THE DISTILLATION SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2023 CHALLENGE

Technical Report

Sang-Ick Kang , Kyongil Cho, Myungchul Keum, Yeonseok Park

KT Corporation, South Korea
{sangick.kang, cho.kyongil, mc.keum, yeonseok.park}@kt.com

ABSTRACT

This report describes our systems submitted to the DCASE2023 challenge task 3: Sound Event Localization and Detection (SELD) with audio-only data and audio-visual data. Audio-visual data consists of multi-channel audio data for sound events and 360-degree video data. To solve the issue of sparsity in the training data, we conducted various augmentations on both audio and video. The proven ResNet-Conformer based architecture in the Sound Event Localization and Detection system is employed, including the augmented data. To effectively improve the performance of the audio network, we applied the Knowledge Distillation technique by training both a teacher model and a student model. In addition, we fused the SELD model and the object detection model YOLOv7 in the audio-visual network. Finally, post-processing strategies involve an ensemble method for both audio-only track and audiovisual track. The experimental results demonstrate that the deep learning-based models trained on the STARSS23 dataset significantly outperform the DCASE challenge baseline in the proposed system.

Index Terms— DCASE2023, Sound Event Localization and Detection, Audiovisual Model, Ensemble Method

1. INTRODUCTION

Sound Event Localization and Detection (SELD) refers to the detection of sound events belonging to specific target classes, tracking their temporal activity, and estimating their directions-of-arrival (DOA) or positions. The challenge for this year involves the detection of sound events by incorporating 360-degree video data and utilizing it as visual information in addition to sound scenes. Autonomous robots equipped with cameras and multi-channel microphones benefit greatly from the SELD system, which processes audio and visual information simultaneously. For example, home service robots can easily distinguish between footsteps, knocks and door sounds by analyzing audio and video data. Also real human voices can be differentiated from sounds recorded by electronic devices.

In DCASE2023 task3, audio-only track and audio-visual track are released, and we are propose robust methods for each tracks. In audio-only track, the SELD system based on conformer model demonstrated outstanding performance [1]. However, the conformer model has the drawbacks of requiring a long training time and a large amount of data. To address these limitations, we propose a distillation technique using resnet-gru and resnet-conformer. For the audio-visual SELD method, an audio-visual conformer is proposed. After going through a method of fusion audio and visual encoders, finally, the audio-visual encoder is used to determine the acoustic event and location.

Audio data augmentation techniques such as data generation, audio channel swapping (ACS) [2], SpecAugment [3] and cutout [4] are used and video data augmentation techniques such as mosaic synthesis and ACS are used.

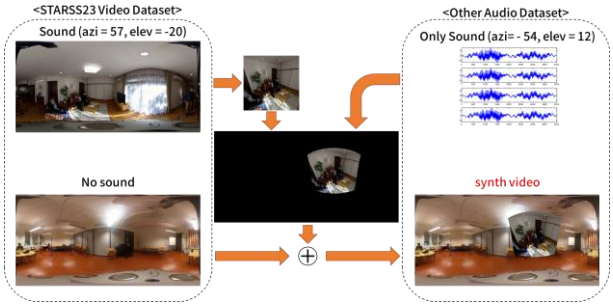

Figure 1. Video mosaic augmentation

2. PROPOSED METHOD

The official STARSS23 Development dataset [5] has been recorded actually audio and video data and provided annotated by real sound scenes. To design more robust deep learning-based models, larger datasets are required. Therefore, we propose audio and video data augmentation approaches for training the SELD system.

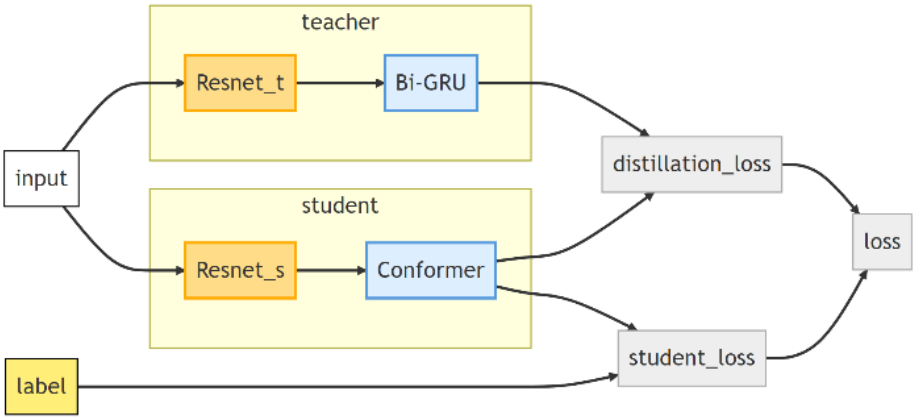

Figure 2. Distillation technique

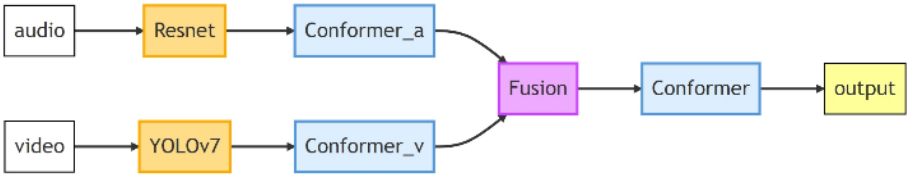

Figure 3. The network architecture of audiovisual fusion model

For Audio-only SELD, we propose a distillation-guided training strategy where we guide different models individually to converge to a much better optimum and show that it has the potential to improve existing works as well. For Audio-Visual SELD, we propose an audio-visual fusion model for audio-visual SELD that combines the resnet-conformer model for audio data and the yolo-v7-based conformer model for visual data.

2.1 Audio Data Augmentation

In research, we adopt audio data augmentation methods including audio data generation method, ACS, Spec Aug and cutout.

The audio data generation method generates synthesized sound data with single-channel sound samples from FSD50K [6], AudioSet strong [7], ESC-50 [8] datasets and SRIRs from TAU-NIGENS Spatial Sound Events Datasets [9, 10]. The generated dataset consists of several hundred hours of audio recordings capturing multichannel sound scenes, where each recording has a duration of 1 minute and contains up to 3 simultaneous sound sources (i.e., maximum polyphony of 3). To obtain a high-quality dataset, clean single-channel sound samples are selected high-score using the first SELD model. Finally, about 180 hours of data are generated as the final training set.

After applying ACS, the provided STARSS23 training data was augmented by a factor of 8. Additionally, some of the generated dataset was also subjected to ACS to ensure that enough data was available for training. Furthermore, Spec Aug and cutout techniques were employed to increase the complexity of the audio data.

2.2 Video Data Augmentation

The proposed video mosaic augmentation method crops and covers an image corresponding to the location of the sound event in the audio dataset, as described in Figure 1. The cropped images are extracted for each sound class from the training video dataset STARSS23. The extraction method converts a 360-degree image in the equivalent format into a cubemap to crop only the region corresponding to the sound event. This cube image is converted to equirectangular format at the location of the sound event in another audio dataset. Next step, the converted image is combined with the no sound event background frame in STASS23. Finally, new video data is synthesized from an audio data set where video did not exist. The new video data has 30 frames per second and is designed to fit the information in the audio set metadata. In addition, horizontal flip was applied to the cube image for the diversity of the data set, and in the case of overlapping sound events, the cube image was reduced to allow as many sound objects as possible.

ACS for video is based on the ACS applied to audio data augmentation, which involves moving and rotating the video without compromising the 360-degree view, similar to the audio data, thereby increasing the training data by a factor of 8.

2.3 Audio-only Network Architecture

We used the Resnet-Conformer student model and a Resnet-GRU-based teacher model to efficiently train the Conformer layer. The figure 2 illustrates the overall structure of distillation technique. The two outputs from the student and teacher models are passed

through distillation loss and student loss, and the weighted sum of the two losses is used as the final loss.

2.4 Audiovisual Network Architecture

we describe our proposed Audio-Visual Conformer network. The model is composed of 4 main components: An audio encoder, a visual encoder, an audio-visual fusion module, and an audio-visual encoder. The audio encoder is pre-trained by the audio-only model and the visual-encoder is pre-trained by yolo v7 [11] image classification model. Efficient Conformer back-end networks are used to model local and global temporal relationships. The overall architecture of the model is shown in Figure 3.

2.5 Model Ensemble

In track-wise output format, predictions of sound events can be arbitrarily assigned to tracks. Averaging or weighted ensembles are unable to make predictions across different tracks, making them unsuitable for a track-wise output format. To address this problem, a track-wise ensemble model has been proposed [12].

The ensemble model architecture, as shown in Fig. 2, takes multiple outputs from distinct models as input. It has CRNN multi-task network decoder but incorporates multiple SELD outputs. The model handles SED and DOA separately. Soft parameter-sharing is utilized to enhance the per-track results. The ensemble model predicts the outcome in the form of a track-wise output.

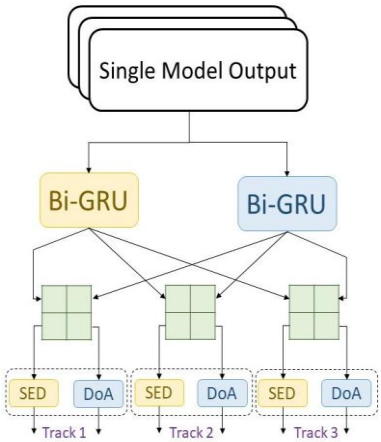

Figure 4. Model Ensemble

3. EXPERIMENTAL RESULTS

The audio data is sampled at a frequency of 24 kHz. For the Short-Time Fourier Transform (STFT), we use a frame length of 20 ms and a frame hop of 10 ms. The log-mel spectrograms and IV features are computed with 64 mel bins. The input to the neural networks consists of frames with a length of 1,000. During training, we utilize a batch size of 64, and each training sample is dynamically generated on-the-fly. The learning rate is gradually increased from 1.0^{-8} to 0.0001. We employ the AdamW optimizer. The video input has been resampled to 640×320px 10fps and equirectangular format.

Table 1 presents the performance of the audio-only SELD network and the ensemble models derived from them. Table 2 displays the performance of the audio-visual SELD network, where the ensemble experiment results are achieved by combining the audiovisual fusion model and the audio-only model.

Table 1. The experiment results of audio-only SELD track

Methods	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}
Baseline (FOA)	0.57	29.9%	22°	47.7%
Baseline (MIC)	0.62	27.8%	27°	44.3%
single model	0.42	55.35%	15°	64.8%
Ensemble #1	0.43	56.9%	15.3°	70.9%
Ensemble #2	0.43	55.8%	15.9°	71.5%
Ensemble #3	0.42	57.5%	15.8°	72.7%
Ensemble #4	0.43	56.4%	15.8°	70.4%

Table 2. The experiment results of audio-visual SELD track

Methods	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}
Baseline (FOA)	1.07	14.3%	48°	35.5%
Baseline (MIC)	1.08	20.6%	62°	29.2%
single model	0.45	49.4%	16°	63.7%
Ensemble #1	0.43	54.5%	15.6°	65.8%
Ensemble #2	0.44	54.1%	15.6°	66.5%

4. CONCLUSION

Our approach solves the sound event localization and detection (SELD) with directional interference in the DCASE2023 challenge. Specifically, in this challenge, the inclusion of 360-degree video requires us to estimate the positions of sound events using both auditory and visual information. This report suggests the Knowledge Distillation technique based on the Resnet-Conformer model for the audio-only track, and a fusion model combining SELD networks and image detection networks for the audiovisual track. In order to enhance the performance of deep learning models, we employed augmentation techniques for both audio and video data. Furthermore, to ensure an abundant amount of data, we generated additional samples. Specifically, A substantial video database was necessary to train the audiovisual model effectively. Experimental results demonstrate that the proposed networks outperform the baseline model for the DCASE2023 challenge task 3.

5. REFERENCES

[1] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng and C. H. Lee, "The NERC-SLIP system for sound event localization and detection of DCASE2022 challenge," in Tech. report of DCASE Challenge, 2022.

[2] 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,” arXiv:2101.02919, 2021.

[3] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Proc. Interspeech, 2019, pp. 2613 – 2617.

[4] T.DeVries, G. W. Taylor, "Improved regularization of convolutional neural networks with cutout,". arXiv preprint arXiv:1708.04552.

[5] A. Politis, K. Shimada, P. Sudarsanam, A. Hakala, S. Takahashi, D. A. Krause, N. Takahashi, S. Adavannem, Y. Koyama, K. Uchida, Y. Mitsufuji, and T. Virtanen, “STARSS23: Sony-TAu Realistic Spatial Soundscapes 2023,” arXiv:2206.01948, 2022.

[6] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “FSD50k: an open dataset of human-labeled sound events,” arXiv:2010.00475, 2020.

[7] J. F. Gemmeke, D. PW 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 Proc. IEEE ICASSP, 2017, pp. 776–780.

[8] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in Proc. 23rd Annual ACM Conference on Multimedia. ACM Press, pp. 1015–1018.

[9] A. Politis, S. Adavanne, and T. Virtanen, “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” in Proc. DCASE2020 Workshop, 2020, pp. 165–169.

[10] 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 Proc. DCASE2021 Workshop, 2021, pp. 125–129.

[11] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". arXiv preprint arXiv:2207.02696.

[12] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, J. Yang, "A Track-Wise Ensemble Event Independent Network for Polyphonic Sound Event Localization and Detection," in Proc. IEEE ICASSP, 2022, pp. 9196-9200.