SOUND EVENT LOCALIZATION AND DETECTION MODEL WITH ATTENTION-BASED NEURAL NETWORKS AND DATA MODELING

Technical Report

Gwantae Kim

Samsung Electronics
Suwon, South Korea

ABSTRACT

The technical report presents our submission system for task 3 of the DCASE 2025 Challenge, which tackles sound event localization and detection(SELD) problems in regular video content and stereo audio contents. In this report, we introduce data preparation and augmentation method, neural network structure, post-processing, and model ensemble strategy.

***Index Terms*—** DCASE2025, sound event localization and detection, attention, multimodal vision transformer, ensemble

1. INTRODUCTION

Human can detect and localize sound sources using stereo audio systems and vision systems. Over the past few years, the Task 3 of Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [1] has defined a more challenging and human-like problem by utilizing stereo audio instead of 4-channel audio in the sound event localization and detection (SELD) task.

SELD is the task that identifying direction of arrival (DOA), distance, and event class from mixture of sounds. Several methods [2–4] have been proposed to solve the problems of the SELD.

In this report, we first introduce data preparation and augmentation methods that improve generalizing performance of the model. We also propose neural networks structure, which is based on vision transformer (ViT), post-processing, and model ensemble strategies for SELD.

2. METHOD

2.1. Data Preparation and Augmentation

The official dataset contains 30000 training samples and 13000 evaluation samples, and each sample has 5-seconds long stereo recordings and labels. The dataset is sampled from Sony-TAu Realistic Spatial Sound-scapes 2023 (STARSS23) [5, 6]. Since the class of the training samples are unbalanced, we generates more samples using SpatialScaper [7] and stereo SELD data generator [8], which are shared in Challenge homepage. We used FSD50K [9] dataset to generate additional samples. We initially synthesize 120000 samples, split into the 4 subsets, and finally select one subset(30000 samples) for training submitted model. Moreover, we use Specmix [10] data augmentation to improve generalization performance. The method reinforces that the attention focus on the input features on the same time position. For the audiovisual task, we trained the model with official audiovisual training set, and 30000 generated audio-only training set.

2.2. Model Structure

The proposed SELD model is based on vision transformer structure, which consists of audio encoder, video encoder, positional embedding, transformer body, and fully-connected heads. The first 50 tokens are output tokens, which are random initialized and trainable. The i-th token denotes i-th output frame, and estimates labels with fully-connected heads. The next N tokens are audio feature tokens, and the next M tokens are video feature tokens, which are embedded by audio encoder and video encoder, respectively. For the audio-only task, the video inputs are zero-masked. The audio encoder, video encoder, and fully-connected heads consist of fully-connected layers. For the transformer body, we used vanilla 12/16 ViT layers [11]. Moreover, 4 ViT layers for each subtasks(DOA, class, distance, onscreen).

2.3. Training

We organized two ground-truth label styles. The first style is based on Multi-ACCDOA [12], but we used cross-entropy loss and mean-sqaured error for each class, instead of the multi-ACCDOA-adpit loss. The second style is track-wise representation. For each frame, up to 3 tracks are parsed into labels. If the number of tracks is less than 3, the labels are filled by -100, which is not back-propagated during training.

Table 1: Comparison of baselines and submissions with the development set (Track A).

	F1	Angle	Dist.	Screen
Baseline	22.8	24.5	0.41	-
Sub2	28.82	18.15	0.34	-

Table 2: Comparison of baselines and submissions with the development set (Track B).

	F1	Angle	Dist.	Screen
Baseline	26.8	23.8	0.40	0.80
Sub1	25.12	16.38	0.36	0.73
Sub2	26.14	19.60	0.30	0.74

We trained two models. The first model adopts the first label style with 50 output tokens, and the second model uses the both styles with 150 output tokens (50 frames, 3 tracks).

2.4. Post-processing and Ensemble

We saved 5 latest models during training, then we finally got 10 models. The candidates from the 10 models are merged by the same label, distance threshold and DOA threshold. Moreover, we applied moving average to distance along frames with window size 10.

3. EXPERIMENTS

3.1. Experimental setup

We use the AdamW optimizer with a learning rate from 1e-4 to 1e-5. The batch size is 4 or 8.

3.2. Results

The experiment results are summarized in Table 1. As shown in the table, each proposed model and ensemble model outperforms the baseline systems.

4. CONCLUSION

This report proposes a SELD model submission of the DCASE 2025 challenge task 3. We changed several factors, such as data, model structure, and post-processing, to improve the SELD performance. We generated synthetic dataset and applied Specmix data augmentation to improve generalization performance. We explored neural networks model structure, and found the suitable forms for SELD. Finally, we implemented post-processing and ensemble method for SELD. The experimental results show that the proposed method outperforms the baseline systems. In the future, we will perform ablation studies in several factors to find the reason of improvements.

Table 3: Style 1 Model Ensemble (Track B).

	F1	Angle	Dist.	Screen
Sub1-1	24.47	15.65	0.40	0.68
Sub1-2	24.86	17.03	0.41	0.77
Sub1-3	24.52	17.60	0.41	0.76
Sub1-4	25.45	18.03	0.41	0.78
Sub1-5	24.64	16.59	0.40	0.79
Sub1	25.12	16.38	0.38	0.73

5. REFERENCES

[1] D. Diaz-Guerra, A. Politis, P. Sudarsanam, K. Shimada, D. A. Krause, K. Uchida, Y. Koyama, N. Takahashi, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in dcase2024 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, Tokyo, Japan, October 2024, pp. 41–45.

[2] G. Kim and H. Ko, "Convnext and conformer for sound event localization and detection."

[3] ——, "Data augmentation, neural networks, and ensemble methods for sound event localization and detection," *Tech. Report of DCASE Challenge*, 2023.

[4] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

[5] 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," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings

[6] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Advances in*

Neural Information Processing Systems, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c9671ed3b3106b71cafda3ba225c1a-Abstract-Datasets_and_Benchmarks.html

[7] 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)*, Seoul, South Korea, April 2024.

[8] J. Wilkins, M. Fuentes, L. Bondi, S. Ghaffarzadegan, A. Abavisani, and J. P. Bello, "Two vs. four-channel sound event localization and detection," *arXiv preprint arXiv:2309.13343*, 2023.

[9] 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.

[10] G. Kim, D. K. Han, and H. Ko, "Specmix: A mixed sample data augmentation method for training withtime-frequency domain features," *arXiv preprint arXiv:2108.03020*, 2021.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] 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.