Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **5**, Number **3**, August **2020**

Speech Denoising Method Based on Improved Least Squares GAN

Yingli Wang

Department of Electronic Engineering
Heilongjiang University
No.74 Xuefu Road, Harbin, Heilongjiang, China
wangyingli@hlju.edu.cn

Honglei Wu

Department of Electronic Engineering
Heilongjiang University
No.74 Xuefu Road, Harbin, Heilongjiang, China
mwhlwhl@163.com

Hongbin Ma*

Department of Electronic Engineering
Heilongjiang University
No.74 Xuefu Road, Harbin, Heilongjiang, China
*Corresponding author: mahongbin@hlju.edu.cn

Qitao Ma

Faculty of Engineering
The Hong Kong Polytechnic University
11 Yucai Road, Hung Hom, Kowloon,Hong Kong,China
Jack_coldsweat@163.com

Qun Ding

Department of Electronic Engineering
Heilongjiang University
No.74 Xuefu Road, Harbin, Heilongjiang, China
qunding@aliyun.com

Received May 2020; revised July 2020

Abstract. *This paper presents a improved least-squares GAN speech denoising method, aiming at training problem of gradient disappeared, training is unable to continue, and to generate against network loss function was improved, using the least squares cross entropy loss function, and behind the loss function add a L1 norm, after pretreatment of speech signal as the input of the model, removing noise signal. By using this method to remove the speech signal containing airport noise, the SNR and PESQ values are improved to some extent, indicating that the speech denoising method based on improved least squares GAN has better denoising performance.*
Keywords: Generation antagonism network; Speech denoising; Least squares; The L1 norm

1. **Introduction.** Voice technology is the most effective way to communicate with people. In recent years, it has been widely used in communication, medical research and other important fields. However, in the complex acoustic system, everywhere is full of noise, which seriously affects the effect and performance of the speech processing system. Traditional speech denoising methods include spectral subtraction, wiener filtering, wavelet transform and MMSE[5, 6, 7]. With the rapid development of artificial intelligence technology, generated countermeasure network has been widely used in the field of image denoising[8].

Generate against network can overcome the shortage of traditional method of speech denoising, don't need noise signals are assumed to be gaussian distribution, to estimate the noise, and Santiago will generate 2017 against network (Generative Adversarial Networks, referred to as GAN) [1, 2, 3, 4, 5, 6, 7, 9] used in the field of speech denoising, gained better denoising effect. In 2019, some literatures proposed a speech denoising algorithm combining wiener filtering[7] and GAN, which preprocessed the speech signal with wiener filtering to provide more speech features for the generation of the counter network, and used GAN for speech denoising[3] in the frequency domain, which significantly improved the speech denoising effect. But in the first generation network, the generator and the loss function is cross entropy loss of discriminant function, when the real data sample and produce samples with tiny, even no overlap between the loss function of value is constant, so the network was trained can disappear so happen gradient, training cannot continue[2]. This paper optimizes the GAN model on the basis of generating the anti-network denoising algorithm, and improves the generation of the anti-network loss function. The least square cross entropy loss function[5] is used to drag the false samples far away from the decision boundary into the decision boundary, which effectively overcomes the problem of the initial generation of the anti-network training. After the least square loss function, an L1 norm[6, 7] is added to reduce the gap between the generated sample and the real sample, which makes the speech denoising effect more obvious and improves the quality of the generated speech signal.

2. Generate antagonistic network.

2.1. **The original generated antagonistic network.** GAN's idea is a two-person zero-sum game. The generation of antagonist network mainly consists of generator G and discriminator D. Its main purpose is to train the false data for many times[9], make it infinitely close to the real data, and generate the most realistic samples as possible. Its working principle is as follows: generator G generates false samples close to real data through feature analysis of real data, x represents pure speech signal, $G(z)$ represents speech signal containing noise distribution, discriminator D determines real data to be true ($D(x) \to 1$), and false data to be false ($D(G(z)) \to 0$). Through the feedback network to continuously optimize the generator G and discriminant D, until the discriminant device cannot judge the generator G D sample data is true or false, namely $D(x)$ tends to 1, $D(G)(z)$ tends to 1, to achieve a Nash equilibrium,[2] according to the network training was basically completed, the false data has infinite close to generate the data.

Discriminant model, the first is the intuitive, is a simple neural network model, enter a voice signal, the output is a probability value (probability is greater than 0.5 is true, the probability is less than 0.5 is false), followed by the generation model, the generation model is composed of the neural network, the input is a set of contains pure speech signal eigenvalue of the random number, the output is the voice signal after training. In the experiment, there are two data sets, one is the pure speech signal, the real data set, and the other is the false data set generated after the network training. The purpose of the

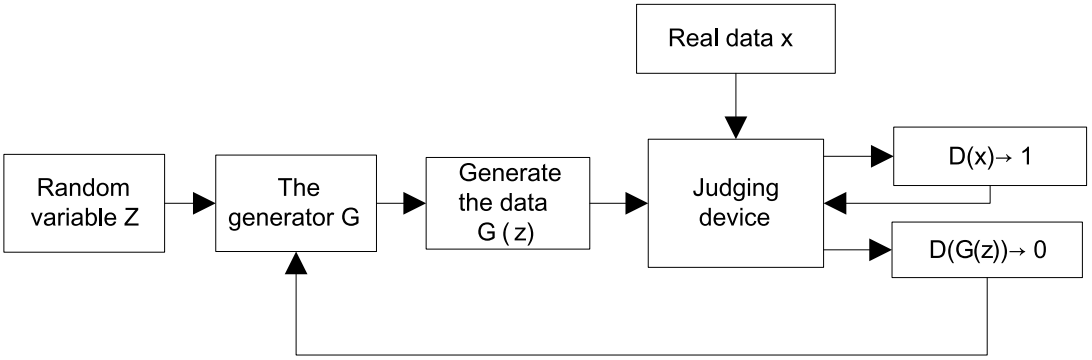

FIGURE 1. GAN model

discriminant network is to determine whether the input data set is from a real sample or a false sample. The purpose of the generation network is to generate samples as strong as possible to distinguish between true and false.

In the process of game, the big method to train the two models is: alternating iterative training. During each training, an important operation is not to make the parameters of the discrimination network D change, but to pass the error to the generated network G, and constantly update the parameters of the generated network, so as to achieve the optimization of the generated network G training, and then the same method to identify the optimization of network D.

The loss function of G optimization process is

$$\mathrm{L^D(D,G)} = \mathrm{E_{Z-P_Z(Z)}[\log(1-D(G(z)))]}. \tag{1}$$

The loss function of D optimization process is

$$\mathrm{L^D(D,G)} = -\frac{1}{2}\mathrm{E_{x-P_{\mathit{data}}(x)}}\left[\log \mathrm{D(x)}\right] - \frac{1}{2}\mathrm{E_{z-P_z(Z)}[\log(1-D(G(z)))]}. \tag{2}$$

Finally, the global optimization is carried out, and the loss function is

$$\min_G \max_D \{f(D,G)\} = E_{x-P_{\mathrm{data}}\,(x)}[\log D(x)] + E_{z-P_z(z)}[\log(1-D(G(z)))]. \tag{3}$$

In order to maximize the discriminant ability of the discriminant network, min is equivalent to maximizing the probability of the generated data as real data. When optimizing, 1 step is used to optimize G and K step is used to optimize D. Where, $P_{data}(x)$ is the distribution of pure speech signal data, and $P_z(z)$ is the data distribution of random noise speech signal.

2.2. **Conditional generation of antagonistic network.** The general training method of game algorithm for generating network does not need to assume in advance, but for some samples with a large data set, or for data with a large dimension, it is difficult to train a stable generator because of the large search sample range of its real data. In order to solve these problems, a conditional generation antagonism network (CGAN)[3] is proposed. In other words, some additional conditional information y is added into the GAN model to generate a sample data set conforming to this condition. The model figure of conditional generation antagonistic network is shown in FIG. 2, and its cross entropy function is:

$$\min_G \max_D \{f(D,G)\} = E_{x-P_{\mathrm{data}}\,(x|y)}[\log D(x|y)] + E_{z-P_z(Z)}[\log(1-D(G(z|y)))]. \tag{4}$$

The generation of countermeasures network for speech denoising can overcome the shortcomings of the traditional methods of speech denoising. Not only can the training

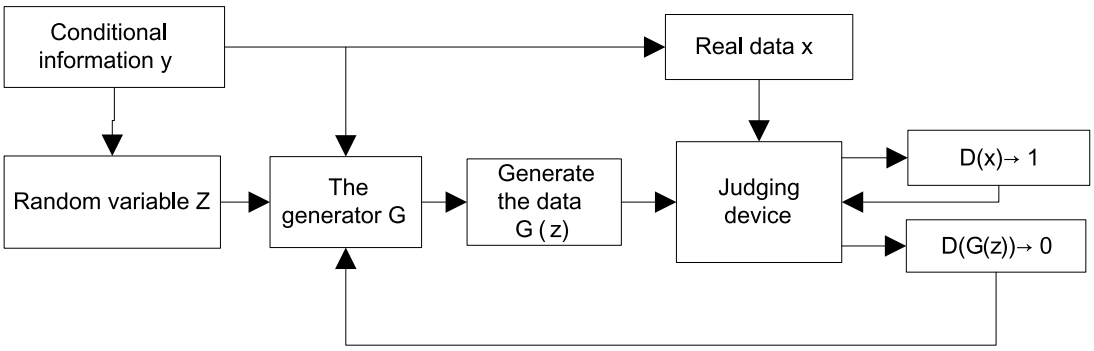

FIGURE 2. CGAN mode

speed of this model be greatly improved, but also the data set of any speech signal can be learned, which improves the generalization ability of the denoising model and does not need to estimate the noise signal in advance. Generative antagonism network has been widely used in video, text and image, and achieved good results. In this paper, the model is mainly applied to the speech denoising system, and the model is optimized, the loss function is improved, the least squares cross entropy loss function is used, and further improved, so as to improve the speech denoising effect of the generated antagonistic network model.

3. Improve the generation of anti-network loss function.

3.1. **Least squares loss function.** The experimental results show that generated by GAN voice quality is not high, mainly because the formation of traditional D to fight discrimination in the network device is using a sigmoid function[4], and because the sigmoid function saturation very rapidly, as shown in FIG. 3, so even a very small data points x, this function will ignore the sample x quickly to the distance of the decision boundary w. This means that the sigmoid function essentially does not penalize samples that are far from the decision boundary, and it also means that we are content to label x correctly, so the gradient of discriminator D will soon drop to 0. The least square loss function will punish the samples that are far away from the decision boundary and drag the false samples far away from the decision boundary into the decision boundary to improve the quality of the generated speech.

Based on this observation, we choose the least square generated antagonistic network (LSGAN). The loss function of discriminator is shown in equation,[5] and the loss function of generator is shown in equation.[6]

$$\min_{\mathrm{D}} \mathrm{V}_{\mathrm{LSGAN}}(\mathrm{D}) = \frac{1}{2}\mathrm{E}_{\mathrm{x}-\mathrm{P}_{\mathrm{data(x)}}}\left[(\mathrm{D}(\mathrm{x})-\mathrm{b})^2\right] + \frac{1}{2}\mathrm{E}_{\mathrm{z}-\mathrm{P}_{\mathrm{z}}(\mathrm{z})}\left[\mathrm{D}(\mathrm{G}(\mathrm{z})-\mathrm{a})^2\right], \tag{5}$$

$$\min_{\mathbf{D}} \mathrm{V}_{\mathrm{LSGAN}}(\mathrm{D}) = \frac{1}{2}\mathrm{E}_{\mathrm{z}-\mathrm{P}_{\mathrm{z}}(\mathrm{Z})}\left[\mathrm{D}(\mathrm{G}(\mathrm{z})-\mathrm{c})^2\right]. \tag{6}$$

In the above formula, by minimizing the loss function of discriminator, the real data is encoded as a, and the generated data is encoded as b. By minimizing the loss function of the generator, the generated data can be confused with the discriminator and coded as c to identify the generated data as real data. During the training of the generator, parameters are often adjusted according to the output result of the discriminator. In order to maximize the output result of the discriminator to be close to the real data sample, parameter b=c is generally set. When b=c=1 and a=0, the loss function is:

$$\min_{\mathrm{D}} \mathrm{V}_{\mathrm{LSGAN}}(\mathrm{D}) = \frac{1}{2}\mathrm{E}_{\mathrm{x}-\mathrm{P}_{\mathrm{data(x)}}}\left[(\mathrm{D}(\mathrm{x})-1)^2\right] + \frac{1}{2}\mathrm{E}_{\mathrm{z}-\mathrm{P}_{\mathrm{z}}(\mathrm{Z})}\left[\mathrm{D}(\mathrm{G}(\mathrm{z}))^2\right], \tag{7}$$

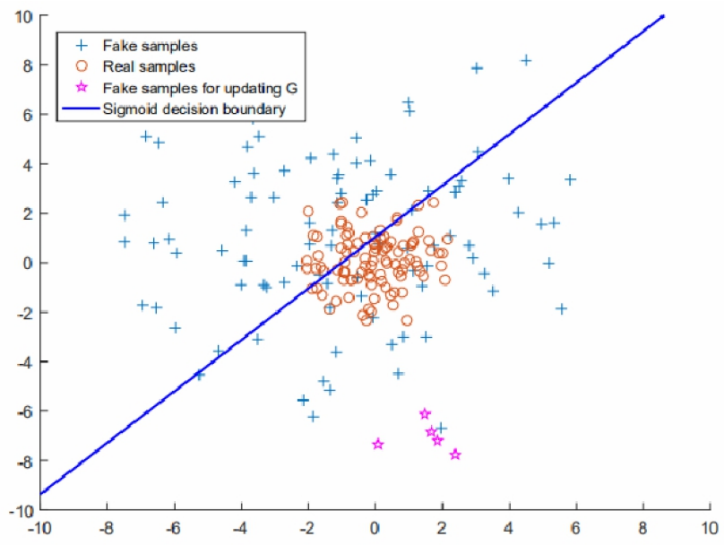

FIGURE 3. Cross entropy loss function

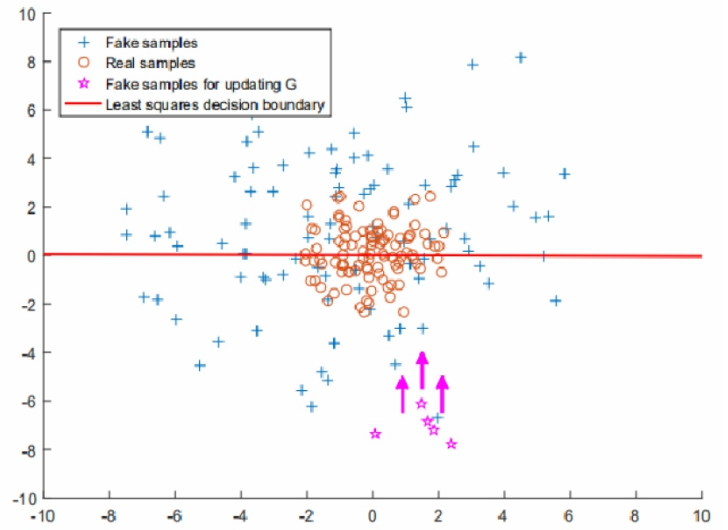

FIGURE 4. Least squares loss function

$$\min_{D} V_{LSGAN}(D) = \frac{1}{2} E_{z-P_z(Z)} \left[D(G(z) - 1)^2 \right]. \tag{8}$$

3.2. **Improvement of minimum loss function.** As the generator and discriminator of generating antagonistic network are composed of convolutional neural network, the parameters of model training increase exponentially with the increase of training structure level, which leads to the gradual increase of model complexity, model training is very difficult, and the speech denoising effect is not ideal. In addition, the generator will learn irrelevant features in the speech signal, resulting in a small error generated by the model, while the actual received speech signal has a large error compared with the pure speech signal. This results in overfitting and reduces the generalization ability of the model.

According to the deficiencies of the above analysis, this paper proposes to add a L1 norm on the basis of the least square loss function [6], which can measure the gap between the generated data and the real data and enhance the quality of the generated speech signal. The improved loss function formula can not only make the discriminator D, which

generates the anti-network, more accurate in distinguishing the generated samples from the real ones, but also make generator G generate the generated samples closer to the real ones. Where P is the weight parameter of L1 norm. The loss function of LSGANs generated network is shown as follows:

$$\min_{\mathbf{D}} \mathrm{V}_{\mathrm{LSGAN}}(\mathrm{D}) = \frac{1}{2}\mathrm{E}_{\mathrm{x}}\left[(\mathrm{D}(\mathrm{x})-1)^2\right] + \frac{1}{2}\mathrm{E}_{\mathrm{z}-\mathrm{P}_{\mathrm{z}}(\mathrm{Z})}\left[\mathrm{D}(\mathrm{G}(\mathrm{z}))^2\right], \tag{9}$$

$$\min_{\mathrm{G}} \mathrm{V}_{\mathrm{LSGAN}}(\mathrm{G}) = \frac{1}{2}\mathrm{E}_{\mathrm{z}\sim\mathrm{P}_{\mathrm{z}}(\mathrm{z}),\mathrm{x}^*\sim\mathrm{P}_{\mathrm{data}}(\mathrm{x}^*)}\left[\left(\mathrm{D}\left(\mathrm{G}\left(\mathrm{z},\mathrm{x}^*\right),\mathrm{x}^*\right)-1\right)^2\right] + \rho\left\|\mathrm{G}\left(\mathrm{z},\mathrm{x}^*\right)-\mathrm{x}\right\|. \tag{10}$$

In this loss function, the first term uses the characteristics of the least square function to represent the square error of the generated data and the real data in the feature space. The second is mainly to prevent the over-fitting from causing the quality decline after speech enhancement, and represents the absolute value of the difference between the enhanced speech and the pure speech, which is continuously adjusted through the weight parameter to improve the quality of the enhanced speech.[7]

3.3. **Improved least squares GAN speech denoising model algorithm.** First because of the influence of the speech signal by the excitation source, the high frequency of the signal attenuation is serious, so before the speech signal processing, generally uses the digital filter for speech signal preprocessing, secondly, speech signal is non-stationary and short-time stationarity, in order to extract the characteristic parameters of the optimal usually need to block or framing the voice. In this paper, hamming window function is used to add window to the speech signal. The speech signal processed by hamming window can be regarded as a stable random signal, which provides good input data for the speech processing system. In this paper, speech is denoised by the improved least squares GAN model, and the block diagram of the algorithm is shown in the figure.

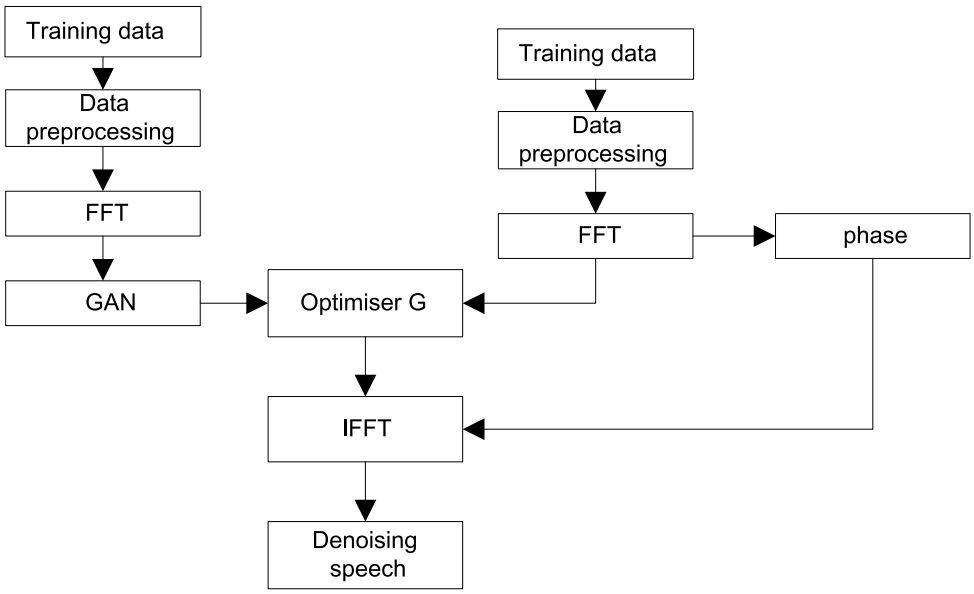

FIGURE 5. Flow chart of improved GAN model algorithm

In the speech denoising model of the optimized least squares GAN, its optimal weight value is mainly obtained by training generator G and discriminator D, and its internal GAN model algorithm is shown as follows

Least squares GAN speech denoising model algorithm
1. Design a function train() with startup and fixation functions
2. Set the number of iterations I, and start D through the function train()
3. Optimize D's objective function and reverse update D's weight coefficient
4. Fix D by the function train(), so that the weight of D is not updated
5. Optimize G and reverse update G's weight coefficient
6. $i = i + 1$
7. Repeat the above steps until the number of iterations I reaches the set value

4. **Experimental process.**

4.1. **Experimental environment.** This experiment takes tensorflow as the framework of network learning on the ubuntu system based on Linux, and constructs the model of generator G and discriminator D through the convolutional neural network. The model was trained and tested using pycharm software.

4.2. **Data preparation.** The data used in the experiment are NOIZEUS sound library and NOISEX-92 noise library. The NOIZEUS library provides pure speech, as well as noisy speech with signal-to-noise ratios of 0dB, 5dB, 10dB and 15dB. Noisex-92 noise library provides noise signal, including Babble noise, Street noise, Airport noise, Exhibition noise, Restaurant noise, Car noise, etc. Each noise includes four SNR: 0dB, 5dB, 10dB, 15dB. These two speech libraries can be used to superimpose the pure speech signal and the noise signal, and set the speech under different SNR and different noise conditions, so as to provide the training data set and test data set for the experiment.

4.3. **Experimental process.** PESQ evaluation methods, this paper build in article contains babble noise of 500 voice data set as the training data, the number of iterations I take 50, input article 20 babble noise voice as test data, to train the model test, using experimental method, choose between [0,150] 7 value test, draw a sparse formula ρ optimal weighting coefficient, the results shown in the following table

Table 1. PESQ scores of different values under babble noise

ρ	0db	5db	10db	15db
0	0.4654	0.4686	0.4596	0.4896
25	0.6354	1.0685	0.8695	1.6789
50	1.5987	1.5358	1.2598	2.3548
75	1.8648	2.0365	1.8965	2.4586
100	1.9828	2.3459	2.3658	2.6982
125	1.8625	2.1587	2.1589	2.3485
150	1.7659	2.0368	2.0385	2.1458

According to the experimental results in the table it is concluded that p take 100 speech denoising effect is best, so this article USES the weight coefficient ofρ equals 100, to deal with the noise of the airport noise, get the waveform diagram as shown in figure 6, the pure represents pure voice signal, noisy on behalf of the speech signal with the airport noise, voice on behalf of by improving GAN after denoising model of speech signal.

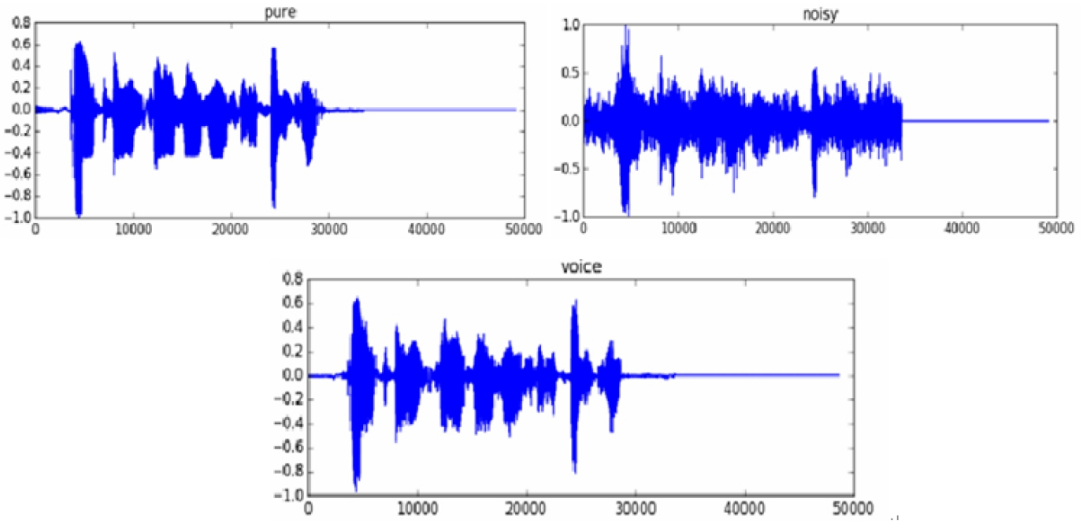

FIGURE 6. Speech signal waveform

In this paper, four speech enhancement algorithms are selected: spectral subtraction, MMSE enhancement algorithm, wiener filtering method, and generated antagonism network (GAN). Airport noise was used for the test. SNR and PESQ evaluation methods were used to calculate the mean of all evaluation scores of each noise under different SNR, and the results are shown in the following table:

TABLE 2. SNR values of speech denoising methods under airport noise

Denoising method	0db	5db	10db	15db
Spectral subtraction	5.15	8.56	10.59	12.68
Wiener filtering	-7.65	-6.64	-5.58	-4.68
MMSE	3.68	7.25	12.58	15.84
GAN	5.15	8.65	12.68	15.48
Improved GAN	5.35	8.96	13.65	16.89

TABLE 3. PESQ values of voice denoising method under airport noise

Denoising method	0db	5db	10db	15db
Spectral subtraction	1.23	1.25	1.58	1.20
Wiener filtering	1.84	1.76	1.85	1.81
MMSE	1.12	1.36	1.24	1.32
GAN	1.51	1.42	1.64	1.36
Improved GAN	1.56	1.48	1.78	1.50

Can be seen from the table above, under the condition of airport noise, this paper adopted the method of speech denoising SNR values are higher than other speech denoising method, PESQ value is lower than the wiener filtering method, is higher than that of the rest of the speech denoising method, shows that the method of speech denoising method of denoising performance better than the other method of speech denoising, voice quality is lower than that of wiener filtering method, but also higher than that of other three kinds of method of speech denoising. In general, the generated antagonistic network is improved in this paper, and its SNR and PESQ values are improved, indicating that the speech denoising model based on optimized GAN has better denoising performance.

REFERENCES

[1] Wang G, Li C, Dong L, Noise estimation using mean square cross prediction error for speech enhancement[J],*Circuit & Systems I Regular Paper IEEE Transactions on*, Springer, Berlin-Heidelberg, Germany, 2007.

[2] Odena A, Olah C, Shlens J, Conditional image synthesis with auxiliary classier GANs, *arXiv preprint arXiv:1610.09585*, 2016.

[3] Gulrajani I, Ahmed F, Arjovsky M, etal, Improved Training of Wasserstein GANs [J], *arXiv preprint arXiv: 1704.00028*, 2017.

[4] Santiago P, Antonio B, Joan S, SEGAN: Speech Enhancement Generative Adversarial Network [C], *INTERSPEECH*, 2017:3642-3646.

[5] Wei Chu, Speech enhancement based on conditional deep convolution generation countermeasure network [J], *Intelligent computer and application*, 2019, 9 (04):82-86.

[6] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, Feiyue Wang, Research progress and Prospect of generative countermeasure network Gan [J], *Journal of automation*, 2017, 43 (03):321-332.

[7] Huafeng Wang, Guixian Wang, A method of speech denoising based on generative antagonism network [J],*Journal of North University of technology*, 2019, 31 (5):84-89, 96.

[8] Yang Sun,Shoulin Yin,Hang Li,A New Wavelet Threshold Function Based on Gaussian Kernel Function for Image De-noising,*Journal of Information Hiding and Multimedia Signal Processing*, vol.10, no.1, pp.91-101, 2019.

[9] Lijian Zhou, Zuowei Wang, Han Guo, Siyuan Hao, Zhao Zhuo,Palm-print Recognition based on CNN against Rotation and Noise,*Journal of Information Hiding and Multimedia Signal Processing*, vol.9, no.6, pp.1603-1612, 2018.