MULTI-SCALE FEATURE FUSION FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Da Mu, Huamei Sun, Haobo Yue, Yuanyuan Jiang, Zehao Wang
Zhicheng Zhang, Jianqin Yin

School of Artificial Intelligence
Beijing University of Posts and Telecommunications, China
{da.mu, shm, hby, jyy, wzhao, zczhang, jqyin}@bupt.edu.cn

ABSTRACT

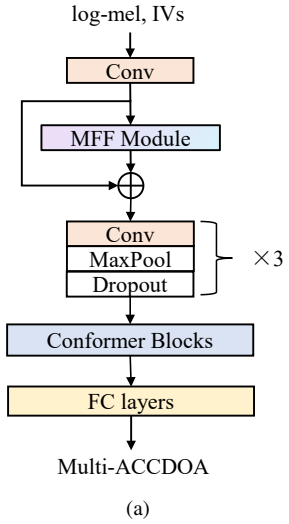
This technical report describes our submission system for task 3 of the DCASE2024 challenge: Sound Event Localization and Detection with Source Distance Estimation. Our experiment specifically focused on analyzing the first-order ambisonics (FOA) dataset. Building upon our previous work, we utilized a three-stage network structure known as the Multi-scale Feature Fusion (MFF) module. This module allowed us to efficiently extract multi-scale features across the spectral, spatial, and temporal domains. In this report, we introduce the implementation of the MFF module as the encoder and Conformer Blocks as the decoder within a single-branch neural network named MFF-Conformer. This configuration enables us to generate Multi-ACCDOA labels as the output. Compared to the baseline system, our approach exhibits significant improvements in F20° and DOAE metrics and demonstrates its effectiveness on the development dataset of DCASE task 3.

Index Terms— sound event localization and detection, multi-scale feature fusion

1. INTRODUCTION

Sound Event Localization and Detection (SELD) aims to use multichannel sound recordings to detect the onset and offset of sound events within specific target classes and estimate their direction of arrival (DoA). Its applications cover various fields, including smart homes, surveillance systems, and human-computer interaction.

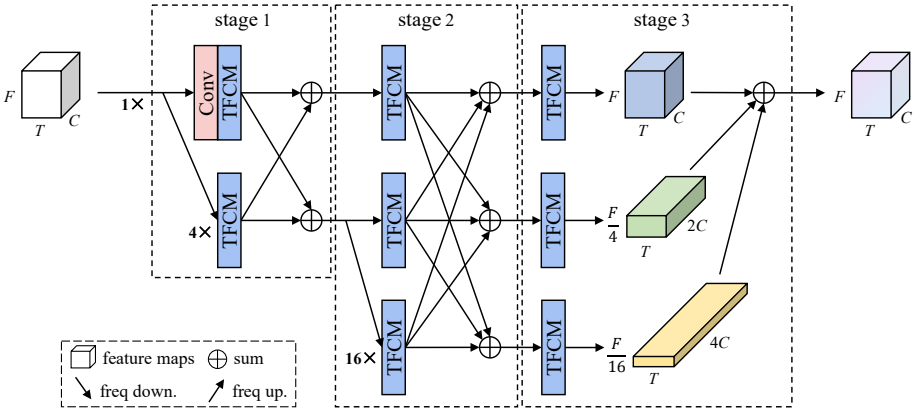
Since its introduction in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge as task 3, deep neural network (DNN) based methods have been extensively explored [1]. Initially, the convolutional recurrent neural network (CRNN) was commonly used for the SELD task. To improve the network's performance, researchers began incorporating ResNet and Conformer architectures [2, 3]. Subsequently, researchers focused on modeling the relationship between the spectral, spatial, and temporal domains. For instance, the CST-former [4] introduced distinct attention mechanisms to process channel, spectral, and temporal information independently. Recently, Zheng et al. proposed Spatial-AST [5], which was the first work to use the audio spectrogram transformer (AST) [6] for studying spatial audio. However, this year's challenge differs from previous ones in that it requires estimating the source distance, which significantly increases the task's difficulty.



(a)

Figure 1: MFF-Conformer network architecture.

In our previous work, we introduced the Multi-scale Feature Fusion (MFF) module[7], which effectively extracts multi-scale features across spectral, spatial, and temporal information. In this report, we present a novel approach called MFF-Conformer, which leverages the MFF module as encoders and integrates Conformer Blocks [8] as decoders. MFF-Conformer takes log-mel spectrogram and intensity vectors (IVs) as input features and generates Multi-ACCDOA labels. Within the MFF module, we employ a parallel subnetwork architecture combined with a TF-Convolution module (TFCM) [9]. This design enables efficient feature extraction by capitalizing on the strengths of each subnetwork. Furthermore, we employ repeated multi-scale fusion to enhance the representation of the subnetworks, thereby further improving the feature extraction performance. These methods help to increase the robustness and generalization capabilities of our system. As a result of these advancements, our system yields a significant improvement in F20° and DOAE metrics over the baseline system, demonstrating its effectiveness for the SELD task.



(a)

Figure 2: An illustration of the details of the MFF module. *C*, *T*, and *F* are the dimension sizes of channel, time, and frequency, respectively. "1×", "4×", and "16×" represent different scales of feature maps. "freq down." and "freq up." refer to frequency downsampling (FD) and frequency upsampling (FU), respectively.

2. PROPOSED METHOD

2.1. Input Features

In our approach, we specifically select audio files in the first-order ambisonics (FOA) format. From the FOA data, we extract two types of features: 4-channel log-mel spectrograms and 3-channel IVs. These two features are then combined to create a 7-channel feature, which serves as the input for our model.

2.2. Network Architecture

Our MFF-Conformer network architecture is depicted in Figure 1, and the details of the MFF module can be found in Figure 2. Initially, we input the feature maps with 7 channels into a convolution block (Conv), which preserves the dimensions of T (number of time frames) and F (number of frequencies) while transforming the channel dimension to 64. The resulting feature maps are then fed into the MFF module to comprehensively extract multi-scale spectral, spatial, and temporal information. To ensure the MFF module retains the original information and avoids extracting irrelevant features, we establish a residual connection between the output of the MFF module and the initial Conv. Subsequently, Conv, MaxPool, and Dropout operations are employed to further encode deeper information within the network. For the decoder, we utilize Conformer Blocks, which integrate convolution layers and multi-head self-attention (MHSA) mechanisms. This integration allows for the extraction of both local and global time context information from the feature sequence simultaneously. Finally, the fully connected (FC) layers generate Multi-ACCDOA labels as the output of our system.

3. EXPERIMENTS

3.1. Dataset

We conducted training and evaluation of our approach using the STARSS23 dataset[10]. This dataset consists of thirteen sound event classes and includes two types of multichannel array signals: FOA and tetrahedral microphone array signals. Our model was trained exclusively on FOA array signals, which consist of four channels: an omnidirectional channel (w) and three directional channels (x, y, and z). In addition to the original dataset, we enriched our training data by incorporating synthetic data[11], enhancing the diversity and robustness of our model.

3.2. Experimental setup and Evaluation Metrics

We follow the same audio feature extraction process as the baseline approach. The audio is sampled at a frequency of 24kHz, and we utilize 64 Mel filters for feature extraction. We apply the Short-Time Fourier Transform (STFT) with a frame length of 40ms and a frame hop of 20ms. The input length is set to 250 frames. For optimization, we employ the Adam optimizer. The batch size is set to 16, and the model is trained for 100 epochs. The learning rate is set to 0.0001. To evaluate the performance of our SELD system, we utilize the official metrics[12] recommended for the DCASE challenge. These metrics include the location-dependent F1-score (F_{20°), class-dependent direction of arrival error (DOAE), and class-dependent relative distance error (RDE).

Table 1: Experimental results of the audio-only SELD system for the development dataset using FOA format data.

Model	macro F_{20°(%)	DOAE(°)	RDE(%)
Baseline	13.1	36.9	**33**
Ours	**19.0**	**27.5**	39

3.3. Experiment Results

Table 1 illustrates the performance of our proposed method on the development dataset. Our model showcases impressive competence in macro F20° and DOAE metrics. Particularly, it displays a noteworthy enhancement of 5.9% in F_{20° and 9.4° in DOAE compared to the baseline. However, it falls short in terms of RDE, exhibiting a disadvantage of 6% compared to the baseline.

4. CONCLUSION

In this report, we present our approach for task 3 of DCASE2024. We propose MFF-Conformer, which incorporates a multi-scale feature fusion mechanism and Conformer Blocks into the SELD system. Through our experiments and evaluations, we demonstrate that our proposed system frameworks outperform the baseline approach. These results indicate the effectiveness and superiority of our approach in improving the performance of SELD.

5. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[2] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "A track-wise ensemble event independent network for polyphonic sound event localization and detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9196–9200.

[4] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," *arXiv preprint arXiv:2312.12821*, 2023.

[5] Z. Zheng, P. Peng, Z. Ma, X. Chen, E. Choi, and D. Harwath, "Bat: Learning to reason about spatial sounds with large language models," *arXiv preprint arXiv:2402.01591*, 2024.

[6] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021.

[7] D. Mu, Z. Zhang, and H. Yue, "Mff-einv2: Multi-scale feature fusion across spectral-spatial-temporal domains for sound event localization and detection," 2024.

[8] 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 *Proc. Interspeech*, 2020, pp. 5036–5040.

[9] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9122–9126.

[10] K. Shimada, A. Politis, P. Sudarsanam, D. 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," *arXiv preprint arXiv:2306.09126*, 2023.

[11] D. A. Krause and A. Politis, "[DCASE2024 Task 3] Synthetic SELD mixtures for baseline training," Apr. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10932241

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