Speech Enhancement Based on Noise-Compensated Phase Spectrum

Md. T. Islam and C. Shahnaz
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka,Bangladesh
tauhid_eee@yahoo.com, celia.shahnaz@gmail.com

***Abstract*—In this paper, a noisy speech enhancement method based on noise compensation performed on short time phase spectrum is presented. Here the noise estimate to be used to modify the noisy speech phase spectrum is proposed to be determined by exploiting the low frequency regions of noisy speech of current frame rather than depending only on the initial silence frames. We argue that this approach of noise estimation offers the capability of the tracking the time variation of the non-stationary noise. By employing the noise estimates thus obtained, a procedure is formulated to compensate the distortion in the phase spectrum, which is kept unchanged in the typical speech enhancement methods. The noise compensated phase spectrum is then recombined with the magnitude spectrum to produce a modified complex spectrum prior to synthesize an enhanced frame. Extensive simulations are carried out using NOIZEUS database in order evaluate the performance of proposed method. It is shown in terms of objective measures, spectrogram analysis and informal subjective listening tests that the proposed method consistently outperforms a state-of-the-art method of speech enhancement from noisy speech corrupted by car or babble noise of very low levels of SNR.**

***Index Terms*—Speech Enhancement, Phase Compensation, Noise Estimation**

I. Introduction

The primary objective of speech enhancement is to suppress the perceivable background noise from the noise-corrupted speech without affecting the speech quality. There has been an increasing interest in noisy speech enhancement in a broad range of speech communication applications, such as mobile telephony, speech coding and recognition, and hearing aid devices [1].

Speech enhancement methods can be divided mainly into three categories based on their domains of operation. Time domain methods include the subspace approach [2], frequency domain methods include the spectral subtraction [3], minimum mean square error (MMSE) estimator [4], and Wiener filtering [5], and time frequency-domain methods involve the employment of family of wavelets [6]-[11]. Time domain subspace method provides a tradeoff between speech distortion and residual noise. On the other hand, frequency domain methods provide the advantage of real-time processing with less computational load. The time-frequency domain methods, namely Universal threshold [6], WPF[7], BayesShrink[8], and SURE[9] use thresholding in the wavelet domain as process of removing noise.

In speech analysis, it is commonly believed that human auditory system is phase-deaf. That is why in the conventional spectral subtraction based speech enhancement methods, for synthesizing a clean speech, operations are performed only on the short-time magnitude spectrum and an unaltered short-time phase spectrum is maintained or vice versa. Recently, it has been shown that the phase spectrum is also useful in speech analysis[12]. Here, noise compensation in the phase spectrum of a frame is performed according to the noise information in that particular frame. The main problem lies in considering an empirical value for the compensation constant which actually determines the amount of compensation. In this paper, we develop a new method for noise compensated phase spectrum in which the value of the compensation constant is determined from a mathematical equation which not only removes the necessary noise but also preserves the signal. In this approach, a noise estimation technique which can update the noise estimate using the information in the previous frame as well as that in the current frame is used .

II. PROPOSED METHOD

A windowed noisy speech frame is expressed in the time domain as

$$y[n] = x[n] + v[n], \tag{1}$$

where, $x[n]$ and $v[n]$ represent the windowed version of the clean speech and that of the noise, respectively. In a transform domain, such as frequency domain, eqn. 1 can be expressed as

$$Y[k] = X[k] + V[k], \tag{2}$$

where, $Y[k]$, $X[k]$ and $V[k]$ are the Fast Fourier transforms (FFTs) for frames of noisy speech, clean speech and noise in that order. The FFT $Y[k]$ of $y[n]$ can be written in polar form as

$$Y[k] = |Y[k]|e^{j\angle Y[k]}. \tag{3}$$

In (3), $Y[k]$ denotes the short-time magnitude spectrum and $\angle Y[k]$ denotes the short-time phase spectrum.

A. Noise Compensated Phase Spectrum

The noisy speech signal in the current frame $y^t[n]$ is a real valued signal and therefore, its FFT is conjugate symmetric, ie.

$$Y^t[k] = \{Y^t[N-K]\}^*. \tag{4}$$

The conjugate arise naturally from the symmetry of the magnitude spectrum and anti-symmetry of the phase spectrum. In our approach, the degree to which the conjugates reinforce or cancel during IFFT operation (needed for clean speech synthesis) is obtained by altering their angular relationship. Moreover, we propose the degree of phase spectrum compensation to be dependent on the SNR as well as magnitude of the noise spectrum estimate of the current frame. Therefore, the phase spectrum compensation function that we formulate is given by

$$\phi[k] = \eta\Lambda[k]|\widehat{D}^t[k]| \quad where, \tag{5}$$

where, t is the frame index, η is a real-valued constant, and $\Lambda[k]$ stands for a weighting function and. An estimate of the short time magnitude spectrum of noise in the current speech frame is represented by $|\widehat{D}^t[k]|$.

1) Noise Estimation: We propose to obtain the estimate $|\widehat{D}^t[k]|$ in (5) as

$$|\widehat{D}^t[k]| = \alpha_t|\widehat{V}^t[k]|, \tag{6}$$

For the t-th speech frame in the speech region that appears after a silence period, $|\widehat{V}^t[k]|^2$ represents an initial estimate of the noise power spectrum, which is updated during each silence period as follows

$$|\widehat{V}^t[k]|^2 = \begin{cases} |Y^t[k]|^2 & \text{for t=1} \\ v_n|V^{t_I}[k]|^2 + (1-v_n)|Y^t[k]|^2 & \text{otherwise} \end{cases} \tag{7}$$

where v_n is the forgetting factor and $|\widehat{V}^{t_I}[k]|^2$ represents the estimated noise power spectrum in the immediate last silence frame t_I before the beginning of speech frame in the speech region. In (6), the over-estimation factor used to prevent the overestimation of the noise power spectrum is symbolized by α_t. Since for the noisy speech, the low frequency band $\Delta = [0, 50]Hz$ of the t-th frame contains only noise, the variation of the noisy speech power spectrum is equivalent to the noise power spectrum of that frame. In view of this fact, in order to change the value of α_t for the t-th frame after a silence period, we propose to use the ratio between the powers of $|Y^t[k]|^2$ and $|\widehat{V}^{t_I}[k]|^2$ in the low frequency band Δ of the corresponding frame as

$$\alpha_t = \frac{\sum_{k\varepsilon\Delta}|Y^t[k]|^2}{\sum_{k\varepsilon\Delta}|\widehat{V}^{t_I}[k]|^2}, \tag{8}$$

Thus, the use of α_t defined in (8) clearly serves as a relative weighing factor for the $|\widehat{V}^t[k]|^2$ while computing $|\widehat{D}^t[k]|$ leading to a reasonable tracking for the time variation of the noise if non-stationary.

2) SNR Dependent η: Unlike [12], instead of considering η as a constant, it is proposed as,

$$\eta = \pi^2 e^{\sqrt{\nu}} \tag{9}$$

where, ν is defined as,

$$\nu = \sqrt{\frac{|\widehat{D}^t[k]|^2}{(|Y^t[k]| - |\widehat{D}^t[k]|)^2}} \tag{10}$$

The RHS of (10) is the inverse of the SNR and the plot of η with SNR is shown in Fig. 1. It is seen from Fig. 1 that if the SNR increases, the value of the constant ν decreases and phase compensation becomes less so that there is no distortion in the signal. On the contrary, when noise increases to a higher level, ν increases which in turn increases η. As a result, the phase compensation on the signal increases and denoising is obtained to a significant extent.

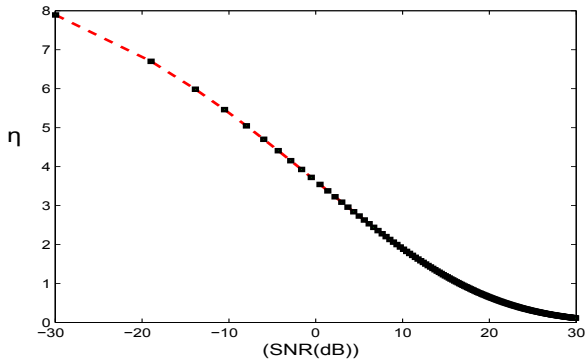

Fig. 1. η as a function of SNR (dB)

The main rational behind such characteristics of η can be justified by the nature of effect of noise on the clean speech signal with SNR. For this purpose, mean square root of difference(MSD) of phase between clean and noisy speech spectrum is exploited, where MSD at a particular SNR can be determined as,

$$MSD = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{P}\frac{(\angle Y^{i,j}[k] - \angle X^{i,j}[k])^2}{MNP}} \tag{11}$$

here $i = 1....M$ denotes file number, $j = 1....N$ denotes frame number and $k = 1....P$ denotes FFT points. The plot of MSD at different SNRs is shown in Fig.2. It is clear from Fig.2 that the noise affect the phase of the clean signal in exponential manner. With increment of signal to noise power, the effect decreases exponentially and at high SNR, effect of noise on the phase reduces to almost a negligible level.

3) Weighting Function $\Lambda[k]$: The weighting function $\Lambda[k]$ in (5) is expressed as

$$\Lambda[k] = \begin{cases} 1 & ,if \quad 0 < \frac{k}{N} < \frac{1}{2} \\ -1 & ,if \quad \frac{1}{2} < \frac{k}{N} < 1 \\ 0 & otherwise \end{cases} \tag{12}$$

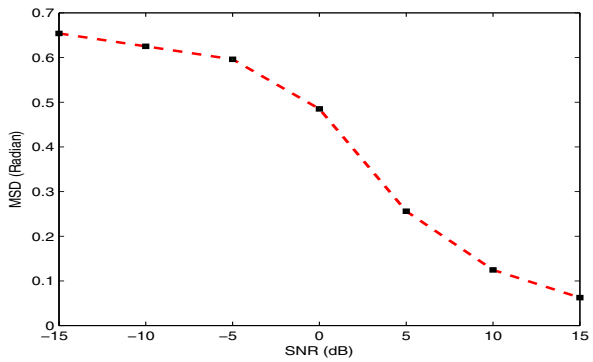

Fig. 2. MSD as a function of SNR

here, zero weighting is assigned to the values of k corresponding to the non-conjugate vectors of FFT, such as $k = 0$ and the value at $k = \frac{N}{2}$ if N is even. Since the estimate of noise magnitude spectrum $|\widehat{D}^t[k]|$ is symmetric, introduction of the weighting function $\Lambda[k]$ defined by (12) produces an anti-symmetric compensation function $\phi[k]$ in (5) that acts as the cause for changing the angular phase relationship in order to achieve noise cancelation during synthesis.

We compute the real value frequency dependent compensating function $\phi[k]$ and utilize it to offset the complex spectrum of the noisy speech as

$$\widehat{X}^t_\phi[k] = Y^t[k] + \phi^t[k]. \tag{13}$$

The strength of the compensation is dependent on the magnitude of both the FFT involving $Y^t[k]$ vectors and the $\phi[k]$ function. Finally , the noise-compensated phase spectrum is obtained from $\widehat{X}^t_\phi[k]$ as

$$\angle\widehat{X}^t_\phi[k] = ARG[\widehat{X}^t_\phi[k], \tag{14}$$

where ARG is a complex angle function. Although $\angle\widehat{X}^t_\phi[k]$ may not possesses the properties of phase spectrum of the clean speech it is capable of tracking the phase compensation required due to noise present in each frame. This is achieved by incorporating the noise estimate $|\widehat{D}^t[k]|$ of the corresponding frame while constructing the compensating function $\phi[k]$ used for computing the noise compensated phase spectrum $\angle\widehat{X}^t_\phi[k]$ in (14).

B. Resynthesis of Enhanced Signal

In the synthesis stage, the noise-compensated phase spectrum is recombined with the magnitude spectrum to produce an enhanced complex spectrum as,

$$\widehat{X}^t[k] = |Y^t[k]|e^{j\angle\widehat{X}^t_\phi[k]}. \tag{15}$$

The enhanced speech frame is synthesized by performing the inverse FFT on the resulting $\widehat{X}^t[k]$,

$$\widehat{x}[n] = IFFT\{\widehat{X}^t[k]\},, \tag{16}$$

where $\widehat{x}[n]$ represents the the enhanced speech frame. The final enhanced speech signal is resynthesized by using the standard overlap and add method.

III. Results

A. Simulation Conditions

Real speech sentences from the NOIZEUS database are employed for the experiments, where the speech data is sampled at 8 KHz [13]. Two different types of noises, such as car and babble are adopted from the NOIZEUS databases [14]. Different signal to noise ratio (SNR) levels ranging from 15 dB to -15 dB are considered. In order to obtain overlapping analysis frames, hamming windowing operation is performed, where the size of each of the frame is 512 samples with 50% overlap between successive frames. To determine the MSD as plotted in Fig. 2, FFT of 50 frames for all thirty files of the the NOIZEUS database for the whole SNR range are used.

B. Comparison Metrics

Standard Objective metrics, namely segmental SNR improvement in dB, Perceptual Evaluation of Speech Quality (PESQ) and Weighted Spectral Slope (WSS) are used for the evaluation of the proposed method[13]. The proposed method is subjectively evaluated in terms of the spectrogram representations of the clean, the noisy and enhanced speech signals. Since we proposed a new approach of speech enhancement exploiting the idea of noise compensation on phase spectrum, it is rational to compare the performance of our method with the existing prominent phase compensation based speech enhancement method such as Phase Spectrum Compensation (PSC) [12] in both objective and subjective senses. However, comparison with the recent time-frequency domain wavelet based speech enhancement methods as proposed in [10] and [11] is beyond the scope and motivation of this present research.

C. Objective Evaluation

Segmental SNR improvement scores for speech signals corrupted with car noise are presented in Fig.3 using PSC and proposed methods. It is found that the proposed method has the higher segmental SNR improvement scores in comparison with the PSC method.

PESQ scores for speech signals corrupted with babble noise are shown in Table.I using PSC and proposed proposed methods. From this table, it is clearly seen that the proposed method outperforms the PSC method.

WSS scores for speech signals corrupted with car noise are plotted in Fig.4 using PSC and proposed methods. It is vivid that the lower values of WSS scores belong to the proposed method which proves its best correlation with the clean signal. The proposed method continue to remain better in terms of WSS even in the babble noise.

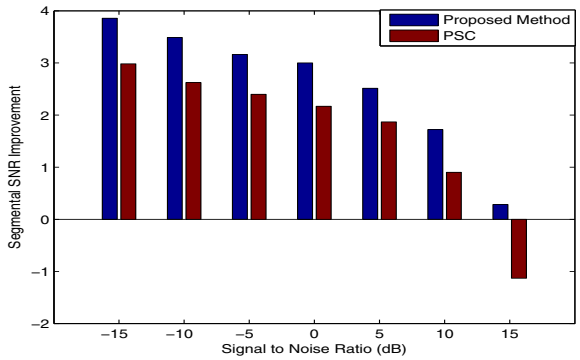

Fig. 3. Segmental SNR Improvement for PSC and proposed methods in the presence of car noise.

TABLE I
COMPARISON OF PESQ SCORES IN BABBLE NOISE.

SNR(dB)	Proposed Method	PSC
-15	1.38	1.29
-10	1.57	1.53
-5	1.73	1.71
0	1.87	1.87
5	2.34	2.12
10	2.84	2.53
15	3.20	2.86

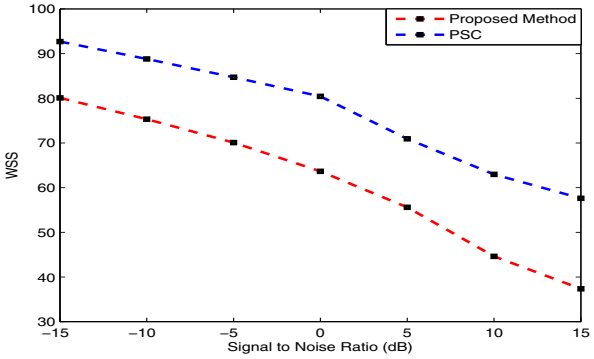

Fig. 4. WSS for PSC and proposed methods in the presence of car noise.

D. Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech obtained by using the proposed method, spectrograms (Time-Frequency plot) of the clean, the noisy, and the enhanced speech signals obtained by using PSC and proposed methods are presented in Fig.5 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are better preserved and the amount of distortion is greatly reduced in the proposed method.

Informal listening tests are also conducted, where the listeners are allowed and arranged to perceptually evaluate the clean, the noisy, and the enhanced speech signals. It is found that the subjective sound quality of the proposed method posses the highest correlation with the objective evaluation in comparison to that of the PSC method in case of car and babble noises considered at different levels of SNR.

IV. CONCLUSIONS

An improved phase compensation approach to solve the problem of speech enhancement has been presented in this paper. We incorporated a better noise estimation approach as well as a noise characteristics driven phase compensation that can track the noise properly irrespective of the stationarity of noise. Phase compensation is a unique method that uses vector summation of noisy signal and noise vector in such a way that the phase as well as magnitude is compensated so that the effect of noise on magnitude and phase both are compensated. But for noise compensation on magnitude spectrum, phase is kept unchanged. At high SNR, we know that the phase is not effected by noise. Therefore, the effect of noise can be ignored at high SNR. But at low SNR, phase of speech signal is heavily affected by noise and the effect must be addressed by a compensation method which is successfully incorporated in our method. Simulation results show that the proposed method yields consistently better results in the sense of higher segmental SNR improvement in dB, higher output PESQ, and lower WSS values than those of the existing phase compensation based method, hence yielding a more enhanced speech.

REFERENCES

[1] OShaughnessy, D., Speech Communications: Human and Machine, 2nd Edition, Wiley-IEEE Press, 1999.
[2] C. H. You, S. N. Koh,, and S. Rahardja, "An invertible frequency eigen domain transformation for masking-based subspace speech enhancement", IEEE Signal Processing Letters, volume.12, no.6, pp. 461-464, June 2005.
[3] K. Yamashita and T. Shimamura, "Nonstationary noise estimation using low-frequency regions for spectral subtractio", Signal Processing Letters, volume. 12, pp. 465-468, 2005.
[4] J. H. L. Hansen, V. Radhakrishnan, and K.H. Arehart, "Speech Enhancement Based on Generalized Minimum Mean Square Error Estimators and Masking Properties of the Auditory System", IEEE Transactions on Audio, Speech, and Language Processing, volume. 14, no.6, pp. 2049-2063, Nov. 2006.
[5] I. Almajai and B. Milner, "Visually Derived Wiener Filters for Speech Enhancement", IEEE Transactions on Audio, Speech, and Language Processing, volume. 19, no.6, pp. 1642-1651, Aug. 2011.
[6] D.L. Donoho, "De-noising by soft-thresholding", IEEE Transactions on Information Theory, volume. 41, pp. 613-627, 1995.
[7] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator", IEEE Signal Processing Letters, vol. 8, pp. 10-12, 2001.
[8] S. G. Chang, B. Yu, and M. Vetterli, " Adaptive wavelet thresholding for image denoising and compression", IEEE Trans. on Image Proc., vol. 9, 1532-1546, 2000.
[9] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Inter-scale orthonormal wavelet thresholding", IEEE Trans. Image Process., vol. 16, no. 3, pp. 593-606, Mar. 2007.
[10] T.F. Sanam and C. Shahnaz, "A Semisoft Thresholding Method based on Teager Energy Operation on Wavelet Packet Coefficients for Enhancing Noisy Speech", EURASIP Journal on Audio, Speech, and Music Processing, doi:10.1186/1687-4722-2013-25, 2013.
[11] T.F. Sanam and C. Shahnaz, "Noisy Speech Enhancement based on an adaptive threshold and a modified hard thresholding function in wavelet packet domain", Digital Signal Processing, Elsevier, vol: 23, pp. 941-951, 2013.

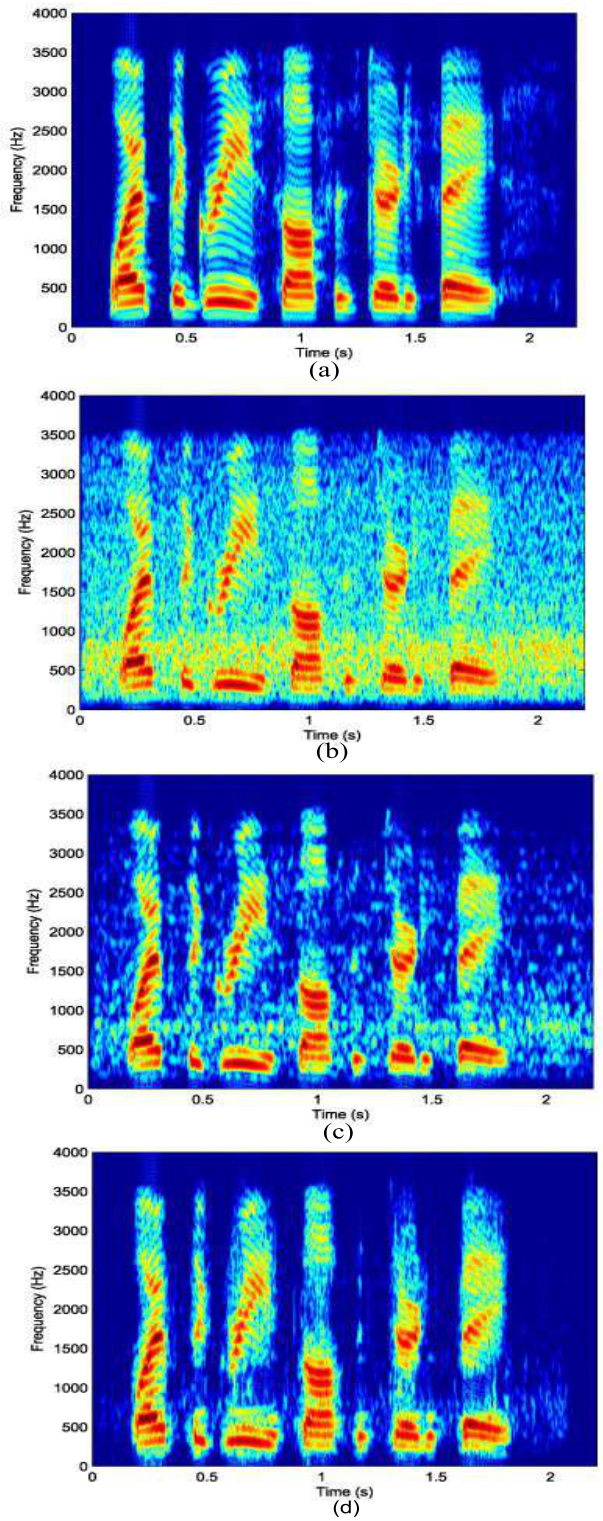

Fig. 5. Spectograms (a) Clean Signal (b) Noisy Signal at 10 dB car noise (c) PSC Method (d) Proposed Method

[12] K. Wjcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, Exploiting Conjugate Symmetry of the Short-Time Fourier Spectrum for Speech Enhancement. IEEE Signal Processing Letters, vol. 15, pp.461-464, 2008.

[13] Y. Hu, and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Transactions on Audio, Speech, and Language Processing, volume. 16, pp. 229-238, 2008.

[14] Y. Lu, P.C. Loizou, Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty, IEEE Trans. Audio, Speech, Lang. Process. 19 (2011) 1123-1137.