On the Importance of Phase and Magnitude Spectra in Speech Enhancement

Erfan Loweimi*, *Member*, *IEEE*, Seyed Mohammad Ahadi**, *Senior Member*, *IEEE*, and Samira Loveymi***

* Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran , eloveimi@aut.ac.ir
** Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran, sma@aut.ac.ir
***Shahid Chamran University, Ahwaz, Iran, s.loveymi@scu.ac.ir

Abstract: *The aim of this paper is to investigate the importance of phase and magnitude spectra in speech enhancement at different conditions with emphasizing on the role of phase spectrum. The speech signal is exposed to additive noise in different SNRs. Then, it is decomposed into different frame lengths from 32 to 1024 ms. In synthesis stage we have used clean phase spectrum along with noisy magnitude spectrum as well as clean magnitude spectrum along with noisy phase spectrum. The quality of speech is evaluated by PESQ objective measure. The maximum speech quality improvement in SNR of 0 dB in case of using clean phase is 1.1 in PESQ scale obtained in frame length of 128 ms and in case of using clean magnitude spectrum is 2.2 in PESQ scale in frame length of 32 ms. Finally, we have shown that phase spectrum cleaning for female speakers leads to more speech quality improvement and in case of male speakers magnitude spectrum enhancement is more useful.*

Keywords: Speech enhancement, phase spectrum cleaning, magnitude spectrum cleaning, PESQ.

1. Introduction

It is well established that the phase spectrum of speech signal does not play a significant role in speech processing. Taking a glance on methods of speech enhancement and feature extraction algorithms for automatic speech recognition (ASR) proves this claim. In most of speech enhancement methods, such as spectral subtraction [1] and MMSE [2], all of the enhancement process is focused on magnitude spectrum. At the end, the phase spectrum of noisy signal is used along with enhanced magnitude spectrum to synthesize the enhanced speech. On the other hand, in most of feature extraction algorithms such as MFCC and LPC it is the magnitude spectrum that plays the major role and the phase spectrum is discarded.

The first study which explored the importance of phase spectrum in speech enhancement was conducted by Wang and Lim [3]. They synthesized the speech signal from phase and magnitude spectra which were extracted in different SNRs. The speech signals were decomposed into frame lengths of 50 and 400 ms. They utilized Hanning window with 50% overlap. Phase spectrum was extracted from speech with higher SNR. The results show that phase spectrum (with higher SNR) in case of 50 ms frame length does not have any remarkable influence on the quality of reconstructed speech. However, in frame length of 400 ms, the quality of reconstructed speech was notably influenced by the phase spectrum.

In [4], Shannon and Paliwal, explored the importance of the phase spectrum in speech enhancement. The clean signal decomposed into frame length of 32 ms along with different type of windows such as Chebyshev, rectangular, and Hamming windows with 87.5% overlap. The noisy speech framed in a similar manner with applying Hamming window. The magnitude spectrum of noisy speech mixed with the phase spectrum of clean signal and the quality of reconstructed speech evaluated with PESQ [5]. Results show that rectangular and Chebyshev windows with dynamic range of 30 to 35 dB result in more quality improvement.

In [6] Wójcicki *et al.* proposed a novel speech enhancement algorithm which was focused only on phase spectrum and passed the noisy magnitude spectrum directly to the output. They proposed to add the short-time Fourier transform (STFT) of noisy speech with Λ which is defined as following

$$\Lambda(n,k) = \begin{cases} +\lambda & 0 \le k < \frac{N}{2} \\ -\lambda & \frac{N}{2} \le k \le N-1 \end{cases}, \qquad (1)$$

where k denotes the kth discrete frequency of N uniformly spaced frequencies. Then, they replaced the magnitude spectrum with its noisy version. Hence, only the phase spectrum was modified. Their proposed method increases the PESQ of speech up to 0.6 in case of applying suitable λ. This method shows

notable ability to attenuate the background noise. The weakness of this method is that the suitable value of λ depends on signal to noise ratio (SNR) and noise type. Dealing with this problem, Stark *et al.* modified the definition of Λ in the following way [7]

$$\hat{\Lambda}(n,k) = \Lambda(n,k)\left|\hat{D}(n,k)\right|, \quad (2)$$

where $\hat{\Lambda}$ is modified form of Λ and $\left|\hat{D}(n,k)\right|$ is an estimation of noise magnitude spectrum. In this case, the maximum improvement of speech in PESQ scale reached to 0.65. The main advantage of this modification is that the suitable value of λ is no longer depends on SNR. In this case, the appropriate value of λ was 3.74 [7]. However, estimating the noise magnitude spectrum is not an easy task in practice. For more details readers can refer to [8].

The aim of this paper is exploring the importance of phase and magnitude spectra in speech enhancement in different situations, i.e. different window lengths, SNRs, and window types. The emphasis of this paper is on the role of phase spectrum. We will contaminate the speech signal with noise, in different SNRs. Then, by reconstructing the signal through combining the noisy magnitude spectrum with phase spectrum of clean speech, we will achieve to a quantitative assessment about the importance of phase spectrum. We will repeat the experiments for clean magnitude spectrum along with noisy phase spectrum. To our knowledge, the relative importance of phase and magnitude spectra in different SNRs and frame lengths is not explored yet. We will show that in higher SNRs and longer frame lengths the relative importance of phase spectrum in comparison with magnitude spectrum will increase.

All of the works cited here are based on Analysis-Modification-Synthesis (AMS) framework. Here, we have used similar framework with what has been applied in [4]. Results show that phase spectrum even in short frame lengths such as 32 ms can have notable effect on the quality of speech signal. However, the maximum quality improvement due to phase enhancement is lower than that of magnitude spectrum enhancement.

The organization of the rest of this paper is as follows. In Section 2 we will briefly review analysis-modification-synthesis (AMS) framework. In Section 3 the utilized objective measure to score the quality of speech will be introduced. In Section 4 the simulation results and their analysis will be presented and Section 5 concludes the paper.

2. Analysis-Modification-Synthesis (AMS) Framework

This framework, as its name shows, is constituted from three main parts. The first step is analysis. Since the speech signal, *x(n)*, is a quasi-stationary signal, it should be analyzed in a frame-wise manner in which stationarity assumption held within each frame through short-time Fourier transform (STFT)

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)\exp\left(-\frac{j2\pi km}{N}\right), \quad (3)$$

where k denotes the kth discrete frequency of N uniformly spaced frequencies, *w(n)* is an analysis window, and n indicates the index of STFT. In speech processing, usually Hamming window with duration of 20-40 ms is applied. $X(n,k)$ is a complex quantity, so it can be rewritten in the polar form as follows

$$X(n,k) = |X(n,k)|\exp\left(j\angle X(n,k)\right), \quad (4)$$

where $|X(n,k)|$ and $\angle X(n,k)$ are short-time magnitude and phase spectra, respectively. Hereafter, the modifier 'short-time' is implied wherever the phase spectrum and/or magnitude spectrum were mentioned.

The next step is modification in which purposely some modifications are made on the spectrum of the signal. For instance, in order to investigate the importance of phase spectrum, one can reconstruct the signal only from its phase spectrum by setting the magnitude spectrum to unity [9-11]. Here, in order to investigate the importance of phase spectrum in speech enhancement, we replaced the phase spectrum of noisy speech with the phase spectrum of its corresponding clean speech

$$\hat{X}(n,k) = |X(n,k)|\exp\left(j\angle S(n,k)\right), \quad (5)$$

where $S(n,k)$ is STFT of clean speech $(s(n))$, $|X(n,k)|$ is the magnitude spectrum of noisy speech, and $\hat{X}(n,k)$ is modified STFT. Similarly, in order to investigate the importance of magnitude spectrum in speech enhancement, the noisy magnitude spectrum is substituted by its clean version

$$\hat{X}(n,k) = |S(n,k)|\exp\left(j\angle X(n,k)\right). \quad (6)$$

The next and last step is synthesis. In this stage, the inverse of Fourier transform should be computed in each frame. Then, by the use of synthesis methods such as overlap-add (OLA) or least square error estimation (LSEE) [12], the signal will be reconstructed. Here, we have used both LSEE and OLA. The advantage of LSEE is that after modifying the spectrum in the previous step (modification), some problems may arise. In fact, there is no guarantee that the modified spectrum remains valid. Thus, there will be no signal in time domain with such spectrum. LSEE tries to find a signal with the most similar spectrum to modified spectrum in sense of mean square error (MSE).

3. Quality Assessment

There are two main approaches for speech quality assessment, subjective and objective. The subjective tests are preferred, but they suffer from some problems. They should be conducted on a large population in order to result in reliable outcomes. This will make them both time-consuming and costly. The objective measures do not suffer from these two problems and can be computed quickly with computers but their reliability is strictly questionable. Here, we used PESQ [5] objective measure. This selection was based on the results which were reported by Hu and Loizou [13] showing that PESQ has the maximum correlation with subjective tests in comparison with other objective measures (correlation coefficient = 0.89).

PESQ was proposed by ITU-T to evaluate the quality of speech in telephone handsets and narrowband speech codecs [5]. It represents an aggregation of PAMS and PSQM99. These two algorithms were the highest performing methods in ITU-T competition that was held to find a more robust objective speech quality measure. Among all of the objective measures PESQ has the most complexity and computational cost. PESQ produces a score between 1.0 and 4.5, with high values indicating better quality.

4. Experiments, Simulation Results, and Analysis

Experiments were conducted on all of the 30 speech utterances of NOIZEUS database [14]. This database is composed of gender and phonetically balanced utterances. The speeches were originally sampled at 25 kHz and downsampled to 8 kHz with 16-bit precision per sample. The speech signals were reconstructed from the frame lengths of 32, 64, 128, 256, 512, and 1024 ms via OLA and LSEE. Frame shift was set to one eighth of the frame length (87.5% overlap) and FFT size is set to $2N$ where N is the number of samples of each frame. These two selections were done to minimize the aliasing error [9]. Magnitude spectrum was extracted from the frames which were windowed with Hamming window because of its compatibility with magnitude spectrum [9-11]. In case of phase spectrum, we have windowed the speech frames with Chebyshev window with 35 dB dynamic range, due to [4]. We have used additive white Gaussian noise (AWGN) in different signal to noise ratios i.e. -5, 0, 5, and 15 dB in order to make the speech signal noisy. The speech signals were reconstructed in two stages, from clean magnitude with corresponding noisy phase spectrum and from noisy magnitude spectrum along with clean phase spectrum in different situations. The baseline for comparison is the quality of noisy signal. The last point is the type of synthesis window. We have used both the Hamming and Chebyshev windows.

Figs. 1, 2, and 3 illustrate that the importance of phase and magnitude spectra in speech enhancement depends directly on the frame length, signal to noise ratio, window type, and synthesis method. It is very important to consider all of the aforementioned factors simultaneously. In [15], Shi, Modir Shanechi, and Aarabi through a subjective recognition tests show that the importance of phase spectrum in lower SNRs becomes more. They showed that synthesizing the noisy speech with clean phase spectrum (along with noisy magnitude spectrum) in lower SNRs have more significant influence on the recognition rate in comparison with higher SNRs. Finally, they inferred that the importance of phase spectrum in lower SNRs becomes more and concluded that phase spectrum can play a significant role in robust speech recognition. There is a point which should be noted. In fact, with decreasing the SNR any information of clean signal, either of phase or magnitude spectra, becomes more valuable. In instance, the clean phase spectrum combined with noisy magnitude spectrum in SNR of 0 dB in comparison with +5 dB, will be more helpful and has more constructive influence on the speech quality and recognition rate. The same is true for magnitude spectrum. Fig. 1 proves this point. Actually, the importance of phase spectrum should be compared relatively with magnitude spectrum. By the way, this comparison must be based on considering window length, window type, and signal to noise ratio. Neglecting each factor would be misleading. We will discuss the role of each factor.

As seen in Figs. 1, 2, and 3, by increasing the length of window the importance of magnitude spectrum decreases due to the non-stationarity of speech signal. In case of phase spectrum cleaning, it was expected that by extending the frame length the importance of phase spectrum increases [9-11], but the maximum quality improvement appears in frame length of 128 and 64 ms, depending on the synthesis method. In [10] and [11] the speech reconstruction was done in an iterative manner but here the synthesis process is not iterative. The quality of speech signal depends on the information which is provided by both phase and magnitude spectra. By increasing the frame length although the phase spectrum becomes more informative, the magnitude spectrum information will be less valuable, so, there is a trade-off. As a result, the maximum information provided by both phase and magnitude spectra is in frame length of 128 and 64 ms, in case of utilizing LSEE and OLA, respectively.

As the Figs. 1, 2, and 3 show, in case of phase cleaning the LSEE has better performance in comparison with OLA. LSEE intensifies the role of the window, consequently it seems that in any application in which the applied window provide a

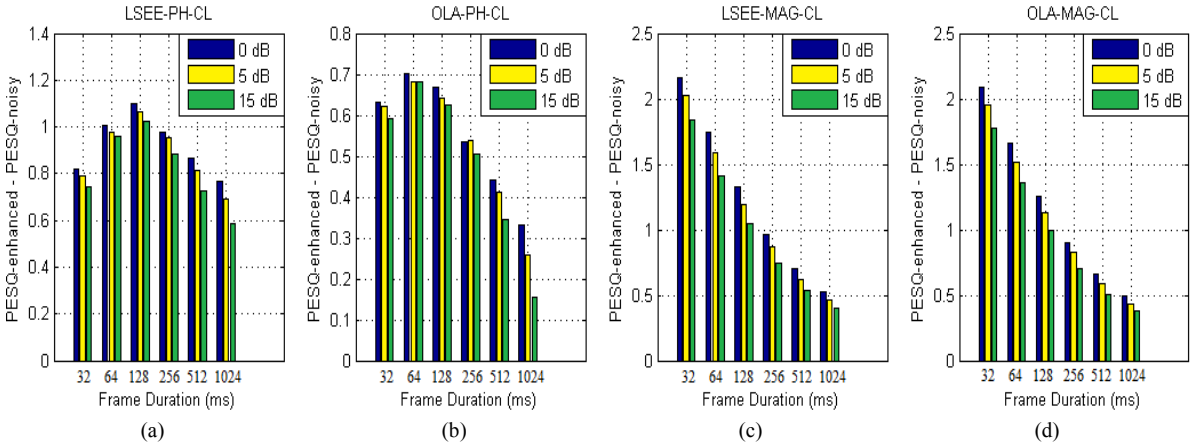

Fig. 1: Effect of cleaning the phase and magnitude spectra at different SNRs on the quality of speech signal. (a) LSEE (Synthesis method) - PH-CL (Phase Cleaning to investigate the importance of phase spectrum), (b) OLA-PH-CL, (c) LSEE-MAG-CL, (d) OLA-MAG-CL.

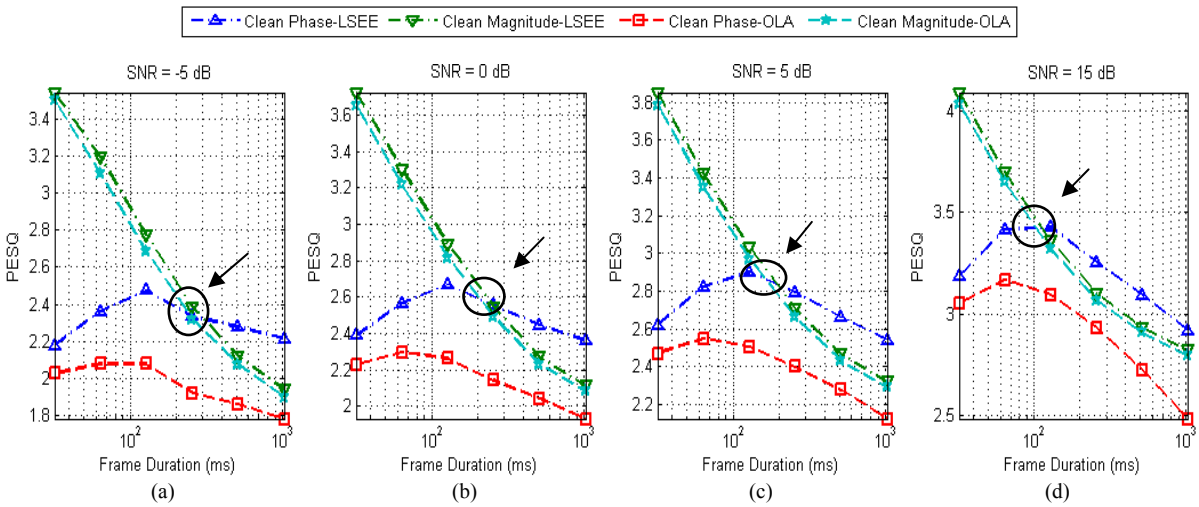

Fig. 2: Effect of phase and magnitude spectra cleaning at different SNRs and frame lengths in case of applying Hamming window in synthesis stage. Apparently, in case of using OLA for synthesis, there will be no crossover point. (a) SNR = -5 dB, (b) SNR = 0 dB, (c) SNR = 5 dB, (d) SNR = 15 dB.

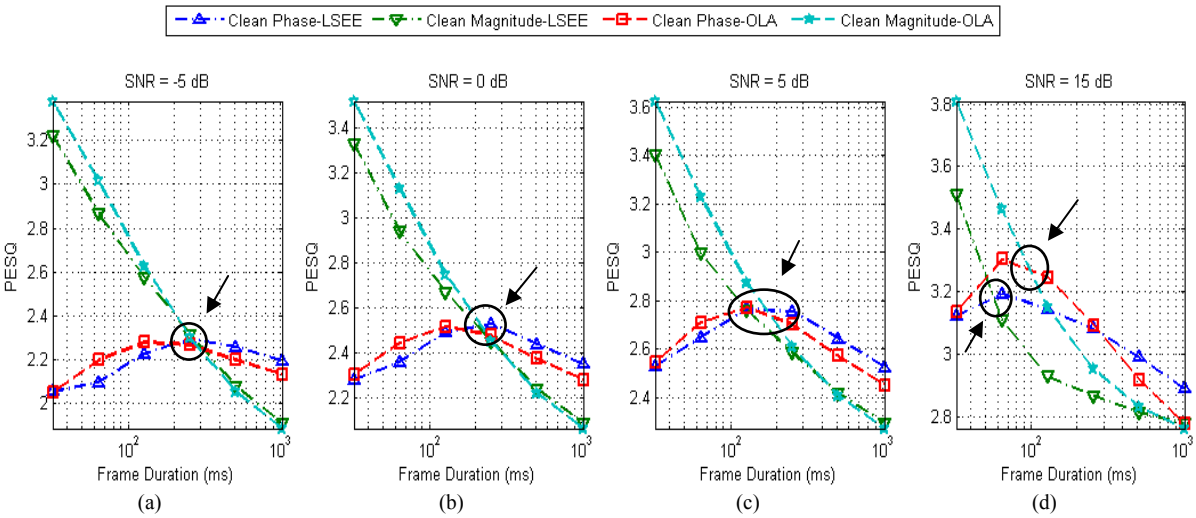

Fig. 3: Effect of phase and magnitude spectra cleaning at different SNRs and frame lengths in case of applying Chebyshev window with dynamic range of 35 dB in synthesis stage. (a) SNR = -5 dB, (b) SNR = 0 dB, (c) SNR = 5 dB, (d) SNR = 15 dB.

better smear-leakage trade-off in comparison with rectangular window LSEE surpasses OLA[1].

As seen in Figs. 2 and 3, the crossover point, in which the role of phase spectrum becomes more important than magnitude spectrum, depends on SNR, synthesis method, and synthesis window. In case of utilizing LSEE, by increasing the SNR the crossover point shifted backward and occurs in shorter frame lengths. This shows that by increasing the SNR, the relative importance of phase spectrum in comparison with magnitude spectrum becomes more. Fig. 2 shows that in case of using OLA along with Hamming window as synthesis window, there is no crossover point. The black circles in Fig. 2 and 3 illustrate the crossover points.

Comparing Fig. 1 (a) and (b) with Fig. 1 (c) and (d) shows that cleaning the magnitude spectrum results in more quality improvement in comparison with cleaning the phase spectrum. However, there is no enhancement method that could return the noisy magnitude spectrum to its clean state. Consequently, in practical circumstances this advantage of magnitude spectrum cleaning is not very beneficial, at least as much as it appears to be. Thus, it is not an adequate reason for putting phase spectrum aside and just focusing on magnitude spectrum.

Another point is the window type. As said before in this framework we have three windows, two for analyzing and one for synthesizing. It seems that for analyzing the speech and working with its magnitude spectrum Hamming window is an appropriate choice [9-11]. In case of analyzing the signal for working with its phase spectrum in speech enhancement it has been shown in [4] and [9-11] that the rectangular or Chebyshev window with dynamic range of 30 to 40 dB are suitable choices. Due to these reports we have utilized Chebyshev window with 35 dB dynamic range. Now a question arises that which window should be used in synthesis stage? We have reconstructed the speech signals with both of them. By comparing Figs. 2 and 3 one can obviously see that the Hamming window in comparison with Chebyshev window is a more appropriate choice for synthesizing the speech signal because it better provides the smear-leakage trade-off which is required in this application.

As a final point, we repeated the experiments in order to study the relationship between the gender of speaker and the importance of phase and magnitude spectrum in speech enhancement. As seen in Figs. 4 and 5, cleaning (enhancing) the phase spectrum leads to more quality improvement in case of female speakers while cleaning the magnitude spectrum results in more quality improvement for male speakers.

[1] In case of applying rectangular window both LSEE and OLA lead to the same results.

5. Conclusion

In this paper we investigated the importance of phase and magnitude spectra of speech signal in different conditions i.e. different frame lengths, window shapes, SNRs, and synthesis methods. We have reconstructed the speech signal from noisy phase spectrum along with the magnitude spectrum of clean speech. Similarly, we reconstructed the speech signal from the magnitude spectrum of noisy speech along with the phase spectrum of clean speech. Our simulation results show that frame length and window type has notable effects on the relative importance of these two spectra. As well, we observed that by increasing the SNR the relative importance of phase spectrum in comparison with the magnitude spectrum increases. LSEE leads to better results in comparison with OLA in case of using appropriate window. Finally, we showed that magnitude spectrum enhancement in case of male speakers is more beneficial whereas in case of female speakers phase spectrum enhancement results in more speech quality improvement.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[3] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 679–681, Aug. 1982.

[4] B. Shannon, and K. Paliwal, "Role of Phase Estimation in Speech Enhancement", in *Proc. Int. Conf. Spoken Language Processing (INTERSPEECH-ICSLP)*, Sep. 2006.

[5] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, UT, 2001, vol. 2, pp. 749–752.

[6] K. Wójcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement," *IEEE Signal Process. Lett.*, vol. 15, pp. 461–464, 2008.

[7] A. Stark, K. Wójcicki, J. Lyons, and K. Paliwal, "Noise driven short-time phase spectrum compensation procedure for speech enhancement," in *Proc. Int. Conf. Spoken Language Processing (INTERSPEECH-ICSLP)*, Sep. 2008.

[8] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech communication*, vol. 53, pp. 465-494, Dec. 2010.

[9] L. Alsteris and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digit. Signal Process.*, vol. 17, pp. 578–616, May 2007.

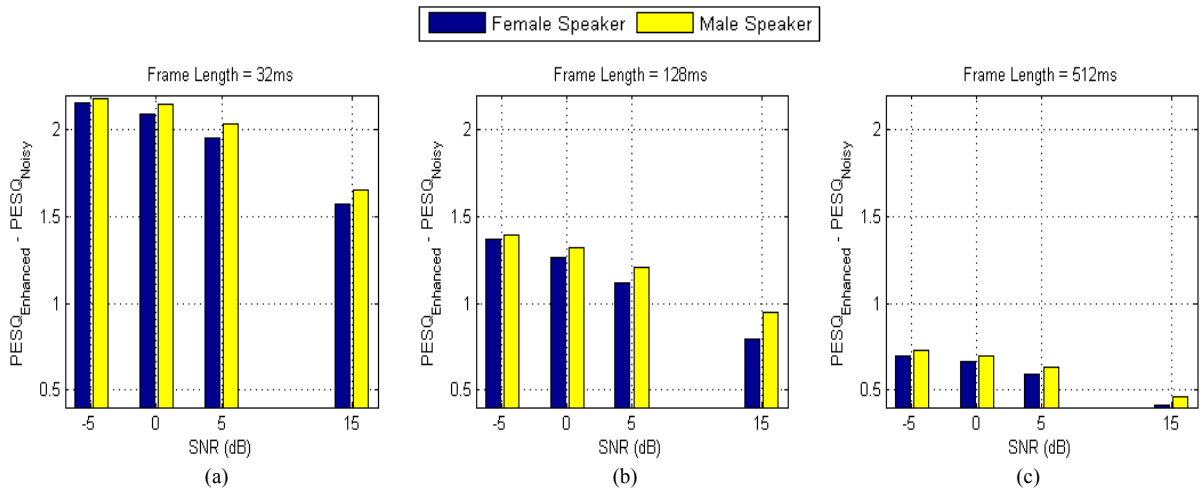

Fig. 4: Speech enhancement through magnitude spectrum cleaning (using the magnitude spectrum of clean signal) in case of male and female speakers for different frame lengths. (a) 32 ms. (b) 128 ms, (c) 512 ms.

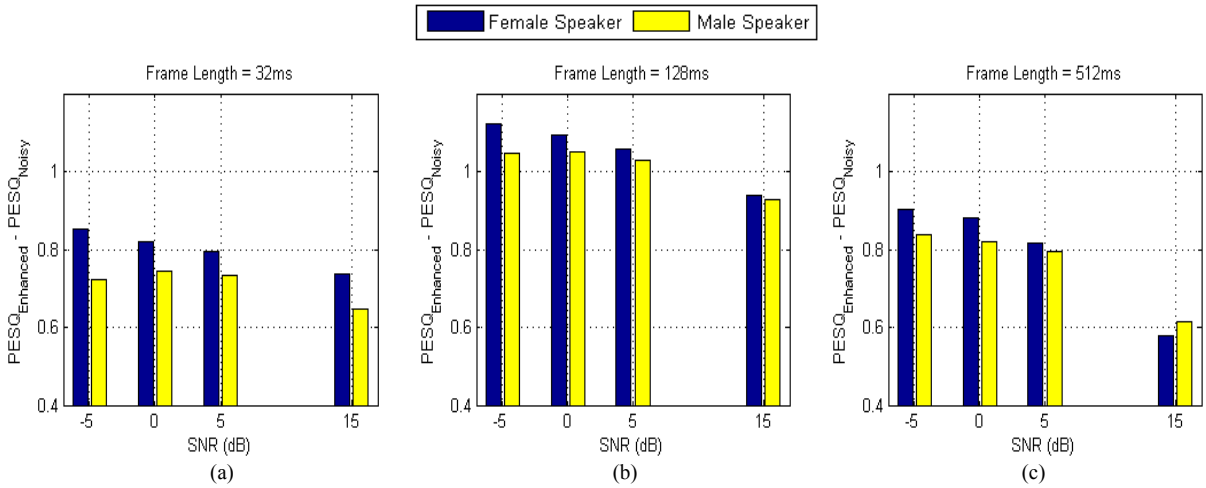

Fig. 5: Speech enhancement through phase cleaning (using the phase spectrum of clean signal) in case of male and female speakers for different frame lengths. (a) 32 ms. (b) 128 ms, (c) 512 ms.

[10] E. Loveimi and S.M. Ahadi, "*Objective Evaluation of Magnitude and Phase Only Spectrum-based Reconstruction of the Speech Signal",* in *Proc. Int. Symp.* On Communications, Control and Signal Processing (*ISCCSP 2010*), Limassol, Cyprus, Mar. 2010.

[11] E. Loveimi and S.M. Ahadi, "*Objective Evaluation of Magnitude and Phase Only Reconstructed Speech: new considerations",* in *Proc. ISSPA 2010.*

[12] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[13] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement*," IEEE Trans. Audio, Speech, Lang. Process.* 16, 229–238, Jan. 2008.

[14] Y. Hu, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," http://www.utdallas.edu/loizou/speech/noizeus, 2005.

[15] G. Shi, M. Modir Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, Vol.14, No.5, pp1867- 1874, Sep. 2006.