JLESS SUBMISSION TO DCASE2023 TASK3: CONFORMER WITH DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION IN REAL SPACE

Technical Report

Dongzhe Zhang[1,2]*, Jisheng Bai*[1,2]*, Siwei Huang*[1]*, Mou Wang*[1]*, Jianfeng Chen*[1,2]

[1] Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
[2] LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{dongzhezhang2022, baijs, hsw838866721, wangmou21}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

In this technical report, we describe our proposed system for DCASE2023 task3: Sound Event Localization and Detection(SELD) Evaluated in Real Spatial Sound Scenes. At first, we review the famous deep learning methods in SELD. Then we apply various data augmentation methods to balance the sound event classes in the dataset, and generate more spatial audio files to augment the training data. Finally, we use different strategies in the training stage to improve the generalization of the system in realistic environment. The results show that the proposed systems outperform the baseline system on the dev-settest of Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23).

Index Terms— sound event localization and detection, Data augmentation, model ensemble, real spatial scenes

1. INTRODUCTION

Sound Event Localisation and Detection (SELD) is a challenging task that aims to detect and locate the sound events as well as to estimate their corresponding direction of arrival (DOA) using multichannel spatial audio[1, 2]. Given the complexity of real-world acoustic environments, SELD systems have the potential to be extremely useful in diverse applications such as surveillance systems, smart homes, public safety, and outdoor navigation for the visually impaired.

The SELD system was first introduced in the DCASE2019 Task3 [3] and involved the use of single static sound sources. Multichannel audio files were synthesised by combining mono audio files and impulse response in real rooms, allowing manual control over factors such as signal-to-noise ratio (SNR), event occurrence and arrival direction. However, subsequent SELD [4, 5, 6] challenges introduced several new factors that made the task much more challenging. These included new impulse responses, moving sources, polyphonic events and overlapping events of the same class. In the latest SELD challenge [7], the task is set in real spatial sound scenes with even more complex environments and lower SNR. The factors for sound events are no longer manageable and depend on the exact layout of the room. Furthermore, video information has also been added to complement multi-channel audio information for target detection and localization.

In this report, we propose a SELDnet based neural network with data augmentation for SELD. The entire framework of the SELD system is built upon two main components: SELDnet and multi-track ACCODA [8]. SELDnet is a neural network architecture that combines spatial information with spectrogram representations to accurately classify and localize sound events. And the multi-track ACCODA algorithm is used to handle same-class overlapping sound events and extract precise localization information for each event. In addition to the existing dataset, we have generated additional synthesized data using the FSDK50[9] and TAU-SRIR DB[10]. To alleviate the class imbalance, we generate more data for several classes with less data or poorer performance. Meanwhile, in order to avoid the model from overfitting on the synthesized data, we employ a strategy to train our model on a combination of the real and synthesized data, and fine-tune the model the real recordings.

2. PROPOSED METHOD

In this section, we first introduce the input features of the proposed SELD system. Then we introduce the data augmentation, network architecture and training procedures.

2.1. Features

Our network's input features consists of the signals from four channels of first-order ambisonics (FOA). We specifically emphasize on using first-order ambisonics (FOA) signals as they don't contain spatial aliasing within the range of 9 kHz. And the FOA format was preferred as it performed better than the MIC format in the baseline system. FOA features contain seven channels, including four log-mel spectrograms and three intensity vectors.

2.2. Data augmentation

Since the dataset provided by DCASE only comprises 1200 synthetic files, we augmented the training data to enhance the model's performance. By using external data provided and TAU-SRIR DB, we synthesized 2400 additional files for training. The synthetic data are FOA format data consisting of 13 classes. Notably, the maximum polyphony was defined at 2, with a duration of 60 seconds and a sampling rate of 44.1 kHz. In order to alleviate the class imbalance and allow the model to have high detection and localization capabilities in each class, we also generated more synthesized data for several classes with poor performance. Also, We introduce three

Table 1: 16 patterns of channel rotation. Swap($-X,Y$) denotes $X' \leftarrow Y$ and $Y' \leftarrow X$.

	$\phi - \pi/2$	ϕ	$\phi + \pi/2$	$\phi + \pi$
θ	Swap($-X,Y$)	-	Swap($X,-Y$)	$Y' \leftarrow -Y$, $X' \leftarrow -X$
$-\theta$	Swap($-X,Y$), $Z' \leftarrow -Z$	$Z' \leftarrow -Z$	Swap($X,-Y$), $Z' \leftarrow -Z$	Swap($-X,-Y$), $Z' \leftarrow -Z$
	$-\phi - \pi/2$	$-\phi$	$-\phi + \pi/2$	$-\phi + \pi$
θ	Swap($-X,-Y$)	$Y' \leftarrow -Y$	Swap(X,Y)	$X' \leftarrow -X$
$-\theta$	Swap($-X,-Y$), $Z' \leftarrow -Z$	$Y' \leftarrow -Y,Z' \leftarrow -Z$	Swap($-X,Y$),$Z' \leftarrow -Z$	$X' \leftarrow -X,Z' \leftarrow -Z$

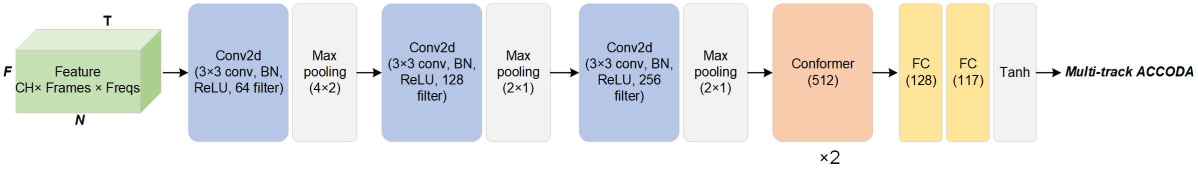

Figure 1: Overall architecture of the proposed network.

data augmentation methods in our SELD system: mixup[11], Random cutout[12] and channel rotation [13]. The mixup technique has gained widespread adoption in the realm of environmental sound recognition. In mixing augmentation method, the data and labels are mixed to generate new training data. And random cutout is also utilized to mask specific feature areas while keeping their original labels. We use 16 spatial ransformation methods, which rotate audio channels and change the spatial labels, to augment the spatial data of FOA format. Table 1 shows 16 patterns of channel rotation.By applying channel rotation, we double the amount of data.

2.3. Network architecture

The model was created based on the baseline CRNN structure,and multi-ACCDOA was applied to predict the SED and DOA ina single branch. Figure 1 shows the overall structure of the proposed model. The primary aim of our network is to extract spatial information from the given input of FOA features. This is achieved by feeding the log-mel spectrograms from all four FOA channels, together with the IV channels, into a convolutional network comprising three layers. Bidirectional GRU layers are replaced with two Conformer[14] blocks.

3. EXPERIMENTS

In this section, we show our results on the development dataset.

3.1. Experimental settings

We evaluated our proposed methods on the Sony-TAu Realistic Spatial Soundscapes 2023 dataset, and compared our systems with the baseline system. The baseline is a multi-ACCDOA-based system using CRNN network. Five metrics are used for evaluation : error rate (ER_{20°), F-score (F_{20°), LE_{CD} , LR_{CD} , SELD_{score} [15]. Except for error rate, the other four metrics are computed per class and macro-averaged. We use only the FOA subset of the dataset for out experiments.

We follow the settings of the baseline during feature extraction and down sampling. 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. We use a batch size of 64. The model is firstly trained on simulated data for 100 epochs with learning rate of 0.0005. In the fine-tune stage, The saved best result is further trained on real recordings for extra 30 epochs with learning rate of 0.1 decay.

3.2. Results

Table 2 shows the performance with the development set for proposed methods. As shown in the table, our proposed method outperforms the baseline by a large margin. For model ensemble, we average outputs from different networks, data, augmentation methods. Model ensemble also has a better performance than single model.

4. CONCLUSION

We present the proposed SELD system of DCASE2023 task3. we apply various data augmentation methods to balance the sound event classes in the dataset, and generate more spatial audio files to augment the training data. Considering the difference between simulated spatial audios and real recordings in exclusive environment, we use different strategies in the training stage to improve the generalization of the system in realistic environment. Finally, we perform model ensemble with different models, augmentation methods. Our proposed system achieve great improvement and significantly outperform the baseline system.

5. REFERENCES

[1] S. Kapka and M. Lewandowski, “Sound source detection, localization and classification using consecutive ensemble of crnn models,” *arXiv preprint arXiv:1908.00766*, 2019.

[2] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, “The ustc-iflytek system for

Table 2: SELD performance of our systems evaluated by using joint metrics for the development set.

system	$\mathrm{ER}_{20^\circ}\downarrow$	$\mathrm{F}_{20^\circ}(macro)\uparrow$	$\mathrm{LE}_{CD}\downarrow$	$\mathrm{LR}_{CD}\uparrow$	SELD↓
baseline-FOA	0.57	29.9%	22°	44.7%	0.487
model1	0.5	49.1%	16.5°	63%	0.368
model2	0.51	46.6%	16.8°	61.7%	0.380
model3	0.51	46.5%	16.8°	60.8%	0.383
ensemble#1	0.47	51.1%	14.6°	60.9%	0.358
ensemble#2	0.47	49.6%	15.6°	58.9%	0.368
ensemble#3	0.46	52.0%	14.0°	59.5%	0.357
ensemble#4	0.48	49.2%	15.1°	61.6%	0.364

sound event localization and detection of dcase2020 challenge," *IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events*, 2020.

[3] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[4] https://dcase.community/challenge2020.

[5] https://dcase.community/challenge2021.

[6] https://dcase.community/challenge2022.

[7] https://dcase.community/challenge2023.

[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 *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 316–320.

[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] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[12] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[13] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[14] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 367–376.

[15] 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.