DATA AUGMENTATION, NEURAL NETWORKS, AND ENSEMBLE METHODS FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Gwantae Kim, Hanseok Ko

Korea University
Department of Electrical Engineering
Seoul, South Korea

ABSTRACT

This technical report describes the system participating in the DCASE 2023, Task3: Sound event localization and detection evaluated in real spatial sound scenes challenge. The system contains data augmentation strategies, neural network models, and ensemble methods. For track A, we adopt rotation and Specmix data augmentation strategies to increase the amount of data samples and improve robustness. The neural network model, which is based on baseline networks, consists of residual convolution neural networks with spatial attention, recurrent neural networks, and multi-head self-attention. Moreover, we propose several ensemble methods, such as windowing, weight averaging, and clustering-based output selection. For track B, we extend the audio-only baseline model to the audio-visual model with 3D convolution layers using raw video, optical flow, and object detection features. Through a series of relevant experiments, the proposed methods achieve competitive results compared to the baseline and state-of-the-art methods.

Index Terms— DCASE2023, data augmentation, attention, ensemble, sound event localization and detection

1. INTRODUCTION

Human can distinguish and localize different sounds coming from several directions. By mimicking this behavior, various approaches for sound event localization and detection(SELD) tasks are introduced and actively explored. SELD is a crucial part of the many applications, including human-computer interaction, robot audition, and scene understanding.

SELD is the task that identifying both the direction of arrival(DOA) and event class from sounds so that machines can have the same capabilities. Since 2019, numerous methods have been proposed to solve problems of the SELD through the DCASE challenge [1–6]. In the challenge, methods focus on various aspects, such as preprocessing, data augmentation, model structure, output format, loss function, post-processing, and ensemble. Mazzon et al. [1] proposed spatial data augmentation method for first-order ambisonic(FOA) domain data. Adavanne et al. [5] proposed a end-to-end convolutional recurrent neural networks(CRNN), named SELDnet, to solve polyphonic SELD problems. It was utilized for a joint task of sound event detection and regression-based DOA estimation. Shimada et al. [7, 8] presented activity-coupled cartesian direction of arrival(ACCDOA) representation, which is suitable to represent overlapped sound events. Although the performance is improved gradually with these researches, it is ambiguous which method makes better results because many papers conducted experiments with different settings,

In this study, we propose several methods for SELD on data augmentation, neural networks, and ensemble aspects. With the series of experiments, we found some data augmentation solutions, such as rotation, and Specmix [9]. We also propose the sound event localization and detection model based on convolution neural networks, recurrent neural networks, and attention. Finally, we present rotation, windowing, and clustering-based ensemble and weight-averaging ensemble methods.

2. TRACK A

The First-order ambisonic(FOA) recordings are used as input signals of the proposed model. The FOA recordings are transformed into the log-mel spectrogram and its intensity vectors. The output labels are converted into the ACCDOA and multi-ACCDOA formats. Two data samples are merged by Specmix mixed sample data augmentation policy.

2.1. Features

The STARSS22 dataset provides 2 recording formats: FOA and microphone array. We used 4-channels 24kHz FOA recording format. First we extract two time-frequency domain features:multichannel log-mel Spectrogram, and Intensity Vector(IV). Every features are computed from the Short-

Time Fourier Transform(STFT). The hop length, window length, nfft, and mel-coefficients of the features are 20ms, 40ms, 1024, and 128, respectively. Let log mel-spectrogram coefficients are $x_{t,f}$, where t, f denote the time frame and the frequency bin. For the FOA format, the intensity vector [10] are

$$I_{t,f,c} = \begin{bmatrix} \Re\{w_{t,f} * x_{t,f}\} \\ \Re\{w_{t,f} * y_{t,f}\} \\ \Re\{w_{t,f} * z_{t,f}\} \end{bmatrix} \quad (1)$$

where w, x, y, and z are ambisonic channels, and subscript c denotes the channel index. Finally, we can extract 7 channels of time-frequency features, which consist of 4 log-mel spectrograms and 3 IVs.

2.2. Data augmentation

Rotation We applied rotation [1] data augmentation to STARSS22 data and DCASE 2022 simulated data. We selected 8 directions that are the same settings of [11].

Specmix We used a mixed sample data augmentation strategy, named Specmix [9], to promote the generalization of the model. The goal of Specmix is to generate a new training sample $(\tilde{x}, \tilde{y})$ by combining two training samples (x_A, y_A) and (x_B, y_B). The combining operation is

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \quad (2)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (3)$$

where $\mathbf{M} \in \{0, 1\}^{F \times T}$ denotes a binary mask indicating where to drop out and fill in from two images, $\mathbf{1}$ is a binary mask filled with ones, and $\odot$ is element-wise multiplication. The combination ratio λ between two data points is the number of pixels of x_A in $\tilde{x}$. For the classification task, which is analyzed in Specmix [9] paper, λ is calculated on the whole spectrogram. In contrast, λ is calculated on each time bin because the outputs are predicted along the time axis. Specmix has frequency masking and time masking. The number of frequency mask f_{times} is 3 and the width of each frequency mask γ is 0.1. The number of time mask t_{times} is 3 and the width of each time mask γ is 0.1.

2.3. Network architecture

Fig. 1 (a) illustrates the overall architecture of our system for track A. The networks consist of SE-ResNet, RNN, MHSA, and Linear layers. The SE-ResNet contains stacks of convolution layer and squeeze-excitation layer [12] with the residual connection. The RNN, MHSA, and linear layers are the same as the challenge baseline.

2.4. Ensemble

We propose a windowing-clustering ensemble, a rotation-clustering ensemble, and a model weight-averaging ensemble. For the windowing-clustering ensemble, a window slides along the time axis in the inference time. The window length is 5s, which is the same as the model input length, and the hop length is 100ms. After the windowing, each timestamp of the label has 50 candidates. We use DBSCAN [13] to find outliers in the candidates. The candidates except outliers are averaged. For the rotation-clustering ensemble, we estimate the candidates with an 8-way rotation augmentation and remove outliers with DBSCAN clustering. The output may not rely on a single estimation thanks to these two ensemble methods.

Since the model is sensitive to recent training samples, we save the 10 best model weights using the evaluation dataset during the training phase and averaging them. This model weight-averaging ensemble improves robustness and overall performance.

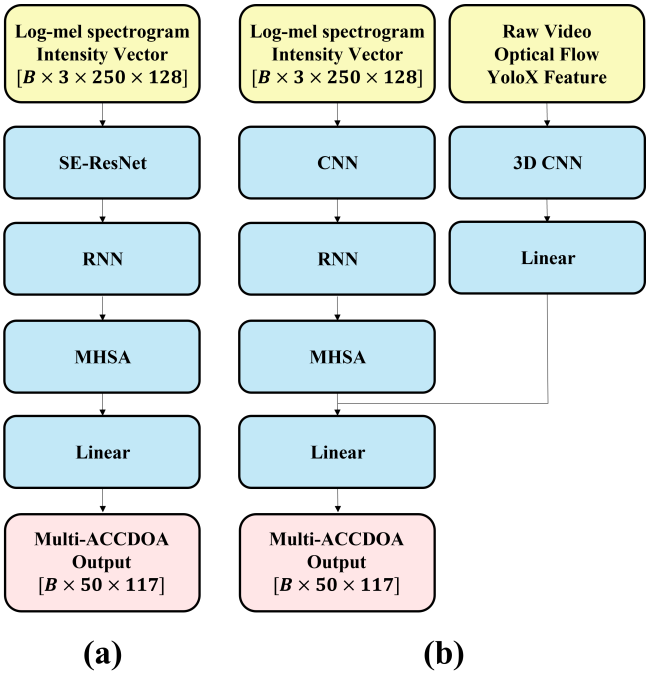

Figure 1: The architecture of the proposed model. (a) Track A (b) Track B

3. TRACK B

Fig. 1 (b) illustrates the architecture of our system for track B. The audio networks are the same as SELDnet, and the visual networks consist of 3D convolution layers and linear layers. The visual features are raw video, optical flow, and YoloX [14] object detection features. Each feature pass through 3D convolution layers and linear layers and add to the MHSA output of the audio networks.

Table 1: Submission configuration.

	Configuration	Params.
Track A 1	Multi-ACCDOA	69M
Track A 2	ACCDOA	69M
Track A 3	Ensemble	138M
Track A 4	Ensemble 2	138M
Track B 1	Multi-ACCDOA	7M

4. EXPERIMENTS

4.1. Dataset

We used STARSS22 [15] Dataset, which contains 3 hours recordings of real scenes. Despite its recording environments being well-formulated, the amount of data is too small to train neural networks robustly. To increase the amount of data while maintaining quality, we used synthetic datasets and several data augmentation methods. We used DCASE 2022 simulated data[1], which contains 20 hours recordings. In addition, recorded sound samples selected from FSD50K [16] dataset and convoluted with SRIRs from TAU-SRIR DB[2] to generate 4000 1-minute long multi-channel scene recordings with a maximum polyphony of 3.

4.2. Experimental setup

We use the Adam [17] optimizer with a learning rate from 5e-3 to 1e-5. The batch size is 32. Table 1 shows the setups of our submitted systems. Submission #1 is multi-accdoa output format, and Submission #2 is accdoa output format with 3 ensemble methods. Submission #3 and #4 integrated candidates of Submission #1 and Submission #2 when the rotation-clustering ensemble is performing.

4.3. Results

We evaluate our proposed method on the development dataset of STARSS22 [15]. The experiment results are summarized in Table 2. As shown in the table, each proposed single model and ensemble model outperforms the baseline systems by a large margin.

5. CONCLUSION

This report proposes an ensemble system to solve the SELD task in DCASE 2023 challenge task 3. We focus on data augmentation and ensemble methods. We adopt data augmentation approaches to expand the official dataset and synthetic dataset. Several model ensemble methods are used to get a more robust SELD estimation. The neural networks are trained to acquire candidates in ACCDOA and multi-ACCDOA representation formats. For the audio-visual SELD task, we additionally used raw video, optical flow, and object detection features, which pass through 3D convolution layers. The experimental results show that the proposed method outperforms the baseline systems by a large gap.

[1] https://zenodo.org/record/6406873

[2] https://zenodo.org/record/6408611

Table 2: Comparison of baselines and submissions with the development set (Track A).

	ER	F1	LE	LR
Baseline	0.57	29.9	22.0	47.7
Sub1	0.47	52.7	15.2	68.8
Sub2	0.49	51.1	15.5	69.7
Sub3	0.47	52.9	15.0	69.3
Sub4	0.47	51.7	15.2	70.2

Table 3: Comparison of baselines and submissions with the development set (Track B).

	ER	F1	LE	LR
Baseline	1.07	14.3	48.0	35.5
Sub1	**0.52**	**45.1**	**17.8**	**59.9**

6. ACKNOWLEDGMENT

7. REFERENCES

[1] 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.

[2] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 119.

[3] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019, pp. 10–14.

[4] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane." in *DCASE*, 2017, pp. 93–102.

[5] 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, 2018.

[6] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," DCASE2017 Challenge, Tech. Rep., September 2017.

[7] 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.

[8] 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.

[9] G. Kim, D. K. Han, and H. Ko, "Specmix: A mixed sample data augmentation method for training withtime-frequency domain features," *arXiv preprint arXiv:2108.03020*, 2021.

[10] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 241–245.

[11] 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.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[14] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[15] 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," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings

[16] 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.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.