ONE AUDIO AUGMENTATION CHAIN PROPOSED FOR SOUND EVENT LOCALIZATION AND DETECTION IN DCASE 2023 TASK3

Technical Report

Shichao Wu

Nankai University
College of Artificial Intelligence
Tianjin, 300350, China
wusc@mail.nankai.edu.cn

ABSTRACT

In this technical report, we propose to give the system details about our submitted results to the sound event localization and detection challenge in DCASE 2023. We concentrate on the audio-only based SELD sub-track, where inference of the SELD labels is performed with multichannel audio input only, as the previous years had done. We only used the audio data for training without any video information, since we think it's hard to fully explore the visual information collected with the 360° video setup. We present three improvements in this work concerning the neural network model, external data generation, and audio augmentation, compared to the baseline system. Specifically, we use one more deep and powerful neural network of the event-independent network (EINV2) in place of CRNN. Second, we propose to augment the audio data with one audio augmentation chain. Third, we synthesize more simulated audio samples for network training. Experiments on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) benchmark dataset showed our system remarkably outperformed the DCASE baseline system.

***Index Terms*—** Sound event localization and detection, sound event detection, direction-of-arrival estimation, audio augmentation chain, audio generation

1. INTRODUCTION

With machine-learning acoustic models performed on multiple channels of audio, SELD (sound event localization and detection) refers to classifying sound categories from a known set of classes and locating their location in the spatial spaces when they are active. Recently, several effective approaches to improve the SELD performance have been investigated, including data generation, audio feature extraction, audio augmentation, advanced and complex deep neural network designing, and post-processing techniques for the system outputs.

In this technical report, we propose to give the system details about our submitted results to the sound event localization and detection challenge in DCASE 2023 [1]. Specifically, we present three improvements in this work concerning the neural network model, external data generation, and audio augmentation, compared to the baseline system. First, we use one more deep and powerful neural network of the event-independent network (EINV2) in place of CRNN. Second, we propose to augment the audio data with one audio augmentation chain. Third, we synthesize more simulated audio samples for network training. Experiments on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) benchmark dataset showed our system remarkably outperformed the DCASE baseline system, concerning the audio-only based SELD sub-track with multichannel audio input only.

[1]https://dcase.community/challenge2023/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes

2. THE PROPOSED SYSTEM

We describe the proposed SELD system from input feature extraction, neural network architecture, audio augmentation chain, and simulated mixture generation.

2.1. Input features

In this report, we focus on the FOA audio representation-based SELD system and hence extract the Mel-spectrograms and intensity vectors as the benchmark systems have done. Firstly, the origin 4 channels of waveforms (W; Y; Z; X) are chunked into fixed-length segments of 5 seconds in a nonoverlap manner. Then, a 4-channel linear spectrum is computed from the segments using a 1024-point STFT with a Hanning window of 1024 points and a stride of 300 points. After that, 128 Mel-bins were used to extract the Mel-spectrograms from the linear spectrograms. The acoustic intensity vectors are computed based on linear spectrograms of pairs of channels W and X, Y, and Z, and then aggregated into the same number of Mel-bins. Finally, these intensity vectors are combined with the Mel-spectrograms and stacked along the channel dimension, resulting in a feature representation of 7×128 for each annotation frame.

2.2. Network architecture

We adopt the advanced EIN (Event-Independent Network V2), built by Hu et al. [1], as the acoustic model to classify the sound categories and locate their locations based on the input features. The EINV2 is extended by Hu et al. [2] in several aspects to improve its performance. First, it was extended to three tracks to address a maximum of three overlapping sound events. Second, the module block of multi-head self-attention (MHSA) blocks in EINV2 is replaced with Conformer, which consists of two feed-forward layers with residual connections sandwiching the MHSA and convolution modules. Third, convolutional blocks are replaced with dense blocks to increase the diversity of the model.

Table 1: Evaluation results of the baseline approaches and the proposed systems on the STARSS23 dataset (development-test split), training with officially and manually synthesized audio samples. 'w/o' means 'without'. $^{\diamond}$ represents the reported scores from the challenge description.

Method	Audio_Agu	Data_Syn	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$\mathcal{E}_{\text{SELD}} \downarrow$
CRNN$^{\diamond}$	w/o	official	0.57	29.90%	22.00°	47.70%	0.479
CRNN	w/o	official	0.57	29.80%	36.30°	45.90%	0.504
EINV2	w/o	official	0.64	31.30%	22.94°	56.20%	0.472
	w/o	user_syn	0.60	38.00%	23.00°	**60.40**%	0.436
	✓	user_syn	**0.54**	**40.40**%	**19.28**°	58.40%	**0.415**

2.3. Audio augmentation for SELD

We use the audio augmentation chain from our former work to augment the acoustic spectrogram and location diversities, since they are helpful to train one more generative model with the limited training samples. Our proposed audio augmentation chain consists of feature map augmentation (SpecAug), audio channel swapping (ACS), and sample mixup (Mixup). These audio augmentation blocks are exquisitely assembled by operating on a single feature map, between feature maps within an audio sample, and across audio samples, hierarchically. It takes the extracted audio features of mel-spectrograms and intensity vectors as input, sequentially augments the feature diversity, simulates new audio locations by swapping audio channels and rotating the corresponding Cartesian axes, and finally produces a mixture of audio feature maps by convex combinations of existing feature maps.

2.4. Simulated mixture generation

To further improve the SELD performance and fairly compare the various systems, we generate simulated audio samples for model training. For data synthesis, the main work has been to select different types of sound events and then to convolve them with SRIRs for specialization at varying SNRs. In this work, a bunch of sound events is chosen from three publicly available audio datasets, i.e., ESC-50[3], FSD50K[4], and AudioSet[5], based on the proximity of the labels in these datasets to the target classes of STARSS23. Around 440, 3,679, and 26,717 audio clips were extracted from ESC-50, FSD50K, and AudioSet, respectively. However, the roughly selected sound events are extremely imbalanced across the target classes. To alleviate the clipping imbalance, we perform resampling to extract the final sound events for data synthesis. Following the baseline system, we set the maximum polyphony of the target class to be three.

3. EXPERIMENTS AND RESULTS

In this work, we concentrate on the audio-only based SELD subtrack, where inference of the SELD labels is performed with multichannel audio input only. We only used the audio data for training without any video information, since we think it's hard to fully explore the visual information, collected with the 360° video setup, due to the structural and semantic gap between panoramic and normal images.

3.1. Dataset description and training settings

The STARSS23 (an extension dataset based on that of STARSS22) benchmark 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 is collected in two different sites, in Tampere, and Tokyo, using a similar setup and annotation procedure.

We train on the development-training split of the STARSS23 benchmark dataset and the external synthesized simulation data, as the baseline work has been done. Then, evaluate the network with the development-testing split of the STARSS23 dataset, and report the corresponding detection and localization scores.

We performed all experiments involved in this report on a single NVIDIA GeForce RTX3090 GPU. The AdamW optimizer with an initial learning rate of 0.0003, paired with the StepLR learning rate scheduler that attenuated 0.1 every 40 epochs, was used for model training. Other training details, remained the same as the baseline work.

3.2. Experiment results

We report the evaluation metrics on the development-testing split of the STARSS23 dataset, including the localization-dependent classification error ER_{20° and F-Score F_{20° with a threshold of 20 degrees, and the classification-dependent localization error LE_{CD} and recall LR_{CD}.

Table 1 summarizes the testing scores of the baseline system and our proposed systems. We re-trained the publicly available baseline system and found that the re-implementation system was inferior to the reported one, with aggregated $\mathcal{E}_{\text{SELD}}$ score of 0.504 to 0.479 (reported).

Concerning our proposed system, when EINV2 was adopted for training with the official data setting without any audio augmentation applied, the $\mathcal{E}_{\text{SELD}}$ is 0.472. When trained with more audio mixtures, as shown in the second line from the bottom, the SELD performance was improved on the sound event detection metric of F_{20° (38.00%), and sound event localization metric of LR_{CD} (60.40%).

Our proposed system, as shown in the last line, trained with more audio samples and adopted audio augmentation, achieved the best $\mathcal{E}_{\text{SELD}}$ score of 0.415. It obtains the best scores on every evaluation indicator, except for LR_{CD}.

4. CONCLUSION

In this technical report, we describe the submitted system details for the SELD task in DCASE 2023. Experimental results show that designing more advanced acoustic models, training with more audio samples, and augmenting audio samples can all improve SELD performance.

5. REFERENCES

[1] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 885–889.

[2] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9196–9200.

[3] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1015–1018. [Online]. Available: https://doi.org/10.1145/2733373.2806390

[4] 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, 2021.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.