

Speech Enhancement Algorithms in Vehicle Environment

Chunli Wang^{a,*}, Yuchen Li^a, and Huaiwei Lu^b

^a*School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China*

^b*School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou, 730070, China*

Abstract

In the actual driving process, the driver is in a complex noise interference environment of the vehicle's own mechanical vibration, the passenger dialogue inside the vehicle, and the sound of other equipment. In order to improve driving efficiency and ensure driving safety, the operation of the vehicle equipment is precisely controlled by the voice control system. Aiming at the residual music noise in traditional spectral subtraction, the improved multi-window spectrum estimation algorithm is applied to improve the estimation accuracy of a priori SNR (signal-to-noise ratio). The experimental results show that the algorithm significantly eliminates the music noise. In the case of low SNR, the signal-to-noise ratio gain is improved by 0.64dB. The waveform similarity and speech naturalness are improved after speech enhancement. Furthermore, the current single-microphone voice de-reverberation technology only takes advantage of the information of time domain and frequency domain with the spatial information limitedly utilized, resulting in a difficulty of achieving a better de-reverberation effect. In light of these insufficiencies, we combine the de-reverberation technique with complex cepstrum blind deconvolution, and a simulation experiment is carried out according to the subjective and objective evaluation indexes of the waveform and the effect of de-reverberated voice, proving that the optimized algorithm improves the intelligibility of the de-reverberated voice.

Keywords: vehicle environment; speech enhancement; noise; spectral subtraction; multi-window spectrum estimation

(Submitted on September 30, 2019; Revised on November 2, 2019; Accepted on November 12, 2019)

© 2019 Totem Publisher, Inc. All rights reserved.

1. Introduction

With the maturity of digital signal processing theory, speech enhancement technology has developed into an important branch of speech signal processing, widely used in areas such as voice control systems, teleconferencing, mobile communication, entertainment systems, multimedia applications, smart home appliances, scene recording, and military eavesdropping [1-5]. In the era of automation and intelligence, the number of cars has exploded. This has brought great convenience to people's lives as well as hidden dangers. A car-mounted voice control system can prevent drivers from using their hands to control related equipment in the car. It can not only improve the driving efficiency of the driver but also ensure the safety of driving [6-8].

2. Acoustic Characteristics of the Vehicle Environment

2.1. Vehicle Phonetic Features

In the vehicle environment, the voice signal is usually an intermittent command voice, which can be in two states:

(1) Noise segment. The driver does not issue commands to the voice control system during this time, and the microphone array system only collects ambient noise. Since the desired speech is not contained, the speech signal in this state can be suppressed [9-12].

(2) Voice segment. During this time, the driver issues a control command to the voice control system. The array system collects a mixed signal of desired speech and noise, and it needs to suppress noise and enhance speech [13-16].

* Corresponding author.

E-mail address: wcl@mail.lzjtu.cn

2.2. Vehicle Noise Characteristics

The noise in the vehicle environment is mainly composed of correlated noise and non-correlated noise. The former refers to the dialogue between the passengers in the vehicle, which can be suppressed in the spatial direction of the sound source through the beam forming process. The latter includes the engine vibration, air disturbance, road friction, and so on. This paper selects a set of pure speech and noise in the vehicle environment collected by Gannot in his related research [17-19] to analyze the distribution of their energy spectrum in the frequency domain, as shown in Figure 1.

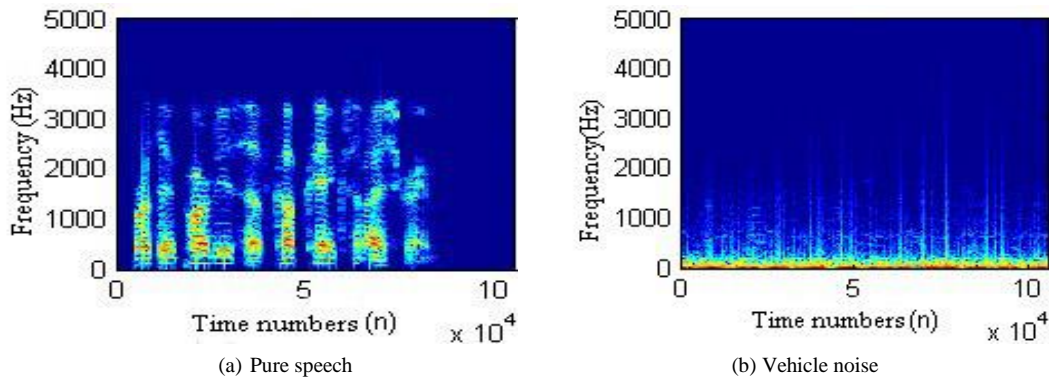


Figure 1. Energy spectrums of two signals

It can be seen from Figure 1 that the energy of pure speech in the vehicle environment is mainly concentrated in the high frequency band, while the non-related noise energy is mainly concentrated in the low frequency band.

2.3. Reverb Characteristics in the Vehicle Environment

In the field of acoustics, reflected waves with a delay time of less than 50 ms are called reverberation [20-23]. In the relatively closed car space, the voice command issued by the driver will be reflected by the inner wall of the medium, such as the car body. The reflected signal is similar to the original voice in waveform, while the amplitude of the reflected signal is larger, and there are some phase changes. The sound received by the microphone is the sum of the direct sound and the reflected sound arriving through all other paths, i.e., reverberation, as shown in Figure 2. The masking effect occurs between the syllables of the aliased speech, so the voice control system cannot receive clear and accurate instructions, resulting in misoperation.

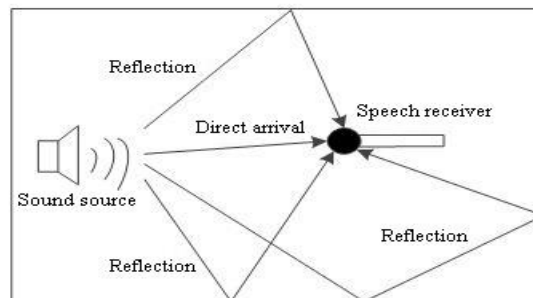


Figure 2. Schematic diagram of the reverberation process in the vehicle environment

3. Simulation of Vehicle Environment Speech Enhancement Algorithm based on Spectral Subtraction

The speech signal can be regarded as a short-term stationary signal for a short period of time (usually 10-30 ms). The noise spectrum of some small time can be calculated by windowing and framing. Then, subtract the noise spectrum with the noisy speech spectrum to obtain the pure desired speech.

It is assumed that the time sequence of the speech signal is $x(n)$, $x_i(m)$ is the i^{th} frame speech signal obtained by windowing and framing preprocessing, and the frame length is N . After performing DFT on any frame signal, we can obtain:

$$X_i(k) = \sum_{m=0}^{N-1} x_i(m) \exp(j \frac{2\pi mk}{N}), k = 0, 1, \dots, N-1 \quad (1)$$

The amplitude and phase angle of $x_i(k)$ is obtained from Equation (1), the amplitude is $|x_i(k)|$, and the phase angle is

$$X_{angle}^i(k) = \arctan \left[\frac{\text{Im}(X_i(k))}{\text{Re}(X_i(k))} \right] \quad (2)$$

It is known that the time length of the leading no speech segment (noise segment) [24] is IS , and the corresponding number of frames is NIS , so the average energy of the noise segment is

$$D(k) = \frac{1}{NIS} \sum_{i=1}^{NIS} |X_i(k)|^2 \quad (3)$$

Furthermore, the spectral subtraction algorithm is

$$\left| \hat{X}_i(k) \right|^2 = \begin{cases} |X_i(k)|^2 - a \times D(k), & |X_i(k)|^2 \geq a \times D(k) \\ b \times D(k), & |X_i(k)|^2 < a \times D(k) \end{cases} \quad (4)$$

The amplitude is obtained after spectral subtraction, the phase angle value obtained by Equation (1) is combined, and then the desired speech sequence after spectral subtraction can be obtained by fast Fourier transform. The principle of spectral subtraction is shown in Figure 3.

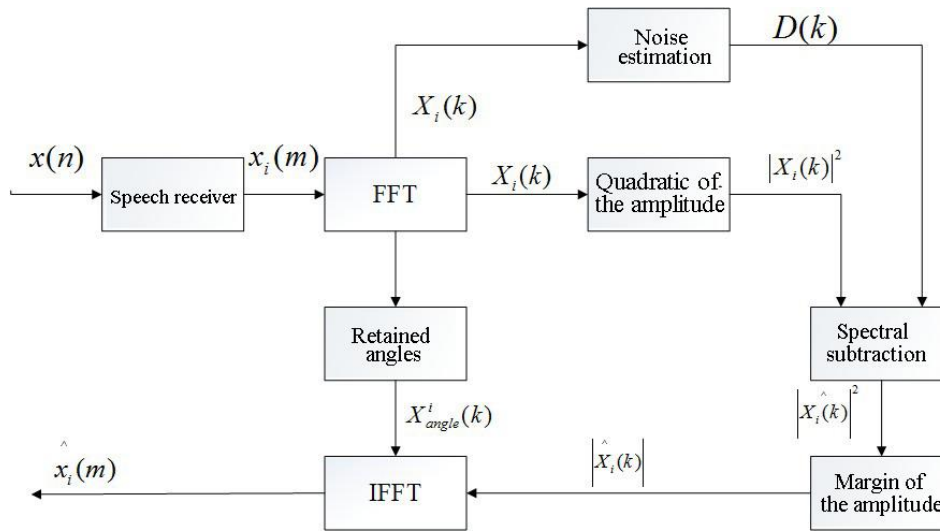


Figure 3. Schematic diagram of the basic spectrum subtraction

Basic spectral subtraction uses the waveforms and signal-to-noise ratio before and after speech enhancement to measure the speech enhancement effect. Figure 4 is a waveform diagram of the sound spectrum after denoising (sound source 1) by the basic spectral subtraction, Table 1 is the signal-to-noise ratio and gain table before and after speech enhancement, and Table 2 is the objective evaluation value based on LLR obtained before and after the basic spectral subtraction speech enhancement. The values are obtained by the short-term definition index evaluation criteria.

The average signal-to-noise ratio gain of the three sets of data is also SNR Original = 8.3681dB. There is obvious "music noise" in the speech after denoising by the basic spectral subtraction, which causes some interference to the speech recognition of the car voice control system. Therefore, it is necessary to study an optimization algorithm that further reduces music noise and improves signal to noise ratio gain.

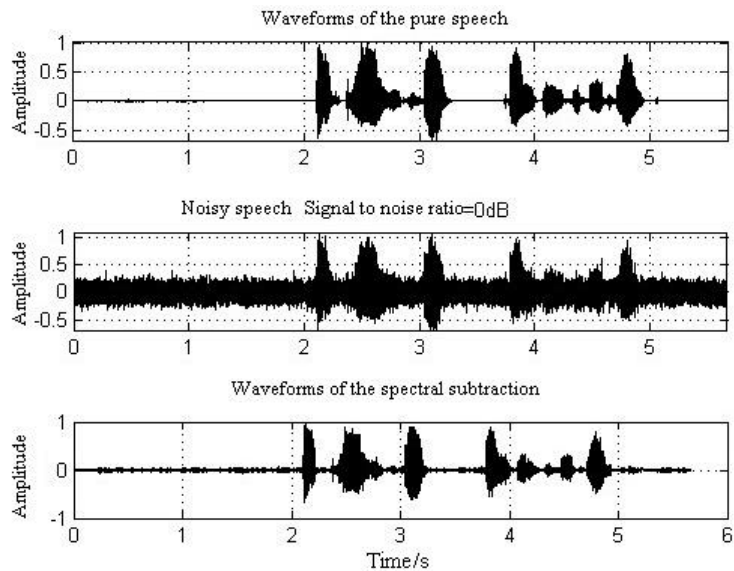


Figure 4. Waveform diagram before and after basic spectral subtraction speech enhancement

Table 1. Spectral subtraction denoising signal to noise ratio change

Sound source	SNR1/dB	SNR2/dB	SNR3/dB
Sound source 1	0	9.0759	9.0759
Sound source 2	0	7.8413	7.8413
Sound source 3	0	8.1873	8.1873

Table 2. Intelligibility and resolution of basic spectrum subtraction speech enhancement

Sound source	Sound source 1	Sound source 2	Sound source 3
LLR	1.5075	1.2743	1.3853
AI-ST	0.0812	0.0760	0.0949

4. Multi-Window Spectrum Estimation Algorithm

The remaining "music noise" in the above speech spectrum subtraction is due to the larger variance generated by the a priori signal-to-noise ratio estimation when noise estimation is performed in the "silent segment" of the speech. Therefore, the accuracy of the noise power spectrum estimation directly affects the quality of the signal obtained after speech enhancement. In order to reduce the "music noise", multi-window spectrum estimation uses multiple orthogonal data windows to directly map the same data sequence. Then, it averages the obtained multiple sets of data to obtain the final spectral estimate, so the estimated variance of this method is smaller.

The multi-window spectrum is defined as follows:

$$S^{mt}(w) = \frac{1}{L} \sum_{k=0}^{L-1} S_k^{mt}(w) \tag{5}$$

Where L represents the number of data windows and S^{mt} represents the spectrum of the k^{th} data window.

$$S_k^{mt}(w) = \left| \sum_{n=0}^{N-1} a_k(n)x(n)e^{-jnw} \right|^2 \tag{6}$$

Where $x(n)$ represents the data sequence, N represents the sequence length, $a_k(n)$ represents the k^{th} data window, and the multiple data windows are orthogonal to each other and satisfy the following formula:

$$\begin{cases} \sum a_k(n)a_j(n) = 0, & k \neq j \\ \sum a_k(n)a_j(n) = 1, & k = j \end{cases} \tag{7}$$

4.1. Multi-Window Spectrum Estimation Algorithm Implementation

Using the MATLAB simulation platform, the PMTM function is called to calculate the power spectral density of the multi-window spectrum, thereby obtaining the gain factor in the spectral subtraction method and realizing the operation of spectral subtraction speech enhancement. The specific steps are as follows:

(1) The overlapping frame method is used to frame the noisy speech $x(n)$ by windowing to obtain $x_i(m)$.

(2) Perform FFT on $x_i(m)$ to obtain the amplitude spectrum $|x_i(k)|$ and the phase spectrum $\theta_i(k)$ respectively, and the smoothing process is performed at the same time between adjacent frames to calculate the average amplitude spectrum $|x(_)_i(k)|$.

$$|\bar{X}_i(k)| = \frac{1}{2M+1} \sum_{j=-M}^M |X_{i+j}(k)| \quad (8)$$

There are $2M+1$ frames to be averaged here, and the actual situation usually involves three frames averaged together, that is, M takes 1.

(3) Use the PMTM function to obtain the framed signal $x_i(m)$ and find the multi-window power spectral density $P(k, i)$ (where i represents the i^{th} frame and k represents the k^{th} spectral line).

$$P(k, i) = \text{PMTM}[x_i(m)] \quad (9)$$

(4) Perform smoothing between adjacent frames on the multi-window power spectral density estimate to calculate the smoothed power spectral density $P_y(k, i)$.

$$P_y(k, i) = \frac{1}{2M+1} \sum_{j=-M}^M P(k, i+j) \quad (10)$$

(5) Calculate the average power spectral density value $P_n(k)$ of the noise based on the number of frames occupied by the leading speechless segment (NIS).

$$P_n(k) = \frac{1}{NIS} \sum_{i=1}^{NIS} P_y(k, i) \quad (11)$$

(6) Calculate the spectral subtraction factor as follows:

$$g(k, i) = \begin{cases} (P_y(k, i) - \alpha P_n(k)) / P_y(k, i), & P_y(k, i) - \alpha P_n(k) \geq 0 \\ \beta P_n(k) / P_y(k, i), & P_y(k, i) - \alpha P_n(k) < 0 \end{cases} \quad (12)$$

In the above formula, α represents the over-subtraction factor and β represents the gain compensation factor.

A proper value selection of α can effectively reduce the music noise, but if the value of α is too large, it may distort the speech.

(7) Obtain the amplitude spectrum [25] after spectral subtraction based on the gain factor $g(k, i)$ and the average amplitude spectrum $|x(_)_i(k)|$.

$$|\hat{X}_i(k)| = g(k, i) \times |\bar{X}_i(k)| \quad (13)$$

IFFT is performed by the amplitude spectrum $|\hat{X}_i(k)|$ and the phase spectrum $\theta_i(k)$ of step 2, and $|x(_)_i(k)|$ is transformed into the time domain to obtain a noise-reduced speech signal $\hat{x}_i(m)$.

$$\hat{x}_i(m) = IDFT[|\hat{X}_i(k)| \exp[j\theta_i(k)]] \tag{14}$$

According to the above steps, the improved spectral subtraction implementation process based on multi-window estimation can be obtained as shown in Figure 5. The waveform diagram and the signal-to-noise ratio before and after speech enhancement are used to measure the effect of speech enhancement [26-27]. Figure 6 is a waveform diagram of the improved algorithm before and after denoising (sound source 1), and Table 3 is the signal-to-noise ratio and gain table before and after speech enhancement. Table 4 shows the objective evaluation values based on LLR obtained before and after the improved algorithm, as well as the values obtained by the short-term definition index evaluation criteria [28].

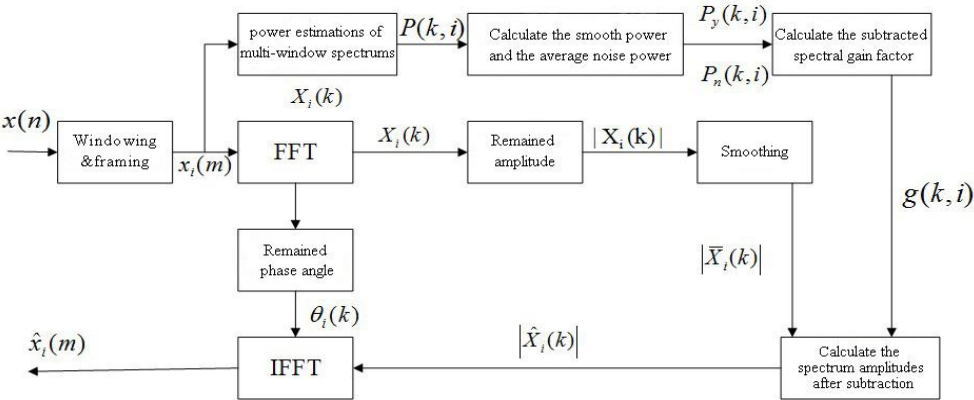


Figure 5. Flow chart of improved spectral subtraction for multi-window spectrum estimation

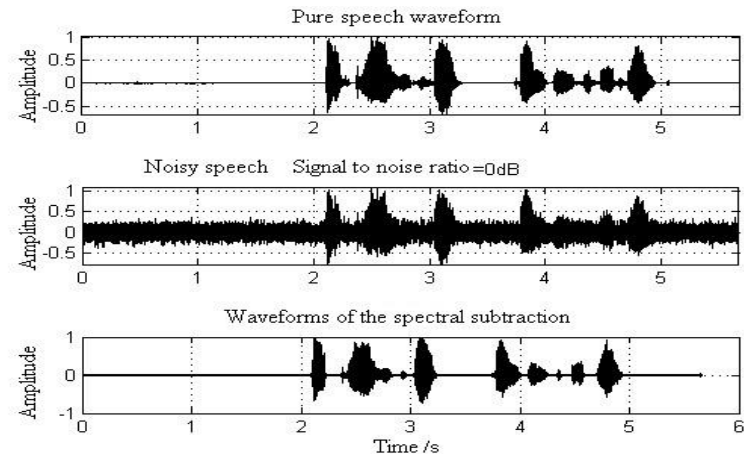


Figure 6. Speech waveform before and after multi-window spectrum estimation algorithm

Table 3. Signal-to-noise ratio gain before and after improved spectral subtraction signal enhancement

Sound source	SNR1/dB	SNR2/dB	SNR3/dB
Sound source 1	0	9.9574	9.9574
Sound source 2	0	8.1679	8.1679
Sound source 3	0	8.9249	8.9249

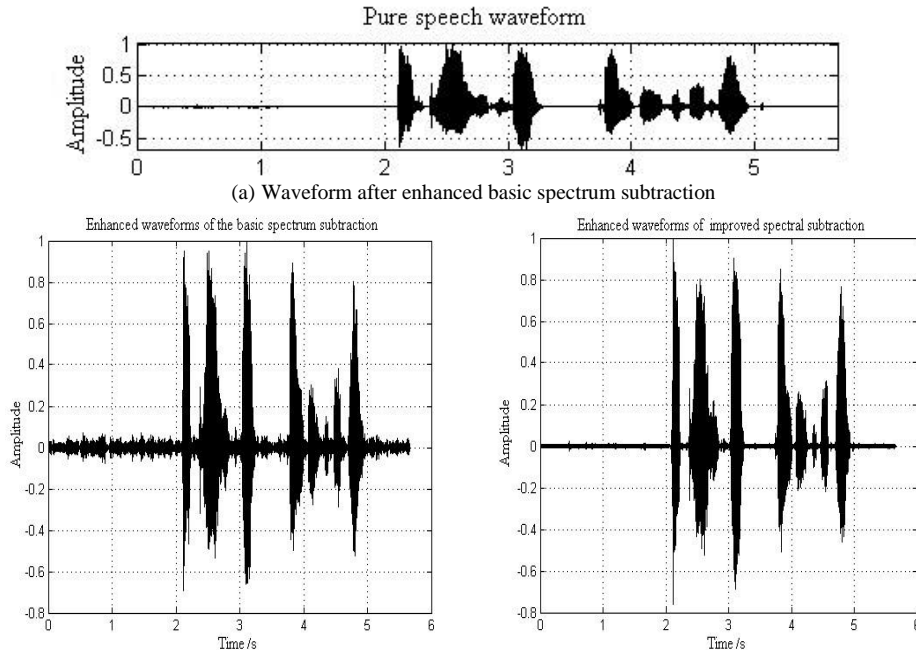
Table 4. Intelligibility and clarity of the improved spectral subtraction speech enhancement

Sound source	Sound source 1	Sound source 2	Sound source 3
LLR	1.6217	1.3848	1.4048
AI-ST	0.0839	0.0965	0.09881

The spectral subtraction based on multi-window spectrum estimation from these three sets of data is SNRImproved = 9.0167dB.

4.2. Simulation Results Analysis

The simulation results under different input signal to noise ratio environments are shown in Figures 7 and 8.



(b) Waveform of improved spectrum subtraction after enhancement
Figure 7. Waveform comparison before and after speech enhancement of two algorithms

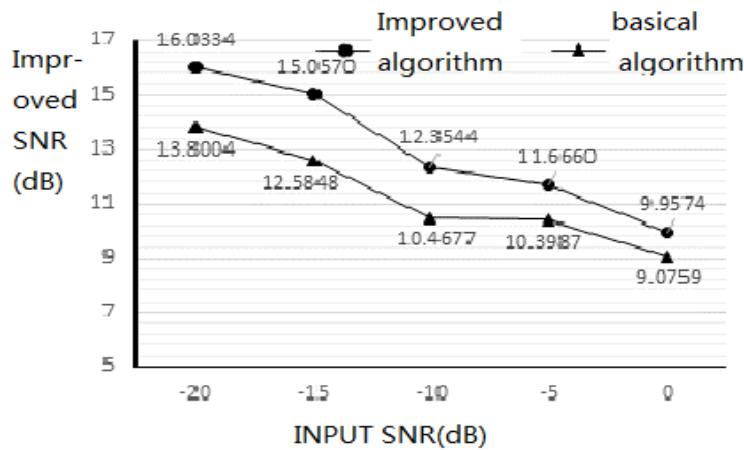


Figure 8. Comparison of the denoising performance of two spectral subtractions in different SNR environments

As shown in Figure 8, for the same sound source in the same overlapped noise environment, the quality of the signal waveform obtained by the improved algorithm is significantly better than that of the traditional spectral subtraction method [29]. Secondly, from the intuitive auditory sensory analysis, the music noise in the speech signal is obviously eliminated with the improved algorithm. As shown in Figure 8, from the robustness analysis of the communication system, the improved algorithm has better speech enhancement effects than the traditional spectral subtraction in various low SNR environments (the input signal SNR is less than 0dB). From the average signal-to-noise ratio gain analysis obtained from multiple sets of data, the average signal-to-noise ratio gain ($SNR_{Improved} = 9.0167$) of the improved algorithm is significantly higher than that of the traditional spectral subtraction ($SNR_{Original} = 8.3681$).

5. Conclusions

In this paper, we study the classical and improved spectral subtractions and multi-windows spectrum estimation algorithms in vehicle environments. Firstly, the SNR and similarity of the signal waveform obtained by the improved multi-windows spectral estimation algorithm is significantly better than those of the traditional spectral subtraction method. Secondly, since the traditional method has very limited ability in controlling the music noise in the enhanced speech, the multi-windows spectrum estimation algorithm has greatly improved the quality of received speech. The subjective and objective evaluation experiments of the de-noising effect verify that the method based on multi-windows spectral estimation improves the overall

noise inhibiting ability of the car-counted voice controlling system.

Acknowledgments

This work was supported by the University Innovation Ability Enhancement Project of Gansu Province (No. 2019B-058), the Construction Science and Technology Soft Science Project of Gansu Province (No. JK2019-31), the Postgraduate Teaching Reforming Project (No. 1600120131), the Undergraduate Teaching Reforming Project (No. JGY201923), and the Innovation and Entrepreneurship Reforming Project (No. 2019CXCKCY09) of Lanzhou Jiaotong University.

References

1. T. N. Sainath, B. Kingsbury, A. Mohamed, G. E. Dahl, G. Saon, H. Soltau, et al., "Improvements to Deep Convolutional Neural Networks for LVCSR," *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on IEEE*, pp. 315-320, 2013
2. O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 10, pp. 1533-1545, 2014
3. O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring Convolutional Neural Networks Structures and Optimization Techniques for Speech Recognition," *Interspeech*, pp. 3366-3370, 2013
4. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors," *Computer Science*, Vol. 3, No. 4, pp. 220-223, 2012
5. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1949-1958, 2014
6. N. Srivastava, "Improving Neural Networks with Dropout," University of Toronto, 2013
7. H. Che, B. Li, and Z. Chen, "Automatic Extracting Event-Related Potentials Within Several Trials using Infomax ICA Algorithm," *Journal of Scientific and Industrial Research*, Vol. 71, No. 7, pp. 468-473, 2012
8. G. Qian and P. Wei, "Stability Analysis of Complex ICA by Negentropy Maximization: A Unique Perspective," *Neural Computing*, Vol. 214, pp. 80-85, 2016
9. A. Narayanan and D. L. Wang, "Joint Noise Adaptive Training for Robust Automatic Speech Recognition," in *Proceedings of ICASSP 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2504-2508, 2014
10. M. Djendi and P. Scalart, "Reducing over- and under-Estimation of the Prior SNR in Speech Enhancement Techniques," *Digital Signal Processing*, Vol. 32, pp. 124-136, 2014
11. R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-Noise-Free Speech Enhancement based on Optimized Iterative Spectral Subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 124-136, 2012
12. T. Mourad, L. Salhi, B. N. Mouhamed, and A. Cherif, "Wiener Filtering Application in the Bionic Wavelet Domain for Speech Enhancement," *International Journal of Advancements in Computing Technology*, Vol. 4, No. 2, pp. 146-160, 2012
13. K. Paliwal, B. Schwerin, and K. Wojcicki, "Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Modulation Magnitude Estimator," *Speech Communication*, Vol. 54, No. 2, pp. 282-305, 2012
14. Y. Xu, J. Du, L. Dai, and C. Lee, "An Experimental Study on Speech Enhancement based on Deep Neural Networks," *Signal Processing Letters*, Vol. 21, No. 1, pp. 65-68, 2014
15. M. A. B. Messaoud and A. Bouzid, "Speech Enhancement based on Wavelet Transform and Improved Subspace Decomposition," *Journal of the Audio Engineering Society*, Vol. 63, No. 12, pp. 990-100, 2016
16. X. M. Liu, C. F. Ban, and X. R. Feng, "A Short-Time Spectrum Estimation Algorithm of Speech Enhancement under the Distortion Control," *Journal of Xi'an Jiaotong University*, Vol. 8, pp. 1-16, 2011
17. Y. Zhang and Y. Zhao, "Real and Imaginary Modulation Spectral Subtraction for Speech Enhancement," *Speech Communication*, Vol. 55, No. 4, pp. 509-522, 2013
18. H. Ding, Y. Soon, and C. K. Yeo, "A DCT-based Speech Enhancement System with Pitch Synchronous Analysis," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 8, pp. 2614-2623, 2011
19. Y. S. Xia and J. Wang, "Low-Dimensional Recurrent Neural Network-based Kalman Filter for Speech Enhancement," *Neural Networks*, Vol. 67, pp. 131-139, 2015
20. A. Saadoune, A. Amrouche, and S. A. Selouani, "MCRA Noise Estimation for KLT-VRE-based Speech Enhancement," *International Journal of Speech Technology*, Vol. 16, No. 3, pp. 333-339, 2013
21. L. E. E. Kyungsun and K. Minseok, "Visual Speech Recognition using Weighted Dynamic Time Warping," *IEICE Transactions on Information and Systems*, Vol. 98, No. 7, pp. 1430-1433, 2015
22. H. Ding, T. Lee, I. Y. Soon, C. K. Yeo, P. Dai, and G. Dan, "Objective Measures for Quality Assessment of Noise-Suppressed Speech," *Speech Communication*, Vol. 71, pp. 62-73, 2015
23. A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, et al., "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on IEEE*, pp. 2332-2337, 2014
24. A. Stark and K. Paliwal, "Use of Speech Presence Uncertainty with MMSE Spectral Energy Estimation for Robust Automatic Speech Recognition," *Speech Communication*, Vol. 53, No. 1, pp. 51-61, 2011
25. E. H. E. Bouchikhi, V. Choqueuse, and M. E. H. Benbouzid, "Current Frequency Spectral Subtraction and its Contribution to Induction Machines' Bearings Condition Monitoring," *IEEE Transactions on Energy Conversion*, Vol. 28, No. 1, pp. 135-144, 2013

26. Y. Tu, Y. Lin, J. Wang, and J. -U. Kim, "Semi-Supervised Learning with Generative Adversarial Networks on Digital Signal Modulation Classification," *CMC-Computers Materials and Continua*, Vol. 55, No. 2, pp. 243-254, 2018
27. Y. Lin, Y. Li, X. Yin, and Z. Dou, "Multisensor Fault Diagnosis Modeling based on the Evidence Theory," *IEEE Transactions on Reliability*, Vol. 67, No. 2, pp. 513-521, 2018
28. Y. Lin, C. Wang, J. Wang, and Z. Dou, "A Novel Dynamic Spectrum Access Framework based on Reinforcement Learning for Cognitive Radio Sensor Networks," *Sensors*, Vol. 16, No. 10, pp. 1-22, 2016
29. Y. Lin, X. Zhu, Z. Zheng, Z. Dou, and R. Zhou, "The Individual Identification Method of Wireless Device based on Dimensionality Reduction and Machine Learning," *Journal of Supercomputing*, Vol. 5, pp. 1-18, 2017

Chunli Wang received her M.S degree in communication and information systems from Lanzhou Jiaotong University in 2008. She is currently a teacher in the Department of Electron and Information Engineering at Lanzhou Jiaotong University. Her research interests include communication and speech signal processing.

Yuchen Li received his B.S degree in communication engineering from Lanzhou Jiaotong University in 2018.

Huaiwei Lu is currently a professor at Lanzhou Jiaotong University. His research interests include signal processing.