DOA AND EVENT GUIDANCE SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION

Technical Report

Hogeon Yu[1*],

[1] Hyundai Motor Company, Robotics Lab, South Korea,
hogeon.yu@hyundai.com

ABSTRACT

This technical report describes the proposed system submitted to the DCASE2024 Task3: Sound Event Localization and Detection with Source Distance Estimation. There are two tracks, and we participate in the audio-only track. At first, we adopt the CST block, a transformer-based network, to extract meaningful features for predicting sub-tasks DOA and SED. Next, DOA and EVENT guidance attention blocks are introduced to boost the performance on a Multi-ACCDOA-based single-task system for the SELD tasks. We only apply the data augmentation method, a multi-channel simulation technique to complement the sparsity of training data provided by the challenge. Tested on the dev-test set of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset, our proposed systems outperform the baseline system.

Index Terms— sound event localization and detection with source distance estimation, attention

1. INTRODUCTION

The goal of Sound Event Localization and Detection (SELD) is to detect the occurrence of sound events in three-dimensional space belonging to specific target classes, track their temporal activity, and estimate their directions-of-arrival(DOA) or positions. In complex real-world acoustic environments where multiple sound events overlap in time and space, humans can individually identify and localize multiple sound events, but this is a very difficult task for machines. Effective SELD systems are highly valuable in various fields. The SELD systems in CCTV can perform rescue missions when detecting gunshots, broken glasses, and screaming at crime scenes and localizing them. In smart homes, they can be used for sound scene analysis and audio monitoring, such as baby crying, doorbell detection, and elderly falls. In SELD-enabled service robots, they can detect the user's voices, and surrounding sound events to interact with the user naturally.

The SELD problem consists of a task that identifies both Sound Event Detection (SED) and Direction of Arrival Estimation (DoAE). First introduced in 2019, the DCASE Challenge[1] specified single static sound source situations and used multichannel audio files synthesized by combining mono audio files and impulse response. Subsequent DCASE Challenges[2, 3, 4, 5] evolved into complex environmental configurations, including moving sources, various impulse responses, polyphonic events and overlapping events of the same class, and furthermore, lower SNR, and real spatial sound scenes.

*Thanks to ABC agency for funding.

Various model architectures have been designed for the SELD task. the models comprised two branches followed by a single network to predict SED and DOA separately[6, 7, 8]. Another method of separating the network into SED and DOA modules and combining the outputs was also introduced[9, 10]. Recent approaches have been moving towards creating single-task systems to solve the SELD task[11, 12, 13, 14]. Multi-ACCDOA[12] was proposed to overcome the limitations of ACCDOA to catch polyphonies of the same event class. Focusing on the model structure, many researchers adopted the model based on attention mechanisms. ResNet-Conformer[15, 16, 17, 18] was used for many participants in DCASE challenges and showed outstanding performance. To further enhance the attention mechanisms, CST-Former [8] which uses separate attention for all three regions of time, frequency, and channel was introduced and achieved higher performance than ResNet-Conformer on the DCASE 2022 challenge dataset. Although the performance is improving gradually over previous DCASE challenges, a new change this year introduces distance estimation of the detected events, which makes the task more challenging. With estimating distance together, the Multi-ACCDOA-based single-task method causes overall performance degradation for SED and DOA. For example, the baseline system for the DCASE 2024 challenge, which has the same model architecture as the DCASE 2023 challenge, shows the lower performance of SED and DOA.

Through the fact that the overall prediction performance of the model decreases as subtasks increase, we considered giving additional guidance to the single-task system that predicts subtasks jointly. In light of human auditory cognitive processes for the SELD task, humans can focus preferentially on the direction of sound sources, and then predict sound events through temporal acoustic information about the direction. In this work, motivated by these processes of human cognition, we propose a method for sequentially extracting DOA and SED features and guiding Multi-ACCDOA-based single-task in the correct direction.

2. PROPOSED METHOD

2.1. Features

In this work, motivated by these processes of human cognition, we propose a method for sequentially extracting DOA and SED features and guiding Multi-ACCDOA-based single-task in the correct direction. The STARSS23 dataset provides two recording formats: first-order ambisonics (FOA) and microphone array. In previous works, the FOA format was preferred as input features because it performed better than the MIC format for the SELD task. We use the FOA format as input features. We convert four channels of audio

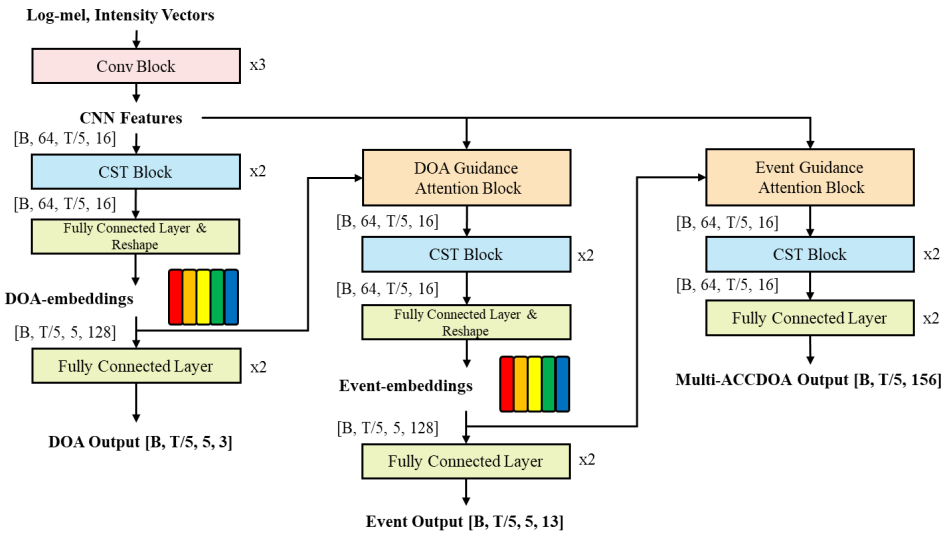

Figure 1: Overall architecture of the proposed network.

data sampled at a frequency of 24kHz into seven channel features containing four log-mel spectrograms and three intensity vectors. For the short-time Fourier transform (STFT), the Hann window with a length of 1024 points(20ms) is used and the length of the shift is 512 points(10ms) to generate a 513-dimensional complex spectral vector. The log-mel spectrograms and intensity vectors are computed with 64-dimensional real-valued vectors.

2.2. Audio Data Augmentation

The official training dataset is the development set and synthetic recordings provided by the DCASE 2024 challenge. However, we analyzed real recordings and found a significant lack of occurrence of specific event classes. Also, To add diversity to the sound sources of the event classes for training the model, we generate the multi-channel audio data. Isolated sound samples are convoluted with spatially room impulse response (SRIRs)[19] using the image method. Samples of male and female speeches are mainly sourced from the VoxCeleb[20] dataset, and the samples of remaining target event classes were extracted from FSD50K[21] and ESC-50[22]. The synthesized audio sample is 1 minute long and contains up to 3 simultaneous sound sources. Eventually, 20 hours of synthetic data is generated and converted from MIC format to FOA format through an ambisonics format converter. We combine these data with the baseline dataset for training. Other augmentation methods such as audio channel swapping[23], time-domain mixing[18], and random cutout[17] are not included in our work.

2.3. Network Architecture

In this study, we adopt the CST block[13] as a transformer-based network containing the local perception unit (LPU), the CST attention layer, and the inverted residual FFN (IRFFN). This method uses distinct attention mechanisms for channel, spectral, and temporal aspects and it has been shown to outperform other top-tier models on the SELD tasks. So, we use the CST block to extract features for several subtasks. The overall network architecture is shown in Fig.1 The architecture is composed of CNN blocks, CST blocks,

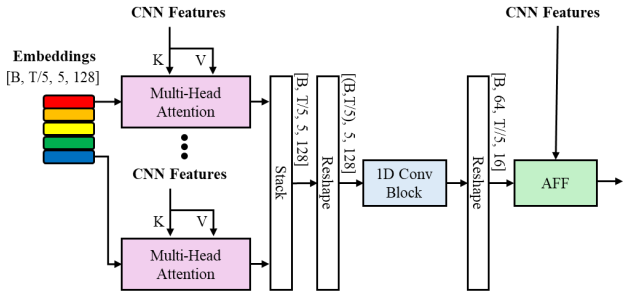

Figure 2: A detailed implementation of the DOA and Event Guidance attention block.

and the guidance(DOA, Event) attention blocks. As can be seen, we first extract the CNN features using convolution blocks. Each of the CNN blocks consists of a 2D convolution, batch normalization, and rectified linear unit same as the challenge baseline CNN blocks. The only difference is that the T-F poolings are applied as (1,1), (1,2), and (5,2) kernels for three convolution blocks.

We design to have two outputs for angle and event estimation in addition to the final Multi-ACCDOA-based single task output. First of all, the model estimates the cartesian coordinates (x,y,z) per frame for the five overlapping events(maximum polyphony). The DOA guidance attention block is then used to allow the model to focus on the corresponding five directions, as shown in Fig.2. We fused the two features using AFF[24]. Next, the model is aimed to focus on sound event detection. The model detects the sound events per frame for the five overlapping events as the second output. Finally, the model predicts the output of Multi-ACCDOA in the same way as CST-Former.

Building upon Permutation-invariant training(PIT)[25], the loss function for DOA and Event output is defined as in (1), where Ψ

denotes the set of all the permutations of overlapping sources, and ψ refers to a permutation. $\mathcal{L}_{\mathrm{DOA}}$ is mean squared error (MSE) loss and $\mathcal{L}_{\mathrm{EVENT}}$ is binary cross entropy loss.

$$\mathcal{L}_{\mathrm{PIT}} = \min_{\psi \in \Psi} \sum_{c=1}^{C} \mathcal{L}_{\psi,\mathrm{DOA}} + \min_{\psi \in \Psi} \sum_{c=1}^{C} \mathcal{L}_{\psi,\mathrm{EVENT}} \quad (1)$$

The final SELD loss is defined as in (2)

$$\mathcal{L}_{\mathrm{SELD}} = \mathcal{L}_{\mathrm{PIT}} + \mathcal{L}_{\mathrm{ACCDOA}} \quad (2)$$

2.4. Network Training

The maximum number of epochs is set to 500 and the batch size is set to 128. The Adam optimizer is used and a tri-stage learning rate scheduler is used with an upper limit of 0.0005. For the first 50 epochs, we warm up the learning rate linearly, hold the learning rate for the 100 updates, and decay the learning rate linearly. In the fine-tuning stage, the saved best result is further trained on real recordings for an extra 50 epochs with a learning rate of 3e-5 and decays linearly. Table 1 shows the final output and model parameters of the submitted system.

Submission	Output Type	Parameter Size
Submission 1-4	Multi-ACCDOA	6M

Table 1: Sumission Configuration.

3. RESULTS ON DEVELOPMENT DATASET

We evaluate our proposed method on the development dataset of STARSS23. The performance of the proposed method is shown in Table 2. 'Baseline-FOA' represents the baseline results presented by the organizers. We submitted a total of four submissions. Our data augmentation method only proceeds with multi-channel audio data generation and other methods are not accessed. Additional performance improvements are expected if we can afford to experiment further using other data augmentation methods used in the SELD tasks.

Model	SELD	macro $\mathbf{F_{20^\circ/1}}$	DOAE	RDE
Baseline-FOA	-	13.1%	36.9°	33%
Sub1	0.35	33.9%	19.5°	28%
Sub2	0.35	34.7%	18.8°	28%
Sub3	0.35	35.0%	19.0°	29%
Sub4	0.35	35.1%	18.9°	29%

Table 2: Experimental results of the proposed SELD systems evaluated by joint metrics and development dataset. The metrics of SELD, macro $\mathrm{F}_{20^\circ/1}$, DOAE, RDE are SELD metric, macro-averaging of the location-dependent F1-score, class-dependent DOA error, and class-dependent relative distance error respectively.

4. CONCLUSION

In this paper, we present the proposed system to solve the SELD task in DCASE 2024 challenge task 3 (audio-only task). The additional task for distance estimation of detected events is included this year. Therefore, building a well-generalized system for SELD tasks is more challenging. We propose DOA and EVENT guidance methods to improve the model generalization by conceiving the human audio-cognitive process. Before the single SELD output, the model goes through the process of estimating the directions of arrival and predicting events for corresponding directions, sequentially. In this way, we aim to reduce complexity by giving additional guidance on DOA and SED. We also employ multi-channel simulation techniques to improve the generalization of the system. The experimental results show that our method outperformed the baseline for the DCASE 2024 challenge task 3.

5. REFERENCES

[1] http://dcase.community/challenge2019/.

[2] http://dcase.community/challenge2020/.

[3] http://dcase.community/challenge2021/.

[4] http://dcase.community/challenge2022/.

[5] http://dcase.community/challenge2023/.

[6] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[7] S. Park, S. Suh, and Y. Jeong, "Sound event localization and detection with various loss functions," in *Proceedings of the Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 2–4.

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

[9] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking." in *DCASE*, 2020, pp. 120–124.

[10] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "Sound event localization and detection for real spatial sound scenes: Event-independent network and data augmentation chains," *arXiv preprint arXiv:2209.01802*, 2022.

[11] 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)*. IEEE, 2021, pp. 915–919.

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

[13] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8686–8690.

[14] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng, Y. Wang, L. Sun, Y. Fang, J. Pan, *et al.*, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," *DCASE2022 Challenge, Tech. Rep.*, 2022.

[16] S.-I. Kang, K. Cho, M. Keum, and Y. Park, "The distillation system for sound event localization and detection of dcase2023 challenge," DCASE2023 Challenge, Tech. Rep., June 2023.

[17] J. Hu, Y. Cao, M. Wu, F. Yang, W. Wang, M. D. Plumbley, and J. Yang, "Jless submission to dcase2023 task3: Conformer with data augmentation for sound event localization and detection in real space," DCASE2023 Challenge, Tech. Rep., June 2023.

[18] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] A. Politis, S. Adavanne, and T. Virtanen, "Tau spatial room impulse response database (tau-srir db)," 2022.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[21] 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.

[22] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[23] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[24] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560–3569.

[25] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.