THE NERC-SLIP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2022 CHALLENGE

Technical Report

Qing Wang[1]*, Li Chai*[2]*, Huaxin Wu*[2]*, Zhaoxu Nian*[1]*, Shutong Niu*[1]*, Siyuan Zheng*[1]*, Yuyang Wang*[1]*, Lei Sun*[2]*, Yi Fang*[2]*, Jia Pan*[2]*, Jun Du*[1]*, Chin-Hui Lee*[3]

[1] University of Science and Technology of China, Hefei, China, {qingwang2, cl122}@ustc.edu.cn
{zxnian, niust, zsy19, wyy201221, sunlei17, panjia}@mail.ustc.edu.cn, {jundu}@ustc.edu.cn
[2] iFLYTEK, Hefei, China, {hxwu2, yifang2}@iflytek.com
[3] Georgia Institute of Technology, Atlanta, USA, {chl}@ece.gatech.edu

ABSTRACT

This technical report describes our submission system for the task 3 of the DCASE2022 challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. Compared with the official baseline system, the improvements of our method mainly lie in three aspects: data augmentation, more powerful network architecture, and model ensemble. First, our previous work shows that the audio channel swapping (ACS) technique [1] can effectively deal with data sparsity problems in the SELD task, which is utilized in our method and provides an effective improvement with limited real training data. In addition, we generate multichannel recordings by using public datasets and perform data cleaning to drop bad data. Then, based on the augmented data, we employ a ResNet-Conformer architecture which can better model the context dependencies within an audio sequence. Specially, we found that time resolution had a significant impact on the model performance: with the time pooling layer moving back, the model can obtain a higher feature resolution and achieve better results. Finally, to attain robust performance, we employ model ensemble of different target representations (e.g., activity-coupled Cartesian direction of arrival (ACCDOA) and multi-ACCDOA) and post-processing strategies. The proposed system is evaluated on the dev-test set of Sony-TAu Realistic Spatial Soundscapes 2022 (STARS2022) dataset.

Index Terms— Sound event localization and detection, data augmentation, model ensemble, conformer

1. INTRODUCTION

The goal of sound event localization and detection (SELD) task is to detect the presence of sound events, and localize them in time and space when active. SELD can be applied in many areas [2]. Environmental types can be recognized and suppressed to improve speech quality during speech communication or to improve performance of robust automatic speech recognition (ASR). In smart cities, audio surveillance system plays an indispensable role.

The SELD task in first 3 DCASE challenges 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 [3, 4, 5]. This year the challenge task changes considerably compared to the previous iterations since it transitions from computationally generated spatial recordings to recordings of real sound scenes, manually annotated. Similarly to the previous iterations of the challenge, a straightforward convolutional recurrent neural network (CRNN) based on SELDnet is used as the baseline system, but with a few important modifications [6]. In DCASE2021 challenge, the baseline adopted the ACCDOA representation [7] for training localization and detection with a single unified regression vector loss. In this challenge, the baseline[1] adopts the strategy of multi-ACCDOA proposed by [8] with the output of the model switched to a track-based format in order to make it suitable for handling simultaneous events of the same class.

In this report, data augmentation approaches are investigated to expand official dataset. We adopt spatial augmentation to simulate new DOA information by audio channel switching (ACS) proposed in [1]. We generate additional synthetic mixtures following the same process as the generation code[2] by using several sound event databases. In order to solve the SELD for overlapping sound events, the augmented data is generated according to a certain proportion of overlapping events. We use the Resnet-Conformer [1] as the training network based on the augmented data, which moves the time pooling layer back to improve model recognition results. In order to make the model more adaptable to the situation of overlapping events of the same class, we make a model ensemble for two different target representations: ACCDOA-based method and multi-ACCDOA-based method. Besides, we use a post-processing method to make the model more accurate.

The rest of the report is organized as follows. In Section 2, the proposed method is described in detail, including data augmentation, network training, model ensemble and post-processing. Experimental results on development dataset is shown in Section 3. Conclusions are summarized in Section 4.

2. PROPOSED METHOD

In the proposed method, some useful data augmentation methods are utilized to generate training data. Then, strong deep neural network (DNN) architectures called Resnet-Conformer are respectively trained for both ACCDOA and multi-ACCDOA formats based on augmented training data. Model ensemble and post-processing are adopted to get the final sound event detection and localization estimation. We will describe the four parts of the method

[1]https://github.com/sharathadavanne/ seld-dcase2022
[2]https://github.com/danielkrause/ DCASE2022-data-generator

in detail: the data augmentation, the network training, the model ensemble, and the post-processing.

2.1. Data Augmentation

The official real training data consists only 3 hours recordings. It is obviously that such a small data set can not make the deep learning-based model robust. So data augmentation approaches are necessary for the training of the SELD system. In the technical report, we adopt two data augmentation methods. One is ACS spatial augmentation proposed in our previous work [1] to increase DOA representations based on the rotational properties of the recorded data set. The other is to simulate new multi-channel data by using provided SRIRs and sound samples selected from several public datasets.

In our submitted system, only FOA format data is used based on the results of preliminary experiment. ACS [1] has been demonstrated as a strong data augmentation strategy in multi-channel SELD tasks. The amount of data can be augmented to 8 times after applying the ACS strategy, which is about 184 hours for the training split in the development dataset. In addition, single-channel sound samples extracted from AudioSet[9], ESC[10], FSD50K[11] datasets are convoluted with SRIRs from TAU-NIGENS Spatial Sound Events Datasets[12, 13] to generate 1-minute long multi-channel scene recordings with a maximum polyphony of 3, which yields several hundreds hours of data.

Besides, a data cleaning strategy is used to improve the quality of the simulated recordings. We firstly train a SELD model on the official development dataset applied by ACS augmentation. Then, we test the simulated scene recordings using the trained model. The data with high SELD scores is considered as poor quality and will be dropped in the cleaning procedure. Finally, about 300 hours of data are generated as the final training set.

2.2. Network Training

In the proposed system, only FOA format data is adopted as the training data. Log-mel spectra features are extracted from multi-channel audio of 24 kHz sampling rate using 1024-point discrete Fourier transform from a 40 msec Hanning window and 20 msec hop length. Then 4-channel mel spectra features and 3-channel intensity vectors are concatenated together to get the 7-channel feature for every input frame of the model.

The Resnet-Conformer [1] as shown in Fig.1 is adopted as the main network architecture in our method. Resnet [14] is a well-known convolutional neural network (CNN) based network that has achieved great performance in different tasks. Here a modified Resnet architecture is used as a feature extractor to get high-level feature representation from the input filter bank features. Conformer [15] is a network widely used in speech recognition areas to replace the Transformer encoder and achieves state-of-the-art (SOTA) performance. The Conformer module performs a combination of self-attention and convolution operation. With self-attention capturing global-level dependencies and convolution learning local-level relationship, Conformer can utilize global and local information at the same time. The output dimension of Resnet is reduced from 512 to 256 before feeding into the Conformer by adopting a linear layer. ACCDOA and multi-ACCDOA are trained through modifying the output dimension of the last linear layer for further fusion. One important improvement in the structure is that we found the Conformer module is suit for a high time resolution. So the pooling on time frames is performed after the Conformer module instead of in the Resnet module, which makes the Conformer learn more complete information and achieve better result.

The training data are cut into 20-second long segments. Adam optimizer is used to train the model and a tri-stage learning rate scheduler is used to control the learning rate. Models are trained on 184 hours of data and then fine-tuned on a larger 300 hours of training data. The upper limit of learning rate is 0.001 when training from scratch and 0.0001 when fine-tuning. Batch size is set to 16 and iteration steps are set to 120,000. The optimization criterion of ACCDOA is mean square error (MSE) while that of multi-ACCDOA is the MSE loss with auxiliary duplicating permutation invariant training (ADPIT) [8].

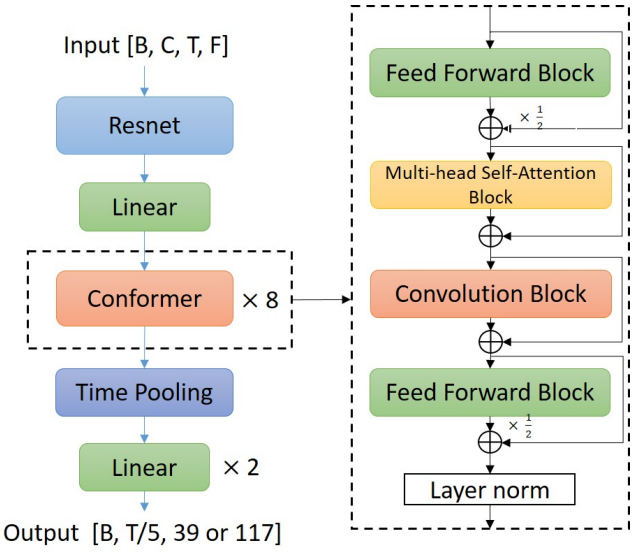

Figure 1: The network architecture of Resnet-Conformer.

2.3. Model Ensemble

Two model ensemble strategies are utilized to improve the generalization ability and achieve better results. Preliminary experiment shows that the set of random seed will bring some fluctuations to the experimental results. So Resnet-Conformer models with ACCDOA target representation trained with different random seeds are fused to reduce randomness. However, for multi-ACCDOA target representation, we do not perform model ensemble between different random seeds.

Besides, a ACCDOA and multi-ACCDOA fusion strategy is proposed to fully utilize the advantages of these two modeling methods. ACCDOA-based modeling method can provide accurate boundary information. Multi-ACCDOA-based method can process the overlap segments of the same event class but may introduce false alarms. Here the boundary information from ACCDOA is combined with the SED and DOA estimation of multi-ACCDOA. Assume class c happened at frame t in ACCDOA estimation. If the SED estimation of multi-ACCDOA at frame t is the same as ACCDOA, we will calculate the angle difference between the two estimated events. If the difference of DOA estimation is higher than a specific threshold, we think the DOA estimated by multi-ACCDOA model is more accurate. Otherwise we think these two events are

Table 1: Experimental results of the proposed method for development dataset.

	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
Baseline-FOA	0.71	21.0%	29.3°	46.0%
Baseline-MIC	0.71	18.0%	32.2°	47.0%
ACCDOA	0.40	65.0%	15.0°	77.0%
Multi-ACCDOA	0.41	61.0%	15.3°	74.0%
Ensemble system	**0.38**	**67.0%**	**14.8°**	**78.0%**

exactly the same event, and the final DOA estimation will be the mean value of these two methods.

2.4. Post-processing

Two post-processing strategies are adopted to further improve the system performance. First, when testing the input data is cut into 20-second long segments with a 1 second hop length. Then the result of each time frame is the mean value of the time-overlapped segments. Tested on time-overlapped segments can further decrease the variance of the results. Second, dynamic threshold is adopted to improve the SED performance. The threshold is chosen on the real validation set.

3. RESULTS ON DEVELOPMENT DATASET

We evaluate our proposed method on the development dataset of Sony-TAu Realistic Spatial Soundscapes 2022. We generate a larger training set with the abovementioned data augmentation approaches, namely audio channel swapping and multi-channle data simulation, which consists of about 300 hours augmented data. Table 1 shows the experimental results of the proposed method for development dataset. "ACCDOA" represents the ACCDOA-based modeling method and "Multi-ACCDOA' represents the multi-ACCDOA-based method, both of which used post-processing strategies. As shown in the table, each proposed single model outperforms the two baseline systems by a large margin. By fusing the SELD results predicted by ACCDOA and multi-ACCDOA methods, further improvements are achieved as shown in the last row of Table 1.

4. CONCLUSION

In this report, we propose an ensemble system to solve the SELD task in DCASE2022 challenge. We first adopt data augmentation approaches to expand the official dataset. Then the ResNet-Conformer architectures are trained to predict SED and DOA results in ACCDOA and multi-ACCDOA representation formats. Finally model ensemble and post-processing strategies are used to get a more robust SELD estimation. The experimental results on the development dataset show that the proposed method outperforms the baseline systems by a significant margin.

5. REFERENCES

[1] 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 preprint arXiv:2101.02919*, 2021.

[2] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

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

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

[5] http://dcase.community/challenge2021/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, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, June 2021.

[8] 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.

[9] 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.

[10] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[11] 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.

[12] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 165–169. [Online]. Available: https://dcase.community/workshop2020/proceedings

[13] 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 *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 125–129. [Online]. Available: https://dcase.community/workshop2021/proceedings

[14] 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.

[15] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.