Zhang *et al.*
*EURASIP Journal on Audio, Speech, and Music Processing* (2024) 2024:20
https://doi.org/10.1186/s13636-024-00341-x

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

# Supervised Attention Multi-Scale Temporal Convolutional Network for monaural speech enhancement

Zehua Zhang[1], Lu Zhang[2], Xuyi Zhuang[1], Yukun Qian[1] and Mingjiang Wang[1*]

**Abstract**

Speech signals are often distorted by reverberation and noise, with a widely distributed signal-to-noise ratio (SNR). To address this, our study develops robust, deep neural network (DNN)-based speech enhancement methods. We reproduce several DNN-based monaural speech enhancement methods and outline a strategy for constructing datasets. This strategy, validated through experimental reproductions, has effectively enhanced the denoising efficiency and robustness of the models. Then, we propose a causal speech enhancement system named Supervised Attention Multi-Scale Temporal Convolutional Network (SA-MSTCN). SA-MSTCN extracts the complex compressed spectrum (CCS) for input encoding and employs complex ratio masking (CRM) for output decoding. The supervised attention module, a lightweight addition to SA-MSTCN, guides feature extraction. Experiment results show that the supervised attention module effectively improves noise reduction performance with a minor increase in computational cost. The multi-scale temporal convolutional network refines the perceptual field and better reconstructs the speech signal. Overall, SA-MSTCN not only achieves state-of-the-art speech quality and intelligibility compared to other methods but also maintains stable denoising performance across various environments.

**Keywords** Supervised attention, Monaural speech enhancement, Complex compressed spectrum, Complex ratio mask, Multi-scale temporal convolutional network

## 1 Introduction

Speech enhancement has numerous applications, including hearing aids, robust speech recognition, and video conferencing. The main objective of speech enhancement is to minimize background noise, thereby improving the quality and intelligibility of the enhanced speech. In real application scenarios such as video conferencing, the signal-to-noise ratio (SNR) of the speech signal is usually not very low, which requires the speech enhancement method to avoid causing distortion. In addition, the speech signal will be affected by reverberation, which requires speech enhancement methods for robust performance. Therefore, this study aims to explore how to build a training dataset for robust speech enhancement and propose a better monaural speech enhancement model with both performance and robustness.

Traditional single-channel speech enhancement methods such as spectral subtraction [1], Wiener filtering [2], and minimum mean squared error speech estimator [3, 4] often require estimation of the noise power spectral density (PSD) or the a priori SNR. These traditional methods are often effective in suppressing stationary noise. Whether using voice activity detector [5], minimum statistics [2, 6], or recursive averaging [7–9], it is difficult to estimate the noise PSD effectively under non-stationary noise conditions. Error in noise PSD estimation leads to enhanced speech containing residual noise

*Correspondence:
Mingjiang Wang
mjwang@hit.edu.cn
[1] Harbin Institute of Technology, Shenzhen, No. 6, Pingshan 1st Road, Taoyuan Street, Shenzhen 518000, Guangdong, China
[2] NIO Automobile Co., LTD, Lane 56, Antuo Road, Anting Town, Jiading District, Shanghai 201800, Shanghai, China

or speech distortion. This results in these methods being unable to process speech signals with non-stationary noise effectively.

Owing to the problems of traditional speech enhancement methods, including low upper-performance limits and difficulty in handling non-stationary noise, some researchers apply deep neural networks (DNNs) to speech enhancement [10–12] and achieve excellent performance. Zhang et al. [11] propose a new a priori SNR estimation structure called Deep Xi-TCN which contains a temporal convolutional network (TCN) [13, 14] with residual connections [15, 16]. For speech enhancement, they [10, 11] substitute the a priori SNR into a noise PSD estimator based on minimum mean square error (MMSE) called DeepMMSE. These methods cleverly combine traditional speech enhancement methods with DNNs and have the advantage of low computational cost. However, these methods [10, 11] do not provide an accurate estimate of the a posteriori SNR, nor do they enhance the noisy phase.

Some masking-based methods incorporating DNNs, such as the ideal binary mask (IBM) [17, 18] and ideal ratio mask (IRM) [19, 20], tend to mask the magnitude spectrum to denoise. Zhang et al. [21] propose joint log-power spectra (LPS) and IRM-based temporal convolutional network called multi-scale TCN (MSTCN). Unlike a traditional TCN, MSTCN stacks the input features forward into each residual block to enlarge and refine the receptive field of the model. Multi-objective learning enables the model to integrate the advantages of IRM and LPS, thereby further enhancing speech enhancement performance.

Magnitude masking-based methods do not consider the effect of phase information on speech enhancement performance, but studies [22–24] show that phase recovery contributes significantly to improving speech enhancement performance. Later, complex ratio mask (CRM) [25–27] and phase-sensitive mask (PSM) [28] estimation are used to enhance the complex spectrum in the frequency domain, to reconstruct the real and imaginary components of noisy speech. Hu et al. [26] propose a deep complex convolution recurrent network (DCCRN) capable of estimating CRM. To simulate the complex multiplication, they improve the convolutional recurrent network (CRN) using complex convolution and complex LSTM. Scale-invariant source-to-noise ratio (SI-SNR) is used as the loss function to replace the mean square error (MSE) loss. DCCRN achieves a very powerful performance and wins first place in the 1st deep noise suppression (DNS) challenge. However, a study [29] shows that within DCCRN, complex-valued DNNs and real-valued DNNs achieve similar performance, although complex-valued DNNs require more computational cost. Le et al. [27] extend the dual-path recurrent neural network (DPRNN) [30] and propose a dual-path convolution recurrent network (DPCRN) for estimating CRM. DPCRN replaced the recurrent neural network (RNN) in CRN with DPRNN modules and captured both temporal and frequency dependence. DPCRN has comparable performance to DCCRN and is ranked third in the 3rd DNS [31]. The advantages of DPCRN are that it includes only 0.8M model parameters and requires a much smaller number of multiply-accumulate operations (MACs) than DCCRN. Incorporating phase information enables the aforementioned models [25–28, 30] to achieve better performance than models using only the magnitude spectrum. Consequently, research on speech enhancement methods involving the phase spectrum and complex spectrum become more widespread.

There are also models [32, 33] that recover both noisy magnitude spectrum and noisy complex spectrum. Li et al. [32] propose a parallel structure for coarse and refined estimation named Glance and Gaze Network (GaGNet). GaGNet contains spectral feature extraction modules and multiple stacked Glance-Gaze modules (GGMs). The GGM is a dual structure in which the glance path masks the magnitude spectrum of noisy speech, and the gaze path compensates for the complex spectrum. Zhang et al. [33] propose a phase-aware dual-path dilated convolutional network (PhaseDCN) that estimates the complex spectrum and IRM. PhaseDCN interacts with information in a dual path using an attention-gating factor. Therefore, PhaseDCN can combine the magnitude and phase information of noisy speech for speech enhancement. Both GaGNet and PhaseDCN achieve good objective performance in the case of their smaller MACs.

Spectral mapping is a more direct way to reconstruct noisy speech. Tan and Wang [34] propose a novel CRN which integrates a convolutional encoder-decoder and LSTM for mapping the clean magnitude spectrum without using future information. Tan and Wang [35] propose an improved model of CRN called a gated convolutional recurrent neural network (GCRN) for mapping the complex spectrum. GCRN still employs the encoder-decoder architecture, with a dual-path decoder for estimating the enhanced complex spectrum. In addition, GCRN replaces 2-D convolution and deconvolution with gated linear unit blocks.

Another class of methods involves end-to-end speech enhancement [36–38] in the time domain, which can avoid additional short-time Fourier transform (STFT) and inverse STFT (iSTFT) operations. Luo and Mesgarani [36] propose Conv-TasNet, which uses dilated 1-D convolutional blocks instead of LSTM to improve model applicability. In Conv-TasNet, the mixture waveform is

modeled using a convolutional encoder-decoder architecture, which consists of an encoder with non-negativity constraints on its output and a linear decoder that inverts the encoder output back to the sound waveform. Evaluated in terms of both objective distortion measurements and listeners' subjective quality assessments, Conv-TasNet exceeds several ideal temporal-frequency amplitude masks in two-speaker speech separation and speech enhancement [39] tasks. As attention has attracted substantial interest in the deep learning field, Pandey and Wang [37] propose a dense CNN with self-attention (DenseCNN). DenseCNN utilizes an encoder-decoder architecture with skip connections and comprises a dense block and attention block at each layer of the encoder-decoder. In addition, sub-pixel convolution is used to avoid checkerboard artifacts in the output signal. Compared to spectral magnitude loss, phase-constrained magnitude loss offered better estimation for both noise spectrum and clean spectrum. Therefore, phase-constrained magnitude loss [37] enhances objective performance while reducing the issue of artifacts.

Our study compares experiments on reverberation, dataset duration, and language type. This leads to the development of a dataset construction strategy that improves model robustness. Following this, we introduce our causal speech enhancement model. Building on our previous research on Multi-Scale Temporal Convolutional Networks (MSTCN) [21], we find that refining the time-frequency (T-F) analysis granularity of features significantly improves both the performance and robustness of speech enhancement models.

We propose a model known as supervised attention multi-scale TCN (SA-MSTCN) for monaural speech enhancement. SA-MSTCN comprises two stages: the masking stage and the compensation stage. In the masking stage, we introduce gated TCN and a novel supervised attention U$^2$-LSTM (SAU$^2$-LSTM) for fixed-length and dynamic long-term modeling. Both the magnitude compressed spectrum (MCS) and complex compressed spectrum (CCS) are inputted into these long-term modeling modules for feature extraction. MSTCN then analyzes the extracted features to obtain CRM, which enhances the complex spectrum. The compensation stage aims to further suppress residual noise and recover spectral details, utilizing another U$^2$-LSTM to refine the masking stage enhanced spectrum. Compared with models like DCCRN, GCRN, and ConvTasNet, our model shows excellent speech quality and intelligibility and exhibits stronger generalization capabilities.

The rest of this paper is organized as follows. In Section 2, the proposed SA-MSTCN is introduced in detail, including the supervised attention network U$^2$-LSTM, multi-scale temporal convolutional module (MSTCM) and CCS. In Section 3, the experimental setup, baseline model, and training strategies are described. Section 4 discusses the effects of language, duration, and reverberation on model robustness. In Section 5, ablation studies and comparative experiments are performed to inform the model design. Finally, conclusions are presented in Section 6.

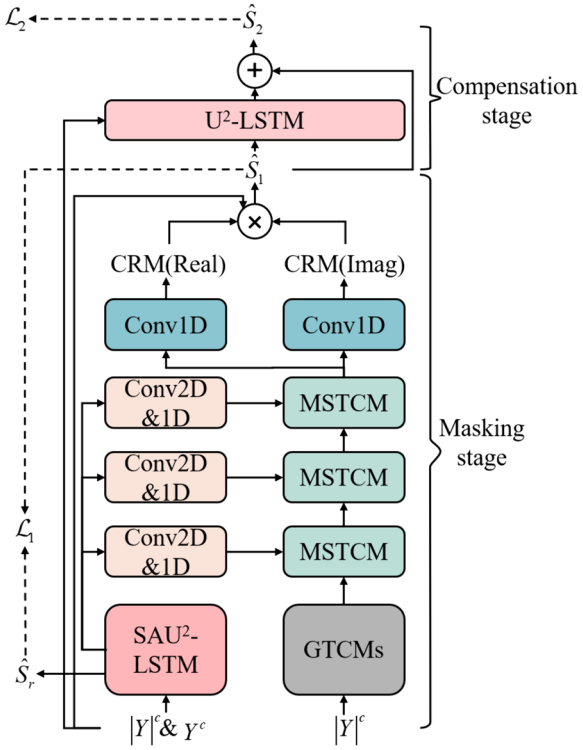
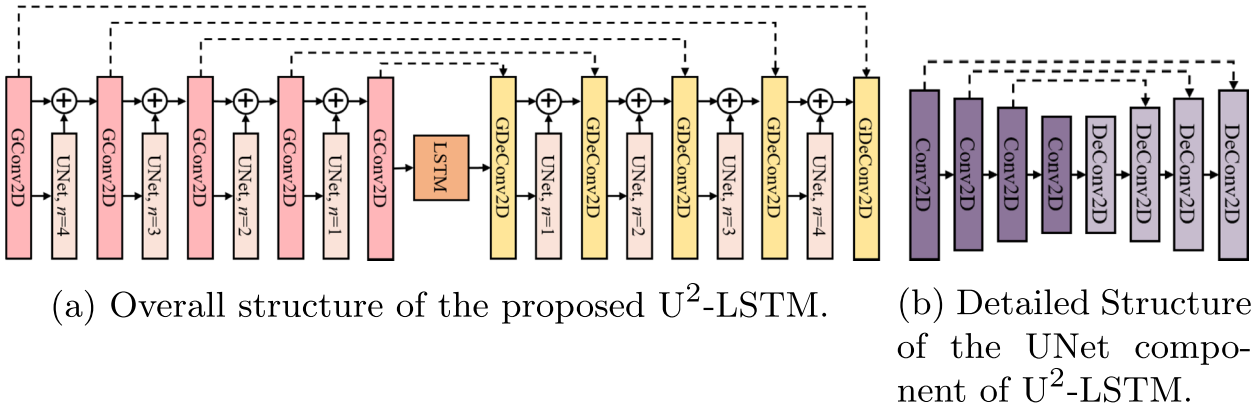



**Fig. 1** Illustration of the proposed two-stage speech enhancement method

## 2 Proposed Supervised Attention Multi-Scale TCN for speech enhancement

In this section, we introduce the details of the proposed SA-MSTCN. As shown in Fig. 1, SA-MSTCN includes a masking stage and a compensation stage and four module types: U$^2$-LSTM, SAU$^2$-LSTM, gated temporal convolutional module (GTCM), and MSTCM. The training process of SA-MSTCN is conducted in two steps. In the first step, only the parameters of the masking stage are updated. In the second step, the parameters of the masking stage are frozen, and then the parameters of the compensation stage are updated. In the masking stage, we implement a time-dependent feature extraction strategy for both CCS and MCS. Here, SAU$^2$-LSTM is utilized for dynamic temporal feature extraction of CCS and MCS, along with fixed-size temporal feature extraction specifically for MCS. The output from the last MSTCM




(a) Overall structure of the proposed U$^2$-LSTM. (b) Detailed Structure of the UNet component of U$^2$-LSTM.

**Fig. 2** Proposed encoder-decoder network U$^2$-LSTM. Dashed lines indicate connected features in the channel dimension

undergoes convolution by two 1-D convolution layers, each with a kernel size of 1, to derive the real and imaginary components of the CRM. With post-processing, we can calculate the enhanced complex spectrum.

Given that the enhanced complex spectrum may still contain residual noise or distortions, the speech quality, and intelligibility are further refined in the compensation stage. In the compensation stage, both the enhanced complex spectrum and the noisy complex spectrum are inputted into U$^2$-LSTM for computing the compensation values. The final compensated complex spectrum is obtained by summing the enhanced complex spectrum with these compensation values.

The specifics of the four modules-CCS, loss function, and post-processing-will be elaborated in the subsequent subsections, providing a comprehensive understanding of each component's role and functionality in the system.

### 2.1 Compressed complex spectrum

Usually, complex spectrum is used as the input for complex spectrum masking or mapping. However, paper [40, 41] found that compressing the complex spectrum resulted in better speech quality and intelligibility. The specific procedure for compressing the complex spectrum is as follows, where $y(t)$, $s(t)$, and $n(t)$ respectively denote noisy speech, clean speech, and noise in the time domain.

Assuming that noise is additive, noisy speech can be obtained according to the following equation:

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

The complex spectrum can be obtained by applying the STFT on Eq. (1).

$$Y(k,l) = S(k,l) + N(k,l) \tag{2}$$

where $k$ and $l$ indicate the frequency and frame index of the STFT. The complex spectrum $Y(k, l)$ can be rewritten as:

$$Y(k,l) = |Y(k,l)|\exp(i\theta_Y(k,l)) \tag{3}$$

where $|Y(k, l)|$ and $\theta_Y$ represent the magnitude spectrum and phase spectrum, respectively. The MCS can be obtained by performing exponential operations on the magnitude spectrum $|Y(k,l)|^c = |Y(k,l)|^{0.3}$. The MCS is used to calculate the CCS via Eq. (4).

$$Y^c(k,l) = |Y(k,l)|^c \frac{Y(k,l)}{\max(|Y(k,l)|, \delta)} \tag{4}$$

where $Y^c(k,l)$ and $\delta$ denote the CCS and a very small constant, respectively. The real and imaginary parts of the CCS and the MCS are used as channels into 2-D convolution.

### 2.2 U$^2$-LSTM

Inspired by U$^2$-Net [42], a similar topology named U$^2$-LSTM is proposed as shown in Fig. 2a to capture the temporal dependence, in which GConv2D and GDeConv2D represent gated 2-D convolution and gated 2-D deconvolution, respectively. The specific structure of the UNet component is shown in Fig. 2b, where $n$ denotes the number of each 2-D convolution or 2-D deconvolution. Instance normalization and a parametric rectified linear unit (PReLU) are added after Conv2D and DeConv2D. The dashed lines between GConv2D/Conv2D and GDeConv2D/DeConv2D in Fig. 2 indicate connections in the channel dimension. The first GConv2D and the last GDeConv2D have a convolution kernel size of $2 \times 5$, and the rest are $2 \times 3$. The convolution kernel size for Conv2D and Deconv2D is $1 \times 3$. The number of output channels for all layers is 64. Following the last GConv2D, RNN

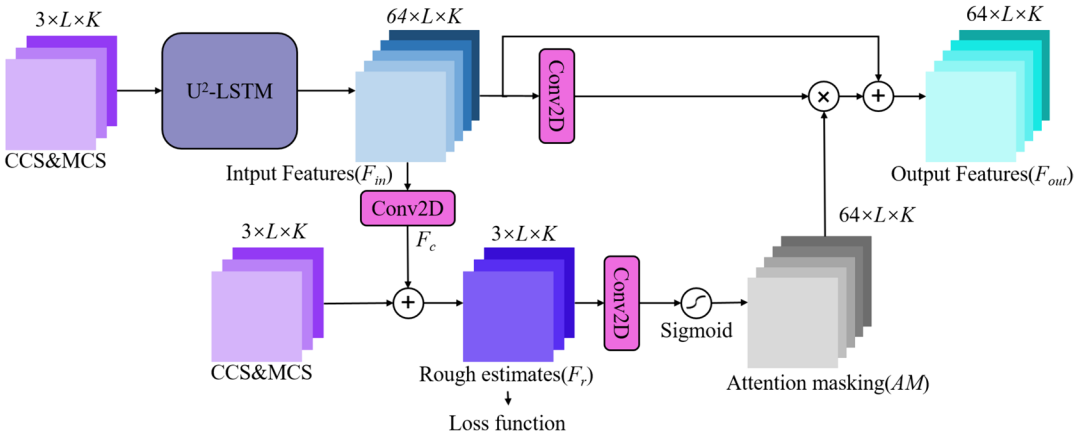



**Fig. 3** Principles of supervised attention in the U$^2$-LSTM

components are used to process the temporal aspects of the audio data. A 4-layer LSTM with a hidden size of 256 is added in the middle of U$^2$-Net.

### 2.3 Supervised attention U$^2$-LSTM

DNNs can be seen as a black box, and the feature information it extracts is often difficult to interpret. We propose a supervised attention structure to ensure that the feature extraction process aligns more closely with our expectations. This structure is modified from Zamir et al. [43], as shown in Fig. 3. U$^2$-LSTM performs feature extraction on the CCS and MCS to obtain input features $F_{in} \in \mathbb{R}^{64\times L\times K}$, where $L$ denotes the number of frames, $K$ denotes the number of frequency bins, and 64 is the number of channels. A 2-D convolution is performed to generate the compensation value $F_c \in \mathbb{R}^{3\times L\times K}$. The compensation value $F_c$ is summed with the CCS and MCS to obtain the rough enhancement spectrum $F_r \in \mathbb{R}^{3\times L\times K}$, which is fed into the loss function for supervision. The attention masking $AM \in \mathbb{R}^{64\times L\times K}$ is generated by performing a 2-D convolution and sigmoid on $F_r$. The result is then used to recalibrate $F_{in}$ to obtain attention-guided features. The calibrated features $F_{out} \in \mathbb{R}^{64\times L\times K}$ are supplied to the next stage for processing. Here, the three 2-D convolutional kernels are all of size $1 \times 1$.

### 2.4 Gated TCM

To compensate for the insufficient dimensionality of the input features of the first MSTCM, GTCMs are added to extract the MCS, as shown in Fig. 4. GTCMs consist of multiple gated TCNs with varying dilation rates. GTCMs are lightweight and easy to implement. In Fig. 4, $k$ denotes the convolution kernel size, $d$ denotes the dilation rate, $I$ denotes the number of input channels, and $O$ denotes the number of output channels. Three GTCMs are used, each of which stacks six gated TCNs with different dilation rates growing exponentially from $2^0$ to $2^5$. For each GTCM, instance normalization and PReLU are applied before the second and subsequent 1-D convolutions.

### 2.5 Multi-scale TCN

Since our previous studies [21] have shown that a multi-scale approach to refine the receptive fields will help improve speech reconstruction, we propose a simple and effective multi-scale subband analysis method as shown in Fig. 5.

Each MSTCM stacks five causal MSTCNs with a convolutional kernel size of $k = 3$ and dilation rates $d = 1, 3, 5, 7, 11$, respectively. To compress the feature dimension of SAU$^2$-LSTM output $F_{out} \in \mathbb{R}^{64\times L\times K}$, Conv2D &1D is used. For 2-D convolution, there are 64 input channels and 6 output channels, and the kernel size is $1 \times 1$, so the output feature is $F_{Conv2D} \in \mathbb{R}^{6\times L\times K}$. We reshape $F_{Conv2D} \in \mathbb{R}^{6\times L\times K}$ as $F_{Conv2DRe} \in \mathbb{R}^{6K\times L}$. The reshaped features $F_{Conv2DRe}$ are supplied as input to a 1-D convolution with 256 output channels and a kernel size of 1 to generate the output feature $F_{Conv1D} \in \mathbb{R}^{256\times L}$. All MSTCNs after the first will receive the output features from the previous MSTCN as input, and these features will be compressed into $F_{Pre} \in \mathbb{R}^{256\times L}$ by a 1-D convolution with a kernel size of 1. $F_{Conv1D}$ concatenates with $F_{Pre}$ to create a new feature $F_{cat} \in \mathbb{R}^{512\times L}$. The concatenated features $F_{cat}$ will be divided into eight subbands of equal length $F_{sub,i=0,1,\ldots,7}$ resulting from multi-scale analysis. As shown in Fig. 5, an MSTCN contains left and right branches, each of which has eight dilated 1-D convolutions [44] with $I$ input channels and $O$ output channels. Each MSTCN receives the output of the previous dilated 1-D convolution, and the current subband features $F_{sub,i}$ as input. Batch normalization [45], a rectified linear unit (ReLU) activation function, and dropout [46]

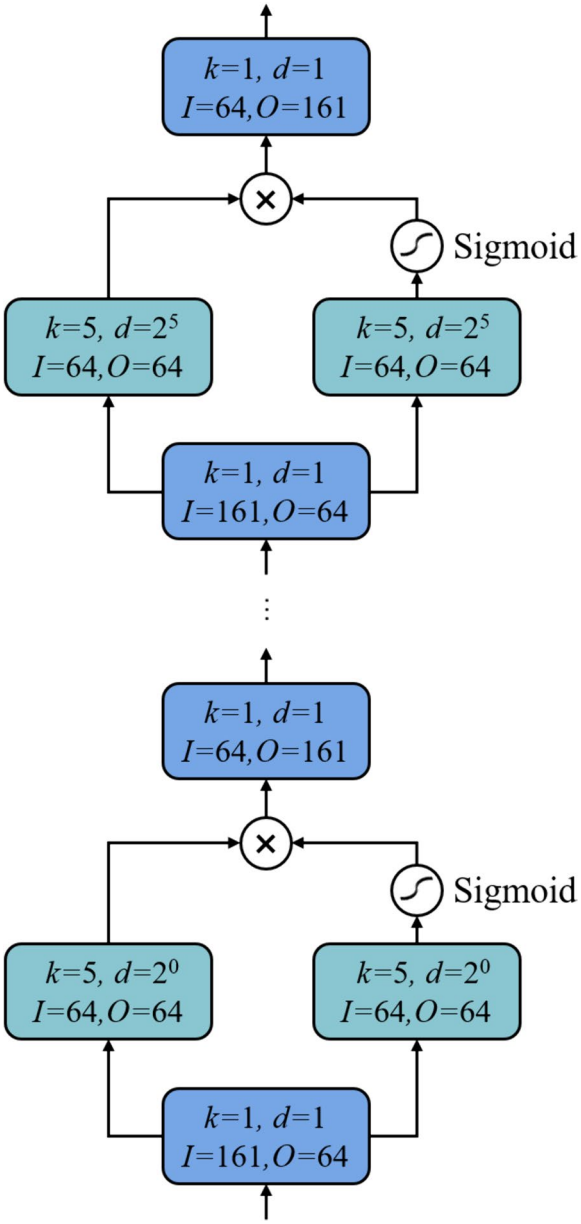



**Fig. 4** Gated temporal convolution module

are used after each dilated 1-D convolution to enhance model capability and avoid overfitting. Before 1-D convolution, the output features of the left and right branches are added to produce the output of the MSTCN.

### 2.6 Loss function

The loss function of many models [27, 32] directly calculates the mean square error between the enhanced complex spectrum and the clean complex spectrum. In the masking stage, to supervise the feature extraction, we train the model using a supervised attention complex compressed loss function :

$$\mathcal{L}_1 = \frac{\beta}{K \times L}\left(\alpha \sum_{k,l}\left|S^c - \hat{S}_1^c\right|^2 + (1-\alpha)\sum_{k,l}\left||S|^c - |\hat{S}_1|^c\right|^2\right) + \frac{1-\beta}{K \times L}\left(\alpha \sum_{k,l}\left|S^c - \hat{S}_r^c\right|^2 + (1-\alpha)\sum_{k,l}\left||S|^c - |\hat{S}_r|^c\right|^2\right) \tag{5}$$

where $\hat{S}_r^c$ denotes the rough estimate of the enhanced CCS via SAU$^2$-LSTM, and $\hat{S}_1^c$ denotes the enhanced CCS after the masking stage. $\alpha$ and $\beta$ are coefficients, which in this study are respectively assigned values of 0.3 and 0.8. In the compensation stage, because supervised attention is no longer required, we use the following loss function:

$$\mathcal{L}_2 = \frac{1}{K \times L}\left(\alpha \sum_{k,l}\left|S^c - \hat{S}_2^c\right|^2 + (1-\alpha)\sum_{k,l}\left||S|^c - |\hat{S}_2|^c\right|^2\right) \tag{6}$$

where $\hat{S}_2^c$ denotes the final enhanced CCS after the masking and compensation stages. In the loss function, $.^c$ represents the compressed spectrum, and its calculation method can refer to Eq. 4.

### 2.7 Post-processing for signal reconstruction

Inspired by Hu et al. [26], instead of multiplying the CRM and CCS directly, we use the following method to enhance the complex spectrum. The estimated CRM can be expressed as:

$$\hat{M}^c(k,l) = |\hat{M}(k,l)|^c \exp(i\hat{\theta}_M^c(k,l)) \tag{7}$$

$\hat{S}_1(k,l)$ represents the enhanced complex spectrum after the masking stage, which can be calculated according to the following equation:

$$\hat{S}_1(k,l) = \tanh(|\hat{M}(k,l)|^c)|Y(k,l)|\exp(i(\hat{\theta}_M^c(k,l) + \theta_Y(k,l))) \tag{8}$$

We use the tanh activation function to limit the magnitude mask to the range 0 to 1 and then compensate for the noisy phase with the masking phase.

The enhanced complex spectrum after the compensation stage can be expressed as follows, where $\hat{C}(k,l)$ denotes the compensation value of the compensation stage:

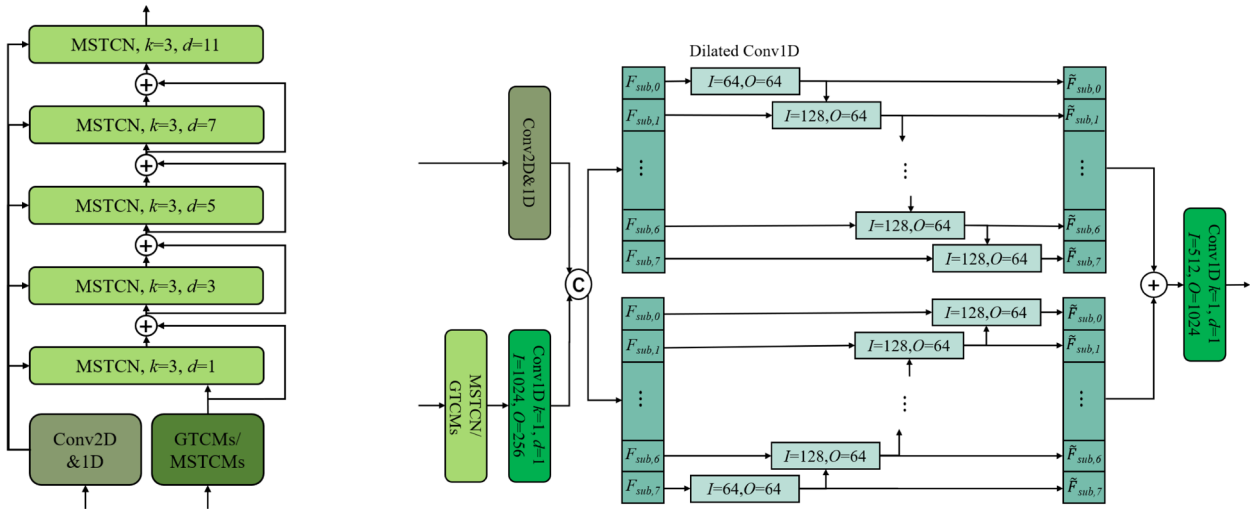
$$\hat{S}_2(k,l) = \hat{S}_1(k,l) + \hat{C}(k,l) \tag{9}$$

This compensation further improves the quality and intelligibility of the enhanced speech.

## 3 Experimental setup

### 3.1 Dataset construction

To ensure that the training data is sufficiently rich, we use the dataset [31] of the 3rd DNS challenge for training and




(a) Overall structure of the proposed MSTCM.

(b) Detailed structure of the MSTCN component of the MSTCM.

**Fig. 5** Proposed multi-scale temporal convolution module

testing. For the dataset construction, we select English and Chinese, two of the most widely spoken languages globally. Because the audio quality of the original DNS dataset is uneven, we clean the DNS dataset. Audio with a lower prior SNR often contains noise. We use a trained Deep-Xi model to estimate the average prior SNR for each audio segment. To enhance the quality of the clean speech dataset, we remove the bottom 20% of audio files with the lowest average prior SNR. The cleaned dataset is divided into three parts: 80% for training, 10% for validation, and 10% for testing. The final training set contains 337 h of English audio, 146 h of Chinese audio, and 147 h of noise audio. The validation set and test set each include 42 h of English audio, 18 h of Chinese audio, and 18 h of noise audio. To simulate a wide range of scenarios from extreme noise conditions to relatively quiet environments, we mix speech and noise at random SNRs ranging from -5 dB to 20 dB. In addition, to account for reverberation effects in the real environment, a portion of the clean speech is blended with both synthetic and realistic room impulse responses (RIRs) provided by the 3rd DNS dataset before being mixed with noise signals. The reverberation time $T_{60}$ is between 0.3 and 1.3 s.

Given the focus of this study on training duration, reverberation, and the impact of different languages on speech enhancement models, we construct multiple datasets in Section 4. In Section 4.1, to investigate the influence of different languages on model robustness, we construct two training datasets, one containing English data and the other containing both Chinese and English data. Both training datasets are 500 h long and include no reverberation. In Section 4.2, we explore the impact of training dataset size on model performance by constructing four datasets of varying durations (100, 500, 1000, and 1500 h) with Chinese and English audio. Training datasets with and without reverberation are constructed to compare the effect of reverberation on model robustness in Section 4.3. The two training datasets contain English and Chinese data with a total duration of 500 h.

### 3.2 Training details

All clean speech and noise audio are sampled at 16 kHz. In SA-MSTCN, the frame length is 20 ms with 50% frameshift, and the Hamming window is used before applying the STFT.

All parameters in the model are randomly initialized, and the Adam algorithm [47] is used as the optimizer. The initial learning rate is 0.001, and the learning rate is halved when the loss stops decreasing for three training epochs. When the learning rate of the masking stage decreases to 0.0001, the parameters of the masking stage are frozen, and the compensation stage is updated. When the compensation stage learning rate decays to 0.0001, the network parameters stop updating.

### 3.3 Baseline models

We selecte eight state-of-the-art models for comparison from current speech enhancement methods, including

magnitude spectral masking, complex spectral masking, and time domain mapping. For magnitude spectral masking, we chose CRN, MSTCN, and LSTM-IRM. LSTM-IRM is the baseline model we build, containing two LSTM layers with a hidden dimension of 1024 and one fully connected layer. GCRN, GaGNet, DPCRN, and DCCRN are chosen for comparison for complex spectral masking speech enhancement. Conv-TasNet, which performed speech enhancement in the time domain, is also chosen as a comparison model. All baseline models are implemented officially.

### 3.4 Evaluation metrics

To verify the validity of the model structure and to compare the performance of the models, we evaluate the models using the following metrics:

PESQ (perceptual evaluation of speech quality) [48]: This is the most commonly used objective metric for evaluating speech quality and uses clean speech as the standard for evaluating enhanced speech. PESQ scores range from − 0.5 to 4.5, with higher scores indicating better voice quality.

STOI (short-time objective intelligibility) [49]: This is a widely used objective metric for evaluating speech intelligibility and has a strong correlation with the intelligibility of speech. STOI scores range from 0 to 1, with higher scores indicating higher intelligibility of speech.

SDR (signal to distortion ratio) [50]: This metric evaluates the distortion of the speech signal in the time domain. It measures the ratio of the energy of clean speech to the energy of distortion, with higher scores indicating smaller amounts of distortion.

OUTE (optimal unit training epoch): This is a new metric we defined to compare the relative time taken to train a model to the optimum with training datasets of varying sizes. We define the time to train a model in the 100-h training dataset for one epoch as the unit training epoch. Similarly, the time to train a model in the 500-h training dataset for one epoch is recorded as five training epochs. With this metric, we can compare the relative time to train each model for different training dataset sizes.

## 4 Building a training dataset for real scenarios

In general, the components of the training dataset have a substantial impact on the application of the model. Many experiments are conducted in this section to make the model robust. The following subsections discuss the effects of language, duration, and reverberation on the training datasets.

### 4.1 Language of the training dataset

As we all know, every place has its own language, and each language features unique characteristics. To prevent speech enhancement models from failing with unseen languages, this subsection discusses the impact of the language used in the training dataset on the model's robustness. We construct two different training datasets for comparison: one exclusively containing English and the other containing both Chinese and English. The models are tested using datasets in English, Chinese, and a mix of English and Chinese, as well as with unseen French, Spanish, and Japanese. The findings are comprehensively presented in Tables 1 and 2. In these tables, "Mix" denotes a dataset that incorporates both English and Chinese. The term "English→Mix" indicates the variation in evaluation metrics as the training dataset shifts from English to a mixed language format. A positive difference indicates that the mixed dataset yields better results, while a negative value indicates that the English dataset performs better. SA-MSTCN[1] and SA-MSTCN[2] denote the enhanced speech after the masking and compensation stages, respectively.

The experimental results show that models trained on the English and Mix training datasets exhibit comparable performance on the English test datasets. However, a noticeable performance gap becomes evident when these models are tested using the Chinese dataset. This is attributed to the fact that the model trained solely on the English dataset has not been exposed to Chinese, resulting in a significant performance decline when processing Chinese speech. On the other hand, the model trained on the Mix dataset, having been exposed to Chinese, maintains robust performance on the Chinese test dataset.

To further explore the impact of language diversity in the training dataset, we test the models on French, Spanish, and Japanese languages, none of which had been previously encountered by the models. Interestingly, likely due to the similar characteristics between French and English, the performance of the models trained on both the English and Mix datasets is nearly identical to the French test dataset. However, when tested with the Spanish and Japanese datasets, the model trained on the Mix dataset performs better than the one trained solely on English.

In conclusion, broadening the linguistic diversity of the training dataset seems to enhance the robustness and generalizability of the model to a certain degree. In subsequent experiments, the training and testing datasets include Chinese and English.

**Table 1** Comparison of average PESQ, STOI, and SDR in different languages

| | Training dataset | English | | | Mix | | | English→Mix | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Test dataset** | **Metrics** | **PESQ** | **STOI (%)** | **SDR** | **PESQ** | **STOI (%)** | **SDR** | **ΔPESQ** | **ΔSTOI (%)** | **ΔSDR** |
| English | Unprocessed | 1.96 | 92.6 | 13.63 | 1.96 | 92.6 | 13.63 | - | - | - |
| | CRN | 2.41 | 94.8 | 18.49 | 2.45 | 95.0 | 18.56 | 0.04 | 0.2 | 0.07 |
| | MSTCN | 2.68 | 94.8 | 16.40 | 2.66 | 95.1 | 16.32 | − 0.02 | 0.3 | − 0.08 |
| | LSTM-IRM | 2.80 | 95.9 | 18.84 | 2.80 | 96.0 | 18.92 | 0.00 | 0.1 | 0.08 |
| | GCRN | 2.76 | 95.5 | 19.98 | 2.76 | 95.5 | 19.98 | 0.00 | 0.0 | 0.00 |
| | GaGNet | 2.89 | 96.1 | 20.43 | 2.92 | 96.1 | 20.35 | 0.03 | 0.0 | − 0.08 |
| | Conv-TasNet | 3.07 | 96.6 | 21.67 | 2.93 | 96.2 | 20.94 | − 0.14 | − 0.4 | − 0.73 |
| | DCCRN | 3.17 | 96.6 | 21.20 | 3.14 | 96.5 | 20.57 | − 0.03 | − 0.1 | − 0.63 |
| | DPCRN | 3.16 | 96.6 | 20.75 | 3.11 | 96.4 | 20.59 | − 0.05 | − 0.2 | − 0.16 |
| | SA-MSTCN[1] | 3.26 | 96.6 | 20.11 | 3.30 | 96.8 | 20.57 | 0.04 | 0.2 | 0.46 |
| | SA-MSTCN[2] | 3.29 | 96.7 | 20.46 | 3.35 | 97.0 | 21.13 | 0.06 | 0.3 | 0.67 |
| Chinese | Unprocessed | 2.31 | 88.2 | 16.81 | 2.31 | 88.2 | 16.81 | - | - | - |
| | CRN | 2.44 | 86.8 | 16.77 | 2.76 | 89.4 | 20.30 | 0.32 | 2.6 | 3.53 |
| | MSTCN | 2.26 | 86.9 | 14.94 | 2.99 | 91.5 | 17.45 | 0.73 | 4.6 | 2.51 |
| | LSTM-IRM | 2.80 | 90.4 | 18.37 | 3.15 | 92.6 | 20.97 | 0.35 | 2.2 | 2.60 |
| | GCRN | 3.09 | 90.2 | 21.46 | 3.09 | 90.2 | 21.45 | 0.00 | 0.0 | − 0.01 |
| | GaGNet | 2.77 | 89.1 | 18.02 | 3.16 | 91.2 | 21.38 | 0.41 | 2.1 | 3.36 |
| | Conv-TasNet | 2.95 | 90.5 | 20.12 | 3.16 | 91.1 | 21.39 | 0.21 | 0.6 | 1.27 |
| | DCCRN | 2.66 | 89.4 | 17.98 | 3.43 | 92.7 | 21.70 | 0.77 | 3.3 | 3.72 |
| | DPCRN | 3.01 | 90.7 | 18.99 | 3.40 | 92.5 | 22.49 | 0.39 | 1.8 | 3.50 |
| | SA-MSTCN[1] | 3.21 | 92.2 | 20.74 | 3.58 | 93.6 | 21.92 | 0.37 | 1.4 | 1.18 |
| | SA-MSTCN[2] | 3.26 | 92.3 | 21.00 | 3.60 | 93.8 | 22.21 | 0.34 | 1.5 | 1.21 |
| Mix | Unprocessed | 2.08 | 91.7 | 14.81 | 2.08 | 91.7 | 14.81 | - | - | - |
| | CRN | 2.55 | 93.8 | 19.29 | 2.55 | 93.8 | 19.29 | 0.00 | 0.0 | 0.00 |
| | MSTCN | 2.63 | 93.5 | 16.40 | 2.77 | 94.3 | 16.82 | 0.14 | 0.8 | 0.42 |
| | LSTM-IRM | 2.82 | 94.6 | 19.26 | 2.90 | 95.2 | 19.90 | 0.08 | 0.6 | 0.64 |
| | GCRN | 2.86 | 94.4 | 20.81 | 2.85 | 94.4 | 20.83 | − 0.01 | 0.0 | 0.02 |
| | GaGNet | 2.83 | 94.0 | 19.84 | 2.98 | 94.9 | 21.04 | 0.15 | 0.9 | 1.20 |
| | Conv-TasNet | 2.96 | 94.6 | 20.55 | 2.99 | 95.0 | 21.50 | 0.03 | 0.4 | 0.95 |
| | DCCRN | 3.08 | 95.0 | 20.65 | 3.22 | 95.7 | 21.48 | 0.14 | 0.7 | 0.83 |
| | DPCRN | 3.14 | 95.2 | 20.98 | 3.19 | 95.6 | 21.53 | 0.05 | 0.4 | 0.55 |
| | SA-MSTCN[1] | 3.24 | 95.7 | 20.98 | 3.38 | 96.1 | 21.45 | 0.14 | 0.4 | 0.47 |
| | SA-MSTCN[2] | 3.26 | 95.8 | 21.30 | 3.41 | 96.2 | 21.95 | 0.15 | 0.4 | 0.65 |

### 4.2 Duration of the training dataset

Most DNNs-based speech enhancement methods are data-driven, and the richness of the dataset greatly impacts model performance. To explore the number of training hours needed to saturate model performance, we train baseline models using datasets containing 100, 500, 1000, and 1500 h of audio data, with the results shown in Table 3.

As expected, smaller training datasets, such as the 100-h dataset, struggle to bring out the model's full potential. Both CRN and the compensation stage of SA-MSTCN prove almost ineffective with small datasets, indicating that such datasets are not suitable for ablation studies. The performance of the models markedly improves when the training dataset reaches 500 h. However, the increment in performance is smaller when expanding the dataset from 500 to 1000 h. When the training dataset is extended to 1500 h, some models continue to show performance improvements, while others have already reached saturation or even show degradation.

The evaluation metric OUTE shows that most models have similar training durations with 500-h and 1000-h datasets. However, with a 1500-h dataset, the models

**Table 2** Comparison of average PESQ, STOI, and SDR in different languages

| | Training dataset | English | | | Mix | | | English→Mix | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Test dataset** | **Metrics** | **PESQ** | **STOI (%)** | **SDR** | **PESQ** | **STOI (%)** | **SDR** | **ΔPESQ** | **ΔSTOI (%)** | **ΔSDR** |
| French | Unprocessed | 2.19 | 92.4 | 15.53 | 2.19 | 92.4 | 15.53 | - | - | - |
| | CRN | 2.54 | 93.1 | 19.28 | 2.63 | 93.8 | 19.77 | 0.09 | 0.7 | 0.49 |
| | MSTCN | 2.91 | 94.2 | 17.51 | 2.89 | 94.3 | 17.54 | − 0.02 | 0.1 | 0.03 |
| | LSTM-IRM | 2.97 | 94.9 | 20.26 | 2.98 | 94.9 | 20.35 | 0.01 | 0.0 | 0.09 |
| | GCRN | 2.93 | 94.2 | 20.98 | 2.93 | 94.2 | 20.98 | 0.00 | 0.0 | 0.00 |
| | GaGNet | 3.03 | 94.8 | 21.50 | 3.08 | 94.9 | 21.66 | 0.05 | 0.1 | 0.16 |
| | Conv-TasNet | 3.14 | 95.2 | 22.34 | 3.06 | 94.7 | 21.95 | − 0.08 | − 0.5 | − 0.39 |
| | DCCRN | 3.26 | 95.6 | 22.42 | 3.28 | 95.8 | 22.56 | 0.02 | 0.2 | 0.14 |
| | DPCRN | 3.24 | 95.6 | 21.84 | 3.25 | 95.7 | 22.22 | 0.01 | 0.1 | 0.36 |
| | SA-MSTCN[1] | 3.35 | 95.5 | 21.34 | 3.38 | 95.8 | 21.93 | 0.03 | 0.3 | 0.59 |
| | SA-MSTCN[2] | 3.36 | 95.7 | 21.56 | 3.40 | 96.0 | 22.28 | 0.04 | 0.3 | 0.72 |
| Spanish | Unprocessed | 2.24 | 93.6 | 15.48 | 2.24 | 93.6 | 15.48 | - | - | - |
| | CRN | 2.59 | 94.2 | 18.39 | 2.67 | 95.1 | 19.21 | 0.08 | 0.9 | 0.82 |
| | MSTCN | 2.84 | 94.2 | 16.35 | 2.85 | 95.2 | 16.69 | 0.01 | 1.0 | 0.34 |
| | LSTM-IRM | 2.91 | 95.5 | 19.17 | 2.95 | 95.7 | 19.68 | 0.04 | 0.2 | 0.51 |
| | GCRN | 2.87 | 95.5 | 20.37 | 2.87 | 95.5 | 20.37 | 0.00 | 0.0 | 0.00 |
| | GaGNet | 2.91 | 95.7 | 20.16 | 3.01 | 96.0 | 20.57 | 0.10 | 0.3 | 0.41 |
| | Conv-TasNet | 3.05 | 96.0 | 20.85 | 2.99 | 93.6 | 20.74 | − 0.06 | − 2.4 | − 0.11 |
| | DCCRN | 3.18 | 96.4 | 21.07 | 3.23 | 96.6 | 21.55 | 0.05 | 0.2 | 0.48 |
| | DPCRN | 3.20 | 96.4 | 21.09 | 3.21 | 96.6 | 21.49 | 0.01 | 0.2 | 0.40 |
| | SA-MSTCN[1] | 3.28 | 96.5 | 21.03 | 3.31 | 96.6 | 21.22 | 0.03 | 0.1 | 0.19 |
| | SA-MSTCN[2] | 3.30 | 96.6 | 21.59 | 3.33 | 96.7 | 21.69 | 0.03 | 0.1 | 0.10 |
| Japanese | Unprocessed | 1.96 | 92.3 | 13.75 | 1.96 | 92.3 | 13.75 | - | - | - |
| | CRN | 2.33 | 92.8 | 17.77 | 2.34 | 93.1 | 17.86 | 0.01 | 0.3 | 0.09 |
| | MSTCN | 2.42 | 93.0 | 15.60 | 2.49 | 93.2 | 15.75 | 0.07 | 0.2 | 0.15 |
| | LSTM-IRM | 2.63 | 94.1 | 18.28 | 2.65 | 94.3 | 18.37 | 0.02 | 0.2 | 0.09 |
| | GCRN | 2.52 | 93.5 | 18.88 | 2.57 | 93.5 | 18.89 | 0.05 | 0.0 | 0.01 |
| | GaGNet | 2.59 | 93.7 | 17.87 | 2.67 | 93.8 | 19.00 | 0.08 | 0.1 | 1.13 |
| | Conv-TasNet | 2.71 | 94.1 | 19.68 | 2.66 | 93.7 | 19.23 | − 0.05 | − 0.4 | − 0.45 |
| | DCCRN | 2.83 | 94.2 | 19.70 | 2.88 | 94.4 | 19.93 | 0.05 | 0.2 | 0.23 |
| | DPCRN | 2.90 | 94.8 | 19.67 | 2.91 | 94.7 | 19.85 | 0.01 | 0.1 | 0.18 |
| | SA-MSTCN[1] | 2.95 | 94.8 | 18.98 | 2.97 | 94.9 | 19.13 | 0.02 | 0.1 | 0.15 |
| | SA-MSTCN[2] | 2.97 | 94.9 | 19.12 | 3.00 | 95.0 | 19.56 | 0.03 | 0.1 | 0.44 |

require longer to converge. Consequently, a training dataset of 500 to 1000 h emerges as a stable and cost-effective choice for constructing speech enhancement models. As models trained with the 500-h dataset sacrifice only minimal performance and require less time to train, it is a preferable option for comparison experiments. Therefore, in this study, the 500-h training dataset is used for all experiments comparing baseline models.

### 4.3 Reverberation of the training dataset

In real environments, such as conference rooms, speech reverberation is a common and unavoidable phenomenon. Room impulse response (RIR) severely disrupts the resonant peak structure of speech, which can render speech enhancement algorithms ineffective. To explore the impact of reverberation on speech enhancement model performance, we train the baseline model using training datasets with reverberation, without reverberation, and with half of the data containing reverberation. We evaluate the model using noisy-reverberant speech and noisy-anechoic speech, with test results as shown in Table 4.

When models trained with anechoic speech are tested on anechoic speech, there is a significant improvement

**Table 3** Comparison of average PESQ, STOI, and SDR for test datasets of different durations

| Metrics | PESQ | | | | STOI (%) | | | | SDR | | | | OUTE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Duration | 100 | 500 | 1000 | 1500 | 100 | 500 | 1000 | 1500 | 100 | 500 | 1000 | 1500 | 100 | 500 | 1000 | 1500 |
| Unprocessed | 2.08 | 2.08 | 2.08 | 2.08 | 91.7 | 91.7 | 91.7 | 91.7 | 14.81 | 14.81 | 14.81 | 14.81 | - | - | - | - |
| CRN | 2.13 | 2.55 | 2.59 | 2.64 | 91.3 | 93.8 | 94.0 | 94.1 | 17.69 | 19.29 | 19.38 | 19.58 | 75 | 335 | 340 | 330 |
| MSTCN | 2.67 | 2.77 | 2.78 | 2.80 | 93.7 | 94.3 | 94.5 | 94.4 | 15.873 | 16.82 | 17.16 | 17.07 | 59 | 230 | 540 | 690 |
| LSTM-IRM | 2.57 | 2.90 | 2.93 | 3.01 | 93.8 | 95.2 | 95.2 | 95.5 | 18.42 | 19.90 | 20.03 | 20.22 | 34 | 120 | 150 | 285 |
| GCRN | 2.55 | 2.85 | 2.91 | 2.96 | 92.9 | 94.4 | 94.6 | 94.9 | 18.84 | 20.83 | 20.99 | 21.40 | 93 | 355 | 400 | 525 |
| GaGNet | 2.67 | 2.98 | 2.98 | 3.02 | 93.4 | 94.9 | 95.0 | 95.1 | 19.51 | 21.04 | 21.14 | 21.44 | 50 | 230 | 260 | 300 |
| Conv-TasNet | 2.62 | 2.99 | 3.12 | 3.09 | 93.4 | 95.0 | 95.6 | 95.4 | 19.58 | 21.50 | 22.15 | 22.02 | 78 | 200 | 260 | 315 |
| DCCRN | 3.06 | 3.22 | 3.28 | 3.25 | 95.1 | 95.7 | 95.8 | 95.8 | 20.73 | 21.48 | 21.75 | 21.56 | 66 | 145 | 210 | 300 |
| DPCRN | 3.15 | 3.19 | 3.27 | 3.24 | 95.4 | 95.6 | 95.9 | 95.7 | 21.23 | 21.53 | 21.84 | 21.74 | 45 | 130 | 210 | 345 |
| SA-MSTCN[1] | 3.16 | 3.38 | 3.44 | 3.44 | 95.4 | 96.1 | 96.3 | 96.3 | 20.53 | 21.45 | 21.70 | 21.74 | 58 | 190 | 340 | 420 |
| SA-MSTCN[2] | 3.16 | 3.41 | 3.50 | 3.48 | 95.4 | 96.2 | 96.6 | 96.4 | 20.53 | 21.95 | 22.31 | 22.15 | 87 | 355 | 640 | 720 |

in objective metric scores across all models. However, when the test set's noisy speech is mixed with reverberation, the effectiveness of all models significantly decreases. Most models struggle in this scenario, showing limited noise suppression capability. As indicated in Table 4, models trained with datasets mixed with reverberation successfully suppress noise in noisy-reverberant speech. Interestingly, such training ensures that the models also performed well on noisy-anechoic speech. Although models trained with reverberant speech don't process noisy-anechoic speech as effectively as those trained with anechoic speech, the performance degradation was within acceptable limits. When the training dataset included half of the data with reverberation, the performance of most models is intermediate compared to those trained exclusively on datasets with or without reverberation. This offers a valuable balance, suggesting that training with half of the data containing reverberation can significantly improve the model's robustness.

## 5 Experiments and analysis

After determining a better strategy for building a training dataset, this section discusses the design of SA-MSTCN. The proposed SA-MSTCN is subjected to ablation studies of different component configurations and the performance is compared with many current state-of-the-art models.

### 5.1 Performance comparison for different component configurations

We perform the ablation study to verify the validity of each part of SA-MSTCN. We use MSTCM as the basis for gradually adding other modules. The training and test datasets are the same as those in Section 4.3, and both contain reverberation. The results of this ablation study are shown in Table 5. MSTCMs using MACs at a rate of only 1.02 G/s outperform GCRN and GaGNet and are comparable to Conv-TasNet, establishing it as a highly competitive module. The addition of the $U^2$-LSTM module compensates for the inability of MSTCMs to capture temporal dependency. With the addition of $U^2$-LSTM, the PESQ increases by 0.33, and STOI increases by 1.6%. The supervised attention mechanism is a very cost-effective module, which significantly improves speech quality and intelligibility in exchange for only 0.01 M increase in model parameters and a 0.08 G/s increase in MACs. Incorporating GTCMs slightly increases the parameters and MACs but continues to improve all three performance metrics. The most resource-intensive configuration with the highest parameters and MACs also shows the best performance metrics. Table 5 also presents an alternative configuration path, starting with MSTCMs, then adding GTCMs, followed by $U^2$-LSTM, supervised attention, and compensation in sequence. Each addition seems to follow a similar trend of increasing computational cost for improved performance metrics.

### 5.2 Performance comparison for different loss functions

The choice of loss function in training speech enhancement models is a critical decision that affects various aspects of model performance. In this subsection, we discuss the impact of the proposed loss function on model performance, with the results presented in Table 6. Here, MSE indicates the substitution of uncompressed spectrum for the compressed spectrum in $\mathcal{L}_1$ and $\mathcal{L}_2$. The versions of SA-MSTCN1 ($\mathcal{L}_1$) and SA-MSTCN2 ($\mathcal{L}_2$) outperform their MSE counterparts, indicating that the choice of loss function has a notable impact on the model's performance.

**Table 4** Comparison of the average PESQ, STOI, and SDR for test datasets with and without reverberation

| | Test dataset | No RIR | | | RIR | | |
|---|---|---|---|---|---|---|---|
| **Training dataset** | **Metrics** | **PESQ** | **STOI (%)** | **SDR** | **PESQ** | **STOI (%)** | **SDR** |
| No RIR | Unprocessed | 2.08 | 91.7 | 14.81 | 2.24 | 88.5 | 14.81 |
| | CRN | 2.55 | 93.8 | 19.29 | 2.18 | 88.7 | 15.76 |
| | MSTCN | 2.77 | 94.3 | 16.82 | 2.52 | 90.1 | 14.36 |
| | LSTM-IRM | 2.90 | 95.2 | 19.90 | 2.71 | 91.6 | 16.81 |
| | GCRN | 2.85 | 94.4 | 20.82 | 2.37 | 89.1 | 16.13 |
| | GaGNet | 2.98 | 94.9 | 21.04 | 2.47 | 89.5 | 16.55 |
| | Conv-TasNet | 2.99 | 95.0 | 21.50 | 2.44 | 89.3 | 16.31 |
| | DCCRN | 3.22 | 95.7 | 21.48 | 2.49 | 90.4 | 16.43 |
| | DPCRN | 3.19 | 95.6 | 21.53 | 2.71 | 91.6 | 17.53 |
| | SA-MSTCN[1] | 3.38 | 96.1 | 21.45 | 2.74 | 91.4 | 17.21 |
| | SA-MSTCN[2] | 3.41 | 96.2 | 21.95 | 2.71 | 91.3 | 17.24 |
| RIR | CRN | 2.43 | 93.3 | 18.75 | 2.59 | 90.7 | 18.39 |
| | MSTCN | 2.59 | 93.6 | 16.19 | 2.75 | 91.6 | 15.93 |
| | LSTM-IRM | 2.83 | 95.0 | 19.70 | 3.02 | 93.2 | 19.31 |
| | GCRN | 2.68 | 93.6 | 19.75 | 2.84 | 91.8 | 19.08 |
| | GaGNet | 2.69 | 93.8 | 19.87 | 2.86 | 91.6 | 19.49 |
| | Conv-TasNet | 2.93 | 94.8 | 21.08 | 3.03 | 92.5 | 20.22 |
| | DCCRN | 3.00 | 94.9 | 21.16 | 3.15 | 93.0 | 20.30 |
| | DPCRN | 2.98 | 94.9 | 20.58 | 3.24 | 93.3 | 20.14 |
| | SA-MSTCN[1] | 3.24 | 95.7 | 20.99 | 3.44 | 94.3 | 20.61 |
| | SA-MSTCN[2] | 3.26 | 95.8 | 21.30 | 3.47 | 94.3 | 20.83 |
| Half RIR and half no RIR | CRN | 2.50 | 93.5 | 19.02 | 2.58 | 90.7 | 18.40 |
| | MSTCN | 2.69 | 94.0 | 16.54 | 2.75 | 91.6 | 15.87 |
| | LSTM-IRM | 2.92 | 95.2 | 19.90 | 3.01 | 93.1 | 19.33 |
| | GCRN | 2.72 | 94.0 | 20.09 | 2.84 | 91.4 | 19.14 |
| | GaGNet | 2.91 | 94.0 | 20.87 | 2.84 | 91.3 | 19.22 |
| | Conv-TasNet | 2.94 | 94.8 | 21.23 | 3.02 | 92.4 | 20.18 |
| | DCCRN | 3.16 | 95.2 | 21.35 | 3.15 | 92.9 | 20.11 |
| | DPCRN | 3.09 | 95.2 | 20.93 | 3.20 | 93.1 | 20.03 |
| | SA-MSTCN[1] | 3.32 | 95.9 | 21.26 | 3.42 | 94.3 | 20.55 |
| | SA-MSTCN[2] | 3.36 | 96.0 | 21.41 | 3.46 | 94.3 | 20.79 |

**Table 5** Experimental results of combining different modules with MSTCM

| | **#Param. (M)** | **MACs (G/s)** | **PESQ** | **STOI (%)** | **SDR** |
|---|---|---|---|---|---|
| Unprocessed | - | - | 2.24 | 88.5 | 14.81 |
| MSTCMs | 10.13 | 1.02 | 2.99 | 92.3 | 19.08 |
| +$U^2$-LSTM | 23.54 | 8.26 | 3.32 | 93.9 | 20.16 |
| +Supervised Attention | 23.55 | 8.34 | 3.39 | 94.1 | 20.45 |
| +GTCMs | 24.66 | 8.45 | 3.44 | 94.3 | 20.61 |
| +Compensation | 28.12 | 12.07 | 3.47 | 94.3 | 20.83 |
| MSTCMs | 10.13 | 1.02 | 2.99 | 92.3 | 19.08 |
| +GTCMs | 11.24 | 1.13 | 3.16 | 92.9 | 19.75 |
| +$U^2$-LSTM | 24.65 | 8.37 | 3.37 | 94.1 | 20.43 |
| +Supervised Attention | 24.66 | 8.45 | 3.44 | 94.3 | 20.61 |
| +Compensation | 28.12 | 12.07 | 3.47 | 94.3 | 20.83 |

**Table 6** Experimental results of different loss function

| | PESQ | STOI (%) | SDR |
|---|---|---|---|
| Unprocessed | 2.24 | 88.5 | 14.81 |
| SA-MSTCN1 (MSE) | 3.35 | 93.8 | 20.02 |
| SA-MSTCN1 ($\mathcal{L}_1$) | 3.44 | 94.3 | 20.61 |
| SA-MSTCN2 (MSE) | 3.39 | 93.9 | 20.24 |
| SA-MSTCN2 ($\mathcal{L}_2$) | 3.47 | 94.3 | 20.83 |

model performance and robustness, two test datasets are constructed, with and without reverberation. The SNR of the two test datasets ranged from − 5 dB to 20 dB in increments of 5 dB, with 1 h of noisy speech at each level. Some test demos are available at the link[1].

As shown in Table 7, compared with the baseline models, the proposed SA-MSTCN shows a significant improvement in PESQ scores, especially in low-SNR conditions. While most models show the greatest improvement in PESQ scores for noisy speech between 0 and 10 dB, the proposed SA-MSTCN demonstrates a substantial PESQ enhancement across this SNR range. ΔAvg. represents the average difference between enhanced speech and unprocessed speech. ΔAvg. with RIR is almost identical to ΔAvg. without RIR, indicating that the training dataset with reverberation allows the model to process

**Table 7** Average PESQ scores of compared methods for noisy and enhanced speech under various SNR conditions

| | No RIR | | | | | | | RIR | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** |
| Unprocessed | 1.36 | 1.57 | 1.82 | 2.23 | 2.73 | 3.22 | - | 1.42 | 1.68 | 1.97 | 2.44 | 2.98 | 3.47 | - |
| CRN | 1.74 | 2.06 | 2.32 | 2.65 | 2.95 | 3.23 | 0.33 | 1.84 | 2.20 | 2.49 | 2.85 | 3.17 | 3.45 | 0.34 |
| MSTCN | 1.77 | 2.14 | 2.46 | 2.87 | 3.27 | 3.60 | 0.53 | 1.86 | 2.26 | 2.62 | 3.07 | 3.49 | 3.83 | 0.53 |
| LSTM-IRM | 1.99 | 2.38 | 2.71 | 3.11 | 3.46 | 3.75 | 0.74 | 2.12 | 2.56 | 2.90 | 3.34 | 3.71 | 4.00 | 0.78 |
| GCRN | 1.98 | 2.34 | 2.62 | 2.93 | 3.20 | 3.42 | 0.59 | 2.04 | 2.44 | 2.76 | 3.13 | 3.42 | 3.65 | 0.60 |
| GaGNet | 1.93 | 2.30 | 2.59 | 2.94 | 3.23 | 3.49 | 0.59 | 2.03 | 2.44 | 2.76 | 3.14 | 3.47 | 3.74 | 0.60 |
| Conv-TasNet | 2.13 | 2.52 | 2.81 | 3.15 | 3.46 | 3.70 | 0.80 | 2.17 | 2.58 | 2.92 | 3.32 | 3.65 | 3.91 | 0.76 |
| DCCRN | 2.16 | 2.59 | 2.92 | 3.29 | 3.59 | 3.85 | 0.91 | 2.24 | 2.70 | 3.07 | 3.48 | 3.81 | 4.06 | 0.90 |
| DPCRN | 2.17 | 2.58 | 2.89 | 3.24 | 3.54 | 3.79 | 0.88 | 2.37 | 2.83 | 3.18 | 3.56 | 3.87 | 4.10 | 0.99 |
| SA-MSTCN[1] | 2.40 | 2.84 | 3.15 | 3.49 | 3.77 | 3.98 | 1.11 | **2.63** | **3.10** | 3.40 | 3.75 | 4.01 | 4.21 | 1.19 |
| SA-MSTCN[2] | **2.43** | **2.87** | **3.18** | **3.51** | **3.78** | **3.99** | **1.13** | **2.63** | **3.10** | **3.43** | **3.77** | **4.03** | **4.22** | **1.20** |

**Table 8** Average STOI (%) scores of compared methods for noisy and enhanced speech under various SNR conditions

| | No RIR | | | | | | | RIR | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** |
| Unprocessed | 83.1 | 88.4 | 91.7 | 94.7 | 96.8 | 98.1 | - | 77.1 | 83.9 | 88.2 | 92.3 | 95.2 | 97.1 | - |
| CRN | 86.7 | 91.1 | 93.5 | 95.4 | 96.8 | 97.6 | 1.4 | 82.2 | 87.7 | 90.7 | 93.4 | 95.3 | 96.5 | 2.0 |
| MSTCN | 87.0 | 91.4 | 93.8 | 95.9 | 97.3 | 98.1 | 1.8 | 83.2 | 88.7 | 91.7 | 94.4 | 96.3 | 97.6 | 3.0 |
| LSTM-IRM | 89.5 | 93.2 | 95.2 | 96.9 | 98.0 | 98.7 | 3.2 | 86.0 | 90.7 | 93.3 | 95.6 | 97.1 | 98.2 | 4.5 |
| GCRN | 87.7 | 91.7 | 93.8 | 95.6 | 96.7 | 97.5 | 1.7 | 83.4 | 88.5 | 91.3 | 93.7 | 95.3 | 96.4 | 2.4 |
| GaGNet | 89.5 | 91.6 | 93.9 | 95.8 | 97.1 | 98.0 | 2.2 | 83.3 | 88.6 | 91.5 | 94.2 | 95.9 | 97.2 | 2.8 |
| Conv-TasNet | 89.7 | 93.2 | 95.0 | 96.6 | 97.6 | 98.4 | 2.9 | 85.4 | 90.1 | 92.6 | 95.0 | 96.6 | 97.7 | 3.9 |
| DCCRN | 89.3 | 93.1 | 95.1 | 96.8 | 97.9 | 98.6 | 3.0 | 85.5 | 90.4 | 93.0 | 95.4 | 96.9 | 98.0 | 4.2 |
| DPCRN | 89.2 | 92.9 | 94.9 | 96.6 | 97.7 | 98.5 | 2.9 | 86.1 | 90.8 | 93.3 | 95.6 | 97.1 | 98.2 | 4.5 |
| SA-MSTCN[1] | 90.6 | 94.0 | 95.8 | 97.2 | 98.2 | **98.8** | 3.7 | **87.9** | **92.1** | 94.3 | 96.2 | **97.6** | 98.4 | 5.4 |
| SA-MSTCN[2] | **90.7** | **942** | **95.9** | **97.3** | **98.3** | **98.8** | **3.8** | **87.9** | **92.1** | **94.4** | **96.3** | **97.6** | **98.5** | **5.5** |

### 5.3 Performance comparison with baseline models

In this subsection, we compare the proposed SA-MSTCN with eight baseline models. All models are trained on a 500-h training dataset with reverberation. To compare

[1] https://hitsziot.github.io/2024/02/20/SAMSTCN/

**Table 9** Average SDR scores of compared methods for noisy and enhanced speech under various SNR conditions

| | No RIR | | | | | | | RIR | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** | **− 5 dB** | **0 dB** | **5 dB** | **10 dB** | **15 dB** | **20 dB** | **ΔAvg.** |
| Unprocessed | 3.79 | 8.84 | 12.84 | 17.84 | 22.84 | 27.84 | - | 3.73 | 8.77 | 12.77 | 17.77 | 22.77 | 27.77 | - |
| CRN | 11.24 | 14.77 | 17.49 | 20.90 | 24.24 | 27.43 | 3.68 | 10.88 | 14.36 | 17.09 | 20.52 | 23.97 | 27.25 | 3.41 |
| MSTCN | 9.79 | 13.22 | 15.74 | 18.51 | 20.71 | 22.23 | 1.03 | 9.44 | 12.88 | 15.41 | 18.26 | 20.60 | 22.27 | 0.88 |
| LSTM-IRM | 11.69 | 15.33 | 18.20 | 21.83 | 25.47 | 29.18 | 4.61 | 11.47 | 15.03 | 17.90 | 21.58 | 25.45 | 29.45 | 4.55 |
| GCRN | 12.71 | 16.06 | 18.64 | 21.79 | 24.81 | 27.55 | 4.59 | 11.91 | 15.27 | 17.89 | 21.09 | 24.22 | 27.01 | 3.97 |
| GaGNet | 12.27 | 15.62 | 18.30 | 21.55 | 24.79 | 28.09 | 4.43 | 12.15 | 15.49 | 18.13 | 21.54 | 25.02 | 28.68 | 4.57 |
| Conv-TasNet | 13.91 | 17.10 | 19.55 | 22.61 | 25.73 | 28.99 | 5.65 | 13.09 | 16.33 | 18.89 | 22.18 | 25.65 | 29.32 | 5.31 |
| DCCRN | 13.65 | **17.22** | **20.04** | **23.44** | **26.87** | **30.45** | **6.28** | 12.74 | 16.22 | 18.96 | 22.43 | 25.93 | 29.68 | 5.39 |
| DPCRN | 13.33 | 16.78 | 19.45 | 22.76 | 26.01 | 29.11 | 5.57 | 12.92 | 16.32 | 19.02 | 22.40 | 25.80 | 29.25 | 5.35 |
| SA-MSTCN[1] | 13.34 | 16.68 | 19.33 | 22.64 | 26.01 | 29.51 | 5.58 | **13.45** | **16.74** | 19.24 | 22.69 | 26.32 | 30.08 | 5.82 |
| SA-MSTCN[2] | **13.66** | 17.05 | 19.63 | 22.91 | 26.26 | 29.62 | 5.85 | **13.45** | **16.74** | **19.47** | **22.87** | **26.46** | **30.15** | **5.92** |

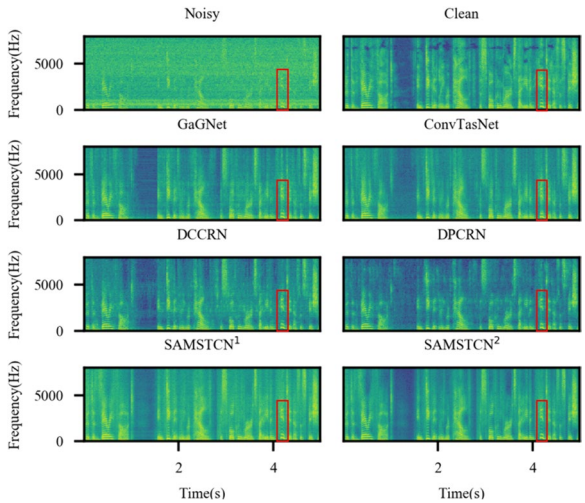



**Fig. 6** Spectrograms of noisy and clean speech, enhanced by GaGNet, Conv-TasNet, DCCRN, DPCRN, and SA-MSTCN

noisy speech with and without reverberation, attaining the same speech quality improvement for both.

From Table 8, it can be concluded that SA-MSTCN achieves a significantly better STOI score than the baseline models. When the SNR of noisy speech is 20 dB, many models can no longer improve speech intelligibility or even reduce it, but under these conditions, the STOI of the proposed SA-MSTCN improves by 0.008. Unlike the PESQ scores, the difference between ΔAvg. with and without RIR indicates a more significant improvement in intelligibility when the model processes speech with reverberation.

As shown in Table 9, DCCRN achieves the highest SDR score for noisy speech without reverberation, and SA-MSTCN reaches a more advantageous SDR score for noisy speech with reverberation. Similar to the PESQ scores, ΔAvg. is very similar with RIR and without RIR for all models.

**Table 10** Comparison of parameter counts and multiply-accumulate operations. Here, ✓ indicates causal models

| | **Causal** | **#Param. (M)** | **MACs (G/s)** |
|---|---|---|---|
| CRN | ✓ | 17.58 | 2.54 |
| MSTCN | ✓ | 4.78 | 0.48 |
| LSTM-IRM | ✓ | 13.24 | 1.34 |
| GCRN | ✓ | 9.77 | 2.47 |
| GaGNet | ✓ | 5.94 | 1.63 |
| Conv-TasNet | ✓ | 5.00 | 5.23 |
| DCCRN | ✓ | 3.67 | 14.38 |
| DPCRN | ✓ | 0.80 | 3.17 |
| SA-MSTCN[1] | ✓ | 24.66 | 8.45 |
| SA-MSTCN[2] | ✓ | 28.12 | 12.07 |

In addition, we plot the spectrograms of clean speech, noisy speech, and speech enhanced by GaGNet, Conv-TasNet, DCCRN, DPCRN, and SA-MSTCN, as shown in Fig. 6. The spectrograms show that Conv-TasNet, DCCRN, and DPCRN suppress the high-frequency part of the noisy speech signal significantly, but the proposed SA-MSTCN recovers better. Compared to the masking stage, the compensation stage enables the effective recovery of over-masked speech signals.

The number of parameters and MACs for each model are shown in Table 10. Because more 1-D convolutions are employed in the proposed SA-MSTCN, the number of parameters is greater than the other models. Compared to DCCRN, the first stage of SA-MSTCN has significantly fewer MACs but achieves better results

for most objective evaluation metrics. The number of MACs of the two-stage SA-MSTCN are similar to those of DCCRN, and the speech quality and distortion are further improved than SA-MSTCN[1]. Combining the number of parameters, MACs, and performance, SA-MSTCN[1] is more cost-effective, and SA-MSTCN[2] has stronger performance.

## 6 Conclusion

Our findings highlight the critical role of the training dataset's composition in enhancing the model's robustness. To improve the performance of the model's robustness, the training dataset should include reverberation, multiple languages, and a duration of more than 500 h. This study proposes a causal monaural speech enhancement method called supervised attention multi-scale temporal convolutional network (SA-MSTCN), which learns the CRM from the complex compressed spectrum. The model takes full advantage of convolution and LSTM in local modeling and long-term modeling. The proposed supervised attention mechanism achieves a performance improvement at a very small cost. SA-MSTCN is associated with significant PESQ and STOI improvement in both high-SNR and low-SNR environments compared to other state-of-the-art models. The robustness and generalizability of SA-MSTCN, bolstered by our proposed dataset construction approach, ensure consistent performance across unseen languages and reverberations. Further reducing the parameters and computational cost, exploring the application of SA-MSTCN in real life is the next step to be studied.

**Acknowledgements**
Thanks to Professor Mingjiang Wang for his support. Thanks to all editors and reviewers for their suggestions and efforts.

**Authors' contributions**
Zhang, Z. conceptualized the study and implemented the codebase, and wrote the manuscript. Zhang, L. further improved the details of the model. Zhuang, X. and Qian, Y. revised the manuscript and integrated experimental data. Wang, M. supervised the work. All authors read and approved the final manuscript.

**Funding**
This work was supported in part by the National Natural Science Foundation of China under Grant No.62276076, the National Natural Science Foundation of China under Grant No.62176102, and the Natural Science Foundation of Guangdong Province under Grant No.2020B1515120004.

**Availability of data and materials**
The datasets are available in the Deep Noise Suppression Challenge [31] repository: https://github.com/microsoft/DNS-Challenge.

## Declarations

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

Received: 13 November 2023 Accepted: 25 March 2024
Published online: 11 April 2024

**References**

1. R. Martin, Spectral subtraction based on minimum statistics. Power **6**(8), 1182–1185 (1994)
2. P. Scalart, J. Filho, in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. Speech enhancement based on a priori signal to noise estimation. IEEE Atlanta (1996), p. 629–632
3. Y. Ephraim, D. Malah, 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. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
5. J.H. Chang, N.S. Kim, S. Mitra, Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. **54**(6), 1965–1976 (2006)
6. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
7. I. Cohen, B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. **9**(1), 12–15 (2002)
8. I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
9. S. Rangachari, P.C. Loizou, A noise-estimation algorithm for highly non-stationary environments. Speech Commun. **48**(2), 220–231 (2006)
10. A. Nicolson, K.K. Paliwal, Deep learning for minimum mean-square error approaches to speech enhancement. Speech Commun. **111**, 44–55 (2019)
11. Q. Zhang, A. Nicolson, M. Wang, K.K. Paliwal, C. Wang, Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1404–1415 (2020)
12. A. Nicolson, K.K. Paliwal, Masked multi-head self-attention for causal speech enhancement. Speech Commun. **125**, 80–96 (2020)
13. 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)
14. 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)
15. Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recogn. **90**, 119–133 (2019)
16. M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K.K. Paliwal, F. Shang, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. Deep residual-dense lattice network for speech enhancement. AAAI, New York (2020), p. 8552–8559
17. Z. Jin, D. Wang, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*. A supervised learning approach to monaural segregation of reverberant speech. IEEE, Honolulu (2007), p. IV–921–IV–924
18. G. Kim, Y. Lu, Y. Hu, P.C. Loizou, An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. **126**, 1486–1494 (2009)
19. S. Srinivasan, N. Roman, D. Wang, Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. **48**, 1486–1501 (2006)
20. A. Narayanan, D. Wang, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ideal ratio mask estimation using deep neural networks for robust speech recognition. IEEE, Vancouver (2013), p. 7092–7096
21. L. Zhang, M. Wang, in *Interspeech 2020*. Multi-Scale TCN: Exploring Better Temporal DNN Model for Causal Speech Enhancement. ISCA, Shanghai (2020), p. 2672–2676
22. K. Paliwal, K. Wójcicki, B. Shannon, The importance of phase in speech enhancement. Speech Commun. **53**, 465–494 (2011)

23. E. Jokinen, M. Takanen, H. Pulakka, P. Alku, in *Interspeech*. Enhancement of speech intelligibility in near-end noise conditions with phase modification. ISCA, Singapore, (2014)
24. P. Mowlaee, J. Kulmer, Phase estimation in single-channel speech enhancement: Limits-potential. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(8), 1283–1294 (2015)
25. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
26. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, L. Xie, in *Interspeech 2020*. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. ISCA, Shanghai (2020), p. 2472–2476
27. X. Le, H. Chen, K. Chen, J. Lu, in *Interspeech 2021*. DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement. ISCA, Brno (2021), p. 2811–2815
28. H. Erdogan, J.R. Hershey, S. Watanabe, J. Le Roux, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. IEEE, Brisbane (2015), p. 708–712
29. H. Wu, K. Tan, B. Xu, A. Kumar, D. Wong, in *Interspeech 2023*. Rethinking complex-valued deep neural networks for monaural speech enhancement. ISCA, Dublin (2023), pp. 3889–3893
30. Y. Luo, Z. Chen, T. Yoshioka, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. IEEE, Virtual Barcelona (2020), p. 46–50
31. C.K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, S. Srinivasan, in *Interspeech 2021*. Interspeech 2021 deep noise suppression challenge. ISCA, Brno (2021), p. 2796–2800
32. 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)
33. L. Zhang, M. Wang, Q. Zhang, X. Wang, M. Liu, Phasedcn: A phase-enhanced dual-path dilated convolutional network for single-channel speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2561–2574 (2021)
34. K. Tan, D. Wang, in *Interspeech 2018*. A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement, ISCA, Hyderabad (2018), p. 3229–3233
35. T. Ke, W. DeLiang, Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 380–390 (2020)
36. 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)
37. 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)
38. S.W. Fu, T.W. Wang, Y. Tsao, X. Lu, H. Kawai, End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(9), 1570–1584 (2018)
39. S. Sonning, C. Schüldt, H. Erdogan, S. Wisdom, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Performance study of a convolutional time-domain audio separation network for real-time speech denoising. IEEE, Virtual Barcelona (2020), p. 831–835
40. S. Wisdom, J.R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, R.A. Saurous, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Differentiable consistency constraints for improved deep speech enhancement. IEEE, Brighton (2019), p. 900–904
41. S. Braun, H. Gamper, in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Effect of noise suppression losses on speech distortion and asr performance. IEEE, Singapore (2022), p. 996–1000
42. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognit. **106**, 107404 (2020)
43. S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.H. Yang, L. Shao, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Multi-stage progressive image restoration. IEEE, Virtual (2021), p. 14816–14826
44. S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint (2018). arXiv:1803.01271
45. S. Ioffe, C. Szegedy, in *32nd International Conference on Machine Learning*. Batch normalization: Accelerating deep network training by reducing internal covariate shift. JMLR, Lille (2015), p. 448–456
46. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
47. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint (2014). arXiv:1412.6980
48. J. Beerends, A. Rix, M. Hollier, A. Hekstra, in 2001 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. IEEE, Salt Lake City (2001), p. 749–752
49. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
50. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)


## Publisher's Note