EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Channel and temporal-frequency attention UNet for monaural speech enhancement

Shiyun Xu[1], Zehua Zhang[1] and Mingjiang Wang[1*]

Abstract

The presence of noise and reverberation significantly impedes speech clarity and intelligibility. To mitigate these effects, numerous deep learning-based network models have been proposed for speech enhancement tasks aimed at improving speech quality. In this study, we propose a monaural speech enhancement model called the channel and temporal-frequency attention UNet (CTFUNet). CTFUNet takes the noisy spectrum as input and produces a complex ideal ratio mask (cIRM) as output. To improve the speech enhancement performance of CTFUNet, we employ multi-scale temporal-frequency processing to extract input speech spectrum features. We also utilize multi-conv head channel attention and residual channel attention to capture temporal-frequency and channel features. Moreover, we introduce the channel temporal-frequency skip connection to alleviate information loss between down-sampling and up-sampling. On the blind test set of the first deep noise suppression challenge, our proposed CTFUNet has better denoising performance than the champion models and the latest models. Furthermore, our model outperforms recent models such as Uformar and MTFAA in both denoising and dereverberation performance.

Keywords Speech enhancement, Neural network, Denoising, Dereverberation

1 Introduction

Speech is vital in various aspects of our daily lives, including mobile communication, audio chat, remote conferences, and speech control. There are many sources of noise, such as car honking, machine noise, rain, and murmurs. Reverberation occurs when sound waves propagate indoors, reflecting and absorbing off walls, ceilings, floors, and other obstacles. Even after the sound source has stopped, the sound wave persists in the room, reflecting and absorbing until it eventually dissipates. Therefore, noise and reverberation frequently disrupt speech, severely affecting the listener's experience. In light of the issues above, removing background noise and reverberation from noisy speech is essential. Because of the user's desire for high-quality speech, speech enhancement, and de-reverberation technologies are increasingly critical.

Speech enhancement techniques can be broadly classified into traditional methods and deep neural network (DNN)-based methods. Traditional methods refer to completing the speech enhancement task via signal processing and certain statistical assumptions. Examples of traditional methods include subspace algorithms [1], spectral subtraction [2], and algorithms based on statistical models [3, 4]. Traditional methods mainly operate under the assumption that noise signals are stationary. However, most noise in natural environments is non-stationary, and most traditional methods have limitations.

Due to significant advancements in computing power, DNN-based speech enhancement methods have become increasingly prevalent. Because DNNs are highly effective in handling non-stationary noise, the research on DNN-based speech enhancement is more and more abundant. DNN-based speech enhancement methods can be categorized into time-domain and time-frequency domain approaches. The time-domain

*Correspondence:
Mingjiang Wang
mjwang@hit.edu.cn
[1] Key Laboratory for Key Technologies of IoT Terminals, Harbin Institute of Technology, Shenzhen, China

approach directly estimates the clean speech signal based on the noisy speech, taking an end-to-end approach. Wavenet [5], the first DNN capable of generating natural human speech and better modeling acoustic features, was instrumental in developing end-to-end denoising methods. Rethage et al. [6] propose an end-to-end denoising method that retains the acoustic feature modeling capability of Wavenet while reducing the algorithm's time complexity by removing its autoregressive features. Speech enhancement models generally process only the amplitude spectrum, ignoring the phase information. To exploit phase information fully, Stoller et al. [7] propose the time-domain end-to-end speech enhancement model Wave-U-Net, which allows the modeling of phase information and avoids fixed spectral transformations. To mitigate high delay and computational cost issues, Luo et al. [8] propose an end-to-end full-convolution time-domain speech separation network (Conv-TasNet).

Scholars have made significant progress in speech enhancement in the time domain; however, speech enhancement in the time-frequency domain is becoming increasingly popular. The time-frequency domain approach offers several advantages, such as the ability to focus on features often overlooked by the time-domain approach, enhanced robustness, and reduced computational cost [9]. The time-frequency domain approach typically involves two main methods: spectral mapping and spectral masking.

Spectral mapping refers to estimating the clean spectrum from the noisy spectrum. In the time-frequency domain-based speech enhancement algorithms, the phase information plays a crucial role in the enhancement performance [9, 10]. However, estimating the phase spectrum directly is challenging since it lacks a clear structure. To address this problem, Tan et al. propose a novel framework using a convolutional recurrent network (CRN) [11] and introduce a gated convolution module [12] to estimate the phase spectrum.

Spectral masking takes the noisy spectrum as input and the mask as a training target. The mask can take various forms, such as ideal binary mask (IBM), ideal ratio mask (IRM), and complex ideal ratio mask (cIRM). Chen et al. [13] propose a separation and enhancement model based on long short-term memory (LSTM) that focuses on the temporal dynamics features of speech to improve speech intelligibility. This model significantly enhances objective speech intelligibility under low delay. Hao et al. propose FullSubNet [14], which uses a combination of a pure full-band model and a pure sub-band model to model the signal smoothly, pay attention to local features, and capture global long-distance features. To address the problems of input and output mismatch and rough handling of the frequency band, the FullSubNet+ is proposed by Chen et al. [15].

In the field of speech enhancement, the attention mechanism plays a pivotal role in dynamically adjusting focus to distinct regions based on the unique characteristics of input signals. As a result, it has gained widespread adoption in speech enhancement models, aiming to enhance the quality and intelligibility of speech signals. However, the vanilla attention approach presents significant challenges due to its high computational complexity, rendering it impractical for speech-processing tasks. Consequently, finding effective strategies to mitigate the complexity of attention remains a significant challenge.

In this work, based on time-frequency domain speech enhancement and spectral masking, we propose a novel model for speech enhancement called the channel and temporal-frequency attention UNet (CTFUNet), which combines channel and time-frequency attention mechanisms to denoising and dereverberation speech signals. CTFUNet takes the noisy complex spectrum as input and produces the cIRM as the output, achieving excellent performance in speech enhancement. Our contributions are summarized as follows:

- To alleviate the computational complexity of vanilla self-attention, we propose the multi-conv head channel attention (MCHCA) module. It enables the extraction of temporal-frequency speech features while maintaining linear complexity calculations for self-attention.
- With the aim of improving the efficiency of channel feature extraction, we introduce the residual channel attention module (RCAM) into our work. This module can selectively highlight the channels with the most features in the neural network.
- In the encoding and decoding framework, the encoding process compresses and loses a large amount of detailed information. To alleviate this problem and further extract features from channel dimensions and temporal-frequency dimensions at multiple scales and levels, we propose the channel temporal-frequency skip connection (CTFSC) between the down-sampling and up-sampling modules.

The remaining contents of this paper are presented as follows: In Section 2, we provide an overview of the related works relevant to our study. Section 3 presents the signal model and the various components of our proposed CTFUNet in detail. Section 4 elaborates on the datasets used in our experiments and the implementation details of our experiments. Section 5 presents the results of our experiments and provides a detailed analysis. Finally, in Section 6, we draw conclusions based on our findings.

2 Related works

2.1 Self-attention

Self-attention is a widely used attention mechanism and a crucial component of transformer. By utilizing self-attention, the network can capture long-range dependencies in the input, and the multi-head structure enables parallel attention calculation. Recently, the effectiveness of self-attention has been demonstrated in various fields such as computer vision, natural language processing, and speech processing [16–18]. The specific calculation process of self-attention is as follows:

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{1}$$

where Q, K, and V denote query, key, and value projection vectors. d_k represents the dimension of K.

However, self-attention takes up a significant amount of computation and graphics memory when calculating the attention map. For example, for an image with $H \times W$ pixels, its complexity is $\mathcal{O}(H^2W^2)$.

To alleviate this problem, Zhao et al. [19] make improvements to the vanilla attention by dividing temporal-frequency attention into temporal attention and frequency attention, reducing the complexity of attention. Zhang et al. [20] propose a axial self-attention (ASA) for speech enhancement. ASA can reduce the need for memory and computation, making it more suitable for speech signals. In our research, we make improvements to self-attention, reducing its complexity by implicitly encoding global information by calculating self-attention on the channel dimension.

2.2 Temporal convolutional network

Previous research has demonstrated that recurrent neural networks (RNNs) excel in addressing sequence-related tasks [21, 22]. However, RNNs operate one time step at a time and process the next step only after completing the previous one. As a result, RNN calculations require significant memory to store all intermediate results. To overcome this challenge, researchers propose a new network for time series processing called temporal convolutional network (TCN) [23]. TCN is based on convolutional neural networks (CNNs) and incorporates causal convolution, dilated convolution, and residual module. In comparison to RNNs, TCN offers several benefits, including:

- Parallelism. Unlike RNN, TCN processes the input time series as a whole without waiting for the last time step to complete processing.
- Flexible receptive field size. Dilated convolution improves the receptive field, so TCN can flexibly change the size of the receptive field by using dilated convolution.
- Because of the introduction of the residual module, TCN has stable gradients, which can avoid gradient explosion or vanishing.
- During training, TCN requires less memory than RNN.

In the field of speech enhancement, TCN is widely used because of its superior ability to process sequence-related tasks than RNN [24]. Pandey et al. [25] insert a TCN between the encoder and decoder and achieve a good performance of speech enhancement with fewer trainable parameters. Lin et al. [26] combine self-attention with TCN and adopt the multi-stage learning method to extract features. In our study, we used temporal-frequency convolutional network (TFCN) [27] instead of TCN. TFCN is an improvement of TCN that can simultaneously utilize features from both temporal and frequency dimensions, resulting in stronger modeling capabilities.

2.3 UNet

UNet [28, 29] is a network model that follows a symmetrical U-shaped structure. It is typically an encoder-decoder structure. The first half of UNet is responsible for feature extraction and continuously reducing the input size, typically achieved through convolution and down-sampling operations. The latter half aims to restore the original input size. Apart from convolution, the crucial steps of this process include up-sampling and skip connections. Skip connections concatenate the location information of the bottom layer with the semantic information of the deep layer to achieve better results. While UNet has a straightforward structure and good performance, its model size is relatively large, and its performance may be affected by the receptive field.

Because the network structure of UNet has local connectivity characteristics, it can be used for speech signal processing. Choi et al. [30] improve UNet by proposing Tiny Recurrent UNet (TRUNet) and propose phase-aware β-sigmoid mask (PHM) for speech enhancement. Fu et al. [29] build a network framework based on UNet and Conformer [31]. In addition, they simultaneously model the real and imaginary parts of the input speech spectrum and calculate self-attention on both temporal-frequency dimensions. In our study, we also improve UNet to focus not only on temporal-frequency dimensional features, but also on channel dimensional features.

3 Method

3.1 Signal model

Assuming that $x(t)$ represents a clean speech signal, the acoustic signal captured by the microphone in a noisy room can be expressed as follows:

$$y(t) = h(t) * x(t) + n(t) \tag{2}$$

where $h(t)$ denotes the room impulse response (RIR), $n(t)$ indicates the background noise, and $*$ denotes convolution operation. Moreover, based on the definition of reverberation, the RIR $h(t)$ can be decomposed into the direct part $h_d(t)$ and the reflection part $h_r(t)$, so $y(t)$ can be re-expressed as:

$$\begin{aligned} y(t) =& h_d(t) * x(t) + h_r(t) * x(t) + n(t) \\ =& d(t) + r(t) + n(t) \end{aligned} \tag{3}$$

where $d(t)$ denotes direct sound, which is the sound that travels directly from the sound source to the listener without reflecting off any surfaces, and $r(t)$ denotes reverberation, which is the sound that is reflected off the surfaces in the room before reaching the listener's ear. The discrete Fourier transform of Eq. 3 is given by:

$$Y(l,f) = D(l,f) + R(l,f) + N(l,f) \tag{4}$$

where l and f denote frame index and frequency bin, respectively. $D(l, f)$ represents the target to be estimated, while $R(l, f)$ and $N(l, f)$ represent the complex spectrum of the reverberation and noise that need to be removed, respectively.

The proposed model in our study takes $Y(l, f)$ as input and outputs the estimated spectrum mask $\hat{M}(l,f)$. Subsequently, we use $\hat{M}(l,f)$ to estimate the desired output $\hat{D}(l,f)$.

3.2 Overall structure

In recent years, UNet has demonstrated its efficacy in feature extraction from data. This architecture has been widely adopted in speech enhancement and has shown remarkable results [28, 29]. The proposed CTFUNet, which follows a typical UNet structure, is presented in Fig. 1. The input to CTFUNet is the complex spectrum of noisy speech. First, a phase encoder (PE) is employed to convert complex spectral features to real spectral features. Then, an 3x3 input convolution layer extracts features and changes the channel number for the later calculations. Following this, three encoders, two neck modules, three decoders, and CTFSC are utilized to construct the main network.

Each encoder mainly comprises a frequency down-sampling (FD) module, a temporal-frequency convolution module (TFCM), a MCHCA module, and a RCAM. The neck module consists of a TFCM, a MCHCA module, and a RCAM. The structure of the decoder is similar to that of the encoder, but with a frequency up-sampling (FU) module replacing the FD module. Furthermore, we utilize the CTFSC to connect the encoder and decoder. Finally, an output convolution layer is employed to obtain the cIRM $\hat{M}(l,f)$, and the masking method proposed by [20] is applied to obtain the enhanced spectrum $\hat{D}(l,f)$.

3.3 Phase encoder and TF-convolution module

Some previous studies [32, 33] have proven that real-valued speech enhancement networks are easier to build and achieve better enhancement effects on various datasets. Inspired by [20], we introduce the PE module into the model to perform the mapping of complex spectral features to real spectral features. The structure of our PE module is similar to that in [20], but it consists of only one complex convolution layer for

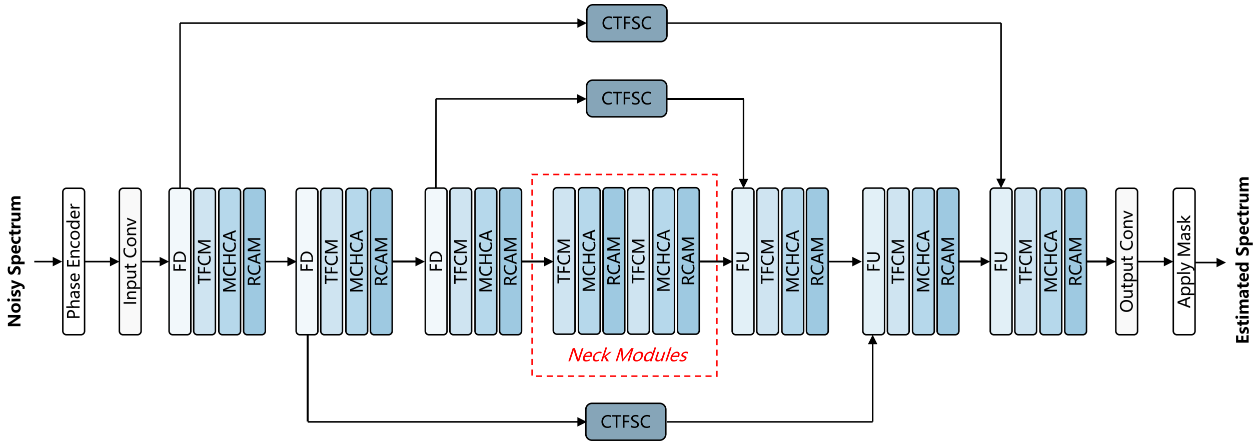

Fig. 1 Overall structure diagram of the proposed channel and temporal-frequency attention UNet

Table 1 Architecture of the *i*th TFCN

Layer name	Input size	Hyperparameters	Output s
conv2d-1	$C \times F \times L$	(1,1),(1,1),(1,1)	$C \times F \times L$
conv2d-2	$C \times F \times L$	(3,3),(1,1),(1,2^{i-1})	$C \times F \times L$
conv2d-3	$C \times F \times L$	(1,1),(1,1),(1,1)	$C \times F \times L$

processing the noisy speech spectrum. The kernel size and the stride of the complex convolution layer are set to (1,3) and (1,1). The feature dynamic range compression layer's power compression ratio [34] is 0.5.

To efficiently extract temporal-frequency features using small parameters and convolution kernels, [27] proposes a TFCN by replacing 1-D convolutions in TCN with 2-D convolutions. Motivated by this study, we introduce TFCM, which contains 6 TFCNs, each consisting of two point-wise convolution layers and a 2-D dilated convolution layer. The kernel size and stride of the 2-D dilated convolution layer are (3,3) and (1,1), respectively. For the ith TFCN, the dilations of the 2-D dilated convolution layer are set to 2^{i-1}. For the ith TFCN, its detailed description is shown in Table 1.

The input size and the output size of each layer are specified in channel_numbers × frequency_frames × time_frames format, and the hyperparameters in (kernelsize, strides, dilations) format.

3.4 Multi-conv head channel attention

Due to its large receptive field, self-attention has been widely used to capture long-term dependencies between features. However, its use in neural networks significantly increases the network's computational complexity. For instance, when calculating the self-attention map of a speech spectrum with size of $C \times F \times L$, the time complexity can be as high as $C \times F^2 \times L^2$. To ease this problem, many scholars put forward their solutions [35, 36]. Motivated by these works, we propose the MCHCA module as illustrated in Fig. 2. After replacing vanilla self-attention with MCHCA, the time complexity of calculating the self-attention map becomes $C^2 \times F \times L$, where C is far less than L. MCHCA can capture long-term information with linear complexity, thanks to its two key features:

- MCHCA avoids calculating self-attention across the temporal-frequency dimension and instead obtains the self-attention map across the channel dimension to encode global information implicitly. This approach effectively reduces the computational complexity of traditional self-attention.

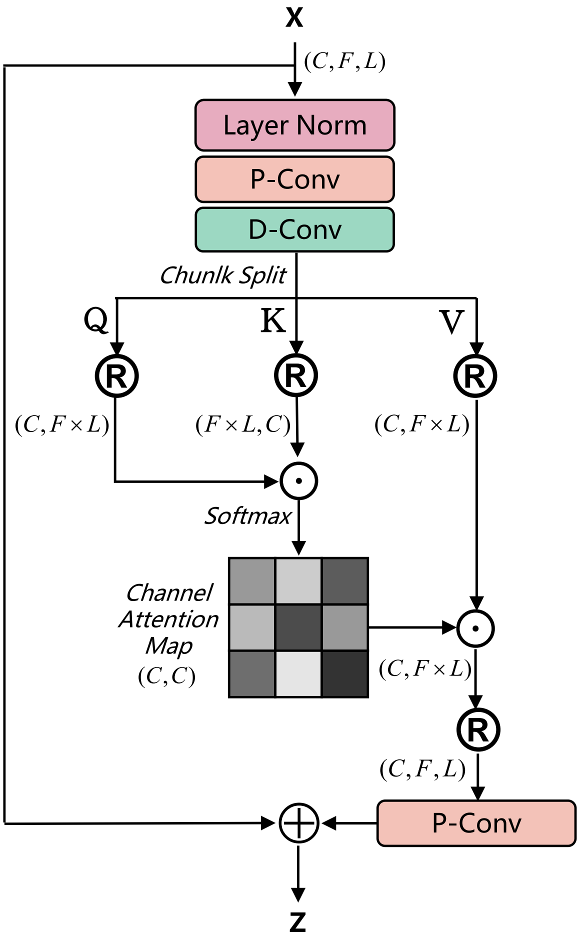

Fig. 2 The structure diagram of multi-conv head channel attention module

- To focus on local information, we incorporate 1x1 point-wise convolutions and 3x3 depth-wise convolutions prior to generating the self-attention map.

In MCHCA, we first apply layer normalization to the input. After the point-wise convolution layer captures the cross-channel information, we use the depth-wise convolution layer to extract the temporal-frequency information and obtain query (Q), key (K), and value (V) projection vectors. The process is mathematically represented as follows:

$$\begin{aligned} Q &= W_D^Q W_P^Q \text{LayerNorm}(\mathrm{x}) \\ K &= W_D^K W_P^K \text{LayerNorm}(\mathrm{x}) \\ V &= W_D^V W_P^V \text{LayerNorm}(\mathrm{x}) \end{aligned} \tag{5}$$

where W_P^* and W_D^* represent the projection matrixes in the point-wise convolution and depth-wise convolution layers. The integration of point-wise and depth-wise convolution layers exploits the features of different channels in the same temporal-frequency position, enabling the network to concentrate on local information.

For subsequent computation, we reshape $Q \in \mathbb{R}^{(C,F\times L)}$, $K \in \mathbb{R}^{(F\times L,C)}$ and $V \in \mathbb{R}^{(C,F\times L)}$ from the original size of $\mathbb{R}^{(C,F,L)}$. Then, we calculate the dot product of Q and K to encode global information across the channel dimension. Following this, we apply the Softmax function to the result to obtain the channel attention map, which has a size of $\mathbb{R}^{(C\times C)}$. Finally, we take the dot product of channel attention map and V to obtain the channel attention. The complete channel attention calculation process is expressed as Eq. 6:

$$\mathrm{ChAtten}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{Q}\cdot\mathrm{K}}{\mu}\right)\cdot\mathrm{V} \tag{6}$$

where μ is a learnable scaling factor to adjust the result of the dot product of Q and K. The overall calculation process of MCHCA is expressed as follows:

$$z = W_p\mathrm{ChAtten}(\mathrm{Q},\mathrm{K},\mathrm{V}) + \mathrm{x} \tag{7}$$

Furthermore, we incorporate multi-head processing on the channel dimension in MCHCA, which enables parallel computation of attention and the capture of features at multiple scales. Table 2 provides a detailed description of MCHCA, the hyperparameters are in (kernelsize, strides) format.

To confirm that our proposed MCHCA indeed reduces time complexity, we compared it with vanilla self-attention (VSA), improved T-F self-attention (ISA) [19], and axial self-attention (ASA) [20]. To ensure the successful operation of vanilla self-attention, the length of the speech is selected as 5 s. The comparison results are shown in Table 3.

Compared to VSA and ASA, although MCHCA has more MACs and the number of parameters, MCHCA greatly shortens the runtimes. Compared to ISA, MCHCA has similar runtime in fewer MACs and the number of parameters.

3.5 Residual channel attention module

Although we use MCHCA to obtain the channel attention map with the size of $\mathbb{R}^{(C\times C)}$, its essence is still temporal-frequency attention. [37] proposes a residual channel attention block, which allows the network to focus more on useful feature channels. Based on it, we introduce RCAM to capture features across different channels. The specific structure of RCAM is illustrated in Fig. 3.

The input first passes through the instance normalization layer and then through the depth convolution-ReLU-depth convolution block (a simple residual block) to obtain the residual features. Subsequently, the residual features are used to obtain the feature information of all channels through 2-D average pooling, down-sampling convolution, ReLU, up-sampling convolution, and sigmoid activation functions. Finally, the residual features are multiplied by the channel feature information and

Table 2 Architecture of MCHCA

Layer name	Input size	Hyperparameters	Output size
conv2d-1	$C \times F \times L$	(1,1),(1,1)	$3C \times F \times L$
conv2d-2	$3C \times F \times L$	(3,3),(1,1)	$3C \times F \times L$
conv2d-3	$C \times F \times L$	(1,1),(1,1)	$C \times F \times L$

Table 3 Comparison results of several self-attentions

	Times(s)	MACs	Para.
VSA	7.015	317.482 M	3.936 K
ISA [19]	0.064	495.581 M	6.144 K
ASA [20]	0.192	**152.933 M**	**1.752 K**
ours.	**0.062**	400.079 M	4.960 K

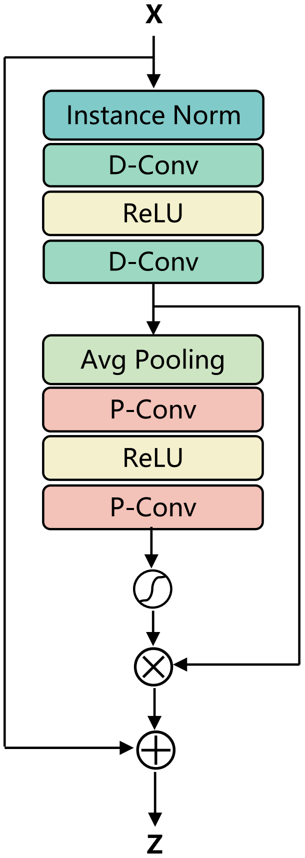

Fig. 3 The structure diagram of residual channel attention module

Table 4 Architecture of RCAM

Layer name	Input size	Hyperparameters	Output size
conv2d-1	$C \times F \times L$	(3,3),(1,1)	$C \times F \times L$
conv2d-2	$C \times F \times L$	(3,3),(1,1)	$C \times F \times L$
avg pooling2d	$C \times F \times L$	-	$C \times 1 \times 1$
conv2d-3	$C \times 1 \times 1$	(1,1),(1,1)	$C/4 \times 1 \times 1$
conv2d-4	$C/4 \times 1 \times 1$	(1,1),(1,1)	$C \times 1 \times 1$

added to the input to enable the network to use channel information fully. Table 4 provides a detailed description of RCAM, the hyperparameters are in (kernelsize, strides) format.

3.6 Frequency down and up sampling

In [20], their works have been demonstrated that FD and FU modules are effective in extracting multi-scale features. Based on their works, we incorporate FD and FU modules into our approach. Additionally, at each scale, we introduce TFCM, MCHCA, and RCAB to enable the network to more effectively capture temporal-frequency and channel features.

The FD and FU modules in our work have similar structures to those proposed in [20]. However, we make modifications by replacing the batch normalization layer with the instance normalization layer, which is better for speech enhancement tasks [38, 39]. Furthermore, we set the kernel size, stride, and groups of both the convolution layer and the transpose convolution layer to (4, 4), (2, 1), and 2, respectively.

3.7 Channel temporal-frequency skip connection

Up to now, significant progress has been made in research on attention mechanisms. The incorporation of attention can not only highlight critical regions but also enhance the representation power of these regions. Woo et al. [40] and Hu et al. [41] calculate attention weights on both the channel and spatial dimensions, highlighting the importance of channel attention. To further exploit the features of the temporal-frequency and channel dimensions at multiple scales and levels, we propose CTFSC, as illustrated in Fig. 4. CTFSC mainly consists of a channel focussing module and a temporal-frequency focussing module.

In the channel focussing module, input first passes through the average pooling layer and the max pooling layer to aggregate the temporal-frequency features of speech and obtain P_{ca} and P_{cm}, respectively. Then, P_{ca} and P_{cm} are fed into a convolution block (CB) with shared parameters. Finally, we merge and output the channel eigenvector F_c using a sigmoid function and element-wise addition. The overall calculation process of the channel focussing module is as follows:

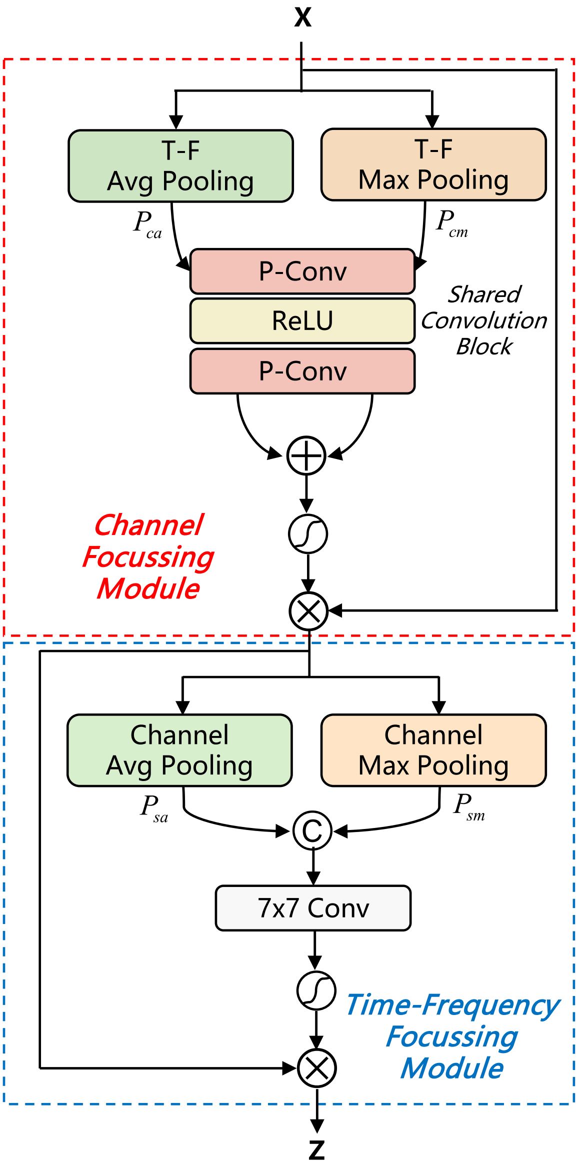

Fig. 4 The structure of channel temporal-frequency skip connection

$$F_c = \sigma(CB(Avg(x)) + CB(Max(x))) \otimes x \tag{8}$$

where σ denotes sigmoid function, $Avg(\cdot)$ represents average pooling calculation, $Max(\cdot)$ denotes max pooling calculation, x represents the input, and $\otimes$ denotes element-wise product.

In the temporal-frequency focussing module, the output of channel focussing module F_c passes through the average pooling layer and the max pooling layer for channel dimension to aggregate the channel features of speech, obtaining P_{sa} and P_{sm}, respectively. Subsequently, the concatenation of P_{sa} and P_{sm} passes through a convolution layer with a kernel size of (7, 7) followed by a sigmoid layer to obtain the output. The calculation process of the temporal-frequency focussing module can be expressed as follows:

$$z = \sigma(W([Avg(F_c); Max(F_c)])) \otimes F_c \tag{9}$$

where W denotes the projection matrix of the 7x7 convolution layer. Table 5 provides a detailed description of CTFSC, the hyperparameters are in (kernelsize, strides) format.

3.8 Loss function

For speech enhancement tasks, both magnitude and phase information are crucial. Therefore, we adopt the complex mean squared error (cMSE) proposed in [42] as our loss function. The cMSE is defined as follows:

$$\mathcal{L} = \frac{1}{l \times f}\left(\alpha \cdot P_{cRI} + \beta \cdot P_{cMag}\right) \tag{10}$$

where P_{cRI} and P_{cMag} can be expressed as follows:

$$\begin{aligned} P_{cRI} &= \sum_{l,f} \left|\hat{S}_{cRI} - S_{cRI}\right|^2 \\ P_{cMag} &= \sum_{l,f} \left|\hat{S}_{cMag} - S_{cMag}\right|^2 \end{aligned} \tag{11}$$

where S_{cRI} and S_{cMag} represent the complex compression spectrum and magnitude compression spectrum of clean speech. $\hat{S}_*$ denotes the estimated speech spectrum. It should be noted that we omit the frame index l and the frequency index f for brevity. α and β are 0.3 and 0.7. S_{cRI}, and S_{cMag} can be specifically expressed as follows:

$$S_{cMag} = \left|S_{Mag}\right|^c, \quad S_{cRI} = S_{cMag} \cdot \frac{S_{RI}}{S_{Mag}} \tag{12}$$

where c denotes the compressibility coefficient, set as 0.3.

Table 5 Architecture of CTFSC

Layer name	Input size	Hyperparameters	Output size
avg pooling2d-1	$C \times F \times L$	-	$C \times 1 \times 1$
max pooling2d-1	$C \times F \times L$	-	$C \times 1 \times 1$
conv2d-1	$C \times 1 \times 1$	(1,1),(1,1)	$C \times 1 \times 1$
conv2d-2	$C \times 1 \times 1$	(1,1),(1,1)	$C \times 1 \times 1$
avg pooling2d-2	$C \times F \times L$	-	$1 \times F \times L$
max pooling2d-2	$C \times F \times L$	-	$1 \times F \times L$
concatenation	$2, 1 \times F \times L$	-	$2 \times F \times L$
conv2d-3	$2 \times F \times L$	(7,7),(1,1)	$1 \times F \times L$

4 Experiment

4.1 Datasets

In our experiment, we utilize three training datasets, all of which are derived from clean speech and noise datasets provided by the first Deep Noise Suppression Challenge [43]. The clean speech datasets include about 500 h of English speech clips from 2150 speakers. The noise datasets are composed of 181 h of clips from 150 classes. To conduct the ablation study, we first use the image source method to obtain 100,000 pairs of RIRs with reverberation time RT60 from 0.3 s to 1.4 s. We convolve 75% of the clean speech with randomly selected RIRs and add noise with a random signal-to-noise ratio (SNR) ranging from −5 to 20 dB to the reverberant speech. Finally, we generate a 100-h train dataset and a 20-h validation dataset. For our test set, we select the blind test set provided by the 1st DNS challenge, which comprises two parts: with reverberation and without reverberation. The SNR of the blind test set ranges from 0 to 20 dB.

We generate the second dataset to compare the denoising performance with other models. All generation processes are the same as above, except that the duration of the dataset is 500h. To ensure the fairness of comparison, we also select the blind test set provided by the 1st DNS challenge as the test set.

To evaluate the denoising and dereverberation performance of the CTFUNet, we still employ the clean speech datasets and the noise datasets provided by the 1st DNS challenge to generate our third dataset. We divide the clean speech and noise datasets into the train, validation, and test datasets according to the proportion of 80%, 10%, and 10%. Subsequently, all clean speech is convolved with the RIRs generated earlier, and noise with SNR range from −5 to 20 dB is randomly added. Finally, we obtain a 100-h train dataset, a 10-h validation dataset, and a 5-h test dataset for our experiments.

The sampling rate of all the above speech is 16 kHz.

4.2 Implementation details

In the experiment, the frame length and hop length of the STFT complex spectrum are 20 ms and 10 ms. The output channel numbers of PE and input convolution layer are 2 and 32. The output channel numbers of the three FDs are 64, 128, and 256. The output channel numbers of the three FUs are 128, 64, and 32. The head numbers of MCHCA are 1, 2, and 4 in encoders and 8 in the neck module. The multiple of down-sampling convolution and up-sampling convolution in RCAM is 4. The output channel number of the output convolution layer is 4. Table 6 provides a detailed description of CTFUNet, the hyperparameters are in (*kernelsize*, *strides*) format.

Table 6 Architecture of CTFUNet

Layer name	Input size	Hyperparameters	Output size
PE	2 × 161 × *L*	-	2 × 161 × *L*
input conv2d	2 × 161 × *L*	(3,3),(1,1)	32 × 160 × *L*
encoder-fd-1	32 × 160 × *L*	(4,4),(2,1)	64 × 80 × *L*
encoder-fd-2	64 × 80 × *L*	(4,4),(2,1)	128 × 40 × *L*
encoder-fd-3	128 × 40 × *L*	(4,4),(2,1)	256 × 20 × *L*
neck-1	256 × 20 × *L*	-	256 × 20 × *L*
neck-2	256 × 20 × *L*	-	256 × 20 × *L*
decoder-fu-1	256 × 20 × *L*	(4,4),(2,1)	128 × 40 × *L*
decoder-fu-2	128 × 40 × *L*	(4,4),(2,1)	64 × 80 × *L*
decoder-fu-3	64 × 80 × *L*	(4,4),(2,1)	32 × 160 × *L*
output conv2d	32 × 160 × *L*	(3,3),(1,1)	4 × 161 × *L*

The optimizer is AdamW, and the initial learning rate is 0.001 decaying exponentially by 0.98 with the training epoch increasing. We train the network for 50 epochs with a batch size of 2.

To evaluate the denoising performance of the CTFUNet, we select the following metrics: WB-PESQ [44], NB-PESQ [45], STOI [46], SI-SDR [47], and DNSMOS [48]. To evaluate the dereverberation performance of the CTFUNet, we select SRMR [49], CD [50], LLR [50], and SNR$_{fw}$ [50] as our objective evaluation metrics.

5 Results and analysis

5.1 Ablation study

In this section, we conduct an ablation study to investigate the impact of key modules in the proposed CTFUNet on performance. We evaluate denoising performance on the blind test set while disregarding dereverberation performance. Specifically, we replace MCHCA with ISA and ASA to demonstrate the superiority of MCHCA. We do not choose VSA because its computational complexity is too large to be used for speech processing tasks. Besides, we individually remove RCAM, MCHCA, and CTFSC modules from the CTFUNet architecture. "+ISA" and "+ASA" refer to replace MCHCA with ISA and ASA. "−CTFSC" refers to simply passing the output of FD into FU, concatenating, and element-wise multiplying with the input of FU without any further processing. Table 7 presents the results of the ablation study. After replacing MCHCA with ISA and ASA, there is a significant decrease in various performances. Combined with Table 3, our proposed MCHCA has significant advantages in both runtimes and performance. CTFSC increases the number of parameters by 0.2 M, but it significantly enhances the denoising performance of the network. The CTFSC module helps alleviate information loss during down-sampling and up-sampling. MCHCA adds 1 M parameters to the network but fully extracts temporal-frequency features and improves denoising capability. RCAM captures channel dimension features, leading to a parameter increase of 1.2 M.

5.2 Denoising performance comparison

Table 8 illustrates the denoising performance comparison between our proposed CTFUNet and other models with similar parameters. Compared with the champion model DCCRN [51] in the real-time track of the 1st DNS challenge, the NB-PESQ scores of CTFUNet increased by 0.664 and 0.373 with and without reverberation. In comparison with the champion model PoCoNet [53] in the non-real-time track of the 1st DNS challenge, the WB-PESQ scores of CTFUNet increased by 0.535 and 0.428 with and without reverberation. Additionally, we also compare with several speech enhancement models proposed in recent years, such as GaGNet [55], FullSubNet+ [15], and FS-CANet [56]. The results demonstrate that the WB-PESQ, NB-PESQ, STOI, and SI-SDR of CTFUNet are significantly better than those of other models, with or without reverberation. Therefore, our proposed CTFUNet can achieve excellent denoising performance with a 500-h train dataset, which is much smaller than the datasets used by other models.

Table 7 Performance of WB-PESQ, NB-PESQ, STOI, and SI-SDR in the ablation study

Model	#Para.	With reverb				Without reverb			
		WB-PESQ	NB-PESQ	STOI(%)	SI-SDR	WB-PESQ	NB-PESQ	STOI(%)	SI-SDR
Noisy	-	1.822	2.753	86.62	9.03	1.582	2.454	91.52	9.07
+ISA	6.5M	2.525 ± 0.089	3.290 ± 0.042	89.36 ± 0.19	13.38 ± 0.50	2.214 ± 0.076	3.023 ± 0.051	92.32 ± 0.15	13.84 ± 0.57
+ASA	5.4M	3.155 ± 0.012	3.658 ± 0.007	93.44 ± 0.03	16.14 ± 0.13	2.945 ± 0.017	3.520 ± 0.008	96.45 ± 0.05	17.35 ± 0.14
CTFUNet	6.1M	**3.196** ± 0.014	**3.673** ± 0.003	**93.63** ± 0.01	**16.36** ± 0.03	**2.979** ± 0.003	**3.540** ± 0.001	**96.64** ± 0.03	**17.60** ± 0.03
−RCAM	4.9M	3.157 ± 0.015	3.648 ± 0.006	93.51 ± 0.07	16.27 ± 0.08	2.951 ± 0.001	3.517 ± 0.002	96.53 ± 0.03	17.52 ± 0.05
−MCHCA	5.1M	3.143 ± 0.007	3.643 ± 0.003	93.38 ± 0.06	16.08 ± 0.02	2.951 ± 0.012	3.510 ± 0.001	96.54 ± 0.02	17.43 ± 0.16
−CTFSC	5.9M	2.996 ± 0.150	3.576 ± 0.054	92.93 ± 0.34	15.57 ± 0.60	2.820 ± 0.117	3.428 ± 0.072	96.08 ± 0.25	16.97 ± 0.48

Table 8 Denoising performance comparison of CTFUNet with other models

Model	**#Para.**	**Year**	**With reverb**				**Without reverb**			
			WB-PESQ	**NB-PESQ**	**STOI(%)**	**SI-SDR**	**WB-PESQ**	**NB-PESQ**	**STOI(%)**	**SI-SDR**
Noisy	-	-	1.822	2.753	86.62	9.033	1.582	2.454	91.52	9.07
DCCRN [51]	3.7M	2020	-	3.077	-	-	-	3.266	-	-
DCCRN+ [52]	4.7M	2021	-	3.30	-	-	-	3.33	-	-
Conv-TasNet [8]	5.1M	2019	2.75	-	-	-	2.73	-	-	-
PoCoNet [53]	50M	2020	2.832	-	-	-	2.748	-	-	-
CTS-Net [54]	4.4M	2021	3.02	3.47	92.7	15.58	2.94	3.42	96.66	17.99
FullSubNet [14]	5.6M	2021	3.057	3.584	92.11	16.04	2.882	3.428	96.32	17.30
GaGNet [55]	5.9M	2022	3.18	3.57	93.22	16.57	3.17	3.56	97.13	18.91
FullSubNet+ [15]	8.7M	2022	3.177	3.648	93.64	16.44	3.002	3.503	96.67	18.00
FS-CANet [56]	4.2M	2022	3.218	3.665	93.93	16.82	3.017	3.513	96.74	18.08
CTFUNet	6.1M	2023	**3.367**	**3.741**	**94.39**	**17.16**	**3.176**	**3.639**	**97.17**	**18.66**

5.3 Denoising and dereverberation performance comparison

In this section, we conduct a comparative analysis of the denoising and dereverberation performance of CTFUNet, Uformer [29], and MTFAA [20]. The datasets and experimental conditions for the three models are identical. The evaluation results are presented in Tables 9, 10, and 11.

To compare the denoising performance separately, we generate five noisy test sets with SNR ranging from −5 to 15 dB in 5 dB intervals. Each test set lasts 1 h, and all speech has no reverberation. We use WB-PESQ, DNSMOS, STOI, and SI-SDR as metrics. Table 9 shows the results of denoising performance. Obviously, our proposed CTFUNet has tremendous advantages over MTFAA and Uformer in denoising performance. On average, CTFUNet improves WB-PESQ by 1.002, DNSMOS by 0.795, STOI by 5.762%, and SI-SDR by 5.161 compared with noisy speech, demonstrating significant denoising capability.

To compare the dereverberation performance separately, we generated six reverberation test sets with RT60 range of 0.4 s to 1.4 s in steps of 0.2 s. WB-PESQ, DNSMOS, STOI, SI-SDR, SRMR, CD, LLR, and SNR_{fw} are selected as metrics, and the results of dereverberation performance are illustrated in Table 10. Compared with unprocessed reverberation speech, the three models significantly improve WB-PESQ, DNSMOS, STOI, SI-SDR, SRMR, and reduce CD, LLR at each reverberation time. For SNR_{fw}, all models decrease in low RT60 and increase in high RT60 compared with reverberation speech, but only SNR_{fw} of CTFUNet is higher than reverberation speech in the end. Overall, CTFUNet has more significant advantages than other models in all metrics except SI-SDR and SRMR. On average, CTFUNet improves WB-PESQ by 0.927 and DNSMOS by 0.978 and decreases CD by 1.297 and LLR by 0.306. Therefore, CTFUNet exhibits an excellent enhancement effect on reverberant speech.

Table 9 Denoising performance on test dataset without reverberation

SNR	**−5 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**Avg.**	**−5 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**Avg.**
	WB-PESQ						**DNSMOS**					
Noisy	1.257	1.379	1.593	1.901	2.323	1.691	1.809	2.135	2.491	2.758	2.936	2.423
Uformer	1.588	1.761	1.192	2.008	2.060	1.723	2.711	2.876	2.994	3.047	3.054	2.936
MTFAA	1.575	1.831	2.153	2.485	2.822	2.173	2.695	2.920	3.101	3.228	3.310	3.051
CTFUNet	**2.005**	**2.341**	**2.711**	**3.052**	**3.358**	**2.693**	**3.000**	**3.145**	**3.252**	**3.323**	**3.368**	**3.218**
	STOI (%)						**SI-SDR**					
Noisy	77.117	84.300	90.089	94.036	96.674	88.443	0.925	5.903	10.953	15.860	20.872	10.903
Uformer	82.139	86.073	88.034	86.620	88.257	86.625	7.615	9.481	10.576	10.900	10.624	9.8392
MTFAA	83.204	89.440	93.281	95.608	96.970	91.701	6.245	9.805	12.674	14.843	16.393	11.992
CTFUNet	**88.186**	**92.517**	**95.329**	**96.987**	**98.006**	**94.205**	**10.288**	**13.331**	**16.309**	**19.001**	**21.390**	**16.064**

Table 10 Dereverberation performance on test dataset without noise

RT60	0.4	0.6	0.8	1.0	1.2	1.4	Avg.	0.4	0.6	0.8	1.0	1.2	1.4	Avg.
	WB-PESQ							**DNSMOS**						
Noisy	2.448	1.923	1.574	1.694	1.382	1.342	1.727	2.922	2.398	1.732	1.849	1.498	1.274	1.946
Uformer	3.279	2.765	2.468	2.663	2.214	2.136	2.588	**3.144**	**2.996**	2.774	2.805	2.788	2.715	2.870
MTFAA	2.923	2.437	2.115	2.372	1.834	1.892	2.262	3.077	2.960	2.681	2.757	2.702	2.650	2.804
CTFUNet	**3.280**	**2.798**	**2.477**	**2.798**	**2.312**	**2.260**	**2.654**	3.106	2.981	**2.818**	**2.910**	**2.893**	**2.834**	**2.924**
	STOI							**SI-SDR**						
Noisy	97.215	93.456	87.571	90.498	87.087	83.495	89.887	11.619	7.071	4.333	6.915	3.874	2.845	6.110
Uformer	97.950	95.279	**93.492**	94.100	93.306	91.338	94.244	13.236	**9.286**	**8.262**	**10.596**	**8.267**	**7.144**	**9.465**
MTFAA	97.468	94.035	91.818	93.170	90.699	90.039	92.872	10.865	7.354	5.437	8.776	5.471	4.533	7.073
CTFUNet	**98.047**	**95.473**	93.228	**94.469**	**93.481**	**91.848**	**94.424**	**13.359**	9.098	7.262	10.328	7.793	6.519	9.060
	SRMR							**CD**						
Noisy	4.870	4.200	3.452	4.692	3.194	3.224	3.939	1.644	2.497	3.278	3.402	3.729	4.206	3.126
Uformer	**5.995**	**5.879**	**5.404**	**7.124**	**6.276**	**5.682**	**6.060**	1.615	2.018	2.115	2.222	2.508	2.756	2.206
MTFAA	5.589	5.309	4.866	6.419	5.146	5.129	5.410	1.629	2.033	2.216	2.248	2.525	2.656	2.218
CTFUNet	5.695	5.493	5.163	6.820	5.662	5.634	5.745	**1.405**	**1.658**	**1.804**	**1.850**	**2.037**	**2.222**	**1.829**
	LLR							**SNR$_{fw}$**						
Noisy	0.142	0.309	0.492	0.537	0.635	0.708	0.471	**22.821**	**17.353**	13.867	14.091	12.258	11.086	15.246
Uformer	0.122	0.209	0.216	0.264	0.298	0.363	0.245	18.674	15.013	14.752	14.889	13.764	12.002	14.849
MTFAA	0.123	0.232	0.286	0.282	0.387	0.450	0.293	16.069	13.704	12.868	12.780	11.146	11.019	12.931
CTFUNet	**0.096**	**0.142**	**0.155**	**0.163**	**0.198**	**0.235**	**0.165**	19.873	16.570	**15.764**	**15.680**	**14.648**	**13.219**	**15.959**

Table 11 Denoising and dereverberation performance on test dataset

	WB-PESQ	STOI(%)	SI-SDR	DNSMOS
Noisy	1.295	78.683	3.250	1.384
Uformer	1.996	86.851	7.745	2.698
MTFAA	1.792	84.869	5.391	2.510
CTFUNet	**2.164**	**88.645**	**8.008**	**2.805**
	SRMR	**CD**	**LLR**	**SNR$_{fw}$**
Noisy	3.149	4.911	0.870	9.814
Uformer	**6.340**	3.350	0.672	9.407
MTFAA	5.333	3.270	0.476	9.818
CTFUNet	5.948	**2.738**	**0.346**	**11.717**

Finally, we generate a test set containing noise and reverberation to evaluate the denoising and dereverberation performance of CTFUNet simultaneously. The test set contains noisy-reverberant speech with RT60 range of 0.4 s to 1.4 s and SNR ranging from −5 to 15 dB. The results of Table 11 show that CTFUNet has notable advantages in all metrics except SRMR, which is basically consistent with previous experimental results. Compared with unprocessed noisy-reverberation speech, CTFUNet improves WB-PESQ by 0.869 and DNSMOS by 1.421 and decreases CD by 2.173 and LLR by 0.524. To observe the speech enhancement effect of CTFUNet more intuitively, Fig. 5 illustrates the comparison results of the unprocessed speech spectrogram and the enhanced speech spectrogram. Obviously, the enhanced speech spectrogram of CTFUNet is similar to the clean speech spectrogram, demonstrating that CTFUNet can effectively suppress noise and reverberation. In addition, we visualize the learned attention matrix of each layer, as shown in Fig. 6. This picture shows that MCHCA has learned the correlation between different channels and is able to utilize global contextual information.

Based on the results of all experiments, we can conclude that CTFUNet can effectively improve the clarity and intelligibility of speech under different noise and reverberation levels.

6 Conclusions

Noise and reverberation seriously affect the quality and intelligibility of speech. To address this issue, we propose CTFUNet, a speech enhancement model that adopts a typical encoder-decoder framework. We mainly use the

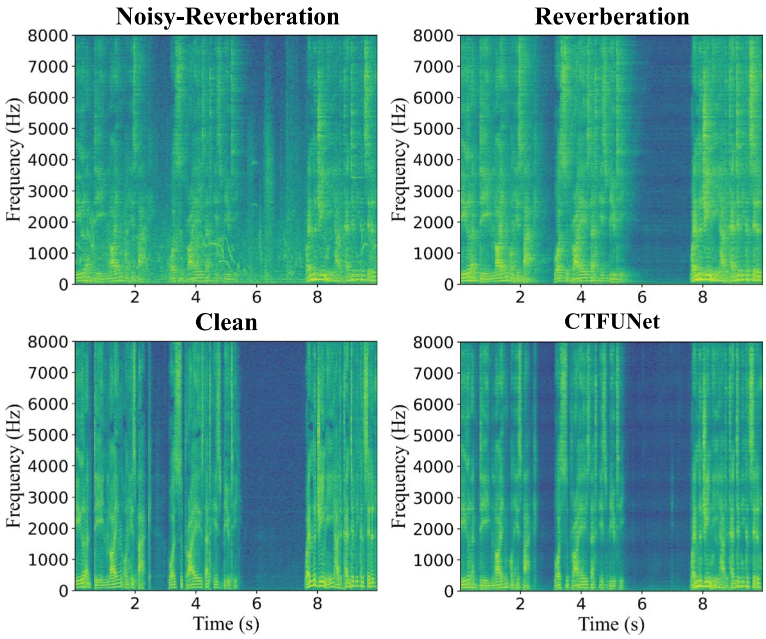

Fig. 5 Comparison of the unprocessed speech spectrogram and the enhanced speech spectrogram

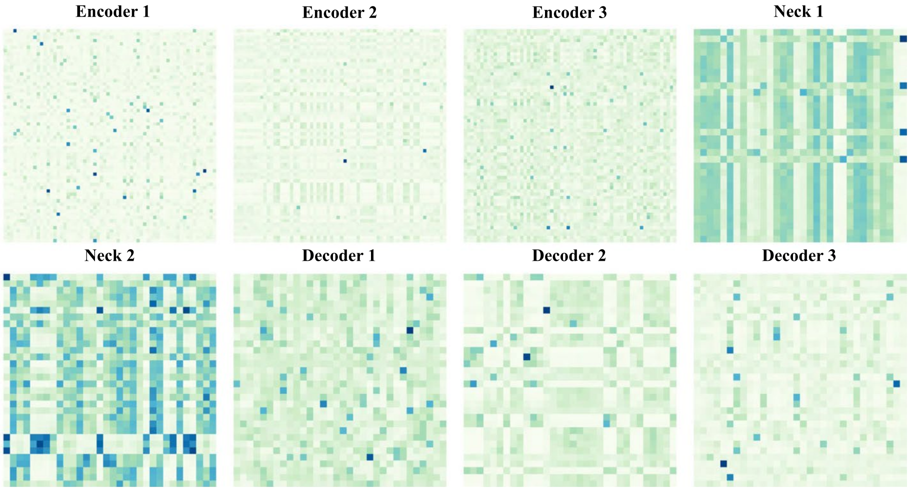

Fig. 6 The learned attention matrix of each layer

temporal-frequency convolution module and the multi-conv head channel attention with linear complexity to extract the temporal-frequency features of the signal. We use the residual channel attention module to capture the signal's channel features. Additionally, we introduce the channel temporal-frequency skip connection to mitigate the information loss problem in the process of down-sampling and up-sampling. Experimental results demonstrate that CTFUNet can effectively suppress different levels of noise and reverberation, exhibiting excellent speech enhancement performance.

Abbreviations

CTFUNet	Channel and temporal-frequency attention UNet
cIRM	Complex ideal ratio mask
DNNs	Deep neural networks
TCNN	Temporal convolutional neural network
Conv-TasNet	Full-convolution time-domain speech separation network
CRN	Convolutional recurrent network
IBM	Ideal binary mask
IRM	Ideal ratio mask
LSTM	Long short-term memory
MCHCA	Multi-conv head channel attention
RCAM	Residual channel attention module
CTFSC	Channel temporal-frequency skip connection
RNNs	Recurrent neural networks

TCN	Temporal convolutional network
CNN	Convolutional neural network
PE	Phase encoder
FD	Frequency down-sampling
FU	Frequency up-sampling
TFCM	Temporal-frequency convolution module
TFCN	Temporal-frequential convolutional network
cMSE	Complex mean squared error
RIRs	Room impulse responses
DNS	Deep noise suppression
SNR	Signal noise ratio
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility
SI-SDR	Scale-invariant signal-to-distortion ratio
DNSMOS	Deep noise suppression mean opinion score
SRMR	Speech-to-reverberation modulation energy ratio
CD	Cepstrum distance
LLR	Log likelihood ratio
SNR_{fw}	Frequency-weighted segmental signal-to-noise ratio

Acknowledgements
The authors are very grateful to the editors and anonymous reviewers for their guidance and useful suggestions.

Authors' contributions
S.X. conceptualized the study and implemented the codebase, run the experiments, and wrote the initial draft of the manuscript. Z.Z. further refined the details of the model. S.X. and Z.Z. revised the manuscript. M.W. supervised the work. The authors read and approved the final manuscript.

Funding
This research was supported by the National Natural Science Foundation of China under Grant No. 62176102, the National Natural Science Foundation of China under Grant No. 62276076, and the Natural Science Foundation of Guangdong Province under Grant No. 2020B1515120.

Availability of data and materials
The datasets generated during and/or analyzed during the current study are available in the deep noise suppression challenge repository, https://github.com/microsoft/DNS-Challenge [43].

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 17 March 2023 Accepted: 24 July 2023
Published online: 14 August 2023

References

1. K. Hermus, P. Wambacq, H.V. Hamme, A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Adv. Signal Process (2007), pp. 1–15. https://doi.org/10.1155/2007/45821
2. S.F. Boll, in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79*, A spectral subtraction algorithm for suppression of acoustic noise in speech (IEEE, 1979)
3. EPHRAIM, Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process. **32**(6), 1109–1121 (1984)
4. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. Acoust. Speech Signal Proc. IEEE Trans. **33**(2), 443–445 (1985)
5. A.v.d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
6. D. Rethage, J. Pons, X. Serra, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, A wavenet for speech denoising (IEEE, 2018), pp. 5069–5073
7. D. Stoller, S. Ewert, S. Dixon, Wave-u-net: a multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185 (2018)
8. Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(8), 1256–1266 (2019)
9. D. Yin, C. Luo, Z. Xiong, W. Zeng, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, Phasen: a phase-and-harmonics-aware speech enhancement network (AAAI, 2020), pp. 9458–9465
10. S. Routray, Q. Mao, Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. Comput. Speech Lang. **71**, 101270 (2022)
11. K. Tan, D. Wang, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement (IEEE, 2019), pp. 6865–6869
12. K. Tan, D. Wang, Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 380–390 (2019)
13. J. Chen, D. Wang, Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. **141**(6), 4705–4714 (2017)
14. X. Hao, X. Su, R. Horaud, X. Li, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement (IEEE, 2021), pp. 6633–6637
15. J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, H. Meng, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement (IEEE, 2022), pp. 7857–7861
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process Syst. **30**, 1–15 (2017)
17. A. Pandey, D. Wang, Dense CNN with self-attention for time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1270–1279 (2021)
18. J. Kim, M. El-Khamy, J. Lee, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, T-GSA: transformer with gaussian-weighted self-attention for speech enhancement (IEEE, 2020), pp. 6649–6653
19. Y. Zhao, D. Wang, in *Proc. Interspeech 2020*, Noisy-Reverberant Speech Enhancement Using DenseUNet with Time-Frequency Attention (2020), pp. 3261–3265. https://doi.org/10.21437/Interspeech.2020-2952
20. G. Zhang, L. Yu, C. Wang, J. Wei, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Multi-scale temporal frequency convolutional network with axial attention for speech enhancement (IEEE, 2022), pp. 9122–9126
21. A. Graves, A.R. Mohamed, G. Hinton, in *IEEE International Conference on Acoustics*, Speech recognition with deep recurrent neural networks (IEEE, 2013)
22. J. Li, M.T. Luong, J. Dan, A hierarchical neural autoencoder for paragraphs and documents. Comput. Sci. **1**, 1106–1115 (2015)
23. S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
24. P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, Y. Liu, Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput. **24**, 16453–16482 (2020)
25. A. Pandey, D. Wang, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain (IEEE, 2019), pp. 6875–6879
26. J. Lin, A.J.d.L. van Wijngaarden, K.C. Wang, M.C. Smith, Speech enhancement using multi-stage self-attentive temporal convolutional networks. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3440–3450 (2021)
27. X. Jia, D. Li, Tfcn: temporal-frequential convolutional network for single-channel speech enhancement. arXiv preprint arXiv:2201.00480 (2022)

28. O. Ronneberger, P. Fischer, T. Brox, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, U-net: Convolutional networks for biomedical image segmentation (Springer, 2015), pp. 234–241
29. Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, L. Xie, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Uformer: a unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation (IEEE, 2022), pp. 7417–7421
30. H.S. Choi, S. Park, J.H. Lee, H. Heo, D. Jeon, K. Lee, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Real-time denoising and dereverberation with tiny recurrent U-Net (IEEE, 2021), pp. 5789–5793
31. A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., Conformer: convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
32. C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, Y. Lu, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, Interactive speech and noise modeling for speech enhancement (AAAI, 2021), pp. 14549–14557
33. X. Xiang, X. Zhang, H. Chen, A nested u-net with self-attention and dense connectivity for monaural speech enhancement. IEEE Signal Process. Lett. **29**, 105–109 (2021)
34. A. Li, C. Zheng, R. Peng, X. Li, On the importance of power compression and phase estimation in monaural speech dereverberation. JASA Express Lett. **1**(1), 014802 (2021)
35. Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Maxim: multi-axis MLP for image processing (IEEE, 2022), pp. 5769–5780
36. S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.H. Yang, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Restormer: efficient transformer for high-resolution image restoration (IEEE, 2022), pp. 5728–5739
37. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, in *Proceedings of the European conference on computer vision (ECCV)*, Image super-resolution using very deep residual channel attention networks (Springer, 2018), pp. 286–301
38. Z. Zhang, S. Xu, X. Zhuang, L. Zhou, H. Li, M. Wang, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Two-stage unet with multi-axis gated multilayer perceptron for monaural noisy-reverberant speech enhancement (IEEE, 2023), pp. 1–5
39. Z. Zhang, S. Xu, X. Zhuang, Y. Qian, L. Zhou, M. Wang, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Half-temporal and half-frequency attention u 2 net for speech signal improvement (IEEE, 2023), pp. 1–2
40. S. Woo, J. Park, J.Y. Lee, I.S. Kweon, in *Proceedings of the European conference on computer vision (ECCV)*, Cbam: convolutional block attention module (Springer, 2018), pp. 3–19
41. J. Hu, L. Shen, G. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Squeeze-and-excitation networks (IEEE, 2018), pp. 7132–7141
42. A. Li, W. Liu, X. Luo, C. Zheng, X. Li, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ICASSP 2021 deep noise suppression challenge: decoupling magnitude and phase optimization with a two-stage deep network (IEEE, 2021), pp. 6628–6632
43. C.K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, et al., The interspeech 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results. arXiv preprint arXiv:2005.13981 (2020)
44. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, in *ICASSP*, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs (IEEE, 2001), pp. 749–752
45. I. Rec, P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. Int. Telecommun. Union (2005)
46. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE/ACM Trans. Audio Speech Lang. Process. **19**, 2125–2136 (2011)
47. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
48. C.K. Reddy, V. Gopal, R. Cutler, in *ICASSP*, DNSMOS p.835: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors (IEEE, 2022)
49. T.H. Falk, C. Zheng, W. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1766–1774 (2010)
50. J. Ma, Y. Hu, P. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. **125**, 3387–405 (2009)
51. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, L. Xie, in *Interspeech*, DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement (ISCA, 2020), pp. 2472–2476
52. S. Lv, Y. Hu, S. Zhang, L. Xie, in *Interspeech*, DCCRN+: channel-wise subband DCCRN with SNR estimation for speech enhancement (ISCA, 2021), pp. 2816–2820
53. U. Isik, R. Giri, N. Phansalkar, J. Valin, H. K., A. Krishnaswamy, in *Interspeech*, PoCoNet: better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss (ISCA, 2020), pp. 2487–2491
54. A. Li, W. Liu, C. Zheng, C. Fan, X. Li, Two heads are better than one: a two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1829–1843 (2021)
55. A. Li, C. Zheng, L. Zhang, X. Li, Glance and gaze: a collaborative learning framework for single-channel speech enhancement. Appl. Acoust. **187**, 108499 (2022)
56. J. Chen, w. Rao, z. Wang, z. Wu, Y. Wang, T. Yu, S. Shang, H. Meng, in *Interspeech*, Speech enhancement with fullband-subband cross-attention network (ISCA, 2022), pp. 976–980

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.