THE SYSTEM USING CONVNEXT, CONFORMER, AND DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Jiahao Li

Beijing Institution of Technology, China
3220230786@bit.edu.cn

ABSTRACT

This technical report details our submission system for DCASE2024 Task 3: Audio and Audiovisual Sound Event Localization and Detection (SELD) with Source Distance Estimation. To address the audio-only task, we initially apply the Audio Channel Swapping (ACS) method to generate augmented data, enhancing the performance of the proposed system. Subsequently, we introduce the ConvNeXt module for feature extraction and processing. To further enhance feature extraction capabilities, we employ the Squeeze-and-Excitation Block (SEBlock) after ConvNeXt. We then utilize the Conformer to extract additional features and ultimately compute the multi-ACCDOA output. The proposed system significantly outperforms the baseline on the development dataset of DCASE2024 Task 3.

***Index Terms*—** DCASE2024, data augmentation, Sound event localization and detection, attention

1. INTRODUCTION

The goal of the sound event localization and detection (SELD) task is to detect occurrences of sound events belonging to specific target classes, track their temporal activity, and estimate their directions-of-arrival (DOA) or positions during those events. SELD systems can be applied in various fields such as robot auditory systems, smart home systems, virtual reality (VR), and augmented reality (AR).

The annual Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge has consistently drawn researchers' attention to SELD, particularly Task 3, leading to significant advancements in this area. From 2019 to 2021, Task 3 utilized a simulated dataset created through spatial room impulse responses (SRIRs) combined with sound events. In 2022, the dataset transitioned to real spatial sound scene recordings. The 2023 challenge further advanced the task by incorporating an audio-visual track, which included simultaneous 360-degree video recordings accompanying the audio recordings and additional source distance information in the labels. This year, Task 3 also introduced distance estimation, adding another layer of complexity to the task.

The DCASE challenge provides a baseline system that employs a neural network architecture combining Convolutional Recurrent Neural Networks (CRNN), Gated Recurrent Units (GRU), and Multi-Head Self-Attention (MHSA) to address the SELD task[1, 2].

In this report, we focus on the Audio-only track and propose our system. We utilize the STARSS23 dataset as our primary training dataset. To enhance the performance of the proposed system, we incorporate simulated data used in baseline training, generated using the FSD50 dataset and TAU-SRIR, and perform Augmented Circular Shift (ACS) data augmentation on the real data during training[3]. Our network structure is modified from the baseline by integrating ConvNeXt[4] and Conformer[5] architectures, with the addition of Squeeze-and-Excitation (SE) blocks[6] after each layer of ConvNeXt to improve the model's feature extraction capabilities. Experimental results demonstrate that our proposed model outperforms the baseline model.

2. PROPOSED METHOD

2.1. Data Augmentation

Although STARSS23 provides a certain amount of real data, the amount of data still cannot meet the robustness requirements of model detection. The proposed method uses the Audio channel swap method to achieve data enhancement. We used 16 rotation methods to rotate the audio channels and update the labels in the same form to enhance the audio data in the FOA format. The proposed system only chooses to perform ACS on real data. This is because performing ACS on simulated data may make the training time too long and will not bring significant improvement. Table 1 shows 16 rotation methods. After implementing ACS through rotation, the amount of real data increases by 16 times.

2.2. Network Architecture

The Baseline system adopts the CRNN structure, which consists of 3 layers of CNN, 1 layer of GRU, 2 layers of MHSA, and 2 layers of FNN. This system makes adjustments based on the framework of the baseline. First, a 5-layer 5×5 convolution ConvNeXt is used to replace CNN, and an 8-layer Conformer is used to replace GRU and MHSA, while the FNN structure remains unchanged. In order to improve the model's feature extraction capability, there is a Res-SEblock after each layer of ConvNeXt to further enhance the network's ability to extract and process features and increase the global receptive field. Figure 1 shows the framework of the proposed system. The input of the model is the log-mel spectrum of the 4-channel FOA, which is connected to the 3-channel acoustic intensity vector. The output is multi-ACCDOA[7], which contains the information of the sound event and location that occurred at a certain time. The loss function of the network is ADPIT MSE Loss, which is the same as the baseline.

Table 1: ACS implemented through 16 rotations.$Swap(X, Y)$ means $Y \leftarrow X, X \leftarrow Y$

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

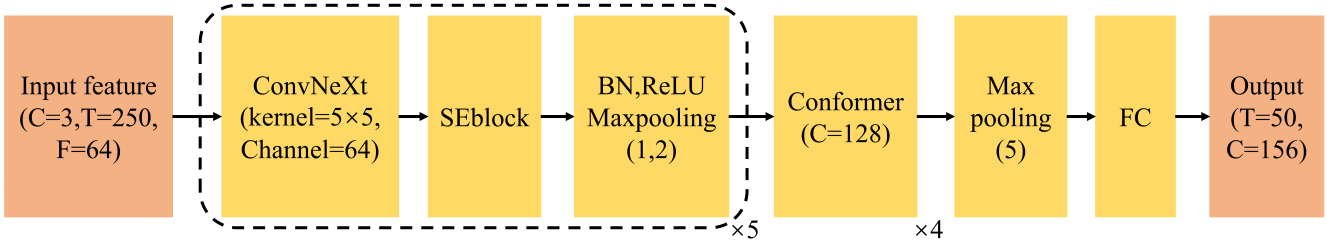

Figure 1: Example of a figure with experimental results.

3. EXPERIMENTS

3.1. Dataset and Settings

The proposed system is trained on STARSS23[8] and simulated data generated by FSD-50K and TAU-SRIRDB[9]. The sampling rate is set to 24kHz, the number of mel filters is set to 64, and the STFT is used with 40ms frame length and 20ms frame hop. The length of the input feature is 250 frames, and the length of the input label is 50 frames. The batch size is 128, the learning rate is 0.001, and the Adam optimizer is used. A total of 400 epochs are trained. The evaluation indicators are location-dependent F1 score (F), DOA error (AE), and relative distance error (RDE), which are consistent with the baseline system. We only use the FOA subset in our experiments.

3.2. Result

Table 2 shows the results of our system for the DCASE2024 Task 3 Audio-only track on the development dataset. As shown in Table 2, the proposed method outperforms the baseline system. This superior performance is because we adopt multiple modules, which enables the proposed system to extract more complex features, and the ACS data augmentation method makes the model more robust.

Table 2: Comparison of the proposed model with the baseline system on the development set

	F1	AE	RDE
Baseline	11.3%	38.4°	36%
Submission	33.9%	21.1°	30%

4. CONCLUSION

This report presents the proposed system for solving the DCASE2024 Task3 Audio-only track. We used the ACS data augmentation method to generate simulated data and expand the official dataset. At the same time, we used modules such as ConvNeXt, SEblock, and Conformer to improve the model's ability to extract features. The experimental results show the improvement of the proposed system over the baseline system.

5. REFERENCES

[1] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022.

[2] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv*, 2024.

[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] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[5] 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.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[7] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022.

[8] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c9671ed3b3106b71cafda3ba225c1a-Abstract-Datasets_and_Benchmarks.html

[9] 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," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, April 2024.