L-TCN Speech Separation Algorithm for Effectively Acquisition IPD Information Based on Attention in Reverberation Environment

Xiyu Song[1], Zhengyi An[1], Shiqi Wang[1], Fangzhi Yao[1], and Mei Wang[1,2(✉)]

[1] Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin University of Electronic Technology, Guilin 541004, China
mwang@glut.edu.cn

[2] School of Information Science and Engineering, Guilin University of Technology, Guilin 541006, China

Abstract. Speech separation aims to separate a target speaker's speech from mixed speech. However, various noises and reverberations in real life make separation difficult. To solve this problem, a multi-channel microphone array is introduced to extract the spatial information of the target speech; however, the number of inter-channel phase differences (IPDs) increases linearly with the square of the number of microphones. Indeed, using all IPDs will impose a massive load on the system; therefore, We use the attention mechanism to effectively acquire IPD information. Moreover, the time convolution network (TCN) exhibits excellent performance in speech separation; however, a large number of parameters of deep dilated convolution results in a huge system burden. In summary, a speech separation method aided by effectively acquisition IPD information based on attention is proposed for a lightweight time convolution network (L-TCN). Compared with the control experiment, the proposed method reduces the parameters by 90% and doubles the utilization rate of the IPD. Based on the premise of reducing the system load, the short-time objective intelligence (STOI) increases by 0.19 and the scale-invariant signal to distortion ratio (SI-SDR) increases by 6.33.

Keyword: reverberation environment · speech separation · time convolution network

1 Introduction

The proposition of target speech separation originated from the cocktail party effect [1, 2], aimed at separating the target speech from the background interference. Background interference includes background noise, the voices of other people, and the reflection of one's voice through various objects. These interferences have significantly influenced the post-processing of speech (automatic speech recognition (ASR), semantic analysis, simultaneous translation, etc.) [3]. To solve this problem, some speech separation

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 162–178, 2025.
https://doi.org/10.1007/978-3-031-60347-1_14

algorithms have attempted to improve the signal-to-noise ratio (SNR) and intelligibility of speech, such as spectral subtraction [4], Wiener filtering [5], and statistical-based methods [6]. However, in a reverberation environment, the reverberation noise caused by reflection is coherent noise to the target speech, and the separation performance of the above algorithms is poor in the face of coherent noise.

In the reverberation environment, reverberation affects the directivity of the target sound source [7], and the spatial information of the corresponding speech is disturbed. To address this problem, some scholars have introduced information on other modes in the speech separation scene. A. Ephrat utilized the visual features of a speaker's lip information to enhance the effect of speech separation [2]. D. E. King calculated the orientation features through face detection and tracking [8]; however, it is obvious that the orientation information based on vision involves privacy concerns. Although acoustic direction information can compensate for this type of defect, it still has a considerable error compared with visual direction information. Rongzhi Gu introduced the speaker recognition system [9]; however, this method requires a substantial amount of prior information and is limited by the current speaker recognition system. Several scholars have proposed a speech separation model based on microphone arrays [10] that uses mixed speech collected by an array to extract the spatial information of the target speech. The method improves the separation performance under reverberation conditions to a certain degree.

When a microphone array features, the inter-channel phase difference (IPD) is typically used to indicate the position where the sound source reaches each microphone [11, 12], and the number of IPDs increases linearly with the square of the number of microphones. It is foreseeable that we can capture additional spatial information about the target speech by using more IPD; however, this will significantly increase the load on the system. Therefore, this paper proposes an effectively acquisition IPD information based on attention method. The proposed method can capture additional spatial information of the target speech by using additional inter-channel phase differences without increasing the system load to enhance the speech separation performance under reverberation conditions.

Speech separation methods based on deep learning have been proposed with wide application of deep learning [13, 14]. Owing to the good performance of deep learning, speech separation is regarded as a supervision problem. The separation method based on deep learning tends to break away from theoretical modeling, adopting a data-driven approach and using deep learning to separate the target speech from the mixed speech. Compared with traditional speech separation algorithms, the performance of speech separation under supervised learning has been significantly improved [15–17]. The methods based on deep learning primarily employ neural networks to fit the mapping relationship between the mixed speech signal and the target speech and to construct T-F masks so that the mixed speech can be separated according to the corresponding mapping relationship.

Recently, the performance of a time convolution network (TCN) has been comparable to that of a long-term and short-term memory network (LSTM) with fewer network parameters, which makes it ideal for speech separation [18–21]. The network was expanded and convolved in one-dimensional form. Following several iterations, the

receptive field of the output layer was significantly increased, and the output information contained a sufficiently wide range of input information. Compared with LSTM, the parameters of the network are relatively small; however, in the iterative process, a large number of parameters are still required to map the relationship between the input and output. The burden of network parameters is considerably high, which seriously affects deployment of the model and post-processing of speech.

To reduce the burden of parameters, this study proposes a lightweight time convolution network (L-TCN) that can reduce the number of parameters and computation by 90%. First, this method does not use deep dilated convolution to double feature dimensions, instead it uses conventional dilated convolution to compress features. Inspired by the gated linear unit (GLU) [21] and LSTM [22], the gated weighted branch was added to the dilated convolution under the activation of the gated mechanism. The ablation experiment proved that the performance of the L-TCN was as good as or even better than that of the control experiment (TCN), on the premise of reducing the parameters and computation.

The primary contributions of this paper are as follows: On the one hand, by utilizing the scoring mechanism of the attention mechanism to extract the spatial information of the phase difference between channels, the proposed method still exhibits excellent performance even without increasing the number of IPDs, which is equivalent to advancing the utilization rate of IPDs to utilize additional spatial information without increasing the system load, so as to achieve a better aliasing and speech separation effect. On the other hand, the L-TCN assisted by the attention mechanism designed in this study can fully exploit the gating function of the weighted branches, reduce the network parameters and computation, and simultaneously improve the effectiveness and reliability of the speech separation task in a reverberation environment. The experimental results reveal that compared with the traditional TCN model, the speech separation framework proposed in this study exhibits better performance in terms of accuracy, parameters, computation, short-time objective intelligence (STOI), and scale-invariant signal-to-distortion ratio (SI-SDR).

The remainder of this paper is organized as follows. The signal model and process of TCN dilated convolution are presented in Section II. Section III, present the methods: Section IV presents the experimental results. Finally, conclusions are presented in Section V.

2 Related Work

In this section, the signal model of mixed speech is introduced. Following that, the acoustic features used in the experiment are introduced, and finally the dilated convolution process of the TCN is introduced.

We assume that the target speech is $\hat{s}(t)$, that is, the reference signal, the noise is $n(t)$; the room impulse response is h, and the inter-channel interference and other noises are n_0. Accordingly, the mixed speech model is:

$$y(t) = h(\hat{s}(t) + n(t)) + n_0 \tag{1}$$

2.1 A Signal Feature

Fourier transform was used to obtain the spectrum feature of the mixed speech signal $y(t)$, and the time-domain mixed signal is converted into a complex frequency domain:

$$y(t) \xrightarrow{STFT} Y(t,f) \tag{2}$$

Following the short-time Fourier transform (STFT), we constructed a logarithmic power spectrum (LPS) feature based on the corresponding spectrum feature. We chose the LPS of the mixed speech received by the channel 0 microphone as the spectrum feature, and we obtained the 257-dimensional LPS feature by using the STFT of 512 points.

The inter-channel phase difference is calculated from the phase difference value of the complex frequency domain between channels, which can reflect subtle changes in the direction of arrival (DOA) of the sound source and construct the spatial features of sound:

$$IPD^m(t,f) = \angle Y^{m_1}(t,f) - \angle Y^{m_2}(t,f) \tag{3}$$

where m denotes the number of microphone pairs, and m_1 and m_2 denote two microphones in the m-th microphone pair. Here, the number of microphone pairs M evidently increases linearly with the square of the number of microphone channels N:

$$m = \frac{N(N-1)}{2} \ (N \geq 2) \tag{4}$$

In the experiment, to obtain additional spatial information about the target speech, it is crucial to use as many IPDs $M \leq m$ as possible. However, constrained by the model and algorithm, it is necessary to reduce the dimension of spatial features, which makes it necessary to reduce the number of IPDs $M \geq 0$, and connect them in series in the form of $IPD^m(t,f)$ to become IPD features:

$$IPD = \underbrace{IPD^1(t,f), IPD^2(t,f), ..., IPD^M(t,f)}_{M} \tag{5}$$

In the case of this contradiction, other scholars often only use IPD with different microphone distances to characterize the spatial features of the signals [23]. Simultaneously, the visual direction feature (DF) is introduced to compensate for the lack of information in spatial features.

Given the direction of the target speaker, the directional features of the target speech can be obtained accordingly to clarify the spatial features of the target speech. It is assumed that the target speech is from the speaker in the θ direction, then $df_\theta(t,f)$ will be close to one; otherwise, it is close to zero, and the DF feature is calculated according to the speaker's direction and expressed as:

$$df_\theta(t,f) = \sum_{m=1}^{M} \left\langle e^{TPD^m(f,\theta_t)}, e^{IPD^m(t,f)} \right\rangle \tag{6}$$

$$TPD^m(\theta_t, f) = \frac{2\pi f \Delta_m \cos\theta_t}{f_s c} \tag{7}$$

The DF feature represents the cosine distance between the steering vector and IPD, where $e^{(\cdot)} = \begin{bmatrix} \cos(\cdot) \\ \sin(\cdot) \end{bmatrix}$,$TPD^m(f, \theta_t)$ target-related phase difference corresponds to the phase delay of a plane wave (frequency is f), and the distance between the m-th pair of microphones, c denotes the speed of sound, and f_s represents the sampling rate. We believe that the speaker movement is a small probability event during the experiment, so $\theta_t = 0$ during the experiment.

2.2 B Time Convolution Network

In speech-related tasks, such as speech recognition or separation, it is crucial to capture hidden information between voices, which requires the correlation of local and contextual information. Dilated convolution effectively increases the receptive field by aggregating different scales, as depicted in Fig. 1, considering a one-dimensional dilated convolution as an example [23].

$$O = \sum_{y+dr=p} F(I)k(r) \tag{8}$$

where $F(I)$ represents the input, and $k(r)$ represents the convolution kernel of size $2r+1$. Additionally, d indicates the dilation rate. When $d = 1$, it degenerates into a regular convolution.

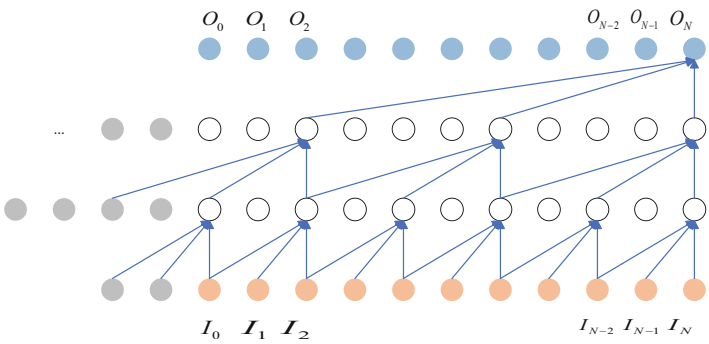

Fig. 1. Calculation diagram of the depth dilated convolution tensor.

2.3 C Long-Short Term Memory and Gating Linear Unit

To deal with the problem that RNN will cause gradient explosion and gradient disappearance, variants such as LSTM introduce a gating mechanism to choose whether to

establish long-term memory, whereas GLU introduces a gating branch:

$$y = (x * W + b_0) \odot \sigma(x * V + b_1) \tag{9}$$

where x and y correspond to the input and output, respectively; W, V represent the convolution kernel weight; b_0, b_1 represent paranoia; σ scales the value range to (0, 1), and the mapping relationship between the input and output can be dynamically controlled by using branches.

3 Proposed Method

In this section, first, the overall framework of the system is introduced; second, the process of attention aggregation IPD is introduced Finally, the L-TCN system proposed in this work is introduced.

3.1 A Model Structure

The basic structure of the system, shown in Fig. 2, was obtained according to the principle of masked speech separation based on deep learning.

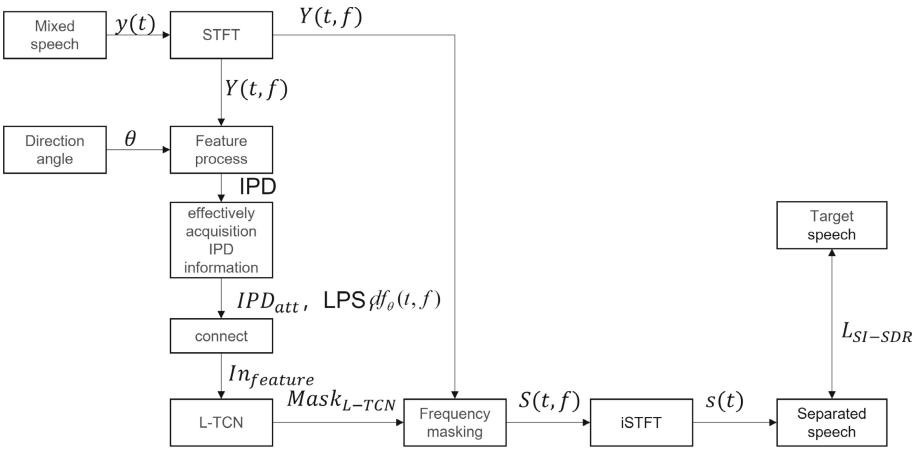

Fig. 2. Basic structure of L-TCN speech separation system by effectively acquisition IPD information based on attention.

First, we obtained complex frequency domain features from multichannel noisy speech by STFT and then obtained three features by feature processing: LPS, IPD, and DF. Following dimension reduction and fusion of IPD features using the attention mechanism, the fused IPD is spliced with LPS and DF to obtain the model input and then sent to the L-TCN to obtain a speech separation mask. The speech separation mask is multiplied by the complex frequency domain feature $Y(t, f)$ to obtain the complex frequency domain feature $S(t, f)$ of the separated speech and obtain the separated speech

$s(t)$ through the inverse short-time Fourier transform (iSTFT). Subsequently, the SI-SDR loss function is used to calculate the loss of the separated speech and the clean target speech, back propagation is performed, and the optimal model is obtained.

3.2 B Attention Mechanism

Note that the wight coefficient is the essence of the mechanism [24], which is obtained by calculating the similarity between the key vector and query vector, and then weighting and summing the value vectors:

$$Attention(Q, K, V) = soft\max(Q * K^{\mathrm{T}}) * V \tag{10}$$

The essence of the attention mechanism is the scoring, weighting, and summing processes. We applied this idea to the aggregation of multiple IPDs, considering different IPDs as inputs, and finally obtained the aggregated IPD_{att}:

$$IPD_{att} = Attention(IPD^1, IPD^2, IPD^3) \tag{11}$$

3.3 C Lightweight Time Convolution Network

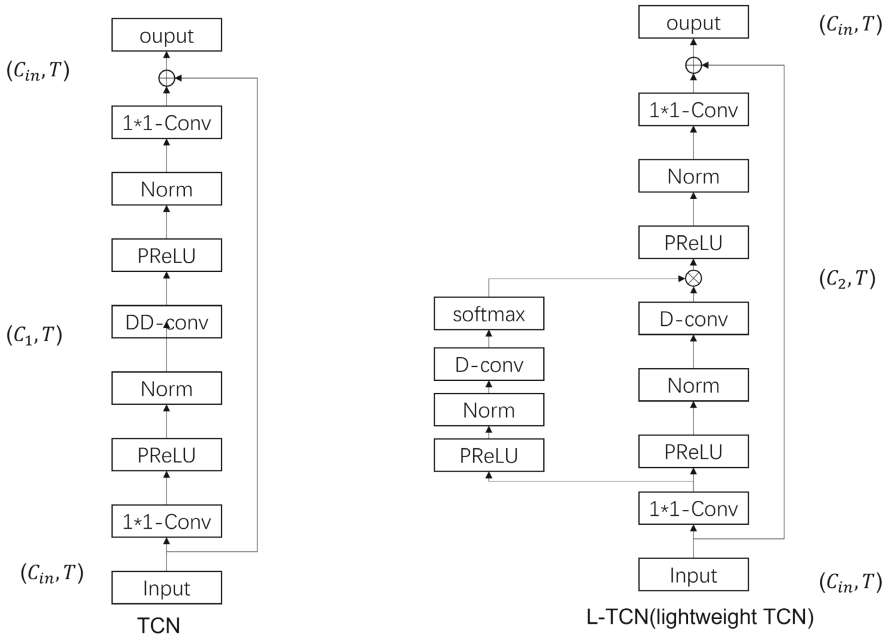

Fig. 3. (a) Deep dilated convolution process of TCN. (b) Conventional dilated convolution process of L-TCN.

TCN is widely utilized in speech separation tasks. The TCN exhibits better performance than LSTM in time-series modeling [25]. Simultaneously, parallel convolution was adopted, which significantly reduced the reasoning time. Figure 3(a) illustrates the dilated convolution process in the TCN, which is primarily composed of three parts. Two one-dimensional convolutions are employed for the input and output, and deep dilated convolution is used to increase the receptive field. Simultaneously, a residual connection is introduced. The input and output dimensions of the system are the same, that is $C_{in} = C_{out}$. Figure 3(b) depicts the convolution process of a lightweight TCN. Inspired by GLU [26], besides the main dilated volume integral branch, a secondary dilated volume integral branch is introduced, which becomes a weighted branch. The weighted branch is similar to the main branch of the structure. Softmax is introduced at the end of the weighted branch to weight the output of the dilated convolution. Thus, information can be spread on a better gradient.

In terms of the parameters, the TCN uses a deep dilated convolution. The dimension of the parameter is mapped from (C_{in}, T) to (C_1, T), and the convolution kernel is K_1, often $C_{in} >> K_1$, so the parameter quantity is:

$$T_1 = 2C_{in}C_1 + K_1C_1 \approx 2C_{in}C_1 \tag{12}$$

Accordingly, the L-TCN parameter is:

$$T_2 = 2C_{in}C_2 + 2K_2C_2^2 \tag{13}$$

Thus, the parameter ratio is given as:

$$\eta = \frac{T_1}{T_2} \tag{14}$$

In lightweight TCN, the features are no longer doubled to C_1 by deep dilated convolution; instead they are only compressed to low-dimensional C_2 by conventional dilated convolution. The principle behind this is that the frequency spectrum generally exhibits a sparse distribution in the T-F domain, and useful information can be well preserved by low dimensions [27].

In the control experiment, $C_1 = 512$, in the experiment, when $C_2 = 32, 64, 128$, and $K_1 = K_2 = 3$, and $\eta = 11.7, 4.60$ and 1.34, respectively. Therefore, the parameters of the system using the L-TCN can be drastically reduced, and the experimental results are superior to those of the TCN.

3.4 D Method for Effectively Acquire IPD Information Based on Attention

In this study, L-TCN is utilized to learn to fit the mask, and the complex frequency map of the separated speech is obtained by multiplying it with the input speech. The separated target speech is obtained by the inverse Fourier transform, as follows:

$$S(t,f) = Mask_{L-TCN} * Y(t,f) \tag{15}$$

$$S(t,f) \xrightarrow{iSTFT} s(t) \tag{16}$$

Among them, the complex frequency diagram of $S(t,f)$ separated speech and the mask obtained by aggregating the features of L-TCN $Mask_{L-TCN}$ in the second part via the attention pool are:

$$Mask_{L-TCN} = L - TCN(In_{feature}) \tag{17}$$

where $In_{feature}$ was obtained by IPD through attention-weighted splicing, as depicted in Fig. 4.

$$In_{feature} = cat(IPD_{att}, LPS, df_{\theta}(t,f)) \tag{18}$$

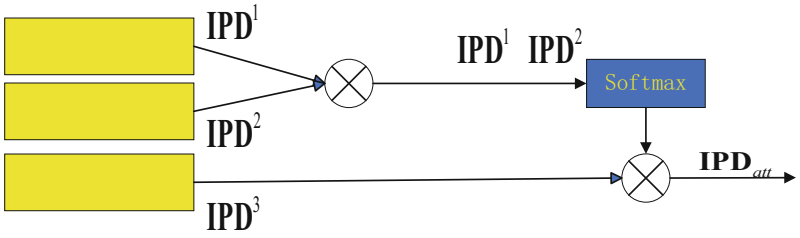

Fig. 4. Method for effectively acquire IPD information based on attention.

The number of IPDs increases linearly with the square of the number of microphones. It is foreseeable that the more IPDs used, the better the separation performance. However, in practical applications, only limited IPDs can be used, primarily owing to the limitations of the network model, computing resources, and data volume. The attention-weighted IPD proposed in this work can reduce the input dimension without increasing the number of IPDs or reducing the input dimension. The corresponding experimental results proved that the proposed method is effective.

4 Experimental Results and Analysis

This section first introduces the construction process of the dataset in the experiment, then the loss function used in the experiment, and finally describes the experimental process.

4.1 A Building Datasets

In the experiment, two identical rooms A and B with length, width, and height of 9, 7.5, and 3.5, respectively, were simulated using the mirror principle. The reverberation reflection coefficient of room A was $rt_{60} = 0.25$, and room B had no echo reflections. A microphone array with channel $N = 4$ was generated in the center of the room. The TIMIT dataset [16] was used for speech in the experiment. In room A, the speaker and noise source positions were randomly generated in the room. Speech was played at the speaker's position, and music noise with a fixed SNR was played at the position of the

noise source. Reverberant speech was collected as mixed speech using a four-channel microphone array. In room B, a coordinate was generated at the same position as the speaker in room A, the voice was played at the coordinate, a microphone was generated in the center of the room, and the microphone was used to collect the voice as the target voice.

According to the microphone and speaker positions, the direction angle was generated, and noise (± 15°) was randomly added to generate the estimated direction angle. Finally, the mixed speech and estimated direction angle were input to the constructed model, and the target speech was used as a label.

4.2 B Loss Function

In fact, the training process of deep learning is the process of minimizing the losses of speech and target speech. Therefore, in selecting the loss function, the maximum SI-SDR was selected in this experiment. Currently, this index has replaced the standard SNR[5,44] and is typically used as an evaluation index of source separation. It is defined as:

$$L_{SI-SDR} = 10\log_{10}\frac{\|\alpha\hat{s}\|^2}{\|\alpha\hat{s}-s\|^2} \tag{19}$$

$$\alpha = \frac{s^{\mathrm{T}}\hat{s}}{\|\hat{s}\|^2} = \arg\min_{\alpha}\|\alpha\hat{s}-s\|^2 \tag{20}$$

Equations (19) and (20) show that SI-SDR represents the SNR of the target speech, and the residual noise is defined as $\|\alpha\hat{s}-s\|^2$.

$$L_{SI-SDR} = 10\log_{10}\frac{\left\|\frac{s^{\mathrm{T}}\hat{s}}{\|\hat{s}\|^2}\hat{s}\right\|^2}{\left\|\frac{s^{\mathrm{T}}\hat{s}}{\|\hat{s}\|^2}\hat{s}-s\right\|^2} = 10\log_{10}\frac{s^{\mathrm{T}}\hat{s}}{\hat{s}^{\mathrm{T}}\hat{s}s^{\mathrm{T}}s-s^{\mathrm{T}}\hat{s}} \tag{21}$$

where s and $\hat{s}$ denote the separated speech and target speech, respectively, and the scaling of the target speech s keeps the amplitude invariance of SI-SDR to the separated speech because speech intelligibility is largely insensitive to amplitude.

Minimizing L_{SI-SDR} maximizes the similarity between the target speech and the separated speech while generating a minimum amount of energy. Similar to the SNR, the SI-SDR is measured in decibels (dB).

4.3 C Experimental Process

In this experiment, we chose four SNR values to represent four scenarios: −15 dB representing a noisy environment, −5 dB representing a noisy environment, 5 dB representing a quiet environment, and 15 dB representing an extremely quiet environment. Experiments were performed under four scenarios. The experiment was conducted in a PC equipped with a GTX1650 graphics card. Limited by the computing power of the experimental equipment, 1 s of mixed speech was resampled to 16 K and 257 × 63

frequency domain features were obtained through STFT with a window length of 257 Hamming windows and extracted LPS, IPD, and DF. In the experiment, an SI-SDR loss function with a learning rate of 0.005 was used to train the network over 350 stages.

The experiment was divided into 9 groups:

Experiment 1. Three pairs of IPD spliced LPS features are used as the input of TCN speech separation model to verify the influence of phase information difference on speech separation performance, that is, TCN (NO-DF);

Experiment 2. Taking 3 pairs of IPD, LPS features and direction features as the input of TCN speech separation model, the influence of direction features on speech separation performance is verified, that is, TCN, and set as the control group;

Experiment 3. Taking 3 pairs of IPD's attention-weighted representation and DF and LPS feature splicing as the input of TCN speech separation model, the influence of IPD's effective information acquisition method based on attention mechanism on speech separation performance is verified, that is, TCN (n = 3);

Experiment 4. Taking the attention-weighted representation of 6 pairs of IPD and the concatenation of DF and LPS features as the input of TCN speech separation model, the influence of more IPD features on speech separation performance is verified, that is, TCN(N = 6).

Experiment 5: Taking three pairs of IPD, LPS features and direction features as the input of the TCN speech separation model with the middle dimension of 256, the influence of reducing the TCN in the middle dimension on the speech separation performance is verified;

Experiment 6: Taking three pairs of IPD, LPS features and direction features as the input of the lightweight TCN speech separation model with the middle dimension of 32, the influence of lightweight TCN on speech separation performance is verified, that is, TCN (C = 32);

Experiment 7: Taking three pairs of IPD, LPS features and direction features as the input of the lightweight TCN speech separation model with the middle dimension of 64, it is verified that the impact of adding lightweight TCN with the middle dimension on the speech separation performance is TCN (C = 64);

Experiment 8: Taking three pairs of IPD, LPS features and direction features as the input of the lightweight TCN speech separation model with the middle dimension of 128, the influence of increasing the middle dimension of lightweight TCN on speech separation performance is verified, that is, TCN (C = 128);

Experiment 9: Taking 6 pairs of IPD's attention-weighted representation and DF and LPS feature splicing as the input of the lightweight TCN speech separation model with the middle dimension of 64, the influence of the lightweight speech separation model based on attention mechanism on the speech separation performance is verified, that is, L-TCN(C = 64N = 6).

4.4 D Evaluation Criteria

The main application scenario of this experiment is the separation of speech and nonverbal speech, and two evaluation criteria, namely STOI and SI-SDR, were adopted.

The STOI is an objective speech intelligibility estimator that is used to evaluate the performance of most speech separation technologies. The STOI score is defined

between 0 and 1; SI-SDR is a common indicator for evaluating the performance of the SNR. Compared to the SNR, the SI-SDR is insensitive to limiting linear transformations.

Table 1. Comparison of STOI results of nine groups from experimental tests. (0 means mixed speech before separation, The yellow TCN is the control group.)

experiment number	System model	−15dB	−5dB	5dB	15dB
1	TCN/No-DF	0.52	0.70	0.82	0.96
2	TCN	0.63	0.76	0.88	0.96
3	TCN(M = 3)	0.67	0.76	0.88	0.96
4	TCN(M = 6)	0.67	0.83	0.91	0.97
5	TCN(C = 256)	0.55	0.72	0.82	0.96
6	L-TCN(C = 32)	0.74	0.80	0.90	0.96
7	L-TCN(C = 64)	0.78	0.82	0.92	0.97
8	L-TCN(C = 128)	0.78	0.83	0.92	0.97
9	L-TCN(C = 64M = 6)	0.82	0.87	0.95	0.97

Table 2. Comparison of SI-SDR results of nine groups from experimental tests. (0 means mixed speech before separation, The yellow TCN is the control group.)

experiment number	System model	−15dB	−5dB	5dB	15dB
1	TCN/No-DF	−3.36	2.92	6.88	20.40
2	TCN	2.26	6.79	14.09	21.53
3	TCN(M = 3)	4.74	6.87	14.09	21.99
4	TCN(M = 6)	4.76	9.78	14.96	22.13
5	TCN(C = 256)	0.25	3.52	8.39	19.98
6	L-TCN(C = 32)	6.60	8.05	14.48	22.03
7	L-TCN(C = 64)	7.60	9.06	14.89	22.30
8	L-TCN(C = 128)	7.55	9.10	14.89	22.18
9	L-TCN(C = 64M = 6)	8.48	13.76	18.98	23.10

All the experimental results of STOI are listed in Table 1, the results of SI-SDR for all experiments are listed in Table 2, and the histograms are developed according to the relevant data, as illustrated in Figs. 5 and 6. The floating-point number calculation times and total number of parameters for each group of network training are listed in Table 3. The experimental results reveal that the speech separation model of the L-TCN by effectively acquisition IPD information based on attention proposed in this work is superior in both STOI and SI-SDR.

Table 3. Comparison of calculations and parameters for nine groups of experimental tests. (The yellow TCN is the control group.).

System model	Floating point number calculation times	Total parameters of network training
TCN/No-DF	99.995 M	1.486 M
TCN	91.608 M	1.404 M
TCN(M = 3)	69.760 M	969.443 K
TCN(M = 6)	89.471 M	1.234 M
TCN(C = 256)	83.351 M	1.222 M
L-TCN(C = 32)	84.512 M	1.337 M
L-TCN(C = 64)	85.802 M	1.353 M
L-TCN(C = 128)	88.383 M	1.384 M
L-TCN(C = 64M = 6)	**75.278 M**	**1.101 M**

4.5 E Analysis of Results

First, according to Experiments 1 and 3, we conclude that directional information is necessary for speech separation under reverberation conditions. Based on this, it laid a theoretical foundation to introduce additional effective spatial information.

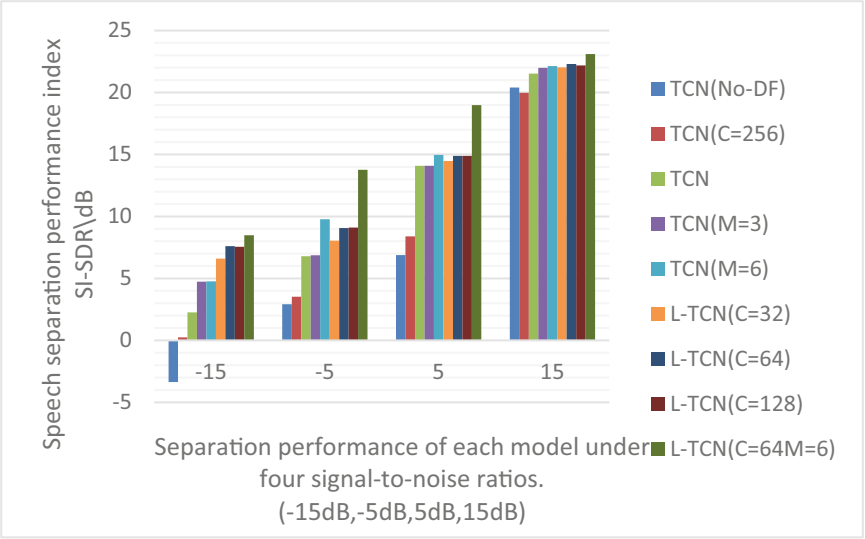

Fig. 5. Comparison of SI-SDR results of nine groups of experimental tests under four SNR ratios

Experiments 3, 4, and 5 proved that the proposed effectively acquisition IPD information based on attention method has an excellent effect on IPD aggregation. Experiments

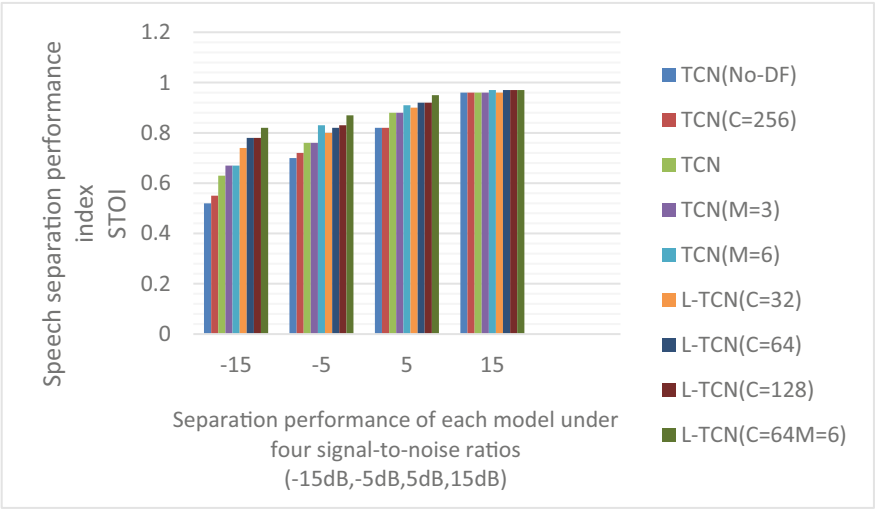

Fig. 6. Comparison of STOI results of nine groups of experimental tests under four signal-to-noise ratios

3 and 4 prove that even if additional IPDs are not introduced, the separation performance of the system can still be improved under the condition of a low SNR. Moreover, following the introduction of more IPDs, the system performance under a high SNR is improved. It has been proven that additional IPDs can be introduced if the SNR is high. The experimental results were close to the original prediction. Since reverberation reflects the voice at a low SNR, it reflects noise. Accordingly, when increasing the IPD, the lifting effect is not obvious, whereas in the case of a high SNR, properly raising the IPD can better capture the spatial information of the speech signal.

Experiments 2 and 3 prove that the network width of the TCN is not wasted when $C_1 = 512$, which provides a theoretical basis for adding weighted branches and reducing the network width. Experiments 3, 6, and 7 prove that the proposed L-TCN can improve the separation performance of the system based on saving parameters. According to the Eq. (14) when the number of expanded convolution channels $C_2 = 32, 64, 128$, the parameters of the original TCN are reduced by factors of ten times, 3.58 times, and 1.15 times respectively, all of which are superior to those of the TCN. With the increase in parameters, the system performance will be considerably improved; however, when the network width is too large $C_2 = 128$, and the system performance is not obviously improved compared with that of $C_2 = 64$, which proves that the network width is sufficient when $C_2 = 64$.

Finally, according to Experiments 5, 7, and 9, it is proven that we can use the attention mechanism to enhance the utilization rate of IPD, while using L-TCN to reduce the parameters, which can still improve the separation performance by simplifying the model.

Figure 7 illustrates the spectrum of mixed speech (−5 dB), target speech, noise, and separated speech. From the spectrum, the proposed L-TCN can evidently filter out most

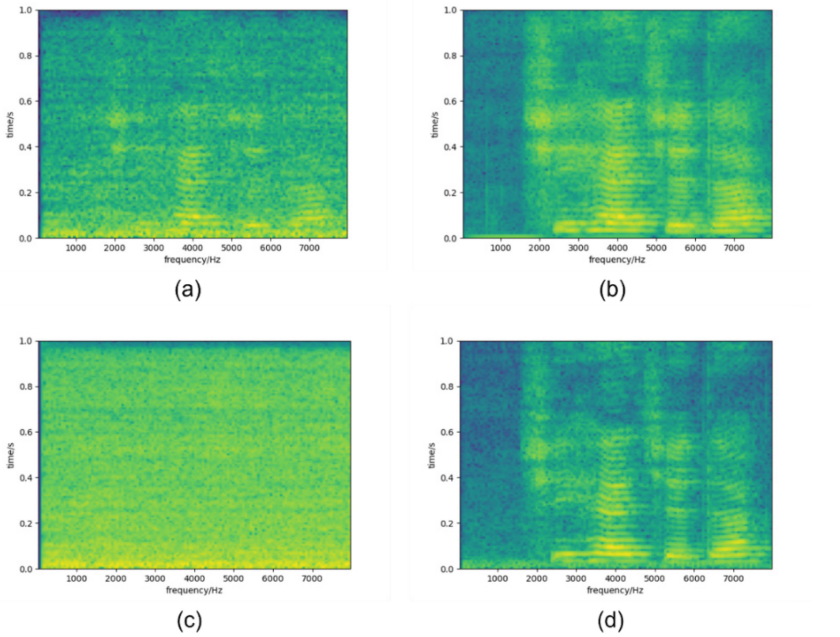

Fig. 7. Comparison before and after separation of -5dB mixed speech ((a) mixed speech, (b) target language, (c) noise, and (d) separated language when L-TCN(C = 64M = 6) and IPD number is 6)

of the noise in the process of speech separation and simultaneously recover the original target speech to a great extent to facilitate the post-processing of separated speech.

Figure 8 presents the spectra of Experiments 5, 7, and 9 and the target speech. According to the spectra, although attention-weighted IPD and L-TCN can both reduce network parameters and improve system performance, their emphasis is different. Using attention-weighted IPD is not as effective as L-TCN in noise removal; however, it exhibits better effects than L-TCN in preserving the original signal and retaining the signal as undistorted as possible because the weighted IPD is introduced in data fusion, which enables the use of additional spatial information in the experiment to recover the target speech more effectively, while the L-TCN focuses on parameter optimization and branch weighting of the network structure, which makes the network more accurate in the process of fitting mixed speech to separated speech, and the noise filtering is more obvious.

Finally, it should be pointed out that in this experiment, under the condition of reverberation, the noise was music with a wide spectrum, which made separation more difficult. Simultaneously, to be more in line with the real scene (the DOA direction of speech is also estimated), the direction error ($\pm 15°$) was artificially added when determining the direction in the experiment. All of these have increased the difficulty in the training of speech-separation systems. However, according to the experimental results,

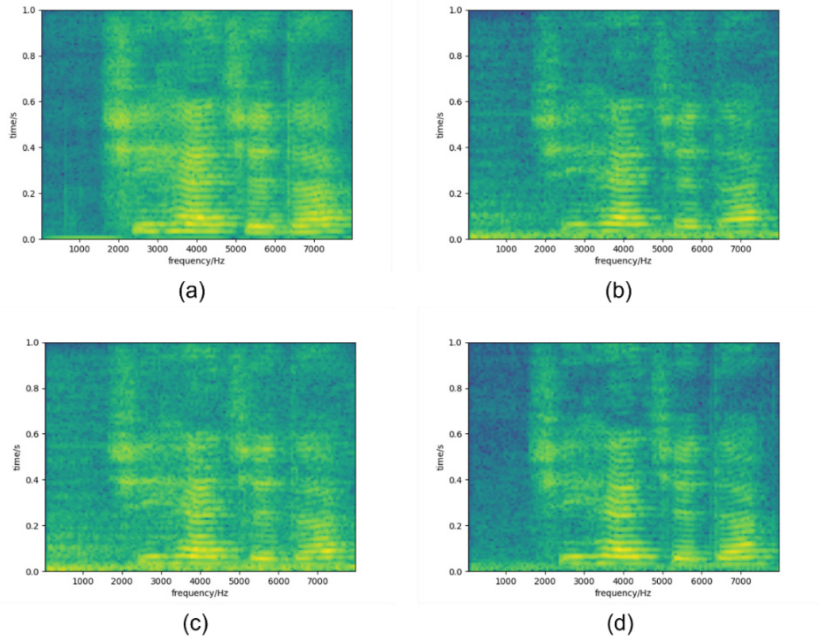

Fig. 8. Three groups of experimental spectrograms of-5dB mixed speech. ((a) the target speech, (b) experiment 7, (c) experiment 5, and (d) separated language of experiment 9).

the speech separation model of the L-TCN by effectively acquisition IPD information based on attention proposed in this study is indeed effective.

5 Conclusion

In this study, we propose an L-TCN model by effectively acquisition IPD information based on attention. By introducing weighted branches, the proposed model uses attention mechanism to reduce the dimension of channel phase difference, which effectively reduces the input dimension, reduces the structural parameters of the model and the overall calculation of the model, and improves the reliability and effectiveness of the speech separation results. We used the TIMIT dataset to evaluate and verify the performance of the proposed speech-separation model. The experimental results reveal that using the attention mechanism to process IPD data from an input perspective can improve the performance of speech separation under various SNRs, particularly at high SNRs. Additionally, by optimizing the structure of L-TCN, the amount of calculations and parameters, and combining it with the strategy of appropriately increasing the network width improves the performance of speech separation.

Although this study reduces the amount of computations while reducing the number of parameters, it increases the spatial complexity of the network. In the future, we will optimize the speech separation performance of the network model from the perspective of network-space complexity.

Acknowledgments. This work was funded by the National Natural Science Foundation of China:62071135, the Project (CRKL200111 and CRKL210110) from Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology) and Innovation Project of GUET Graduate Education (2021YCXS037).

References

1. Cherry and C. J. J. o. t. A. S. o. America, F.: Article title. 2(5), 99–110 (2016)
2. Ephrat, A. et al.,: Article title. 2(5), 99–110 (2016)
3. Bronkhorst, A.W.J.A.: Article title. 2(5), 99–110 (2016)
4. S. F. J. A. S. Boll and S. P. I. T. on, F.: Article title. 2(5), 99–110 (2016)
5. Hu, X., Wang, S., Zheng, C., Li, X.: Appl. Acoust. 2(5), 99–110 (2016)
6. E. J. I. T. A. S. S. Process, F.: Article title. 2(5), 99–110 (2016)
7. Michelsanti,D. et al.: Article title. IEEE/ACM Trans. Audio Speech Langu. Process. 2(5), 99–110 (2016)
8. D. E. J. J. O. M. L. R. King, F.: Article title. 2(5), 99–110 (2016)
9. Gu, R., Et al.: Article title. 2(5), 99–110 (2016)
10. Wang,D.L., Chen, J.: Article title. 2(5), 99–110 (2016)
11. Chen,L., Yu, M., Su, D., Yu, D.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
12. Wang, Z.Q., Roux, J.L., Hershey, J.R.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
13. Xu, Y., et al.,: Article title. 2(5), 99–110 (2016)
14. Zhang, Z., et al.,: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
15. Z. Zhang, Y. Xu, M. Yu, S. X. Zhang, L. Chen, and D. Yu, F.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
16. Mack, W., Habets, E.: Article title. 2(5), 99–110 (2016)
17. Luo,Y., Mesgarani, N.: Article title. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2(5), 99–110 (2016)
18. C. S. Lea, R. Vidal, A. Reiter, and G. J. A. Hager, F.: Article title. 2(5), 99–110 (2016)
19. C. H. Taal, R. C. Hendriks, R. Heusdens, J. J. I. T. o. A. S. Jensen, and L. Processing, F.: Article title. 2(5), 99–110 (2016)
20. Chollet, F.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
21. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Presented at the Proceedings of the 34th International Conference on Machine Learning. Vol 70, Sydney, NSW, Australia (2017)
22. Chen, J., Wang, D.L.: Article title. 2(5), 99–110 (2016)
23. Luo, Y., Mesgarani, N.: 2(5), 99–110 (2016)
24. Vaswani, A., et al.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
25. Xu, C., Rao, W., Chng, E.S., Li, H.: Speech Lang. Process. **2**(5), 99–110 (2016)
26. He,K., Zhang, X., Ren, S., Sun, J.: In: 9th International Proceedings on Proceedings, pp. 1–2. Location (2010)
27. Li, A., Liu, W., Zheng, C., Fan, C., Li, X.: IEEE/ACM Trans. Audio, Speech, Lang. Proc. **2**(5), 99–110 (2016)