Data Augmentation and Cross-Fusion for Audiovisual Sound Event Localization and Detection with Source Distance Estimation

Technical Report

Yongbo Li

Shanghai University
1398518787@shu.edu.cn

Chuan Wang

Shanghai University
wangchuan1101@shu.edu.cn

Qinghua Huang

Shanghai University
qinghua@shu.edu.cn

ABSTRACT

This technical report describes a system participating in the DCASE2024 challenge Task 3: Sound Event Localization and Detection with Source Distance Estimation-Track B: Audio-Visual Reasoning. A system based on the official baseline system is developed and improved in terms of network architecture and data augmentation. The convolutional recurrent neural network (CRNN) is substituted by a ResNet-Conformer block pre-trained on an audio-only network. Audio Channel Swapping (ACS) is applied to the DCASE 2024 official audio dataset to generate more audio data. A simulated audio dataset is also created. Video Pixel Swapping (VPS) is performed on the original video data to obtain more video data. Experimental results show that our system outperforms the baseline method on the Sony-TAU Real Spatial Soundscape 2024 (STARSS24) development dataset. A series of experiments are implemented only on the First-Order Ambisonics (FOA) dataset.

Index Terms— Data augmentation, resnet-conformer, sound event localization and detection, sound distance estimation

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is a combined task involving Sound Event Detection (SED) and Direction of Arrival (DOA) estimation. It identifies the categories of sound events and their corresponding pitch and azimuth angles in three-dimensional space over time. As an intelligent system, SELD has diverse applications in video surveillance, robotics, autonomous driving, scene visualization, and acoustic monitoring.

The SELD task was first introduced as Task 3 of the DCASE challenge in 2019 [1]. It was based on emulated multi-channel recordings, generated from event sample banks spatialized with spatial room impulse responses (SRIRs) captured in various rooms and mixed with spatial ambient noise recorded at the same locations. Initially, the DCASE 2019 included only stationary sound sources. To enhance the task, moving sound sources and unknown directional inferences were introduced in the subsequent 2 DCASE challenges [2] [3]. DCASE 2022 TASK 3 marked a significant departure from previous iterations, transitioning from computationally generated spatial recordings to recordings of real sound scenes that are manually annotated [4]. DCASE 2023, maintained all the recordings from STARSS22 while incorporating an additional 4 hours of material captured at Tampere University. This additional data is distributed between the training and evaluation sets. Furthermore, STARSS23 included simultaneous 360° video recordings for all audio recordings, offering a more comprehensive view of the sound scenes [5]. Additionally, the dataset augments the respective labels with source distance information, in addition to the direction of arrival. However, this approach does not take advantage of full spatial information by limiting it to the DOA only. In many cases, performing Sound Distance Estimation (SDE) would be also important to obtain the explicit position of the sound source in space.

This year the challenge task resembles the previous iteration, evaluating SELD models with audio-only input (Track A) or audiovisual input (Track B) on manually annotated recordings of real interior sound scenes. However, this year the task introduces distance estimation of the detected events [6], which makes the task significantly more challenging.

In this report we particularly focus on Track B, which consists of audio and video data for the Sound Event Localization and Detection with Source Distance Estimation system. Methods based on data augmentation and neural networks to improve accuracy are introduced. Experimental results show that our system outperforms the baseline method on the development dataset of Sony-TAU Realistic Spatial Soundscapes 2024 (STARSS24).

This report is organized as follows: Section 2 introduces our proposed method. Experiments and discussion are shown in Section 3. Finally, the conclusion is shown in Section 4.

2. METHODS

2.1. Feature

For audio features, the STARSS24 dataset provides two recording formats: the first-order ambisonics (FOA) and tetrahedral microphone array (MIC). The 4 channel 24kHz FOA recording format audio is used. For FOA, features need to be extracted from the audio. First, two time-frequency domain features are extracted through short-time Fourier transform (STFT): Log-Mel Spectrogram and Intensity Vector (IV). The Log-Mel Spectrogram and IV are then concatenated as input audio features. For video features, images are extracted at 10 fps. These images are then passed through a pre-trained ResNet-50 to obtain advanced video features.

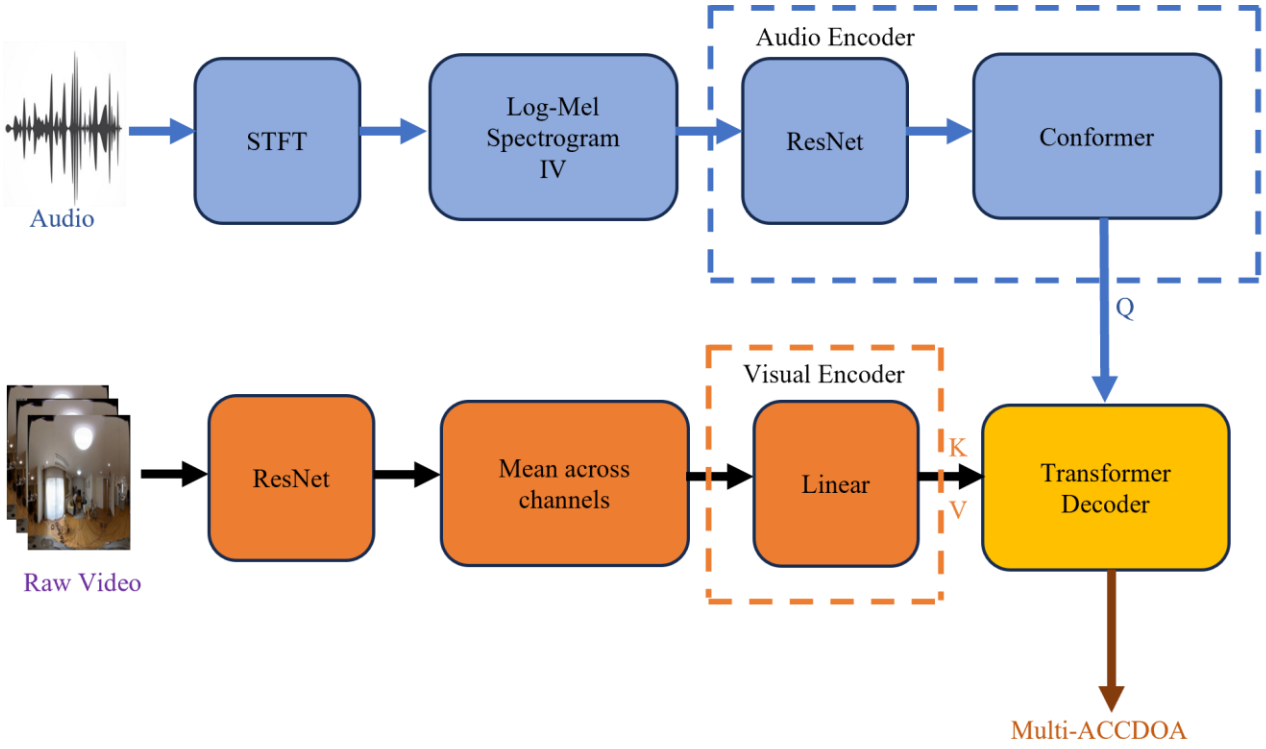

Figure 1: The architecture of the model

2.2. Network architecture

The official baseline architecture consists of an audio encoder, a video encoder, and a feature fusion module. The audio encoder is a CRNN, the video encoder is a linear layer, and the fusion module is a Transformer Decoder. We have improved the audio encoder based on the baseline network architecture, as shown in Figure 1.

2.3. Audio Encoder

The audio encoder input of the baseline system consists of 250 frames of 7-channel data. A CNN layer is used to downsample the frames to 50, matching the video and label information. Then, audio data is transmitted to the GRU. We improved the CRNN in the baseline with ResNet-Conformer blocks pre-trained on the audio-only network. We mix the augmented audio data with the official audio data for training. This helps us obtain a better audio-only model as the audio encoder. In the ResNet setting, we use four residual blocks. Each block contains a 1×1 convolution kernel and two 3×3 convolution kernels. In the Conformer setting, we utilize 2 Conformer blocks. Each Conformer block is configured with the following parameters: an input dimension of 256, 8 attention heads, a feed-forward dimension of 256, a depthwise convolution kernel size of 31, and a dropout rate of 0.1.

2.3.1. Visual Encoder

Advanced video features from ResNet-50 into the video encoder. The video encoder consists of a linear layer to obtain video features of the same shape as the audio features.

2.3.2. Feature Fusion

The audio features and video features are input together into the Transformer Decoder for feature fusion utilizing cross attention. The audio features are denoted as Q, while the video features are denoted as K and V.

2.4. Data augmentation

2.4.1. Audio Data Augmentation

The official audio development dataset has 8 hours of real data recorded in real sound environments, of which about 4.5 hours are used as training sets. Therefore, data augmentation techniques are crucial. Three methods are used to expand the audio data. The first method is ACS spatial augmentation, which is proposed by [7] in their previous work. This technique uses the rotation characteristics of the recorded dataset to improve the DOA representation. Additionally, an external dataset is adopted for training, specifically 750 recordings from FSD50K along with the 1200 recordings provided by the official dataset. The third method [8] simulates new multi-channel data using SRIR and sound samples extracted from FSD50K and AudioSet. Spe-

cifically, the single-channel sound samples in these external datasets are convolved with SRIR to create a 1-minute-long multi-channel scene recording with a maximum polyphony of 2. This method effectively generates additional synthetic audio data.

2.4.2. Video Data Augmentation

Similarly, 4.5 hours of video data is not enough to extract accurate features. We use VPS proposed by Wang et al [9] to consistently enhance the visual modality. Completely new video frames are generated by flipping and rotating the original video frames.

3. EXPERIMENTS AND DISCUSSION

3.1. Training Procedure and Evaluation Metrics

In order to compare the baseline and proposed network fairly, the hyperparameter settings have hardly changed, except for the modified learning rate of 1e-4. The evaluation metrics have changed this year, from the previous 4 metrics (location-dependent F1 score, location dependent error rate (LE), DOA localization error, and localization recall (LR)), to the use of location-dependent F1 score and DOA error (DOAE), and the addition of a new relative distance error (RDE). This year F1 score is different from previous years. The F1 score is spatially thresholded not only on the angular distance of predictions from the reference events, but also on the distances from the references. This year the task performs macro-averaging mode, which computes the metrics for each class and then averages them along the class.

3.2. Results and Discussion

Approximately 90 hours of audio data are used to train the audio-only model, which is then fine-tuned for the audio-visual model based on the audio pre-trained parameters Table 1 showed the experimental results of the proposed audio-visual method on the development dataset. As shown in Table 1, the performance of the proposed audio-visual system is significantly better than the baseline system.

Table 1: The performance comparison for different methods on development dataset.

	$F_{20°/1}$(%) ↑	DOAE (°)↓	RDE (%)↓
AO Baseline	13.1	36.9	33
AV Baseline	11.3	38.4	36
Our Proposed	**39.2**	**18.7**	**31**

4. CONCLUSION

This report proposes a system to solve the audio-visual SELD task in Task 3 of the DCASE 2024 challenge. We focus on data augmentation and pre-trained method. Data augmentation methods are used to extend both official and synthetic datasets. Pre-trained method is used to obtain more robust SELD estimates. Experimental results show that the proposed method achieves significant improvements over the baseline system.

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 J. Selected Topics in Sig. Proc.*, vol. 13, no. 1, pp. 34–48, 2018.

[2] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in Work. *on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 165–169, 2020.

[3] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in Work. *on Detection and Classification of Acoustic Scenes and Events (DCASE)*\, pp. 125–129, 2021.

[4] 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. *arXiv:2206.01948*, 2022.

[5] K. Shimada et al., "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances Neural Inf. Process. Syst., vol. 36*, 2024.

[6] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv:2403.11827*, 2024.

[7] 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.

[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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[9] Q. Wang et al., "The NERC-SLIP system for sound event localization and detection of DCASE2023 challenge," *Tech. Report of DCASE Challenge*, 2023.