JLESS SUBMISSION TO DCASE2022 TASK3: DYNAMIC KERNEL CONVOLUTION NETWORK WITH DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION IN REAL SPACE

Technical Report

Siwei Huang[1], *Jisheng Bai*[1,2], *Yafei Jia*[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
{hsw838866721, baijs, jyf2020260709, wangmou21}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

This technical report describes our proposed system for DCASE2022 task3: Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes. In our approach, we first introduce a dynamic kernel convolution module after the convolution blocks to dynamically model the channel-wise features with different receptive fields. We then incorporate the SELDnet and EINV2 framework into the proposed SELD system with multi-track ACCDOA output. Finally, we use different strategies in the training stage to improve the generalization of the system in realistic environment. Moreover, we apply data augmentation methods to balance the sound event classes in the dataset, and generate more spatial audio files to augment the training data. Experimental results show that the proposed systems outperform the baseline on the development dataset of DCASE2022 task3.

***Index Terms*—** Data augmentation, dynamic convolution, sound event localization and detection, real spatial scenes

1. INTRODUCTION

Sound event localization and detection (SELD) is a task that involves sound onset, offset detection (SED) and the corresponding direction of arrival (DOA) estimation using multi-channel spatial audios. SELD systems can be potentially used in many applications, such as surveillance systems and outdoor navigation.

SELD system was firstly proposed in DCASE2019 task3 [1], using single static sound sources. The multi-channel audio files were synthesized using monopoly sound audio files and impulse response in real rooms, which means SNR, the occurrence of each class of events and direction of arrival can be manually set. During the previous SELD challenges, sort of the factors including new impulse response, moving sources, polyphony events, same-class overlapping events, have made the SELD task more difficult in various aspects. SELD task in DCASE2022 [2] is put in real spatial sound scenes with more complex environment and lower SNR. The factors of sound events are no more manageable, which depend on the layout of the exact rooms. The imbalanced presence of each event class will make it harder to detect the sound event accurately.

In the previous challenges [3, 4, 5], participants were provided with audio in both ambisonics (FOA) and microphone array (MIC) format. Main features used in the SELD systems are log-Mel spectrogram and intensity vectors (IVs) for FOA data, and log-Mel spectrogram and GCC-PHAT for MIC data [6]. FOA features are with energy consistency, which shows better SELD performance than MIC features. However, Nguyen et al. [7] used SALSA-Lite for MIC data and achieved state-of-the-art performance for SELD.

In this study, we propose a dynamic kernel (DK) convolutional based neural network with data augmentation for SELD. The whole framework is based on SELDnet [1], ENIV2 framework [8] and multi-track ACCODA [9], which has the ability to detect same-class overlapping sound events. The DK convolution module is applied for modeling channel-wise features of different receptive fileds. Moreover, we generate more synthesized data using FSDK50 [10] and TAU-SRIR DB [11] to augment the development dataset. Considering that the synthesized data and real recordings in the development dataset have different characteristics, we conduct two strategies for training, which achieve better performance in real spatial scenes.

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

SALSA-Lite of microphone array format has shown its efficiency in DCASE2022 baseline system. In order to reduce the size of the input features, we multiply both spectrogram and SALSA-Lite spatial features with Mel filters. The feature fed into the model is combined with log-Mel spectrogram and SALSA-Lite, which is called SALSA-Mel.

2.2. Data augmentation

Given that we have less spatial data in this year's challenge, to increase the generalizability of the model, We introduce four data augmentation methods in our SELD system: FMix [13], mixup [14], Random cutout [15] and MIC format channel rotation [12]. We use mixing augmentation method, i.e., FMix and mixup, which have

Table 1: MIC data rotation/swapping original by[12], here we make some adjustments. ϕ and θ denote the azimuth angle and elevation angle, C_M^{new} and C_M denote the new channel and original channel.

DOA transformation	MIC channel swapping
ϕ = ϕ-π/2, θ= -θ	C_1^{new}=C_3,C_2^{new}=C_1,C_3^{new}=C_4,C_4^{new}=C_2
ϕ = -ϕ-π/2, θ=θ	C_1^{new}=C_4,C_2^{new}=C_2,C_3^{new}=C_3,C_4^{new}=C_1
ϕ = ϕ, θ=θ	C_1^{new}=C_1,C_2^{new}=C_2,C_3^{new}=C_3,C_4^{new}=C_4
ϕ = -ϕ, θ= -θ	C_1^{new}=C_2,C_2^{new}=C_1,C_3^{new}=C_4,C_4^{new}=C_3
ϕ = ϕ+π/2, θ= -θ	C_1^{new}=C_2,C_2^{new}=C_4,C_3^{new}=C_1,C_4^{new}=C_3
ϕ = -ϕ+π/2, θ= θ	C_1^{new}=C_1,C_2^{new}=C_3,C_3^{new}=C_2,C_4^{new}=C_4
ϕ = ϕ+π, θ= θ	C_1^{new}=C_4,C_2^{new}=C_3,C_3^{new}=C_2,C_4^{new}=C_1
ϕ = -ϕ+π, θ= -θ	C_1^{new}=C_3,C_2^{new}=C_4,C_3^{new}=C_1,C_4^{new}=C_2

been widely used for environmental sound recognition. In mixing augmentation method, the data and labels are mixed to generate new training data. Random cutout is used to mask areas of features without changing its original labels. We use 8 spatial ransformation methods, which rotate audio channels and change the spatial labels, to augment the spatial data of MIC format. In addition, we further use SELD data generator provided and sound event audio files in FSD50K to generate more spatial audio files and balance the occurrence for each event class.

2.3. Network architecture

We propose three networks with DK convolution modules [16], which achieve significant performance in speech recognition. The architecture of the DK convolution module is illustrated in Figure 1. In the proposed system, we introduce DK convolution module in SELDnet and ENIV2 framework to adaptively model the acustic features.

In baseline SELD network, we apply the following changes: The amount of hidden size is extended, two DK convolution modules follows with residual connection are applied after the final convolutional blocks. Bidirectional GRU layers are replaced with two Conformer [17] blocks.

ENIV2 framework has outperformed the pervious SELD systems. We introduce two various updates with DK convolution modules: The first is that the DK convolution modules are only for DOA branch and we only focus on one-branch multi-track ACCDOA outputs. The second is that we both apply DK convolution modules after two branches, and use deeper CNN and other soft parameter sharing block during the output of CNN embeddings and Conformer embeddings. Then, the embeddings are concatenated and we get multi-track ACCDOA outputs using one fully-connected layer. Network architectures in detail can be found in Figure 2.

2.4. Training strategy

Considering that there are differences between simulated data and real data, including impulse response and SNRs. To make model more sensitive to real spatial recordings, we have tried the two training strategies.

- The model is firstly pretrained only with simulated spatial audios, and then trained with real spatial recordings with a lower learning rate.

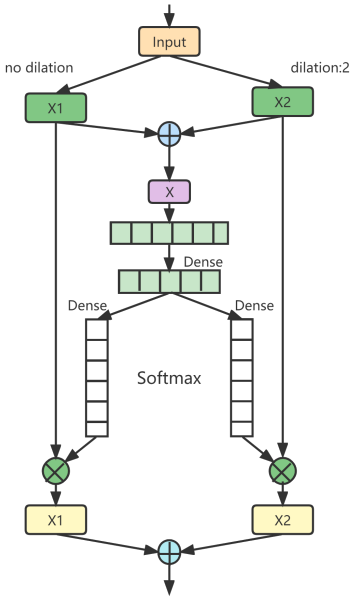

Figure 1: The architecture of DK convolution module. $\bigoplus$ denotes element-wise addition.$\bigotimes$ denotes element-wise multiplication.

- Mix all audio files and achieve the model of best validation results, then further train the model using only real spatial recordings with a lower learning rate.

2.5. Post-processing

During the inference, we generate inference data with different hop frames, then we sum the overlapping outputs with different weights. Furthermore, we rotate MIC data, estimate the multi-track ACCDOA vector, and rotate it back [18] with 8 rotation methods. The final results are averaged on these outputs.

3. EXPERIMENTS

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

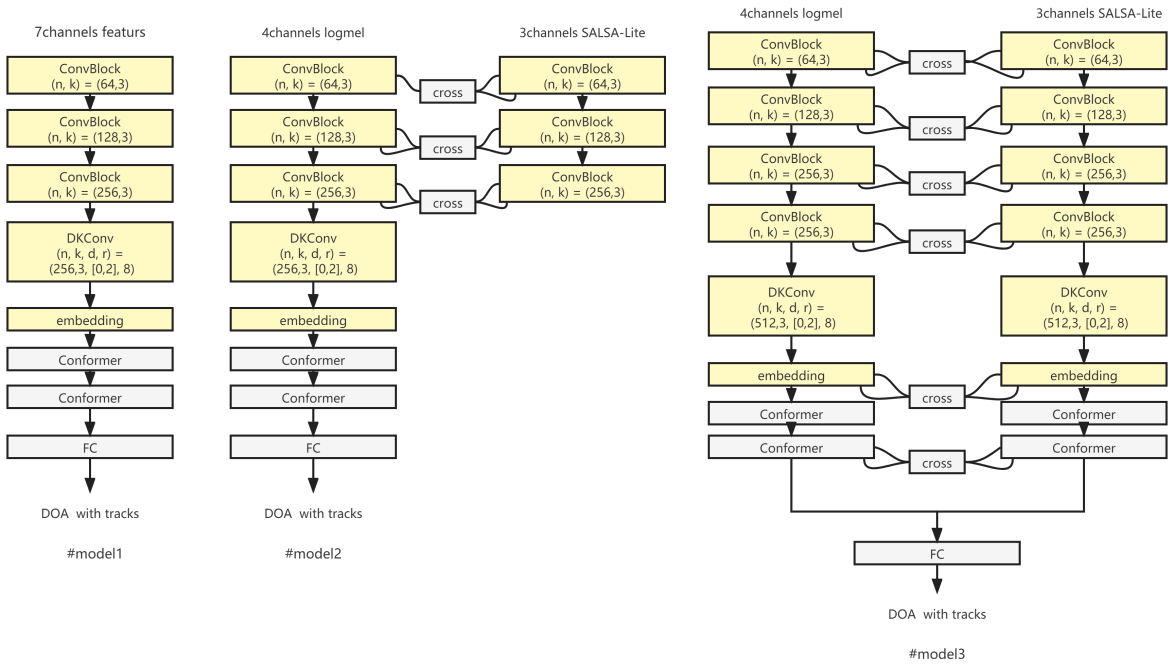

Figure 2: The constructions of threes models, model1 is based on SELDnet, model2 and model3 are based on ENIV2 framework.Parameters n, k ,d ,r denote number of filters, kernel size, dilation rate, reduction rate.Cross denotes Cross-stitch parameter sharing block in ENIV2 framework.

3.1. Experimental settings

We evaluated our proposed methods on the Sony-TAu Realistic Spatial Soundscapes 2022 dataset, and compared our systems with the baseline. The baseline is an ACCDOA-based system using CRNN, with changes of multi-track ACCDOA and SALSA-Lite features for MIC. Five metrics are used for evaluation[19]: error rate (ER_{20°), F-score (F_{20°), LE_{CD}, LR_{CD}, $SELD_{score}$. Except for error rate, the other four metrics are computed per class and macro-averaged.

We use only the MIC 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. For second training strategy mentioned, model is firstly trained on simulated data for 100 epochs with learning rate of 0.0005. The saved best result is further trained on real recordings for extra 25 epochs with learning rate of 0.1 decay.

By applying data augmentation, we generate two more folds of new data, and triple the amount of data using channel rotation. During training on real recordings, the amount of recordings is extended by 8 times using channel rotation.

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 and training strategies on each track. Model ensemble also has a better performance than single model. The results of using two training strategies are very close, which outperform mixture training in the baseline. And we only use the second training strategy.

4. CONCLUSION

We present the proposed SELD system of DCASE2022 task3. We apply DK convolution modules and Conformer block into SELD systems, then incorporate different networks with DK convolution modules based on multi-track ACCDOA systems and ENIV2 systems. Considering the difference between simulated spatial audios and real recordings in exclusive environment, we use different training strategies. Finally, we perform model ensemble with different models, augmentation methods, audio files in different distributions, and training strategies. Our proposed system achieve great improvement and significantly outperform the baseline system.

5. REFERENCES

[1] 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, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[2] https://dcase.community/challenge2022/task-sound-event-localization-and-detection.

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

system	$ER_{20^\circ}\downarrow$	$F_{20^\circ}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$SELD_{score}\downarrow$
baseline-MIC	0.71	18	32.2°	47	-
model1	0.50	49.3	18.2°	57.6	0.383
model2	0.50	45.4	18.1°	60.3	0.385
model3	0.50	45.4	17.5°	59.0	0.388
ensemble#1	0.47	52.4	16.1°	62.1	0.355
ensemble#2	0.51	50.0	17.1°	68.1	0.358
ensemble#3	0.44	54.2	16.0°	65.4	0.333
ensemble#4	0.48	52.2	16.9°	70.7	0.335

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

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

[5] https://dcase.community/challenge2021/task-sound-event-localization-and-detection.

[6] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," 2019. [Online]. Available: https://arxiv.org/abs/1904.03476

[7] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022.

[8] 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," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 885–889.

[9] 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.

[10] 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.

[11] 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.

[12] 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," 2021. [Online]. Available: https://arxiv.org/abs/2101.02919

[13] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, "Fmix: Enhancing mixed sample data augmentation," 2020. [Online]. Available: https://arxiv.org/abs/2002.12047

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017. [Online]. Available: https://arxiv.org/abs/1710.09412

[15] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017. [Online]. Available: https://arxiv.org/abs/1708.04896

[16] T. Kong, S. Yin, D. Zhang, W. Geng, X. Wang, D. Song, J. Huang, H. Shi, and X. Wang, "Dynamic multi-scale convolution for dialect identification," 2021. [Online]. Available: https://arxiv.org/abs/2108.07787

[17] 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.

[18] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 915–919.

[19] 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. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9306885