scientific reports

OPEN

Mixed T-domain and TF-domain Magnitude and Phase representations for GAN-based speech enhancement

Lin Xin✉, Zhang Yang & Wang Shiyuan

Deep learning has made significant advancements in speech enhancement, which plays a crucial role in improving the quality of speech signals in noisy conditions. In this paper, we propose a new approach called M-DGAN, which introduces a time (T)-domain encoder-decoder structure with rich channel representations into the time-frequency (TF)-domain generator framework, resulting in a new generator structure with mixed magnitude and phase representations in the T and TF-domains. The proposed mixed T-domain and TF-domain generator, incorporating the cascaded reworked conformer (CRC) structure, exhibits improved modeling capability and adaptability. Test results on the Voice Bank + DEMAND public dataset show that our method achieves the highest score with $PSEQ = 3.52$ and performs well on all the remaining metrics when compared to the current state-of-the-art methods. In addition, tests on the NISQA_TEST_LIVETALK real dataset of the NISQA Corpus show the breadth and robustness of our model on speech enhancement tasks.

Speech enhancement is a fundamental research area within the field of speech signal processing. Its primary objective is to enhance the quality and intelligibility of speech signals, particularly in the presence of background noise or other adverse conditions[1]. It is widely used in areas such as communication systems[2], hearing aids[3] and speech recognition systems[4]. The rapid development of deep learning techniques has driven the evolution of speech enhancement methods from the three key directions: Time (T) domain[5–8], Frequency (F) domain[9,10], and Time-Frequency (TF) domain[11–13].

As a representative of traditional T-domain methods, Kalman filtering[14] reduces noise effects by directly manipulating the time-domain signal. However, it suffers from performance degeneration in the face of non-stationary noise and transient signal variations[15]. With the rise of deep learning, the neural network structures such as self-encoders[16] and long-short-term memory network (LSTM)[7] have become a hot spot in the research of T-domain methods. Among them, speech enhancement generative adversarial network (SEGAN)[8], is an innovative T-domain speech enhancement method by introducing a generative adversarial network (GAN) for the first time. Through adversarial training, SEGAN is capable of processing raw audio rapidly in an end-to-end manner[8], without the need for manual feature extraction.

Although these deep learning methods improve the ability of T-domain models for modelling complex signal variations, they still suffer from insufficient robustness against noise in the presence of complex nonlinear noise. In addition, since T-domain methods focus more on T-domain variations, they cannot capture the spectral information comprehensively enough[12], potentially leading to information loss.

F-domain methods have shown the superiorities in noise reduction of speech signal, e.g., Wiener filtering[17]. Especially excelling in the handling of complex noise[9], F-domain methods have achieved some success in the improvement of performance and robustness. However, F-domain methods face the issue of computational complexity when dealing with large-scale data[18]. In addition, the frequency domain deep learning methods under complex conditions also have the issue of robustness[10].

In contrast to T-domain and F-domain methods, TF-domain methods focus on the analysis and optimization of speech signals and noise in the frequency domain. Traditional methods such as short-time Fourier transform (STFT) and spectral subtraction[19] suppress noise using the manipulation of spectral information. However, these methods cannot perform well in dealing with nonlinear noise and time-varying signals. Deep learning methods in the TF-domain are categorized into the magnitude-domain and complex-domain methods. The

College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China. ✉email: lx354swu@163.com

magnitude-domain methods directly regress the magnitude spectrum of clean speech, reconstruct the phase of noisy speech and enhance the T-domain signal of the magnitude spectrum by inverse short-time Fourier transform (iSTFT)[20]. However, the magnitude-domain method does not consider the speech phase information. And it has been demonstrated that the phase information plays an important role in enhancing speech for better speech quality[21]. Several studies point out that although the complex-domain method considers both amplitude and phase information, its structure implementation is still challenging due to problems such as the randomness of the phase component[22] and the difficulty in estimating the complex mask value[6].

With the advancement of attentional mechanisms, numerous studies[23,24] have skillfully integrated them into various tasks of speech processing. The incorporation of the Conformer model in TF-domain deep learning methods[24], has significantly contributed to the advancement of spectral information analysis and research. This model has further enhanced our understanding of spectral data by incorporating powerful techniques from deep learning. The Conformer model integrates a convolutional neural network (CNN) and self-attention mechanism[25] to effectively capture long-range dependencies and spectral features in TF-domain speech enhancement tasks[24]. This combination leads to significant performance improvements in analyzing and enhancing speech signals in the TF-domain. In addition, MetricGAN[26], as a generative adversarial network-based speech enhancement model in the TF-domain, introduces an innovative approach by incorporating evaluation metrics as supervision during the training process. This enables MetricGAN to optimize speech quality and intelligibility metrics more directly, leading to the generation of speech that is more compatible with human hearing. Although these deep learning methods in the TF-domain have strong modeling capabilities in the field of speech enhancement, they still suffer from problems such as loss of information[27] when compared to deep learning methods in the T-domain, especially when dealing with short-duration noise or transient signal changes. It is worth noting that Zezario et al.[28] proposed a deep denoising autoencoder structure as a post-filter to compensate for the effects of already enhanced speech, and the method did further enhance speech. However, the Zezario et al. study did not show strong enhancement advantages over the many deep learning speech enhancement models.

Inspired by the aforementioned studies, this paper proposes a cascaded reworked conformer (CRC) based GAN speech enhancement structure with magnitude and phase representations in mixed T-domain and TF-domain, named M-DGAN. M-DGAN employs a generative adversarial network for speech enhancement, introduces a T-domain encoder-decoder structure with rich channel representations into the TF-domain two-way speech enhancement structure, and generates enhanced speech by mixing the outputs of the two in both magnitude and phase to further improve the modeling capability. This innovative generator structure is expected to restore the T-domain characteristics of the original signal more comprehensively and accurately by estimating the T-domain residuals of the speech signal accurately. Meanwhile, we also introduce the cascaded reworked conformer structure as an improved TF-domain generator component with more powerful modeling capabilities and adaptability. We employ a metric discriminator that introduces evaluation metric into the loss function of the discriminator to help the generator produce speech that is more in line with human listener perception.

Specifically, there are four areas of research in this paper. First, the generator of M-DGAN adopts a T-domain encoder-decoder structure with rich channel representations, which is mainly used to estimate the residual noisy data that is ignored in the TF-domain, and a TF-domain two-way speech enhancement structure to realize speech enhancement by combining and reintegrating the two paths in amplitude and phase. Secondly, we introduce the CRC module with optimized parameter tuning into the two-way speech enhancement structure in the TF-domain to further complement the captured speech features in the TF-domain in both time and frequency dimensions, which further achieves the effect of speech enhancement. Then, we test the proposed M-DGAN in this paper on the Voice Bank + DEMAND public dataset[29]. The results show that our model outperforms current T-domain and TF-domain SOTA methods. Finally, we also test the actual processing effect of the M-DGAN proposed in this paper on the NISQA_TEST_LIVETALK[30] dataset of the NISQA Corpus[30], and the test results show that our model still has good performance even in the complex and changeable real noise environment, which suggests that our model can be applied to a wide range of real-life scenarios, for example, in the fields of call speech enhancement, hearing aids, and speech recognition.

Methods

Our M-DGAN is based on the GAN, where the generator and discriminator will be trained alternately during the training process. The generator is used to generate enhanced speech samples and update the model parameters of the generator through a series of loss functions to improve the quality of the generated enhanced speech. The discriminator tries to accurately distinguish between the enhanced speech samples generated by the generator and the clean speech samples, and uses the discriminator loss function to continuously optimize and improve the performance of the discriminator, in order to assist the generator in generating enhanced speech that is closer to clean speech. In the testing phase, the trained generator receives the input noisy speech samples to generate the enhanced speech samples, and the discriminator does not work. In this section, we introduce M-DGAN in detail. It includes the generator of amplitude and phase representation from mixed T and TF domains, the metric discriminator, and the loss function of the model.

Generator structure

The detailed generator structure of M-DGAN is shown in Fig. 1. The M-DGAN generator consists of TF-domain two-way speech enhancement structure and T-domain residual noise estimation encoder-decoder structure. The M-DGAN generator can enhance the input speech effectively by using the speech amplitude and phase representation mixed with T-domain and TF-domain.

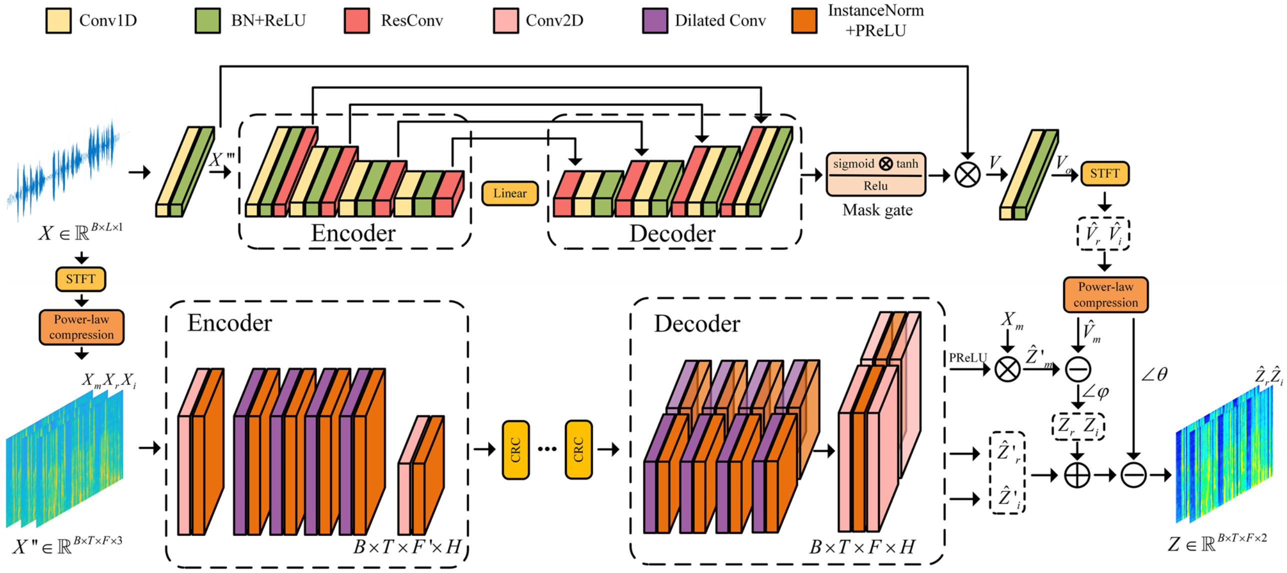


Figure 1. Generator structure of M-DGAN. The M-DGAN generator consists of T-domain residual noise estimation encoder-decoder structure and TF-domain two-way speech enhancement structure. In the TF-domain two-way speech enhancement structure, a cascaded reworked conformer structure is introduced to enhance model performance.

TF-domain two-way speech enhancement structure

As shown in Fig. 1, our TF-domain two-way speech enhancement structure consists of an encoder-decoder based on convolution block composition and a cascaded reworked conformer structure. Before the noisy speech signal $X \in \mathbb{R}^{B\times L\times 1}$ is input into the TF-domain two-way speech enhancement structure, it needs to be converted into a time-spectrogram $X' \in \mathbb{R}^{B\times T\times F\times 2}$ by the short-time Fourier transform (STFT)[31], where B denotes the batch size of the TF-domain two-way speech enhancement structure and the T-domain residual noise estimation encoder-decoder structure, L denotes the number of sample points of the input noisy speech, T stands for the time dimension, F stands for the frequency dimension, and 2 denotes the real component and the imaginary component of the complex value of X produced by the STFT. In our training, we take $B = 5$ and $L = 32000$. Given that the power-law compression of the magnitude equalizes the importance of quieter sounds relative to loud ones, which is more in line with human perception of sound[32]. Therefore, for $X' \in \mathbb{R}^{B\times T\times F\times 2}$, it is recommended to undergo the power-law compression to obtain the compressed spectrogram X_o, i.e.,

$$X_o = \left|X'\right|^k e^{j\varphi} = X_m e^{j\varphi} = X_r + jX_i \tag{1}$$

where X_m, X_r, X_i, and φ represent the magnitude, the real component, the imaginary component X_i, and phase of the compressed spectrogram, respectively; k represents the compression index, and according to Braun et al.[33], the commonly used value of the compression index $k = 0.3$ is set. Finally, the imaginary and real components X_i, X_r and the magnitude X_m are connected as input features $X'' \in \mathbb{R}^{B\times T\times F\times 3}$ to the encoder.

The encoder of the TF-domain two-way speech enhancement structure consists of 2 convolution blocks and 4 dilated convolution blocks. The convolution block consists of 2D down convolution, instance normalization[34], and a parameter rectified linear unit (PReLU)[35]. The dilated convolution block consists of dilated convolution[36], instance normalization, and PReLU. The dilation factors of the 4 dilated convolution blocks in the encoder and decoder are $\{1, 2, 4, 8\}$. In the encoder, the first 2D down convolution is used to extract deeper speech feature information. The middle 4 dilated convolution are used to expand the receptive field of the redspeech feature map. Their main purpose is to aggregate the features within the speech feature map output obtained from the first convolution block. The final convolution block reduces the frequency dimension to $F' = \frac{F}{2}$, which helps to reduce the overall complexity of the model. This reduction in frequency dimension can also lead to a smaller model size and faster inference speed.

The decoder consists of 2 decoder blocks including a magnitude mask estimation decoder block and a complex estimation decoder block. The magnitude mask estimation decoder block consists of 4 dilated convolution blocks and 2 convolution blocks. The 4 dilated convolution blocks as well as the first convolution block are similar to the ones in the encoder, however the second convolution block consists of only 2D up convolution and PReLU. The magnitude mask estimation decoder block will output a multiplicative mask of the magnitude, which will be multiplied by the input magnitude of the encoder X_m to predict the enhanced speech magnitude $\hat{Z}'_m$ in the TF-domain. The complex estimation decoder block has a similar structure to the magnitude mask estimation decoder block, but contains only a 2D up convolution in the second convolution block. The first convolution block of both decoder blocks performs up-sampling of the frequency dimension to restore it back to its original size of F. The complex estimation decoder block is used to predict the real and imaginary components, and its output is $\hat{Z}'_r + j\hat{Z}'_i$.

The final decoder outputs $\hat{Z}'_m$ and $\hat{Z}'_r + j\hat{Z}'_i$ will be fused with the output of the T-domain residual noise estimation encoder-decoder structure to obtain the final enhanced speech signal spectrogram.

Cascaded reworked conformer (CRC)

The standard Conformer[24] was designed by Gulati et al, based on Macaron-Net[37] and contains two Feed Forward module (FFN) sandwiching the Multi-Headed Self-Attention module (MHSA) and the Convolution module, which is similar to the sandwich structure. The study conducted by Gulati et al.[24] demonstrated that the standard Conformer model achieved excellent results in speech enhancement. By combining the strengths of Transformer and CNN networks, the standard Conformer effectively captures content-based global interactions while also leveraging local features of the input information[24]. This hybrid approach enables the Conformer model to excel in various tasks by effectively modeling long-range dependencies and local patterns simultaneously. In this paper, we perform the following set of reworking to the standard Conformer in the expectation of further improving the speech enhancement performance of the structure.

- Modify layer normalization (LN) to basic normalization (BasicNorm)[38] to address the issue of the mechanism of length removal by resistant LN.
- Change swish to doubleswish[38] to improve performance and reduce memory in training.
- Introduce ActivationBalancer[38] to adjust the activation value range to avoid the issue of abnormal activation value.

Finally we introduce learnable parameter scales in FFN, MHSA, and Convolution to form scaled_FFN, scaled_MHSA, and scaled_Conv to weigh the contribution of modules.

After parameter tuning, the new reworked conformer structure used in this paper is formed. We then join the two reworked conformers together by shaping block to form cascaded reworked conformer (CRC). We use a total of four CRC in this paper M-DGAN, as shown in Fig. 2, which shows the detailed structure of the CRC.

Every CRC consists of a shaping block, a time convolutional enhancement (CRC-T), and a frequency convolutional enhancement (CRC-F), where both the CRC-T and CRC-F modules are composed of scaled_FFN, scaled_MHSA, and scaled_Conv. First, the feature map $S \in \mathbb{R}^{B \times T \times F' \times H}$ needs to be reshaped into $S' \in \mathbb{R}^{BF' \times T \times H}$ before it is input into CRC-T for extracting the time dependency, where H denotes the number of feature channels after the expansion of the TF-domain speech enhancement module encoder. Here, $H = 64$ is chosen. In CRC-T, input $S' \in \mathbb{R}^{BF' \times T \times H}$ is processed to output the feature map $S'' \in \mathbb{R}^{BT \times F' \times H}$. $S'' \in \mathbb{R}^{BT \times F' \times H}$ will be passed into CRC-F for extracting the frequency dependence. $S'' \in \mathbb{R}^{BT \times F' \times H}$ is passed through CRC-F to achieve the final output $S_{final} \in \mathbb{R}^{B \times T \times F' \times H}$.

In order to show the effectiveness for the improved CRC structure, we perform ablation experiments to verify it in the Results and Discussion of this paper. Based on the model M-DGAN proposed in this paper, we replace the CRC structure in the generator with a cascade structure composed of unimproved standard conformer[24] structures, and name it SM-GAN. All other settings remain unchanged, and we compare the speech enhancement performances of SM-DGAN and M-DGAN. In addition, we performed ablation experiments on the CRC-T and CRC-F setups. And the experimental results results all indicate that that the CRC structure is effective.

T-domain residual noise estimation encoder-decoder structure

Although the processing of the TF-domain speech enhancement structure has been able to predict cleaner enhanced speech from noisy speech. However, considering the limitations of the TF domain speech enhancement structure, such as inaccurate estimation of the complex-valued masks[6], there is still some noise in the generated enhanced speech, which is difficult to eliminate completely. Therefore, as shown in Fig. 1, we propose a T-domain residual noise estimation encoder-decoder structure with rich channel representations for processing the residual noise of the TF-domain speech enhancement module, which is used to further optimize the quality of the generated speech and enhance the speech enhancement performance of the model. And the T-domain structure has the advantages of low training complexity and simpler structure compared to the TF-domain. Furthermore, we note that ResConv[6] block can efficiently represent multiscale features at a finer level and has fewer parameters.

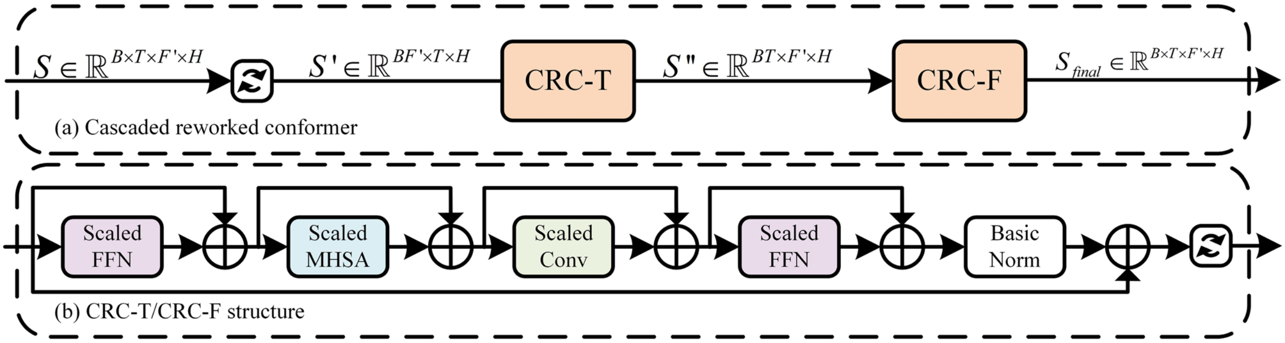

$S \in \mathbb{R}^{B \times T \times F' \times H}$
$S' \in \mathbb{R}^{BF' \times T \times H}$
CRC-T
$S'' \in \mathbb{R}^{BT \times F' \times H}$
CRC-F
$S_{final} \in \mathbb{R}^{B \times T \times F' \times H}$
(a) Cascaded reworked conformer
Scaled FFN
Scaled MHSA
Scaled Conv
Scaled FFN
Basic Norm
(b) CRC-T/CRC-F structure

Figure 2. Cascaded reworked Conformer architecture. Every CRC consists of a shaping block, a time convolutional enhancement (CRC-T), and a frequency convolutional enhancement (CRC-F). Both the CRC-T and CRC-F modules are composed of scaled_FFN, scaled_MHSA, and scaled_Conv.

Therefore, to further enhance the effectiveness of T-domain residual noise estimation encoder-decoder structure, we add ResConv block to the structure.

As shown in Fig. 1, the T-domain residual noise estimation encoder-decoder structure consists of two 1D convolution blocks, an encoder and decoder connected by a linear transformation, a mask gate[6], and the output of the encoder layer is residually connected to the input of the corresponding decoder layer. Both encoder and decoder contain 4 encoder and decoder layers, respectively. Both the encoder and decoder layer consist of a convolutional layer, a batch normalization (BN)[39] and a ResConv[6] block. In the M-DGAN of this paper, the kernel size of the convolutional layer in T-domain residual noise estimation encoder or decoder layers is set to 8 and the stride is set to 4.

In the T-domain residual noise estimation encoder-decoder structure, the noise-laden speech $X \in \mathbb{R}^{B\times L\times 1}$ needs to be downsampled first by a 1D convolution block to obtain $X''' \in \mathbb{R}^{B\times L\times N}$, where N represents the number of channels. In M-DGAN, we set $N = 60$. Then it goes into the encoder to extract the deep residual noise features that are not removed by the TF-domain speech enhancement module and goes through the decoder to recover these features. Finally the T-domain mask is output by the mask gate. The predicted output $V \in \mathbb{R}^{B\times L\times N}$ is obtained by multiplying the T-domain mask with the output of the first 1D convolution block by element-wise multiplication. $V \in \mathbb{R}^{B\times L\times N}$ is upsampled by the second 1D convolution block to obtain the residual noise data $V_o \in \mathbb{R}^{B\times L\times 1}$, and $V_o \in \mathbb{R}^{B\times L\times 1}$ is therefore passed through the STFT to obtain the number of complexes $\hat{V}_r + j\hat{V}_i$ and the residual noise amplitude of $\hat{V}_m$.

Then, as shown in equations (2) and (3), we subtract the TF-domain amplitude decoder output $\hat{Z}'_m$ from the T-domain residual noise amplitude $\hat{V}_m$ and combine it with the original noisy speech phase φ to obtain an amplitude-enhanced complex speech spectrum (Z_r, Z_i), i.e,

$$Z_r = (\hat{Z}'_m - \hat{V}_m)\cos(\varphi) \tag{2}$$

$$Z_i = (\hat{Z}'_m - \hat{V}_m)\sin(\varphi) \tag{3}$$

Finally as shown in equations (5) and (6), (Z_r, Z_i) are summed with the output of the complex estimation decoder in the TF-domain $\hat{Z}'_r + j\hat{Z}'_i$, and the unit complex $(\cos(\theta), \sin(\theta))$ of the residual noise in the T-domain is subtracted to compensate for the phases that cannot be compensated in the TF-domain, yielding the final complex speech spectrogram $Z \in \mathbb{R}^{B\times T\times F\times 2}$, where 2 denotes the two components of complex values $(\hat{Z}_r, \hat{Z}_i)$, i.e.,

$$\hat{V}_m e^{j\theta} = \hat{V}_r + j\hat{V}_i \tag{4}$$

$$\hat{Z}_r = Z_r + \hat{Z}'_r - \cos(\theta) \tag{5}$$

$$\hat{Z}_i = Z_i + \hat{Z}'_i - \sin(\theta) \tag{6}$$

where θ denotes the residual noise phase. The inverse power-law compression is then performed on the final spectrogram $Z \in \mathbb{R}^{B\times T\times F\times 2}$ and the inverse short-time Fourier transform (iSTFT) is applied to obtain the T-domain signal z.

In order to prove the effectiveness of our proposed T-domain residual noise estimation structure, we perform ablation experiments to validate it in the Results and Discussion of this paper. The results of tests fully prove the effectiveness of the proposed T-domain residual noise estimation structure.

Metric discriminator

In GAN-based speech enhancement tasks[8], the objective function is not linked to the evaluation metrics, which leads to low evaluation scores even if this function is well optimized. However, some commonly used evaluation metrics, such as the perceptual evaluation of speech quality (PESQ)[40] and the short-time objective intelligibility (STOI)[41], are usually non-trivial and cannot be directly used as loss functions. A method for associating the discriminator with the evaluation metric PESQ is proposed in MetricGAN[26], where the evaluation metric is used as part of the loss function to optimize the generator. In this paper, we take the same approach as MetricGAN and use normalized PESQ scores as labels for training.

As shown in Fig. 3, our metric discriminator consists of 4 convolution blocks, global average pooling, 2 linear layers, and a sigmoid activation. Among them, each of the four convolution blocks has a different number of channels. In our experiments, we set the number of channels of the convolution block to $\{16, 32, 64, 128\}$. Each convolution block consists of a convolutional layer, instance normalization, and a PReLU activation. In training the discriminator, training is performed for two magnitude spectrum inputs from the same clean speech, and for inputs consisting of a clean magnitude spectrum and an enhanced magnitude spectrum. For the inputs of the same clean speech magnitude spectrum, the discriminator is trained so that its enhanced PESQ score is close to the labeled value 1. For the inputs of clean and enhanced magnitude spectra, the metric discriminator is trained so that its estimated PESQ score is constantly close to the actual one. Thus in constant training, the augmented speech generated by the generator can be constantly close to the clean speech, and the PESQ predicted by the metric discriminator can be constantly close to 1.

Loss function in the entire model

The generator's loss function contains 4 parts, shown as follows.

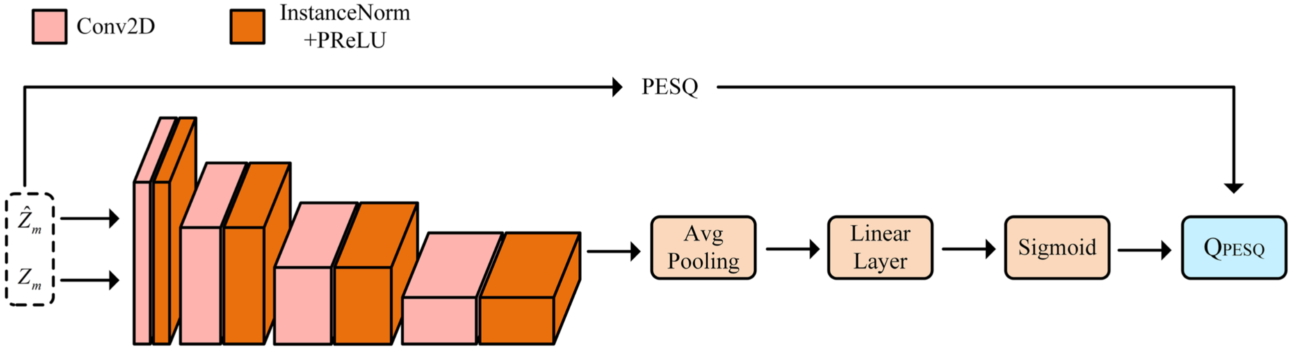


Figure 3. Metric discrimination architecture. The metric discriminator uses the PESQ score of the input signal as a label to learn alternative functions for the PESQ metric, then guides updates to the generator.

$$loss_G = \alpha_1 loss_{mag} + \alpha_2 loss_{RI} + \alpha_3 loss_{GAN} + \alpha_4 loss_{Time} \tag{7}$$

where α_1, α_2, α_3, and α_4 represent the weighting coefficients of the four losses respectively. Here, we take α_1 as 0.4, α_2 as 0.6, α_3 as 0.9, and α_4 as 0.07. $loss_{mag}$ represents the magnitude loss; $loss_{RI}$ represents the compound loss; $loss_{GAN}$ represents the antagonistic loss, i.e.,

$$loss_{mag} = E_{Z_m,\hat{Z}_m}\left[\left\|Z_m - \hat{Z}_m\right\|^2\right] \tag{8}$$

$$loss_{RI} = E_{Z_r,\hat{Z}_r}\left[\left\|Z_r - \hat{Z}_r\right\|^2\right] + E_{Z_i,\hat{Z}_i}\left[\left\|Z_i - \hat{Z}_i\right\|^2\right] \tag{9}$$

$$loss_{GAN} = E_{Z_m,\hat{Z}_m}\left[\left\|D(Z_m, \hat{Z}_m) - 1\right\|^2\right] \tag{10}$$

where D denotes the discriminator. The loss of the discriminator is denoted as:

$$loss_D = E_{Z_m}\left[\left\|D(Z_m, \hat{Z}_m) - 1\right\|^2\right] + E_{Z_m,\hat{Z}_m}\left[\left\|D(Z_m, \hat{Z}_m) - Q_{PESQ}\right\|^2\right] \tag{11}$$

where Q_{PESQ} denotes the normalized PESQ score located in the interval of $[0, 1]$. It has been shown that the extra time loss of $loss_{Time}$ improves speech enhancement quality, we introduced this loss as:

$$loss_{Time} = E_{Z_m,\hat{Z}_m}\left[\left\|Z - \hat{Z}\right\|_1\right] \tag{12}$$

Experimental settings

Dataset and evaluation metrics

In order to compare the proposed model M-DGAN with other speech enhancement model, we first test our module on the Voice Bank + DEMAND public dataset[29]. The training set consists of 28 speakers with 11,572 utterance from 14 men and 14 women mixed with 10 noise data with 4 signal-to-noise ratios (SNRs) of 0dB, 5dB, 10dB, and 15dB. The test set consists of 824 male and female speeches and two speakers mixed with five types of noise at four SNRs of 17.5dB, 12.5dB, 7.5dB, and 2.5dB. Since the PESQ[40] ranging form -0.5 to 4.5 is a commonly used index to evaluate enhanced speech, in this paper, we take PESQ score as the most important performance evaluation index, and also use the segmental signal-to-noise ratio (SSNR) values ranging from 0 to infinity, the short-time objective intelligibility (STOI)[41] values ranging from 0-1, mean opinion score (MOS) prediction of speech distortion (CSIG)[42] values ranging from 0-1, MOS prediction of the intrusiveness of background noise (CBAK)[42] values ranging from 0-1, and MOS prediction of overall effect (COVL)[42] values ranging from 0-1 to comprehensively evaluate the performance of the model. For these six indexes, the higher the value, the better the model performance, vice versa. All data are resamped from 48kHz to 16kHz for the experiment.

In addition, to further validate the effectiveness and extensiveness of M-DGAN, we test M-DGAN without prior training on the NISQA_TEST_LIVETALK[30] dataset of the NISQA Corpus[30]. The NISQA Corpus includes more than 14,000 speech samples with simulated (e.g. codec, packet-loss, background noise) and live (e.g. mobile phone, Zoom, Skype, WhatsApp) conditions. Each file is labeled with subjective ratings for overall quality and quality dimensions noise, coloration, discontinuity, and loudness. NISQA_TEST_LIVETALK is a key test set for the NISQA Corpus, which contains recordings of real phone calls and VoIP calls. The NISQA model[30] is a deep learning model for speech quality prediction, and it can be used to predict the quality of the voice samples sent over the communication system, such as a phone call or a video call. The output of the NISQA model is the nisq[30] metric, which has five dimensions that are used together to evaluate speech quality. Among them, Dimension 1: mos_pred[30]: speech quality indicator, higher means better sound quality. Dimension 2: ngi_pred[30]: noise level metric, higher means less noise. Dimension 3: dis_pred[30]: speech coherence indicator, the higher the better the speech coherence. Dimension 4: col_pred[30]: timbre indicator, the higher the better the sound. Dimension 5: loud_pred[30]: loudness indicator, higher means louder volume. The nisq[30] mertic will be an important reference for us to evaluate the model's effectiveness in real-world testing.

Implementation details

In the training process on the Voice Bank + DEMAND public dataset, we cut the discourse into 2 seconds each, with a batch size of $B = 5$, but in the test set, no slicing operation is performed and the length remains variable. When operating in the TF domain, the Hamming window function is used for analysis, and the spectrogram is obtained by using 400 STFT for a total of 25ms on each frame, with an overlap of 18.75ms (75.0%). In the TF-domain voice enhancement module of the generator, there are 4 CRC blocks, and the number of channels is $H = 64$. The dilation factors of the 4 dilated convolution blocks in the encoder and decoder are $\{1, 2, 4, 8\}$. In the T-domain residual noise estimation structure, the kernel size of the convolutional layer in encoder or decoder layers is set to 8 and the stride is set to 4. In the metric discriminator, the number of channels for the convolutional block is set to $\{16, 32, 64, 128\}$. Besides, the generator and discriminator are optimized using the adaptive moment estimation with decoupled weight decay (AdamW) optimizer and reduce learning rate on plateau (ReduceLROnPlateau) scheduler. The learning rate of the generator is set to 5×10^{-4}, and the learning rate of the discriminator is set to 1×10^{-3}. The decay factor of the generator's scheduler is set to 0.5, and the cumulative number of times patience is set to 2. The decay factor of the discriminator's scheduler is set to 0.45, and the cumulative number of time patience is set to 2. We train a total of 120 epochs.

Note that our real-world testing of M-DGAN on the NISQA_TEST_LIVETALK[30] dataset is done without any prior training on the NISQA_TEST_LIVETALK dataset. Instead, the test is conducted directly with the M-DGAN trained on the Voice Bank + DEMAND[29] dataset in order to use this experiment to demonstrate the robustness of our model.

Results and discussion

Comparison of different models

We compare the M-DGAN proposed in this paper with other state-of-the-art (SOTA) methods of different time periods, which include both T-domain and TF-domain methods. As shown in Table 1, the test results of our model M-DGAN on the public dataset Voice Bank + DEMAND are compared with five T-domain speech enhancement methods including standard SEGAN[8] and six TF-domain speech enhancement methods including MetricGAN[26]. Among them, SEGAN[8] is the first end-to-end GAN framework speech enhancement model. DEMUCS[43] is based on an encoder-decoder architecture with skip-connections. Phasen[13] contains a two-stream network, where amplitude stream and phase stream are dedicated to amplitude and phase prediction. And SN-Net[44] models speech and noise simultaneously in a two-branch convolutional neural network. SEGAN, SerGAN[45], MetricGAN[26], MetricGAN+[46], MAMGAN[5], CMGAN[12], and CGA-MGAN[11] are speech enhancement models based on the GAN framework. And MANNER[6] , MAMGAN, CMGAN, and CGA-MGAN are speech enhancement models that introduce the attention mechanism.

It can be seen that our proposed model M-DGAN performs optimally on six evaluation metrics, including PESQ, SSNR, STOI, CSIG, CBAK, and COVL, when compared to SOTA in recent years. Among them, our model M-DGAN achieves a score of 3.52 on PESQ, which is an improvement of 1.36 compared to the standard SEGAN and 0.66 compared to MetricGAN. Compared to the optimal model CGA-MGAN[11] that participates in the comparison, our method M-DGAN has a higher PESQ score by 0.05, and the scores on SSNR, STOI, CSIG, CBAK, and COVL are 0.93, 0.24, 0.12, 0.19, and 0.15 higher than CGA-MGAN on the five metrics, respectively.

Methods	Year	Domain	Par. (Million)	PESQ	SSNR	STOI(%)	CSIG	CBAK	COVL
Noisy	–	–	–	1.97	1.68	91.00	3.35	2.44	2.63
SEGAN[8]	2017	T	97.47	2.16	7.73	92.50	3.48	2.94	2.80
SerGAN[45]	2019	T	–	2.62	–	94.00	–	–	–
MetricGAN[26]	2019	TF	–	2.86	–	–	3.99	3.18	3.42
PHASEN[13]	2020	TF	–	2.99	10.08	–	4.21	3.55	3.62
DEMUCS[43]	2021	T	18.90	3.07	–	95.00	4.31	3.40	3.63
SN-Net[44]	2021	TF	–	3.12	9.83	–	4.39	3.60	3.77
MetricGAN+[46]	2021	TF	–	3.15	–	95.00	4.14	3.16	3.64
MANNER[6]	2022	T	24.10	3.21	–	95.00	4.53	3.65	3.91
MAMGAN[5]	2022	T	17.29	3.30	–	95.04	4.53	3.64	3.95
CMGAN[12]	2022	TF	1.83	3.41	11.10	96.00	4.63	3.94	4.12
CGA-MGAN[11]	2023	TF	1.14	3.47	11.09	96.00	4.56	3.86	4.06
M-DGAN	2023	TF	1.40	**3.52**	**12.02**	**96.24**	**4.68**	**4.05**	**4.21**

Table 1. Performance comparison on the Voice Bank + DEMAND dataset. "–" indicates not provided in the original article. Bolded fonts represent optimal results, and all data are retained to two decimal places. "Par." indicates the number of trainable parameters.

In addition, although our model adopts the strategy of combining TF and T domains, the number of parameters of our method is only 1.40M compared to T-domain SOTA methods, which is much lower than that of various T-domain SOTA methods, which shows that the M-DGAN method combining T and TF domains not only reduces the number of parameters of the model but also improves the performance of the model compared to the T-domain-only method. Compared to the TF-domain SOTA methods, our method is optimal in terms of PESQ and other scores with relatively fewer parameters.

Effectiveness of the M-DGAN design

As in Table 2, we perform the following ablation experiments to demonstrate the validity of each component selection in our design.

First we verify the validity of the CRC structure. According to the Methods section of this paper, CRC is able to focus on the time and frequency dimensions of the input features. Therefore, we change the CRC-T and CRC-F of the CRC module from sequential connection to parallel connection, and all other settings remain unchanged. We named the model PM-DGAN, and we compared M-DGAN with PM-DGAN. The results show that PM-DGAN with parallel CRC is not as good as M-DGAN for speech enhancement, and the PESQ score is 0.08 worse compared to M-DGAN's, and the remaining five metrics are also worse than the M-DGAN. This is due to the fact that parallel CRC leads to possible learning redundancy when learning features of feature blocks without the synergy of directly connected CRC. Then we flip the CRC-T and CRC-F positions and named the flipped overall model as EM-DGAN, then we train and test it. The results are shown in Table 2. It can be seen that the scores of EM-DGAN and M-DGAN are almost similar. In order to show the effectiveness of our CRC structure, we compare the performance of SM-DGAN, a model based on a cascaded standard conformer[24] structure, with M-DGAN. From Table 2, it can be seen that in terms of the scores on the six assessment metrics including PESQ, the SM-DGAN scores are all lower than those of the M-DGAN. The PESQ score of M-DGAN is 3.52, while the PESQ score of SM-DGAN is only 3.43, with a difference of 0.09. The scores of the evaluation metrics regarding the models also further indicate that the model utilizing the CRC structure has a better performance in speech enhancement than the model utilizing the standard Conformer structure. As shown in Figs. 4 and 5, We plot the training of the two models on the Voice Bank + DEMAND dataset and compare the results for the first 77 epochs in order to compare the two more clearly.

It can be seen that in the pre-training period, the generator of M-DGAN converges faster compared to the generator of SM-DGAN. As the number of training epochs increases, the generator loss curve of M-DGAN is more stable and converges to a lower value compared to SM-DGAN. As can be seen from the discriminator loss curve, the discriminator of M-DGAN performs more robustly in general during the training process.

It is worth noting that, the use of a single Conformer to attend over frequency and time dimension information is theoretically possible, but it leads to an exponential growth in complexity[47]. Therefore, in this paper we do not consider ablation experiments with a single Conformer structure.

In the second part, we validate the T-domain residual noise estimation structure. As shown in Table 2, we compare the results of the effectiveness of this module on the Voice Bank + DEMAND dataset. The complete M-DGAN is included, of which both amplitude and phase are combined with the output of the T-domain residual noise estimation encoder-decoder structure. TFGAN: removing the T-domain residual noise estimation encoder-decoder structure. M-TFGAN: optimizing only the magnitude using the T-domain residual noise estimation encoder-decoder structure without considering the phase. P-TFGAN: optimizes only the phase using the T-domain residual noise estimation encoder-decoder structure without considering the magnitude. We also explored the T-domain residual noise estimation encoder-decoder structure without the ResConv[6] block to estimate the residual noise, and we name the model DRM-DGAN.

As shown in Table 2, the method TFGAN with the removal of the T-domain residual noise estimation encoder-decoder structure has a difference of 0.45 in PESQ and all other scores are lower compared to M-DGAN with the CRC module. In addition, we also validate the ablation together with the model that utilizes the T-domain residual noise estimation encoder-decoder structure with only amplitude optimization. For

Methods	PESQ	SSNR	STOI(%)	CSIG	CBAK	COVL
M-DGAN	**3.52**	**12.02**	**96.24**	**4.68**	**4.05**	**4.21**
PM-DGAN	3.44	11.56	96.22	4.52	4.03	4.15
EM-DGAN	3.49	12.01	96.23	4.65	4.04	4.18
SM-DGAN	3.43	11.95	96.16	4.65	4.01	4.14
TFGAN	3.07	10.84	95.51	4.38	3.76	3.80
M-TFGAN	3.23	11.13	95.49	4.49	3.85	3.95
P-TFGAN	3.08	10.97	95.64	4.27	3.46	3.72
DRM-DGAN	3.50	11.99	96.23	4.64	4.03	4.17
RCM-DGAN	3.32	9.65	96.04	4.55	3.87	4.05
RMM-DGAN	3.25	9.58	95.89	4.46	3.84	4.02
RTM-DGAN	3.54	9.63	96.25	4.56	3.97	4.18

Table 2. Performance comparison on the Voice Bank + DEMAND dataset. M-DGAN scores are bolded, and all data are retained in two decimal places.

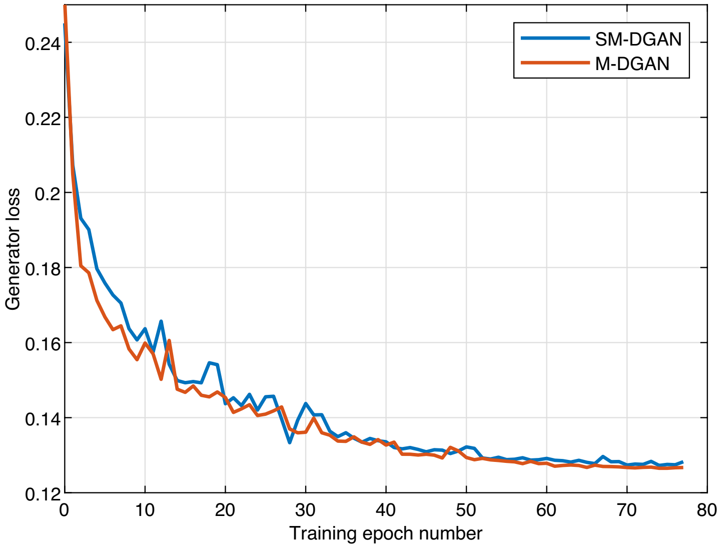


Figure 4. Training loss curves for the model generator consisting of the standard Conformer and CRC respectively.

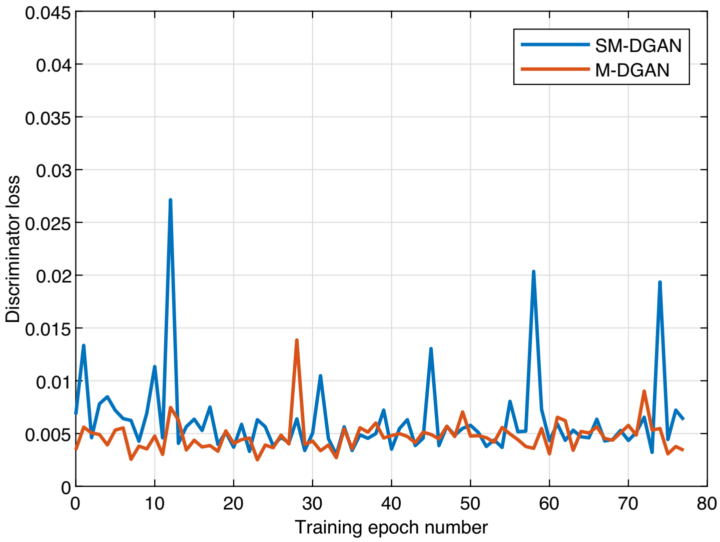
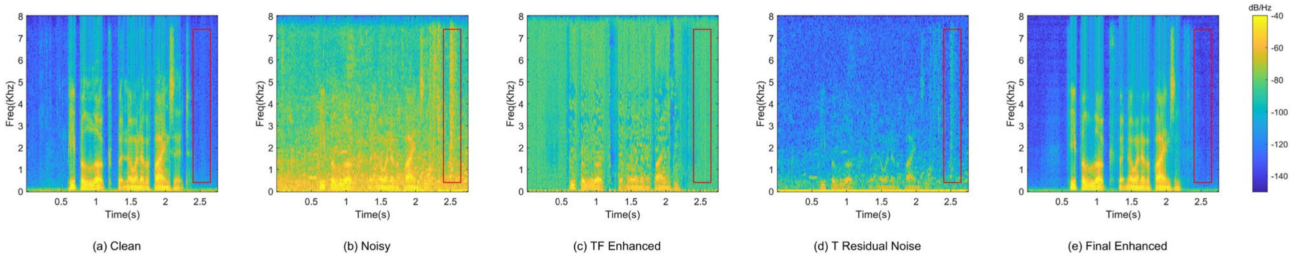


Figure 5. Training loss curves for the model discriminator consisting of the standard Conformer and CRC respectively.

M-TFGAN, it can be seen that there is an improvement of 0.16 in PESQ for the magnitude-only optimization compared to TFGAN with the removal of the T-domain residual noise estimation encoder-decoder structure, but there is still a difference of 0.29 in PESQ when compared to M-DGAN, which is optimized for both magnitude and phase. For P-TFGAN, the test results do not have a significant advantage over TFGAN. However, the PESQ is 0.15 worse compared to M-TFGAN, and other metrics also show the limitations of P-TFGAN. It shows that it is not enough to consider only the speech-compensated phase for the enhanced output of the TF domain. For DRM-DGAN, it can be seen that the scores of DRM-DGAN are very similar to those of M-DGAN, but there is a certain gap in speech enhancement compared to that of M-DGAN with the introduction of ResConv block, and there is still a drop in each score.

As shown in Fig. 6, to visualize the T-domain residual estimation noise validity, we randomly select a piece of test data on the Voice Bank + DEMAND test set and plot the spectrogram of the processing of this test data on M-DGAN.

Significant background noise exists in the entire spectral range of Fig. 6b. Figure 6c embodies the speech signal after the initial enhancement performed in the TF domain, and although the main frequency components of the speech become clearer, there is still incompletely eliminated noise in the high-frequency region and some of the low-frequency regions, especially in the red boxed region. Figure 6d shows the spectral characteristics of



(a) Clean (b) Noisy (c) TF Enhanced (d) T Residual Noise (e) Final Enhanced

Figure 6. Comparison of one randomly selected piece of speech data from the M-DGAN test set. (**a**) denotes clean speech, (**b**) denotes noisy speech, (**c**) denotes the output of the TF-domain two-way speech enhancement structure, (**d**) denotes the output of the T-domain residual noise estimation structure, and (**e**) denotes the final enhanced speech generated by mixing (**c**,**d**).

this residual noise, especially in the red box-selected region. These noises are the parts that are not completely eliminated in Fig. 6c, and these residual noise components can be efficiently identified by the T-domain residual noise estimation structure. Eventually, as can be seen from Fig. 6e, the speech enhancement result after combining the information from Fig. 6c,d shows a significant reduction of high-frequency noise in Fig. 6e compared to Fig. 6c, and the speech signal is purer and clearer. This indicates that by eliminating the residual noise in Fig. 6d, our model is able to further enhance speech and achieve superior speech enhancement results.

Third, we validate the design of the TF-domain two-way speech enhancement structure. We train and test RCM-DGAN with the complex decoder ignored and RMM-DGAN with the magnitude mask decoder ignored, respectively. As shown in Table 2, both RCM-DGAN and RMM-DGAN, have a gap with the speech enhancement performance of M-DGAN. And the PESQ of RCM-DGAN is reduced by 0.07 compared to RMM-DGAN. This suggests that phase information plays an important role in improving the quality of speech enhancement.

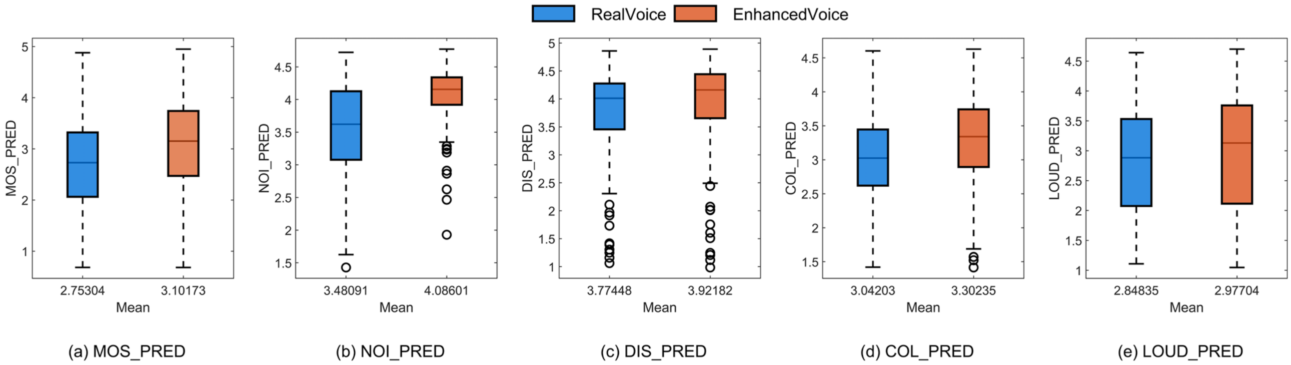
Fourth, we performed ablation experiments on the the extra time loss of the generator. We denote the model with the Time loss removed by RTM-DGAN, and we compare M-DGAN with RTM-DGAN. From the test results, we can see that the pesq of RTM-DGAN with Time loss removed increases 0.02 compared to M-DGAN, but there is some decrease in SSNR, which may be due to the fact that Time loss balances the performance of SSNR and PESQ.

Practical test experiments of M-DGAN

In this section, we compare the performance of the enhanced speech with the original speech in five dimensions to verify the effectiveness and robustness of M-DGAN in real data. The specific results are shown by Fig. 7.

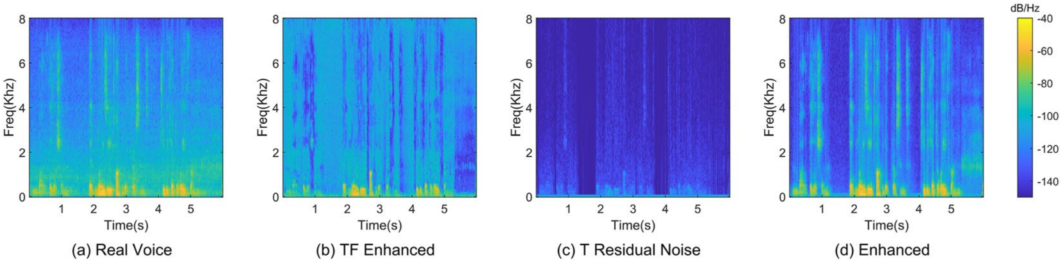
As shown in Fig. 7a, from the point of view of subjective perception, the MOS_PRED of the enhanced speech is significantly higher than that of the original speech, and the mean value is improved from 2.75304 to 3.10173, which indicates that the speech quality is significantly improved by our modeling process. As can be seen in Fig. 7b, the noise suppression effect is also verified. The mean value of NOI_PRED is improved from 3.48091 to 4.08601, showing a significant reduction in the noise level, and the enhanced speech is clearer and purer. The enhancement of speech coherence is demonstrated in Fig. 7c, where the DIS_PRED shows a significant increase in the coherence of the enhanced speech, with the mean value increasing from 3.77448 to 3.92182, indicating that the model improves the coherence of the speech along with the enhancement. In addition, as can be seen in Fig. 7d, the COL_PRED shows an improvement in timbre after enhancement, with a mean increase of 0.26032. Figure 7e shows that the average value of LOUD_PRED improves from 2.84835 to 2.97704, which indicates that M-DGAN enhances speech while maintaining an appropriate loudness level. These results validate the superior performance of our model in speech enhancement reality tasks.

As in Fig. 8, we also randomly selected a copy of speech data from the test set and plotted its spectrum.



(a) MOS_PRED (b) NOI_PRED (c) DIS_PRED (d) COL_PRED (e) LOUD_PRED

Figure 7. Box plot of the distribution of nisq metric[30] for the real speech set before and after M-DGAN processing. The graph shows five dimensions from left to right, including MOS_PRED, NOI_PRED, DIS_PRED, COL_PRED, and LOUD_PRED.



(a) Real Voice (b) TF Enhanced (c) T Residual Noise (d) Enhanced

Figure 8. Comparison of one randomly selected piece of speech data from the NISQA_TEST_LIVETALK[30] real test set. (**a**) denotes real noisy speech, (**b**) denotes the output of the TF-domain two-way speech enhancement structure, (**c**) denotes the output of the T-domain residual noise estimation structure, and (**d**) denotes the final enhanced speech generated by mixing (**b**,**c**).

Comparing Fig. 8a,d, it can be seen that the model effectively reduces the noise components in the spectrogram, making the signal clearer. While reducing the noise, the model retains the key frequency components of the speech signal to ensure the intelligibility and clarity of the speech.

Conclusion

In this paper, we propose a speech-enhanced GAN structure (M-DGAN) based on a cascaded reworked conformer structure with magnitude and phase representations in mixed T and TF domains. In our method, we innovatively introduce a generator structure that mixes the magnitude and phase representations of the T and TF domains to maximize the possibility of avoiding the drawbacks of a single T or TF domain. In addition, the cascaded reworked conformer structure of our method can effectively capture the long-term dependency of information as well as the consistency of local features in both time and frequency dimensions. Experiments show that our method outperforms current SOTA methods in tests on the public dataset Voice Bank + DEMAND. And the test results on the real NISQA dataset again show the robustness and effectiveness of our M-DGAN.

Data availibility

In order to promote open collaboration and enable fellow researchers to access, reproduce, and build upon our work, we have uploaded the reproducible code to GitHub. The link is provided here: https://github.com/1lingyin/M-DGAN.

Received: 12 January 2024; Accepted: 26 July 2024
Published online: 31 July 2024

References

1. Cui, Z. & Bao, C. Power exponent based weighting criterion for DNN-based mask approximation in speech enhancement. *IEEE Signal Process. Lett.* **28**, 618–622 (2021).
2. Das, N., Chakraborty, S., Chaki, J., Padhy, N. & Dey, N. Fundamentals, present and future perspectives of speech enhancement. *Int. J. Speech Technol.* **24**, 883–901 (2021).
3. Diehl, P. U. *et al.* Restoring speech intelligibility for hearing aid users with deep learning. *Sci. Rep.* **13**, 2719 (2023).
4. Donahue, C., Li, B. & Prabhavalkar, R. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5024–5028 (IEEE, 2018).
5. Guo, H. *et al.* Mamgan: Multiscale attention metric GAN for monaural speech enhancement in the time domain. *Appl. Acoust.* **209**, 109385 (2023).
6. Park, H. J., Kang, B. H., Shin, W., Kim, J. S. & Han, S. W. Manner: Multi-view attention network for noise erasure. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7842–7846 (IEEE, 2022).
7. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270 (2019).
8. Pascual, S., Bonafonte, A. & Serra, J. Segan: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452 (2017).
9. Ribas, D., Miguel, A., Ortega, A. & Lleida, E. Wiener filter and deep neural networks: A well-balanced pair for speech enhancement. *Appl. Sci.* **9000** (2022).
10. Roy, S. K., Nicolson, A. & Paliwal, K. K. A deep learning-based kalman filter for speech enhancement. In *Interspeech*, pp. 2692–2696 (2020).
11. Chen, H. & Zhang, X. CGA-MGAN: Metric GAN based on convolution-augmented gated attention for speech enhancement. *Entropy* **25**, 628 (2023).
12. Cao, R., Abdulatif, S. & Yang, B. Cmgan: Conformer-based metric GAN for speech enhancement. arXiv preprint arXiv:2203.15149 (2022).
13. Yin, D., Luo, C., Xiong, Z. & Zeng, W. Phasen: A phase-and-harmonics-aware speech enhancement network. *In Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 9458–9465 (2020).
14. Chui, C. K., Chen, G. *et al.Kalman filtering* (Springer, 2017).
15. Cohen, I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.* **11**, 466–475 (2003).
16. Wang, W., Huang, Y., Wang, Y. & Wang, L. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 490–497 (2014).
17. Chen, J., Benesty, J., Huang, Y. & Doclo, S. New insights into the noise reduction Wiener filter. *IEEE Trans. Audio Speech Lang. Process.* **14**, 1218–1234 (2006).

18. Coto-Jimenez, M., Goddard-Close, J., Di Persia, L. & Leonardo Rufiner, H. Hybrid speech enhancement with Wiener filters and deep lstm denoising autoencoders. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1–8, https://doi.org/10.1109/IWOBI.2018.8464132 (2018).
19. Lebart, K., Boucher, J.-M. & Denbigh, P. N. A new method based on spectral subtraction for speech dereverberation. *Acta Acust. Acust.* **87**, 359–366 (2001).
20. Fan, C. *et al.* Specmnet: Spectrum mend network for monaural speech enhancement. *Appl. Acoust.* **194**, 108792 (2022).
21. Paliwal, K., Wójcicki, K. & Shannon, B. The importance of phase in speech enhancement. *Speech Commun.* **53**, 465–494 (2011).
22. Hu, Y. *et al.* DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020).
23. Sun, C. *et al.* A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. *Sci. Rep.* **11**, 1434 (2021).
24. Gulati, A. *et al.* Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020).
25. Han, K. *et al.* Transformer in transformer. *Adv. Neural. Inf. Process. Syst.* **34**, 15908–15919 (2021).
26. Fu, S.-W., Liao, C.-F., Tsao, Y. & Lin, S.-D. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, 2031–2041 (PMLR, 2019).
27. Nossier, S. A., Wall, J., Moniri, M., Glackin, C. & Cannings, N. A comparative study of time and frequency domain approaches to deep learning based speech enhancement. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2020).
28. Zezario, R. E. *et al.* Deep denoising autoencoder based post filtering for speech enhancement. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 373–377 (IEEE, 2018).
29. Valentini-Botinhao, C., Wang, X., Takaki, S. & Yamagishi, J. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, 146–152 (2016).
30. Mittag, G., Naderi, B., Chehadi, A. & Möller, S. Nisqa: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494 (2021).
31. Durak, L. & Arikan, O. Short-time Fourier transform: Two fundamental properties and an optimal implementation. *IEEE Trans. Signal Process.* **51**, 1231–1242 (2003).
32. Wilson, K. *et al.* Exploring tradeoffs in models for low-latency speech enhancement. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 366–370 (IEEE, 2018).
33. Braun, S. & Tashev, I. A consolidated view of loss functions for supervised deep learning-based speech enhancement. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, 72–76 (IEEE, 2021).
34. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016).
35. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034 (2015).
36. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015).
37. Lu, Y. *et al.* Understanding and improving transformer from a multi-particle dynamic system point of view. arXiv preprint arXiv:1906.02762 (2019).
38. The next-gen Kaldi. Available online at:https://github.com/k2-fsa/icefall/blob/master/egs/librispeech/ASR/pruned_transducer_stateless2/conformer.py.
39. Santurkar, S., Tsipras, D., Ilyas, A. & Madry, A. How does batch normalization help optimization? *Adv. Neural Inf. Process. Syst.* **31** (2018).
40. Rix, A. W., Beerends, J. G., Hollier, M. P. & Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, 749–752 (IEEE, 2001).
41. Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, 4214–4217 (IEEE, 2010).
42. Hu, Y. & Loizou, P. C. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun.* **49**, 588–601 (2007).
43. Defossez, A., Synnaeve, G. & Adi, Y. Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847 (2020).
44. Zheng, C., Peng, X., Zhang, Y., Srinivasan, S. & Lu, Y. Interactive speech and noise modeling for speech enhancement. *In Proceedings of the AAAI conference on artificial intelligence* **35**, 14549–14557 (2021).
45. Baby, D. & Verhulst, S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 106–110 (IEEE, 2019).
46. Fu, S.-W. *et al.* Metricgan+: An improved version of metricgan for speech enhancement. arXiv preprint arXiv:2104.03538 (2021).
47. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

Acknowledgements

The Voice Bank + DEMAND dataset (accessed on 17 December 2021) can be found here: https://datashare.ed.ac.uk/handle/10283/2791. We would like to express our gratitude to the editorial team at Scientific Reports for their contributions to the refinement of this manuscript.

Author contributions

X.L. and Y.Z. conducted the experiment. SY.W. revised the manuscript and provided the technical and writing guidance. All the block diagrams of the models presented in this paper were drawn by X.L. using Office Visio 2021 professional version. The manuscript was written by X.L. and reviewed by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.X.

Reprints and permissions information is available at www.nature.com/reprints.

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