SOUND EVENT LOCALIZATION AND DETECTION COMBINED CONVOLUTIONAL CONFORMER STRUCTURE AND MULTI-ACCDOA STRATEGIES

Technical Report

Zhaoyu Yan, Jin Wang, Lin Yang, Junjie Wang

Lenovo Research, Beijing, China

ABSTRACT

Sound event localization and detection (SELD) task aims to identify audio sources' direction-of-arrival (DOA) and the corresponding class. The SELD task was originally considered as a multi-task learning problem, with DOA and sound event detection (SED) estimation branches. The single target methods were introduced recently as more end-to-end solutions and achieves better SELD performance. The activity-coupled Cartesian DOA (ACCDOA) vectors was firstly introduced as a single SELD training target, and multi-ACCDOA with auxiliary duplicating permutation invariant training (ADPIT) loss overcame the situation that the same event class from multiple locations. In this challenge, we combined the convolutional conformer structure with the multi-ACCDOA training target and ADPIT strategy. With multiple methods of data augmentation adapted, the proposed method achieves promising SELD improvement compared to the baseline CRNN result.

Index Terms— Sound event localization and detection, conformer, multi activity-couples Cartesian direction of arrival, permutation invariant training

1. INTRODUCTION

Sound event localization and detection (SELD) task aims to identify audio sources' direction-of-arrival (DOA) and the corresponding event class simultaneously from multi-channel inputs. In early studies, SELD was regarded as a multi-task learning with two network branches[1]. DOA and sound event detection (SED) can be evaluated in each branch, and treated as a regression task and a classification task, respectively. By combining the mean square error (MSE) loss and cross-entropy (CE) loss, the network can be trained to convergence.

Recently, the activity-coupled Cartesian direction of arrival (ACCDOA) target was introduced in DCASE 2021[2]. By combining the existence of audio source and corresponding DOA coordinates in a normalized three-dimensional cartesian representation, this single target method achieved state-of-the-art performance in DCASE 2021 task3 challenge with the combination of multiple techniques of data augmentation and model ensemble. Moreover, to overcome the situation that sound sources of the same class from multiple locations exist in one frame, which can significantly reduce the current SELD system performance, the multi-ACCDOA method was proposed with auxiliary duplicating permutation invariant training (ADPIT) strategy[3]. By calculating the ACCDOA loss for different sound sources' arrangement orders and settled for the minimum loss sequence, the class-wise ADPIT scheme enables the model to detect and localize overlapping sound sources from the same class.

The Conformer module was initially proposed in automatic speech recognition (ASR) task as a combination of convolutional layers and Transformer[4]. With a sandwich structure consisting Multi-Head Self Attention Module (MHSA) and convolution module, conformer block achieved better modeling performance of both local and global features of a speech sequence. And recently conformer block is also adopted in SELD tasks and obtain SELD performance improvement[5].

In this challenge, we firstly adopted several data augmentation methods to DCASE 2022 task3 dataset to expand the diversity of data distribution. Secondly, we fine-tuned the CRNN baseline structure based on the augmented dataset, and utilized conformer blocks to replace the bidirectional GRU layers, to extract global information from the abstract features extracted from CNN blocks.

2. PROPOSED METHOD

2.1. Data Augmentation

To expand data distribution diversity, and cover a larger acoustic conditions and sound source locations, following data augmentation methods are taken into consideration on first order ambisonics (FOA) dataset: audio channel swapping (ACS), time-frequency masking (TFM), and multi-channel simulation (MCS)[6].

The FOA format signal consists of four channels (*W*, *X*, *Y*, *Z*). Channel *W* corresponds to the omnidirectional microphone and (*X*, *Y*, *Z*) correspond to three bidirectional microphone aligned on Cartesian axes. In time-frequency domain, consider a spatial point $p(t,f)$, the four FOA channels can be described as follows:

$$\begin{bmatrix} W(t,f) \\ X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ \cos(\varphi)\cos(\theta) \\ \sin(\varphi)\cos(\theta) \\ \sin(\theta) \end{bmatrix} p(t,f). \quad (1)$$

where φ indicates the sound source's azimuth and θ indicates the elevation. Therefore, it is convenient to generate a set of sound source signals from different directions through simple trigonometric transform and audio channel exchange. For the azimuth

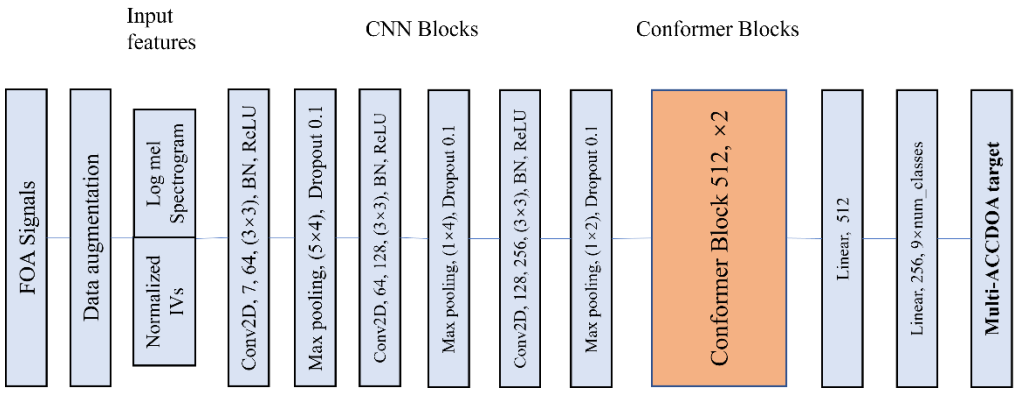

Figure 1 Network structure of proposed method

angle φ, we adopt additional factors {-π, -π/2, 0, π/2, π} and multiplicative coefficients {1, -1}. For the elevation, considering the situation that most sound sources are located near the horizontal plane, we only adopt multiplicative coefficients {1, -1}. Therefore, we can generate 20 sound sources from the original source, and 19 sources are located in new positions.

To further increase the data diversity, time-frequency masking is adopted for multi-channel signals. Frequency SpecAugment[7] and GridMask[8] are adopted as frequency domain masking methods. The frequency mask is added to the continuous spectrogram of frequency f, and a mask with bandwidth $[f_0, f_0+f]$ is applied, where f is a randomly chosen uniform parameter from 0 to frequency mask parameter F, and f_0 is chosen from $[0, v-f]$. v indicates the maximum frequency of input signal. For GridMask, a mask M according to given parameters (r, d, δ_x, δ_y) is produced and multiplied with the spectrogram amplitude, where r denotes the ratio of the mask unit's edge, d is the length of a unit, and δ_x and δ_y are the distances between the first intact unit and boundary of the spectrogram unit. And to generate sound sources of different environments, we used TAU spatial room impulse response database to simulate reverb conditions.

2.2. Features

Two types of input signal dataset are provided in this task: the FOA set and 4-mic array set, with the sound sources were recorded with a 24kHz sampling frequency. As in the data augmentation process, ACS is the only methods that generate sound sources of different locations from the original dataset, and we only utilized FOA as augmented signal for mathematical convenience. Hence FOA dev set is also firstly selected for feature extraction.

In this task, the augmented FOA dev set is firstly extracted 64-dimension log-Mel spectrogram for each channel, with frame length of 40ms and hop length of 20ms. Secondly, for better presentation of phase information, normalized intensity vectors (IV) is extracted from the omnidirectional FOA channel W and each of the other channels. Then log-Mel spectrogram and IV features are concatenated along the channel dimension to form a 7-channel input feature presentation for FOA input signal.

2.3. Network Structure

The network structure we adapted in this task is illustrated in Fiture 1. The input features extracted from augmented FOA dataset have the channel number of 7, 4 of the log-Mel spectrogram and 3 of the intensity vector, therefore form the 7-channel input features. Frame concatenate was adopted in time dimension to get more adjacent information. Firstly, the features are input to stacked CNN blocks to extract local information. Feature maps' channel number increased to 64, 128 and 256 following batch normalization and ReLU activation function. Meanwhile the feature's dimension is gradually reduced to 16, 4, 2 for higher level feature representation by maxpooling layers. Dropout rate of 0.1 is adopted after maxpooling layer. Then after feature reshaping, Conformer blocks are utilized to model global dependencies. The conformer encoder block has a cascade structure of four modules with residual connection , and followed by a post layernorm.

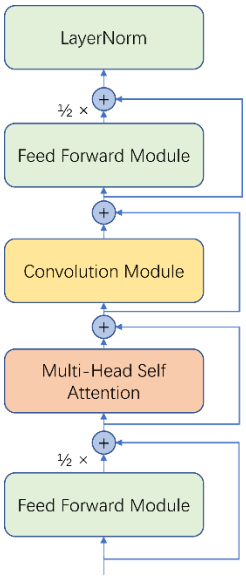

Figure 2 Conformer encoder model architecture

Mathematically, for input x_i to a Conformer block i, the output y_i of the block can be described as following:

$$\tilde{x}_i = x_i + \frac{1}{2}FFN(x_i) \quad (2)$$

$$x'_i = \tilde{x}_i + MHSA(\tilde{x}_i) \quad (3)$$

$$x''_i = x'_i + Conv(x'_i) \quad (4)$$

$$y_i = Layernorm\left(x''_i + \frac{1}{2}FFN(x''_i)\right) \quad (5)$$

where FFN refers to the Feed forward module, MHSA refers to the Multi-Head Self-Attention module, and Conv refers to the Convolution module. We utilized two layers of Conformer blocks, each layer has an encoder dim of 512, to keep the feature dimension constant. After Conformer blocks we set two linear layer to transform the features to training targets. The first linear layer output a 512 dimensional vector, and the last layer output the prediction results. In this task, we utilize the multi-ACCDOA training target with class-wise ADPIT loss, which enables the model to solve the same class overlapping conditions. The multi-ACCDOA target can be seen as a multi-track ACCDOA representation. For each single ACCDOA format, a normalized cartesian coordinate is used to present audio source's existence and DOA, the norm of the coordinate is set to be {0, 1}, 1 indicates the occurrence of the sound source class and 0 indicates the absence. In each track only one event class is detected with corresponding location, which is equivalent to a single-ACCDOA target. In this task we used 3-track 13-class multi-ACCDOA format, set $\boldsymbol{P} \in \mathbb{R}^{3\times N\times C\times T}$ as a training target representation in one time frame. We set N=3 to consider up to 3 sources' overlapping conditions, and C indicates the class number, which is 13 in this task. Therefore, the training target is a vector of 117 dimension for each time frame, which is corresponding to the last linear layer's output dimension.

Multi-ACCDOA format has a class dimension C, which led to a permutation problem in the training process. Therefore, we adapted the class-wise PIT loss for the multi-ACCDOA format:

$$L^{PIT} = \frac{1}{CT}\sum_{c}^{C}\sum_{t}^{T}\min_{a\in Perm(ct)} l_{a,ct}^{ACCDOA} \quad (6)$$

$$l_{a,ct}^{ACCDOA} = \frac{1}{N}\sum_{n}^{N} MSE\left(P_{a,nct}^{*}, \hat{P}_{nct}\right) \quad (7)$$

In the training process of multi-ACCDOA, we calculate the MSE loss of every permutate of sound classes, and select the minimum loss as the optimization object. To further overcome the duplicate of sound classes in single time frame, auxiliary duplicating was introduced to PIT framework. With the class-wise ADPIT, each track is trained to output an original target or a duplicated target. The possible permutations for class-wise AIPIT K_{ct} can be calculated as:

$$K_{ct} = \begin{cases} {}_{N}P_{M_{ct}} \times M_{ct}^{N-M_{ct}}, & M_{ct} > 0, \\ 1, & M_{ct} = 0. \end{cases} \quad (8)$$

where ${}_{N}P_{M_{ct}}$ denotes M_{ct}-permutations of N.

3. EXPERIMENT RESULTS

In the experiments, we evaluated our model on the development set of STARSS 22 and DCASE 2022 simulated data for baseline training. The data augmentation methods are performed on the original dataset with a random sequence. The baseline system was the ACCDOA-based convolutional recurrent neural network (CRNN), and achieved ER_{20}, F-score, LE_{CD}, LR_{CD} performance of 0.72, 0.33, 43.45 and 0.34, respectively. With a SELD-score of 0.56. Then the Conv-Conformer network with multi-ACCDOA is trained in the same condition with CRNN, and achieved the metrics performance of 0.58, 0.35, 22.53 and 0.42, with a SELD-score of 0.48. Therefore, the Conv-Conformer network structure can get better SELD performance on all metrics compared to the CRNN network, especially on the DOA estimation accuracy, and got 14% overall SELD-score performance improvement.

4. CONCLUSION

This technical report described the SELD system participating in the DCASE challenge 2022 task3. We combined the Conv-Conformer network structure with the multi-ACCDOA training target and ADPIT strategy to improve the baseline system's SELD performance in multiple sound source conditions. The experiments show that the proposed system achieved better results in four main metrics than the baseline CRNN system, especially in the DOA estimation accuracy. The proposed method also verified the performance of Conformer structure in modeling global information of high-level features extracted from CNN blocks.

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] Shimada K, Takahashi N, Koyama Y, et al. Ensemble of ACCDOA-and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection[J]. arXiv preprint arXiv:2106.10806, 2021.

[3] Shimada K, Koyama Y, Takahashi S, et al. Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 316-320.

[4] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.

[5] Zhang Y, Wang S, Li Z, et al. Data Augmentation and Class-Based Ensembled CNN-Conformer Networks for Sound Event Localization and Detection[R]. Technical Report of DCASE Challenge. 2021. Available online: http://dcase. community/documents/challenge2021/technical_reports/DCASE2021_Zhang _67_t3. pdf (accessed on 8 May 2022), 2021.

[6] Wang Q, Du J, Wu H X, et al. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection[J]. arXiv preprint arXiv:2101.02919, 2021.

[7] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.
[8] Chen P, Liu S, Zhao H, et al. Gridmask data augmentation[J]. arXiv preprint arXiv:2001.04086, 2020.