ATTENTION MECHANISM NETWORK AND DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Lihua Xue[1], *Hongqing Liu*[1], *Yi Zhou*[1]

[1] School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
S210131279@stu.cqupt.edu.cn, {hongqingliu, zhouy}@cqupt.edu.cn

ABSTRACT

This technical report describes our submission systems for the task 3 of the DCASE2023 challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. In our approach, we firstly generate more spatial audio files for training. To improve the generalization of the model, we employ random cutout, time-frequency masking, frequency shifting and augmix. Secondly, we utilize Resnet-Conformer network as the main body of our model, while we merge the Resnet-Conformer network and EINV2 framework with multi-ACCDOA output. To extract more effective features, we introduce a multi-scale channel attention mechanism and attentive statistics pooling. At last, we adopt model ensemble of different models with the same output format and post-processing strategies. The experimental results show that our proposed systems outperform the baseline system on the development dataset of DCASE2023 task3.

***Index Terms*—**Sound event localization and detection, Data augmentation, Attention mechanism, Ensemble

1. INTRODUCTION

The purpose of Sound Event Localization and Detection (SELD) task is to detect the activity time of a category of interest sound event (SED) while estimating its corresponding direction of arrival (DOA). The SELD system has great potential in many areas [1], such as machine listening, smart homes, and wildlife detection.

With the annual DCASE challenge held, the SELD task has gained the attention of a wide range of researchers. The SELD task has also made significant progress. Currently, the solutions to the SELD problem can be broadly classified into two categories. The first solution is the model architecture that input and output are single. In the DCASE2020 challenge, activity-coupled Cartesian DOA (ACCDOA) representation is proposed in [2],which assigns a sound event activity to the length of a corresponding Cartesian DOA vector. Thereby, the SED task and the DOA task are combined into a regression task in Cartesian coordinates. However, ACCDOA representation cannot solve the problem of similar events occurring at the same time. To address this problem, in [3], ACCCDOA representation is extended to multi-ACCDOA, employing auxiliary duplicating permutation invariant training (ADPIT) [4]. In both DCASE22 and DCASE23 challenges, the baseline system uses multi-ACCDOA output. The second solution is a two-branch structure model. In 2019, a two-step strategy is proposed in [5], where the SED model is trained first before the DOA model, using the output of the SED as a mask to guide the selection of the DOA output. Meanwhile, An event independent network V2 (EINV2) is proposed in [6], introducing soft parameter-sharing and multi-head self-attention (MHSA) to decode the output of SELD. Both schemes achieve excellent performance in the SELD task.

In this report, we propose a multi-scale channel attention mechanism network, attentive statistics pooling [7] and data augmentation for SELD. Our system is based on the ENIV2 framework and multi-ACCODA, which are able to efficiently detect overlapping problems of same or different classes of events. The multi-scale channel attention mechanism [8] is used to efficiently model the correlation information between channels. Meanwhile, attentive statistics pooling is used to improve the identifiability of important features. Data augmentation includes random cutout [9], time-frequency masking [10], frequency shifting [11] and augmix [12]. This challenge agrees with the use of external datasets. We utilize Audioset [13], FSDK50 [14] and TAU Spatial Room Impulse Responses Database (TAU-SRIR DB) [15] to generate more synthetic data to enhance the development dataset. Due to the imbalance between real and simulated data, we always keep half of the training data from the real dataset and half from the simulated dataset during the training phase. Experiments on the development dataset show that our systems have a significant improvement over the baseline system.

2. PROPOSED METHOD

2.1. Input Features

In our approach, we choose to use audio files in FOA format. We extract two kinds of features from FOA including 4-channel log-mel spectrograms and 3-channel sound intensity vectors. Finally, the two features are combined to form a 7-channel feature as the input feature for the model.

2.2. Network Architecture

The overall structure of our model is shown in Fig. 1(a), which contains one stem block, four MS-CAM Resblocks, conformers [16] and FCs.

Compared to the baseline model, firstly, we employ Resnet instead of CNN. Secondly, we introduce the multi-scale channel attention mechanism, which aggregates local and global feature contexts in the channel dimension. Then, we employ conformers

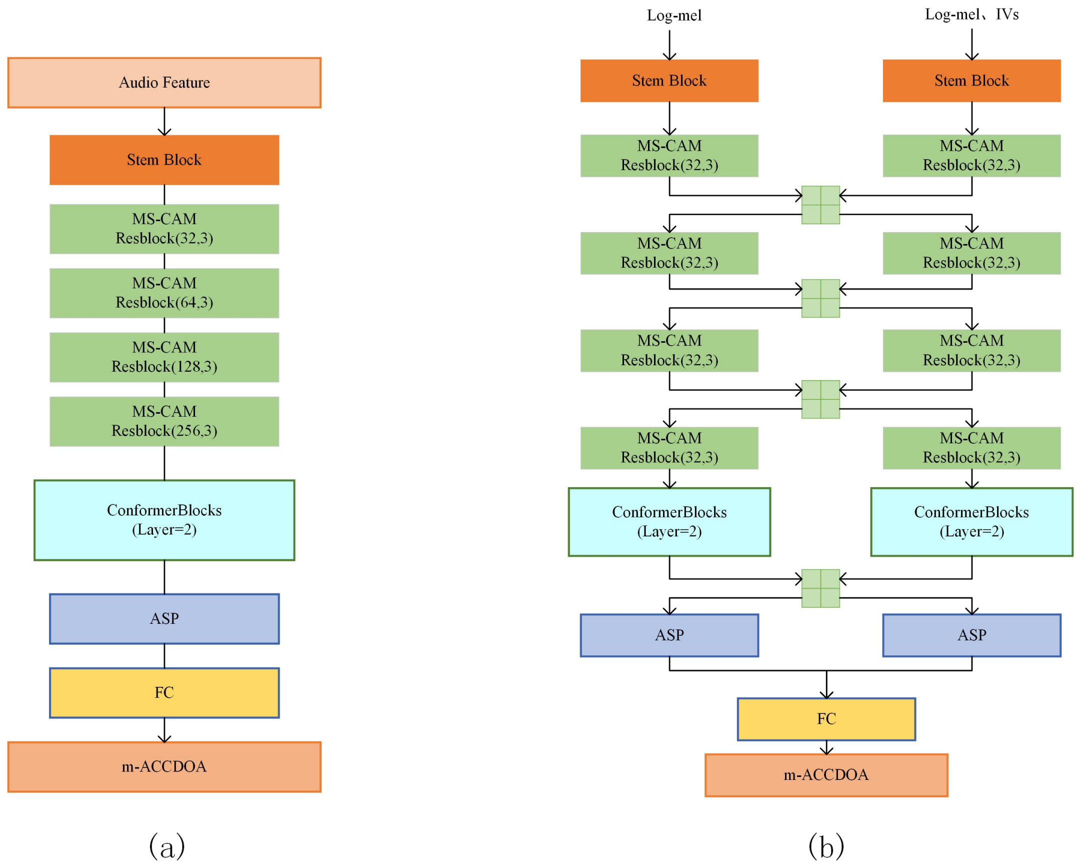

Figure 1: The overview of two models, (a) Resnet-Conformer network with MS-CAM (Model 1), (b) EINV2-based mutli-ACCDOA network (Model 2).

to replace GRU and MHSA in the baseline in order to extract local features and global features in the time dimension more efficiently. In addition, we employ attentive statistics pooling in the temporal dimension, which have recently been applied to Automatic Speech Recognition (ASR) [17], instead of simple maximum pooling or average pooling. Attentive statistics pooling can improve the discriminability of features.

The ENIV2 framework has shown excellent performance in SELD task. In view of this, we combine Model 1 with the EINV2 framework to form Model 2. The major difference between Model 2 and the EINV2 framework is that the output format of Model 2 is multi-ACCDOA, while the output format of the EINV2 framework is a two-branch structure. The structure of Model 2 is shown in Fig. 1(b).

2.3. Data Augmentation

In this year's challenge, the officially provided dataset Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23) has been increased from last year, but it still does not make the model robust. To address this issue, we firstly generate more spatial audio files with a duration of 1 minute and a maximum polyphony of 3 for training using Audioset, FSD50k and SRIRs from TAU-SRIR DB. At the same time, we use audio channel swap technique (ACS) [18] on the real dataset, which expands the existing real dataset up to eight times. Then, random cutout, time-frequency masking and frequency shifting are used to enhance the generalization of the model. Finally, we employ augmix, which is widely used in the image field, to add the original data and the enhanced data with certain weights to form the new data for training.

2.4. Post-processing

In the inference phase, to further improve the performance of each training model, we apply the test time augmentation (TTA) [19]. We apply the audio channel swap technique to the test audio files, and then estimate the output for each audio file while inverting the output accordingly. Finally, the average of the 8 outputs is taken as the final output.

3. EXPERIMENTS

3.1. Dataset

We validate the proposed systems on STARSS23, which consists of a development set (dev-set) for the development phase and an evaluation set (eval-set) for the validation phase. The dev-set consists of a train-set with 90 real recordings and a test-set with

78 real recordings. The total duration of real recordings in the train-set and test-set is about 4 hours and 3 hours, respectively. The duration of each real recording varies from 30 seconds to 10 minutes. There are also 1200 audio recordings of 1 minute duration in the synthetic set. The number of sound events is 13. The common number of overlapping sound events is 3.

3.2. Experimental setup and Evaluation Metrics

Only the dataset in FOA format is used for our experiments. We extract audio features in the same way as the baseline. The sampling frequency is set to 24kHz, the number of Mel filters is set to 64, and the STFT is used with 40ms frame length and 20ms frame hop. The length of input is 250 frames. Adam optimizer is used. The batch size is 256. The model is trained for 100 epochs. The learning rate is set to 0.001 in the first 60 epochs, decreasing from 0.001 to 0.0001 in the 60th to 90th epochs, and 0.00001 in the last 10 epochs. Meanwhile, in order to solve the problem of imbalance between the number of real and synthetic datasets, we set half of the datasets from real datasets and half from synthetic datasets for each training round.

We employ the official metrics [20] to evaluate our SELD system. The official metrics are location-dependent error rate ($\mathrm{ER}_{\leq 20^\circ}$), location-dependent F-score ($\mathrm{F}_{\leq 20^\circ}$), class-dependent localization error ($\mathrm{LE}_{\mathrm{CD}}$), and localization recall ($\mathrm{LR}_{\mathrm{CD}}$).

3.3. Experiment Results

Table 1 shows the performance of our proposed method on the development dataset. As shown in Table 1, both Model 1 and Model 2 outperform the baseline system. To obtain a better performance, we ensemble Model 1 and Model 2 and average their outputs. The difference between Ensemble1 and Ensemble2 is the selection of their sub-models applying different training strategies and different distributions of the datasets.

Table 1: The SELD performance of our system for dev-test set.

System	$\mathrm{ER}_{\leq 20^\circ}$	$\mathrm{F}_{\leq 20^\circ}$	$\mathrm{LE}_{\mathrm{CD}}$	$\mathrm{LR}_{\mathrm{CD}}$
Baseline(FOA)	0.57	29.9%	22°	47.7%
Model 1	0.43	54.8%	14.7°	68%
Model 2	0.44	54.2%	13.9°	67.9%
Ensemble 1	0.42	55.7%	13.9°	67.7%
Ensemble 2	0.41	56.4%	13.7°	67.8%

4. CONCLUSION

In this report, we present our approach for DCASE2023 task 3. We apply the multi-scale channel attention mechanism and attentive statistics pooling to the SELD system. At the same time, we utilize a novel data augmentation approach augmix, which combines random cutout, time-frequency masking and frequency shifting. Then, we merge our proposed network and EINV2 framework with multi-ACCDOA output format. Considering the problem of unbalanced dataset, we employ different strategies in the training phase. In the inference phase, we introduce TTA and Ensemble to improve the performance of the system. The experimental results show that our proposed system frameworks have better performance than the baseline.

5. REFERENCES

[1] Virtanen, Tuomas, Mark D. Plumbley, and Dan Ellis, eds. Computational analysis of sound scenes and events. Berlin, Germany: Springer International Publishing, 2018.

[2] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi and Y. Mitsufuji, "Accdoa: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization And Detection," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 915-919.

[3] 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," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 316-320.

[4] Y. Liu and D. Wang, "Permutation Invariant Training for Speaker-Independent Multi-Pitch Tracking," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5594-5598.

[5] Cao, Yin, et al. "Event-independent network for polyphonic sound event localization and detection." arXiv preprint arXiv:2010.00140 (2020).

[6] 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," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 885-889.

[7] Okabe, Koji, Takafumi Koshinaka, and Koichi Shinoda. "Attentive statistics pooling for deep speaker embedding." arXiv preprint arXiv:1803.10963 (2018).

[8] Dai, Yimian, et al. "Attentional feature fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[9] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, ″Random erasing data augmentation," in Proc. of AAAI 2020, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[10] 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, 2019, pp. 2613 – 2617.

[11] T. T. N. Nguyen, K. N. Watcharasupat, K. N. Nguyen, D. L. Jones, and W.-S. Gan, “SALSA: Spatial cue-augmented logspectrogram features for polyphonic sound event localization and detection,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process., vol. 30, pp. 1749–1762, 2022.

[12] Hendrycks, Dan, et al. "Augmix: A simple data processing method to improve robustness and uncertainty." arXiv preprint arXiv:1912.02781 (2019).

[13] 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 Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

[14] 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, 2022.

[15] A. Politis, S. Adavanne, and T. Virtanen, “TAU Spatial Room Impulse Response Database (TAU- SRIR DB),” Apr. 2022.[Online].Available:https://doi.org/10.5281/zenodo.6408611

[16] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, “Conformer: Local features coupling global representations for visual recognition,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 357–366.

[17] D. Liao, T. Jiang, F. Wang, L. Li and Q. Hong, "Towards A Unified Conformer Structure: from ASR to ASV Task," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5.

[18] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, ״A fourstage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection,” arXiv preprint arXiv:2101.02919, 2021.

[19] J. Cohen, E. Rosenfeld, and Z. Kolter, ״Certified adversarial robustness via randomized smoothing,” in International Conference on Machine Learning. PMLR, 2019, pp. 1310–1320.

[20] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, “Overview and evaluation of sound event localization and detection in dcase 2019,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 684–698, 2020.