Speech Denoising Based on Noise Reshaping

Enqing Dong Xiaoxiang Pu
School of Information Engineering, Shandong University at Weihai, 264209 China

Abstract - **A new algorithm of speech denoising for non-stationary environments based on noise reshaping is proposed. There are two key points in the algorithm. Firstly, the real-time noise estimation technique can continuously update the noise spectrum with a smoothing factor by tracing the minimum of the noisy speech in each frame. Secondly, the linear prediction technique is adopted to compute the perceptual weighting function by extracting characteristic parameters of speech and using the perception properties of human auditory. The perceptual weighting function is used to shape the noise so that the noise spectrum can be redistributed according to the real speech. The simulation result indicates that the algorithm can reduce the background noise effectively and improve the quality of speech.**

Keywords: **speech denoising; noise reshaping; perceptual weighting function**

I. INTRODUCTION

The quality of speech occurs to degrade under the presence of a great variety of additive background noise. To improve the quality of speech communication and reduce the influence of the noise on the performance of speech communication, the speech denoising technique has become a hot subject in speech signal processing. Recently, many kinds of algorithms on speech enhancement have been proposed, such as the traditional spectrum subtraction algorithm, the modern subspace method[1-4], the parameter method and wavelet transform method[5-8]. In this paper, we focus on the perception properties of human auditory system to do some discussions.

In 1988, J. D. Johnston et al firstly proposed a perceptual model based on psychoacoustics[9]. In the recent years, human auditory perceptual properties have been widely applied in speech denoising[10-14] due to the deep study on it.

In 1993, D. E. Tsoukalas et al proposed a speech enhancement algorithm based on auditory masking property[10]. The algorithm utilizes the property of strong signals masking weak signals so that the signals below the masking threshold of hearing are imperceptible. Only the signals over the threshold are processed by neglecting masked weak signals to improve the quality of speech. So the drawback of this algorithm is that the denoising performance relies too much on the estimation of the masking thresholds of the background noise.

In 2003, Y. Hu proposed a new denoising algorithm[14] based on perceptual weighting. In view of the perceptual weighting of the prediction error, the prediction error is weighted for minimizing the perceptual mean square error so that the influence of noise is reduced. But the algorithm in the literature estimates noise only in silent frames and can not estimate the abrupt non-stationary noise effectively. Experimental results show that some musical noise exists in the enhanced speech for this algorithm.

So, in 2006, Jong Won shin proposed another algorithm[15] for speech enhancement based on noise reshaping. While reshaping the noise, the algorithm introduces comfortable spectrum and adopts minimum distance criteria on both noise spectrum and comfortable spectrum. Although the algorithm avoids complicated computation, but the effect of the algorithm is not better than the above algorithm.

The differences of the algorithm proposed in this paper with the algorithm in Ref. [14] is as follows. The new proposed algorithm directly adopts perceptual function to reshape the noise spectrum, and then deduce the filter equation. In the meantime, a real-time noise tracking technique for non-stationary noise environments is adopted to achieve better smooth estimation of noise. The denoising of noisy speech for non-stationary noise environments is further improved.

II. HUMAN AUDITORY PROPERTIES AND PERCEPTUAL WEIGHTING FUNCTION

Human auditory system has masking characteristic which represents that speech sense mainly relies on loudness level and frequency, which is insensitive to phase. Herein we consider about loudness level. When two sounds with different loudness levels affect on human auditory system simultaneously, the frequency component with higher loudness level will restrain that with lower loudness level. On the other hand, speech signal has a characteristic of energy accumulation of low frequency formants. The speech energy is relatively accumulated in low frequency formants. The speech with higher loudness level can mask noise components with relative low loudness level. Human auditory system is insensitive to the noise in the area. However, the system is sensitive to the noise in spectral valley where the speech energy is relatively low so that the noise has obvious influence on speech.

In order to match the decoded speech with original speech well in speech coding system, Atal et al.[16] proposed the perceptual weighted minimum mean square error criterion which reduces the quantization error of coding by using human auditory properties. So far, the perceptual weighting filter equation of speech coding system is still based on the above criterion:

$$P(z)=\frac{A(z/r_1)}{A(z/r_2)}=\frac{1-\sum_{k=1}^{p}a_k r_1^k z^{-k}}{1-\sum_{k=1}^{p}a_k r_2^k z^{-k}} \quad (1)$$

where $A(z)$ is the linear predicting function, p is the prediction order, a_k is the linear predicting coefficient, r_1 and r_2 are the perceptual weighting coefficients, the frequency response of $P(z)$ is opposite to the power spectrum of original speech. In this way, when the predicting error is weighted, the predicting error in the valley will be transferred. In this paper, the distributing characteristic of noise can be changed by adopting $P(z)$ to weight the noise. So the noises centralize more in formant areas and less in the areas where speech energy is low.

III. ANALYSIS OF THE PROPOSED ALGORITHM

A. Principle of the Proposed Algorithm

At first, we assume the approach is for the additive environments. The noise signal is additive and uncorrelated with the clean speech, i.e,

$$y(n) = x(n) + d(n) \tag{2}$$

where $y(n)$, $x(n)$ and $d(n)$ are the noisy speech, clean speech and noise signal respectively. Let $\hat{x}(n)$ be a linear estimation of the clean speech $x(n)$. By applying Fourier Transform, we can obtain

$$Y(\omega) = X(\omega) + D(\omega) \tag{3}$$

$$\hat{X}(\omega) = H(\omega)X(\omega) \tag{4}$$

where $H(\omega)$ is defined as a filter function. By using Wiener filter theory based on the Minimum Mean Square Error criterion, we can get the standard equation of Wiener filter in frequency domain under stationary noise environments.

$$H(\omega) = \frac{S_s(\omega)}{S_s(\omega) + S_d(\omega)} \tag{5}$$

Obviously, Wiener filter is a constant parameter linear system that can be only applied to stationary signals and can not handle non-stationary noise signals. A frequency compensation factor μ is introduced to modify the filter equation as follows:

$$H(\omega) = \frac{S_s(\omega)}{S_s(\omega) + \mu S_d(\omega)} \tag{6}$$

When we adopt the masking properties of human auditory and introduce the perceptual weighting function $p(\omega)$ to reshape the noise spectrum $S_d(\omega)$ in the Equation (6), we can obtain perceptual weighted noise spectrum $S'_d(\omega)$.

$$S'_d(\omega) = \left|p(\omega)\right|^2 S_d(\omega) \tag{7}$$

Substitute the Equation (7) into the Equation (6), the filter equation $H(\omega)$ in frequency domain based on noise reshaping can be rewritten as

$$H(\omega) = \frac{S_s(\omega)}{S_s(\omega) + \mu\left|p(\omega)\right|^2 S_d(\omega)} \tag{8}$$

Obviously, the Equation (8) is formally the same as the gain function $g(\omega)$ proposed by Yi Hu in Ref. [14]. Furthermore, Yi Hu et al has demonstrated the conclusion via perceptual weighting error. So the noise reshaping algorithm proposed in this paper is valid and reasonable.

B. Analysis of Frequency Compensation Factor

Analysing the Equation (6) with Ref. [3] and Ref. [4], we can see that the Equation (6) is actually another expression about multi-band spectrum subtraction method. Here, μ has the same meaning as the over-subtraction factor α_i in Ref. [3], which objectively reflects that the non-stationary background noise nonuniformly affects the speech spectrum. As the calculation of μ in Ref. [14] is too complicated, so we introduce the method in Ref. [3] to calculate μ as follows:

$$\mu_i = \begin{cases} 5 & SNR_i \le -5 \\ 4 - \dfrac{3}{20} SNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i \ge 20 \end{cases} \tag{9}$$

where SNR_i is the SNR of *i*th segment of the frequency band.

C. Noise Estimation

From the Equation (8), we can see that the estimation of the noise spectrum is important in the speech denoising. For most traditional speech enhancement algorithms, the simplest approach of noise estimation is to estimate and update the noise spectrum at the silent segments using voice activity detection (VAD). But the accuracy of this method is confined by the performance of the VAD. Especially when the SNR is very low, the overall accuracy of algorithm is affected.

So we choose a new noise estimation algorithm[17] to estimate the noise in noisy speech, which updates the noise in each frame with a smoothing factor by tracing the minimum spectrum of the noisy speech. Without adopting voice activity detection, the algorithm can estimate the presence probability of speech by using the correlation of power spectral components of neighboring frames. The algorithm can fast track the variation of noise.

At first, the smoothed power spectrum of noisy speech is computed using the following first-order recursive equation:

$$P(\lambda,k) = \eta \times P(\lambda-1,k) + (1-\eta) \times \left|Y(\lambda,k)\right|^2 \tag{10}$$

where $P(\lambda,k)$ is the smoothed power spectrum, λ is the frame index, k is the frequency index, $\left|Y(\lambda,k)\right|^2$ is the short-time power spectrum of the noisy speech and η is the smoothing constant (the value of η is set to 0.94 ~ 0.98 [18]).

Secondly, track the minimum of noisy speech power spectrum and update it in each frame. Then compute the frequency dependent smoothing constant $\alpha_s(\lambda,k)$ according to the speech presence probability and update the noise spectrum $D(\lambda,k)$ according to the following equation (11) (Refer to [17]):

$$D(\lambda,k) = \alpha_s(\lambda,k) \times D(\lambda-1,k) + (1-\alpha_s(\lambda,k)) \times \left|Y(\lambda,k)\right|^2 \tag{11}$$

In addition, as the power spectrum of clean speech $S_s(\omega)$ in the equation (8) can not be obtained in advance in practice, so which is replaced by $S'_s(\omega)$ in multi-band spectral subtraction algorithm (Refer to [3]).

Based on the above analysis, the final expression of noise reshaping filter equation in frequency domain is defined as follows:

$$H'(\omega) = \frac{S'_s(\omega)}{S'_s(\omega) + \mu |p(\omega)|^2 S_d(\omega)} \qquad (12)$$

IV. FLOW DIAGRAM AND IMPLEMENTATION STEPS OF THE PROPOSED ALGORITHM

A. *Flow Diagram of the Proposed Algorithm*

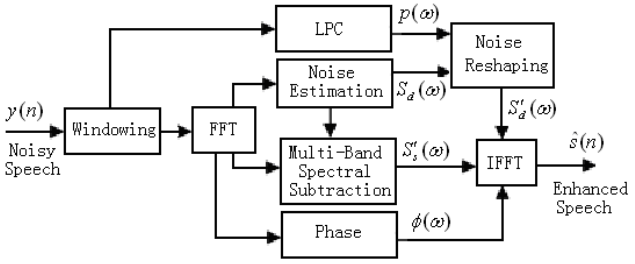

Figure 1. Flow diagram of the proposed algorithm

B. *Implementation Steps of the Proposed Algorithm*

Step1, compute the perceptual weighting function $P(z)$ using the linear prediction technique and the equation (1).

Step2, adopt the real-time noise tracing method to estimate the noise spectrum.

Step3, reshape the noise spectrum to obtain the weighted noise spectrum.

Step4, estimate the clean speech spectrum with the multi-band subtraction method.

Step5, compute the value of the frequency compensation factor μ according to the equation (9).

Step6, compute the gain function in frequency domain according to the equation (12).

Step7, compute the signal power spectrum according to the equation (4).

Step8, obtain the denoised speech with the inverse Fourier transform.

V. EXPERIMENTAL RESULTS

To assess the performance of the proposed algorithm, we choose four kinds of noises (F16 plane noise, Factory noise, M109 tank noise, Pink noise) from the standard noise-92 database as the additive background noise and the clean speech from UT Dallas University speech lab. The speech and the noise are sampled by 8kHz and quantized into digital signal by 16bit. The noisy speech signals are divided into 20 segments according to 5 different SNR. The frame length of the noisy speech is 256 samples. There are 128 samples overlapping between adjacent frames. The experimental results show that the performance is the best when the values of r1、r2 in the equation (1) are 0.9 and 0.1 respectively.

In the experiments, we adopt informal speech test. The 20 speech segments are processed by three algorithms (multi-band spectrum subtraction(MBSS), perceptual weighting (PW) algorithm and the new proposed algorithm) respectively. 5 listeners are invited to test the processed speech for three times. The informal test result indicates that the quality of the proposed algorithm is the best.

Fig.2 is a combined waveform graph obtained by the four methods in m109 tank background noise with -5dB SNR. The experimental results also indicate that the proposed algorithm in this paper is the best in denoising, especially the algorithm can obviously constrain the residual musical noise. The main reason is that the algorithm adopts a better real-time noise tracking estimation technique and a predominant noise reshaping technique.

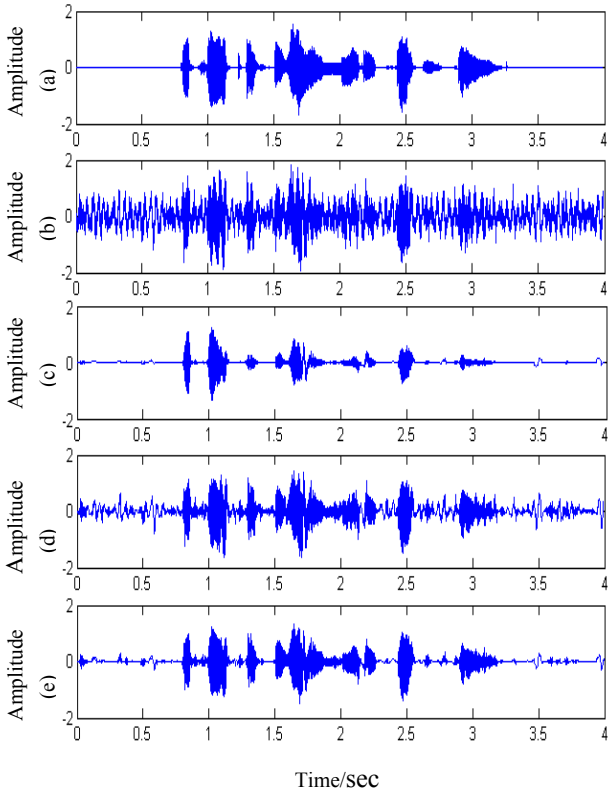

Figure 2. The speech waveform comparison graph
(a) Original clean speech (b)Noisy signal(additive M109 tank noise with -5dB SNR) (c) enhanced speech by MBSS algorithm (d) enhanced speech by PW algorithm (e) enhanced speech by proposed algorithm

VI. CONCLUSIONS

In this paper, a speech denoising algorithm for additive non-stationary environments based on noise reshaping is proposed, which adopts the perception properties of human audition. The real-time noise minimum power spectrum tracking technique is used to update the non-stationary noise frame by frame effectively and estimate the noise effectively. By means of reshaping the noise spectrum with the perceptual weighting function, the distribution characteristic of noise is changed. A great deal of experimental results shows that the proposed algorithm can effectively constrain background noise, reduce speech distortion and improve the definition of the denoised speech.

ACKNOWLEDGMENT

This work was supported in part by the national natural science foundation of China under Grant 60572074.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP, vol. 27(2): 113 - 120, 1979.

[2] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", Proceeding of IEEE ICASSP, pp: 208 - 211, 1979.

[3] Kamath S., Loizou P., "A multi-band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise", In: Proc. ICASSP, IV-4164, 2002.

[4] Ephraim Yharry L., Trees V., "A signal subspace approach for speech enhancement". IEEE Transactions on Speech and Audio Processing, vol. 3(4):251- 266, 1995.

[5] Epharim Y., Malah D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", IEEE. Trans. Acoustic, Speech Signal Processing, vol. 32(6):1109-1121, 1984.

[6] Martin R., "Statistical methods for the enhancement of noisy speech",.IWAENC'2003, pp:1 - 6, 2003.

[7] M. Bahoura and J. Rouat, "New approach for wavelet speech enhancement", Eurospeech, Aalborg, Denmark, pp: 1937 – 1940, 2001.

[8] J. Seok and K. Bae, "Speech enhancement with reduction of noise components in the wavelet domain", ICASSP 97, Munich, Germany, pp: 1223 – 1326, 1997.

[9] J. D. johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE J. on Select. Areas Commun., vol. 6: 314 - 323, 1988.

[10] D. E. Tsoukalas, Paraskevasa M., "Speech Enhancement Using Psicho-acoustic Criteria", Proc. IEEE ICASSP Minneapolis, pp: 359 - 361, 1993.

[11] R. Sarikaya and J. H. L. Hansen, "Auditory masking threshold estimation for broadband noise sources with application to speech enhancement", in Proc Euro speech, pp: 2571 - 2574, 1999.

[12] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression,". IEEE Trans. Speech and Audio Processing, vol.5: 497 - 514, 1997.

[13] Nathalie Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech and Audio Process, vol. 7(2): 126 - 137, 1999.

[14] Y. Hu and P. C. Loizou, "A Perceptually Motivated approach for Speech Enhancement",. IEEE transactions on speech and signal Processing, vol.11 (5): 457 - 465, 2003.

[15] Jong Won Shin, Seung Yeol Lee, "Speech enhancement based on residual noise shaping," ICSLP, Proceedings of International Conference on Spoken Language, pp: 1415 – 1418, 2006.

[16] Schroeder M., Atal B S., "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," J Acoust Soc Am, 66:1647-1652, 1979.

[17] S. Rangachari, P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environ- ments", .Speech Communication 48, pp: 220 – 231, 2006.

[18]Haibin Liu, Zhenyang Wu, Li Zhao, "Speech enhancement based on human auditory masking proprieties under non-stationary environments," Signal Processing, vol.19 (4): 303 - 307, 2003.

[19]Tianren Yao, Hong Sun, "Modern Digital Signal Processing",.Huazhong University of Science and Technology Press,pp:19–27,1999.