# Sound Event Localization and Detection Based on Time-Frequency Separable Convolutional Compression Network

Shih-Tsung Yang[1], Fong-Ci Jhou[1], Jia-Ching Wang[2], and Pao-Chi Chang[1]
[1] Dept. Communication Engineering, National Central University, Taiwan
[2] Dept. Computer Science and Information Engineering, National Central University, Taiwan
{styang.vaplab@gmail.com, fcjhou.vaplab@gmail.com, jcw@csie.ncu.edu.tw, pcchang@ce.ncu.edu.tw}

***Abstract*— This work proposes a Time-Frequency Separable Convolutional Compression Network (TFSCCN) as a system architecture for sound event localization and detection. It utilizes 1-D convolution kernels of different dimensions to extract features of time and frequency components separately, and also reduces the amount of model parameters by controlling the increase or decrease of the number of channels in the neural network. In addition, the model combines multi-head self-attention (MHSA) to obtain global and local information in time series features, and uses dual-branch tracking technology to effectively locate and detect the same or different overlapping sound events.**

***Keywords*— sound event localization and detection, time-frequency separable convolutional compression network, multi-head self-attention, dual-branch tracking.**

## I. INTRODUCTION

In the audio research field, joint sound event localization and detection (SELD) is one of the fast growing research topics. By simulating the hearing ability of human ears, it can distinguish various sound events in the environment and their locations and movement trajectories. The most famous research SELDnet, proposed by S. Adavanne *et al* [1], was chosen as the baseline for DCASE 2019/2020 Task 3, which used the Convolutional Recurrent Neural Network (CRNN) structure for training and prediction.

Different from the commonly used 2-D convolution, we use 1-D convolution to extract features of a single time or frequency component. It can distinguish each sound event class according to the different characteristics of the frequency distribution of different sound events. Meanwhile, it can also track the spatial location and movement trajectory. However, in order to achieve better performance, most studies take the approach to adopting higher complexity models to improve performance. In this study, by controlling the timing of the increase and decrease of the number of channels, the number of parameters is effectively reduced significantly while maintaining better performance. In addition, to detect overlapping sound events of the same or different event classes, a dual-branch tracking method is used to track individual sound events.

## II. METHODOLOGY

### A. Features

The features that are sent to the proposed system can be obtained from the four-channel First-order of Ambisonics (FOA) audio. It can be converted into a time-frequency domain representation through STFT operations with $K$-point DFT and a $K$-point Hamming window. The dimensions of the magnitude and the phase spectrums are $T \times (K/2+1) \times 4$, where $T$ is the time frame of the output feature.

In addition, since the sound intensity vector (IV) is a vector with magnitude and direction, it can be used as a feature of sound event localization, as shown in (1),

$$I(f,t) = \Re\left\{ W^*(f,t) \begin{bmatrix} X(f,t) \\ Y(f,t) \\ Z(f,t) \end{bmatrix} \right\} \quad (1)$$

where $W$, $X$, $Y$, $Z$ are the representations of the time domain signals of the four channels after STFT, $\Re\{\cdot\}$ indicates the real part, and $*$ denotes the conjugate. The output dimension is $T \times (K/2+1) \times 3$. Finally, the three features are normalized using $M$ mel-band filter banks, converted to log-mel spectrum by taking the logarithm, and stacked as the input features of the model. The overall dimensions are $T \times M \times 11$.

### B. Network architecture

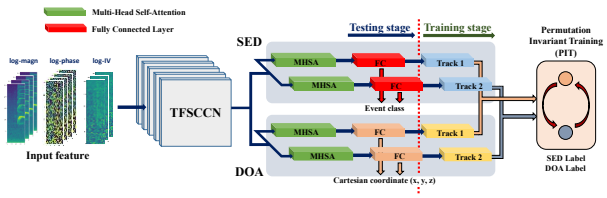


Fig. 1. System architecture diagram

As shown in the network architecture of Fig 1, a Time-Frequency Separable Convolutional Compression Network (TFSCCN) is constructed for feature extraction. TFSCCN adopts the design concept of SqueezeNet to construct a convolutional network with the same number of module layers as SqueezeNet, and adds a skip connection method to combine the input and output results of each module when transferring, as shown in Fig 2.

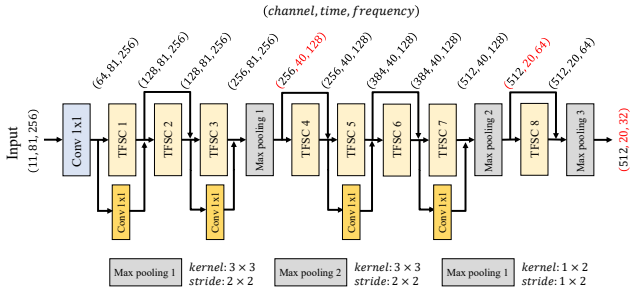


Fig. 2. The block diagram of TFSCCN

The TFSC module is composed of 1-D convolutions of different sizes, as shown in Fig 3, which uses the following design concepts:

(1) After the feature map passes through the TFSC module, the size $(time, frequency)$ of the feature map remains unchanged, only the number of output channels $(channel)$ is changed.

(2) Reduce the input channel to $1/16$ times the output channel through $1 \times 1$ convolution.

(3) Through different sizes of 1-D convolutions along the frequency and time axes to extract the features of the

frequency and time components, and increase the number of channels of the feature map.

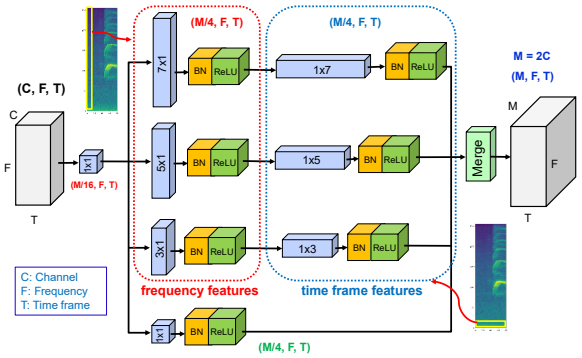


Fig. 3. TFSC module

After the TFSCCN, the final output is fed to the two branches of localization and detection. Each branch is divided into two tracks. Each track is used to identify a single sound event, and the time series features are further processed through the multi-head self-attention (MHSA) mechanism to obtain the relevance of each timestep.

### C. *Dual-branch tracking*

Dual-branch tracking adopts the concept of track-wise proposed by Y. Cao *et al* [2]. The sound event localization and detection operations use mean squared error (MSE) and binary cross entropy (BCE) as the loss functions, respectively. Both the localization and detection branches calculate the loss of the two tracks and average them to obtain the average loss of each branch $(L^{DOA}, L^{SED})$. After the two sets of loss $(L_1^{DOA}, L_1^{SED})$ and $(L_2^{DOA}, L_2^{SED})$ are calculated, the permutation invariant training (PIT) method is used to select the combination with the smallest loss to determine the final sorting for backpropagation.

$$Loss_1 = L_1^{DOA} + L_1^{SED} \qquad Loss_2 = L_2^{DOA} + L_2^{SED} \qquad (2)$$

As shown in (2), when $Loss_1 \leq Loss_2$, it takes $(L_1^{DOA}, L_1^{SED})$ as the loss for back propagation, that is, uses the original track sorting. Conversely, when $Loss_1 > Loss_2$, the set of loss $(L_2^{DOA}, L_2^{SED})$ is used, that is, it flips the original track.

## III. EXPERIMENT

### A. *Dataset and experiment setup*

We use *TAU Spatial Sound Events 2020 Dataset-FOA* for experiments. Each audio file is a 60-second 4-channel FOA audio format, and contains 14 types of sound events. The dataset contains a development dataset with 600 audio files and an evaluation dataset with 200 audio files. The sampling rate of each audio file is 24kHz, and at most two overlapping sound events can occur at the same time.

In the pre-processing of features, every 2 seconds is used as a time frame. In addition, 1024-point DFT and a 1024-length Hamming window are used for STFT, and 256 mel-band filter banks are used to convert to log-mel spectrum. In the prediction stage, we set the threshold of SED to 0.5 to determine whether the sound event class is active.

### B. *Experimental results*

The experimental result uses the evaluation metrics proposed by DCASE 2020 Task 3 to conduct joint evaluation of sound event localization and detection [3], including evaluation of Error rate (ER), F-score (F), Localization error (LE) and Localization recall (LR). Table 1 compares the complexity and performance differences between TFSCCN and various neural networks, including the baseline of DCASE 2020 Task 3 (SELDnet) and several lightweight models, such as depthwise separable convolution neural network (DSCNet), ResNet, and SqueezeNet.

From the results, TFSCCN achieves the best overall performance with the least amount of model parameters compared with other lightweight models.

**Table 1. Use the evaluation dataset to compare the performance of different CNNs**

| Network architecture | Detection | | Localization | | Total Param |
|---|---|---|---|---|---|
| | ER(20°) | F(20°) | LE(°) | LR | |
| Baseline | 0.75 | 32.5% | 26.7 | 57.4% | 513k |
| DSCNet | 0.54 | 56.5% | 15.2 | 67.4% | 17M |
| ResNet | 0.47 | 62.8% | 12.1 | 70.2% | 17.7M |
| SqueezeNet | 0.405 | 68.8% | 10.5 | 74.4% | 14.2M |
| **TFSCCN** | **0.385** | **69.5%** | **12.7** | **79.4%** | **11.5M** |

### C. *Visualization of results*

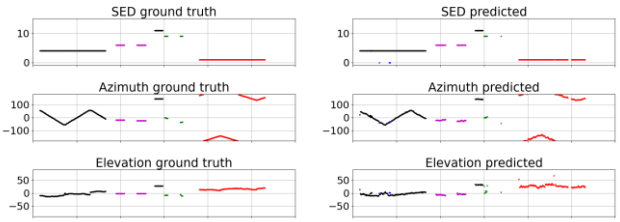


Fig. 4. Sound event localization and detection result

Figure 4 shows the ground truth and prediction results of the sound event class, azimuth, and elevation. Different colors indicate different types of sound events, and show the length of time they appear and the corresponding spatial location and movement trajectory. From the results, it is clear that it can accurately predict the sound events that occur at different times no matter it is a stationary or moving sound source. In addition, it can also accurately locate its corresponding angle information.

## IV. CONCLUSION

We have proposed TFSCCN system that uses 1-D convolutions of different dimensions to extract features of time and frequency components can greatly reduce the amount of model parameters. The experimental results exhibit that it can effectively improve the sound event localization and detection performance. Compared with other CNN models, it has the fewest parameters and the best performance in our experiments.

## V. 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*, 13(1):34–48, March 2018.

[2] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-Independent Network for Polyphonic Sound Event Localization and Detection," *DCASE 2020 Workshop*, November 2020.

[3] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, Oct 2019.