# DATA AUGMENTATION AND CLASS-BASED ENSEMBLED CNN-CONFORMER NETWORKS FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

*Yuxuan Zhang, Shuo Wang, Zihao Li*,*
*Kejian Guo, Shijin Chen, Yan Pang,*

University of Chinese Academy of Sciences, Beijing 100049, China
* Corresponding email: lizihao183@mails.ucas.edu.cn

## ABSTRACT

In this technical report, we describe the system participating in the DCASE 2021, Task 3: Sound Event Localization and Detection (SELD) challenge. We introduce Conformer block into the baseline system to make better use of temporal context information for the SELD task. To expand the official training dataset, we use Audio Channel Swapping (ACS), Speed Perturbation (SP), and Time-Frequency Masking (TFM) as augmentation techniques. In addition, we proposed a class-based ensemble method to attain a more robust sound event detection (SED) and sound source localization (SSL) estimation result for each sound event. After evaluating our best-proposed system on DCASE 2021 Challenge Task 3 Development Dataset, we approximately achieve $44\%$ and $37\%$ relative improvements on the SELD scores, respectively.

***Index Terms*—** Sound event localization and detection, data augmentation, Conformer, model ensemble

## 1. INTRODUCTION

Sound event localization and detection (SELD) task aims to detect individual sound events of specific classes and estimate their locations simultaneously [1]. SELD can be seen as a multitask learning task including sound event detection (SED) and sound source localization (SSL). Similar to other DCASE tasks, neural-network (NN)-based methods exhibit significant progress in the SELD research area.

The baseline system in 2020 and a series of similar state-of-the-art systems consist of a High-level Feature Representation module and a Temporal Context Representation module which is followed by two parallel branches contain two fully connected (FC) layers, each performing individual SED and SSL subtasks [1, 2, 3]. The main difference between these systems is that the above two modules are replaced by different NN architectures. In the SELD task of DCASE 2021, the baseline system (called SELDnet) is no longer a multitask learning system since it eliminates event classification output branch by setting training target as activity-coupled Cartesian DOA (ACCDOA) representation [4]. This modification improves SELD performance and reduces the network size in the meanwhile.

The Conformer, firstly proposed in [5], has achieved excellent results in many automatic speech recognition (ASR) competitions. [6] combined Conformer and ResNet/Xception, and obtained the state-of-the-art SELD performance. In this report, we introduce the Conformer module into SELDnet firstly since its better capability of modeling the temporal dependencies than recurrent neural network (RNN). To overcome the lack and unbalance of training data, we utilize several data augmentation approaches, including Audio Channel Swapping (ACS) [6], Speed Perturbation (SP) [7] and Time-Frequency Masking (TFM) [8]. Finally, by assuming different architectures exhibit different abilities in localization and detection for each sound event, we proposed a class-based ensemble method to attain more accurate SED and SSL results for each event.

## 2. PROPOSED METHOD

### 2.1. Features

Task 3 provides two types of 4-channel spatial sound format: First-Order of Ambisonics (FOA) and tetrahedral microphone array (MIC) [9] while the sound data was recorded with a 24 kHz sampling frequency. In this report, both datasets, say FOA and MIC, are utilized for training our model. Firstly, we extract 64 log-Mel magnitude spectrogram for each audio file using short-term Fourier transform (STFT) with the configuration of 40 ms frame length and 20 ms hop length. Then GCC with phase transform (GCC-PHAT) feature is computed for MIC format data and acoustic intensity vector (IV) is used for FOA format data [10]. Consequently, 17 input feature maps are used to train our model, including 7 feature maps for FOA format signal (4 channels of log-Mel magnitude and 3 channels of IVs) and 10 feature maps for MIC format signal (4 channels of log-Mel magnitude and 6 channels of GCC-PHAT).

### 2.2. Network architecture

The overall architecture of our system is illustrated in Figure 1. The main difference compared with the baseline system is that the bidirectional gated recurrent unit (Bi-GRU) module is substitute by Conformer block, which is proposed to model both local and global dependencies of an audio sequence by combining convolution neural networks and transformers.

The 17 feature maps are fed into the convolutional neural networks (CNN) blocks firstly to extract high-level features. Each CNN block consists of a 2D convolution layer, a rectified linear unit (ReLU) process, a batch normalization layer and a max-pooling operation. The detailed parameters of the CNN block are shown in Figure 1.

The output activation from CNN is further reshaped to a 60 frame sequence of length 512 feature vectors and fed to Conformer block which is used to learn both the position-wise local features and the temporal context information from the CNN output activations. The Conformer block is composed of two Feed Forward

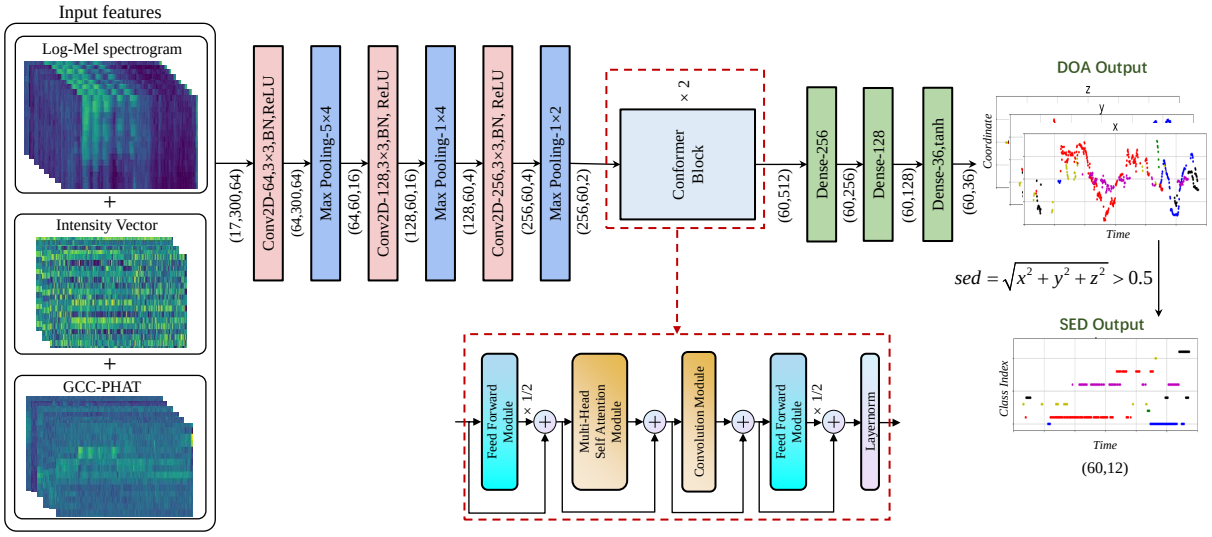


Figure 1: The overview of proposed model architecture

modules sandwiching the Multi-Headed Self-Attention module and the Convolution module. A residual connection is added behind each block. Mathematically, for input $x$ to a Conformer block, the output $y$ can be written as

$$\tilde{x} = x + \frac{1}{2}\mathrm{FFN}(x) \tag{1}$$

$$x' = \tilde{x} + \mathrm{MHSA}(\tilde{x}) \tag{2}$$

$$x'' = x' + \mathrm{Conv}(x') \tag{3}$$

$$y = \mathrm{Layernorm}(x'' + \frac{1}{2}\mathrm{FFN}(x'')) \tag{4}$$

where $\mathrm{FFN}(\cdot)$ refers to the Feed Forward module, $\mathrm{MHSA}(\cdot)$ refers to the Multi-Head Self-Attention module and $\mathrm{Conv}(\cdot)$ refers to the Convolution module. The structure of the Feed Forward module and Convolution module are shown as in Figure 2 and Figure 3, respectively. More details about Conformer can be found in [5]. In our submission system, the dimension of the attention vector is set to 512, the number of attention heads is set to 8 and the kernel size of the depthwise convolution is set to 31. The number of the Conformer block is set to 2 as shown in Figure 1. The Conformer block is followed by FC layers, the activations of the first and the second FC layers are linear while the activation of the last FC layer is set to be tanh since the ACCDOA training target is adopted.

In addition, we have also tried to substitute the CNN layer to other high-level feature representation modules, including SERes-Net, ResNet and Xception. These architectures increased the number of the model parameter, however, showed no improvement in SELD results. The training target is tried to be set as mean square error (MSE) and masked-MSE and the results showed the better performance with ACCDOA.

### 2.3. Data Augmentation

Compared with the scale of model parameters, only 600 recordings in the TAU-NIGENS dataset are obviously insufficient. To overcome this problem and promote the generalization of the model,

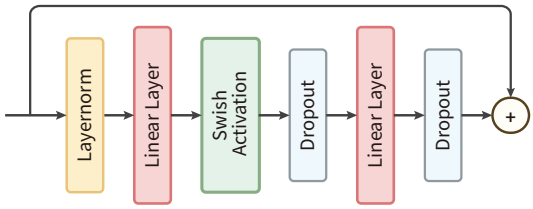


Figure 2: Feed forward module in the Conformer block

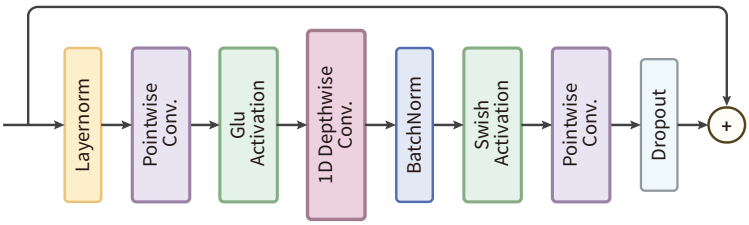


Figure 3: Convolution module in the Conformer block

we use the following data augmentation techniques to expand the dataset.

#### 2.3.1. Audio Channel Swapping

The ACS method has been widely used in the challenges of previous years. The basic idea is to swap channels of the recordings and generate new DOA representations according to the spherically symmetrical geometry of the microphone array. This method is suitable for both MIC and FOA formats. Each DOA label can be transformed and generate seven new labels (90-degree rotation and

Table 1: The performance of the submitted models on the development set

| Index | Model | Loss | Ensemble | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE_{CD}$ | $LR_{CD}$ | $SELD_{score}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Baseline-SELDnet | ACCDOA | None | 0.62 | 39.3% | 22.0° | 47.1% | 0.47 |
| 2 | CNN-Conformer | MSE | None | 0.50 | 60.3% | 18.1° | **68.8%** | 0.33 |
| 3 | CNN-Conformer | ACCDOA | None | 0.49 | 60.4% | 16.0° | 64.0% | 0.33 |
| 4 | CNN-Conformer<br>ResNet-Conformer<br>SEResNet-Conformer | ACCDOA | Yes | 0.47 | 61.3% | 14.0° | 59.0% | 0.33 |
| 5 | CNN-Conformer | ACCDOA | Yes | **0.46** | **63.2%** | **13.9°** | 62.9% | **0.32** |

Table 2: A performance comparision for different models with ACCDOA loss on the development set

| Model | Augmentation | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE_{CD}$ | $LR_{CD}$ | $SELD_{score}$ |
|---|---|---|---|---|---|---|
| Baseline-SELDnet | None | 0.62 | 39.3% | 22.0° | 47.1% | 0.47 |
| Baseline-SELDnet | ACS | 0.59 | 44.1% | 21.4° | 54.9% | 0.42 |
| CNN-Conformer | ACS | 0.52 | 58.7% | 17.3° | 65.7% | 0.34 |
| Baseline-SELDnet | ACS;TFM | 0.56 | 47.6% | 20.5° | 58.2% | 0.40 |
| CNN-Conformer | ACS;TFM | 0.49 | **60.4%** | **16.0°** | 64.0% | **0.33** |
| CNN-Conformer | ACS;TFM;SP | **0.49** | 59.4% | 16.5° | **64.5%** | 0.33 |

mirroring in azimuth, mirroring in elevation).

#### 2.3.2. *Speed Perturbation*

As a data augmentation technique for ASR, SP is proved to be effective. For the SELD task, we use variable speed to stretch the audio and apply it to multi-channel recordings. Specifically, we select a cut-off point in a 60-second recording, increase the sampling rate on one side of the cut-off point, and reduce the sampling rate on the other side to make the length of the entire sequence unchanged. The generated time sequence was used to perform linear interpolation on all features (including MIC and FOA formats) and corresponding labels to obtain a new set of features and labels with the same dimensions.

#### 2.3.3. *Time Frequency Masking*

We apply the TFM method commonly used in ASR to the SELD task. In the Mel-spectrogram features of MIC and FOA recordings, masks are randomly used in the time and frequency domains. Unlike the above two approaches, the masks are generated in each batch, and no new labels will be added.

### 2.4. Model ensemble

As an ensemble method, we use a class-based weighted mean of the output predicted by different models as shown in Figure 4 under the assumption that different models exhibit different capabilities in detection and localization for each sound event. More precisely, let x denotes the coordinates prediction of model m and sound event e. The output of the ensemble model for sound event e is x, where

As an ensemble method, we use a class-based weighted mean of the output predicted by different models as shown in Figure 4 under the assumption that different models exhibit different capabilities in detection and localization for each sound event. More precisely, let $\{x_{e,m}, y_{e,m}, z_{e,m}\}$ denotes the coordinates prediction of model $m$ and sound event $e$. The output of the ensemble model for sound event $e$ is $\{x_e, y_e, z_e\}$, where

$$\begin{cases} x_e = \sum_m w_{e,m} x_{e,m} \\ y_e = \sum_m w_{e,m} y_{e,m} \\ z_e = \sum_m w_{e,m} z_{e,m} \end{cases} \quad (5)$$

with $w_{e,m}$ is the weight for model $m$ and sound event $e$.

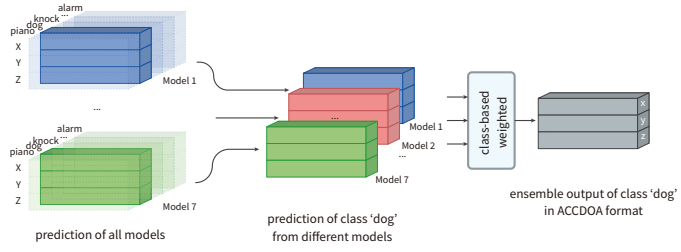


Figure 4: Illustration of the class-based ensemble method

## 3. EXPERIMENTS

The results obtained by the proposed system on the development set are shown in Table 1. The network is trained by the Adam op-

timizer. We train the network for 400 epochs with a minibatch size of 256 and initialize the learning rate as 0.001, which will be decreased by 31.6% if the optimization criterion cannot reduce in 10 consecutive epochs. Finally, the DOA regression determines the prediction of SED, and the threshold is 0.5, which is shown in Figure 1. To train the weight parameters of the ensemble model, the predictions of the training set were used by adopting the cross-validation setup. For submission system index 4, there are 4 CNN-Conformer, 1 SEResNet-Conformer and 1 ResNet-Conformer in the ensemble model. As for submission system index 5, there are 7 CNN-Conformer in the ensemble model while different CNN-Conformer in these systems are obtained by early stopping. The training data set is generated by the abovementioned data augmentation approaches. It can be seen from Table 1 that our proposed model outperforms the baseline in terms of all metrics.

In order to provide more information about the contribution of data augmentation and substituting Bi-GRU block with Conformer block, we provide ablation experiment results in Table 2. It can be shown that ACS, TFM and Conformer block have greatly improved the performance of the system while the benefits brought by the SP data augmentation approach are not obvious.

## 4. CONCLUSION

In this technical report, we described the system participating in the DCASE challenge 2021 task 3. We introduce Conformer into SELDnet to improve the ability to model the temporal dependencies and use several data augmentation approaches to expand the training data. Then we proposed a class-based ensemble model to get a more accurate SELD estimation result. The experiments show that the proposed system achieves better results across the evaluation metrics with a large margin.

## 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, 2018.

[2] 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*, 2020.

[3] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "A model ensemble approach for sound event localization and detection," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[4] 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.

[5] 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," *arXiv preprint arXiv:2005.08100*, 2020.

[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," jan 2021. [Online]. Available: http://arxiv.org/abs/2101.02919

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, pp. 3586–3589, 2015.

[8] 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," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 2613–2617, 2019.

[9] 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.

[10] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," *arXiv preprint arXiv:1905.00268*, 2019.