Polyphonic Sound Event Localization and Detection Using Convolutional Neural Networks and Self-Attention with Synthetic and Real Data

Technical Report

*Yeongseo Shin, Kangmin Kim, Chanjun Chun**

Chosun University
Department of Computer Engineering, Pilmun-daero, Dong-gu, Gwangju, Korea
{ys070400, 20164284}@chosun.kr, cjchun@chosun.ac.kr

ABSTRACT

This technical report describes the system submitted to DCASE 2022 Task 3: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. The goal of Task 3 is to detect the occurrence of sound events belonging to a specific target class in a real spatial sound scene, track temporal activity, and estimate the direction or location of arrival. In a given dataset, synthetic and real data exist together, and only a very small amount of real data exists compared with synthetic data. In this study, we developed a method utilizing a multi-generator and another applying SpecAugment as a data augmentation method to address the problem of imbalance in the amount of data. In addition, in our network architecture, the Transformer encoder was applied to the Convolutional Recurrent Neural Network (CRNN) structure that is mainly used in SELD. In addition, as a result of training with a single model and applying an ensemble, it was confirmed that the performance improved compared to the baseline system.

Index Terms— sound event localization and detection, CRNN, deep learning, Transformer, model ensemble

1. INTRODUCTION

Sound event localization and detection (SELD) involves identifying a sound class and estimating the onset, offset, and direction of arrival (DOA) of the corresponding sound event. In DCASE, several solutions have been proposed based on SELD (polyphonic SELD using only fixed sound sources, the introduction of moving sound sources, and the introduction of unknown directional interferences) in various environments [1, 2, 3, 4]. In 2022, the SELD task was performed using real spatial sound scenes. Because SELD is performed in real space sound scenes, a small amount of real data recorded in the real space and several synth data for training are provided [5].

Fundamental problems must be addressed to complete the SELD task, including sound event detection (SED) [6, 7] and sound source localization (SSL) [8, 9]. In this study, to solve this problem, we proceeded based on the SELDNet model proposed in [10, 11]. In particular, in the aforementioned study, a convolutional recurrent neural network (CRNN) was proposed for the SELD of multiple overlapping sound events in the three-dimensional space. The experiment was conducted in the first-order Ambisonic (FOA) format and multichannel log-mel spectrograms, and intensity vectors were used as the input features.

In addition, a Transformer encoder was used to train temporal context information. The output of the Transformer encoder is fed to a fully connected block. Multi-Activity Combined Cartesian Direction of Arrival (Multi-ACCDOA) in the output format [10]. Finally, to maximize the performance of the system, several SELD models trained with slightly different structures and conditions were combined into an ensemble.

The remainder of this paper is organized as follows. In Section 2, the proposed method is introduced. A comparison of the results obtained using the experimental setup and development data is described in Section 3. Finally, Section 4 closes the paper.

2. PROPOSED METHOD

2.1. Data generation

The dataset provided by DCASE consists of 1200 synthetic datasets and spatial recordings of real scenes with 121 spatiotemporal annotations [5]. The amount of training data was increased to improve the performance of the model. Using the provided external data and spatial room impulse response (SRIR), 5100 synthetic data were used for training. Synthetic data are FOA format data consisting of 13 classes ('Female Speech', 'Male Speech', 'Clapping', 'Telephone', 'Laughter', 'Domestic Sounds', 'Footsteps', 'Door Cupboard', 'Music', 'Music Instrument', 'Water Tap', 'Bell', 'Knock'). As hyperparameters for data generation, the number of maximum polyphony was set to 2, the duration was set to 60 s, and the sampling rate was set to 44.1 kHz. In addition, the signal-to-noise ratio (SNR) was set to have a value of 6 to 31 at random, the speed of the event was set to have values of 10.0, 20.0, and 40.0, and there was no directional interference. 'Female Speech', 'Male Speech', 'Clapping', 'Telephone', 'Laughter', and 'Domestic Sounds' were set to random values as to whether the source event was dynamic, 'Footsteps' was set to be dynamic. The other classes were set as static

2.2. Input feature extraction

We used the same approach as the baseline to obtain the features [5]. The FOA format was used in this study because it performed better at the baseline than the MIC format. Seven channels are present in first-order Ambisonic (FOA) audio files: four log-mel

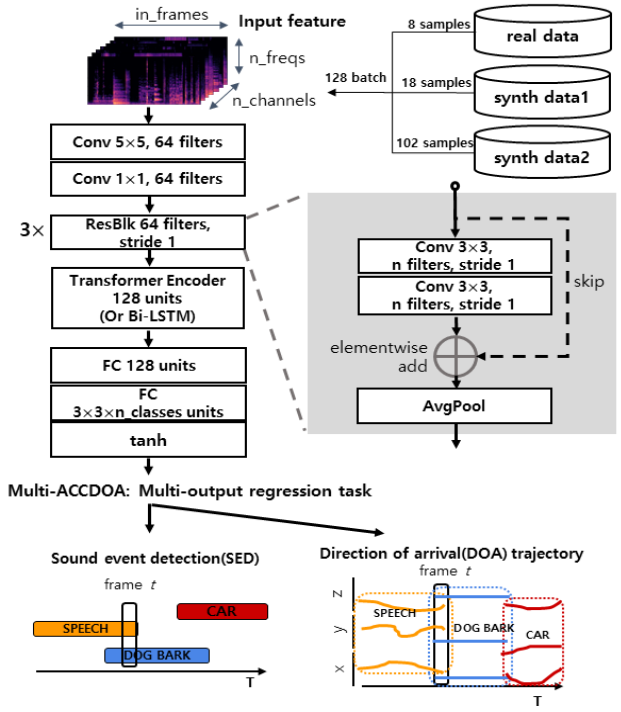

Figure 1: Overall architecture of the proposed network

spectrograms and three intensity vectors. The same settings as the baseline (64 mel bands, 40 ms window and 20 ms hop length at 24kHz) were used to extract the features. We followed the same approach for normalization.

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 in a single branch [10]. To improve the performance of the model, various attempts have been made to change RNN layer to Bi-GRU [12] or Bi-LSTM [13]. The model was modified after several attempts. Figure 1 shows the overall structure of the proposed model. If each batch receives data at a certain rate using a multi-generator, the shape of the input data was set to (128, 250, 64, 7). (128, 250, 64, 7) represent (b, t, fq, ch), respectively. When the input data arrives, it first passes through the stem CNN layer and changes according to the input shape of the Residual Block [14]. It then passes through the Residual Block on the 3rd floor; the residual block passes through the 2nd convolutional layer, and finally, the Pooling layer, following which Dropout is applied. Subsequently, the input data passes through the FC layer through the Transformer encoder [15] and Hyper tangent is applied as an activation function to generate a Multi-ACCDOA- type output [10].

2.4. Training method

The training dataset consists of 121 real scene recorded datasets (real data) provided by DCASE, 1200 synthetic datasets (synth data1), and synthetic datasets (synth data2). At this time, an imbalance in the amount of real data and synth data 1 and 2 appears.

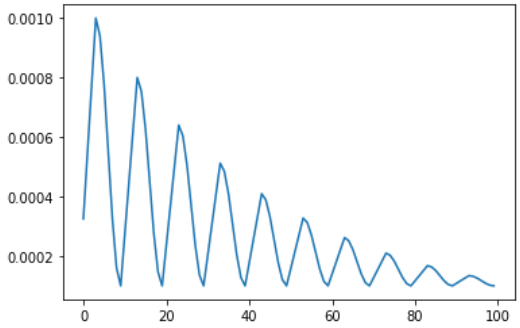

Figure 2: After 30 epochs variation of the learning rate per epoch

To address this problem, this study used a multi-generator to set the ratio of data for each batch, and training was carried out including real data for every batch.

The amount of training data was increased by synthesizing data, and real data were used for each batch, however, the data augmentation method must be applied because the data are still limited. SpecAugment [16], an augmentation technique widely used in speech recognition, was used as a data augmentation technique, in which two types of masking were used: frequency masking and time masking. Masking was applied to all the channels of the training data (real data, synth data1, synth data2).

In addition, training was performed by setting various training parameters. As the optimizer, we changed the performance from Adam [17] to Nesterov Momentum Adam (NAdam) [18] and compared the performance trends. The dropout value was set to 0.2, the learning rates set to 20^{-3} and10^{-3}, and the results were compared. In addition, after training by reducing the learning rate at regular epochs up to 30 epochs using the scheduler, a method of changing the learning rate according to the period of the cosine periodic function was used. Figure 2 show the change in the learning rate after 30 epochs. The cycle was repeated every 10 epochs, with the largest learning rate value at the 4th epoch of the cycle, which then slowly decreases to 10^{-4}.

Table 1: Test results for a single model: using a development dataset

Model	ER	F	LE	LR	SELD
model 1	0.67	0.33	24.63	00.61	0.47
model 2	0.69	0.30	24.03	0.58	0.48
model 3	0.65	0.31	24.81	0.59	0.47
model 4	0.66	0.30	25.72	0.62	0.47
model 5	0.65	0.31	22.24	0.53	0.48
model 6	0.65	0.30	30.01	0.59	0.49
model 7	0.65	0.31	26.41	0.62	0.47

3. EXPERIMENTS

Several models were trained using two types of models, and the ensemble [19] was applied by selecting some models. The difference between the two models is whether the RNN layer is a Bi-LSTM or Transformer. Table 1 presents the results for each model. Several models were trained using two types of models, and then ensemble training was applied by selecting a few values. The difference between the two models lies in whether the RNN layer is

a Bi-LSTM or Transformer. After training, seven single and four ensemble models were selected. The results of each model are listed in Table 1, and the ensemble results are listed in Table 2.

Table 2: Test results for an ensemble model: using a development dataset

Model	ER	F	LE	LR	SELD
1+2+3+4+5+6+7	0.59	0.35	20.68	0.57	0.45
1 +3+4+5+6+7	0.59	0.34	24.75	0.58	0.45
1+2 +4+5+6+7	0.59	0.35	33.75	0.57	0.46
1+2+3+4 +6+7	0.59	0.34	23.00	0.59	0.44

4. CONCLUSION

In this technical report, polyphonic sound event localization and detection method using convolutional neural networks and self-attention was proposed, and the performance of this method was evaluated in real spatial sound scenes. For training such network, the real and synthetic data were employed. Because the imbalance between synthetic and real data was very severe, we utilized a multi-generator for synthetic and real data. In addition, the proposed model using a residual block and transformer encoder was employed. As a result of applying an ensemble, an improvement in performance was confirmed as compared to the baseline model.

5. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2019R1C1C101159).

6. REFERENCES

[1] S. Kapka and L. Mateusz, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 119-123, Aug. 2019.

[2] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and L. Chin-Hui, "The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge." in *Proc. IEEE Audio and Acoustic Signal Processing (AASP)*, Jul. 2020.

[3] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA-and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection," DCASE2021 Challenge, technical report, Nov. 2021.

[4] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2021 Task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," in *Proc. IEEE Audio and Acoustic Signal Processing (AASP)*, Nov. 2021.

[5] 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," *arXiv preprint arXiv:2206.01948*, Jun. 2022.

[6] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 286-290, May 2020.

[7] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 326-330, May 2020.

[8] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136-140, Oct. 2017.

[9] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131, Nov. 2018.

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

[11] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716-720, May 2022.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Neural Information Processing Systems (NIPS) 2014 Workshop on Deep Learning*, Dec. 2014.

[13] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, Dec. 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE conference on computer vision and pattern recognition*, pp. 770-778, Jun. 2016.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 6000-6010, Dec. 2017.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613-2617, Sep. 2019.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, May 2015.

[18] T. Dozat, "Incorporating nesterov momentum into Adam," in *Proc. International Conference on Learning Representations (ICLR)*, pp. 1-4, Feb. 2016.

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. International workshop on multiple classifier systems*, vol. 1857, pp. 1-15, Springer, Dec. 2000.