THE IASP SUBMISSION FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2024 CHALLENGE

Technical Report

Yuanhang Qian, Tianqin Zheng, Yichen Zeng, Gongping Huang

School of Electronic Information, Wuhan University, 430072, Wuhan, China,
{qianyuanhang, 2020302121085, zengyichen, gongpinghuang}@whu.edu.cn

ABSTRACT

The technical report describes the submission systems developed for task 3a of the DCASE2024 challenge: Audio Sound Event Localization and Detection with Source Distance Estimation. To enhance the performance of the audio-only task, we implement audio channel swapping as a data augmentation technique. We adopt the Resnet-Conformer model for the network architecture, which is well-suited for capturing First-Order Ambisonics (FOA) format data patterns. Additionally, the approach utilizes the Multi-ACCDDOA method to concurrently predict the event type and estimate the source distance. This comprehensive strategy yielded superior results compared to the baseline system.

Index Terms— Sound event localization and detection, data augmentation, First-Order Ambisonics, Resnet-Conformer

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is a critical task in the field of audio signal processing, aiming to detect various activating sound events and accurately localize them in both temporal and spatial domains. Since the inception of the DCASE challenge in 2019, the requirements and datasets for SELD have evolved significantly. Initially, the challenge provided emulated multichannel recordings, where spatialized event sample banks were combined with spatial room impulse responses (SRIRs) and mixed with spatial ambient noise. This approach persisted through the first three years (2019-2021). In 2022, the dataset shifted to real sound scene recordings with manual annotations. The DCASE 2024 challenge resembles the previous iteration, evaluating SELD models on manually annotated recordings of real interior sound scenes and introduces distance estimation of the detected events.

In our work, we propose enhancements to the SELD system by integrating Resnet-Conformer neural network architecture and Audio Channel Swapping (ACS) data augmentation techniques [9]. We employ several conformer structures, which combines convolutional neural networks (CNNs) with self-attention mechanisms to better capture and utilize the spatial and temporal features of sound events. The developed system also leverags the Multi-ACCDDOA output format [7] to simultaneously predict the class, location, and distance estimation of events.

The rest of the report is organized as follows: Section 2 describes the proposed method and training process in detail. Section 3 presents the experimental results on the development dataset. Finally, Section 4 concludes the report, summarizing the findings and outlining future work.

2. PROPOSED METHOD

2.1. Features

In our method, we selected FOA format audio as the input. We chose two kinds of features. First, we utilized the log-mel energy spectrograms, transforming the input into a 4-channel mel energy representation. Second, to capture the spatial characteristics of the audio, we employed a 3-channel intensity vector. These two features are concatenated into a 7-channel feature matrix as the input of the model.

2.2. Data augmentation

The official dataset contains 7.5 hours of audio data. To increase the amount of data, we employed the ACS data augmentation method proposed in [6], expanding the dataset by eight times. ACS is a spatial augmentation method for FOA datasets, where new Direction of Arrival (DOA) information is generated by systematically swapping the four FOA channels. This process maintains the spatial characteristics of the original sound field while creating valid augmented data.

2.3. Network Architecture

We employed the ResNet-Conformer architecture proposed in [9] as the backbone of our model. The network architecture is illustrated in Figure 1. The first component is the ResNet module, which comprises multiple residual blocks. Each residual block is capable of learning hierarchical features, thereby effectively extracting both spatial and temporal characteristics. Additionally, we incorporated four frequency domain pooling layers within the ResNet structure to reduce the dimensionality of the feature maps. Subsequently, 8 Conformer modules are utilized for sequential feature extraction and global dependency modeling. The Conformer integrates CNN with the self-attention mechanism of Transformers, enabling it to capture both local and global features, which is particularly advantageous for SELD tasks. The output of the Conformer is processed through temporal pooling and subsequently passed through two fully connected layers, with intermediate dimensions of 256 and 128, respectively.

3. EXPERIMENTS

3.1. Dataset

We evaluated our approach using the official STARSS2023 dataset. This dataset provides 4-channel 3-dimensional FOA and MIC for-

Table 1: The evsluation results of our system for dev-test set

Model	Data Description	$F_{20}(\%)$	$AE(°)$	$RDE(\%)$
Baseline	Base	13.1%	36.9°	**33**%
Baseline	Base + Syn Data + ACS	15.2%	26.1°	55%
ResNet-Conformer	Base + Syn Data	16.3%	30.6°	42%
ResNet-Conformer	Base + Syn Data + ACS	**23.0**%	**25.1**°	42%

mats and is divided into a training set with 90 recording clips and a test set with 78 recording clips [8]. Additionally, we utilized 1200 one-minute synthetic recordings, which are generated through convolution of isolated sound samples with real spatial room impulse responses (SRIRs). All audio data were sampled at 24 kHz, encompassing 13 target event categories with an common overlap of 3 events.

3.2. Evaluation metrics and results

We evaluated our project using the official metrics provided. The evaluation metrics include the location-dependent F1 score, DOA error (AE), and relative distance error (RDE).

Table 1 presents a performance comparison between our model and the baseline under different dataset configurations. It is evident that when using the ResNet-Conformer network and ACS data augmentation, our system achieves the highest F1 score and the best AE performance. However, the RDE metric is lower compared to the baseline RDE. We will further extend our work and address the issues related to distance estimation in the future.

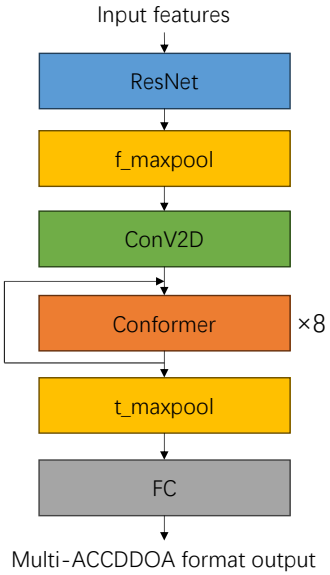

Figure 1: Architecture of our model.

4. CONCLUSION

In this report, we presented the performance of our system for the DCASE2024 Challenge Task 3. Utilizing the ResNet-Conformer architecture and ACS data augmentation method significantly enhances the performance of SELD tasks in terms of event classification and localization. However, the system still has limitations in distance estimation. Therefore, in the next step, we will explore more suitable features and network architectures to address this issue.

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,"* in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, *"Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,"* CoRR, vol. abs/1807.00129, 2018. [Online]. Available: http://arxiv.org/abs/1807.00129

[3] P. Sudarsanam, A. Politis, and K. Drossos, *"Assessment of self-attention on learned features for sound event localization and detection,"* CoRR, vol. abs/2107.09388, 2021. [Online]. Available: https://arxiv.org/abs/2107.09388

[4] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, *"SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays,"* in *ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022. [Online]. Available: https://doi.org/10.1109/icassp43922.2022.9746132

[5] C. Schymura, B. T. Böninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, *"PILOT: Introducing transformers for probabilistic sound event localization,"* CoRR, vol. abs/2106.03903, 2021. [Online]. Available: https://arxiv.org/abs/2106.03903

[6] 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,"* arXiv preprint arXiv:2101.02919, 2021. [Online]. Available: https://arxiv.org/abs/2101.02919

[7] 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.

[8] 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 (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72931–72957. Curran Associates, Inc., 2023.

[9] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, *"An experimental study on sound event localization and detection under realistic testing conditions,"* in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] `https://dcase.community/challenge2024/`.