DCASE comprises only 1200 synthetic files, we augmented the training data to enhance the model's performance. Table 1 shows 16 patterns of channel rota-

Table 1: 16 patterns of channel rotation. Swap(X, Y) denotes $X' \leftarrow Y$ and $Y' \leftarrow X$

	ϕ - $\pi/2$	ϕ	$\phi + \pi/2$	$\phi + \pi$
θ	Swap(-X,Y)	-	Swap(X,-Y)	$Y' \leftarrow -Y$, $X' \leftarrow -X$
-θ	Swap(-X,Y), $Z' \leftarrow -Z$	$Z' \leftarrow -Z$	Swap(X,-Y), $Z' \leftarrow -Z$	Swap(-X,-Y), $Z' \leftarrow -Z$
	-ϕ - $\pi/2$	-ϕ	-$\phi + \pi/2$	-$\phi + \pi$
θ	Swap(-X,-Y)	$Y' \leftarrow -Y$	Swap(X,Y)	$X' \leftarrow -X$
-θ	Swap(-X,-Y), $Z' \leftarrow -Z$	$Y' \leftarrow -Y$,$Z' \leftarrow -Z$	Swap(-X,Y),$Z' \leftarrow -Z$	$X' \leftarrow -X$,$Z' \leftarrow -Z$

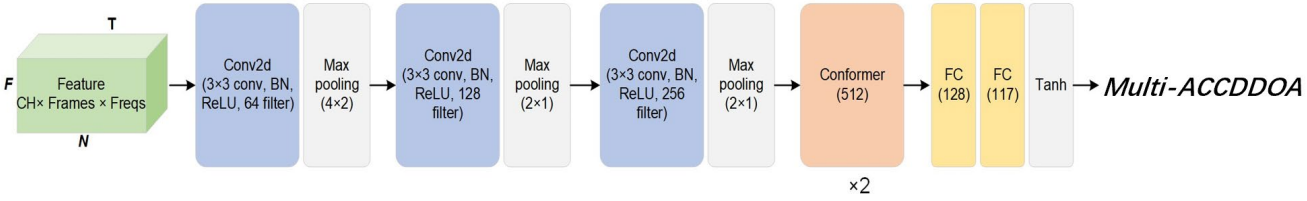

Figure1: Overall architecture of the proposed network.

tion [12], but we applied eight spatial transformation methods, chosen to rotate the audio channels in a way that transforms θ to −θ, resulting in a more uniform distribution of. This approach ensures a more balanced spatial representation. During the channel rotation, we also adjusted the spatial labels accordingly, but the distance labels remained unchanged. Because we performed two different channel rotations for each audio file, we tripled the amount of training data.

2.3. Network architecture

The model was created based on the baseline CRNN structure, incorporating the multi-activity-coupled Cartesian Distance and DOA (multi-ACCDDOA) format. This format extends the known multi-ACCDOA format by including distance in the estimated vector. Figure 1 shows the overall structure of the proposed model. The primary aim of our network is to extract spatial information from the given input of FOA features. This is achieved by feeding the log-mel spectrograms from all four FOA channels, along with the IV channels and sin-IPD, into a convolutional network comprising three layers. The multi-ACCDDOA format can predict Sound Event Detection (SED), Direction of Arrival (DOA), and distance through a single branch. Additionally, bidirectional GRU layers are replaced with two Conformer blocks [13].

3. EXPERIMENTS

In this section, we show our results on the development dataset.

3.1. Experimental settings

We evaluated our proposed methods on the Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23) dataset and compared our systems with the baseline system. The baseline is a multi-ACCDDOA-based system using a CRNN network. Three metrics are used for evaluation: macro $F_{20^\circ/1}$, DOAE, RDE. We use only the FOA subset of the dataset for out experiments.

We followed the baseline settings during feature extraction. The sampling frequency was set to 24kHz, the number of Mel filters was set to 64, and the STFT was applied with a 40ms frame length and a 20ms frame hop. The input length was set to 250 frames, and we used a batch size of 64. In addition to these settings, we incorporated the computation of the sine of the IPD values to smooth phase variations. The training process involved initially training on the synthesized dataset for 150 epochs with a learning rate of 0.001. The learning rate decayed by 0.5 every 50 epochs. During the fine-tuning phase, the model was trained on the real dataset for 30 epochs with a learning rate of 0.00005.

3.2. Results.

Table 2 shows the performance of our proposed methods on the development set. As indicated in the table, our proposed method significantly outperforms the baseline in terms of $F_{20^\circ/1}$ and DOAE. However, it is important to note that the RDE metric did not show an improvement over the baseline.

4. CONCLUSION

We present the proposed SELD system of DCASE2024 task3. We apply data augmentation methods and three additional channels of phase information to augment the training data. Considering the differences between simulated spatial audios and real recordings in exclusive environments, we employed different strategies during the training stage to improve the system's generalization in realistic environments. Our proposed system achieved substantial improvements and significantly outperformed the baseline system across multiple metrics.

SELD performance of our systems evaluated by using joint metrics for the development set

system	*macro* $F_{20^\circ/1}$ (%)↑	DOAE(°) ↓	RDE(%)↓
baseline-FOA	13.1%	36.9°	33%
model1	29.2%	20.7°	47%
model2	21.7%	26.5°	48%

5. REFERENCES

[1] Shimada K, Koyama Y, Takahashi N, et al. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 915-919.

[2] Krause D, Politis A, Kowalczyk K. Comparison of convolution types in CNN-based feature extraction for sound source localization[C]//2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021: 820-824.

[3] Mesaros A, Diment A, Elizalde B, et al. Sound event detection in the DCASE 2017 challenge[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(6): 992-1006.

[4] https://dcase.community/challenge2020.

[5] https://dcase.community/challenge2021.

[6] https://dcase.community/challenge2022.

[7] https://dcase.community/challenge2023.

[8] http://dcase.community/workshop2024/.

[9] Shimada K, Koyama Y, Takahashi S, et al. Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 316-320.

[10] Krause D A, Politis A, Mesaros A. Sound Event Detection and Localization with Distance Estimation[J]. arXiv preprint arXiv:2403.11827, 2024.

[11] Krause D, Politis A, Kowalczyk K. Feature overview for joint modeling of sound event detection and localization using a microphone array[C]//2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021: 31-35.

[12] Mazzon L, Koizumi Y, Yasuda M, et al. First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation[J]. arXiv preprint arXiv:1910.04388, 2019.

[13] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, “Conformer: Local features coupling global representations for visual recognition,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 367–376.