2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) | 978-1-6654-2839-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/AUTEEE52864.2021.9668622

A Speech Enhancement Algorithm combining Spectral Subtraction and Wavelet Transform

Yi Yang
School of Mechanical and Electrical Engineering
Chengdu University of Technology
Chengdu, China
616545806@qq.com

Peipei Liu
School of Mechanical and Electrical Engineering
Chengdu University of Technology
Chengdu, China
xpiy@163.com
*Corresponding author

Huili Zhou
School of Mechanical and Electrical Engineering
Chengdu University of Technology
Chengdu, China
1357031585 @qq.com

Ye Tian
School of Mechanical and Electrical Engineering
Chengdu University of Technology
Chengdu, China
1007865496 @qq.com

***Abstract*—Spectral subtraction is widely used in the research of speech enhancement because of its small calculation amount and fast processing speed. However, traditional spectral subtraction has a large amount of residual music noise in the process of speech noise reduction. In order to solve this problem, a speech enhancement algorithm based on improved spectral subtraction and improved threshold wavelet transform is proposed. The method first uses spectral subtraction to denoise the noisy speech signal, and then the processed signal is subjected to wavelet transform perform voice enhancement. By using MATLAB software to simulate the system, the experimental results show that compared with the traditional spectrum subtraction method, the improved method can effectively suppress the music noise generated by the single spectrum subtraction method, improve the output signal-to-noise ratio and recognizability of the speech signal, and enhance the speech The effect is remarkable.**

Keywords—spectral subtraction,wavelet transform, music noise, speech enhancement

I. INTRODUCTION

Voice is the most efficient way of communication for people to communicate emotionally and transfer information. However, in the process of signal transmission, voice will be interfered by noise signals from the environment, transmission media, and communication equipment. In order to reduce the influence of noise in the transmission process, for different noise sources, many scholars at home and abroad have conducted research on speech enhancement methods. Currently widely used speech enhancement methods include: spectral subtraction, wavelet transform, adaptive filtering, and Kalman filtering, Wiener filtering, etc[1],[2].

Spectral subtraction, because of its advantages of simple calculation and good real-time performance, is widely used in the research of speech noise reduction processing, but the denoised speech has music noise [3], which reduces the intelligibility of the speech signal. In response to this problem, researchers have proposed improvements on traditional spectral subtraction, such as the nonlinear spectral subtraction proposed by Berouti [4], the adaptive gain average spectral subtraction proposed by Gustasson [5], probability spectral subtraction, etc[6]. These improved algorithms can suppress music noise to varying degrees and improve the performance of traditional spectrum subtraction, but the single algorithm has limited speech enhancement capabilities.

With the continuous development and maturity of wavelet analysis theory, wavelet transform is widely used in speech signal processing research. In noisy speech signals, speech signals are usually concentrated in the low frequency domain, and noise signals are usually concentrated in the high frequency domain. Wavelet transform, because of its variable time and frequency resolution characteristics, can better compensate for the fixed resolution of spectral subtraction and other defects , so it is of research significance to use wavelet transform to analyze noisy speech.

This paper analyzes the principles of traditional spectral subtraction and wavelet transformation. On this basis, the traditional algorithm is improved, and a speech enhancement algorithm based on the combination of improved spectral subtraction and improved threshold wavelet transform is proposed. Comparing the experimental results, it can be found that this method can better process time-varying and non-stationary signals such as speech signals, effectively suppress music noise, and improve the intelligibility of speech signals.

II. SPECTRUM SUBTRACTION

A. The principle of traditional spectrum subtraction

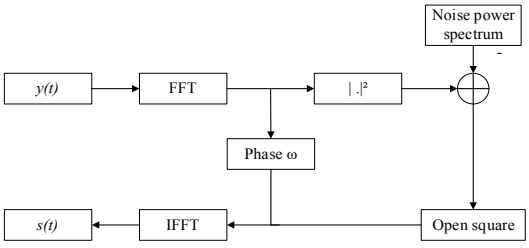

Fig. 1. Schematic diagram of traditional spectrum subtraction

Spectral subtraction is widely used in speech enhancement. This method takes advantage of the statistical stability of noise and the independence of speech and additive noise to enhance noisy speech[7]. This method has no reference noise source, assuming that the noise is statistically stable, that is, the expected value of the amplitude of the noise during the period of speech is equal to the expected value of the amplitude of the noise during the period of no speech, and the spectrum of the noise during the period of speech can be estimated by the noise spectrum calculated during the period of no speech., And finally

subtract the frequency spectrum of the noisy speech with the noise frequency spectrum to obtain the estimated value of the speech signal frequency spectrum[8]. The schematic diagram is shown in Fig. 1.

Suppose *y(t)* is the noisy speech, *s(t)* is the pure speech signal, *n(t)* is the additive white Gaussian noise, and *s(t)* and *n(t)* are independent of each other, from this we can get the relationship of the three:

$$y(t) = s(t) + n(t) \quad (1)$$

Perform Fourier changes on *y(t)*, *s(t)*, and *n(t)* to obtain *Y(ω)*, *S(ω)*, and *N(ω)* respectively,then the Fourier change of (1) can be expressed for:

$$Y(\omega) = S(\omega) + N(\omega) \quad (2)$$

In spectral subtraction, since the pure speech signal and the noise signal are not correlated with each other, we can get:

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 \quad (3)$$

Assuming that the noise is a stationary random process, the noise signal spectrum when there is speech is equal to when there is no speech, the power spectrum of noise can be estimated from *N(ω)*, which can be derived:

$$|S(\omega)|^2 = |Y(\omega)|^2 - |N(\omega)|^2 \quad (4)$$

Take the square root of (4) to obtain the estimated value of the speech signal after spectral subtraction enhancement:

$$|\hat{S}(\omega)| = \left[|Y(\omega)|^2 - |N(\omega)|^2\right]^{\frac{1}{2}} \quad (5)$$

When the value obtained by the above difference is negative, it is set to zero.

The human ear's ability to distinguish the phase of each spectral component of the speech signal is low, and the speech is mainly recognized by the amplitude of each spectral component [9]. Therefore, the phase of the noisy speech can be used to estimate the phase of the pure speech, that is, the enhanced speech after spectral subtraction is obtained by performing IFFT on *S(ω)*, as shown in (6):

$$\hat{S}(t) = IFFT\left[|S(\omega)| e^{j\varphi(\omega)}\right] \quad (6)$$

B. Improved spectrum subtraction

The traditional spectral subtraction method has a small amount of calculation and simple principle, but the traditional spectral subtraction method has music noise in the speech after the noisy speech enhancement, which affects the speech intelligibility. In order to reduce the influence of music noise on the speech enhancement effect, Berouti improved the traditional spectral subtraction method, by introducing two parameters a and b to control the size of the noise power spectrum and limit the minimum value of the speech signal power spectrum, a is Spectral subtraction noise coefficient, b is the spectral subtraction power correction coefficient, the expression after improvement is defined as follows:

$$|\hat{S}(\omega)| = \begin{cases} \left[|Y(\omega)|^b - a|N(\omega)|^b\right]^{\frac{1}{b}} & |Y(\omega)|^b \geq a|N(\omega)|^b \end{cases} \quad (7)$$

In spectral subtraction, the noise spectrum is estimated in the silent period when there is no speech, but in the actual process, the noise spectrum obeys the Gaussian distribution, and its amplitude can dynamically change within a wide range. When the actual noise spectrum is large When the frequency spectrum is subtracted, there will be noise residue. Therefore, the noise coefficient a can be appropriately increased to reduce the noise spectrum residue. Generally, a>1 is used to enhance the speech spectrum by subtracting more noise spectrum. b is the power correction coefficient of spectral subtraction. By increasing the value of b, the output signal-to-noise ratio can be increased, but at the same time it will increase the distortion of the pure speech signal. When the input signal-to-noise ratio is low, the power correction coefficient b of spectral subtraction is effective When the input signal-to-noise ratio is high, b has a small effect on the denoising effect. In order to facilitate the calculation, take b=2 in the subsequent experiments. In summary, we can get:

$$|\hat{S}(\omega)| = \begin{cases} \left[|Y(\omega)|^2 - a|N(\omega)|^2\right]^{\frac{1}{2}} & |Y(\omega)|^2 \geq a|N(\omega)|^2 \end{cases} \quad (8)$$

In the actual speech enhancement process, the values of various coefficients should be dynamically adjusted according to the situation of the noisy signal to better reduce music noise and improve the signal-to-noise ratio.

III. WAVELET TRANSFORM

A. Wavelet transform denoising

In the early 1980s, Morlet proposed wavelet transform [10], which is a local time-frequency analysis method with multi-resolution [11], which can make up for the fixed resolution of Fourier transform to a certain extent. The noisy speech signal is wavelet decomposed on various scales, and the obtained wavelet coefficients represent the information of the signal at different resolutions. This information reflects the non-stationary characteristics of the signal, and at the same time, because of its low entropy value and decorrelation features[12], so that the wavelet transform is conducive to the extraction of local features of the speech, suitable for noisy speech processing, the process of denoising is shown in Fig. 2.

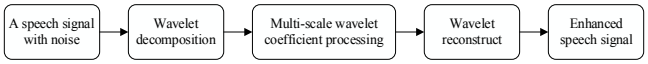

Fig. 2. Principle diagram of wavelet transform denoising

First, perform wavelet decomposition on the noisy speech to obtain wavelet decomposition coefficients at each scale. Then process the wavelet coefficients at each scale to eliminate the wavelet coefficients of the noise signal, and retain the wavelet coefficients of the speech signal at different scales. Finally use the wavelet coefficients The signal is reconstructed, and the estimated value of the speech signal after noise removal is calculated.

B. Wavelet threshold denoising

Wavelet transform and its application have developed rapidly. In 1991, Mallat proposed the theory of singularity detection and derived the modulus maximum denoising method [13]. In 1994, Donoho proposed the nonlinear wavelet transform threshold denoising method [14]. In addition, there are de-stationary invariant wavelet de-noising method, correlation de-noising method and so on. Compared with several other methods, the wavelet threshold denoising method has become one of the popular research directions because of its strong denoising ability and simple operation.

The principle of wavelet threshold denoising: After the noisy speech is transformed by wavelet, the energy of the pure speech signal is mainly concentrated in a few wavelet coefficients with large absolute value and in a specific frequency range, while the noise is distributed in the wavelet domain of various scales. In general, the amplitude of the wavelet coefficients of noise is smaller than the absolute value of the amplitude of the wavelet coefficients of speech. With this feature, by choosing a reasonable wavelet threshold, the wavelet coefficient of the speech component and the wavelet coefficient of the noise signal can be set to 0, and finally use different scales to be effective the wavelet coefficients of the signal reconstruct the enhanced speech.

C. Classical threshold function

The selection of the threshold function determines the effect of signal reconstruction. A threshold λ will be selected during the threshold processing. The coefficient of the speech signal greater than λ is reserved for speech signal reconstruction, and the noise coefficient is set to 0 if the noise coefficient is less than λ. There are two classic threshold functions: hard threshold function and soft threshold function [15].

1) Hard threshold function:

$$\begin{cases} \omega_1 = \omega & |\omega| \geq \lambda \\ \omega_1 = 0 & |\omega| < \lambda \end{cases} \tag{9}$$

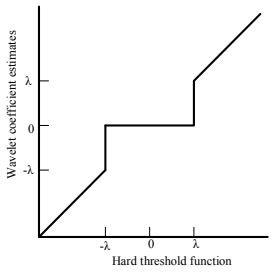

Fig. 3. Hard threshold function

ω is the wavelet coefficient after decomposition, ω_1 is the wavelet coefficient estimated after denoising, and λ is the noise threshold. Equation (9) shows the principle of the hard threshold function: when the decomposed wavelet coefficient ω of the noisy signal is greater than or equal to the threshold λ, the estimated wavelet coefficient ω_1 is equal to ω, and when the decomposed wavelet coefficient ω is less than the threshold λ, ω_1 is equal to 0, The function diagram is shown in Fig. 3.

It can be seen from Fig. 3 that the wavelet coefficients of the hard threshold function are not continuous, and there are discontinuities at ±λ. Therefore, after the signal is reconstructed, new oscillating noises will be generated to cause signal mutations, namely the pseudo-Gibbs phenomernon, which affects the noise removal effect.

2) Soft threshold function

$$\begin{cases} \omega_1 = \mathrm{sgn}(\omega)\left(|\omega| - \lambda\right) & |\omega| \geq \lambda \\ \omega_1 = 0 & |\omega| < \lambda \end{cases} \tag{10}$$

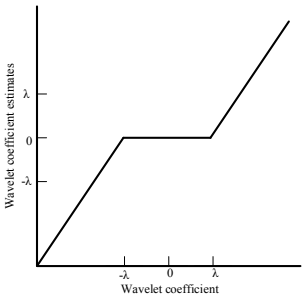

Fig. 4. Soft threshold function

sgn(ω) is a sign function. The other signs in the above expression have the same meaning as the hard threshold function. When the wavelet coefficient ω after decomposing the noisy signal is greater than or equal to the threshold λ, ω_1 is calculated according to the above expression. When the coefficient ω is less than λ, ω_1 is equal to 0. The soft threshold function is shown in Fig. 4.

Analyzed from Fig. 4, different from the hard threshold function, the wavelet coefficients processed by the soft threshold function are smoother and have continuous uninterrupted points at ±λ. Overcome the shortcomings of the hard threshold function with discontinuous wavelet coefficients at ±λ, but there is a constant deviation of λ between the soft threshold function ω_1 and ω, which will affect the degree of approximation between the reconstructed signal and the pure speech signal, resulting in the reconstructed speech and pure speech. There are deviations between voices.

D. Improved threshold function

According to the previous analysis, the above two classic threshold functions have inherent defects in the processing process. A large number of experiments have shown that the denoising of the speech signal processed by the hard threshold function is not clean, and there is still a small amount of music noise in the processed speech signal. Although the speech processed by the soft threshold function eliminates the music noise, the speech signal amplitude is weakened, which reduces the intelligibility of the original voice signal and the clarity of the voice. In order to reduce the influence of the threshold function in the process of speech processing and improve the intelligibility and naturalness of speech, this article adopts the following improved threshold function:

$$\begin{cases} \omega_1 = \mathrm{sgn}(\omega)\left(|\omega| - \lambda\right) & |\omega| \geq \lambda \\ \omega_1 = 0 & |\omega| < \lambda \end{cases} \tag{11}$$

The threshold function uses the modulus square function. In some noisy speech signals, the difference between the wavelet coefficients of the speech signal and the noise is small. After modulus square processing, the difference between the wavelet coefficients of the speech signal and the noise signal can be increased. It is more conducive to the separation between speech and noise.

Introduce the coefficient a, when a=0, it is a hard threshold function. From the above analysis, it can be seen that when $|\omega|>=\lambda$, the difference between the wavelet coefficients ω_1 and ω estimated by the soft threshold function is λ, which makes the reconstructed speech deviate from the pure speech signal. Lead to poor denoising effect. In order to reduce the deviation and avoid reducing the deviation to 0 at the same time, the size of a can be adjusted appropriately to change the degree of approximation between the reconstructed speech signal and the pure speech signal to improve the noise reduction performance. Normally, $0<a<1$, when a is 0.5, the threshold function is shown in Fig. 5.

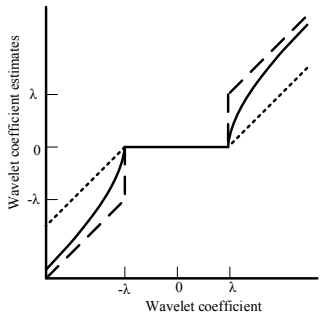

Fig. 5. Improved threshold function

The threshold function is a joint improvement of the hard and soft threshold functions, and the improved function combines their respective advantages. It not only overcomes the problem that the hard threshold function is discontinuous at $\pm\lambda$, but also overcomes the constant deviation λ of the wavelet coefficients after the soft threshold function is estimated. With the continuous increase of the absolute value of the wavelet coefficients, the gap between the estimated value of the wavelet coefficients and the actual value gradually decreases, and the degree of approximation between the improved function curve and the hard threshold function is getting higher and higher.

IV. Improved Speech Enhancement Algorithm

A single speech enhancement method has limited denoising capabilities. In actual applications, different denoising methods are usually used in combination according to different speech and environmental noise characteristics, and their respective advantages are used to obtain better speech enhancement effects. In this article, in order to suppress the residual music noise after spectral subtraction denoising, a speech enhancement algorithm combining improved spectral subtraction and improved threshold wavelet analysis method (hereinafter referred to as: improved speech enhancement algorithm) is adopted. The method steps are shown in Fig. 6.

The Fig.6 shows the processing flow of the improved speech enhancement algorithm:

Step 1 Use the improved spectral subtraction method to preprocess the noisy speech signal to obtain the speech signal with music noise.

Step 2 Use the coif wavelet as the fundamental wave function to decompose the speech signal after spectral subtraction processing.

Step 3 Use an improved threshold function to extract the wavelet coefficient components of the speech signal on each scale, and eliminate the noise wavelet coefficient components on each scale.

Step 4 Use the wavelet coefficients retained after threshold processing to reconstruct the signal to obtain an enhanced speech.

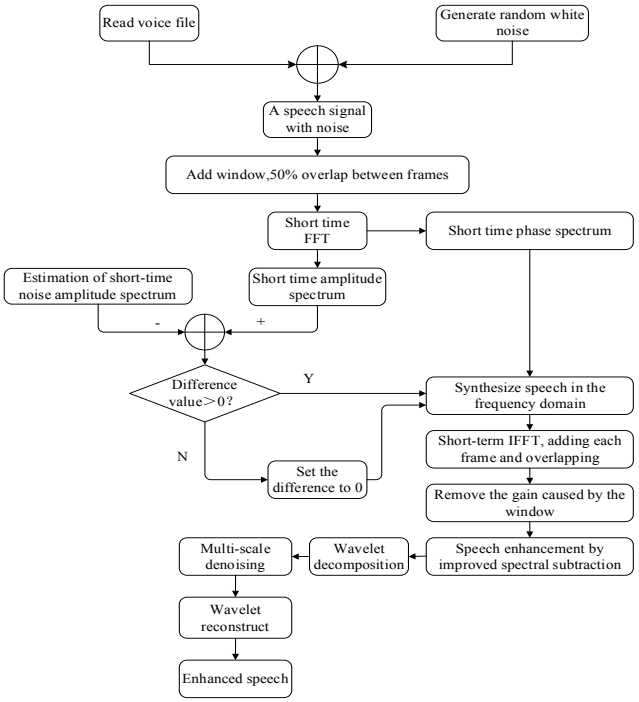

Fig. 6. Improved spectral subtraction and improved threshold wavelet transform combined algorithm

V. Simulation Result Analysis

This experiment simulates the system model through MATLAB, the sampling frequency is 8kHz, the frame length is 1024, and the overlap between frames is 50%. The pure voice signal is the voice recorded under the condition of quiet and no interference, using Gaussian white noise as the input noise of the system. Adding Gaussian white noise with signal-to-noise ratios of -5dB, 0dB, 5dB, and 10dB to the collected pure speech signal to obtain the noisy speech signal to be processed. Perform traditional spectral subtraction, improved spectral subtraction and improved speech enhancement algorithm processing on noisy speech signals under different background noise conditions. The results are shown in Table I, Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11. Table I shows the output signal-to-noise ratio of the above three processing methods under different input signal-to-noise ratio conditions.

It can be seen from the data in Table I: Compared with the traditional spectral subtraction, the output signal-to-noise ratio of the improved spectral subtraction has been improved under different input conditions. The speech processed by the improved spectral subtraction is further processed by wavelet transform, which can further improve the output signal-to-

noise ratio, and when the input signal-to-noise ratio is lower, the performance improvement is greater and the speech enhancement effect is more obvious.

TABLE I. SNR COMPARISON OF DIFFERENT ALGORITHMS

Output signal to noise ratio	Input signal to noise ratio			
	-5dB	0dB	5dB	10dB
Traditional spectral subtraction	2.355	4.314	8.390	11.770
Improved spectrum subtraction	3.690	5.252	9.250	12.482
Improved speech enhancement algorithm	5.569	6.553	9.925	12.936

Fig. 7 shows the pure speech signal, and Fig. 8 shows the noisy speech signal after adding 0dB Gaussian white noise. Fig. 9, Fig. 10 and Fig. 11 respectively show the waveforms of 0dB noisy speech signal after traditional spectral subtraction, improved spectral subtraction and improved speech enhancement algorithm processing.

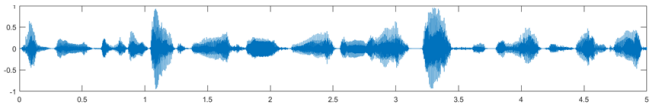

Fig. 7. Pure voice signal

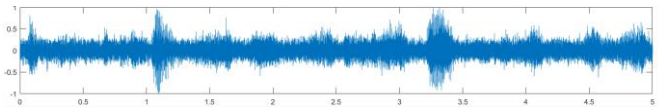

Fig. 8. A speech signal with noise

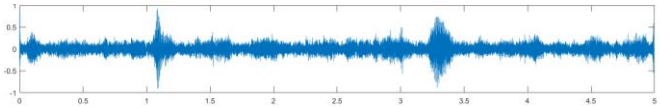

Fig. 9. Traditional spectrum subtraction

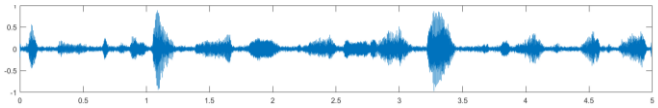

Fig. 10. Improved speech after spectrum subtraction enhancement

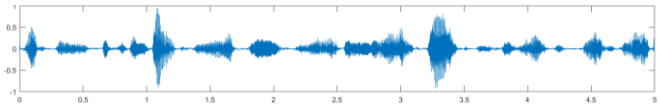

Fig. 11. Improved speech enhancement algorithm

After analyzing the above results, it can be seen from Fig. 8 and Fig. 9 that the speech enhancement ability of traditional spectrum subtraction is weak, and there are still a lot of noise residues in the speech processed by traditional spectrum subtraction.Compared with the traditional spectral subtraction, the improved spectral subtraction has better denoising effect and can better reduce the noise residue, but it can't completely remove the music noise, and the denoising effect is not ideal. Comparing the waveforms, it is found that the improved speech enhancement method proposed in this paper can greatly reduce the noise amplitude and has good denoising effect. Through experimental playback of the denoised speech signal and comparison with the pure speech signal, the improved spectral subtraction still has a small amount of music noise residue, which reduces the speech intelligibility. The improved speech enhancement algorithm proposed in this paper greatly weakens the music noise residue after spectral subtraction speech enhancement, has less damage to the signal, and the enhanced speech has high ear recognition, it can make up for the defect of spectral subtraction speech enhancement.

VI. CONCLUSION

Noise processing is one of the main research contents of speech enhancement. Spectral subtraction is widely used due to its advantages such as simplicity and small amount of calculation. However, the speech denoised by spectral subtraction has obvious musical noise, which affects the recognizability of speech. Wavelet transform causes Its low-entropy, multi-resolution characteristics and other features are suitable for extracting voice features and have a good denoising effect. This paper analyzes the principles of traditional spectral subtraction and wavelet transform, and at the same time improves the traditional spectral subtraction and wavelet threshold function, and proposes a speech enhancement algorithm that combines improved spectral subtraction and improved threshold wavelet transform. This method is based on spectral subtraction, using the different distribution of wavelet coefficients of speech signal and noise signal on various scales, and extracting wavelet coefficients of speech signal through reasonable threshold function selection to realize speech reconstruction. According to the analysis of the experimental results of noisy speech under four different noise backgrounds, compared with the speech enhancement only using spectral subtraction, the improved speech enhancement algorithm can effectively eliminate the residual music noise of spectral subtraction, improve the output signal-to-noise ratio, have better intelligibility and naturalness of speech signal, and have remarkable speech enhancement effect.

REFERENCES

[1] Wang Jing, Fu Fenglin, Zhang Yunwei. Overview of speech enhancement algorithms [J]. Acoustics and Electronic Engineering, 2005(01): 22-26.

[2] Zheng Zhanheng, Zeng Qingning. Research and improvement of speech enhancement algorithm[J]. Modern Electronic Technology, 2020, 43(21): 27-30.

[3] Ding Wu, Zhang Xiaobing, Wu Peng. Speech enhancement combined with spectral subtraction and post-masking processing[J]. Electroacoustic Technology, 2018, 42(03): 52-54.

[4] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.1979,4:208-211.

[5] Gustsfsson H., Nordholm S, Claesson I. Spectral subtraction using reduced delay convolution and adaptive averaging[J]. IEEE Trans. On Speech and Audio Processing,2001,9(8):799-807.

[6] Lockwood P,Boudy J. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars[J]. Speech Communication. 1992,11:215-228.

[7] Zhang Xueying. Digital speech processing and MATLAB simulation [M]. Publishing House of Electronics Industry, 2010: 207-215.

[8] Liu Yaqin, Gan Wenli. A speech enhancement algorithm based on spectral subtraction [J]. Microcomputer Applications, 2020, 36(12): 56-57+64.

[9] Wu Weipeng. Research on Speech Enhancement Algorithm Based on Improved Spectral Subtraction [D]. Nanjing University of Posts and Telecommunications, 2019.

[10] Morlet. Wave propagation and sampling theory and complex waves. Geophysics,1982,47(2):801-813.

[11] Lin Yirong, Xu Shilin. Multi-resolution speech denoising based on wavelet transform [J]. Journal of Hefei University of Technology Natural Science Edition, 2000, 23(5): 688-693.

[12] Wang Yongtao. Research on Speech Signal Denoising Based on Wavelet Transform [D]. Nanjing University of Posts and Telecommunications, 2014.

[13] S. Mallat and W. L. Hwang.Singularity detection and processing with wavelets[J] IEEE Transactions on Information Theory,1992,38(2): 617-643.

[14] Donoho D. L. and Johnstone LM. Adapting to Unknown Smoothness via Wavelet Shrinkage. Journal of American Stat. Assoc. 1995, 12(90):1200-1224.

[15] Donobo D L,Johnstone I M. Ideal denoising in an orthogonal basis chosen from a library of bases[J]. C R Acad Sci I-Math,1994, 319: 1317-1322.