TELKOMNIKA, Vol. 11, No. 9, September 2013, pp. 5359~5364
ISSN: 2302-4046

Improved a Priori SNR Estimation for Speech Enhancement Incorporating Speech Distortion Component

Shifeng Ou*, Chao Geng, Ying Gao
Institute of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005, Shandong, China
*Corresponding author, e-mail: 250800719@qq.com

Abstract

The well known decision-directed (DD) approach drastically limits the level of musical noise, but the estimated a priori SNR matches the previous frame rather than the current one. Plapous introduced a novel method called two-step noise reduction (TSNR) technique to refine the a priori SNR estimation of the DD approach. However, the performance of this method depends on the accurateness of the estimated speech in its second step. In this paper, we propose an improved approach for the a priori SNR estimation in DCT domain with two steps like the TSNR method. While in the second step, considering the two state components of the estimation error between speech signal and its estimation, the speech distortion component and residual noise component, we make the estimated speech subtracted by its speech distortion as a refined estimation for the clean speech signal. Because the speech distortion component is offset, the estimated a priori SNR is more accurate. A number of objective tests results show the improved performance of the proposed approach.

Keywords: *speech enhancement, signal to noise ratio, speech distortion, noise reduction*

1. Introduction

Most of the existing voice communication systems are designed for processing of noise free speech. However, speech signals used as an input to these systems are often degraded by additive noise. So the problem of enhancing speech degraded by uncorrelated additive noise, when only the noisy speech is available, has been widely addressed in the past few decades and it still provides an active field of research. Many approaches have been investigated in order to gain spectral enhancement, including hard or soft decision estimation, spectral subtraction, Wiener filtering, and minimum mean square error (MMSE) estimation [1-4]. Widespread use of these methods in due to the fact that they are fairly straightforward to implement, effective in removing various background noises and have low computational load. Almost all of these speech enhancement approaches rely on the estimation of a short-time spectral gain, which is a function of the a priori SNR. So the estimation of the a priori SNR is a crucial part of speech enhancement algorithm [5]. An erroneous estimation of this parameter leads to speech distortion, musical noise, or reduced noise reduction. In the meantime, piracy becomes increasingly rampant as the customers can easily duplicate and redistribute the received multimedia content to a large audience

Many of the existing a priori SNR estimation techniques require either experimentally pre-specified weighting factors or prior assumptions of the parameter in the signal model. The well established decision-directed (DD) approach is computationally efficient and performs quite well in noise reduction applications [6, 7], but this approach has a serious drawback that the estimated a priori SNR follows the shape of the instantaneous SNR with a simple delay of one short time frame. To suppress the problem of the decision directed approach, a novel method, called two-step noise reduction technique (TSNR) is presented to refine the estimation of the a priori SNR [5, 8]. It is also reported that several a priori SNR estimation approaches have been proposed based on higher order moments[9], which have shown promising results in a number of applications and are of particular value when dealing with a mixture of normal-Laplace processes.

Received March 6, 2013; Revised June 12, 2013; Accepted June 24, 2013

In this paper, an effective a priori SNR estimation for noisy speech enhancement is proposed by incorporating the speech distortion component to the estimated speech to refine the estimated a priori SNR. The proposed algorithm not only retains better noise reduction, but also improves deficiency in terms of suppressing musical noise and further reduces the echo because of the enhancing tracking speed of a posteriori SNR. In simulations with speech signals degraded by diverse noises, the proposed method shows improved performance over the other two methods for a number of measures.

The paper is organized as follows: In Section 2 we review the speech enhancement problem and the decision-directed approach which is most frequently used in speech communication systems. In Section 3 we present a novel SNR estimation approach by employing speech distortion component. In Section 4 we show that the proposed approach outperforms the decision-directed and TSNR approaches for non-stationary noise as well as for stationary noise in terms of several instrumental measures. Finally in Section 5, we give our conclusions.

2. Problem Formulation

It is assumed that the noise signal $v(t)$ is additive, i.e. $y(t) = x(t) + v(t)$ with $x(t)$, $y(t)$ the clean speech and noisy speech at time *t*. Taking the DCT to the observed signal gives us:

$$Y_{n,k} = X_{n,k} + V_{n,k}, \ k = 0, \cdots, K-1 \tag{1}$$

Where $X_{n,k}, Y_{n,k}$ and $V_{n,k}$ denote the DCT transformed components of the clean speech, noisy speech and noise signals respectively, *K* is the total number of frequency components, *k* and *n* represent the frequency and frame index. The objective is to find an estimator $\hat{X}_{n,k}$ which minimizes the expected value of a given distortion measure conditionally to a set of spectral noisy features. Since the statistical model is generally nonlinear, and because no direct solution for the spectral estimation exists, we first derive a priori SNR estimate from the noisy features. An estimation of $X_{n,k}$ is subsequently obtained by applying a spectral gain $G(n,k)$ to each short time spectral component $Y_{n,k}$. The choice of the distortion measure determines the gain behavior, i.e., the tradeoff between noise reduction and speech distortion. However, the key parameter is the estimated a priori SNR because it determines the efficiency of the speech enhancement for a given noise power spectrum density.

With the assumption that different DCT components on index *k* are statistically independent, the estimation for clean speech component can be obtained as follows:

$$\hat{X}_{n,k} = G(n,k) \cdot Y_{n,k} \tag{2}$$

Where $G(n,k)$ is the gain function. In general, it can be expressed as a function of the a posteriori SNR and a priori SNR defined as follows:

$$\mathrm{SNR}_{post}(n,k) = \frac{Y_{n,k}^2}{\gamma_V(n,k)} \tag{3}$$

$$\mathrm{SNR}_{prio}(n,k) = \frac{E\left\{X_{n,k}^2\right\}}{\gamma_V(n,k)} \tag{4}$$

Where $\gamma_V(n,k) = E\left\{V_{n,k}^2\right\}$ is assumed to be known since it can be easily computed during speech pauses. An estimation of the a priori SNR is made according to the so-called DD approach [4]:

$$\hat{\mathrm{SNR}}_{prio}^{DD}(n,k) = \beta \frac{\left|\hat{X}_{n-1,k}\right|^2}{\gamma_V(n,k)} + (1-\beta)\max\left\{\mathrm{SNR}_{post}(n,k)-1,0\right\} \quad (5)$$

Where $\hat{X}_{n-1,k}$ is the estimated speech component at previous frame.

Several variants of the gain function $G(n,k)$ have been reported in the literature, such as Wiener, spectral subtraction, or Maximum Likelihood estimates. But, without loss of generality, here the gain function is chosen as the Wiener filter similar to [5].

$$G(n,k) = \frac{\mathrm{SNR}_{prio}(n,k)}{\mathrm{SNR}_{prio}(n,k)+1} \quad (6)$$

3. A Priori SNR Estimation Incorporating Speech Distortion Component

Some analysis of DD approach behavior has been reported in [5, 6] and the results indicated that DD approach can drastically limit the level of musical noise, but the estimated a priori SNR follows the instantaneous SNR with a frame delay. Consequently, since gain function $G(n,k)$ depends on the a priori SNR, $G(n,k)$ computed at current frame matches the previous frame, and thus the performance of the speech enhancement system is degraded. In order to remove the drawbacks of DD approach while maintaining its advantages, Plapous proposed to compute the a priori SNR for the next frame using DD approach and to apply it to the current frame because of the frame delay. This leads to the TSNR approach which is composed of two steps to refine the estimation of the a priori SNR [5].

In the first step, the gain function $G(n,k)$ is computed using DD approach as described in the previous section.

$$G(n,k) = \frac{\mathrm{SNR}_{prio}^{DD}(n,k)}{\mathrm{SNR}_{prio}^{DD}(n,k)+1} \quad (7)$$

In the second step, the gain is then used to refine the estimated a priori SNR of DD approach, and the estimation of TSNR is obtained using the following equation:

$$\mathrm{SNR}_{prio}^{TSNR}(n,k) = \frac{\left|G(n,k)Y_{n,k}\right|^2}{\gamma_V(n,k)} \quad (8)$$

The a priori SNR estimated using DD approach shows good properties but suffers from a frame delay which is removed by the second step of the TSNR algorithm. Therefore, this technique can provide fast response to an abrupt increase in the speech signal without introducing musical noise.

However, as we can see from Equation (8), the TSNR estimation for a priori SNR depends on the estimation $\hat{X}_{n,k}$ for the clean speech component just as Equation (2) showed $\hat{X}_{n,k} = G(n,k)\cdot Y_{n,k}$. Moreover, it is known that the estimation error $e_{n,k}$ between the estimated speech component $\hat{X}_{n,k}$ and its corresponding actual speech component $X_{n,k}$ includes two parts. This can be formulated as:

$$e_{n,k} = \hat{X}_{n,k} - X_{n,k} = G(n,k)\cdot Y_{n,k} - X_{n,k} \quad (9)$$

Substituting $Y_{n,k} = X_{n,k} + V_{n,k}$ into (9) gives us:

$$e_{n,k} = (G(n,k)-1) * X_{n,k} + G(n,k) * V_{n,k} \quad (10)$$

Where $[G(n,k)-1]\cdot X_{n,k}$ represents the speech distortion component and $G(n,k)\cdot V_{n,k}$ is the residual noise component. In order to improve the accurateness of $\hat{X}_{n,k}$, it is inspired to make the estimation $\hat{X}_{n,k}$ subtracted by speech distortion $[G(n,k)-1]\cdot X_{n,k}$ as the refined estimation $\hat{X}_{R,n,k}$ for the actual clean speech component.

$$\hat{X}_{R,n,k} = \hat{X}_{n,k} - [G(n,k)-1]\cdot X_{n,k} \tag{11}$$

The refined estimation is closer to the actual speech component than the former estimation in Equation (9). From the above equation, we substitute $\hat{X}_{n,k}$ for $X_{n,k}$, then we get:

$$\hat{X}_{R,n,k} = [2-G(n,k)]\cdot \hat{X}_{n,k} = [2-G(n,k)]\cdot G(n,k)\cdot Y_{n,k} \tag{12}$$

Then the refined TSNR (R-TSNR) approach for the a priori SNR estimation is obtained in our paper, which can be described as following two steps:

$$\hat{\mathrm{SNR}}_{prio}^{DD}(n,k) = \beta\frac{\left|\hat{X}_{n-1,k}\right|^2}{\gamma_V(n,k)} + (1-\beta)\max\left\{\mathrm{SNR}_{post}(n,k)-1,0\right\} \tag{13}$$

$$\mathrm{SNR}_{prio}^{R-TSNR}(n,k) = \frac{\left|[2-G(n,k)]\cdot G(n,k)\cdot Y_{n,k}\right|^2}{\gamma_V(n,k)} \tag{14}$$

Where β is the controller parameter which can be adjusted to achieve best result, and in our experimental the parameter is chosen as β=0.98.

4. Experimental Results

In this section, the performance of the proposed R-TSNR approach is tested for noisy speech enhancement, and compared to that of DD as well as TSRN approach. The speech material used for tests consists of six sentences spoken by three males and three females. The number of samples per frame is *K*=256 with an overlap of 128 samples. The noise signals used in our evaluation include white noise (White), High frequency channel noise (HF), Destroyer engine room noise (Destroyer), and Babble noise (Babble). The speech signal is sampled at 8 kHz and degraded by these noises at the SNR of 0dB, 5dB, and 10dB.

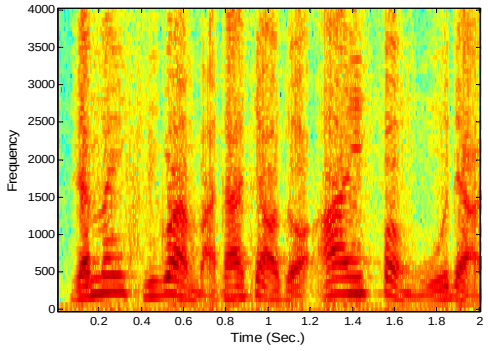

Figure 1. The Original Speech

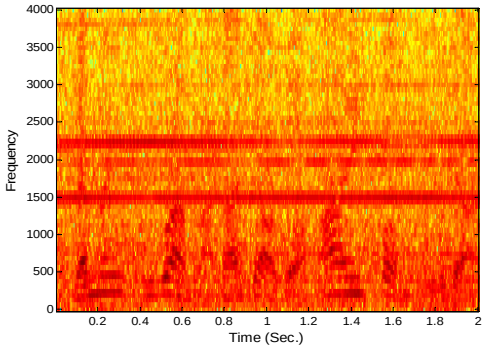

Figure 2. The Noisy Speech

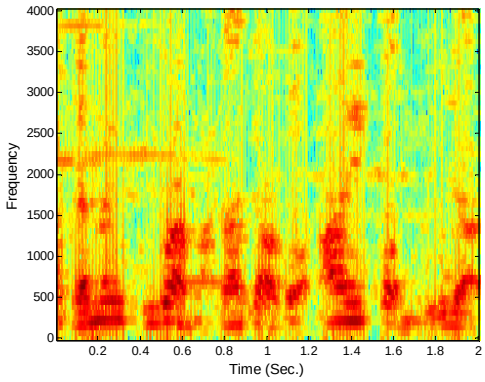

Figure 3. Enhanced Speech by DD Method

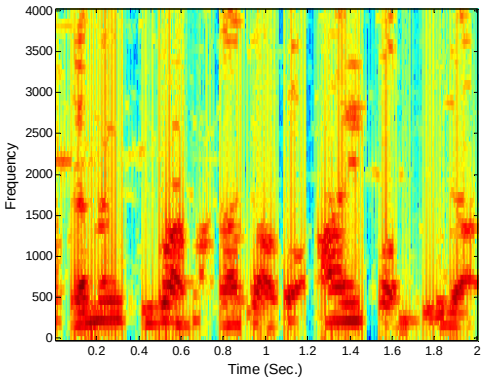

Figure 4. Enhanced Speech by TSNR Method

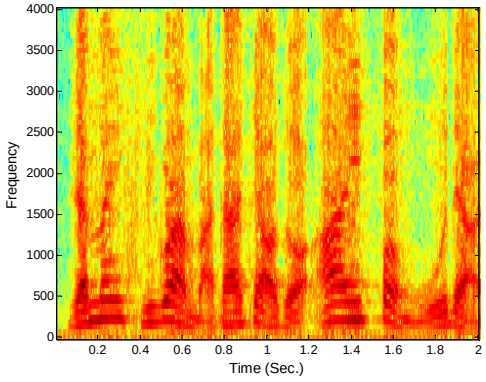

Figure 5. Enhanced Speech by R-TSNR Method

Firstly, the results of the three algorithms for speech enhancement are compared in the frequency domain by means of the speech spectrogram. Figure 1, Figure 2, Figure 3, Figure 4 and Figrue 5 shows clean speech, noisy speech corrupted by the Destroyer engine room noise with 0 dB and the results of enhanced speeches using DD, TSNR, and our proposed methods, respectively. From the obtained results, it can be seen that the R-TSNR approach has a better noise reduction capability; it has less residual noise while keeping more of the speech signals energy unchanged than the other two approaches.

Table 1. Comparison of SEGSNR of Enhanced Signal in Various Noise Conditions

Noise type	Input SNR	Output SEGSNR		
		DD	TSNR	R-TSNR
White	0 dB	4.72	4.90	5.29
	5 dB	7.04	7.29	7.69
	10 dB	8.93	8.96	9.32
HF	0 dB	4.66	4.75	5.11
	5 dB	7.09	7.21	7.59
	10 dB	9.16	9.25	9.50
Destroye	0 dB	4.29	4.47	5.01
	5 dB	6.96	7.08	7.34
	10 dB	9.26	9.39	9.88
Babble	0 dB	2.78	2.86	2.96
	5 dB	4.72	4.81	5.11
	10 dB	7.13	7.30	7.57

Table 2. Comparison of LSD of Enhanced Signal in Various Noise Conditions

Noise type	Input SNR	Output LSD		
		DD	TSNR	R-TSNR
White	0 dB	8.32	7.94	7.73
	5 dB	7.32	7.11	6.89
	10 dB	7.00	6.67	6.42
HF	0 dB	8.07	7.82	7.61
	5 dB	6.62	6.42	5.98
	10 dB	5.96	5.84	5.58
Destroye	0 dB	8.13	7.84	7.64
	5 dB	6.41	6.18	5.81
	10 dB	5.06	4.82	4.54
Babble	0 dB	8.15	7.92	7.66
	5 dB	6.18	5.85	5.61
	10 dB	4.85	4.64	4.37

The segmental SNR (SEGSNR) and log-spectral distortion (LSD) measures are adopted for the objective evaluation [10]. For the segmental SNR, only frames with segmental SNR values greater than -10 dB and less than 35 dB are considered. Table 1 gives the output SEGSNR results of the enhanced speech signals obtained using DD, TSNR and the proposed R-TSNR algorithm in various noise conditions and levels. The results of the LSD are showed in Table 2, also in various noise conditions and levels. From the two tables, we can observe that the proposed algorithm always has a higher SEGSNR and lower LSD as compared to the other algorithms under all tested environmental conditions.

5. Conclusion

An improved expression for the a priori SNR estimation for speech enhancement in DCT domain has been proposed in this paper. Unlike the traditional estimation approaches, we have deliberately considered the speech distortion component in the estimated speech. By incorporating this component to refine the estimator, the improved performance is obtained. The performance of the proposed estimator has been examined using a real speech signal in several noise environments. The experimental results were compared with the DD and TSNR methods and showed that the proposed estimator performed better noise reduction performance than the other estimators.

Acknowledgements

This work was supported by NSFC under Grant Nos. 61005021, 61201457 and A Project of Shandong Province Higher Educational Science and Technology Program under contract J12LN27.

References

[1] MK Hasan, MSA Zilany, MR Khan. DCT Speech Enhancement with Hard and Soft Thresholding Criteria, *Electronics Letters*. 2002; 38(13): 669-670.
[2] T Inoue, H Saruwatari, Y Takahashi, K Shikano, K Kondo. Theoretical Analysis of Musical Noise in Generalized Spectral Subtraction Based on Higher Order Statistics. *IEEE Transactions on Audio, Speech, and Language Processing*. 2011; 19(6): 1770-1779.
[3] H Ding, I Soon, S Koh, C Yeo. A Spectral Filtering Method Based on Hybrid Wiener Filters for Speech eEnhancement. *Speech Communication*. 2009; 51(3): 259-267.
[4] Y Ephraim, D Malah. Speech Enhancement Using a Minimum Mean-square Error Short-time Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1984; 32(6): 1109-1121.
[5] C Plapous, C Marro, P Scalart. Improved Signal-to-noise Ratio Estimation for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*. 2006; 14(6): 2098-2108.
[6] O Cappé. Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor. I*EEE Transactions on Speech Audio Processing*. 1994; 2(2): 345-349.
[7] K Suzumi, S Hiroshi, M Ryoichi, S Kiyohiro, K Kazunobu. *Theoretical Analysis of Musical Noise Generation in Noise Reduction Methods with Decision-Directed a Priori SNR Estimator*. Proceedings of International Workshop on Acoustic Signal Enhancement. Aachen. 2012; 1-4.
[8] X Zhang, H Jiang, J Zhang. *Improved priori SNR estimation for sound enhancement with Gaussian statistical model.* Proceedings of International Conference on Computer Science and Education. Melbourne. 2012; 1307-1310.
[9] T Moazzeni, A Amei, J Ma, Y Jiang, Statistical Model Based SNR Estimation Method for Speech Signals. *Electronics Letters*. 2012; 48(12): 727-729.
[10] K Kondo. *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Berlin: Springer Press. 2012.