DISCRETE S-TRANSFORM BASED SPEECH ENHANCEMENT

Guo-hua Hu[1, 2], Rui Li[1] , and Liang Tao[1*]

[1] MOE Key Laboratory of Intelligent Computing and Signal Processing,
Anhui University, Hefei 230601, China

[2] Department of Electronics and Electrical Engineering,
Hefei University, Hefei 230601, China

[*]Email: taoliang@ahu.edu.cn

Submitted: Aug. 30, 2015 ***Accepted: Nov. 15, 2015*** ***Published: Dec. 1, 2015***

Abstract: S-transform (ST) is an effective time-frequency representation method with the advantages of short-time Fourier transform and wavelet transform. This paper utilizes the advantages of the power spectral subtraction method and the speech enhancement method to presents a novel speech enhancement algorithm based on discrete ST. Firstly, the speech in time domain is transformed by ST to joint time-frequency domain for spectral subtraction so that the clear speech spectrum can be obtained and then the inverse ST is performed to acquire the enhanced speech in time domain. Simulation experiment is done to verify the validity of the method. The experiment results show that the proposed method can effectively enhance the de-noising ability and improve the signal-to-noise ratio.

Index terms*: Discrete s-transform, time-frequency domain, speech enhancement.*

I. INTRODUCTION

Recognition, coding and enhancement of speech signals in the presence of noise and reverberation remain a challenging problem in many applications. Hence, the researchers gave a large attention to speech processing and enhancement [1, 2]. Speech processing technology is used in a wide variety of applications such as speech pre-processing, speech coding and speech recognition. In a noisy environment, speech enhancement can be used to improve the quality, decrease the hearing fatigue, improve the performance, and increase intelligibility of the speech communication systems [3, 4].

Several methods have been proposed for this purpose such as the spectral subtraction method, the signal subspace method, the Wiener filtering method, the wavelet denoising method [5-8], the adaptive Wiener filtering method [9], the LMS method [10-12] and the RLS method [13, 14]. The goal of most techniques is to improve the Signal-to-Noise-Ratio (SNR) of speech. Spectral subtraction is one of the effective methods [15-17]. But it is rather difficult for the speech signal to be immune to the pollution of different noises. One important purpose of speech enhancement is to eliminate the noise from the noisy speech as much as possible.

The speech enhancement technology plays an important role in speech encoding, recognition and other speech processing fields. There are mainly three categories of current mainstream study of speech processing: the time-domain processing method, the frequency-domain processing method and the joint time-frequency domain processing method, which normally utilize the enhancement algorithms based on short-time Fourier transform or wavelet transform based spectrum estimation, perception characteristics, and subspace, etc. The short-time Fourier transform is the most widely used and effective linear transform so that the spectrum subtraction method based on it is widely utilized in actual applications. The spectrum subtraction method requires transforming the signal from the time domain to the frequency domain by using the window Fourier transform. But its unchangeable window width debases the effect of this method. The S-transform has the advantages of both short time Fourier transform and wavelet transform, which can use changed window functions and Fourier transform nucleus [18-21]. Besides, the s-transform can keep the phase position information of the signal and provide the changed time-frequency accuracy. As a linear transform, it can be used as the effective tool for signal analysis and synthesis, unlike the method of Wigner-Ville distribution-WVD which has a lot of cross

terms, nor like the Fourier transform which uses fixed window function [22, 23]. The S-transform is performed for speech signal with noise so that the signal is transformed from the time domain to the time-frequency domain to conduct spectral subtraction. Since the S-transform has unfixed time-frequency resolution, better effect can be achieved in the time-frequency domain.

In this paper, firstly, we review the speech enhancement problem, the s-transform and the Gabor transform. Secondly, we study and utilize the advantages of the power spectral subtraction method and the speech enhancement method to present a new speech enhancement algorithm based on discrete s-transform. The speech in time domain is transformed by s-transform to joint time-frequency domain for spectral subtraction so that the clear speech spectrum can be obtained, and then the inverse s-transform is performed to acquire the enhanced speech. Simulation experiment will be done to verify the validity of the proposed algorithm.

II. REVIEW OF S-TRANSFORM

The definition of standard s-transform (ST) for continuous time function $x(t)$ is given in [2] as:

$$S(t,f)=\int_{-\infty}^{+\infty}x(\tau)w(t-\tau,f)e^{-j2\pi f\tau}d\tau \tag{1}$$

where $j=\sqrt{-1}$, τ and t are time variables, f is a frequency variable and $w(t-\tau,f)$ is a Gaussian function as follows:

$$w(t-\tau,f)=\frac{1}{\sigma(f)\sqrt{2\pi}}e^{-\frac{(t-\tau)^2}{2\sigma(f)^2}} \tag{2}$$

The advantage of ST relative to STFT is that the width (variance) $\sigma(f)$ of Gaussian function is not fixed any more, which is given by:

$$\sigma(f)=\frac{1}{|f|} \tag{3}$$

Substituting (2) and (3) into (1) leads to:

$$S(t,f)=\int_{-\infty}^{+\infty}x(\tau)\frac{|f|}{\sqrt{2\pi}}e^{-\frac{(t-\tau)^2|f|^2}{2}}e^{-j2\pi f\tau}d\tau \tag{4}$$

As shown in (2), the window function is the function of time t and frequency f. The width $\sigma(f)$ of window function is in inverse proportion to frequency f , i.e., the smaller the frequency,

the larger the width of window function; The larger the frequency, the smaller the width of the window function. Therefore, the s-transform has changeable time-frequency accuracy. The window width in the s-transform is the frequency function, and the window standard deviation is fixed to be the reciprocal of *f*. Sometimes the aggregation is bad in different signal analyses, which limits its applications. To facilitate the calculation, Stockwell proposed in [18] another kind of equivalence type:

$$S(t,f)=\int_{-\infty}^{+\infty}X(\alpha+f)e^{-\frac{2\pi^2\alpha^2}{f^2}}e^{j2\pi\alpha t}d\alpha,\qquad f\neq 0 \tag{5}$$

where $X(f)$ is the Fourier spectrum. From (5), we know that the ST can turn to fast Fourier transform and convolution theory for calculation. Based on (1), the integration of time *t* can lead to:

$$\begin{aligned}&\int_{-\infty}^{+\infty}S(t,f)dt=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x(\tau)w(t-\tau,f)e^{-j2\pi f\tau}d\tau dt\\&=\int_{-\infty}^{+\infty}x(\tau)\left[\int_{-\infty}^{+\infty}w(t-\tau,f)dt\right]e^{-j2\pi f\tau}d\tau=X(f)\end{aligned} \tag{6}$$

where

$$\int_{-\infty}^{+\infty}w(t-\tau,f)dt=1 \tag{7}$$

The inverse transform of ST is given by

$$x(\tau)=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}S(t,f)dt\right]e^{j2\pi f\tau}df \tag{8}$$

where the discrete forms of (4) and (5) are shown in (9), (10) and (11). The discretization of (8) is shown in (12).

$$S(m,n)=\sum_{l=0}^{N-1}x(l)\frac{|n|}{N\sqrt{2\pi}}e^{\frac{-n^2(m-l)^2}{2N^2}}e^{\frac{-j2\pi ml}{N}},n\neq 0 \tag{9}$$

$$S(m,n)=\sum_{\alpha=\frac{N}{2}}^{\frac{N}{2}-1}X(\alpha+n)e^{\frac{-2\pi^2\alpha^2}{n^2}}e^{\frac{j2\pi\alpha m}{N}},n\neq 0 \tag{10}$$

where $n\neq 0; n=1,2,3,\cdots,N-1;\ m=0,1,2,\cdots,N-1.$

$$S(m,0)=\frac{1}{N}\sum_{l=0}^{N-1}x(l) \tag{11}$$

$$x(l)=\sum_{n=0}^{N-1}\left(\frac{1}{N}\sum_{m=0}^{N-1}S(m,n)\right)e^{\frac{j2\pi ml}{N}} \tag{12}$$

where N is the length of the discrete-time signal.

III. DISCRETE S-TRANSFORM BASED POWER SPECTRAL SUBTRACTION FOR SPEECH DENOISING

Assume that the speech signal $x(n)$ with noise can be represented as:

$$y(n)=x(n)+d(n) \tag{13}$$

where $x(n)$ is a pure speech signal, $d(n)$ is additive noise, and the statistics of $x(n)$ and $d(n)$ are irrelevant. (14) can be obtained by s-transform of both sides of (13).

$$Y_s(m,n)=X_s(m,n)+D_s(m,n) \tag{14}$$

where $Y_s(m,n)$, $X_s(m,n)$ and $D_s(m,n)$ are the ST coefficients obtained from $y(n)$, $x(n)$ and $d(n)$ based on (9) and (11). m is the m-th time sampling point of certain frame of speech, n represents the n-th frequency sampling point and the length of the frame is N. $Y_s(m,n)$ and $X_s(m,n)$ can be written to be the forms of amplitude and phase position as shown in (15) and (16), where $A_{m,n}$ and φ_n, $R_{m,n}$ and θ_n are respectively the amplitude and phase position of $Y_s(m,n)$ and $X_s(m,n)$. Since the ears of human beings are not sensitive to phase positions, φ_n and θ_n can be considered equal.

$$Y_s(m,n)=A_{m,n}e^{j\varphi_n} \tag{15}$$

$$X_s(m,n)=R_{m,n}e^{j\theta_n} \tag{16}$$

Formula (17) can be obtained based on the principle that the energy keeps unchanged in time domain and in joint time-frequency domain.

$$\left|X_s(m,n)\right|^2=\left|Y_s(m,n)\right|^2-\left|D_s(m,n)\right|^2 \tag{17}$$

where $Y_s(m,n)$ can be obtained by direct discrete s-transform. However, the noise energy spectrum $\left|D_s(m,n)\right|^2$ cannot be obtained accurately and has to be obtained approximately by the way of estimation. The normal estimation method in [24, 25] is to take the statistical average $E[\left|D_s(m,n)\right|^2]$ in the absence of speech as $\left|D_s(m,n)\right|^2$. In order to verify that the discrete s-

transform has better denoising ability and changeable time-frequency resolution, the noise estimation algorithm in the absence of speech is adopted, then the estimation $\left|\widetilde{X}_s(m,n)\right|^2$ of power spectrum of pure speech can be obtained from (18).

$$\left|\widetilde{X}_s(m,n)\right|^2 = \left|Y_s(m,n)\right|^2 - E[\left|D_s(m,n)\right|^2] \tag{18}$$

Thus, $\widetilde{X}_s(m,n)$ can be obtained by (16), where $R_{m,n} = \left|\widetilde{X}_s(m,n)\right|$ and $\theta_n = \varphi_n$, which is known. Enhanced speech can be obtained by the following inverse discrete s-transform,

$$\widetilde{x}(k) = \sum_{n=0}^{N-1}\left(\frac{1}{N}\sum_{m=0}^{N-1}\widetilde{X}_s(m,n)\right)e^{\frac{j2\pi mk}{N}}, \qquad 0 \le k \le N-1 \tag{19}$$

where $\widetilde{x}(k)$ is the speech after denoising. The diagram of the proposed algorithm is shown in Figure 1.

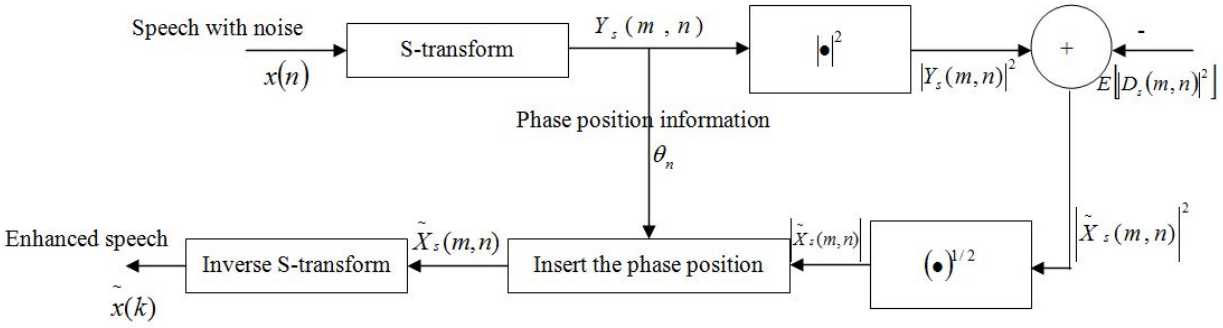


Figure 1. Principle diagram for discrete ST based spectral subtraction

A procedure of the spectral subtraction algorithm based on discrete ST is given as follows:

Step 1. Using (9), take the discrete ST of the speech signal $x(k)$ to obtain the coefficients $Y_s(m,n)$. $\theta_n = \varphi_n$, which can be computed from $Y_s(m,n)$.

Step 2. Computing $E[\left|D_s(m,n)\right|^2]$, take it as the noise power spectrum.

Step 3. Using (18) and (16) to get the spectrum $\widetilde{X}_s(m,n)$.

Step 4. Using (19), take the inverse ST of $\widetilde{X}_s(m,n)$ to obtain the enhanced speech signal $\widetilde{x}(k)$.

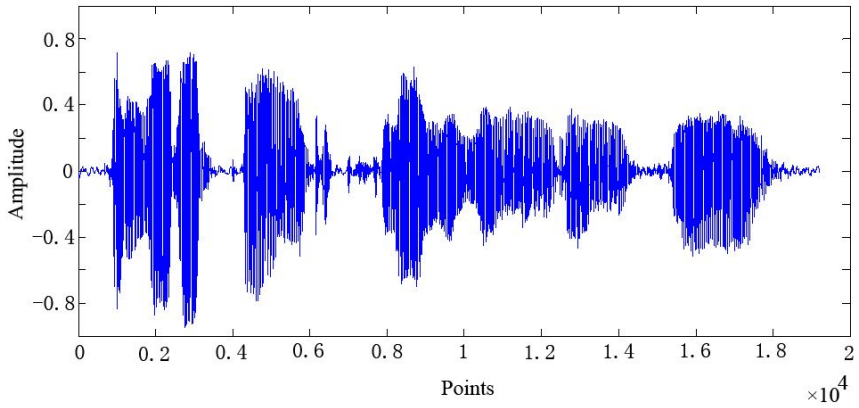
In order to evaluate the performance of the algorithm objectively, the segmental signal-to-noise ratio (SNR_{seg}) is usually used to measure the performance.

$$SNR_{seg} = \frac{1}{K}\sum_{k=0}^{K-1} 10\lg\left\{\sum_{n=0}^{N-1} x_k^2(n) \Big/ \sum_{n=0}^{N-1}\left[\tilde{x}_k(n) - x_k(n)\right]^2\right\} \tag{20}$$

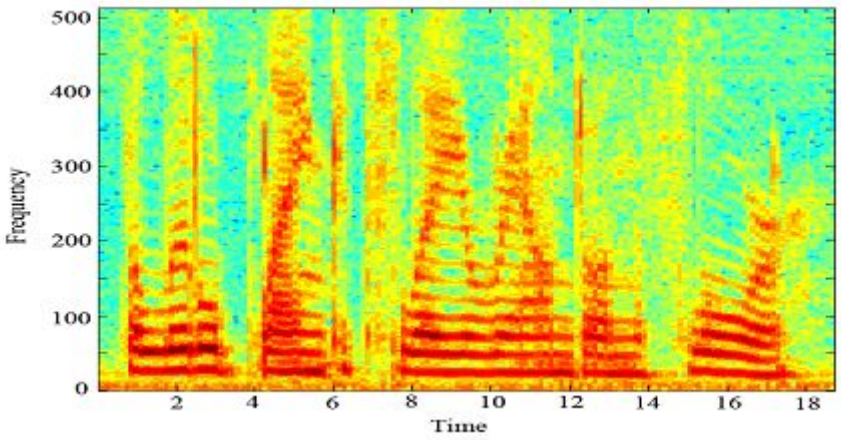
where $x_k(n)$ and $\tilde{x}_k(n)$ are respectively the pure speech signal and the output speech signal. N is the length of the speech frame and K is the number of frames.

IV. EXPERIMENT AND PERFORMANCE ANALYSIS

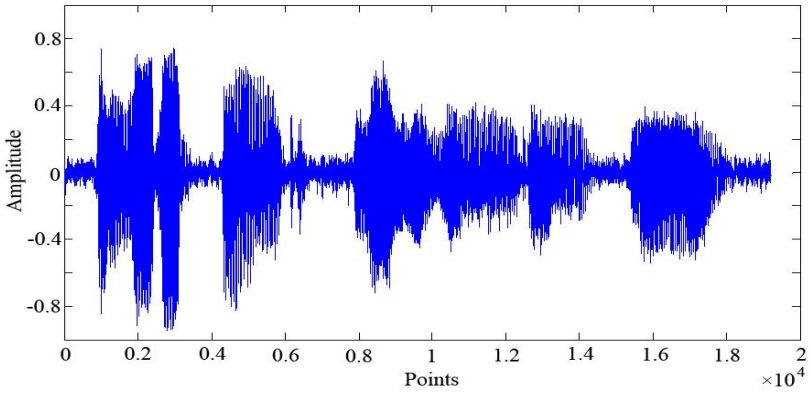
In order to verify the speech denoising effect and speech enhancement performance of the proposed algorithm, The Matlab is used for the simulation experiment. During the experiment, the statistical average in the absence of speech is used in the noise estimation as the power spectrum of the noise and the proposed algorithm is compared with the classical spectral subtraction method (SS) and the Gabor method in terms of performance. The standard speech signal used in the experiment is the pure speech sample of TIMIT database and the noise used is the classical Gaussian white noise. The sampling rate of the pure speech used in the experiment is 8kHz. Noise with certain signal-to-noise ratio (SNR) is added, and then the standard spectral subtraction method [5], the Gabor spectral subtraction method [24], the ST spectral subtraction method, the LMS spectral subtraction method [12] and the RLS spectral subtraction method [14] are applied to de-noise.



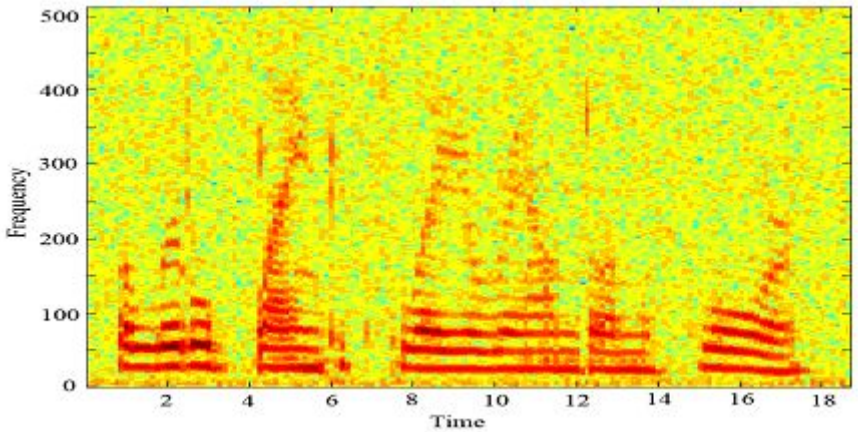
(a) Pure speech



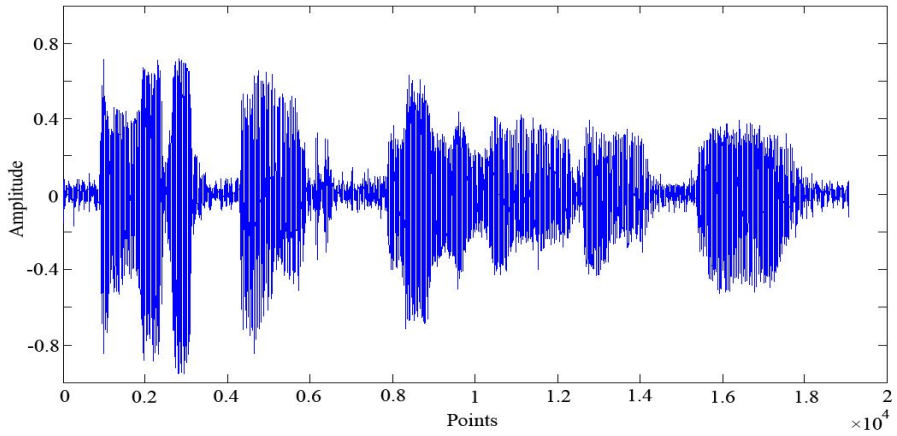
(b) Clear speech spectrum



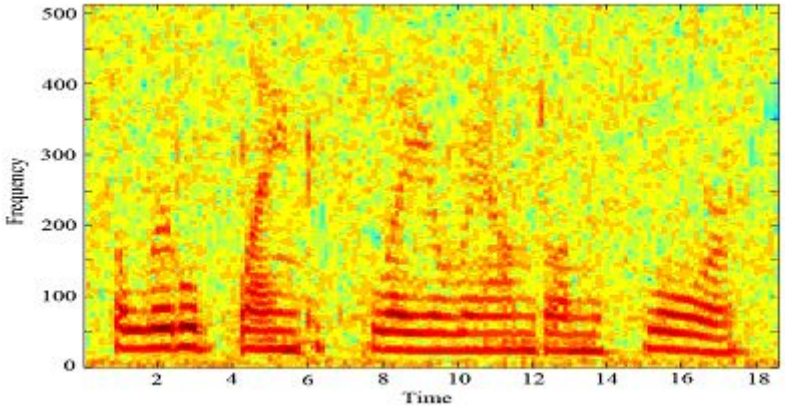
(c) Noisy speech



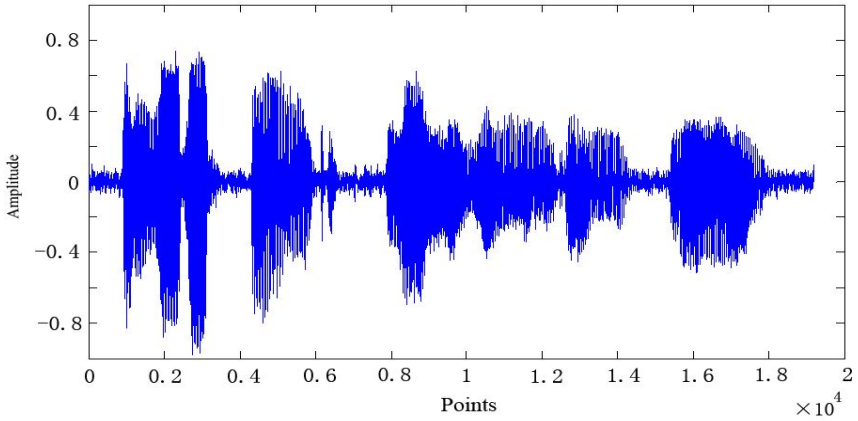
(d) Noisy speech spectrum



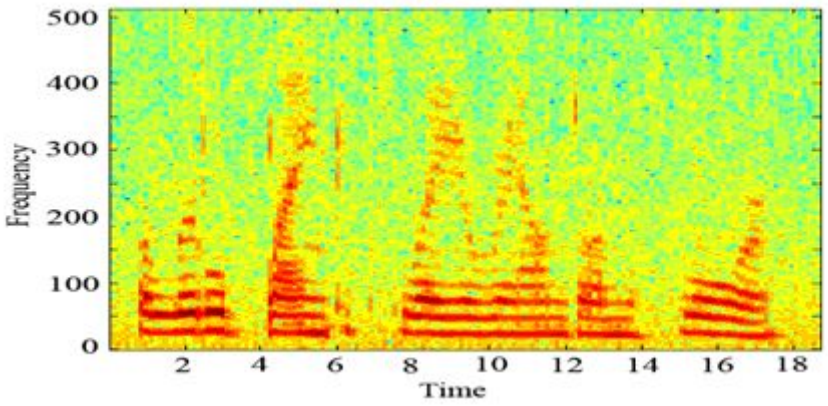
(e) Speech after de-noising by using the standard SS method.



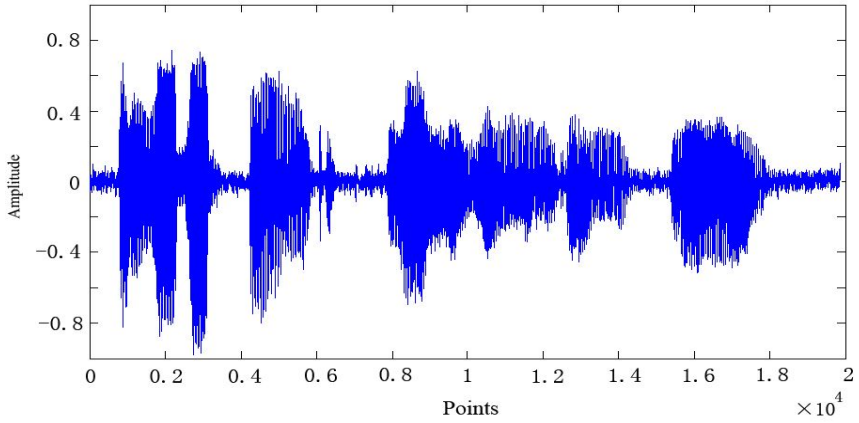
(f) Speech spectrum after standard SS method.



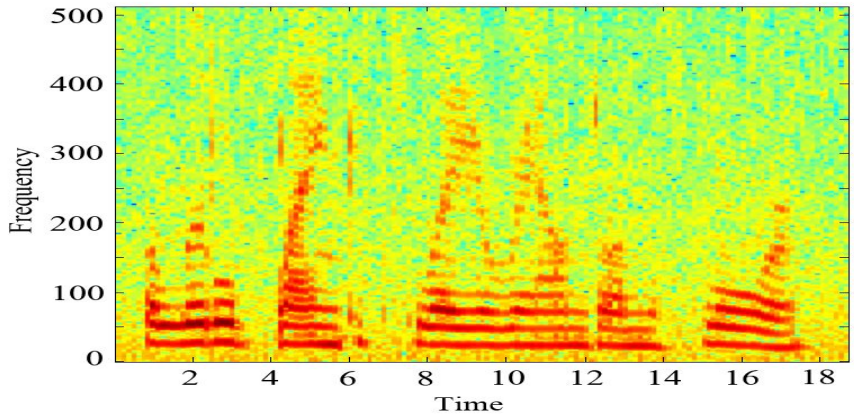
(g) Speech after de-noising by using the Gabor-SS method.



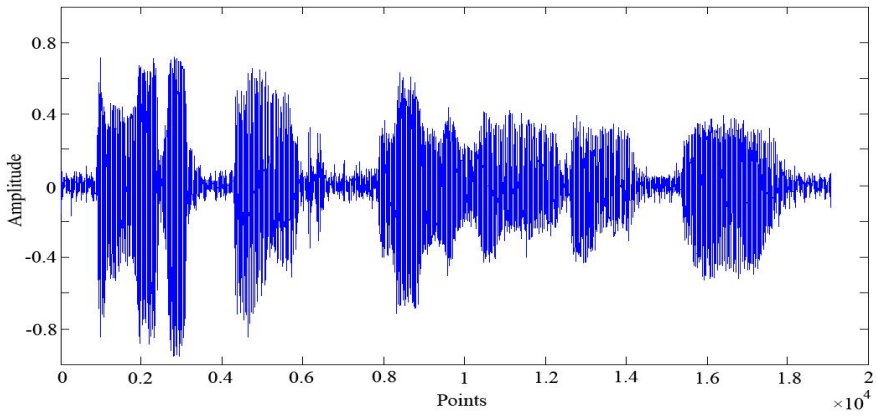
(h) Speech spectrum after de-noising by using the Gabor-SS method.



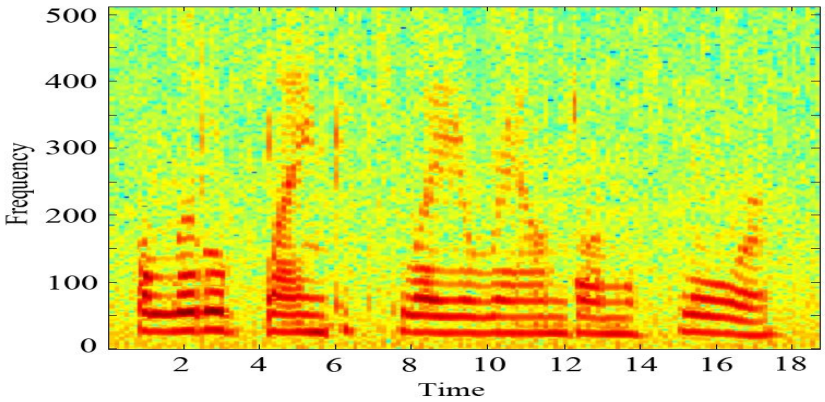
(i) Speech after de-noising by using the LMS method.



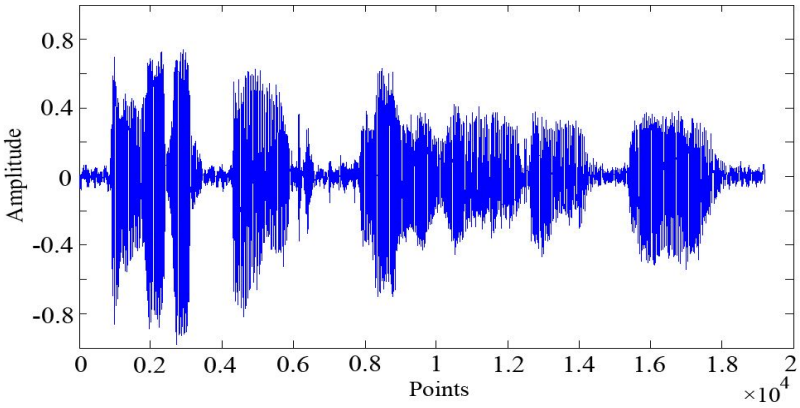
(j) Speech spectrum after de-noising by using the LMS method.



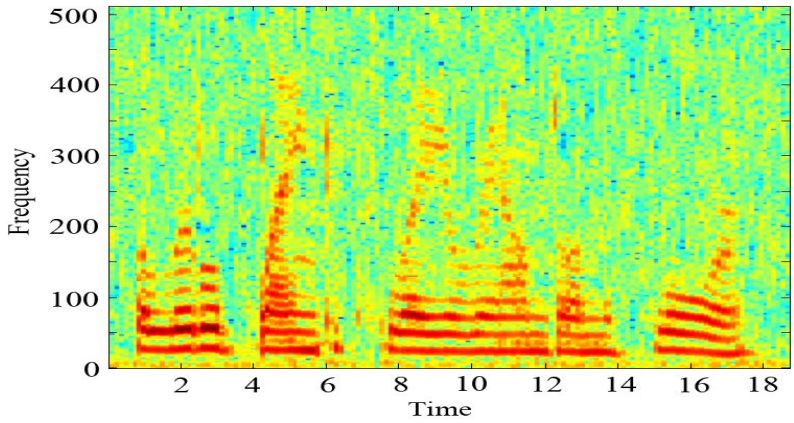
(k) Speech after de-noising by using the RLS method.



(l) Speech spectrum after de-noising by using the RLS method.



(m) Speech after de-noising by using the proposed algorithm.



(n) Speech spectrum after de-noising by using the proposed algorithm.

Figure 2. Comparison of simulation experiment results

Table 1. Denoising comparison of the five algorithms

Input SNR (dB)	Output SNR_{seg} (dB)				
	Boll-SS [5]	Gabor-SS[19]	LMS[12]	RLS[14]	DST-SS
5	1.1167	1.6818	1.8326	1.8415	2.3280
10	5.6391	6.3878	6.9214	7.0133	7.4936
15	10.8302	11.0793	11.8654	11.9104	12.3792
20	15.4403	16.1724	16.9105	17.1023	17.5056

The results of five different denoising algorithms are shown in (e), (g), (i) ,(k), (m) of Figure 2 and the corresponding spectrums are displayed in (f), (h), (j), (l), (n) of Figure 2. As shown in Figure 2(m) and Figure 2(n), the proposed algorithm has better de-noising effect as compared with the other algorithms. As shown in speech spectrum, the noise in the speech is greatly reduced. The objective comparison of the results of these five denoising algorithms are shown in Table 1.

As shown in Table 1, the segmental SNR obtained by the proposed DST-SS algorithm is obviously higher than those obtained by the classic spectral subtraction method, the Gabor spectral subtraction method, the LMS spectral subtraction method and the RLS spectral subtraction method in the same input SNR of speech. The experiment also indicates that the

proposed DST-SS algorithm can maintain well the detailed characteristics of the original speech after de-noising and the distortion is reduced, and verifies that the proposed algorithm is effective.

V. CONCLUSION

The conventional spectral subtraction method uses the window Fourier transform or Gabor transform, where the window function width keeps unchanged during transform so that it has a fixed time-frequency accuracy in the time-frequency domain. Since the ST has the advantages of both wavelet transform and shor-time Fourier transform, it has a changeable time-frequency accuracy in the joint time-frequency domain. The discrete ST based spectral subtraction method proposed in this paper is obviously better than the spectral subtraction method based on linear time-frequency transform. The results of large quantity of comparison experiments show that the proposed DST-SS spectral subtraction method has a distinct effect on the speech enhancement, which greatly reduces the noise, improves the speech quality, and effectively enhance the de-noising ability.

ACKNOWLEDGEMENTS

This work was supported by the National Nature Science Foundation of China (Grant No. 61071169, Grant No. 61372137), the Natural Science Foundation of Anhui Provincial Education Department (No.KJ2015A164), the Natural Characteristics Specialty Construction of Anhui Province (No. 2014tszy028) and Key Disciplines at Hefei University (No. 2014xk06).

REFERENCES

[1] J. Ying, G. Li, and Z. Wang. “A novel approach to speech signal synthesis”, In Proc. International Conference on Audio, Language and Image Processing, 2008, pp. 680–685, 2008.

[2] C. Z. Wu, M. T. Guo, S. B. Xiong, and Q. Li. “Studies on a speech signal testing method based on adaptive filtering feedback technology”, In Proc. International Conference on Signal Processing, 2000, vol. 1, pp. 543–546, 2000.

[3] T. F. Quatieri. “Discrete-time speech signal processing: principles and practice”, Pearson Education India, 2002.

[4] A. Kusumoto, T. Arai, K. Kinshita, and N. Hodoshima. “Modulation Enhancement of speech by a preprocessing algorithm for improving intelligibility in reverberant environments”, Speech Communication, vol. 45, no. 2, pp. 101–113, 2005.

[5] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Transactions on Acoustics Speech and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

[6] M. Berouti, R. Schwartz, and J. Makhoul. “Enhancement of speech corrupted by acoustic noise”, In Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1979, pp. 208–211, 1979.

[7] Y. Ephriam, H. L. Van Trees. “A signal subspace approach for speech enhancement”, In Proc. International Conference on Acoustic, Speech and Signal processing, 1993, Detroit, MI, USA, vol. 2, pp. 355–358, 1993.

[8] F. Huang, T. Lee, W. B. Kleijn, et al., “A method of speech periodicity enhancement using transform-domain signal decomposition”, Speech Communication, vol. 67, pp.102-112, 2015.

[9] MAA El-Fattah, MI Dessouky, AM Abbas, et al., “Speech enhancement with an adaptive Wiener filter”, International Journal of Speech Technology, vol. 17, no.1, pp.53-64, 2014.

[10] C. Caraiscos, B. Liu. “A roundoff error analysis of the LMS adaptive algorithm”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 32, no. 1, pp. 34-41, 1984.

[11] V. Udayashankara. “Fast Recursive DCT-LMS Speech enhancement For Performance Enhancement Of Digital Hearing Aid”, Academic Open Internet Journal, no.18, 2006.

[12] B. M. Soujanya, C. R. Rao, D. V. L. N. Sastry. “Speech Enhancement using Combinational Adaptive LMS Algorithms”, International Journal of Advanced Computer Research, vol. 18, no. 5, 2015.

[13] J. Chen. “A study of speech enhancement based on RLS filter”, Computer Applications & Software, (in Chinese), vol. 19, no. 10, pp. 40-42, 2002.

[14] Pogula Rakesh, T. Kishore Kumar. “A Novel RLS Based Adaptive Filtering Method for Speech Enhancement”, World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Electronics and Communication Engineering, vol. 9, no. 2, pp. 624-628, 2015.

[15] Zhang Qiu-yu，Liu Yang-wei，Huang Yi-bo，et al., “Perceptual Hashing Algorithm for Speech Content Identification Based on Spectrum Entropy in Compressed Domain”, International Journal on Smart Sensing and Intelligent Systems, vol. 7, no. 1, pp. 283–300, 2014.

[16] Y. Ghanbari, M. Karami. “Spectral subtraction in the wavelet domain for speech enhancement”, International Journal of Software & Information Technology, no.1, pp. 26–30, 2004.

[17] Y. Ghanbari, M. Karami, and B. Amelifard. “Improved multiband spectral subtraction method for speech enhancement”, In Proc. 6th IASTED International Conf. on Signal and Image Processing, USA, pp. 225–230, 2004.

[18] R. G. Stockwell, L. Mansinha, and R. P. Lowe. “Localization of the complex spectrum: the S-transform”, IEEE Transactions on Signal Processing, vol. 44, no. 4, pp. 998–1001, 1996.

[19] I. Djurović, E. Sejdić and J. Jiang. “Frequency-based window width optimization for S-transform”, AEU International Journal of Electronics and Communications, vol. 62, no. 4, pp. 245–250, 2008.

[20] E. Sejdić, I. Djurović and J. Jiang. “A window width optimized S-transform”, EURASIP Journal on Advances in Signal Processing, 2008:59, 2008.

[21] Wang Lin, Meng Xiaofeng, “An adaptive Generalized S-transform for instantaneous frequency estimation”, Signal Processing, vol. 91, no. 8, pp.1876–1886, 2011.

[22] Zenton Goh, Kah-Chye Tan and B. T. G. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction”, IEEE Transactions on Speech and Audio Processing, vol. 6, no. 3, pp.287-292, 1998.

[23] K. Paliwal, K. Wójcicki and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain”, Speech Communication, vol. 52, no. 5, pp. 450–475, 2010.

[24] Zhou Jian， Huang Cheng， Zhang Man， et al., “Whisper denoising in joint time-frequency domain based on real valued discrete Gabor transform”, Applied Mechanics and Materials，vol. 152, no. 8, pp.1091-1096, 2012.

[25] L. Debnath, F. A. Shah, “The Gabor Transform and Time–Frequency Signal Analysis”, Wavelet Transforms and Their Applications, Birkhäuser Boston, pp.243-286, 2015.