TACCNN: TIME-ALIGNMENT COMPLEX CONVOLUTIONAL NEURAL NETWORK

Technical Report

Kaibin Guo, Runyu Shi, Tianrui He, Nian Liu, Junfei Yu

Multimedia Technology Department Xiaomi INC.
Beijing 100085, CHN

ABSTRACT

In this technical report, we show our system submitted to the DCASE2022 challenge task3: Sound Event Localization and Detection(SELD) Evaluated in Real Spatial Sound Scenes. At first, we review the famous deep learning methods in SELD, and point out that these works have ignored the time alignment from the perspective the arrival time of the signal, and the amplitude and the phase are modelling in the separate way. Therefore, we put forward a new model, Time Alignment Complex Convolutional Neural Network(TACCNN). In our model, we suggest to use 3DCNN or ConvLSTM to align the feature from different mics. Moreover, we propose to compile the mel spectrogram with intensity vector as the complex vector, and then extract salient feature on the new feature by using complex convolutional neural network. Lastly, we apply Bi-GRU with self-attention to extract the relative information about sound event to determine the rotation of the sound event. The results show that the time alignment block greatly improve the performance of CNN-GRU model. Complex convolutional neural network has the similar result compared with the real convolutional neural network. It seems that we need more experiments to discover the role of complex convolutional neural network.

Index Terms— time alignment, complex convolutional neural network, sound event localization and detection, feature fusion

1. INTRODUCTION

Sound event localization and detection(SELD) consists of two kinds of task: Sound event detection(SED) and Direction of arrival(DOA) estimation. The goal of SED is to recognize the label of sounds given the audio. DOA estimation can be seen a regression task, which aims at estimating the direction of arrival of the sound source. In dcase2022[1] task3, the doa is equivalent to the Cartesian coordinate.

Recently, deep learning method, such as Convolutional Recurrent Neural Network(CRNN)[2], are mainly applied to solve SELD problem. Guirguis K et al. proposed to replace RNN with TCNN to model the sequence information. In [3], Shimada K et al. suggested replacing lots of the normal convolution with the dilated convolution so as to improve the receptive field. While in [4], Lee S H et al. regarded SELD task as multi-task learning problem, and applied transformer to fuse the feature extracted from the encoder of SED task and DOA task.

However, mostly the arrival time from the source varies from mic to mic. The work mentioned above didn't not take this factor into account. That is, the salient feature about sound event could not extracted from the sequence properly. Moreover, most work consider the mel spectrogram and intensity vector as separate feature, and stack two type of feature along the channel axes. When it comes to the learning two task at the same without using multi-task learning method like [2], stacking these feature may lead to loss critical information by convolution.

Therefore, we propose a new neural network, called Time Alignment Complex Convolutional Neural Network(**TACCNN**). TACCNN mainly consist of three parts, **Time alignment block**, **Feature fused block** and **Time series modeling block**. In time alignment block, we mainly employ 3DCNN or ConvLSTM to align the feature from different mics and compress feature map. As for Feature fused block, the aligned feature would pass through a complex convolutional neural network to extract salient feature. Then, we use Bi-GRU layer with self-attention to extract relative feature about sound event, and two fully-connected layers to determine the rotation of the sound event.

The core contributions are summarized as follows:

- To the best of our knowledge, this is the first work to apply convlstm and 3DCNN to align the time information from different mics.
- To the best of our knowledge, this is the first work to apply complex convolutional neural network for SELD task.

2. PROPOSED METHOD

In this section, we first introduce the data augmentation for SELD. Then, we would introduce our method TACCNN in detail.

2.1. Data augmentation

Considering the training samples provided by the development dataset is insufficient, which resulting in poor generalizability of the network.Thus,we apply a novel data augmentation method called rotation to expand the number of labels which representing different DOA information[5].

2.1.1. Channel swapping

For the FOA format,the channel swapping method contains 15 fixed rotation patterns in addition to raw data.Either azimuth ϕ and elevation θ is transformed or both are transformed for each pattern.The same transformation is applied also to the FOA channels directly,in order to guarantee the relationship between steering vectors and observations is maintained.The 16 transformations to the label:the elevation θ' =(θ,$-\theta$),the azimuth ϕ'=(ϕ,$\phi - \pi/2$,$\phi + \pi/2$,$\phi + \pi$,$-\phi$,$-\phi - \pi/2$,$-\phi + \pi/2$,$-\phi + \pi$) for each transformed θ.The algorithm is suitable for the FOA and MIC recordings.

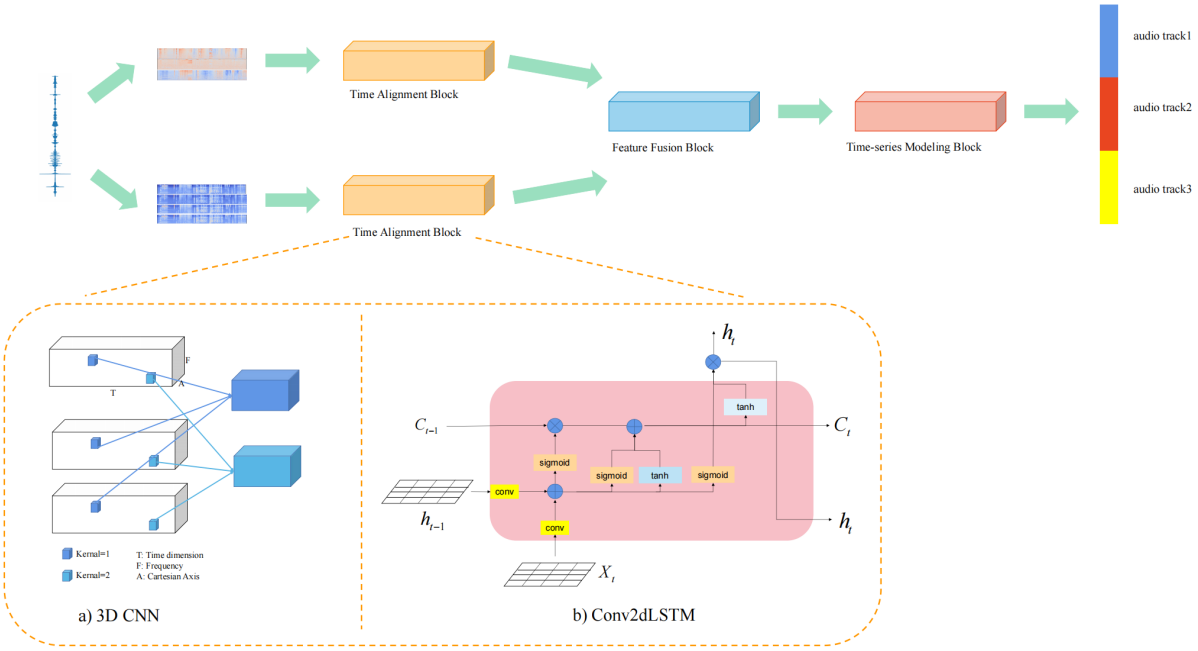

Figure 1: **The framework of our method, TACCNN**. The framework mainly consist of time alignment block, feature fusion block and time-series modeling block. 1) We extract the mel spectrogram and intensity vector from the original audio 2) and get three audio track from the output of the model. 3) At last, we get the SELD result from prediction of these audio tracks. On the bottom of the figure, we show 3DCNN and Conv2dLSTM as alternative time alignment block

2.1.2. Labels first

Unlike the previous method,the augmented labels of this algorithm are not fixed. A randomly generated angle is added to the azimuth and elevation to obtain new DOA labels.For convenience, the azimuth and elevation transformation are divided into two steps.For the azimuth,the augmentation angle α is selected randomly and the rotation matrix is computed, then it is applied to the original channels to generate azimuth-augmented channels.The augmented elevation is similar to the azimuth.Differently,the Rodrigues' rotation formula is applied to obtain the final-augmented channels. However,this method is not suitable for overlapping sound events.To solve this problem, we apply the channel swapping method for the elevation augmentation.

2.1.3. Channels first

The basic ideal of this method is to apply a transformation to the channels firstly, than to labels.An orthonormal matrix is selected and the augmented channels are computed.Then the same orthonormal matrix is applied to the labels in Cartesian coordinate.This algorithm generates the most number of augmented channels but loses control over the labels.

2.2. TACCNN

In this section, we will introduce the architecture of our proposed method, which consist of three parts, time alignment block, feature fusion block, and time series modeling block. First, the mel feature and intensity feature would pass through the different 3DCNN or Conv2dLSTM to align the signal from different audio mic. Then, feature aligned at time axis would be feed into the complex/real convolution neural network which acts like the feature fusion block to extract salient feature. At last, a time-series model with attention block are employed to extract the acoustic event information. The framework of our method are shown in Figure 1.

2.2.1. Time alignment

Since the arrival time of different mics varies, we can't apply Convolutional Neural Network(CNN) or Recurrent Neural Network(RNN) on the original audio feature directly as regular sound event detection task. Traditionally, time delay compensation is employed to align the received signal of each mics in signal processing. In our work, we propose to apply deep learning method to align the signal from the perspective of the feature.

3DCNN are mainly applied to video-based tasks, as it can extracts more information by taking time into account. Following the 3DCNN, 3DMaxpooling or 3DAveragePooling is often applied to compress the feature map, and the frame length is also shortened at the same time. As for spatio-temporal modelling, in [6], Shi et al. put forward ConvLSTM to handle the phenomenon that the spatio-temporal data has to be unfolded into the shape of 1D vectors before processing. Compared with custom LSTM, ConvLSTM replaces the fully-connected operator with convolution operator in LSTM. that is, ConvLSTM can model the time-series information at each pixel as well as model the spatial information by convolution operator. Considering the advantage of 3DCNN and ConvLSTM, we apply one of the blocks as the time alignment block.

Supposed the shape of the audio feature extracted from the mics are $[T, n, d]$, where T is the frame length, n is the number of the audio mics while the d is the feature dim at each time and each mic,

we first add a new axis to expand the size of the feature map. The new axis can be regarded as channel, and the feature size would be $[T, 1, n, d]$. When applying 3DCNN as time alignment block, we make the size of convolutional kernel along the audio axes be equal equivalent to n, and while the size along the time axis should be large enough. In this way, we can extract elaborate feature from different audio mics, and align the time information by pooling method further. As for ConvLSTM, the condition of kernel size is similar to that's of 3DCNN, and we also apply pooling method to compress the feature map.

2.2.2. Feature fusion

After time alignment block, we can get salient feature from the origin mel spectrogram and intensity vector respectively. Usually, convolutional neural network is applied as encoder to extract refined feature by stacking two feature along the channel axis. From the perspective of the signal processing, mel spectrogram represents the amplitude of the signal while intensity vector shows the phase of the signal. But in CNN, the amplitude and phase will be computed respectively since they are in different channels, and they are only compiled by adding at last. In this way, both feature could not help the other to learn task-related information.

Considering the meaning of the amplitude and phase in signal processing, we propose to employ complex convolution neural network (ComplexCNN)[7] as feature extractor, in which we can combine the the amplitude feature with phase feature as complex tensor. Compared with the real-value convolutional neural network, ComplexCNN will multiply the amplitude and phase to extract more information when using convolution. After the complex convolution neural network, we can also apply complex batch normalization, complex maxpooling/ averagepooling and RELU activation function just like real-value deep neural network. Through some complex neural network, we compute the mode of the feature map at each pixel, and reshape it to 1D vector.

2.2.3. Time-series modeling

In our work, we apply Bi-GRU to extract the information related to the sound event. Considering the possibility that different type of the event would appear at the same audio, we also employ self attention block to extract refined feature for each sound event. After time-series modeling, we employ two fully-connected network (FCN) to detect the rotation for each sound event at each time and each audio track.

3. EXPERIMENT

In this section, we show the experimental settings and our results on the development set.

3.1. Experimental settings

The mel spectrogram from each channels and the intensity vector are extracted as the input features.The sampling frequency is 24 kHz.Using a Hanning window of length 20 ms to the STFT and the frame hop is 10 ms.Here,the mel feature consists of four channels and the intensity vector consists of three channels, e.g. x,y,z-axis.

In our work, we explore the effect of the following models, that is:

1) Tcresnet
2) RealCNN+GRU
3) 3DCNN+RealCNN+GRU+FC
3) 3DCNN+ComplexCNN+GRU+FC(TACCNN)

We apply Mean-square-Error as loss function for each audio tracks. The more detail about the loss function can be referred to [8]. We use Adam as optimizer, and the learning rate is setting as 0.001.

3.2. Experimental Results

in this seciton, we should the result on development set. As it can be seen from table 1, 3DCNN+realCNN+GRU leads to great improvement in all the evaluation compared with the CRNN. This shows the effectiveness of the time alignment block. Compared TACCNN with 3DCNN+realCNN+GRU, we can see the similar performance between them. It need more epochs or more experiment to show the difference.

Table 1: Comparisons with baseline models on development set(epoch=30)

Model	**ER**	**F1**	**LE**	**LR**
Tcresnet	0.80	0.18	31.15	0.50
RealCNN+GRU	0.69	0.23	35.27	0.34
3DCNN+realCNN+GRU	0.61	0.29	23.51	0.49
TACCNN	0.63	0.25	23.93	0.48

4. CONCLUSION

In this technical report, we present our results on DCASE2022 task3: sound event localization and detection(SELD) in Real Spatial Sound Scenes. In our system, we put forward a new learning framework, Time Alignment Complex Neural Network. In our work, we propose to apply 3DCNN or ConvLSTM to align the time information as we take the arrival time of signal into account. Besides, we also suggest using complex convolutioal neural network to compile mel spectrogram and intensity vector as to learn more information about SED and DOA. The experimental results show that the importance of time alignment and the role of complex neural network need to be explore further.

5. REFERENCES

[1] http://dcase.community/challenge2022/.

[2] 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*, 2022.

[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," *arXiv preprint arXiv:2106.10806*, 2021.

[4] S.-H. Lee, J.-W. Hwang, S.-B. Seo, and H.-M. Park, "Sound event localization and detection using cross-modal attention and parameter sharing for dcase2021 challenge," *DCASE2021 Challenge, Tech. Rep.*, 2021.

[5] Y. M. e. a. Mazzon L, Koizumi Y, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[7] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=H1T2hmZAb

[8] https://github.com/sharathadavanne/seld-dcase2022.