AUDIO-VISUAL SOUND EVENT LOCALIZATION AND DETECTION BASED ON CRNN USING DEPTH-WISE SEPARABLE CONVOLUTION

Technical Report

Yi Wang[1], Hongqing Liu[1], Yi Zhou[1]

[1]School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
S210101136@stu.cqupt.edu.cn, {hongqingliu, zhouy}@cqupt.edu.cn

ABSTRACT

This technical report describes the systems submitted to the DCASE2023 challenge task 3: sound event localization and detection (SELD) -- track B: audio-visual inference. The goal of the sound event localization and detection task is to detect occurrences of sound events belonging to specific target classes, track their temporal activity, and estimate their directions-of-arrival or positions during it. Compared with the official baseline system, the improvements of our submitted system based on CRNN [1] mainly contain two parts: more powerful audio feature processing network architecture, additional visual feature module. For audio network, we utilize depth-wise separable convolution with multi-scale kernel size to better learn the relevant information of different sound event categories in audio features. Then, we modify the pooling stage and some residual operation is added to prevent information loss. Besides, we use the corresponding image at the start frame of the audio feature sequence processed by a pretrained ResNet-18 model as additional visual feature. Experimental results show that our system outperforms the baseline method on the development dataset of Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23).

Index Terms— Sound event localization and detection, audio-visual fusion, depth-wise separable convolution, ResNet-18

1. INTRODUCTION

Sound event localization and detection (SELD) aims at detecting types of sound and their corresponding temporal activities also spatial position. Polyphonic SELD refers to cases where there are multiple sound events overlapping in time. Due to its ability to characterize sound sources spatially-temporally, SELD can be used to automatically describe social and human activities and assist the hearing impaired to visualize sounds.

The SELD task first as task 3 of DCASE was in 2019 [2] which were based on emulated multichannel 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. In 2019, the SELD challenge of DCASE only includes stationary sound sources. To further improve the task, moving sound sources and unknown directional inferences are introduced in the following two 2 DCASE challenges respectively [3, 4]. This challenge task of 2022 [5] changes considerably compared to the previous iterations since it transitions from computationally generated spatial recordings to recordings of real sound scenes, manually annotated, called Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) [6]. Based on STARSS22, the dataset used in this year called STARSS23 maintains all the recordings of STARSS22, while it adds an additional 4hrs of material captured in Tampere University distributed between the training and evaluation sets. It further includes simultaneous 360° video recordings for all the audio recordings and it augments the respective labels with source distance information, apart from the direction-of-arrival.

By hearing and seeing, human brain is able to perceive surroundings and extract complementary information. For this, the SELD task of DCASE 2023 prepare an audiovisual track to stimulate further developments on SELD research. The video data has the potential to mitigate difficulties and ambiguities of the spatiotemporal characterization of the acoustic scene solely through audio data. So compared with the audio-only baseline takes only the audio input, the baseline method for the audio-visual SELD task takes both the audio and a visual input and the network architecture is based on CRNN with a Multi-ACCDOA [7] sequence output.

In this work, we improve the baseline CRNN network with a more effective audio encoder and utilize raw video frame embedding corresponding with the start frame of the audio feature sequence to enrich visual feature. The detail of our proposed method is described in section 2. Experimental results show that our method outperforms the DCASE 2023 challenge audio-visual baseline model on development dataset.

2. THE PROPOSED METHOD

In this part, we introduce our proposed approach for audio-visual SELD based on CRNN. Figure 1 shows the overall process of our framework which consists of feature extraction, audio encoder, video encoder, decoder.

2.1 Feature Extraction

For audio feature, we first utilize STFT transform the raw audio signal into complex spectrograms. Then, the amplitude of complex spectrograms and the frequency-normalized inter-

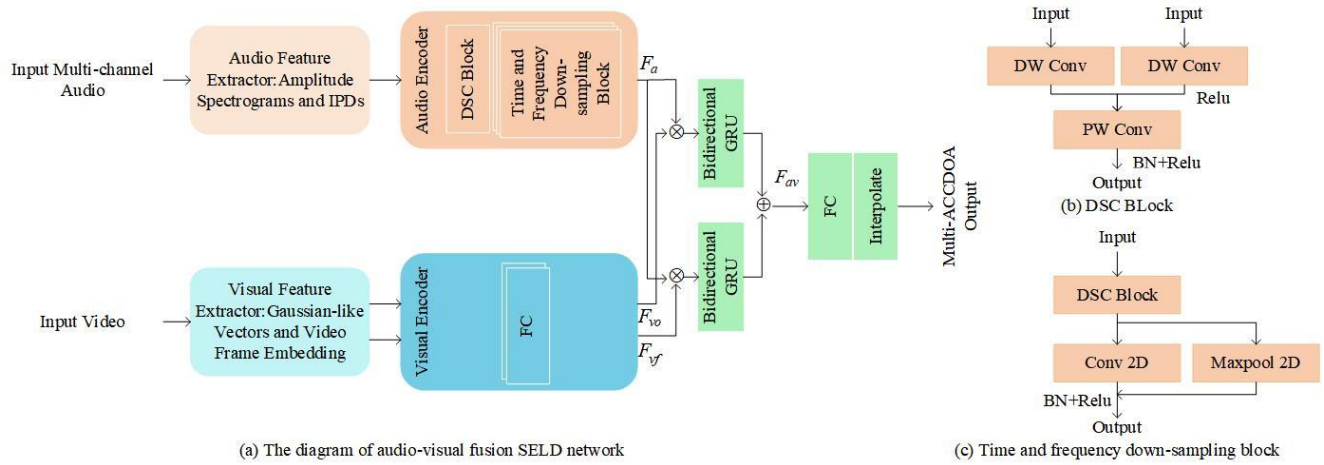

Figure 1: Overall view of proposed framework

channel phase differences (IPDs) are calculated respectively. Finally, we concatenate the amplitude of complex spectrograms and IPDs in channel-wise as the input audio feature.

For visual feature, we use YOLOX-Tiny [8], a light model of YOLOX pre-trained on COCO dataset, as object detector to detect person in the corresponding image at the start frame of the audio feature sequence. Then the bounding boxes of these people are transformed to a concatenation of two Gaussian-like vectors as visual feature, where they represent likelihoods of objects present along the image's horizontal axis and vertical axis [9]. Considering there are some sound event class in STARSS23 which are not relate to people, we pass raw video frame which is also the start frame of the audio feature sequence into a pre-trained ResNet18 [10] model from torch-vision to get a visual embedding as additional visual feature.

2.2 Network Architecture

2.2.1 Audio Encoder

The audio encoder architecture in our audio-visual fusion SELD network is showed in figure 1. Unlike form baseline, we sperate the channel up-sampling operation from first convolution part and modify the time and frequency block. To better catch the relevant information of different sound event categories in audio features and in view of model parameters, we introduce a depth-wise separable convolution [11] (DSC) block composed of two depth-wise 2D convolution with 3 and 5 kernel size respectively and a point-wise 2D convolution with 1 kernel size to perform channel fusion showed in Figure 1(b). First, a DSC block is used to up sample the input audio channel. Then, three time and frequency down-sampling blocks are used to reduce dimension size for feature fusion. For time and frequency down-sampling block showed in figure 1(c), a DSC block is first used to further encode the audio feature. Then we use both 2D convolution with different stride and kernel-size and max-pooling to down sample audio feature in time and frequency dimension. The feature after 2D convolution and max-pooling are added as a residual connection. For every layer, rectified linear unit activation and batch normalization are added to introduce nonlinear characteristics. Through audio encoder, we can get audio embedding F_a.

2.2.2 Visual Encoder

As the video object Gaussian-like vectors and raw video frame embedding are both pre-processed, we use two full connect layers to encode them into visual embedding F_{vo} , F_{vf} respectively. Then, we expand and repeat their time dimension to match audio embedding.

2.2.3 Decoder

As showed in figure 1, we straightly concatenate F_a with F_{vo} and F_{vf} respectively and pass them into two bidirectional GRU for temporal modeling. After that, the two part are added to get the fusion embedding F_{av}. Then, a full connected layer is used to map the F_{av} into output dimension. Due to time down-sampling used in audio encoder, an up-sampling operation called interpolate [12] in the temporal dimension is conducted to ensure the output size is consistent with label temporal dimension. At last, the output result is reshaped into Multi-ACCDOA format.

3. EXPERIMENTS

3.1 Dataset

The Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset contains multichannel recordings of sound scenes in various rooms and environments, together with temporal and spatial annotations of prominent events belonging to a set of target classes. The dataset provides two formats of audio data: 1) First-Order of Ambisonics; 2) tetrahedral microphone array. Besides, STARSS23 further includes simultaneous 360 ° video recordings which are spatially and temporally aligned with the microphone array recordings. During development stage, we train our proposed model on foa/video-dev-train-sony/tau of STARSS23, and evaluate those systems using foa/video-dev-test-sony/tau of STARSS23.

3.2 Hyper-parameters

The audio data sampling frequency of the dataset is 24 kHz. A 1.27 seconds audio clip is first random selected to do Short-time Fourier Transform with a hop size of 240 and 512 Fast Fourier Transform size for the amplitude of complex spectrograms and IPDs in training. And the audio clips in testing is are segmented to have a fix length of 1.2 seconds with no overlap. The video frames per second is 29.97. The first video frame of every audio clip is used for object detection and raw frame input. Adam optimizer is used with a 1e-6 weight decay. The learning rate is set to 0.001, reducing by 0.5 times every 10000 epochs. The max epochs for training the model is 40000.

3.3 Experimental results

For evaluation, we use official evaluation metrics to evaluate the SELD performance. The SELD score is computed as,

$$\text{SELD}=\frac{1}{4}(\text{ER}+(1-\text{F})+\frac{\text{LE}}{180}+(1-\text{LR})) \quad (1)$$

where ER, F, LE, LR are the official SELD metrics. Table1 shows the performance of our submit systems test on foa/video-dev-test-sony/tau of STARSS23. System #2 is the best checkpoint of our proposed method. For system #1, we average the result of two checkpoints with lowest SELD score based on system #2. Compared to system #2, we use pitch-shift additionally for system #4. And system #3 use the same way as system #1 but based on system #4.

Table 1. SELD performance of our systems.

System	ER_{20°	F_{20° (macro)	LE_{CD}	LR_{CD}
FOA_Baseline	1.07	14.3%	48.4%	35.5%
System #1	0.97	17.1%	44.1%	42.7%
System #2	0.97	15.9%	44.9%	41.7%
System #3	0.98	17.9%	41.9%	40.6%
System #4	1.00	17.4%	42.3%	42.0%

4. CONCLUSION

We have introduced our proposed audio-visual fusion method based on CRNN for sound event localization and detection. For audio encoder, we devise a depth-wise separable convolution with multi kernel size and modify the time and frequency down-sampling block used in baseline method. And we introduce raw video frame corresponds to the start frame of the audio feature sequence as additional visual feature. Experiment results show that our proposed method outperforms the baseline method.

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. Sel. Topics Signal Process., vol. 13, no. 1, pp. 34–48, March 2019.

[2] http://dcase.community/challenge2019/ task-sound-event-localization-and-detection.

[3] http://dcase.community/challenge2020/ task-sound-event-localization-and-detection.

[4] http://dcase.community/challenge2021/ task-sound-event-localization-and-detection.

[5] http://dcase.community/challenge2022/ task-sound-event-localization-and-detection.

[6] 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,” 2022. [Online]. Available: https://arxiv.org/abs/2206.01948

[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] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YOLOX: Exceeding YOLO series in 2021,” 2021, arXiv:2107.08430.

[9] X. Qian, Z. Wang, J. Wang, G. Guan and H. Li, "Audio-Visual Cross-Attention Network for Robotic Speaker Tracking," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 550-562, 2023, doi: 10.1109/TASLP.2022.3226330.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[11] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

[12] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” in Proc. Detect. Classification Acoust. Scenes Events Workshop, 2019.