World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:11, No:11, 2017

A Sparse Representation Speech Denoising Method Based on Adapted Stopping Residue Error

Qianhua He, Weili Zhou, Aiwu Chen

***Abstract*—**A sparse representation speech denoising method based on adapted stopping residue error was presented in this paper. Firstly, the cross-correlation between the clean speech spectrum and the noise spectrum was analyzed, and an estimation method was proposed. In the denoising method, an over-complete dictionary of the clean speech power spectrum was learned with the K-singular value decomposition (K-SVD) algorithm. In the sparse representation stage, the stopping residue error was adaptively achieved according to the estimated cross-correlation and the adjusted noise spectrum, and the orthogonal matching pursuit (OMP) approach was applied to reconstruct the clean speech spectrum from the noisy speech. Finally, the clean speech was re-synthesised via the inverse Fourier transform with the reconstructed speech spectrum and the noisy speech phase. The experiment results show that the proposed method outperforms the conventional methods in terms of subjective and objective measure.

***Keywords*—**Speech denoising, sparse representation, K-singular value decomposition, orthogonal matching pursuit.

I. Introduction

IN real communication applications, the speech signal is inevitably corrupted by the environmental noise. This not only degrades the speech quality and intelligibility, but also reduces the performance of the signal processing system. Therefore, reducing noise in the corrupted speech is significant and has a wide range of applications [1]. Many speech denoising approaches have been investigated in the literatures, such as the spectral subtraction (SS) [2], model based [3], wiener filtering (WF) [4], and so on. The SS is widely used in the signal processing system due to good performance and low computational complexity. However, the traditional SS has some problems that affect the noise estimation performance, such as the cross-correlation errors and the magnitude errors. Some attempts (e.g. [6], [7]) have taken these problem into account, but most of these studies were focused on speech recognition.

Recently, sparse representation has been drawing more and more attention and widely used in compressed sensing, image and audio signal processing [8]. The objective of sparse representation is to represent most information of a signal with a linear combination of only a small number of atoms. Recent results have indicated that many signals, including the speech signal, can be approximated sparsely, which provides a new avenue for speech denoising [9]. The clean speech is achieved by reducing noise in the corrupted signal in the traditional methods, while the clean speech components in the noisy mixture are exploited by the sparse representation, and the purpose of denoising is achieved by reconstructing the clean speech from the noisy signal. At present, a handful of sparse representation based speech denoising methods have been reported. In [10], the K-singular value decomposition (K-SVD) method was employed to train the over-complete dictionary in time domain with the noisy speech signal, and the clean speech is reconstructed by the OMP algorithm. The dictionary of speech spectrum in [11] was learned with the approximation K-SVD, and the least angle regression (LARS) algorithm is used to obtain the sparse representation of the clean speech spectrum. A generative dictionary method is proposed in [12], where the dictionary combines both the speech and the noise spectrum dictionaries. In the speech denoising stage, the speech spectrum is estimated by means of batch LARS with coherence criterion (LARC) approach.

In sparse representation based speech denoising, the estimation of the stopping residue error in the sparse coding (e.g. MP, OMP) is important because the reconstructed signal needs to be approximated to the original clean signal rather than the noisy speech [13]. Since the stopping residue error is related to the noise signal, it is acquired in most of the developed algorithms via estimating the noise variance with the voice activity detection (VAD) methods (e.g. [10]), or the noise spectrum in the beginning segment of the noisy speech (e.g. [11]). However, most of the noise signals are non-stationary in real situation, and the construction of a robust VAD at low SNRs is still an open task. On the other hand, the noise type is unpredictable in real application, therefore, training dictionary for each noise is not realistic.

This paper presents a sparse representation speech denoising method based on adapted stopping residue error. An over-complete dictionary of the clean speech power spectrum is trained by the K-SVD algorithm, and the OMP approach is applied to the reconstruction of the clean speech spectrum. To achieve the adapted stopping residue error, the noise spectrum is estimated by the noise tracking algorithm, and adjusted by a posteriori SNR weighted factor for a continuous update. Meanwhile, the cross-correlation is analyzed, and an estimation method is proposed to obtain the approximate calculation. Then, the stopping residue error is adaptively calculated according the adjusted noise spectrum and the cross-correlation. Finally, the clean speech is gained by the

Qianhua He is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China (corresponding author, e-mail: eeqhhe@scut.edu.cn).

Weili Zhou was with the School of Electronic and Information Engineering, South China University of Technology. He is now with the Department of Electronic and Information Engineering, Foshan University, Foshan, China (e-mail: willychow@163.com).

Aiwu Chen is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China.

estimated clean speech spectrum and the noisy speech phase via the inverse Fourier transform.

The rest of the paper is organized as follows. In Section II, the principle of the sparse representation based speech denoising is introduced, and the proposed method is described in detail. The performance of the proposed method is evaluated on the TIMIT dataset in Section III. Section IV is the conclusion of this paper.

II. Method

A. Sparse Representation Based Speech Denoising

Let the noisy speech be described as:

$$y = x + d \tag{1}$$

where y, x, d denote noisy speech, clean speech, and noise, respectively. Consider the clean signal $x \in R^M$ has a sparse representation over $\psi \in R^{M\times N}$, and then the signal can be represented as:

$$x = \psi C, \quad \|C\|_0 \le T \ll N \tag{2}$$

where ψ is an over-complete dictionary($M \ll N$), which has to be calculated beforehand. Each column in ψ is called an atom in sparse representation. $\|\bullet\|_0$ is the l_0 norm, C is a N length sparse coefficient vector with T nonzero elements. By exploiting the sparse coefficient vector of the clean speech in the noisy signal under the premise of the sparse restriction:

$$\hat{C} = \arg\min \|C\|_0 \quad \text{s.t. } \|y - \psi C\|_2 \le \varepsilon \tag{3}$$

where $\|\bullet\|_2$ is the l_2 norm, ε is the stopping residue error which is related to the noise, the clean signal can be reconstructed via (2).

B. Cross-Correlation Analyze and Estimate

Let $y(n) = x(n) + d(n)$ be the noisy input signal in time domain, which is composed of the clean signal $x(n)$ and the noise signal $d(n)$. Take the Fourier transform of $y(n)$:

$$Y(\omega) = X(\omega) + D(\omega) \tag{4}$$

Squaring both sides of (4), the power spectrum of $Y(\omega)$ can be computed as:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2 + X(\omega)\bullet D^*(\omega) + X^*(\omega)\bullet D(\omega) \tag{5}$$

where $X^*(\omega)$ and $D^*(\omega)$ denote the conjugate of $X(\omega)$ and $D(\omega)$, respectively. As we can see, the power spectrum $|Y(\omega)|^2$ is composed of $|X(\omega)|^2$, $|D(\omega)|^2$ and their cross-correlation

$$R(\omega) = X(\omega)\bullet D^*(\omega) + X^*(\omega)\bullet D(\omega) \tag{6}$$

In the literature, the cross-correlation $X(\omega)\bullet D^*(\omega)$ and $X^*(\omega)\bullet D(\omega)$ are assumed to be zero based on the hypothesis that the speech signal $x(n)$ is uncorrelated with the interfering noise $d(n)$. However, this hypothesis is incorrect because the cross-correlations are not necessarily zero if the speech and noise are correlated and can be quite large relative to $|Y(\omega)|^2$ in some cases. The study in [5] assessed the effect of neglecting the cross-correlation on the power spectrum estimation of clean speech signal, and the conclusion was noted that large estimation errors could be resulted from the zero assumption of the cross-correlation at low SNR levels (especially SNR levels near 0 dB). Fig. 1 plots the values of the speech power spectrum with white noise at 0 dB SNR and the related cross-correlation, the plot is taken over 256 samples which would be considered in a 32-ms windowed frame at 8 kHz sampling rate. It can be seen that, at least in the low frequency part, the cross-correlations are not negligible compared with the noisy speech power spectrum.

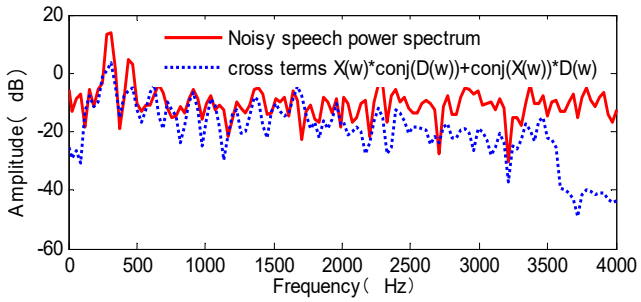

Fig. 1 Plots of the values of the noisy speech power spectrum and the related cross-correlation

Based on the above analysis, the compensation for the cross-correlation is needed in order to improve the accuracy of the noise spectrum estimation. Note that the noise speech spectrum can be expressed as the polar coordinates form via the amplitude and phase:

$$Y(\omega) = |Y(\omega)| e^{j\theta_{Y(\omega)}} \tag{7}$$

where $\theta_Y(\omega)$ denotes the phase of the noisy speech spectrum. Similarly, the noise spectrum can be denoted as $D(\omega) = |D(\omega)| e^{j\theta_{D(\omega)}}$. Since the noise amplitude is unknown, the average amplitude spectrum can be accessed by the noise estimation method.

In speech signal processing, it is generally believed that the human auditory system is not sensitive to the speech phases, which means that the phase does not affect the speech intelligibility, and the noise phase $\theta_D(\omega)$ can be replaced by noisy speech phase $\theta_Y(\omega)$ [5]. That is why in most of the speech denoising methods, calculations are performed only on

the short-time magnitude spectrum, and a fix short-time phase spectrum is maintained for the clean speech re-synthesis. Based on this principle, an estimation of the clean speech complex spectrum can be obtained as:

$$\hat{X}(\omega) = [|Y(\omega)| - |\hat{D}(\omega)|]e^{j\theta_{Y(\omega)}} \tag{8}$$

Substitute (8) into (6), the cross-correlation can be approximately reached:

$$\begin{aligned} R(\omega) &= 2\left||Y(\omega)| - |\hat{D}(\omega)|\right||\hat{D}(\omega)|(e^{j\theta_{Y(\omega)}} \bullet e^{-j\theta_{Y(\omega)}}) \\ &= 2\left||Y(\omega)| - |\hat{D}(\omega)|\right||\hat{D}(\omega)| \end{aligned} \tag{9}$$

C. Denoising Based on Adapted Stopping Residue Error

Consider the (5) can be denoted as the i^{th} frame signal:

$$|Y_i|^2 = |X_i|^2 + \beta_i |\hat{D}_i|^2 + R_i \tag{10}$$

where R_i is the cross-correlation. Based on sparse representation, the above equation can be denoted as:

$$|Y_i|^2 = \psi_{\text{ps}} C_i + \beta_i |\hat{D}_i|^2 + R_i \tag{11}$$

where ψ_{ps} is an over-complete dictionary, which is related to the clean signal power spectrum, and C_i is the sparse coefficient vector, R_i is calculated by (9). $|\hat{D}_i|^2$ is estimated using the continuous noise tracking algorithm [14], and is adjusted by a posteriori SNR weight factor β_i [15]. β_i is proposed to achieve a better estimation of the noise spectrum. When the SNR is low (such as non-speech frame or voice energy is low), the noise spectrum is attenuated more, and vice versa. β_i can be determined as:

$$\beta_i = \begin{cases} 3 & SNR_i < -5dB \\ 2 - \dfrac{1}{10} SNR_i & -5dB \le SNR_i \le 10dB \\ 1 & SNR_i > 10dB \end{cases} \tag{12}$$

where $SNR_i(dB) = 10\log_{10}\left(\sum_{k=0}^{N-1}|Y_i(k)|^2 / \sum_{k=0}^{N-1}|\hat{D}_i(k)|^2\right)$. N is the frame length in samples.

Thus, by setting the l_2 norm of $E_i = \beta_i |\hat{D}_i|^2 + R_i$ as the adapted stopping residue error ε_i, the sparse coefficient vector of the clean signal power spectrum is exploited as:

$$\hat{C}_i = \arg\min \|C_i\|_0 \quad \text{s.t.} \ \left\| |Y_i|^2 - \psi_{\text{ps}} C_i \right\|_2 \le \varepsilon_i \tag{13}$$

where $\varepsilon_i = \|E_i\|_2 = \left\| \beta_i |\hat{D}_i|^2 + R_i \right\|_2$.

With the exploited sparse coefficient vector $\hat{C}_i$ and ψ_{ps}, the power spectrum can be reconstructed as:

$$|\hat{X}_i|^2 = \psi_{\text{ps}} \hat{C}_i \tag{14}$$

A noise suppression filter [10] constructed from $|\hat{X}_i|^2$ and E_i is applied to the mixture spectrum $|Y_i|$, and the final clean magnitude spectrum is obtained:

$$|\hat{X}_i| = h_i |Y_i| \tag{15}$$

h_i is the noise suppression filter at the i^{th} frame($0 \le h_i \le 1$), and is derived with:

$$h_i = \frac{|X_i|}{|Y_i|} = \sqrt{\frac{1 - \dfrac{(\gamma_i + 1 - \xi_i)^2}{4\gamma_i}}{1 - \dfrac{(\gamma_i - 1 - \xi_i)^2}{4\xi_i}}} \tag{16}$$

where $\xi_i = \dfrac{|\hat{X}_i|^2}{E_i}$ is the instantaneous a priori SNR, and $\gamma_i = \dfrac{|Y_i|^2}{E_i}$ is the instantaneous a posteriori SNR.

Finally, $|\hat{X}_i|$ is re-synthesized to the time-domain signal via the inverse discrete Fourier transformation with the noisy speech phase.

Since ψ_{ps} is an over-complete dictionary, the sparse representation is known to be NP-hard problem. The solution of this approximation problem is divided into two classes: the greedy methods and the convex optimization methods [8]. Compared with the convex optimization methods, the greedy methods have lower complexity and provide a comparable performance. Moreover, the OMP algorithm orthogonalises the residue error and all the other selected atoms, which guarantees the convergence of a finite number of iterations [16]. So, the OMP algorithm is applied to the sparse reconstruction stage in this paper.

The above analysis is based on the assumption that the clean speech dictionary is obtained beforehand. Therefore, a suitable over-complete dictionary has to be predefined before the signal reconstruction step. In the proposed method, the clean speech data are used to train the dictionary because the data-driven learning has a better adaption to the signal itself. Moreover, the K-SVD algorithm [17] is employed to the dictionary learning due to its efficiency and great performance. The method is summarized as Algorithm 1.

Algorithm 1.Proposed speech denoising method.

Input: Noisy speech y, dictionary ψ_{ps}

Output: Reconstructed clean speech $\hat{x}$.
For each utterance y

1. split y into overlap frames $y_1,...,y_M$.
2. calculate the magnitude spectrums $Y_1,...,Y_M$ and the noisy speech phases $\zeta_1,...,\zeta_M$.
3. apply the continuous noise tracking algorithm to estimate the noisy power spectrum $\left|\hat{D}_i\right|^2, i=1:M$
4. calculate the adapted residue error:
   a. $E_i=\beta_i\left|\hat{D}_i\right|^2+R_i$, where $R_i=2\left|Y_i-\hat{D}_i\right|\bullet\left|\hat{D}_i\right|$
5. sparse decompose with the OMP algorithm using the adapted residue error:
6. $\hat{C}_i=\arg\min\left\|C_i\right\|_0 \quad \text{s.t.} \left\|\left|Y_i\right|^2-\psi_{ps}C_i\right\|_2\le\left\|E_i\right\|_2$
7. reconstruct the clean speech magnitude spectrum with (15) and (16).
8. Obtain the reconstructed speech signal $\hat{x}$ in time domain via IFFT with the reconstructed magnitude and the noisy speech phases $\zeta_1,...,\zeta_M$.

III. Experiments

A. Experiment Setting

The TIMIT dataset is used to evaluate the performance of the proposed method with all utterances down-sampled at 8 kHz. The frame length is set to 32 ms with 50% overlap, and the DFT number is 256. 300 utterances in total 50000 frames from the TIMIT train set are used to train the power spectrum dictionary of the clean signal. The size of the over-complete dictionary is 256*1024, and it is initialized randomly by the training utterances. In the experiment, the K-SVD toolbox [18] is employed to the dictionary learning and the number of K-SVD iteration is 40. To acquire the noisy speech, 200 utterances in total 33000 frames from the TIMIT test set are corrupted by four different noises; namely, White, Babble, Pink and F16, from the NOISEX-92 database at -5 dB, 0 dB, 5 dB, 10 dB SNR. The standard SS algorithm [2] and the spectral domain sparse representation based speech denoising (SRDN) method [11] are presented for the comparison.

B. Experiment Results and Analysis

Fig. 2 shows the waveforms of the clean speech an utterance from the TIMIT dataset (Fig. 2 (a)), the noisy speech (Fig. 2 (b), clean speech corrupted by the white noise at 10 dB) and the speech denoised by SS (Fig. 2 (c)), SRDN (Fig. 2 (d)) and the proposed method (Fig. 2 (e)). Figs. 3 and 4 show the corresponding spectrograms of the clean speech (Fig. 3 (a)), noisy speech (Fig. 3 (b)), and the speech reconstructed by SS (Fig. 4 (a)), SRDN (Fig. 4 (b)), the proposed method (Fig. 4 (c)).

In terms of waveforms, Fig. 2 (e) (the proposed method) seems much cleaner than Figs. 2 (c) (SS) and (d) (SRDN), and is more similar to Fig. 2 (a) (clean speech). As for the spectrograms of the waveforms, Figs. 4 (a) (SS) and (b) (SRDN) still have much residue noise in contrast to Fig. 3 (a) (clean speech). On the contrary, Fig. 4 (c) (the proposed method) appears to have less residue noise, and the voice part is cleaner than those of Figs. 4 (a) and (b). The above results demonstrate that the proposed method outperforms the other two methods. Note that some unvoiced parts (such as the last unvoiced phoneme "s") in Figs. 4 (b) and (c) are missed in contrast to Fig. 3 (a). A possible explanation is that the phoneme "s" is similar to the white noise, the atoms used for representing "s" are ignored, leading to an energy omitting in the reconstructed speech.

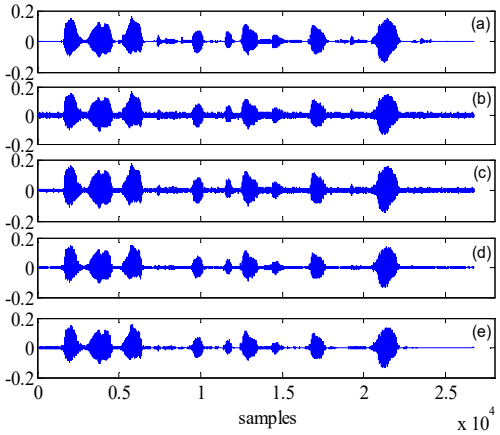

Fig. 2 The utterance "Her wardrobe consists of only skirts and blouses"

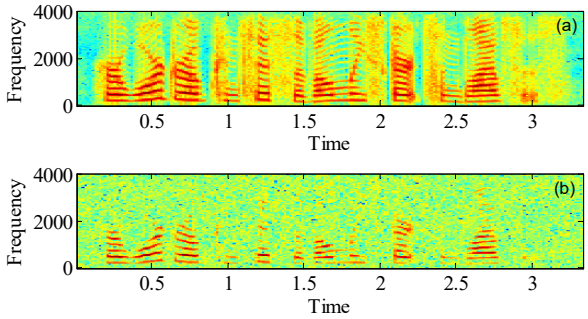

Fig. 3 Spectrogram of the clean speech and noisy speech

Furthermore, Fig. 5 shows another comparison for different denoising algorithms. The denoising performance is measured by the difference of the PESQ score of noisy speech to clean speech and the PESQ scores of denoised speech to clean speech. As we can see, the proposed method outperforms the comparison methods in all noise scenarios at -5 dB, 0 dB, 5 dB SNR, and in 3 out of 4 noise scenarios at 10 dB. The proposed method achieves a mean PESQ score improvement of 0.26 in all noise scenarios at 10 dB, a mean improvement of 0.38 at 5 dB, a mean improvement of 0.40 at 0 dB, and a mean improvement of 0.31 at -5 dB. The results indicate that the performance of the proposed method outperforms the comparison methods in most conditions, and a more significant performance can be reached at low SNRs. The reason may be that the estimation of the stopping residue error is more accurate by the cross term compensation and the noise spectrum adjustment when the SNR is low, resulting in the fact that the exploited atoms have a better representation of the

original clean signal. Therefore, the reconstructed speech of the proposed method is more similar to the clean speech and has less residue noise. On the other hand, compared to white noise, pink noise and F16, the PESQ score improvement of the proposed method is less significant for Babble noise scenarios, as the noisy type becomes more structured likely.

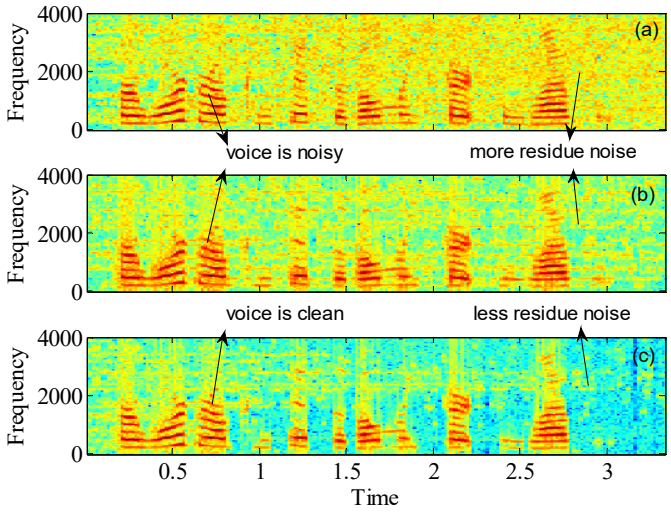

Fig. 4 Spectrogram of the reconstructed speech

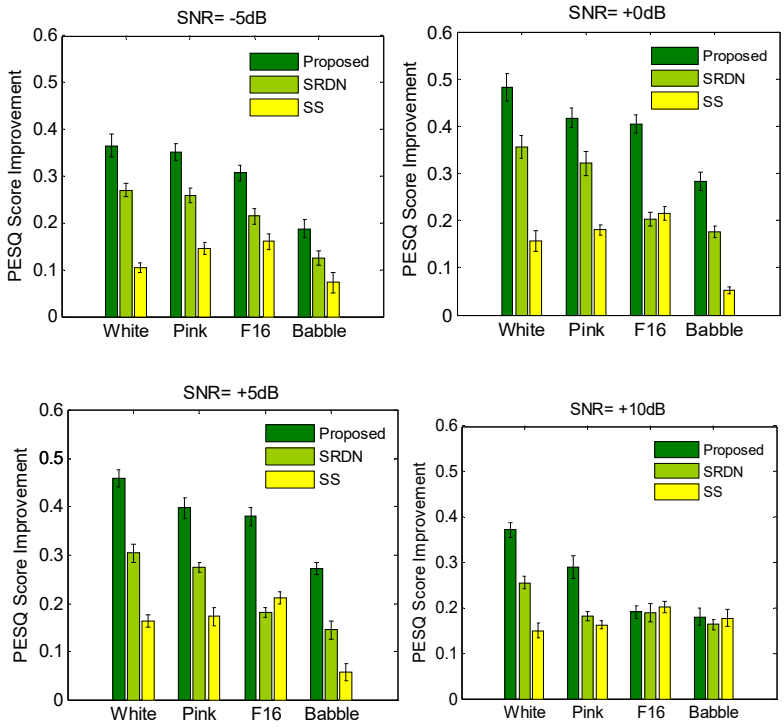

Fig. 5 The comparison of the PESQ scores improvement in four noises at the SNRs of -5 dB, 0 dB, 5 dB, 10 dB. Filled bars denote meaökösSQ scores improvements, error bars denote 95% confidence interval of the mean improvements

IV. Conclusion

In this paper, speech denoising is regarded as a sparse signal recovery problem. An over-complete dictionary of the clean speech power spectrum is learned by the K-SVD algorithm, and the OMP approach is applied to the reconstruction of the clean speech spectrum. The cross-correlations between the clean speech spectrum and the noise spectrum are compensated, and the stopping residue error in the OMP algorithm is adaptively selected according to the cross-correlation and the adjusted noise spectrum. Finally, the clean speech spectrum is reconstructed, and the clean speech in time domain is attained via the inverse Fourier transform. The experiment results show that the proposed method outperforms the standard SS method and the sparse representation based speech denoising method under the conditions of different noises and SNRs.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Item No. 61571192).

REFERENCES

[1] Gaikwad V M, Vasekar S S. Survey on quality and intelligibility offered by speech enhancement algorithms(C). *2015 International Conference on Computing Communication Control and Automation*, Pune, 2015: 694-697.

[2] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics (J).*IEEE Transactions on Audio, Speech, Language Processing*, 2001, 9(5): 504-512.

[3] Kodrasi I, Marquardt D, Doclo S. Curvature-based optimization of the trade-off parameter in the speech distortion weighted multichannel wiener filter (C). *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, 2015: 315-319.

[4] T.Gerkmann. MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase (C). *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, 2014: 4478-4482.

[5] Loizou P C. Speech enhancement: theory and practice (M). Florida: CRC Press, 2013: 1-5.

[6] Evans N, Mason J, Liu W, et al.. An assessement on the fundamental limitations of spectral subtraction (C). *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulous, 2006: 145-148.

[7] Hilman F, Koji I, Koichi S. Feature normalization based on non-extensive statistics for speech recognition (J). *Speech Communication*, 2013, 55(5): 587-599.

[8] Wohlberg B. Efficient algorithms for convolutional sparse representations (J). *IEEE Transactions on Image Processing*, 2016, 25(1): 301-315.

[9] He Yong-jun, Han Ji-qing, Deng Shi-men, et al.. A solution to residual noise in speech denoising with sparse representation (C). *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, 2012: 4653-4656.

[10] Zhao Nan, Xu Xin, Yang Yi. Sparse Representations for Speech Enhancement (J). *Chinese Journal of Electronics*, 2011, 19(2): 268-272.

[11] Zhao Yan-ping, Zhao Xiao-hui, Wang Bo. A speech enhancement method employing sparse representation of power spectral density (J). *Journal of Information and Computational Science*, 2013, 10(6): 1705-1714.

[12] Sigg C D, Dikk T, Buhmann J M. Speech enhancement using generative dictionary learning (J).*IEEE Transactions on Audio, Speech, Language Processing*, 2012, 20(6): 1698-1712.

[13] Sun Lin-hui, Yang Zhen. Speech enhancement based on data-driven dictionary and sparse representation (J). *Signal Processing*, 2011, 27(12): 1793-1800.

[14] Rangachari, S. and Loizou, P, "A noise estimation algorithm for highly nonstationary environments," *Speech Communication*, vol. 28, pp. 220–231, 2006.

[15] Berouti M, Schwartz M, Makhoul J. Enhancement of speech corrupted by acoustic noise (C). *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1979: 208-211.

[16] Pati Y C, Rezaiifar R, Krishnaprasad P S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition (C). *IEEE Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993, 40-44.

[17] Aharon M, Elad M, Bruckstein A. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation (J).*IEEE Transactions on Audio, Speech, Language Processing*, 2006, 54(11): 4311-4322.

[18] Rubinstein R, Zibulevsky M, Elad M. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit(R). Science Department Technical Report CS, 2008.