Single-Channel Speech Quality Enhancement in Mobile Networks Based on Generative Adversarial Networks

WU, G.; HERENCSÁR, N.

Mobile Networks and Applications

vol. 2024

ISSN: 1572-8153

DOI: http://dx.doi.org/10.1007/s11036-024-02300-4

Accepted manuscript

Single-Channel Speech Quality Enhancement in Mobile Networks Based on Generative Adversarial Networks

Guifen Wu[1], Norbert Herencsar[2]

[1]Department of Computer Science and Technology, Lyuliang University, Lv Liang 033001, China
20051014@llu.edu.cn
[2]Department of Telecommunications, Faculty of Electrical Engineering and Communication,
Brno University of Technology, Technicka 12, 616 00 Brno, Czechia
herencsn@ieee.org

Abstract: A large amount of randomly generated noise in mobile networks leads to a lack of targeting and gaming processes in the speech enhancement process, and the enhancement process from the perspective of acoustic features alone suffers from major drawbacks. Propose a single-channel speech quality enhancement method based on generative adversarial networks in mobile networks. Explain the principle of generative adversarial network to realize single-channel speech quality enhancement in mobile networks and clarify its shortcomings. Design an improved Mel frequency cepstral coefficient extraction method to extract 12 base features as the enhancement basis. Use the relative average least squares loss instead of the traditional loss function to enhance the training efficiency, use the hybrid penalty term to enhance the generator's ability to generate single-channel speech, and optimize the discriminator through the multi-layer convolution and the addition of fully connected layers to enhance the speech quality enhancement ability of adversarial generative networks in various aspects, forming a relative average generative adversarial network (RaGAN) based on hybrid penalty term to realize single-channel speech quality enhancement processing. Through the experiment, when the discriminator is applied with the size of a 3*3 convolutional kernel, the best effect of speech quality enhancement is achieved in the mobile network. This method can complete the enhancement of single-channel speech quality in the mobile network, and the effect is significant, which can effectively reduce the noise in the original single-channel speech.

Keywords: generative adversarial networks; RaGAN; hybrid penalty term; single-channel; speech quality; discriminator; mobile networks

1 Introduction

Speech, as an efficient information interaction carrier, is one of the most frequent and basic methods used by people during face-to-face and telephone communication in daily life. However, many different types of noise exist in both real-life and mobile networks [1]. For example, street and restaurant noise or natural random noise such as wind and running water in people's daily life scenarios. Due to noise interference, the auditory quality and intelligibility of speech signals are damaged to different degrees, reducing the efficiency and accuracy of human-computer interaction [2], and affecting the user's trust and satisfaction with the interaction system. To solve the problem of noise interference of speech signals in mobile network propagation, speech enhancement technology has emerged [3,4]. As a front-end processing technology to remove noise interference, speech enhancement aims to efficiently suppress the background noise in the speech signal after it has been contaminated by various types of noise, and improve the clarity and intelligibility of the speech signal, to extract the purest possible speech signal [5]. Speech enhancement techniques have been widely used in automatic speech processing tasks such as automatic speech recognition,

emotion recognition, and hearing repair.

Speech enhancement tasks can be categorized into two main types: single-channel speech enhancement and multi-channel speech enhancement [6,7], which are divided mainly based on the number of microphones. However, in many practical scenarios, such as telephones, recording devices, mobile devices, etc., they are usually equipped with only one microphone. Therefore, single-channel speech enhancement technology has a wider application prospect. Therefore, the research is carried out for the single-channel speech quality enhancement processing method. Traditional single-channel speech enhancement algorithms are proposed based on speech signal processing, which is mainly divided into time-domain methods and frequency-domain methods. The filter design method [8,9] in the time domain method and the short-time spectral estimation method [10] in the frequency domain method can realize single-channel speech enhancement processing. The filter design method mainly estimates the channel parameters and excitation parameters through the filter and constructs the filter to remove the noise. Short- and medium-time spectral estimation methods decompose the mixed speech signal in the frequency domain, obtain the amplitude and phase information of each frequency component, and realize speech enhancement by estimating the power spectrum of the speech signal and the power spectrum of the noise to obtain the denoised speech signal. However, the human vocal organs have complex physiological characteristics, and the speech signal is continuous; the channel parameters are difficult to estimate accurately, which leads to the poor noise reduction effect of the filter. It is impossible to ensure the continuity of the speech signal, resulting in abrupt changes in the transition band between frames, affecting the quality of the enhanced speech signal.

With the development of science and technology, the research of related methods has achieved certain results; for example, Kajla P et al. proposed the use of a two-channel sparse adaptive filtering method to enhance the quality of speech [11], using two mixed sound signals as inputs to estimate the original signals that produce these mixes, using a sparse learning strategy of the two-channel blind source separation method, mixing the sparse characteristics of the impulse response of the modeled acoustic path, and outputting the enhanced speech signal. However, this method is ineffective in single-channel speech enhancement tasks due to its inability to fully utilize the sparse characteristics of speech features. Garg A et al. proposed a deep convolutional neural network-based enhancement method for speech signals with a wide range of speech features [12], which will be used to decompose input speech signals into a series of overlapping frames using a Hanning window for frame splitting in the preprocessing stage. Multiple features such as improved Mel frequency cepstral coefficients (IMFCCs), fractional order delta AMS, and improved STFT (M-STFT) are extracted from these individual frames. However, when this method is used for single-channel speech enhancement, the final enhancement effect is affected due to the presence of insufficient training and overfitting. The residual gated recurrent neural network augmented Kalman filtering for the speech enhancement method proposed by Saleem N. et al. [13], in which clean speech and noise signals are modeled as an autoregressive process, with the parameters consisting of the linear prediction coefficients and the driving noise variance A recurrent neural network is trained to estimate the line spectral frequency (lfs) to obtain the noise variance, thus minimizing the difference between the modeled and predicted autoregressive spectra of noise-polluted speech. However, when this method is used for single-channel speech enhancement, it results in poor intelligibility of the processed speech due to poor training of the network.

Valentini-Botinhao et al. proposed a speech enhancement method based on RNN [14]. A

recurrent neural network (RNN) is trained to map acoustic features extracted from noisy speech to features describing clean speech, and then the enhanced data is used to train text-to-speech constructed from noisy speech to complete the enhancement of speech. However, due to the limited features that RNN can capture, the effect of speech enhancement obtained by this method in practical application is not ideal. Fu et al. proposed an improved MetricGAN speech enhancement method [15]. Considering that the objective evaluation index of human perception can be used as a bridge to narrow the gap, three kinds of training techniques integrating the knowledge of the speech processing domain are proposed to achieve discriminator optimization and complete speech enhancement. However, this method considers the optimization of the generator, resulting in its poor generalization ability in practical applications, which limits its performance. Wang et al. proposed a neural cascade structure to solve the problem of monaural speech enhancement [16]. The cascade structure consists of three modules that, in turn, optimize enhanced speech in terms of amplitude spectrum, time domain signal, and complex spectrum. Each module takes the noisy speech, and the output obtained from the previous module as input and generates a prediction of the corresponding target. Moreover, trained in an end-to-end manner, the three-domain loss function is used to interpret the three domains represented by the signal to complete monaural speech enhancement. However, this side involves combining and optimizing multiple modules in the implementation process, which can result in a more complex overall approach than a single-module approach.

At the same time, a neural network's computation may be large, which has certain requirements for hardware resources and real-time. Fan proposed a complementary single-channel voice enhancement network (CompNet) [17]. First, the noise speech is enhanced through the time-domain network, but when the problem is reconsidered in the frequency domain, the distribution of the time-frequency box may still be partially different from the target spectrum. To solve this problem, a dedicated dual-path network is designed as a post-processing module that independently filters amplitude and refines phase. This further pushes the estimated spectrum closer to the target spectrum in the time-frequency domain, thus completing single-channel speech enhancement. However, when the time-frequency conversion is not considered in the implementation process, there may be a certain degree of spectrum distortion, which leads to the difference between the recovered speech and the original target speech in some frequency ranges. Li et al. proposed a universal expansion framework for both single and multi-channel voice enhancement tasks [18]. The complex spectrum recovery is transformed into the spectral amplitude mapping of noise-mixed neighborhood space, in which an unknown sparse term is introduced, and phase correction is carried out in advance. On this basis, the mapping function is decomposed into a superposition of 0-order polynomials and higher-order polynomials in the Taylor series, where the former coarsely interferes in the amplitude domain, and the latter gradually fills in the remaining spectral details in the complex spectral domain. In addition, the relationship between adjacent-order terms is studied, revealing that each higher-order term can recursively estimate its lower-order terms, which are then expanded using the Taylor series and evaluated by proxy functions during the implementation. This can lead to greater overall complexity and higher demands on computing resources. Yu et al. used an improved band split recurrent neural network (BSRNN) and multi-resolution spectrogram discriminator to improve the perception quality and complete speech enhancement [19]. However, it does not consider the generator, resulting in its good performance only on specific data sets and tasks, lack of generalization ability, and poor application effect.

Therefore, in response to the problems of the above methods, a single-channel speech quality

enhancement processing method based on generative adversarial networks is proposed. Using relative average least squares loss instead of traditional loss function to enhance training efficiency and utilizing a mixed penalty term composed of L1 regularization and mean square error to enhance the generator's ability to generate single-channel speech, that is, while ensuring the stability of generator G, considering sparsity and prediction accuracy comprehensively, adding regularization terms to the model training process to reduce the number of parameters and computational complexity, to ensure real-time performance, To further assist generator G in filtering noise and redundant features, and improve the model's generalization ability. Moreover, receive the improved MFCC as input to learn from the features and structure of the original speech and preserve its prosodic, intonation, and spectral features as much as possible when generating enhanced speech. By adding mean square error, the similarity between the generated speech signal and the clear speech is maximized, helping the network learn to more accurately fit the target value and make the generated result as close as possible to the real target, improving the quality of single-channel generated speech; based on the discriminator, a fully connected layer is added to transform the output of the convolutional layer into global features, providing more nonlinear expression ability and decision boundaries. This not only improves the stability of network training, reduces competition and imbalance between the generative and discriminative networks, but also enhances the discriminator's ability to distinguish between real and generated samples in order to evaluate the quality of reconstructed speech signals better, enabling generative adversarial networks to achieve better results in speech enhancement tasks. In summary, the proposed method enhances the speech quality enhancement ability of adversarial generative networks in multiple aspects. Therefore, adopting an improved generative adversarial network can greatly improve the quality of single-channel speech and the comprehensibility of speech, enabling the recipient of speech to recognize speech information quickly.

2 Design of single-channel speech quality enhancement methods in mobile networks

Generative Adversarial Network (GAN) is a generative model [20,21] that can generate pure speech by learning a mapping from noisy speech samples to pure speech samples. It is capable of generating data with very high fidelity and has powerful learning and generative capabilities. It can also handle data with complex distributions and applies to a wide range of task scenarios. Therefore, this network is employed to realize the enhancement of single-channel speech signals in mobile networks. However, the original production adversarial network suffers from drawbacks such as instability and gradient vanishing during training and has poor generalization ability. Therefore, to improve the single-channel speech quality enhancement effect, the improvement processing of the production adversarial network is initiated. To carry out effective and targeted improvement of the production adversarial network, the principle of its realization of single-channel speech signal enhancement is first elaborated, and the basic framework of the production adversarial network is understood to provide a clear direction and goal for the subsequent improvement.

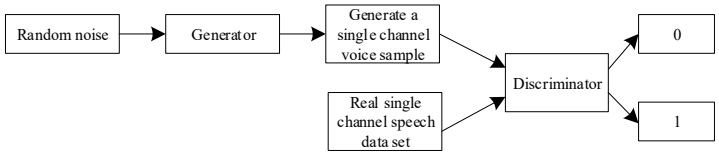

Fig.1 Generates the adversarial network structure

2.1 Principle of single-channel speech signal enhancement for generative adversarial networks

The basic idea of GAN is similar to building a game between two players: generator G and discriminator D [22]. The generator G models the single-channel speech distribution and generates a single-channel speech sample, which is usually low-dimensional random noise, and Discriminator D (usually a binary classifier) learns to classify from the training single-channel speech, distinguishing between the real single-channel speech or the fake single-channel speech generated by Generator G. During training, G aims to learn accurate mappings so that the generated single-channel speech can mimic the real number of single-channel speech well enough to fool D. On the other hand, D learns to better discriminate whether it is the real single-channel speech data or the fake single-channel speech data generated by G. The adversarial training allows the two models to improve each other's accuracy until the discriminator D cannot distinguish between the real sample speech and the generated speech until [23]. The structure of a conventional generative adversarial network is shown in Fig. 1.

In the backpropagation process. D backpropagates a batch of real single-channel speech, denoted as "1", then D backpropagates a batch of fake single-channel speech generated by G, denoted as "0", and finally, G backpropagates to make D classify incorrectly. However, the generative adversarial network suffers from unstable training and poor generalization ability in the single-channel speech signal enhancement task. Therefore, improvements need to be made to address these shortcomings to improve the performance of generative adversarial networks for better single-channel speech signal enhancement processing.

2.2 **Improved Mel frequency cepstral coefficients**

Among the many methods for transforming human speech signals into different speech feature parameters, the most common and widely used is the Mel-scale Frequency Cepstral Coefficients (MFCC). However, due to the nonlinear correspondence of Hz-Mel frequencies, the number of filters used in the low-frequency region is large and densely distributed, while the number of filters used in the middle and high-frequency regions is small and sparsely distributed. In this paper, MFCC, IMFCC, and MidMFCC are used to solve the problem of computational accuracy in the low, high, and medium frequency bands, respectively, and frequency masking filtering algorithms are added to reduce the effect of noise. The Meier frequency cepstral coefficients (MFCC) are extracted from the speech signal in the following steps:

(1) Pre-enhancement

Passes the speech signal $s(n)$ through a high-pass filter:

$$H(z)=1-a\times z^{-1}\ (0.1\le a\le 0.9\ ) \tag{1}$$

The time-domain expression of the pre-enhanced signal is given by

$$s_1(n)=s(n)-a\times s(n-1) \tag{2}$$

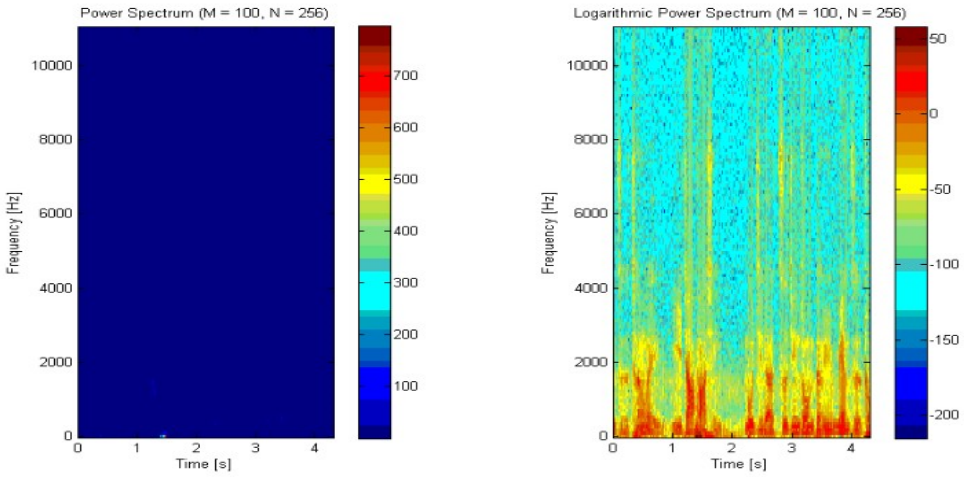

Fig.2 (for M=100, N=256) energy spectrum and logarithmic energy spectrum

(2) Frame

The N sampling points are assembled into one observation unit called a tone frame (the value of N is taken as 256 in this paper). To avoid too much variation between two neighboring frames and get smoother short-time speech features and spectral sequences, in this paper, we let there be a section of the overlapping region between two neighboring frames, and this overlapping region contains M sampling points (the value of M is about 1/2 of N).

(3) Hamming Window

Multiply each tone frame by the Hamming window, assuming that the signal of the tone framing is $s(n)$, n=0,1, ..., N-1. then the Hamming window after multiplication is:

$$x(n) = s(n) \times W(n) \tag{3}$$

$W(n)$ the form is as follows:

$$W(n,\alpha) = (1-\alpha) - \alpha\cos(2\pi n/(N-1)), 0 \le n \le N-1 \tag{4}$$

Different values produce different Hamming windows, and this system takes the $\alpha = 0.46$

(4) Fast Fourier Transform

The Fast Fourier Transform is performed on the $x(n)$ signal and the windowed frames are subjected to FFT to find the spectral parameters of each frame.

(5) Frequency mask filtering

Filtering a signal through a frequency masking filter. Research on the mechanism of human hearing has found that when two tones of similar frequency are emitted at the same time, only one tone can be heard. Critical bandwidth refers to the boundary of the bandwidth where the subjective perception changes abruptly. When the frequency of the two tones is less than the critical bandwidth, people hear the two tones as one, which is the masking effect. The Frequency Masking Filter (FMF) algorithm employs a nonlinear bi-directional filter to simulate the masking function of the human ear, which can effectively improve the robustness of the system. The filter model is simplified into a triangular filter, and the linear relationship between the slope, frequency, and spectrum of the triangle is shown in Fig. 3 below:

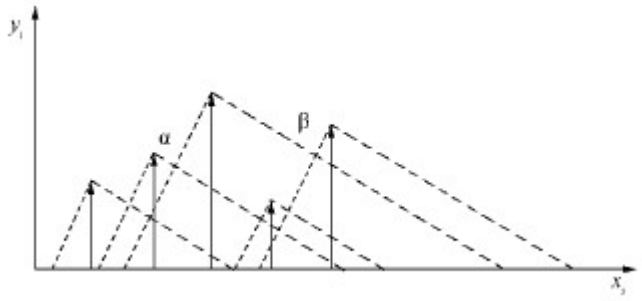

Fig. 3 Frequency masking model is schematic

The specific steps of the frequency masking algorithm are as follows: first, calculate the power spectrum of the corresponding frequency of the speech frame x_i ; then, filter the power spectrum using the following algorithm. First, the spectrum of the original speech signal is calculated, and the results are sorted to obtain the power spectrum sequence

$$\begin{aligned} y_{i-1} &= \alpha y_i \\ y_{i-1} &= y_{i-1} \quad (y_{i-1} > x_{i-1}) \\ y_{i-1} &= x_{i-1} \quad (y_{i-1} \le x_{i-1}) \end{aligned} \tag{5}$$

Where x_i is the power spectrum of the original signal at frequency index $i\ (0 \le i \le N)$; y_i is the filtered power spectrum and α is the low-frequency mask value. The initial condition of this equation is $y_N = x_N$, and the execution direction is from high to low-frequency index i .

$$\begin{aligned} y_i &= \beta y_{i-1} \\ y_i &= y_i \quad (y_i > x_i) \\ y_i &= x_i \quad (y_i \le x_i) \end{aligned} \tag{6}$$

Where: β is the high-frequency masking threshold. The initial condition for the execution of this equation is $y_0 = x_0$ and the frequency index of the execution direction i is from low to high.

(6) Triangle Filter

The spectral energy is multiplied by a set of 20 triangular bandpass filters (1st-7th order low-frequency MFCC, 8th-13th order MidMFCC,14th-20th order IMFCC) to find the logarithmic energy of the output of each filter. The frequencies of the two base points of the triangle of each filter are each equal to the center frequencies of the two adjacent filters, i.e., the transition bands of each two adjacent filters lap each other. The Hz-Mel frequency correspondence is shown below:

$$f_{MFCC} = 2595 \times \log_{10}(1 + f / 700)$$

$$f_{IMFCC} = 2146.1 - 1127 \times \ln\left(1 + \frac{4000 - f}{700}\right) \tag{7}$$

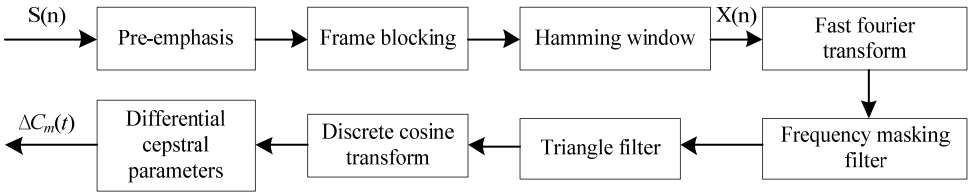

Fig. 4 Improved MFCC flow

$$f_{MIDMFCC}=\begin{cases}1073.05-527\times\ln\left(1+\dfrac{2000-f}{300}\right), 0<f\le 2000\\ 1073.05+527\times\ln\left(1+\dfrac{f-2000}{300}\right), 2000<f\le 4000\end{cases}$$

(7) Discrete cosine conversion

The 20 logarithmic energies $E_k(k=0,2,...,19)$ mentioned above are brought into the discrete cosine conversion formula to find the Mel-scale Cepstral parameter of L order. Here L is usually taken to be 12 and N to be 20. The discrete cosine transformation formula is as follows:

$$y(k)=\alpha(k)\sum_{n=0}^{N-1}x(n)\cos(\frac{\pi(2n+1)k}{2N}), k=0,1,...,L-1 \quad (8)$$

included among these $\alpha(k)=\begin{cases}\sqrt{\dfrac{1}{N}} & k=0\\ \sqrt{\dfrac{2}{N}} & k\neq 0\end{cases}$

(8) Delta cepstral

Although the 12 feature parameters have been derived, however, in practical applications for voiceprint recognition, the different cepstral parameter is usually added selectively to show the change of the cepstral parameter concerning time. Its meaning is the rate of change of the cepstral parameter for time, that is, it represents the dynamic change of the cepstral parameter in time, and the formula is as follows.

$$\Delta C_m(t)=\left[\sum_{j-M}^{M}C_m(t+\tau)\right]/\left[\sum_{j-M}^{M}\tau^2\right] \quad (9)$$

At this point, the MFCC feature extraction is finished. From its overall framework, it is essentially a short-time Fourier transform of the speech signal, followed by bandpass filtering of the energy spectrum with a filter bank, and finally, cepstral calculation. In the actual application process, feature vectors with different dimensions can be selected according to the needs and experimental tests.

The improved MFCC flowchart is shown in Fig. 4.

2.3 Improvements in generative adversarial networks

Due to the shortcomings of the original production adversarial network such as instability and gradient vanishing during training, it leads to the poor quality of the generated speech. To address the above problems, this paper proposes a Relativistic Average GAN with Mixed Penalty (RaGAN-MP).RaGAN can effectively enhance the stability of GAN model training and alleviate the gradient

vanishing of the GAN model; introduce the Relativistic average Least Squares loss (Relativistic average LSGAN, RaLSGAN) to replace the cross-entropy adversarial loss of traditional GAN to optimize the generator and discriminator, which can accelerate the network convergence. The hybrid penalty term, consisting of the L1 regularization and the mean-square error term, can more accurately measure the distance between the generated and real speech. Minimizing the value of the hybrid penalty term can improve the quality and intelligibility of the generated speech. The improved single-channel speech quality enhancement processing generative adversarial network model is shown in Fig. 5.

(1) Loss function improvement

Since the training of the original GAN is a process of a very large and very small game, the G network and the D network run alternately and finally reach a state of Nash equilibrium, and this very large and very small game is mainly realized by the loss function. The loss functions of the D network and G network of the original GAN are shown below, respectively:

$$L(D)=E_{x\sim p(x)}\left[\log D(x)\right]+E_{z\sim p(G)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (10)$$

$$L(G)=E_{z\sim p(G)}\left[\log\left(1-D\left(G(z)\right)\right)\right]+L_b \quad (11)$$

In the formula, $E_{x\sim p(x)}$ and $E_{z\sim p(z)}$ denote the corresponding network expectation; p_x denotes the probability distribution of the real single-channel speech data; p_G denotes the probability distribution of the generated single-channel speech data; $D(x)$ and $D\left(G(Z)\right)$ are the outputs of the real single-channel speech data and the generated single-channel speech data through the discriminator, respectively; and L_b is the penalty term. The D-network estimates the probability of the input single-channel speech data to be the real one, and the G-network will increase the probability of generating the single-channel speech data through training. Samples $G(z)$ as the probability of true single-channel speech. However, in the training process of most GANs, as the probability of the generated data being real gradually increases, the probability of the real single-channel speech data x being real never decreases, and it is always 1. The real training situation of the antigenerative network and the ideal training situation are shown in Fig. 6.

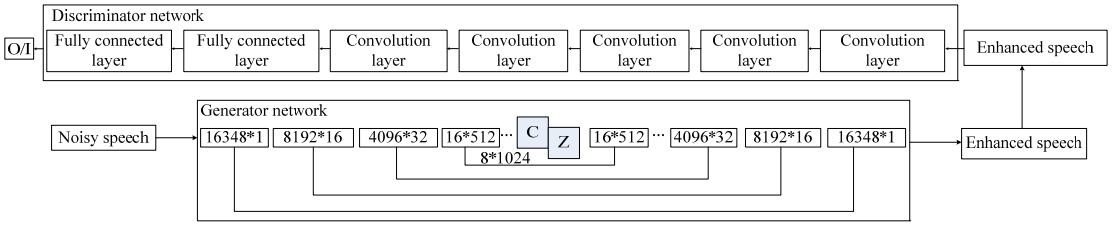

Fig.5 Improved generative adversarial network

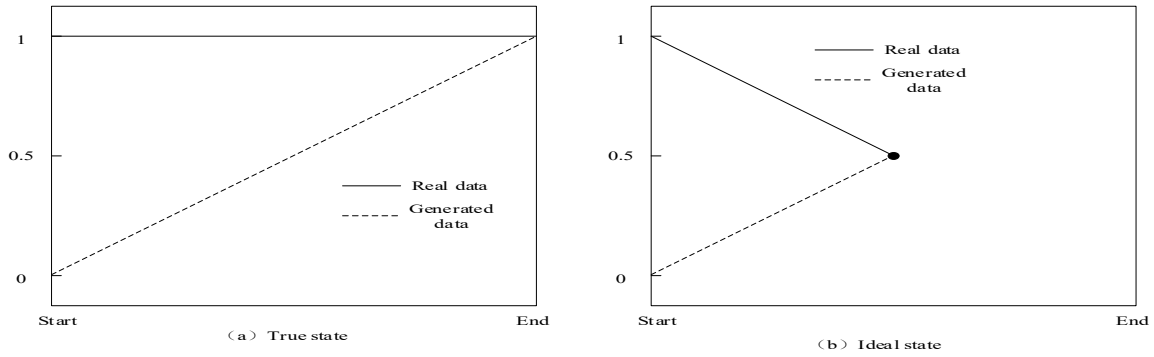

Fig.6 Training state of the adversarial generation network

Ideally, the GAN reaches an equilibrium state through training, at which time both the real single-channel speech data x and the generated single-channel speech data $G(z)$ are input to the D network, and the output single-channel speech discrimination probability value is 0.5, which indicates that the G-generated single-channel speech data is enough to be fake. However, this is not the case for GAN training in practice. As D and G keep training iteratively, the single-channel speech generation sample $G(z)$ is more and more similar to the real single-channel speech sample x, and the single-channel speech discrimination probability value given by D is also increasing. However, the probability score of the real data x is always the same, which is determined by the loss function of GAN. In this case, training with fixed labels makes D saturate very easily, resulting in the first term in Eq. (1) $L(D)$ no longer changing, while the second term cannot fall. So much so that when the second term has no room to improve the training of the model, it still focuses on that term, but ignores the role of the first term, which affects the validity of the JS (Jensen-Shannon distance) distance calculation and reduces the stability of model training. In addition, theoretically, after the convergence of the GAN model, the probability scores of the single-channel speech-generated data $G(z)$ and the single-channel speech-true data x are the same, which is inconsistent with the a priori knowledge. For this reason, this paper introduces Relativistic average least squares loss (Relativistic average LSGAN, RaLSGAN) to replace the traditional cross-entropy adversarial loss to optimize the generator and discriminator, which can accelerate the network convergence, then after the improvement of the above Equation (1) and Equation (2), the training model of the more stable GAN is more stable for the RaLS adversarial loss of the generator G function, is expressed as:

$$\begin{aligned} L_{RaLs}(G) = & E_{x\sim p_x(x)}\left[\left(D\left(G(x)\right)-E_{z\sim p_z(z)}\left(D\left(G(z)\right)\right)-1\right)^2\right] \\ & +\mathrm{E}_{z\sim p_z(z)}\left[\left(D(z)-E_{x\sim p_x(x)}\left(D\left(G(z)\right)\right)+1\right)^2\right]+L_b \end{aligned} \tag{12}$$

The loss function of the discriminator D is:

$$L_{RaLs}(D)=E_{z\sim p_z(z)}\left[\left(D\left(G(z)\right)-E_{x\sim p_x(x)}\left(D\left(G(x)\right)\right)-1\right)^2\right] \\ +E_{x\sim p_x(x)}\left[\left(D(x)-E_{z\sim p_z(z)}\left(D\left(G(z)\right)\right)+1\right)^2\right] \quad (13)$$

(2) Mixed penalties

For the optimization of generator G, Baby D et al. [24] achieved optimization by adding gradient penalty terms, which promoted the smoothness and stability of the network learning process. Since this minimax game training method of GAN is difficult to achieve the true Nash equilibrium state in practice, it also means that even if the model training stops, the single-channel speech data generated by G will be different from the real speech data to varying degrees [25]. However, the optimization achieved by adding gradient penalty term in the existing studies only considers the smoothness and stability of the network learning process, which cannot overcome the above problems in practical application, resulting in poor generalization performance and accuracy. Therefore, on the basis of ensuring the stability of the generator G, the sparsity and prediction accuracy are comprehensively considered, and regularization terms are added to the training process of the model to reduce the number of parameters and computational complexity, so as to ensure real-time performance, further help the generator G filter noise and redundant features, and improve the generalization ability of the model. Moreover, the improved MFCC is received as input to learn from the features and structure of the original speech, and retain the prosodical, intonation and spectral characteristics of the enhanced speech as much as possible when generating the enhanced speech, and the mean square error is added to maximize the similarity between the generated speech signal and the clear speech, helping the network to learn more accurate fitting of the target value. The result is as close to the real target as possible to improve the quality of single-channel speech generation. Therefore, a Mixed Penalty consisting of L1 regularization and Mean Square Error (MSE) is proposed to optimize the training of the G network. L_1 and L_{MSE} are two different distance metric terms defined as:

$$L_1=\left\|G\left(z,x_c\right)-x\right\|_1 \quad (14)$$

$$L_{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left(G^i\left(z,x_c\right)-x^{(i)}\right)^2 \quad (15)$$

where x_c is the condition variable and N is the single-channel speech real data. The combination of these two terms instead of the penalty term L_b in G can more accurately measure the distance between the generated speech samples and the real speech samples, and guide G to generate speech data in a more realistic direction, to achieve the purpose of improving the quality of the generated speech samples by minimizing the difference between the two data distributions. When the model is being trained, L_1 and L_{MSE} calculate the expected value of the distance between generated samples and real samples in each batch, is used as a penalty term to guide the training of the generator.

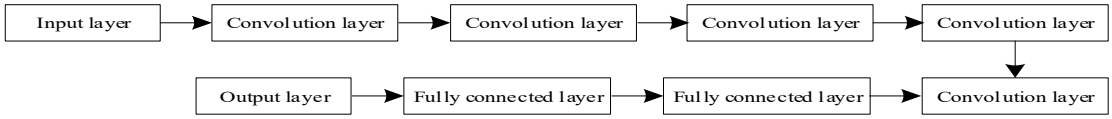

Fig.7 Improved discriminator network structure

(3) Discriminator improvement

In this paper, a generator is used to reconstruct a single-channel speech signal, and a discriminator is required to judge whether the reconstructed single-channel speech signal conforms to the conditional information. The goal of the discriminator is to judge the truth and falsity of the input single-channel speech signal. For the optimization of discriminator D, Pascual et al. [26] completed the optimization of the discriminator through end-to-end training and relative average discriminator, taking into account the pattern collapse in the training process to improve the balance between the generated network and the discriminator network. However, it is difficult to capture the global characteristics of speech signals in practical applications, which leads to the accuracy of discriminator evaluation is not ideal. Therefore, to solve this problem, this paper adds a fully connected layer on the basis of the discriminator, which can convert the output of the convolutional layer into global features [27], to provide more nonlinear expression capabilities and decision boundaries. Based on improving the stability of network training and reducing the competition and imbalance between the generated network and the discriminant network. The discriminator's ability to distinguish between real and generated samples is also improved so that the quality of reconstructed speech signals can be better evaluated, and the generated adversarial network can achieve better results in speech enhancement tasks. So, a fully connected layer is added on top of the generative adversarial network in this paper, and the structure is shown in Fig. 7.

The improved generator halves the dimensions of the single-channel speech data layer by layer by layer convolution, maps the speech signal feature dimensions to 1*8 dimensions as conditional information input to the generator structure through a first fully connected layer, and maps a 1*1 dimensional value as an output through a second fully connected layer, which is used by the discriminator as a basis for evaluating the quality of the reconstructed speech signal. As a result, based on the above optimization, the performance of the generator and the discriminator is effectively improved to form a relative average generative adversarial network (RaGAN) based on the hybrid penalty term; using the improved generative adversarial network, the generator generates samples that are indistinguishable from real speech signals based on the input conditional information, and the discriminator determines whether the input samples are real or not, and evaluates the quality of the reconstructed speech signal based on the conditional information, and the two work in concert until the discriminator assesses the quality of the reconstructed speech signal based on it. Quality and the two work together until the discriminator cannot distinguish between the real sample speech and the generated speech, completing the enhancement processing of single-channel speech signals.

3 Experimental analyses

3.1 Experimental Objects

The speech datasets used for the experiments in this chapter are Valentini 2016 and Valentini 2017, both of which are selected from the publicly available corpus in the mobile network at the Center for Speech Technology Research (CSTR) of the University of Edinburgh’s contains many different types of CSTR contains many different types of speech datasets, which are widely used in

the field of speech signal processing. The Valentini series of datasets selected for the experiments in this paper are mainly oriented to the study of speech enhancement and speech recognition in the context of noise. The noise data is selected from the DEMAND library, which contains a large number of single-channel noise data, and the single-channel speech quality enhancement experiment is completed by adding noise. To ensure that the speech quality enhancement method in this paper has some significance, the enhanced speech quality is scored in two ways, which are divided into two evaluation criteria: subjective and objective.

Table 1 MOS scoring criteria

MOS score	Quality grade	Distortion degree level
5	Optimal	Imperceptible
4	Good	Have the slightest inkling
3	Intermediate	Perceptible, slightly annoying
2	Poor	The perception is obvious.
1	Range	Can't stand

Subjective Evaluation Criteria: The subjective evaluation criterion of speech enhancement is to let the testers directly evaluate the enhanced speech subjectively, reflecting a subjective impression of the quality of the enhanced speech. The basic idea is to let the testers grade the enhanced speech by comparing and listening to the original and enhanced speech according to a pre-agreed scale. Currently, the most commonly used subjective judgment criterion for speech enhancement is the Mean Opinion Score (MOS), and the scoring criteria of the MOS method are shown in Table 1.

The MOS uses a 5-level scoring scale, and the specific scoring rules are shown in Table 1. The MOS evaluation process is divided into two phases: training and evaluation. In the training phase, the participants receive different levels of signals as a reference, and in the evaluation phase, the test speech is rated according to the scoring criteria. The participants need to be in the same environment, and the test results are weighted and averaged, and the final value is the MOS score of the tested voice, and the weighted average formula is:

$$MOS = \frac{1}{M}\sum_{k=1}^{5} S_k \quad (16)$$

where M denotes the number of testers, k is the score, and S_k denotes the total number of ratings given to the speech.

To keep the MOS scores less fluctuating and not affect the pair scores because of the personal preference of the listener, it is necessary to use as many speech professionals as possible, and the test environment is kept the same. This judging method is simple and direct, and the evaluation results are very reliable, but it also has high requirements for the environment and personnel, and the repeatability is very poor, so there is a need to have objective judging criteria for correlation to evaluate the performance of speech enhancement methods.

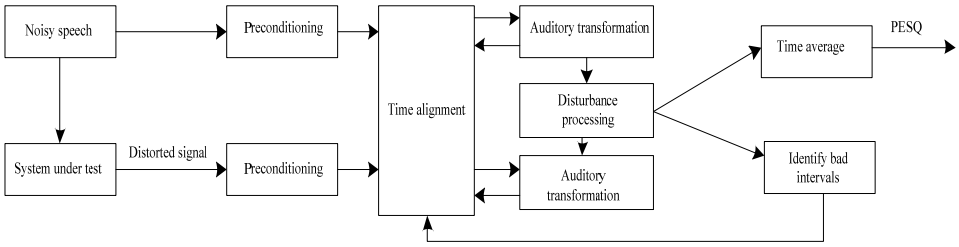

Fig.8 PESQ scoring process

Objective Evaluation Criteria: Objective evaluation criteria for speech enhancement is the most basic and accurate objective judgment of speech quality, which is not influenced by personal preference and can be repeated. Commonly used methods include SNR, Segmental SNR, Log Likelihood Ratio, Perceptual Evaluation of Speech Quality (PESQ), and so on. In this paper, the Log Likelihood Ratio and PESQ are chosen to evaluate the enhanced single-channel speech quality. The logarithmic spectral measure is a common method that can be used to measure the effectiveness of single-channel speech signal enhancement. It maps the original spectrum onto a logarithmic scale by taking the logarithm of the spectrum of the speech signal, which can better reflect the differences and variations in the signal. Specifically, the effectiveness of the proposed method is assessed by comparing the original speech signal with the processed speech signal and calculating the logarithmic spectral measure value between them. The smaller the value, the more similar the original signal is to the processed signal, i.e., the more effective the enhancement is. The log-spectral measure is calculated by the formula:

$$LSD(x,y)=\frac{10\lg\left(\sum_{i=1}^{M}\min(x_i,y_i)^2\right)}{\sum_{i=1}^{M}\left(x_i^2\right)+\sum_{i=1}^{M}\left(y_i^2\right)-2\sum_{i=1}^{M}\min(x_i,y_i)^2} \quad (17)$$

Where x、y denotes the spectra of the original speech signal and the processed speech signal, respectively, and M is the number of signals.

The Perceived Evaluation of Speech Quality PESQ is a new objective measure that has relatively high reliability under a variety of codecs and network conditions and is suitable for most environments. The computational flow of the PESQ measure is shown in Fig. 8.

The test signal and the pure signal were first level-adjusted by the system under test, then time-aligned to correct for the effect of delay, and then calculated to obtain a loudness spectrum after auditory transformation. The difference in the loudness spectra is utilized to find the PESQ score of the test speech with the formula:

$$PESQ = a + bd_{mpv} + cd_{ampv} \quad (18)$$

Where a is the intercept term, b is the symmetric perturbation coefficient, c is the asymmetric perturbation coefficient, d_{mpv} is the symmetric perturbation value, d_{ampv} is the asymmetric perturbation value, b and c are used for adjusting the perturbation value positively or negatively to change the PESQ score.

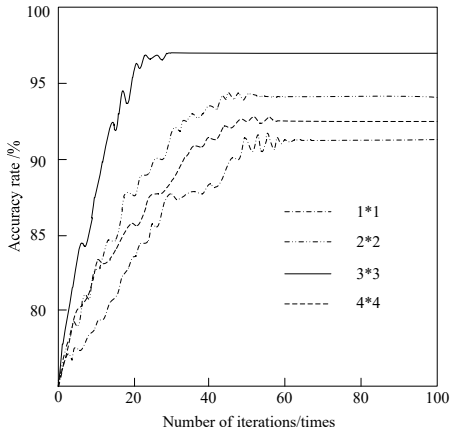

Fig.9 Training accuracy of convolution kernel of different sizes

3.2 Experimental results and analysis

To ensure the single-channel speech quality enhancement effect of the generative adversarial network, this paper needs to train the convolutional layer in the discriminator, choose the appropriate convolutional kernel size to ensure that there are the best test results, based on the 1*1 convolutional kernel and gradually increase the size of the convolutional kernel until it is increased to 4*4, the accuracy of different convolutional kernel sizes under the discriminator is shown in Fig. 9.

The effect of convolution kernel size on the performance of the generative adversarial network can be seen by looking at Fig. 9. When the convolution kernel is gradually increased from 1*1, the accuracy shows an increasing trend, which indicates the effectiveness of the convolution kernel in extracting features. When the convolutional kernel size is 3*3, the accuracy reaches a peak of about 97%, which indicates that a convolutional kernel size of 3*3 can best capture and recognize speech features in this speech quality enhancement processing method. However, as the convolutional kernel is further increased to 4*4, the accuracy shows a significant drop. This phenomenon may be because too large a convolutional kernel introduces too much noise and irrelevant information, which interferes with the judgment of the discriminator. Therefore, it can be concluded that a convolutional kernel size of 3*3 is optimal for the discriminator of this generative adversarial network.

To verify the effectiveness of the proposed method, ablation experiments were conducted on the loss function, generator G, and discriminator D before and after improvement. Three indicators, namely entropy value, distance coefficient, and discriminant accuracy, were selected to test the effectiveness of the loss function, generator G, and discriminator D before and after improvement. Among them, the higher the entropy value and discriminant accuracy indicators, the better. The distance coefficient measures the distance between the generated speech sample and the real speech sample. The smaller the distance between the two, the closer the generated speech sample is to the real speech sample, So the smaller the value of this indicator, the better. At the same time, the expected baseline values for each indicator were set separately. The test results are shown in Table 2.

Table 2 Test results before and after optimization

Loss function before and after optimization	Entropy value	Expected baseline value
Relative mean least squares loss	0.96	0.90
Cross entropy versus loss	0.91	
Generator G before and after optimization	Distance coefficient	Expected baseline value
Original generator G	3.5	1.0
L1 regularization and mean square error mixed penalty term	0.24	
Discriminator D before and after optimization	Discriminant accuracy	Expected baseline value
Primary discriminator D	91%	95%
Add a fully connected layer	98%	

According to the analysis of Table 2, it can be seen that the optimized loss function, generator G, and discriminator D of the proposed method have significantly improved indicator results compared to the pre optimized loss function, generator G, and discriminator D, and all are higher than the set expected baseline values, indicating the effectiveness of its optimization. This is because the proposed method uses L1 regularization and a mixed penalty term of mean square error to optimize generator G. While ensuring the stability of generator G, it comprehensively considers sparsity and prediction accuracy. Through L1 regularization, the sparsity of network parameters is promoted, improving the network's generalization ability and interpretability. The mean square error focuses on the square difference between the predicted value and the target value, in order to help the network learn to fit the target value more accurately and make its predicted results as close as possible to the real target. And the proposed method optimizes discriminator D by adding a fully connected layer, fully considering the processing of local features by convolutional layers, which cannot fully capture the global features of speech signals. By adding fully connected layers, the output of the convolutional layer is transformed into global features to provide more nonlinear expression ability and decision boundaries. On the basis of improving the stability of network training, reducing competition and imbalance between generative and discriminative networks, the discriminator's ability to distinguish between real and generated samples is also improved. Therefore, the proposed method further improves the performance of the generator and discriminator and will have a better effect on achieving single-channel speech quality enhancement.

In order to verify the speech quality enhancement effect of the proposed method, a piece of clean speech is selected and environmental noise of home, office, street, factory, shopping mall, agriculture, construction site, nature and other types are added to it respectively. The spectrogram of clean speech is shown in Figure 10.

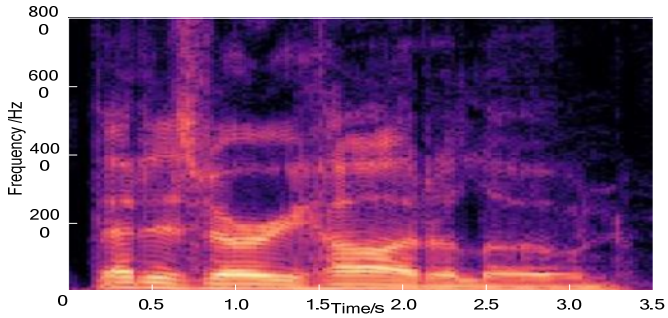

Fig.10 Clean speech spectrogram

The before and after results of speech signal enhancement through the adversarial network are shown in Fig. 11 below:

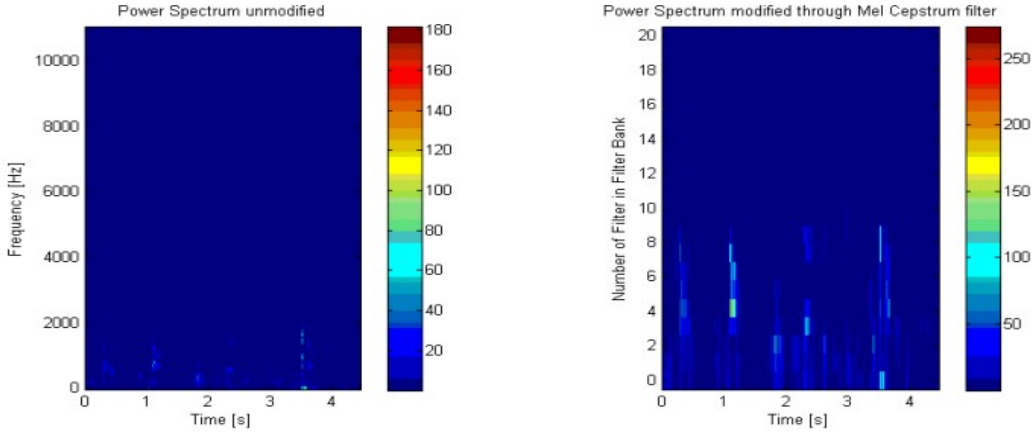

Fig. 11 Power spectra before and after speech signal enhancement

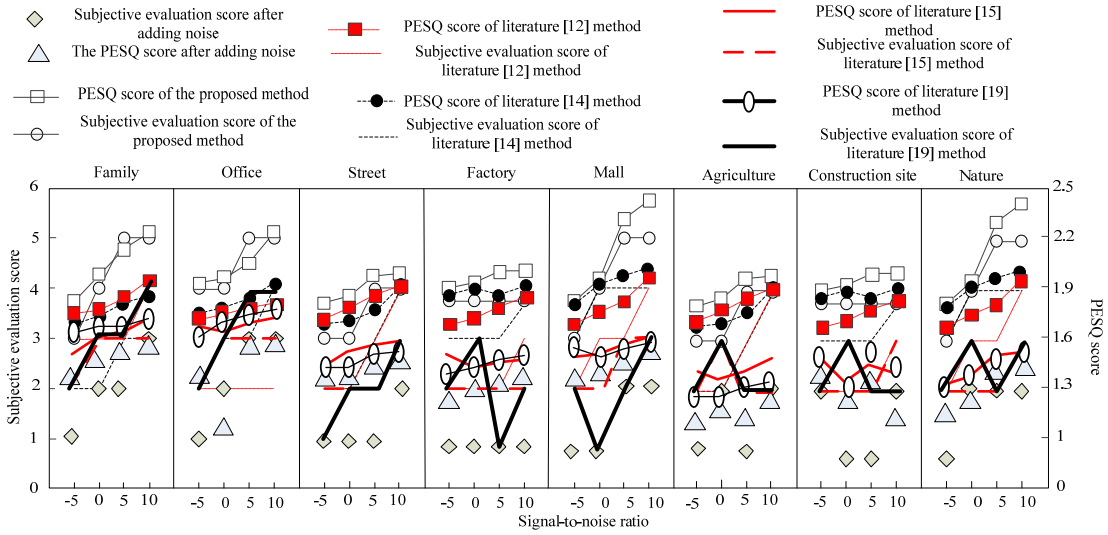

Fig. 12 Subjective evaluation scores as well as PESQ scores

The comparison shows that the method of this paper can accomplish the enhancement of speech signal and overcome the related interference.

The PESQ scores and subjective evaluation scores after adding noise, as well as the speech enhancement using the method in this paper, the method [12], the method [14], the method [15] and the method [19], the subjective evaluation scores and PESQ scores are shown in Figure 12.

From the results obtained in Fig. 12, it can be clearly seen that when environmental noise is added to clean speech, both subjective and objective evaluation get very low scores, which proves

that these five kinds of environmental noise are unbearable. Four kinds of literature are used to enhance the speech with added environmental noise, which can enhance the speech to a certain extent, making it easier to tolerate and distinguish compared with the original speech full of noise, but the performance is different in different environments. The literature [12] method performed moderately in the home and street, while the literature [14] method performed better in the factory and shopping mall environment, and the literature [15] method and the literature [19] method both performed better in the home and office. However, through the scores of subjective evaluation and objective evaluation, it can be seen that the enhancement effect of the four literature methods is not ideal compared with the method in this paper, and there is an obvious gap between the scores. The single-channel speech quality enhancement method proposed in this paper has achieved remarkable effect in reducing speech noise. The subjective evaluation results show that it has a high score, can effectively improve the speech quality, make the noise almost imperceptible, and can better improve the clarity of the speech content. At the same time, the PESQ score was improved compared with the two literature methods, which further verified the effectiveness of the proposed method. This shows that the proposed method has significant advantages in the aspect of single-channel speech quality enhancement. It can reduce noise more effectively, improve the clarity and intelligibility of speech, and provide users with a better voice experience.

On the basis of the above tests, in order to further verify the enhancement effect of the proposed method, 100 test signal samples were randomly selected in the Valentini series data set, which were integrated into the environmental noise of homes, offices, streets, factories, shopping malls, agriculture, construction sites, nature and other types. Each sample contained 100 signal data, which were taken as test objects. In order to reflect the superiority of the proposed method, the method [12], the method [14], the method [15] and the method [19] are compared with the proposed method, and the logarithmic spectral measure of the five methods is calculated using formula (17), and the results obtained by the three methods are specifically shown in Figure 13 below.

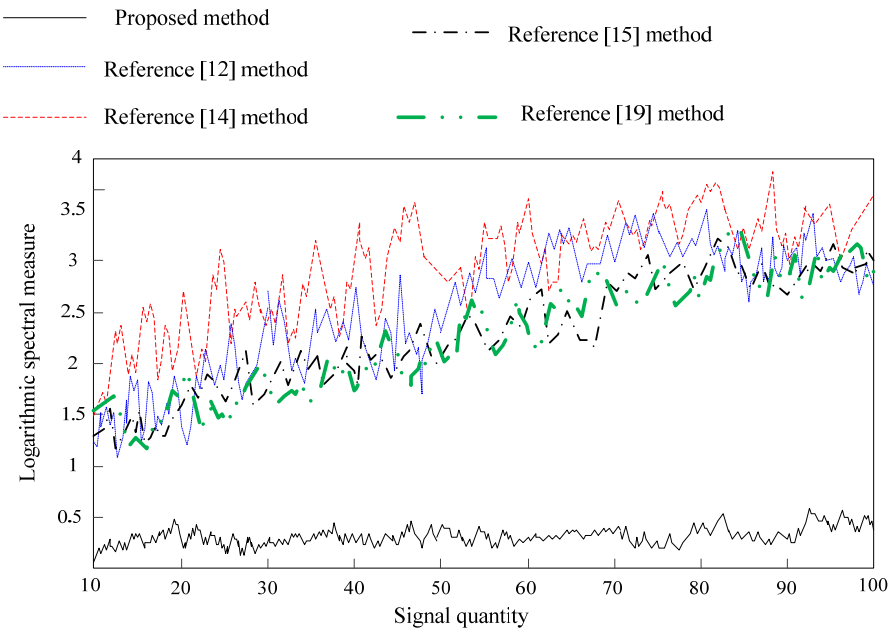

Fig.13 Comparison of logarithmic spectral measurement results

According to the results obtained in FIG. 13, it can be seen that among the five methods, the log-spectrum measure of the speech signal processed by the proposed method is the smallest, which is maintained below 0.5, while the log-spectrum measure of the speech signal processed by the reference [12] method, the reference [14] method, the reference [15] method and the reference [19] method are all above 1, and the number of processed signals increases. There is an upward trend. By comparing the log-spectrum measurement results obtained by the three methods, it can be concluded that after the speech signal enhancement processing by the proposed method, it is more similar to the original signal, and the better the enhancement effect is, the better it is, it can effectively improve the clarity of speech, enable listeners to hear the speech content more clearly, reduce the possibility of mishearing, and improve the quality and efficiency of voice communication.

4 Conclusion

To effectively improve the quality of single-channel speech, a single-channel speech quality enhancement processing method in mobile networks based on generative adversarial networks is proposed. The improvement of the generative adversarial network is achieved by implementing the loss function, optimized generator, and discriminator from three aspects to form the relative average generative adversarial network (RaGAN) based on the hybrid penalty term to improve its performance and effectively complete the single-channel speech quality enhancement processing. Through experimental observation, it can be seen that the performance of this paper's method for speech quality enhancement in different environments has achieved significant improvement, and the obvious increase in its subjective evaluation score and objective evaluation score fully verifies the effectiveness of the method in reducing single-channel speech noise. The logarithmic spectral measure of the speech signal after processing with this paper's method is the smallest, which is maintained below 0.5, which further indicates that this paper's method can ensure that the user receiving speech can obtain the speech information more clearly and reduce or eliminate the noise interference in speech, which can provide a reliable and valuable reference for many fields such as speech communication and speech recognition.

In future, coarse-grained features fusion (fuzzy system and etc.) or deep feature fusion will be introduced intothe quality prediction and enhancement to reach a higher and reliable application in real world [28-30].

References

[1] Shah S A. A., Bais A., Alashaikh A., et al. (2023). Discrete wavelet transform based branched deep hybrid network for environmental noise classification. Computational Intelligence, 39(3):478- 498.

[2] Dwyer, Robert T., Kessler D., et al. (2021). Contralateral Routing of Signal Yields Significant Speech in Noise Benefit for Unilateral Cochlear Implant Recipients. Journal of the American Academy of Audiology, 30(3):235-242.

[3] Zhang Y., Dong Z., Wang S., et al. (2020) Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation, Information Fusion, 64: 149-187.

[4] Jassim, W. A., & Harte, N. (2022). Comparison of discrete transforms for deep-neural-networks-based speech enhancement. IET Signal Process., 16(4), 438-448.

[5] Li Y., Zhang X. & Sun M. (2023). A unified speech enhancement approach to mitigate both background noises and adversarial perturbations. Information Fusion, 95(4):372-383.

[6] Ranjbaryan, R., & Abutalebi, H. R. (2021). Multiframe maximum a posteriori estimators for single-microphone speech enhancement. IET Signal Processing, 15(7), 467-481.

[7] Malek, J., & Bohac, M. (2020). Block-online multi-channel speech enhancement using deep neural network-supported relative transfer function estimates. IET Signal Processing, 14(3), 124-133.

[8] Roy, S. K., & Paliwal, K. K. (2022). Robustness and sensitivity metrics-based tuning of the augmented Kalman filter for single-channel speech enhancement. Applied Acoustics, 185(1), 108335.

[9] Sivapatham, S., Kar, A., & Christensen, M. G. (2022). Gammatone filter bank-deep neural network-based monaural speech enhancement for unseen conditions. Applied Acoustics, 194(6), 108784.

[10] Shi S., Paliwal K. & Busch A. (2023) On DCT-based MMSE estimation of short time spectral amplitude for single-channel speech enhancement. Applied Acoustics, 202(1):1-23.

[11] Kajla, P., & George, N. V. (2020). Speech quality enhancement using a two channel sparse adaptive filtering approach. Applied Acoustics, 158(1), 107035.1-107035.6.

[12] Garg, A., & Sahu, O. P. (2021). Deep convolutional neural network-based speech signal enhancement using extensive speech features. International Journal of Computational Methods, 19(8), 2142005.

[13] Saleem, N., Gao, J., Khattak, M. I., et al. (2022). DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition. Knowledge-Based Systems. 238(28), 107914.

[14] Valentini-Botinhao C., Wang X., Takaki S., et al. (2016). Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. ISCA-SSW, 146–152. DOI: 10.21437/ssw.2016-24.

[15] Fu S W., Yu C., Hsieh T A., et al. (2021). MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. Cornell University - arXiv,Cornell University - arXiv..DOI: 10.48550/arxiv.2104.03538.

[16] Wang H., & Wang D. (2022). Neural Cascade Architecture With Triple-Domain Loss for Speech Enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 734–743.

[17] Fan C., Zhang H., Li A., et al. (2023). CompNet: Complementary network for single-channel speech enhancement[J]. Neural Networks, 168: 508-517.

[18] Li A., Yu G., Zheng C., et al. (2023). A General Unfolding Speech Enhancement Method Motivated by Taylor's Theorem. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:3629-3646.

[19] Yu J., Chen H., Luo Y., et al. (2023). TSpeech-AI System Description to the 5th Deep Noise Suppression (DNS) Challenge. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/icassp49357.2023.10097210.

[20] Ruika, M., Voloin, M., Gazda, J., et al. (2022). Fast and computationally efficient generative adversarial network algorithm for unmanned aerial vehicle-based network coverage optimization. International Journal of Distributed Sensor Networks, 18(3), 3417-3442.

[21] Huang S., Fu W., Zhang Z., et al. (2024) Global-local fusion based on adversarial sample generation for image-text matching, Information Fusion, 103: 102084

[22] Li Y., Sun M. & Zhang X. (2022). Perception-guided generative adversarial network for

end-to-end speech enhancement. Applied Soft Computing, 29 (7),73504.1-73504.9.

[23] Zhou L., Zhong Q., Wang T., et al. (2021). Speech Enhancement via Residual Dense Generative Adversarial Network. International Journal of Computer Systems Science & Engineering, 38(3):279-289.

[24] Baby D., & Verhulst S. (2019). Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/icassp.2019.8683799.

[25] Li Z., Dong Z., Wen-Hua Chen, et al. (2022). On the game-theoretic analysis of distributed generative adversarial networks. International Journal of Intelligent Systems, 37(1):516-534.

[26] Pascual S., Bonafonte A., & Serrà J. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. Interspeech 2017. DOI: 10.21437/interspeech.2017-1428.

[27] Wang S., Nayak D. R., Guttery D. S., et al. (2021) COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis, Information Fusion, 68: 131-148

[28] Liu S., Huang S., Xu X., et al. (2023) Efficient Visual Tracking Based on Fuzzy Inference for Intelligent Transportation Systems, IEEE Transactions on Intelligent Transportation Systems, 24(12): 15795-15806

[29] Liu S, Wang S, Liu X, et al. (2021) Fuzzy Detection aided Real-time and Robust Visual Tracking under Complex Environments. IEEE Transactions on Fuzzy Systems, 29(1), 90-102

[30] Wang S., Govindaraj V. V., Gorriz J. M., et al. (2021) Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Information Fusion, 67: 208-229