POWER CUE ENHANCED NETWORK AND AUDIO-VISUAL FUISON FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2024 CHALLENGE

Technical Report

Xin Guan[1]*, Yi Zhou*[1]*, Hongqing Liu*[1]*, Yin Cao*[2]*,*

[1] Chongqing University of Posts and Telecommunications,
School of Communication and Information Engineering, Chongqing, China,
S220101035@stu.cqupt.edu.cn, {zhouy, hongqingliu}@cqupt.edu.cn
[2] Department of Intelligent Science, Xi'an Jiaotong Liverpool University, China,
yin.k.cao@gmail.com

ABSTRACT

This technical report describes our submission systems for Task 3 of the DCASE2024 challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. To address the audio-only SELD task, we utilize a Resnet-Conformer as the main network. Additionally, we introduce a branch to receive power cue features, specifically log root mean square (log-rms). We employ various data augmentation techniques, including audio channel swapping (ACS), random cutout, time-frequency masking, frequency shifting, and AugMix, to enhance the model's generalization. For the audio-visual SELD task, we also augment the visual modality in alignment with ACS. The audio and visual embeddings are sent to parallel Cross-Modal Attentive Fusion (CMAF) blocks before concatenation. We evaluate our approach on the dev-test set of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset.

***Index Terms*—** Sound event localization and detection, data augmentation, ensemble, audio-visual fusion

1. AUDIO-ONLY TRACK

This part will provide a detailed description about the four parts of the audio-only approach: input features, network architecture, data augmentation and post-processing.

1.1. input features

We first followed the setup of the baseline model and extracted features from First Order Ambisonics (FOA) audio, including 7 channels: 4 channels of logmel spectrograms and 3 channels of intensity vector (iv). Additionally, we extracted the root-mean-square (RMS) value for each frame and took its logarithm as an additional feature to improve sound source distance estimation, named as log-rms.

1.2. network architecture

We used Resnet-Conformer[1] as the main network for the SELD task, with the output format adopting multi-ACCDOA[2]. This format is used to predict sound event detection (SED), direction of arrival (DOA), and distance in the case of using a single-branch network, To better extract audio information, we introduced Attentional Feature Fusion (AFF) in the resblock. The AFF has also been proven effective in audio pattern recognition[3]. The Conformer structure is more sensitive to higher time resolution[1]. Therefore, we moved the temporal pooling in ResNet to after the Conformer, allowing the Conformer to extract more detailed information. For pooling, we used Attentive Statistics Pooling[4] on the time dimension instead of maxpool or avgpool, which is also applied in Automatic Speech Recognition (ASR). The structure of this model is shown in Fig. 1(a)

To get better performance in sound source distance estimation, we introduced an additional branch based on the main network. This branch accepts log-rms feature input. The high-level representations are concatenated before send to the subsequent conformer. Since the log-rms feature merges the frequency dimension, we used 1D operations (e.g., conv1d) in these side resblocks. The structure of additional side branch model is shown in Fig. 1(b)

1.3. data augmentation

To increase the diversity of our dataset, we convolved sound samples from FSD50K[5] and FMA[6] with the TAU-SRIR DB[7], generating more spatial audio samples. Using SpatialScaper[8], we have generated a total of 3000 one-minute audio clips, with a maximum overlap of three events per clip. Additionally, we used audio channel swap technique [9] (ACS) to enhance the accuracy of DoA localization. ACS achieved an 8-fold increase in sample quantity by swapping audio channels and adjusting their corresponding spatial positions. After extracting the spectrograms of the samples, we applied AugMix[10], which is widely used in the image processing domain, to the spectrograms. The spectrogram operations include cutout[11], time-frequency masking[12], and frequency shifting[13]. AugMix was sampled with a width k of 3 and a depth of 3.

1.4. post-processing

Firstly, we applied ACS to the test samples in the same way as with the training data, rotating the results back to their original orientation based on the swap order. We averaged the total of 8 samples generated from the same sample. Additionally, we used a 1-second hop length on an input length of 5 seconds. This overlap generated 5 results for each time-overlapping input, except for the start and end parts of the audio. By simultaneously applying ACS and overlap addition, we could generate 40 results from a single input. Averag-

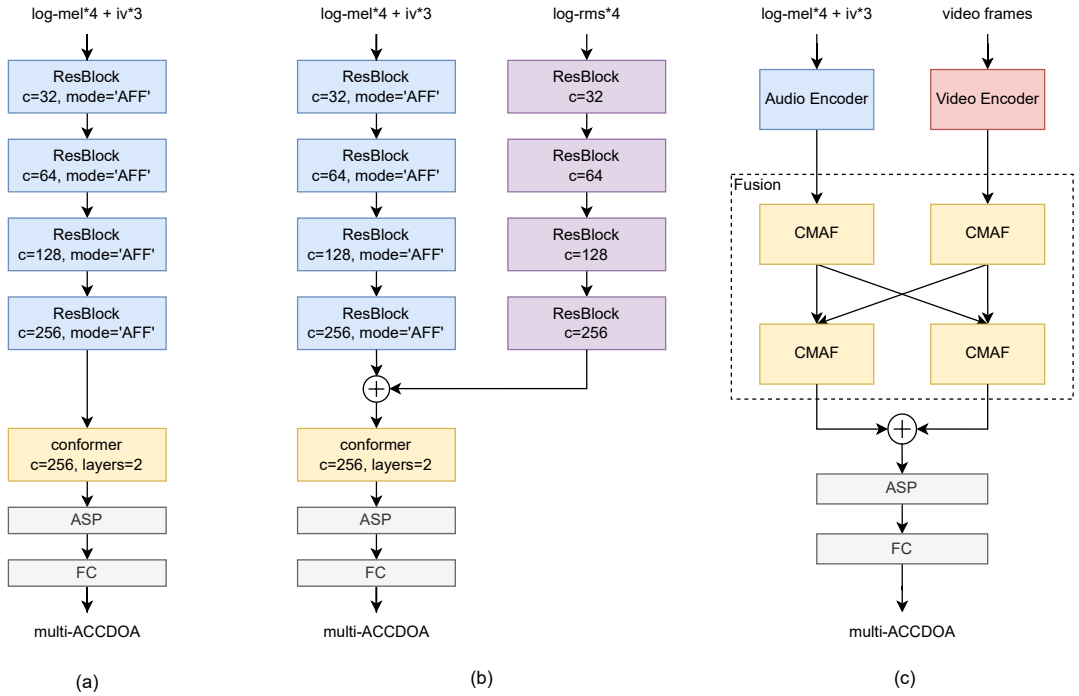

Figure 1: The overview of our models, (a) Resnet-Conformer network, (b) Resnet-Conformer network with side branch, (c) Audio-Visual Fusion Network

ing these 40 results effectively reduced the variability. Lastly, unlike the default SED threshold of 0.5, we used different SED thresholds for different classes.

2. AUDIO-VISUAL TRACK

This part will provide a detailed description about the four parts of the audio-visual approach: input features, network architecture, data augmentation and post-processing.

2.1. input features

In the audio-visual section, the audio features follow the audio-only track setup, extracting 7-channel features from FOA audio, including 4-channel logmel spectrograms and 3-channel iv. The corresponding video is segmented into frames at a frame rate of 10 fps. Each frame is processed using ResNet50[14] with default weights to extract embedding vectors. Unlike the baseline, the embedding vectors retain the channel dimension, and the final pooling layer is modified to perform avgpooling on both the width and height.

2.2. network architecture

The model simultaneously receives audio and corresponding video feature inputs. In the audio encoder part, we use the same structure as in the audio-only section, with 4 resblocks and 2 conformers. The overall structure of audio-visual model is shown in Fig. 1(c)

In the video encoder part, the features extracted by ResNet50 are dimensionally reduced through a linear layer, followed by 2 layers of conformers, resulting in video embedding vectors. The time dimension of the audio vectors differs from that of the video vectors. Consequently, the video vectors are replicated to match the time dimension of the audio vectors before entering the fusion layer.

The fusion layer consists of 2 layers of Cross-Modal Attentive Fusion[15] (CMAF) blocks. The fused audio and video vectors are then concatenated to form fused features, which are fed into a fully connected layer to output the multi-ACCDOA vectors.

2.3. data augmentation

ACS applies channel swapping to the audio, resulting in an 8-fold increase in augmented data. Correspondingly, Audio-Visual Channel Swap[16] (AVCS) rotates or flips video frames to achieve video data augmentation.

2.4. post-processing

For audio and video, we follow the post-processing setup used in the audio-only track. We first apply ACS and AVCS to generate more results by a factor of 8. Then, we apply overlap addition to get more results. Finally, we average these results.

3. EXPERIMENTS

3.1. Experimental settings

In our experiments, we only used FOA format audio. The sampling rate was set to 24 kHz, the STFT frame length was 40 ms, and the

Table 1: The SELD performance of our system for dev-test set.

model	modality	$F_{20^\circ/1}$	DOAE	RDE	SELD_{score}
model#a	audio	39.2%	15.5°	0.30	0.318
model#b	audio	41.3%	15.4°	0.29	0.316
ensemble#1	audio	43.2%	14.6°	0.29	0.313
ensemble#2	audio	43.7%	14.0°	0.30	0.301
ensemble#3	audio	44.1%	13.7°	0.30	0.298
ensemble#4[1]	audio	-	-	-	-
ensemble#5	av	44.4%	15.2°	0.27	0.305
ensemble#6	av	46.7%	14.2°	0.28	0.297

[1] Ensemble#4 used the entire development set, so specific metrics are not provided.

hop length was 20 ms. We used 64 mel filters. The input length was 5 seconds, or 250 frames. The model was trained with the Adam optimizer for 120 epochs, with a learning rate set to 0.001, gradually decaying to 0.0001 after 80 epochs.

We evaluated our SELD system using official metrics, including the location-dependent F1 score, direction-of-arrival error (DOAE), and relative distance error (RDE).

3.2. results

Tabel 1 shows the performance of our system on the development set. In the model ensemble part, we averaged the outputs from different networks and augmentation methods.

4. REFERENCES

[1] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng, Y. Wang, L. Sun, Y. Fang, J. Pan, J. Du, and C.-H. Lee, "The NERC-SLIP System For Sound Event Localization And Detection Of DCASE2022 Challenge," DCASE2022 Challenge, Tech. Rep., June 2022.

[2] 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)*, May 2022, pp. 316–320, iSSN: 2379-190X.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Interspeech 2018*, Sept. 2018, pp. 2252–2256.

[5] 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, 2022.

[6] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A Dataset for Music Analysis," Dec. 2016.

[7] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 165–169.

[8] 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)*. IEEE, 2024.

[9] 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.

[10] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," Feb. 2020, arXiv:1912.02781 [cs, stat].

[11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 001–13 008, Apr. 2020.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019*. ISCA, Sept. 2019, pp. 2613–2617.

[13] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022, publisher: Institute of Electrical and Electronics Engineers (IEEE).

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016, conference Name: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781467388511 Place: Las Vegas, NV, USA Publisher: IEEE.

[15] X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, "Audio-Visual Cross-Attention Network for Robotic Speaker Tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2023.

[16] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, April 2024.