CONVNEXT AND CONFORMER 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 to the DCASE 2022, Task3 : Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes challenge. The system consists of a convolution neural networks and multi-head self attention mechanism. The convolution neural networks consist of depth-wise convolution and point-wise convolution layers like the ConvNeXt block. The structure with the multi-head self attention mechanism is based on Conformer model, which contains combination of the convolution layer and the multi-head self attention mechanism. In the training phase, some regularization methods, such as Specmix, Droppath, and Dropout, are used to improve generalization performance. Multi-ACCDOA, which is output format for the sound event localization and detection task, is used to represent more suitable output format for the task. Our systems demonstrate a improvement over the baseline system.

Index Terms— ConvNeXt, Conformer, DCASE2022, Sound event localization and detection, Specmix

1. INTRODUCTION

Humans can distinguish and localize different sounds coming from various directions. Sound Event Localization and Detection(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. The task is a crucial part of the many applications, including human-computer interaction, robot audition, and scene understanding.

Since 2019, numerous methods have been solving problems of the SELD through the DCASE challenge [1–7]. [6] proposed the Convolutional Recurrent Neural Networks(CRNN) model with a new training procedure. [7] proposed multi-ACCDOA output format with RD3Net based neural networks structure.

Recently, many new convolution layers and attention layers are proposed to solve various problems. ConvNeXt [8] is presented to solve image classification. It consists of depthwise convolution and pointwise convolution layers. Conformer [9] is a kind of self attention mechanism based structure, which is presented to solve the speech recognition problem. We adopt the ConvNeXt block and Conformer block to solve our task, SELD.

In this study, we proposed the sound event localization and detection model based on ConvNeXt [8] and Conformer [9]. First, The mixtures of signals are transformed into time frequency domain features, which are log-mel spectrogram(log-mel), Intensity Vector(IV), and Inter-Phase Difference(IPD). Second, We applied Specmix [10] to improve generalization performance. The augmented features passed through the proposed neural networks, and predicted label. We used multi-ACCDOA [7] as output format. To evaluate the novelty of the our model, we trained and tested the model with the DCASE 2022 Task 3 dataset, named STARSS22 [11]. To deliver the detailed description, the remainder of the paper is organized as follows. The proposed SELD model are described in Section 2 The experimental process and results are presented in Section 3. Conclusions are drawn in Section 4.

2. PROPOSED MODEL

The First-Order Ambisonic (FOA) recordings are used as input signals of the proposed model. First, the FOA recordings are transformed to the three frequency domain features, and the output labels are converted into the nulti-ACCDOA [7] format. Second, the input features are converted into the mixed sample following Specmix [10] mixed sample data augmentation policy. The generated input-output pair is used to train the proposed SELD model. Since the recordings are only given in the test phase, we only converts the recordings into frequency domain features, put the features into the model. The output of the model is converted into original output format, and calculate the scores.

2.1. Features

Multichannel log-mel Spectrogram, Intensity Vector(IV), and inter-channel phase differences (IPDs) are used as frame-wise features. Every features are computed from the Short-Time Fourier Transform(STFT) coefficients $x_{t,f}$, where t, f denote the time frame and the frequency bin. In the FOA format, the intensity vector [12] are

$$I_{t,f,p} = \begin{bmatrix} \Re\{w_{t,f} * x_{t,f}\} \\ \Re\{w_{t,f} * y_{t,f}\} \\ \Re\{w_{t,f} * z_{t,f}\} \end{bmatrix}$$

where w, x, y, z are Ambisonic channel. The IPDs are computed by $IPD_{t,f,p,q} = \angle x_{t,f,p} - \angle x_{t,f,q}$, where p, q denote the Ambisonic channel pairs. We fix $p = 0$ to compute relative IPDs between all other channels, $q \neq 0$. We use the both cosIPDs and sinIPDs as IPD features. Since the input consists of four Ambisonic channel signals, we can extract four log-mel spectrogram, three IVs, and six IPDs.

The authors of Korea University were supported by Deep Machine Lab (Q2128981). Corresponding Author:Hanseok Ko.

2.2. Data augmentation

We used mixed sample data augmentation strategy, named Specmix [10], 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

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

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (2)$$

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 $\tilde{x}$. Specmix has frequency masking and time masking but we used frequency masking only. The number of frequency mask f_{times} is 3 and the width of each frequency mask γ is 0.1.

2.3. Network architecture

Fig. 1 and Fig.2 illustrate the overall architecture of our system. The input features have corresponding ConvNeXt-SELD module and they pass through their ConvNeXt-SELD module. The details of the ConvNeXt-SELD module are described in Fig. 2(a) and Fig. 2(c). One ConvNeXt-SELD block has depthwise convolution layer with C channels, batch normalization, pointwise convolution layer with 4C channels, GELU activation function, and pointwise convolution layer with C channels. It has residual connection with residual scale factor 0.1 and droppath trick [8]. The ConvNeXt-SELD block is stacked N times, where N is iterator number of the Fig. 2(c) block. In our settings, the ConvNeXt-SELD module has 4 sub-blocks, then first block has one ConvNeXt-SELD block, second block has two ConvNeXt-SELD block, third block has three ConvNeXt-SELD block, and fourth block has four ConvNeXt-SELD block.

The outputs of the each ConvNeXt-SELD module are concatenated along channel axis and reshape to [Batch, Time, Feature] shape. The reshaped feature now pass through Conformer-SELD module, which is described in Fig. 2(b) and Fig. 2(d). The two Conformer-SELD blocks are used to build COnformer-SELD module. Each Conformer-SELD block has fully-connected layer, residual connection, Multi-Head Self Attention(MHSA), residual connection, convolution layer, residual connection, fully-connected layer, residual connection, and layer normalization. The model structure, activation function, and tricks are same as Conformer settings [9]. We tried to increase the number of blocks, but it makes the model too heavy.

2.4. Hyper-parameters

The sampling frequency is set to 24kHz. The STFT is used with a 20ms frame length and 10 ms frame hop. The nfft of input to the networks is 512 frames. The frame shift length is set to 100ms during the inference. We use a batch size of 128. The learning rate is decreased from 1e-3 to 1e-4. We use the Adam [13] and PAdam [14] optimizer. We validate and save model weights every 1 epochs and select the model with best validation accuracy.

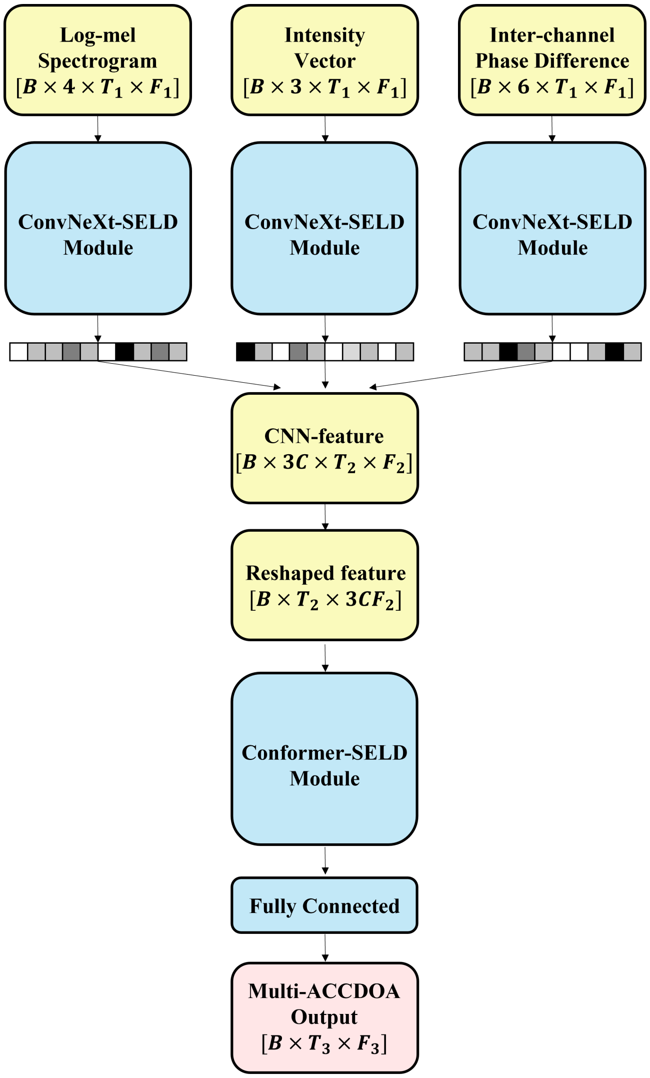
Log-mel Spectrogram $[B \times 4 \times T_1 \times F_1]$
Intensity Vector $[B \times 3 \times T_1 \times F_1]$
Inter-channel Phase Difference $[B \times 6 \times T_1 \times F_1]$
ConvNeXt-SELD Module
ConvNeXt-SELD Module
ConvNeXt-SELD Module
CNN-feature $[B \times 3C \times T_2 \times F_2]$
Reshaped feature $[B \times T_2 \times 3CF_2]$
Conformer-SELD Module
Fully Connected
Multi-ACCDOA Output $[B \times T_3 \times F_3]$

Figure 1: The architecture of the proposed model.

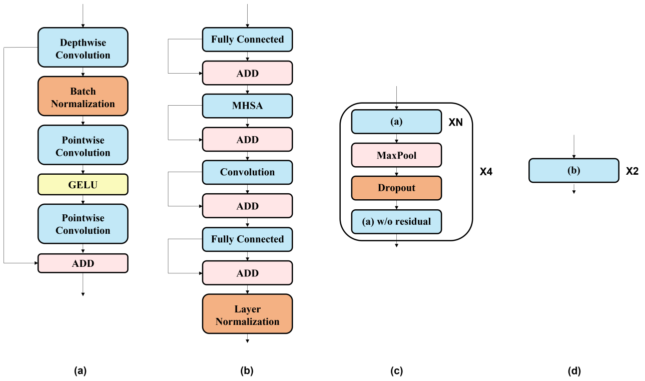

Figure 2: The detailed architecture of the proposed modules. (a) ConvNeXt-SELD Block. (b) Conformer Block. (c) ConvNeXt-SELD Module. (d) Conformer-SELD Module.

Table 1: Submission configuration.

	Configuration
Submission 1	Adam optimizer w/o Specmix
Submission 2	Adam optimizer w Specmix
Submission 3	PAdam optimizer w Specmix

Table 2: Comparison of baselines and submissions with the development set.

	ER	F1	LE	LR
[6]	0.67	0.32	26.20	0.52
[7]	0.71	0.36	29.30	**0.46**
Sub1	0.66	0.31	21.68	0.51
Sub2	0.66	**0.30**	22.51	0.49
Sub3	**0.65**	0.33	**20.39**	0.51

3. EXPERIMENTS

3.1. Experimental Settings

We evaluated our approach on the development set of STARSS22 dataset - Ambisonic settings. The baseline was an multi ACCDOA-based system with a CRNN [6] and RD3Net [15]. In the setup, four metrics were used for the evaluation [16]. The metrics are basd on True Positive(TP) and False Positives(FP) determined not only by correct or wrong detection, but also based on if they are closer or further than a distance threshold T = 20 degrees from the reference. The challenge form the location-dependent F1-score(F1), Error Rate(ER), a class-dependent localization error per class LE_{CD}, and a localization recall per class LR_{CD}. Table 1 describes configurations of the submissions. Submission 1 used Adam optimizer and did not use Specmix data augmentation, Submission 2 used Adam optimizer and Specmix, and Submission 3 used PAdam optimizer and Specmix in the training phase. Other settings, such as networks structure and learning rate, are same.

3.2. Experimental results

Table 2 shows the performance with the development set for baselines and our system. As shown in the table, the ER, F1, and LE are improved with our system compared to the baseline. However, there were no significant improvement between baseline and our system although the number of parameters are increased(1M vs. 100M). We tried to solve the problem with another training procedure, but the loss was not converged with our training procedure on the same model. We think that not only the model structures but also training procedure are important to achieve good results.

4. CONCLUSION

We presented our approach to DCASE2022, Task3: Sound Event Localization and Detection Evaluated in Real Spatial SOund Scenes challenge. Our system use the ConvNeXt and Conformer structure to improve SELD performance. We also used Specmix, PAdam, and multi-ACCDOA tricks to achieve good results. As a result, our systems performed slightly better than baseline system. In the future, we build a new training procedure to bring out the best potential of the proposed model.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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," DCASE2020 Challenge, Tech. Rep., July 2020.

[4] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," DCASE2020 Challenge, Tech. Rep., June 2020.

[5] 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, Tech. Rep., November 2021.

[6] 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.

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

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

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

[11] 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," 2022. [Online]. Available: https://arxiv.org/abs/2206.01948

[12] 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.

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

[14] J. Chen, D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu, "Closing the generalization gap of adaptive gradient methods in training deep neural networks," *arXiv preprint arXiv:1806.06763*, 2018.

[15] 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.

[16] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.