Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.021453

check for updates

ARTICLE

Using Hybrid Penalty and Gated Linear Units to Improve Wasserstein Generative Adversarial Networks for Single-Channel Speech Enhancement

Xiaojun Zhu[1,2,3] and Heming Huang[1,2,*]

[1]School of Computer Science, Qinghai Normal University, Xining, 810008, China

[2]The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining, 810008, China

[3]School of Electronic and Information Engineering, Lanzhou City University, Lanzhou, 730000, China

*Corresponding Author: Heming Huang. Email: huanghm@qhnu.edu.cn

Received: 15 January 2022 Accepted: 06 July 2022

ABSTRACT

Recently, speech enhancement methods based on Generative Adversarial Networks have achieved good performance in time-domain noisy signals. However, the training of Generative Adversarial Networks has such problems as convergence difficulty, model collapse, etc. In this work, an end-to-end speech enhancement model based on Wasserstein Generative Adversarial Networks is proposed, and some improvements have been made in order to get faster convergence speed and better generated speech quality. Specifically, in the generator coding part, each convolution layer adopts different convolution kernel sizes to conduct convolution operations for obtaining speech coding information from multiple scales; a gated linear unit is introduced to alleviate the vanishing gradient problem with the increase of network depth; the gradient penalty of the discriminator is replaced with spectral normalization to accelerate the convergence rate of the model; a hybrid penalty term composed of L1 regularization and a scale-invariant signal-to-distortion ratio is introduced into the loss function of the generator to improve the quality of generated speech. The experimental results on both TIMIT corpus and Tibetan corpus show that the proposed model improves the speech quality significantly and accelerates the convergence speed of the model.

KEYWORDS

Speech enhancement; generative adversarial networks; hybrid penalty; gated linear units; multi-scale convolution

1 Introduction

In practical application scenarios, speech signals will inevitably be disturbed by many interference factors such as noise, echo, and reverberation. Therefore, speech enhancement technology has been widely used in household appliances, communications, speech recognition, automotive electronics, hearing aids, and other fields [1–3]. Traditional speech enhancement methods, based on signal processing and statistical modeling, have good performance for stationary noise. However, their enhancement performance is not ideal for non-stationary noises and low signal-to-noise ratio (SNR) situations. With the advent of the era of "big data" and the great improvement of computing resources, deep learning technology has made great progress and has been widely used in such fields

as data processing, software development, automatic speech recognition, and other related fields [4–7]. Similarly, the development of deep learning technology has also brought new opportunities and breakthroughs for speech enhancement technology. Due to their powerful modeling ability and fewer hypothesis requirements for signals, speech enhancement methods based on deep learning technology have gradually replaced traditional speech enhancement methods as the mainstream technology for speech enhancement [8–10].

Generative adversarial network (GAN) is a deep learning model, and it is one of the most promising methods for unsupervised learning in recent years. Its emergence has pushed the research of deep learning technology to a deeper level: the focus of research has shifted from perception problems such as vision and hearing to cognitive problems such as decision-making and generation [11]. The essence of GAN is to transform the solution problem of the likelihood function into the training problem of neural networks [12]. This improvement makes GAN have strong applicability and plasticity. Not only can the structure of the generator and discriminator be changed according to different needs, but GAN can explore better problem-solving strategies through antagonistic training between the generator and discriminator. In recent years, with the improvement and maturity of GAN, speech enhancement technology based on GAN has emerged. Pascual et al. [13] came up with a GAN-based end-to-end time domain speech enhancement model to learn the mapping from the raw waveform of noisy speech to that of clean speech. Michelsanti et al. [14] proposed a frequency domain speech enhancement model based on conditional GAN (CGAN) to learn a mapping from the spectrogram of noisy speech to an enhanced counterpart. Shah et al. [15] proposed a time-frequency masking speech enhancement algorithm based on convolutional GAN, and it achieves good effects on noise suppression. Adiga et al. [16] proposed a new speech enhancement model based on Wasserstein Generative Adversarial Networks (WGAN), and it improves the stability of model training by introducing a gradient penalty into the loss function. Qin et al. [17] introduced the elastic network into the objective function of a speech enhancement system based on WGAN and improve the performance of the model in low-resource data.

Because they are neither a need to make any assumptions about the original data nor a need to extract manual features, the GAN-based speech enhancement methods have achieved good performance results in time-domain noisy signals. However, there are still many problems to be improved in the generation effect and training stability of the model based on GAN. In this work, an end-to-end speech enhancement model based on WGAN, and some improvements have been made in the proposed model to make the model converge faster and more suitable for speech enhancement tasks. Specifically, in the generator coding part, each convolution layer adopts different convolution kernel sizes to conduct convolution operations for obtaining speech coding information from multiple scales. A gated linear unit (GLU) is introduced to alleviate the vanishing gradient problem with the increase in network depth. To accelerate the convergence of the model, spectral normalization is applied to the weights of the discriminator. To improve the quality and clarity of the generated speech, a hybrid penalty term consisting of L1 regularization and a scale-invariant signal-to-distortion ratio (SI-SDR) is introduced into the loss function of the generator. The whole network, abbreviated as SEGWAGN-HP, processes signals at the waveform level. Experiments show that, on both TIMIT corpus and Tibetan corpus, in contrast to other methods under the condition of unknown noise and low SNR, SEGWAGN-HP improves the quality of generated speech significantly, and it could accelerate the convergence speed of GAN.

The rest of the paper is organized as follows. Section 2 reviews the structure of WGAN briefly. Section 3 describes the proposed model. The experimental details are described in Section 4, and the experimental results and analysis are presented in Section 5. Finally, conclusions are made in Section 6.

2 Wasserstein Generative Adversarial Network

The GAN was proposed by Ian Goodfellow in 2014 [18]. It is originally used to generate images and is composed of two independent models: a generator and a discriminator. Their parameters are optimized iteratively with the process of adversarial learning, and their loss function can be represented as

$$\min_{G}\max_{D} V(D,G)=\mathbb{E}_{x\sim P_{data}(x)}\left[\log D(x)\right]+\mathbb{E}_{x\sim P_{z}(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right], \tag{1}$$

where D represents the discriminator, G represents the generator, $P_{data}(x)$ represents the real sample distribution, $P_{z}(z)$ represents the generated sample distribution, and $\mathbb{E}(\cdot)$ represent the mean of $\bullet$.

In the speech enhancement task, an enhanced speech signal is generated by the generator of the GAN. In order to make the enhanced speech content consistent with the noisy speech content, the noise information is added to the loss function of GAN as an additional condition [19]. It can be represented as

$$\begin{aligned}\min_{G}\max_{D} V(D,G)=&\mathbb{E}_{x,x_c\sim P_{data}(x,x_c)}\left[\log D(x,x_c)\right]\\&+\mathbb{E}_{x\sim P_{z}(z),x_c\sim P_{data}(x_c)}\left[\log\left(1-D\left(G\left(z,x_c\right),x_c\right)\right)\right],\end{aligned} \tag{2}$$

where x represents the clean speech, x_c represents the noisy speech, z represents the Random Gaussian noise.

The application of GAN brings a new idea for time-domain speech enhancement. However, the training of GAN needs to find the Nash equilibrium between the generator and the discriminator. It leads to such problems as convergence difficulty and model collapse. Therefore, in WGAN, Wasserstein distance (WD) is employed in the objective functions of both the generator and the discriminator, and a Lipschitz constraint is imposed on the discriminator to limit its gradient [20,21]. It makes the GAN training process stable. The objective functions of the generator and the discriminator of WGAN can be expressed as

$$\begin{aligned}\min_{D} V(D)=&\frac{1}{2}\mathbb{E}_{x,x_c\sim P_{data(x,x_c)}}\left[\left(D\left(x,x_c\right)-1\right)^2\right]\\&+\frac{1}{2}\mathbb{E}_{z\sim P_{z}(z),x_c\sim P_{data(x_c)}}\left[D\left(G\left(z,x_c\right),x_c\right)^2\right],\end{aligned} \tag{3}$$

and

$$\min_{G} V(G)=\frac{1}{2}\mathbb{E}_{z\sim P_{z}(z),x_c\sim P_{data}(x_c)}\left[\left(D\left(G\left(z,x_c\right),x_c\right)-1\right)^2\right] \tag{4}$$

In WGAN, to make the loss function of the discriminator satisfy 1-lipschitz continuity, the weight-clipping technique is adopted. However, it causes easily all the parameters of the discriminator to tend to extreme values, and it makes the parameter adjustment process more difficult [22,23]. To solve the above problems, Ishaan et al. proposed a model called WGAN-GP. It replaces the weight-clipping technique with a gradient penalty to satisfy 1-lipschitz continuity [24]. The loss function of its discriminator is changed to

$$\begin{aligned}\min_{D} V(D)=&\frac{1}{2}\mathbb{E}_{x,x_c\sim P_{data(x,x_c)}}\left[\left(D\left(x,x_c\right)-1\right)^2\right]\\&+\frac{1}{2}\mathbb{E}_{z\sim P_{z}(z),x_c\sim P_{data(x_c)}}\left[D\left(G\left(z,x_c\right),x_c\right)^2\right]+\lambda\left(\left|\left|\nabla_{x_t\sim P(x_t)}D\left(x_t\right)\right|\right|_2-1\right)^2,\end{aligned} \tag{5}$$

where λ is gradient penalty factor, ∇ represents gradient of $D(x_t)$, $||\bullet||_2$ represents the 2-norm of $\bullet$, $P(x_t)$ is the probability distribution of the whole sample space.

3 Proposed Model

The proposed model draws lessons from the SEGAN in reference [13]. In order to make the model converge faster and more suitable for speech enhancement tasks, the following improvements have been made: Firstly, in the generator coding part, each convolution layer adopts different convolution kernel sizes to conduct convolution operations for obtaining speech coding information from multiple scales. Secondly, a gated linear unit (GLU) is introduced to alleviate the vanishing gradient problem with the increase of network depth. Thirdly, spectral normalization is used to replace the gradient penalty of the discriminator to accelerate the convergence rate of the model. Finally, a hybrid penalty term composed of L1 regularization and SI-SDR is introduced into the loss function of the generator to improve the quality of generated speech.

3.1 Multi-Scale Convolution

In 2015, a deep convolutional neural network architecture codenamed InceptionNet was proposed by Szegedy et al. [25]. It uses convolution kernels of different sizes in the same layer network to improve the perception of the model. Therefore, to obtain speech coding information from multiple scales, multiple convolution kernels with different sizes are employed in each convolution layer in the SEGWAGN-HP. Its diagram is shown in Fig. 1.

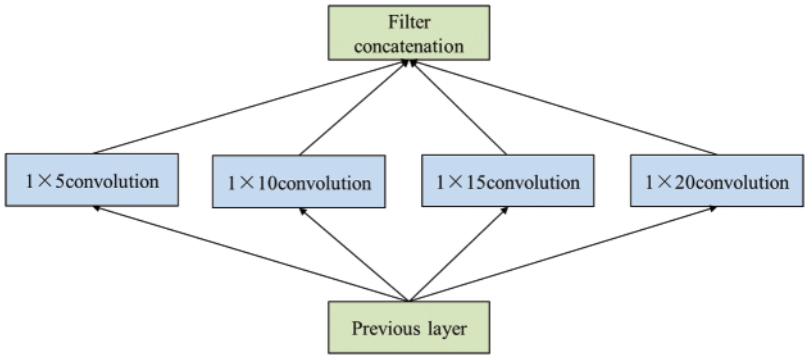

Figure 1: Schematic diagram of multi-scale convolution

As shown in Fig. 1, convolution is performed on speech information using convolution kernels of different sizes. The results of different convolutions are connected along the channel dimension as the input of the next layer. The output of each convolution layer is expressed as

$$out_{conv-layer} = concat\left(out_{conv1}, out_{conv2}, out_{conv3}, out_{conv4}\right), \tag{6}$$

where out_{conv1}, out_{conv2}, out_{conv3} and out_{conv4} are the results of convolution operation with multiple convolution kernels with different sizes.

3.2 Gated Linear Units

Adding gating mechanism into neural network can be used to control the transmission of information in neural network. It could alleviate the problems of gradient disappearance and gradient explosion in neural networks by adding a linear dependency between the front and back layers of the

neural network [26]. A gated linear unit is a simplified gating mechanism. It can alleviate the vanishing gradient problem by having linear units coupled to the gates, and it can be expressed as

$$y = (x * w_i + b_i) \otimes \sigma\,(x * w_i + b_i)\,, \tag{7}$$

where w_i and b_i denote filters and biases of the convolution layer, respectively, σ indicates the sigmoid activation function.

A convolutional gated linear units (GLU) block and a de-convolutional GLU block are introduced in the coding part and decoding part of SEGWAGN-HP to control the information flows throughout the model. Its diagram is shown in Fig. 2.

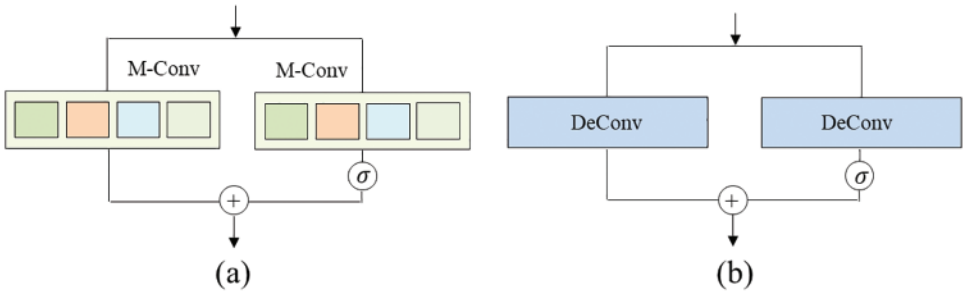

Figure 2: Diagrams of a convolutional GLU block and a de-convolutional GLU block

3.3 Spectral Normalization

By introducing a gradient penalty into the loss function of the discriminator, WGAN-GP avoids the phenomenon of gradient disappearance or gradient explosion. However, it takes effect only locally. 1-Lipschitz continuity will be invalid if there are too many kinds of samples. In this work, spectral normalization is adapted to make a function satisfy 1-Lipschitz continuity, and it is realized by dividing all parameters in the discriminator by its spectral norm [27]. The operation is expressed as

$$W_{SN} = \frac{W}{||W||_2}, \tag{8}$$

where $||W||_2$ is the spectral norm of the weight, and it equals to the maximum singular value of the weight matrix.

3.4 Hybrid Penalty

Regularization is a common technology in machine learning, and the basic method is to add a penalty term to the original objective function. In order to improve the generalization ability of the model and the auditory quality of the generated speech, a hybrid penalty term consisting of L1-normand SI-SDR is introduced into the loss function of the generator.

The L1-norm term accurately measures the distance between the generated samples and the real samples [28]. It is defined as

$$L_1 = ||G(z, x_c) - x||_1 \tag{9}$$

SI-SDR is an improved version of the traditional SDR. It evaluates the distortion and quality of speech by calculating the ratio of pure speech to residue noise [29]. The larger the value is, the smaller

the distortion is. Compared with the traditional SDR, it has better robustness and more accurate calculation results. It can be defined as

$$SI - SDR = 10 \log_{10} \left(\frac{||\alpha x||^2}{||\alpha x - \widetilde{x}||^2} \right), \tag{10}$$

where α is the best scale factor and $\widetilde{x}$ represents generated samples.

By introducing L1-norm into the loss function of the generator, it can improve the generalization ability of the model. By introducing SI-SDR into the loss function of the generator, the generator is guided to improve the auditory quality and intelligibility of the generated speech. The loss function of the generator with hybrid penalty term can be expressed as

$$\begin{aligned} \min_G V(G) = & \frac{1}{2} \mathbb{E}_{z \sim P_z(z), x_c \sim P_{data}(x_c)} \left[(D(G(z, x_c), x_c) - 1)^2 \right] \\ & + \lambda_1 L_1 - \lambda_2 (SI - SDR), \end{aligned} \tag{11}$$

where λ_1 and λ_2 are the control factors of the L1-norm term and the SI-SDR term.

3.5 Network Structure

In SEGWGAN-HP, the generator is an encoder-decoder structure. The encoder is used to compress the noisy speech signal into a condensed representation. The decoder consists of de-convolution layers, and it is used to recover the original data. In the generator, a gated linear unit (GLU) is introduced to alleviate the vanishing gradient problem for the increase of network depth. Meanwhile, each convolution layer adopts different convolution kernel sizes in the encoder to obtain speech coding information from multiple scales. In addition, a skip connection is added between each encoding layer and its homologous decoding layer. The skip connection directly transfers the waveform information of speech from the encoding layer to the corresponding decoding layer, and it improves the accuracy of speech reconstruction. The schematic diagram of the generator is shown in Fig. 3.

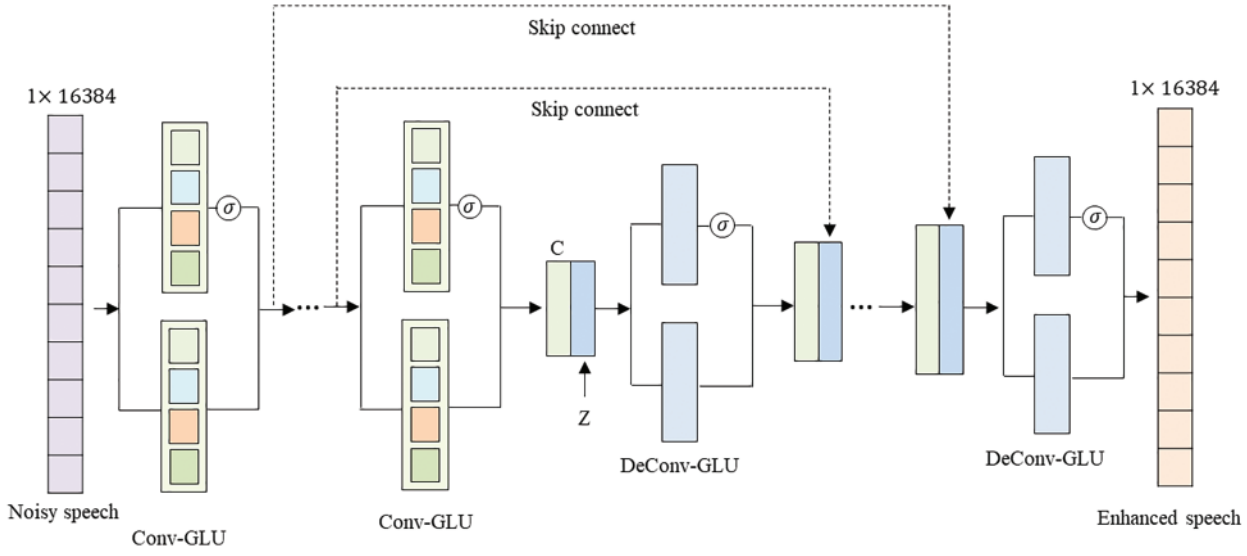

Figure 3: Schematic diagram of the generator

The discriminator is a binary classifier used to identify the authenticity of input data. It consists of 11 layers of one-dimensional stride convolution modules. In order to speed up the training speed of the model and alleviate the gradient disappearance problem, a Batch-Normalization layer and a

SELU activation function layer are added after each convolution layer. The schematic diagram of the convolution module and the discriminator are shown in Figs. 4a and 4b.

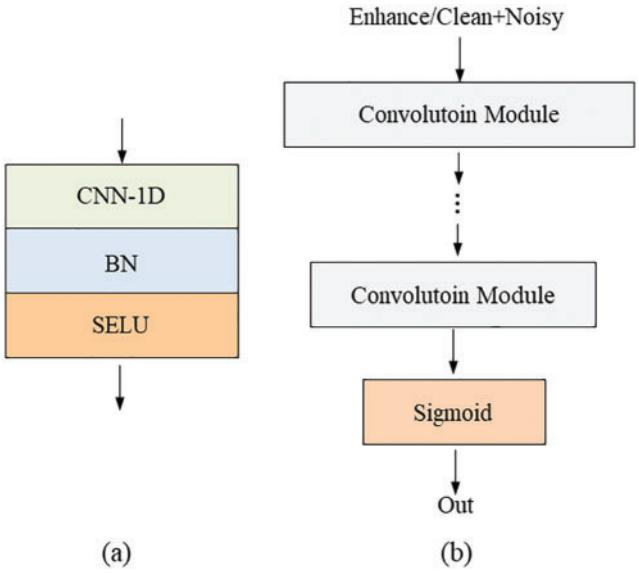

Figure 4: Schematic diagram of the convolution module and discriminator

4 Experimental Setup

4.1 Database

The experiments are conducted on two datasets[1], whose clean speeches are chosen from TIMIT corpus and Tibetan corpus. TIMIT corpus is a speech corpus jointly developed by Texas Instruments (TI), Massachusetts Institute of Technology (MIT) and Stanford Research Institute (SRI) and is one of the most widely used corpora in the field speech signal processing. It contains 6300 sentences spoken by 630 people of different genders from eight dialect areas in the United States. Among them, 4620 sentences are designated as the training set, while the remaining 1680 sentences constitute the test set. The Tibetan corpus is recorded by 12 speakers in a laboratory environment, and the sampling rate is 16 kHz with the mono channel. This corpus contains 2968 pieces of speech, and the total length reaches 6 h. 2500 sentences are chosen as the training set while the remaining 468 sentences constitute the test set. In this work, both the stationary and non-stationary noise conditions are considered. The noise data are chosen from datasets 100 Non-Speech Sounds and NOISEX-92, which contain 100 and 15 different noises, respectively.

To reduce the volume of experimental data, all speeches are down sampled to 8 KHz. In the training process, each speech of the training data is divided into several frames with a length of 16,384 sample points and a shift of 8192 sample points. In the test process, there is no overlapping during frame splitting. Each frame of noisy speech is enhanced and stitched to get the final enhanced signal. The creation details of two experimental datasets are as follows: 4500 clean speeches selected from the TIMIT training set are corrupted with 10 types of noise from 100 Non-Speech Sounds at 4 levels of SNR, i.e., 10, 5, 0 and −5.0 dB. As a result, a training data set of about 80 h is constituted. 500 clean speeches selected from the TIMIT test set are corrupted with 4 types of noise (Babble, Leopard, Volvo,

[1] https://pan.baidu.com/s/1DZRrlMToTXYkGS7U0iGhRw?pwd=aaaa.

and factory2) from NOISEX-92 at 4 levels of SNR, i.e., 1, 7.5, 2.5, and −2.5 dB. As a consequence, a testing data set of about 20 h is constituted. When constructing the training and testing sets from the Tibetan corpus, the same type of noise and the same method of adding noise are adopted.

4.2 Network Setup and Training Details

There are 22 layers in the generator of SEGWGAN-HP. The encoder and decoder each have 11 layers. The specific parameter settings of the generator are shown in Table 1.

Table 1: The specific parameter settings of generator

Layer name	Input size	Output size	Layer name	Input size	Output size
Conv-GLU-1	[B, 16384, 1]	[B, 8192, 16]	DeConv-GLU-11	[B, 8192, 16 × 2]	[B, 16384, 1]
Conv-GLU-2	[B, 8192, 16]	[B, 4096, 32]	DeConv-GLU-10	[B, 4096, 32 × 2]	[B, 8192, 16]
Conv-GLU-3	[B, 4096, 32]	[B, 2048, 32]	DeConv-GLU-9	[B, 2048, 32 × 2]	[B, 4096, 32]
Conv-GLU-4	[B, 2048, 32]	[B, 1024, 64]	DeConv-GLU-8	[B, 1024, 64 × 2]	[B, 2048, 32]
Conv-GLU-5	[B, 1024, 64]	[B, 512, 64]	DeConv-GLU-7	[B, 512, 64 × 2]	[B, 1024, 64]
Conv-GLU-6	[B, 512, 64]	[B, 256, 128]	DeConv-GLU-6	[B, 256, 128 × 2]	[B, 512, 64]
Conv-GLU-7	[B, 256, 128]	[B, 128, 128]	DeConv-GLU-5	[B, 128, 128 × 2]	[B, 256, 128]
Conv-GLU-8	[B, 128, 128]	[B, 64, 256]	DeConv-GLU-4	[B, 64, 256 × 2]	[B, 128, 128]
Conv-GLU-9	[B, 64, 256]	[B, 32, 256]	DeConv-GLU-3	[B, 32, 256 × 2]	[B, 64, 256]
Conv-GLU-10	[B, 32, 256]	[B, 16, 512]	DeConv-GLU-2	[B, 16, 512 × 2]	[B, 32, 256]
Conv-GLU-11	[B, 16, 512]	[B, 8, 1024]	DeConv-GLU-1	[B, 8, 1024 × 2]	[B, 16, 512]

There are 11 convolution module in the discriminator of SEGWGAN-HP. The specific parameter settings of the discriminator are shown in Table 2.

Table 2: The specific parameter settings of discriminator

Layer name	Input size	Output size	Layer name	Input size	Output size
Conv-module-1	[B, 16384, 2]	[B, 8192, 32]	Conv-module-8	[B, 128, 256]	[B, 64, 512]
Conv-module-2	[B, 8192, 32]	[B, 4096, 64]	Conv-module-9	[B, 64, 512]	[B, 32, 512]
Conv-module-3	[B, 4096, 64]	[B, 2048, 64]	Conv-module-10	[B, 32, 512]	[B, 16, 1024]
Conv-module-4	[B, 2048, 64]	[B, 1024, 128]	Conv-module-11	[B, 16, 1024]	[B, 8, 2048]
Conv-module-5	[B, 1024, 128]	[B, 512, 128]	Conv-12	[B, 8, 2048]	[B, 8, 1]
Conv-module-6	[B, 512, 128]	[B, 256, 256]	Dense-13	[B, 8, 1]	[B, 1, 1]
Conv-module-7	[B, 256, 256]	[B, 128, 256]	Sigmoid-14	[B, 1, 1]	[B, 1, 1]

For the generator, the last layer uses TanH as an activation function, while the others use SELU; for the discriminator, each layer employs SELU as the activation function. The initial learning rate of the network is set to 0.0002, and the model parameters are optimized by Root Mean Square Prop (RMSProp).

To compare the effects of noise reduction and the generalization capability of the proposed model, five methods are selected as comparative experiments: unprocessed noise baseline (unprocessed), Ideal

Ratio Mask (IRM) [30], SEGAN [13], Minimum Mean Square Error-Spectrum Power estimator based on zero cross-terms (MMSE-SPZC) [31], and DCCRN [32]. PESQ, STOI, and SegSNR are used to evaluate the performance of speech enhancement.

5 Results and Analysis

This work only considers speech enhancement in the case of additive noise. Four types of experiments are carried out to verify the performance of the proposed model. They are:

(1) Experiment of super-parameters;

(2) Performance comparison of different enhancement approach on TIMIT corpus;

(3) Ablation experiments of the proposed model;

(4) Performance comparison of different enhancement approach on Tibetan corpus.

5.1 Experiment of Super-Parameters

To improve the convergence speed of the model and the quality of the generated speech, a hybrid penalty is introduced into SEGWGAN-HP. The loss function of the generator is calculated with Eq. (11) and λ_1 and λ_2 are the adjustable hyper-parameters. In this section, experiments are designed to explore the influence of the super parameters.

5.1.1 Control Factor of L_1 Regularization

In this experiment, only the item of L_1 regularization is introduced to the loss function of the generator. The value of λ_1 is increased gradually from 0 to 140 with an interval of 20. The corresponding values of PESQ, STOI, and SegSNR are shown in Fig. 5.

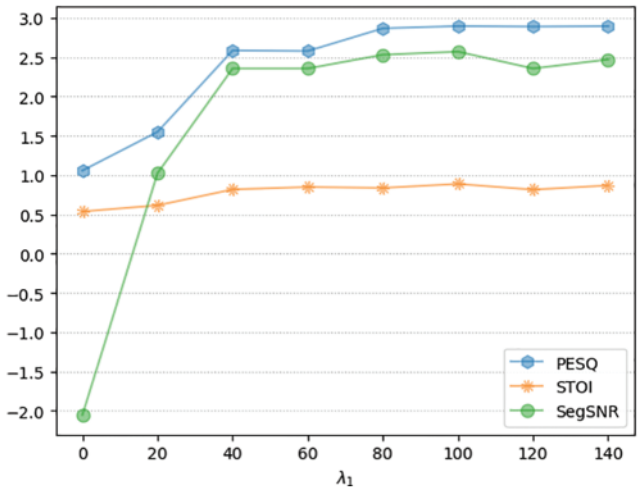

Figure 5: The average values of PESQ, STOI and SegSNR as the value of increasing gradually from 0 to 140 with an interval of 20

It can be seen from Fig. 5 that the average values of the indicators PESQ, STOI, and SegSNR increase sharply as the value of λ_1 changes from 0 to 40; with the further increase of λ_1, the values of the three indicators increase gently; the values of the three indicators reach their maximum while

the value of λ_1 is equal to 100. Therefore, it is appropriate to set the value of λ_1 to 100 in subsequent experiments.

5.1.2 Control Factor of SI-SDR

Based on the previous experiment, the SI-SDR item is introduced to the loss function of the generator. As the value of λ_2, the control factor of SI-SDR, increases gradually from 0 to 50 with an interval of 10, the average values of the three indicators PESQ, STOI, and SegSNR are shown in Fig. 6.

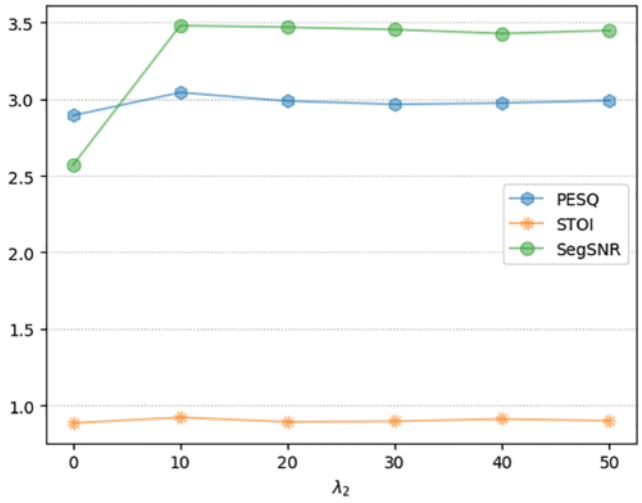

Figure 6: The average values of the three indicators PESQ, STOI and SegSNR as the value of increasing gradually from 0 to 50 with an interval of 10

From Fig. 6, it is evident that the values of the three indicators PESQ, STOI, and SegSNR increase as the value of λ_2 increases gradually from 0 to 50. In particular, the increase of the indicator SegSNR is significant. This shows that the loss function optimized by the SI-SDR term could evaluate the quality of the generated speech, and it forces the generator to be trained in the direction of generating high quality speech. The overall performance is optimal while the value of λ_2 is equal to 10. Therefore, the value of λ_2 is set to be 10 in subsequent experiments.

5.2 Performance Comparison of Different Enhancement Approach on TIMIT Corpus

5.2.1 Evaluation of Speech Quality

PESQ is one of the most popular objective measures in the field of speech enhancement, and its score range is [−0.5, 4.5]. It has the highest correlation with the score of subjective audiometry (Mean Opinion Score). The higher the score is, the better the performance of the model gets. Table 3 shows the average PESQ results of the proposed model and several other competing models on the TIMIT dataset.

As shown in Table 3, in the cases where SNR is 7.5 or 12.5 dB, MMSE-SPZC achieves almost the same performance as other methods; in the cases where SNR is −2.5 and 2.5 dB, the performances of the two speech enhancement methods based on GAN are significantly higher than those of MMSE-SPZC and IRM. The PESQ score obtained by SEGWGAN-HP is higher than other models except that it is slightly lower than DCCRN model at 12.5 dB SNR. In particular, in the cases where SNR

is −2.5 and 2.5 dB, SEGWGAN-HP gains more obvious improvement. It proves that the proposed model has better de-noising performance in the case of low SNR.

Table 3: Performance measurement: Average PESQ on TIMIT dataset

Model	PESQ scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	1.756	2.093	2.433	2.764
MMSE-SPZC	2.081	2.425	2.767	3.084
IRM	2.114	2.426	2.805	3.123
SEGAN	2.248	2.527	2.785	3.101
DCCRN	2.346	2.594	2.791	3.204
SEGWGAN-HP	2.447	2.631	2.804	3.121

5.2.2 Evaluation of Speech Intelligibility

STOI is a common measure to evaluate speech intelligibility and clarity, and its value range is [0, 1]. The higher the value is, the better the intelligibility of speech is. Table 4 shows the average results of the STOI of the proposed model and several other competing models on the TIMIT dataset.

Table 4: Performance measurement: Average STOI on TIMIT dataset

Model	STOI scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	66.81	73.56	79.63	85.09
MMSE-SPZC	70.66	77.61	83.44	88.30
IRM	73.55	81.33	86.24	89.87
SEGAN	75.59	82.31	86.28	89.15
DCCRN	76.42	82.52	86.33	89.37
SEGWGAN-HP	76.13	82.34	86.32	89.56

It can be seen from Table 4 that the proposed model does not gain the best performance in terms of STOI. In the case where SNR is equal to 12.5 dB, the STOI score obtained by SEGWGAN-HP is higher than those of the other models. In the cases where the SNR increases from −2.5 to 7.5 dB, the STOI scores obtained by SEGWGAN-HP are slightly lower than those of other models. Comparing the results of Tables 3 and 4, it could be concluded that the STOI score of SEGWGAN-HP has no evident difference to that of other comparison methods, but the score of PESQ is much higher than that of other methods.

5.2.3 Evaluation of Signal Distortion Degree

SegSNR is an evaluation indicator of speech enhancement based on the time domain. Compared with SNR, it can better represent the noise distortion degree of the speech signal because speech is a

short-term stable signal. The SegSNR values obtained by various speech enhancement methods are shown in Table 5.

Table 5: Performance measurement: Average SegSNR on TIMIT dataset

Model	SegSNR scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	−3.847	−2.124	1.107	2.746
MMSE-SPZC	−2.053	0.622	3.354	5.569
IRM	0.175	1.296	1.813	3.123
SEGAN	0.605	1.954	2.228	3.229
DCCRN	0.892	3.217	4.324	6.177
SEGWGAN-HP	1.114	3.421	4.339	6.143

It can be seen from Table 5 that the IRM and SEGAN models achieve lower SegSNR results than that of MMSE-SPZC at high SNR. The reason is that these two models remove a lot of real speech signals while removing noise signals. In the SEGWGAN-HP model, an evaluation criterion for speech quality is introduced to the loss function. The value of SegSNR has significantly improved. This shows that the hybrid penalty term can greatly improve the de-noising performance of the model without damaging the quality and intelligibility of the speech.

5.3 Ablation Experiments of the Proposed Model

To investigate the effectiveness of each component in our proposed model, such as Multi-scale Convolution, GLU, Spectral Normalization, and Hybrid Penalty, ablation experiments are carried out on datasets that are constructed based on the TIMIT corpus. Here, the base model is SEWGAN-GP, which refers to the speech enhancement system constructed by WGAN with a gradient penalty term; SEWGAN-SN indicates that the gradient penalty term in SEWGAN-GP is replaced by spectral normalization. SEGWGAN-SN indicates the addition of Multi-scale Convolution and GLU on the basis of SEWGAN-SN. SEGWGAN-SN-HP indicates that a mixed penalty term is added to the loss function of the generator in SEWGAN-SN. The average evaluation results of the four ablation experiments are shown in Table 6.

Table 6: The average evaluation results of the four ablation experiments

Model	PESQ	STOI	SegSNR
SEWGAN-GP	2.3478	0.7914	3.3469
SEWGAN-SN	2.3513	0.7926	3.3983
SEGWGAN-SN	2.5213	0.8298	3.6436
SEGWGAN-SN-HP	2.7507	0.8358	3.7542

From Table 6, when spectral normalization is used instead of gradient punishment, three indexes are slightly improved, and its effect is mainly reflected in the convergence speed of the model. The addition of Multi-scale Convolution, GLU, and Hybrid Penalty has significantly improved three

indicators. In particular, the scores of PESQ have significantly improved since the introduction of the hybrid penalty term.

5.4 Performance Comparison of Different Enhancement Approach on Tibetan Corpus

5.4.1 Performance Evaluation on Tibetan Corpus

The effect of speech enhancement depends on the recovery of effective sound units in noisy speech. However, different languages have different phoneme distributions. Therefore, the same speech enhancement algorithm may yield different improvement performances for different languages. To verify the enhancement performance of the proposed model on the different speech corpora, the same experiments are carried out on the Tibetan corpus. The experimental results in terms of three objective metrics are shown in Tables 7–9, respectively.

Table 7: Performance measurement: Average PESQ on Tibetan dataset

Model	PESQ scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	1.646	1.973	2.303	2.617
MMSE-SPZC	1.951	2.215	2.617	2.924
IRM	1.944	2.266	2.695	2.983
SEGAN	1.968	2.247	2.665	2.991
DCCRN	1.974	2.268	2.672	3.037
SEGWGAN-HP	2.007	2.311	2.664	2.983

Table 8: Performance measurement: Average STOI on Tibetan dataset

Model	STOI scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	65.76	72.35	77.81	83.63
MMSE-SPZC	69.34	76.56	81.74	85.89
IRM	72.21	80.69	83.63	87.84
SEGAN	74.28	82.18	84.16	88.13
DCCRN	75.87	82.47	84.87	88.65
SEGWGAN-HP	75.67	82.32	84.97	88.72

Table 9: Performance measurement: Average SegSNR on Tibetan dataset

Model	SegSNR scores at different SNR			
	−2.5	2.5	7.5	12.5
Unprocessed	−6.177	−3.494	0.073	3.376
MMSE-SPZC	−2.432	0.401	3.034	6.218
IRM	0.085	1.056	1.442	2.793

(Continued)

Table 9 (continued)

Model	SegSNR scores at different SNR			
	−2.5	2.5	7.5	12.5
SEGAN	0.164	1.174	1.557	2.868
DCCRN	0.887	2.946	4.786	6.473
SEGWGAN-HP	0.904	3.011	5.029	6.802

Table 7 depicts the average PESQ scores obtained by the SEGWGAN-HP model and several other competing models on the Tibetan datasets.

Table 8 depicts the average STOI scores obtained by the SEGWGAN-HP model and several other competing models on the Tibetan datasets.

Table 9 depicts the average SegSNR scores obtained by the SEGWGAN-HP model and several other competing models on the Tibetan datasets. From the above three tables, it can be observed that the experimental results on the TIMIT dataset are consistent with the experimental results on the TIMIT dataset in terms of three objective metrics. SEGWGAN-HP also has good de-noising performance on the Tibetan dataset. Among them, the improvement in term of the SegSNR metrics is significant. It proves that SEGWGAN-HP has good language generalization ability.

5.4.2 Comparison and Analysis of Spectrograms

To analyze intuitively the performance of speech enhancement in the proposed model, a noisy speech is selected from the Tibetan test set whose noise has been added at 2.5 dB SNR condition, and it is used for speech enhancement in different algorithms. The spectrogram is shown in Fig. 7.

It is easy to see that the overall outline obtained by MMSE-SPZC is fuzzier than that of other methods. The enhancement effect of the proposed model is much better than that of the other methods, and the obtained overall contour and energy distribution are the closest to the real speech spectrum. Although the de-noising effects of the IRM and SEGAN are as good as those of SEGWGAN-HP, as shown in the black box in Fig. 7, they destroy the energy texture of some real speech signals while removing noise. The absolute value of the difference between enhanced speech and clean speech is inversely proportional to the score of SegSNR. The destruction of speech energy texture reduces the SegSNR value of the models IRM and SEGAN, and it is consistent with the conclusion of the previous experiment.

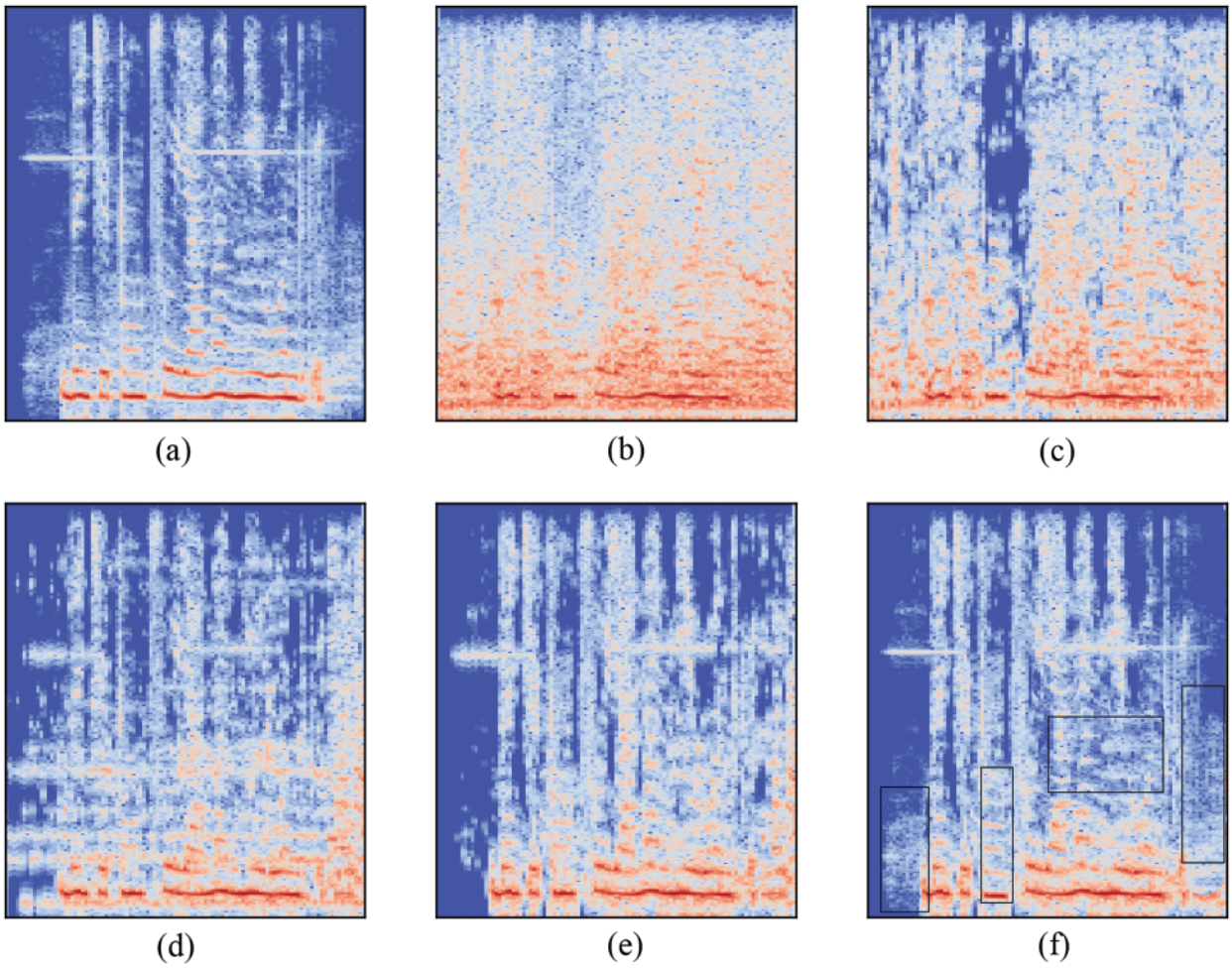

Figure 7: Comparison and analysis of spectrograms. (a) Clear. (b) Unprocessed. (c) MMSE-SPZC. (d) IRM. (e) SEGAN (f) SEGWGAN-HP

6 Conclusions

In this work, SEGWGAN-HP, a novel end-to-end speech enhancement model based on WGAN, is proposed. In the proposed model, some improvements have been made to make the model converge faster and more suitable for speech enhancement tasks. Specifically, a gated linear unit (GLU) is introduced in the generator to alleviate the vanishing gradient problem with the increase of network depth. Meanwhile, each convolution layer adopts different convolution kernel sizes in the encoder part of the generator to obtain speech coding information at multiple scales. In addition, a hybrid penalty term combining regularization and speech quality measures is introduced to the loss function of the model to get better generated speech quality. The experimental results on both TIMIT corpus and Tibetan corpus show that SEGWGAN-HP improves the speech quality significantly and it can be well applied to different speech corpora.

In this paper, the case of additive noise is only considered when constructing the noisy sound database. However, reverberation and echo problems usually exist in the real scenario. Therefore,

adding the de-reverberation capability to the model is one of the future research directions. In addition, there are 22 convolutional layers in the generator of the SEGWGAN-HP. There are 11 convolution modules in the discriminator of SEGWGAN-HP. It makes the model spend more time in the training process. Therefore, simplifying the model structure while ensuring performance is another research direction in the future.

Funding Statement: This work is partially supported by the National Science Foundation under Grant No. 62066039. This work is also partially supported by the State Key Laboratory of Tibetan Intelligent Information Processing and Application.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Chai, L., Du, J., Liu, Q. F., Lee, C. H. (2019). Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12),* 1919–1931. DOI 10.1109/TASLP.6570655.
2. Jyoshna, G., Rahman, M. Z. U., Koteswararao, L. (2022). An efficient reference free adaptive learning process for speech enhancement applications. *Computers, Materials & Continua, 70(2),* 3067–3080. DOI 10.32604/cmc.2022.020160.
3. Zhu, X., Huang, H. (2020). End-to-end Amdo-Tibetan speech recognition based on knowledge transfer. *IEEE Access, 8,* 170991–171000. DOI 10.1109/ACCESS.2020.3023783.
4. Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B. et al. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.
5. Hamed, I., Denisov, P., Li, C. Y., Elmahdy, M., Abdennadher, S. et al. (2022). Investigations on speech recognition systems for low-resource dialectal arabic–English code-switching speech. *Computer Speech & Language, 72,* 101278. DOI 10.1016/j.csl.2021.101278.
6. Ullah, F., Naeem, M. R., Naeem, H., Cheng, X., Alazab, M. (2022). CroLSSim: Cross-language software similarity detector using hybrid approach of LSA-based AST-MDrep features and CNN-LSTM model. *International Journal of Intelligent Systems, 2022,* 1–28. DOI 10.1002/int.22813.
7. Abad-Segura, E., Infante-Moro, A., González-Zamar, M. D., López-Meneses, E. (2021). Blockchain technology for secure accounting management: Research trends analysis. *Mathematics, 9(14),* 1631. DOI 10.3390/math9141631.
8. Xu, Y., Du, J., Dai, L. R., Lee, C. H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1),* 7–19. DOI 10.1109/TASLP.6570655.
9. Huang, P. S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P. (2015). Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(12),* 2136–2147. DOI 10.1109/TASLP.2015.2468583.
10. Pandey, A., Wang, D. (2019). TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6875–6879. Brighton, UK.
11. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. et al. (2017). Least squares generative adversarial networks. *International Conference on Computer Vision (ICCV)*, pp. 2813–2821. Venice, Italy.
12. Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X. et al. (2017). Generative adversarial networks: Introduction and outlook. *IEEE/CAA Journal of Automatica Sinica, 4(4),* 588–598. DOI 10.1109/JAS.6570654.

13. Pascual, S., Bonafonte, A., Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452.
14. Michelsanti, D., Tan, Z. H. (2017). Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. arXiv preprint arXiv:1709.01703.
15. Shah, N., Patil, H. A., Soni, M. H. (2018). Time-frequency mask-based speech enhancement using convolutional generative adversarial network. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1246–1251. Honolulu, USA.
16. Adiga, N., Pantazis, Y., Tsiaras, V., Stylianou, Y. (2019). Speech enhancement for noise-robust speech synthesis using wasserstein GAN. *Interspeech*, pp. 1821–1825. Graz, Austria.
17. Qin, S., Jiang, T. (2018). Improved wasserstein conditional generative adversarial network speech enhancement. *EURASIP Journal on Wireless Communications and Networking, 2018(1),* 1–10. DOI 10.1186/s13638-018-1196-0.
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D. et al. (2014). Generative adversarial nets. In: *Neural information processing systems*, vol. 27, pp. 2672–2680, Montreal, Canada.
19. Mirza, M., Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
20. Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein generative adversarial networks. *International Conference on Machine Learning (ICML)*, pp. 214–223. Sydney, Australia.
21. Donahue, C., Li, B., Prabhavalkar, R. (2018). Exploring speech enhancement with generative adversarial networks for robust speech recognition. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5024–5028. Calgary, Canada.
22. Yang, F., Wang, Z., Li, J., Xia, R., Yan, Y. (2020). Improving generative adversarial networks for speech enhancement through regularization of latent representations. *Speech Communication, 118,* 1–9. DOI 10.1016/j.specom.2020.02.001.
23. Baby, D., Verhulst, S. (2019). Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106–110. Brighton, UK.
24. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. (2017). Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028.
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. et al. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. Boston, USA.
26. Wang, Y., Yu, G., Wang, J., Wang, H., Zhang, Q. (2020). Improved relativistic cycle-consistent GAN with dilated residual network and multi-attention for speech enhancement. *IEEE Access, 8,* 183272–183285. DOI 10.1109/Access.6287639.
27. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.
28. Li, S., Liu, H., Zhou, Y., Luo, Z. (2020). A SI-SDR loss function based monaural source separation. *International Conference on Signal Processing (ICSP)*, pp. 356–360. Suzhou, China.
29. Phan, H., McLoughlin, I. V., Pham, L., Chén, O. Y., Koch, P. et al. (2020). Improving GANs for speech enhancement. *IEEE Signal Processing Letters, 27,* 1700–1704. DOI 10.1109/LSP.97.
30. Tu, M., Zhang, X. (2017). Speech enhancement based on deep neural networks with skip connections. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5565–5569. New Orleans, USA.

31. Lu, Y., Loizou, P. C. (2010). Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. *IEEE Transactions on Audio, Speech, and Language Processing, 19(5),* 1123–1137. DOI 10.1109/TASL.2010.2082531.
32. Tan, K., Wang, D. (2019). Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6865–6869. Brighton, UK.