# Decoupling Temporal Convolutional Networks Model in Sound Event Detection and Localization

*Shen Song, Cong Zhang*, Xinyuan You*

*School of Mathematics and Computer, Wuhan Polytechnic University, China*
*1403826619@qq.com, hb_wh_zc@163.com, 1272388955@qq.com*

## Abstract

Sound event detection is sensitive to the network depth, and the increase of the network depth will lead to a decrease in the event detection ability. However, event localization has a deeper requirement for the network depth. In this paper, the accuracy of the joint task of event detection and localization is improved by decoupling SELD-TCN. The joint task is reflected in the early fusion of primary features and the enhancement of the generalization ability of the sound event detection branch as the DOA branch mask, while the advanced feature extraction and recognition of the two branches are carried out in different ways separately. The primary features extracted by resnet16-dilated instead of CNN-Pool. The SED branch adopts linear temporal convolution to realize sound event detection by imitating the linear classifier, and ED-TCN is used for the localization detection branch.

The joint training of the DOA branch and the SED branch will affect each other badly. Using the most appropriate way for both branches and masking the DOA branch with the SED branch can improve the performance of both. In the TUT Sound Events 2019 dataset, the DOA error achieved an error effect of 6.73, 8.8 and 30.7 with no overlapping source data, with two and three overlapping sources, respectively. The SED accuracy has been significantly improved, and the DOA error has been significantly reduced.

**Keywords:** Decoupling, Dilated convolution, Causal convolution, Mask, Temporal convolutional network

## 1 Introduction

Sound Event Localization and Detection (SELD) is divided into two separate tasks: Sound Event Detection (SED) and Sound Event Localization what is called Time difference of arrival (DOA). Sound event detection is a multi-label problem whose main role is to identify the start time and end time of a sound event as well as the event itself. In addition to this, audio is divided into stereo and mono. For the mono audio, only one sound time can be identified at any given time. However, stereo channel sound can detect multiple overlapping sound events at any given time [1]. SELD can not only provide assistance to other sensors and improve safety, but can also be more efficient than visual perception [2].

In sound event monitoring (SED). Mono SED systems can only detect up to one acoustic event at any given time. If the purpose of the system is to detect all events occurring at once, multiple sound events are likely to overlap in time. For example, an audio recording from a busy street may contain footsteps, speech, and car horns, all in the form of a mixture of events. This is shown in Figure 1.

Sound source localization (DOA) focuses on identifying the location of sound sources. It plays an important role in applications such as robotic hearing, speech enhancement, source separation and acoustic visualization [3-5]. DOA has been studied using two main approaches: parameter-based and learning-based methods. Parameter-based DOA methods are divided into three categories: time difference of arrival (TDOA) estimation, maximized steered response power (SRP) of a beamformer device, and high-resolution spectral estimation (time difference of arrival (TDoA) estimation, maximized steered response power (SRP) of a beamformer, and high-resolution spectral estimation.) The generalized correlation (GCC) method [6] is the most widely used method in time difference estimation. Since the time difference information is transmitted in the form of phase rather than cross-spectral amplitude, the GCC phase transform (GCC-PHAT) is proposed, which eliminates the effect of amplitude while retaining only the phase.

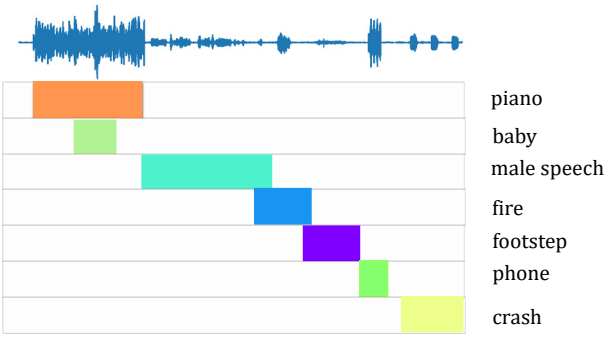


**Figure 1.** Multi-label sound events

Adavanne et al. [7] used CRNN to learn the features of the generalized cross-correlation (GCC-PHAT) for mono signal. Adavanne [8] arranges features from each channel as different layers of a multi-layered input volume enables

*Corresponding Author: Cong Zhang; E-mail: hb_wh_zc@163.com
DOI: 10.53106/160792642023012401009

the network to learn the sound events in multichannel audio better than a simple concatenation of the features. Adavanne et al. [9] use a 3D CNN for multichannel audio input that will learn features with phase-transformed generalized correlation (GCC-PHAT) among individual channels. CNNs can learn time and frequency-domain invariance by learning fixed connections between the input and hidden units. Parascandolo [10] proposed to use BLSTM to realize sound event localization and detection, which can realize the exchange of sequence information across windows, and achieve better recognition effect in Polyphonic SED. However, RNN and BLSTM have problems such as weak feature extraction ability, small receptive field and weak parallel ability. Therefore, CRNN is proposed based on their respective advantages. Cakır [11] compares the CRNN method performance against that of CNN and RNN to analyze the nature of CNN's ability to extract higher-level features that are invariant to local spectral and temporal variations; and the very powerful properties of RNN in learning long-term temporal context in audio signals, combining the advantages of both. The results show a considerable improvement in the performance of the CRNN approach. Guirguis [12] proposed SELD-TCN which introduces causal convolution and dilated convolution into SELD, using this time series of causal convolution to achieve parallel training, which combines the advantages of CNN. The parallel ability of SELD-TCN has been significantly improved and combined the advantages of CNN, but the difference in feature granularity leads to inconsistent learning methods for the network. It needs to be decoupled to achieve further optimization and implementation.

In this paper, we implemented the joint extraction of SED and DOA with mel spectrum and GCC-PHAT as input features. Our experimental results show that a more satisfactory result can be obtained by using 15% masking, so the DOA is masked from the results of SED to improve the performance of DOA, while the joint training of SED and DOA will reduce the performance of both.

The main contributions of this paper are concluded as follows:

1. The primary features extracted by resnet16-dilated instead of CNN-Pool.
2. The SED branch and the DOA branch are trained separately according to the task. When training the SED branch, a layer of linear temporal convolution is used to achieve sound event detection. When training the DOA branch, a deeper ED-TCN is used to achieve sound localization.
3. The upsampling groundtruth obtained by the SED branch is used to mask the DOA branch by 15% to improve the generalization ability of the DOA branch.

The model of DOA branch uses a time-domain convolutional network TCN to acquire features and increase the robustness and feature learning capability of GCC-PHAT. The features extracted from GCC-PHAT can in turn act on the acquired classification features and interact with each other to both decouple and prevent the increase of common loss caused by shared weights, and to make full use of individual channel features to improve recognition and localization accuracy. In view of the above problems, the model proposed in this paper has achieved better results than the previous models.

The rest of the paper is organized as follows. Section 2 shows the feature extraction include MFCC and GCC-PHAT. Section 3 describes the proposed Network Structure of SED branch, DOA branch and its joint part. Section 4 is the joint task to merge the branches. Section 5 is experimental results and analysis on the network. Section 6 depicts the main work of the paper and gives some suggestions for further work.

## 2 Feature Extraction

### 2.1 MFCC Spectrogram

The structure of the Mel filter bank. The MFCC process is mainly divided into pre-weighting, framing, windowing, fast Fourier transform, Mel filter bank, logarithmic compression, discrete cosine transform, and dynamic feature extraction. As shown in the Figure 2.

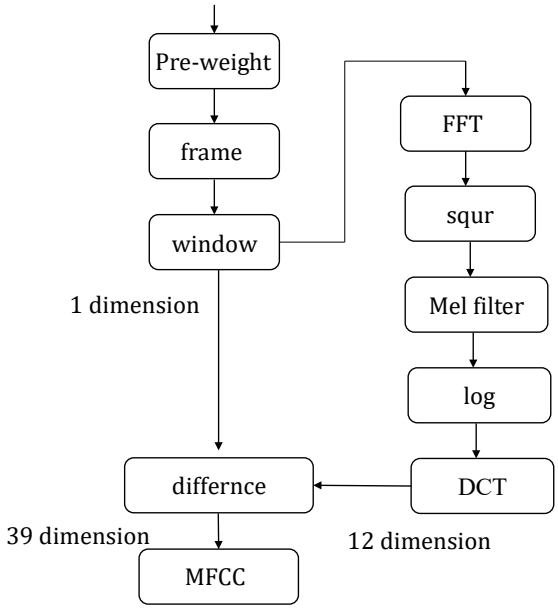


**Figure 2.** Mel scale Frequency Cepstral Coefficients

Mel scale Frequency Cepstral Coefficients (MFCC) features [13] may be used when 40-dimensional features or 27 last in which the first 13 are taken and the second to last is retained to form 39-dimensional features or 26-dimensional features or 12-dimensional. The specific process is to first pre-emphasize, split the frames, add windows, and then do FFT on each frame to convert the time domain signal into frequency domain signal, and stack each frame of frequency domain signal (spectrogram) in time to form an energy spectrum to obtain the acoustic spectrogram. After applying Mel filter to the acoustic spectrum, logarithmic compression is performed to obtain the compressed log Meier acoustic spectrum, and the discrete cosine transform is done to the log Meier acoustic spectrum. Taking the 2nd to 13th coefficients, the MFCC, a 12-dimensional characteristic inverse spectral parameter, is formed.

However, the standard MFCC is only a static feature of the signal, and the differential spectrum of static features can

be used to express the dynamic features of the audio signal. The general feature extraction up to 39 dimensions is sufficient to achieve the recognition capability of the audio system. The components can be expressed as $N/3$ MFCC coefficients + $N/3$ first-order differential + $N/3$ of second-order differential + frame energy (short-time feature representation can be chosen).

### 2.2 GCC-PHAT

Sound sources with different spatial locations have different intensities in the binaural channel. Most overlapping sound events have different frequency propagation. This combination of intensity differences in different frequency bands can be used to distinguish overlapping sound events. This idea originates from the binaural intensity difference (IID) used by humans [14]. MFCC band energies extracted from two binaural channels using 40 mel bands in a 40 ms Hamming window are used as features. The neural network can learn to obtain IID information from these channel energies.

The generalized cross-correlation function delay estimation algorithm estimates the delay value based on the peak of the cross-correlation function of the two microphone signals. In the sound source localization system, each array element of the microphone array receives the target signal from the same source. The signals of each channel have a strong correlation with each other. Ideally, the time delay between two microphone observation signals can be determined by calculating the correlation function between each two signals.

The microphone array signal is.

$$\begin{aligned} x_1(t) &= \alpha_1 s(t-\tau_1) + n_1(t) \\ x_2(t) &= \alpha_2 s(t-\tau_2) + n_2(t). \end{aligned} \quad (1)$$

$s(t)$ denotes the source signal, $n_1(t)$ and $n_2(t)$ and is the ambient noise. $\tau_1$ and $\tau_2$ is the propagation time of the signal from the source to the two microphone array elements.

The cross-correlation algorithm is often used for time delay estimation.

$$R_{x_1x_2}(\tau) = E(x_1(t)x_2(t-\tau)). \quad (2)$$

Relationship between the cross-correlation function and the cross-power spectrum.

$$\begin{aligned} R_{x_1x_2}(\tau) &= \int_0^{\pi} G_{x_1x_2}(\omega) e^{-jwt} d\omega \\ &= \int_0^{\pi} X_1(\omega) X_2^*(\omega) e^{-jwt} d\omega. \end{aligned} \quad (3)$$

In the microphone array signal processing, the reverberation and noise effects cause the $R_{x_1x_2}(\tau)$ peaks to be less prominent and decline the accurancy of the delay estimation. In order to sharpen the peaks $R_{x_1x_2}(\tau)$, the power spectrum can be weighted in the frequency domain based on the a priori information of the signal and noise, which can suppress noise and reverberation interference. Finally, the Fourier inverse transform is performed to obtain the generalized cross-correlation function $R_{x_1x_2}(\tau)$.

$$R_{x_1x_2}(\tau) = \int_0^{\pi} \varphi_{12}(\omega) X_1(\omega) X_2^*(\omega) e^{-jwt} d\omega. \quad (4)$$

Where denotes the frequency domain weighting function. The block diagram of the generalized cross- correlation time delay estimation algorithm is as Figure 3.

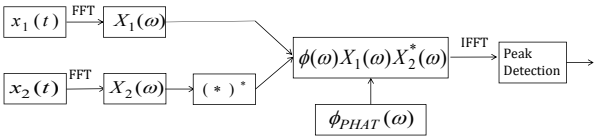


**Figure 3.** Generalized cross-correlation delay

Its weighting function uses PHAT weighting, which can achieve better results for large noise, and it is mainly used under the near-field model. The distance of the sound source needs to be considered relative to the distance in the microphone array when the sound source is close. The near-field model considers the acoustic wave received by the microphone as a spherical wave. The near-field model is more in line with the actual application and can provide more information about the location of the sound source and improve the accuracy of localization. The direct signal from the sound source, the reflected signal passing through walls or obstacles, and the ambient noise signal.

$$\varphi_{PHAT}(\omega) = \frac{1}{\left|G_{x_1x_2}(\omega)\right|} = \frac{1}{\left|X_1(\omega)X_2^*(\omega)\right|}. \quad (5)$$

GCC is widely used for time difference estimation by maximizing the correlation function to obtain the delay time between two microphones. The cross-correlation function is usually calculated by the inverse FFT of the reciprocal power spectrum. GCC-PHAT can eliminate the effect of amplitude in this way, leaving only the phase.

$$GCC_{ij}(t,\tau) = F_{f\to\tau}^{-1} \frac{X_i(f,t)X_j^*(f,t)}{\left|X_i(f,t)X_j^*(f,t)\right|}. \quad (6)$$

$GCC_{ij}(t,\tau)$ is the inverse Fourier transform from $f$ to $\tau$. $X_i(f,t)$ is the fast Fourier change of the $i$ th mic signal and * represents the conjugate time difference between the two mic signals. The conjugate time difference can be estimated by maximizing the $\tau$ of GCC. However, this estimation is usually unstable, especially in high reverberation and low signal-to-noise environments, and cannot be used directly for

multiple sources. However $GCC_{ij}(t,\tau)$ contains all time-delayed signals, which are usually short-time smooth. $GCC_{ij}(t,\tau)$ is considered as a GCC spectrogram. $\tau$ Corresponds to the number of melband filters.

# 3 Network Structure

## 3.1 Temporal Component

A spatiotemporal CNN [15] is used to model the multi-label recognition problem. RNNs with LSTMs use a pick-and-pass mechanism to implicitly understand how the underlying states inside transition. Although they perform well, they are hard-to-interpret black box models. In contrast, the temporal part of spatiotemporal CNNs explicitly understands how potential states transition and is easy to interpret and visualize. And better results are obtained compared to RNNs.

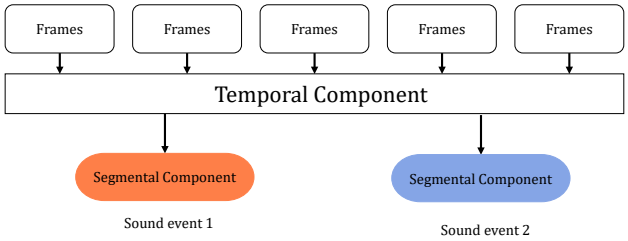


**Figure 4.** Segmental spatiotemporal CNN

For audio $T$, let $\{h_t\}_{t=1}^{T}$ be a series of features and $y_t \in \{1,...,C\}$ be the audio frame label for time frame $t$. The temporal filter weights $W = \{W_1,...,W_{F_e}\}$ shared between the frames and the deviation $b = \{b_1,...,b_{F_e}\}$. The $i$ th filter result at time $t$ is given by the 1D convolution between the spatial feature $h$ and the temporal filter using the ReLU nonlinearity.

$$a_{t,j} = RELU(\sum_{t'=1}^{d} W_{j,t'} h_{t+d-t'} + b_j). \quad (7)$$

The fractional vector $s_t \in \mathbb{R}^C$ is $sigmoid$ function of the fully connected layer.

$$s_t = sigmoid(W \cdot a_t + b). \quad (8)$$

It is shown in Figure 4. A clear identification of the situation will be made for each time frame. The input is the audio features of each frame and the output is the sound events of each frame. Figure 4 represents the complete temporal convolution model.

## 3.2 Dilated Temporal Convolutional Network

Temporal convolutional network [16] (TCN), mainly consists of causal convolution and dilated convolution. TCN uses causal convolution, for the value at moment t in the previous layer, t can only depend on the value at moment t and before in the next layer, so causal convolution cannot see future data and is a one-way structure; dilated convolution enables the convolutional network to see farther and is no longer limited by the size of the convolutional kernel, which allows the CNN to handle longer time series with certain parameters. In contrast, standard convolution is used to obtain a larger perceptual field by adding pooling layers, however, there is some information loss in this way. Dilated convolution is to inject dilated holes into the standard convolution to increase the perceptual field. Dilated convolution has an additional hyperparameter dilation rate, which refers to the number of intervals in the kernel (the standard CNN dilation rate is equal to 1). In this paper, the dilation rate $= \{2^0, 2^1, ..., 2^9\}$.

Figure 5 shows an overview of the TCN architecture. Normally, TCNs utilize causal convolutions, where the output at time t depends solely on the current and past elements. This can be achieved by zero paddings.

However, to mimic the bidirectional RNNs' use of future knowledge, we modify all convolutions within the TCN block to be non-causal.

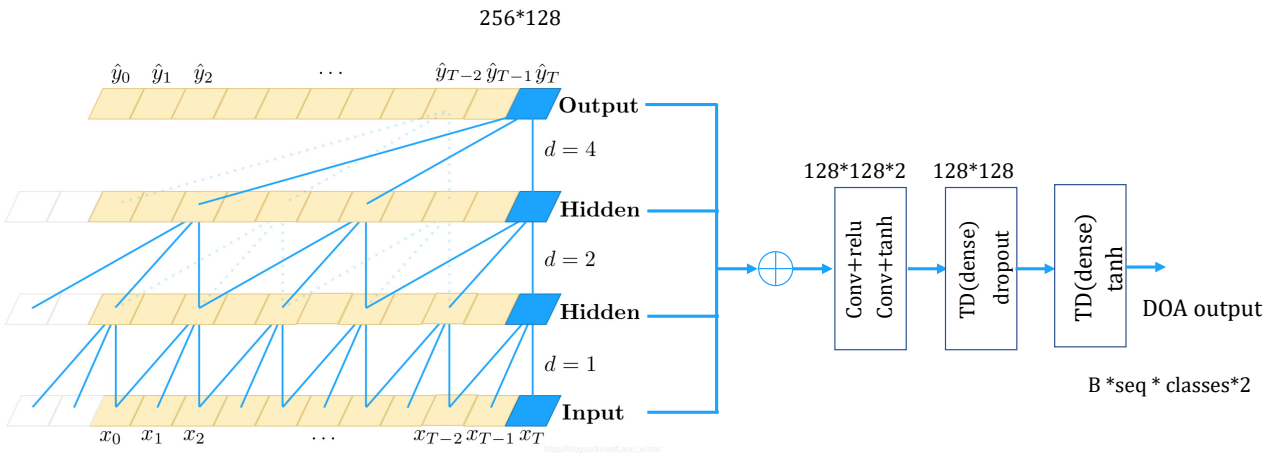


**Figure 5.** TCN architecture in DOA branch

A TCN is used to predict the DOA part. The TCN uses the magnitude spectrum to predict the time- frequen-cy mask, which produces a mask [17] on the time-frequency band. Recurring causal convolution units can retain longer history information [18], which is often required when processing sequential data such as speech. Predicting the next time-difference value while making the previous time-difference output available helps to avoid spurious noise and interference-generated peaks while keeping the time-difference values smooth.

### 3.3 Masked-based DOA

The masking process is shown in Figure 6, where the time-difference space input feature $R_{ii'}(\tau,t,b)$ is obtained by multiplying the feature $R_{ii'}(\tau,t,k)$ of the cross-correlation feature GCC-PHAT with the weights $\eta_{ii'}(t,k)$ obtained from the mel filters.

The phase components of the signal STFT coefficients are given in the form of a matrix of size $M \times K$, where M and K are the number of microphone and frequency subbands, respectively.

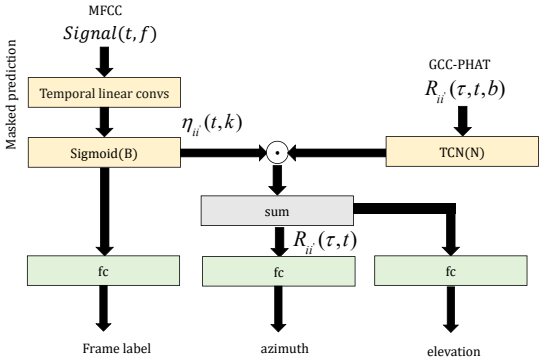


**Figure 6.** Masking process

Local filters of size $2\times1$ are used throughout the convolution process, which allows each filter to learn from the phase correlation between neighboring micro-phones in each frequency subband separately. This is due to the disjoint activity of the speech signal to be used for localization.

The learning-based Boda orientation methods have good generalization capabilities at different levels of reverberation and noise. They aim to enable the system to understand the connection between input features and DOA. There have been a series of studies using deep neural networks to solve DOA problems [19-25]. It is comparable to parametric methods. However, these neural network-based methods are mainly based on static signal sources. In addition to spectrum-based features, GCC-based features that can effectively provide time difference information are also used as input features In order to further improve the learning-based methods, more practical sources need to be considered.

## 4 Method

For a specific source array setup, the corresponding different DOAs are concatenated along the time axis to form two multi-channel signals for training.

During the training process, there are two branches, the SED branch and the DOA branch, as shown in Figure 7. Firstly, the features of shape are $C\times T\times F$ fed into the SED branch. $C$ denotes the number of feature mapping, $T$ denotes the size of time unit, and $F$ denotes the frequency bins. the whole network extracts features to make four groups of 2D residual blocks, each residual block consists of two 2D Convs with Receptive Field of $3\times3$, step size of $1\times1$,padding size of $1\times1$, and dilated rate of 2 to replace the effect of pool and reduce information loss. The last layer can perform feature filtering based on the input features, or do feature mapping for the last layer to learn the information between each channel, after which there is an $2\times2$ average pool for flatten operation. After the above layers, the data is fed to the global pooling layer by the global data shape can be expressed as $C_{out}\times T/16\times F/16$ , and the data is reshaped as $C_{out}\times T/16$ , and fed to the DOA branch and SED branch, respectively.

The process of SED branch is different from the process of DOA in terms of the audio signal features required, which needs to be decoupled. And reasonably choose the appropriate extraction method and classification method for the features. For the multi-label problem, its timing signal timing correlation is weak and focuses on a completely different situation compared with GCC-PHAT features.

Therefore, instead of training the network in one way for the features fed into one network, it will lead to the two networks interfering with each other and the reverse process will affect each other. However, it is necessary to retain the features between the two networks and use a mask approach for feature learning.

In the SED part, linear time convolution is used. The core of SED is to extract the features of the spectral envelope for recognition and analysis, and some of its problems are that its spatial correlation is small and there are not many advanced features needed. Therefore, a relatively simple network structure should be used for identification and analysis, rather than extracting with a complex network. In the ablation experiments, it was found that a shallower network could be used to obtain better recognition results, so linear time convolution was used for the implementation of the classifier. As well as feeding its prediction results into DOA for mask enhancement operation.

After being sent to the SED branch, it is sent to the linear time convolution, keeping its output size, and sent to the fully connected layer $T/16\times N$ where N is the number of events, after which a sigmoid activation function is used to upsample in the time dimension to ensure that the output size is consistent with T. The SED prediction is then obtained by activation thresholding. Binary cross-entropy is used for the multi-label classification task.

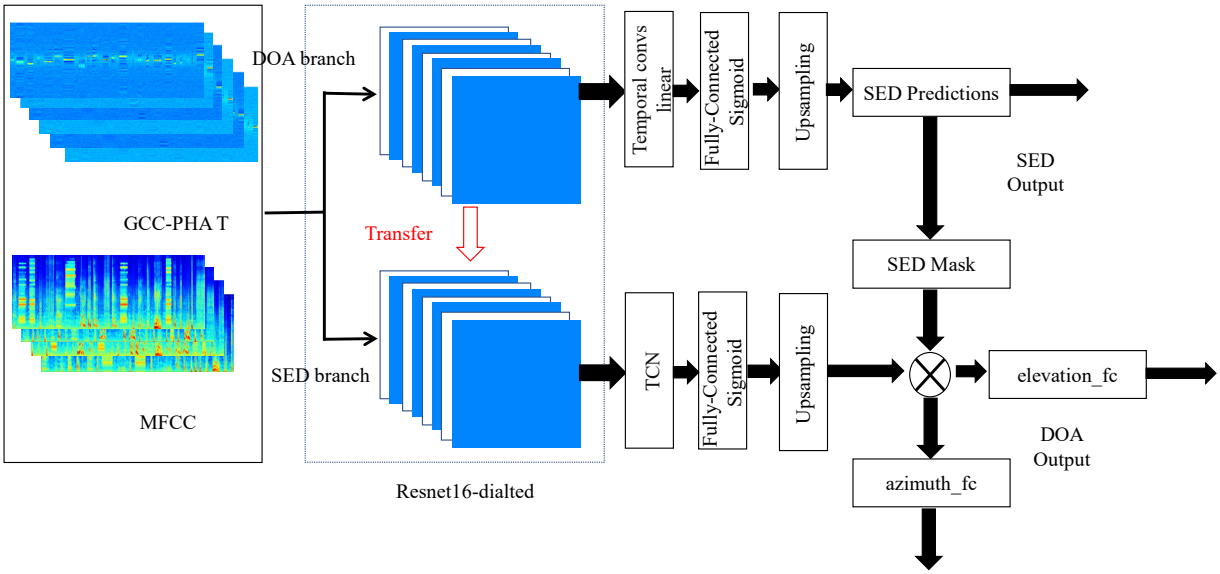


**Figure 7.** Decoupled Seld-TCN network architecture

The GCC-PHAT is then fed into the TCN, which with its dilated convolution can handle a larger range of inputs, where the dilated rate $d=1$ performs the normal convolution. The larger the expansion factor, the larger the perceptual field. The output of the fully connected layer of the DOA branch is a vector of $N\times 2$, i.e., the azimuth and elevation angles of the N events. They are then masked by the SED ground truth during the training period to determine whether the corresponding angles are currently active. Finally, the mean absolute error is chosen as the DOA regression loss.

During the inference process, the SED branch first computes the SED predictions, which are then used as SED masks to obtain the DOA. the reason for constructing this network architecture is to enhance the representation capability of a single network so that each branch is responsible for only one task, while the DOA branch can still absorb the benefits from the SED.

# 5 Metrics

## 5.1 Loss

SELD uses a binary cross entropy binary crossentropy loss function in calculating the SED loss, which calculates the error on each time frame, as shown in the following equation.

$$Loss_{SED}=-\frac{1}{N}\sum_{i=1}^{m}\begin{matrix} y_i\cdot\log(\frac{1}{1+\exp(-\hat{y}_i)}) \\ +(1-y_i)\cdot\log(\frac{\exp(-\hat{y}_i)}{1+\exp(-\hat{y}_i)}) \end{matrix}. \quad (9)$$

Where $N$ denotes the number of categories, $y_i$ denotes the true label and $\hat{y}_i$ denotes the output value of the unactivated network. $m$ denotes the vector length in the labels.

To calculate the DOA loss, the loss of the sampled MSE, the mean of the squared sum of the difference between the predicted target value and the predicted value is predicted. The DOA loss is divided into elevation loss and azimuth loss formulas as follows.

$$Loss_{DOA}=\frac{1}{n}(\sum_{i=1}^{n}(f_{azi}(x)-y_{azi})^2+\sum_{i=1}^{n}(f_{ele}(x)-y_{ele})^2). \quad (10)$$

The joint loss is expressed as, where $\alpha=\frac{1}{50}$:

$$Loss=\alpha\cdot Loss_{SED}+(1-\alpha)\cdot Loss_{DOA}. \quad (11)$$

The SED branch and DOA losses are retained to represent the performance of each component.

## 5.2 Hyperparameters

To test the proposed method. We evaluate the results by comparing the system output with the annotated reference. The sampling rate fs is set to 44.1 KHz, nfft is set to 512 points, window size is set to 512, hop size to 256. frame size is $fs/hop\_len$, frame length is 512, epoch is set to 1000, batch is set to 64, and added noise size is 30 db. The network model is optimized by Adam optimizer with a learning rate of 0.001 and loss weights accounting for ratio [1.,50.].

## 5.3 Dataset

The development set consists of 400 one-minute long recordings, sampled at 48,000 Hz, divided into four cross-validation splits of 100 recordings each. The evaluation set consists of 100 one-minute recordings. These recordings were synthesized using spatial room impulse responses (IR) collected from five indoor locations with 504 unique azimuth- elevation-distance combinations.

Real impulse response synthesis was used for the dataset, among others. The natural ambient noise collected at the IR recording locations was added to the synthesized recordings with an average SNR of 30 dB for the sound events.

To evaluate the proposed method in real life, we used TUT Sound Events 2019. This proprietary dataset contains real recordings from 11 different scenes.

The Hop size was set to 20 ms, the FFT size was set to 2048, each minibatch contained 16 feature sequences, and the length of the feature sequences was set to 128. The number of mel-band filters and the delay of the GCC-PHAT was set to 64. For 4 signal channels, the network had a maximum of 10 signal input channels. The audio clips were re-segmented with a fixed length of 2 seconds and an overlap of 1 second for training. The metrics [26] used to evaluate the model include: F-score, SED error rate (ER), DOA error, and frame recall. Based on these four metrics, an overall evaluation of the model performance is formed.

The DOA error is the average angular error between the predicted DOA and the referenced DOA. For a time frame of length $T$ in the record, let $DOA_R^t$ as the actual value of the time frame $t$, $DOA_E^t$ as the reference value. The DOA error can be expressed as

$$DOA_{error} = \frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathrm{H}(DOA_R^t, DOA_E^t). \quad (12)$$

Where $D_E^t$ denotes the number of DOAs in the $i$ th frame and $H$ is the Hungarian algorithm that matches the individual predicted DOAs with the corresponding actual DOAs. The Hungarian algorithm estimates the error between the individual predicted and actual DOAs by using the central angle between them.

Where DOA is represented by azimuth $\phi_R \in [-\pi, \pi)$ and elevation angles $\lambda_R \in [\frac{-\pi}{2}, \frac{\pi}{2})$, and DOA is denoted by $(\phi_R, \lambda_R)$.

To evaluate the number of time frames in which the predicted DOA differs from the reference DOA, a second metric is used: frame recall.

$$Frame\_recall = \frac{\sum_{t=1}^{T} l(D_R^t = D_E^t)}{T}. \quad (13)$$

Where $D_R^t$ is the number of DOAs in the $t$ th frame $DOA_R^t$, $l()$ is the indicator function, and if $D_R^t = D_E^t$, outputs is 1 and 0 otherwise.

# 6 Experiments

From the results of the two stages in Table 1, we see that the F-score decreases and the DOA error improves after a synthetic mixture of noise and impulse response at 30 db. Based on this, we propose a new training strategy that tries to obtain better results in both SED and DOA branches. This is achieved by applying the hybrid data enhancement technique only on the DOA branch during the training period. Comparing the two phases and the improved results in Table 1 and Table 3, it can be seen that all four evaluation metrics improved in both phases.

**Table 1.** Results for TUT Sound Events 2019 with no MixUp

| | ER | FScore | DE | FR |
|---|---|---|---|---|
| Baseline | 0.350 | 0.800 | 30.8 | 0.854 |
| Twostage | 0.166 | 0.908 | 9.85 | 0.863 |
| Seldnet | 0.213 | 0.879 | 11.3 | 0.847 |
| TS (res) | 0.147 | 0.924 | 9.4 | 0.881 |
| SELDTCN | 0.12 | 0.91 | 10.5 | 0.868 |
| Ours | 0.091 | 0.938 | 9.22 | 0.894 |

It is clear that SELD-TCN is vagueness in event detection than two stage strategy due to its complex network structure in the SED branch. However, the DOA error decreases significantly, which can indicate that the accuracy of TCN in localization has been improved significantly. It indicates that TCN can achieve significa-nt localization accuracy. The extraction using resnet16 for the two stage strategy network can obtain 2% improvement in event detection accuracy and about 0.4 improvement in localization accuracy.

**Table 2.** Results for mixup TUT sound events 2019

| | ER | FScore | DE | FR |
|---|---|---|---|---|
| Baseline | - | - | - | - |
| Twostage | 0.194 | 0.888 | 8.901 | 0.847 |
| Seldnet | - | - | - | - |
| TS (res) | 0.16 | 0.87 | 8.9 | 0.881 |
| SELDTCN | 0.11 | 0.903 | 7.2 | 0.825 |
| Ours | 0.063 | 0.92 | 6.12 | 0.897 |

In the Table 2, after mixup, it was found that the SED branch of mixup enhancement did not have a significant or even a decline, but the DOA branch showed a significant increase in vagueness. The analysis of this process is thought to be that the process of its mixing with natural noise plays a role of feature enhancement, so that there is no effect in the subsequent mixup, or even a significant decrease.

The real recordings were synthesized by convolving isolated real sound events with real impulse responses collected at different spatial locations in the room. There were no time-overlapping sources O1, a maximum of two time-overlapping sources O2, and a maximum of three time-overlapping sources O3.

When adjusting the input sequence length, it was observed that a sequence of 256 frames gave the best score for the reverberant dataset and 512 frames gave the best results for the dataset.

**Table 3**. Evaluation in the three overlapping sources

| | O1 | O2 | O3 |
|---|---|---|---|
| DE | 6.73 | 8.8 | 30.7 |
| F | 97.7 | 89 | 85.6 |
| FR | 98.2 | 89.4 | 78.8 |
| ER | 0.04 | 0.16 | 0.29 |

It can be seen from Table 3 that the higher the number of overlapping sources, the lower the performance for multi-source tracking. DOA error is found to improve significantly with the use of TCN on the dataset, but at the cost of lower frame recall, a significant decrease in frame recall occurs. The recall metric decreases significantly for the reverberation and moving source scenes datasets.

It can also be seen that the error increases rapidly in the case of three temporally overlapping sources. It indicates that the GCC method performs quite poorly for the localization error of multiple overlapping sources.

The Figure 8 and Figure 9 shows the predictions and corresponding references for each 1000-frame test sequence in the O2 dataset and O3 dataset. Each sound category has a unique color on the subplot. We find that the detected sound events are accurate compared to the reference. It can be seen that the DOA predicted values vary around the reference trajectory with very little deviation. This indicates that two overlapping and moving sources can be successfully tracked and identified.

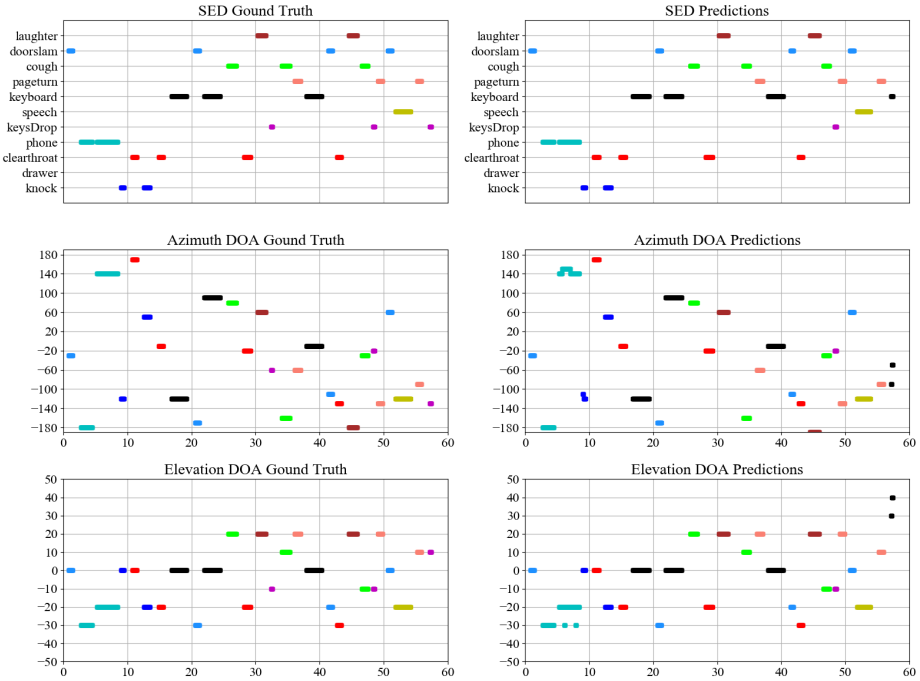


**Figure 8.** Predicted and reference comparison of elevation and azimuth angles for Two overlapping sources

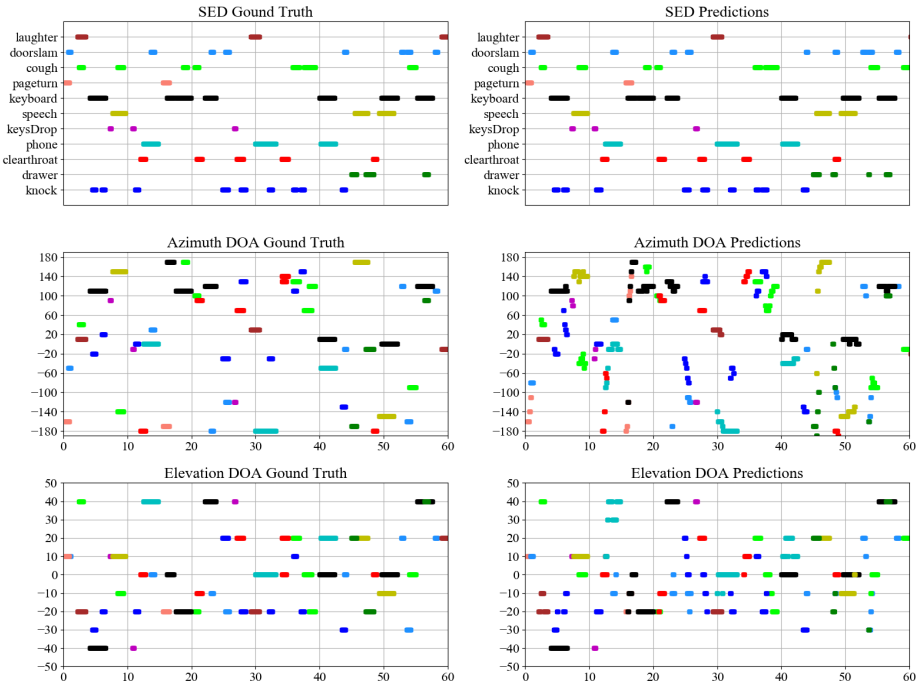


**Figure 9.** Predicted and reference comparison of elevation and azimuth angles for Three overlapping sources

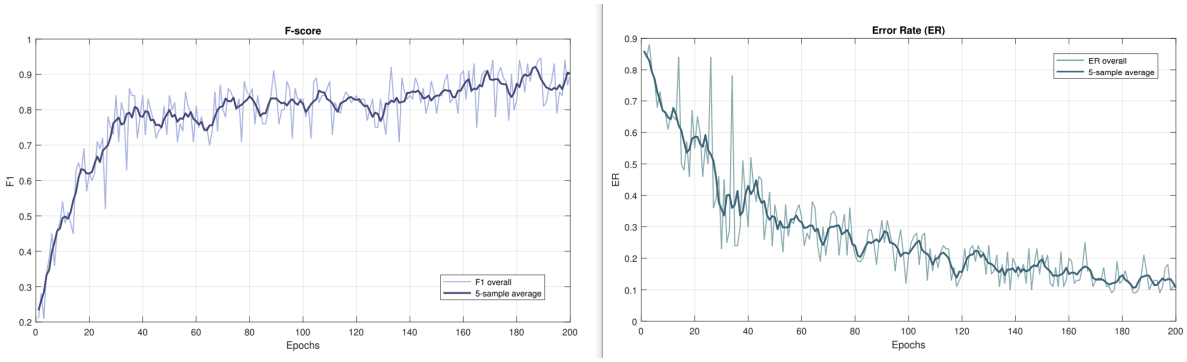


**Figure 10.** High variance leads to large error rate

The more overlapping sources the more significant the performance degradation is. From the visualization, it is observed that there are problems such as spurious DOA tracking and high variance, which leads to large DOA errors in the dataset.

The more overlapping sources the more significant the performance degradation is. From the visualization, it is observed that there are problems such as spurious DOA tracking and high variance, which leads to large DOA errors in the dataset. Show in the Figure 10.

## 7 Conclusion

In the research, regarding sound temporal detection and localization as a joint task has the problem of performance degradation caused by joint training. Sound event detection is sensitive to the network depth, and the increase of the network depth will lead to a decrease in the event detection ability and an increase in the event location ability. However, Events localization has deeper requirements for the network depth. It is necessary to analyze the performance of the two tasks on a single task, reduce the coupling degree, and improve the performance. So it is necessary to absorb the advantages of Seld-TCN and perform decoupling. On this basis, dilated convolution is introduced to improve the feature extraction ability of the model, and causal convolution is used in the SED part to improve the parallel ability. The DOAE branch uses the SED ground truth to train the DOAE mask, thereby realizing the enhancement of the weak feature representation branch to the strong feature representation branch. Although the experiment has achieved good results.

However, it can be seen that GCC-PHAT has corresponding problems.it cannot adapt to the work of ternary and above overlapping sources, resulting in a surge in error rate and poor positioning effect. Therefore, it is necessary to find a more suitable expression of audio position information. Secondly, the result of high variance indicates that the feature extraction ability of the model is relatively weak, and it is necessary to find a more powerful feature extraction method to replace the existing feature expression to further improve the accuracy. In the future, we will continue to conduct in-depth research on these aspects of the problem.

## References

[1] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, T. Virtanen, Sound event detection in the DCASE 2017 challenge, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 6, pp. 992-1006, June, 2019.

[2] T. Hirvonen, Classification of spatial audio location and content using convolutional neural networks, *Audio Engineering Society Convention 138. Audio Engineering Society*, Warsaw, Poland, 2015, pp. 9294.

[3] C. J. Grobler, C. P. Kruger, B. J. Silva, G. P. Hancke, Sound based localization and identification in industrial environments, *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, Beijing, China, 2017, pp. 6119-6124.

[4] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, T. Virtanen, Joint measurement of localization and detection of sound events, *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 333-337.

[5] H. Cordourier, P. Lopez Meyer, J. Huang, J. Del Hoyo Ontiveros, H. Lu, GCC-PHAT cross-correlation audio features for simultaneous sound event localization and detection (SELD) on multiple rooms, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, NY, USA, 2019, pp. 55-58.

[6] B. Kwon, Y. Park, Y. Park, Analysis of the GCC-PHAT technique for multiple sources, *ICCAS 2010*, Gyeonggi-do, Korea, 2010, pp. 2070-2073.

[7] S. Adavanne, A. Politis, T. Virtanen, Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features, *2018 international joint conference on neural networks (IJCNN)*, Rio de Janeiro, Brazil, 2018, pp. 1-7.

[8] S. Adavanne, P. Pertilä, T. Virtanen, Sound event detection using spatial features and convolutional recurrent neural network, *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 771-775.

[9] S. Adavanne, A. Politis, J. Nikunen, 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, March, 2019.
[10] G. Parascandolo, H. Huttunen, T. Virtanen, Recurrent neural networks for polyphonic sound event detection in real life recordings, *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Shanghai, China, 2016, pp. 6440-6444.
[11] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 6, pp. 1291-1303, June, 2017.
[12] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, B. Yang, SELD-TCN: Sound event localization & detection via temporal convolutional networks, *2020 28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands, 2021, pp. 16-20.
[13] F. Zheng, G. Zhang, Z. Song, Comparison of different implementations of MFCC, *Journal of Computer science and Technology*, Vol. 16, No. 6, pp. 582-589, November, 2001.
[14] T. Rodemann, G. Ince, F. Joublin, C. Goerick, Using binaural and spectral cues for azimuth and elevation localization, *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, 2008, pp. 2185-2190.
[15] C. Lea, A. Reiter, R. Vidal, G. D. Hager, Segmental spatiotemporal cnns for fine-grained action segmentation, *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 36-52.
[16] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, Temporal convolutional networks for action segmentation and detection, *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, 2017, pp. 1003-1012.
[17] J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, *IEEE Signal processing letters*, Vol. 24, No. 3, pp. 279-283, March, 2017.
[18] G. Singh, F. Cuzzolin, Recurrent convolutions for causal 3d cnns, *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Korea , 2019, pp. 1-10.
[19] M. Ma, T. May, G. J. Brown, Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 12, pp. 2444-2453, December, 2017.
[20] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, H. Li, A learning-based approach to direction of arrival estimation in noisy and reverberant environments, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015, pp. 2814-2818.
[21] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, A neural network based algorithm for speaker localization in a multi-room environment, *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Salerno, Italy, 2016, pp. 1-6.
[22] S. Chakrabarty, E. A. P. Habets, Broadband DOA estimation using convolutional neural networks trained with noise signals, *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2017, pp. 136-140.
[23] W. He, P. Motlicek, J. M. Odobez, Deep neural networks for multiple speaker detection and localization, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018, pp. 74-79.
[24] S. Adavanne, A Politis, T Virtanen, Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network, *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 1462-1466.
[25] Y. Luo, N. Mesgarani, Tasnet: time-domain audio separation network for real-time, single-channel speech separation, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 696-700.
[26] A. Mesaros, T. Heittola, T. Virtanen, Metrics for polyphonic sound event detection, *Applied Sciences*, Vol. 6, No. 6, Article No. 162, June, 2016.

## Biographies

**Shen Song** received the B.E. degree in Computer Science and Technology from the Wuhan Institute of Technology, Wuhan, China, in 2018.He is currently pursuing the M.S. degree in computer technology with Wuhan Polytechnic University, Wuhan, China. His research interest includes artificial intelligence Technology and its application.

**Cong Zhang** received the bachelor's degree in automation engineering from the Huazhong University of Science and Technology, in 1993, the master's degree in computer application technology from the Wuhan University of Technology, in 1999, and the Ph.D. degree in computer application technology from Wuhan University, in 2010. He is currently a Professor with the School of electrical and Electronic Engineering, Wuhan Polytechnic University. His research interests include multimedia signal processing, multimedia communication system theory and application, and pattern recognition.

**Xinyuan You** received the B.E. degree in Internet of Things Engineering from Luoyang Normal University, Luoyang, China, in 2021.She is currently pursuing the M.S. degree in computer technology with Wuhan Polytechnic University, Wuhan, China. Her research interest includes artificial intelligence Technology and its application.