A Powerful Method for Extracting the Original Signal from the Noisy Input Signal by using the Iterative Reconstruction Framework of the Short Time Fourier Transformation

Saeed Karimi

***Abstract*— In the current system for the reconstruction of speech, it is used iterative reconstruction framework of short time Fourier transformation (STFT). When a lot of noise is added to the input speech, the iterative reconstruction framework of STFT can not properly reconstruct the input speech. In this paper, we offer a method that by using it, we will be able to extract the original signal from the noisy input signal. For achieving this goal, in each step of the iterative reconstruction of input speech, by taking a threshold value For magnitude spectrum, we remove the smaller amounts of threshold and for preserving the original signal features, we consider the phase spectrum then with combination of the modified magnitude spectrum information and the phase spectrum information, we reconstruction part of the signal in each iteration. Two experiments were examined. In the first experiment, we evaluated the reconstruction of input speech with changing the threshold value and in the second experiment, we evaluated the reconstruction of input speech with different numbers of iterations. The results showed that by using our method, when a lot of noise is added to the input speech, we can reconstruct the original signal very well.**

***Keywords*: Iterative reconstruction, threshold, magnitude spectrum, phase spectrum.**

I. Introduction

In automatic speech recognition (ASR), the speech is processed frame-wise using temporal window duration of 20–40 ms. The STFT is normally used for the signal analysis of each frame. The resulting signal spectrum can be decomposed into the magnitude spectrum and the phase spectrum. At such small temporal window durations, it is generally believed that the phase spectrum does not contribute much to speech intelligibility [1] and, as a result, state-of-the-art ASR systems generally discard the phase spectrum in favor of features that are derived only from the magnitude spectrum [2]. have determined that such signals can be uniquely specified by the signed magnitude spectrum (magnitude spectrum with one bit of phase spectrum information) [3]. Phase spectrum, Including two independent variables, frequency and time. According to these variables the phase spectrum, other researchers were able to derive the frequency and time of the phase spectrum, the signal to reconstruct [4]. Other experiments indicate that the STFT with a long window is more effective for automatic speech recognition [5]. other results were obtained from using a combination of information of the phase spectrum and magnitude spectrum. so that, if smaller modulation frame durations improve intelligibility when processing the modulation magnitude spectrum, while longer frame durations improve intelligibility when processing the modulation phase spectrum that showed both components the magnitude spectrum and phase spectrum are important For the better reconstructing of speech [6].

In this paper, we show that by modifying any part of the speech input at each iteration, by increasing threshold value and the number of iterations for the reconstruction the input signal, we can extract the original signal from input noisy signal. For achieving this goal, we offer a framework, using a combination of iterative reconstruction STFT, it was introduced in [7] and Analysis–modification–synthesis (AMS) that was introduced in [8]. First, they introduced and then we modify it to be able to present your method. This paper is organized as follows: in section II description is given the proposed method, in section III, it is surveyed several experiments and the results of each one and finally, conclusions are given in section IV.

II. The Proposed Method

For making the proposed method, first, we briefly describe the iterative reconstruction framework STFT and AMS framework, and then we introduce it with combination of these two frameworks.

A. *Reconstruction within the STFT framework*

In this method, the short-time sections are reconstructed in the order determined by their positions on the time axis. Each section is determined by its known spectral information as well as the known samples in the region of overlap with previous sections. The framework for the method we use is illustrated in Fig. 1.

Manuscript received May 10, 2016

Saeed Karimi, Department of Computer, Islamic Azad university, Dehloran Branch, iran, Tel NO 00988433720222

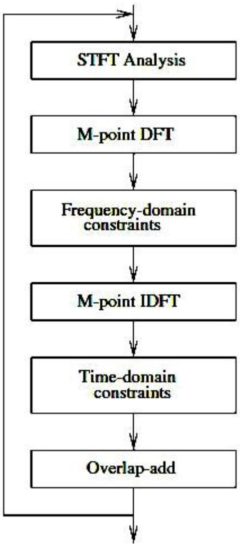

Fig 1: STFT-based iterative reconstruction framework.

B. Analysis–modification–synthesis

Traditional acoustic-domain short-time Fourier AMS framework consists of three stages: (1) the analysis stage, where the input speech is processed using STFT analysis; (2) the modification stage, where the noisy spectrum undergoes some kind of modification; and (3) the synthesis stage, where the inverse STFT is followed by overlap-add synthesis (OLA) to reconstruct the output signal. For a discrete-time signal x (n), the STFT is given by

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)e^{-j2\pi kl/N} \quad (1)$$

Where n refers to the discrete-time index, k is the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples), and w(n) is the acoustic analysis window function. In speech processing, an acoustic frame duration of 20–40 ms is typically used [9], [10], with a hamming window (of the same duration) as the analysis window function. In polar form, the STFT of the speech signal can be written as

$$X(n,k) = |X(n,k)|e^{j\angle X(n,k)} \quad (2)$$

Where $|X(n,k)|$ denotes the acoustic magnitude spectrum and $\angle X(n,k)$ denotes the acoustic phase spectrum. In the modification stage of the AMS framework, either the acoustic magnitude or the acoustic phase spectrum or both can be modified. Let $|Y(n,k)|$ denote the modified acoustic magnitude spectrum, and $\angle Y(n,k)$ denote the modified acoustic phase spectrum. Then, the modified STFT is given by

$$Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)} \quad (3)$$

Finally, the synthesis stage reconstructs the speech by applying the inverse STFT to the modified acoustic spectrum, followed by least-squares overlap-add synthesis [11]. Here, the modified Hamming window [7] given by

$$w_s(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi(n+0.5)}{N}\right), & 0 \le n < N \\ 0, & otherwise \end{cases} \quad (4)$$

is used as the synthesis window function. A block diagram of the acoustic AMS procedure is shown in Fig. 2.

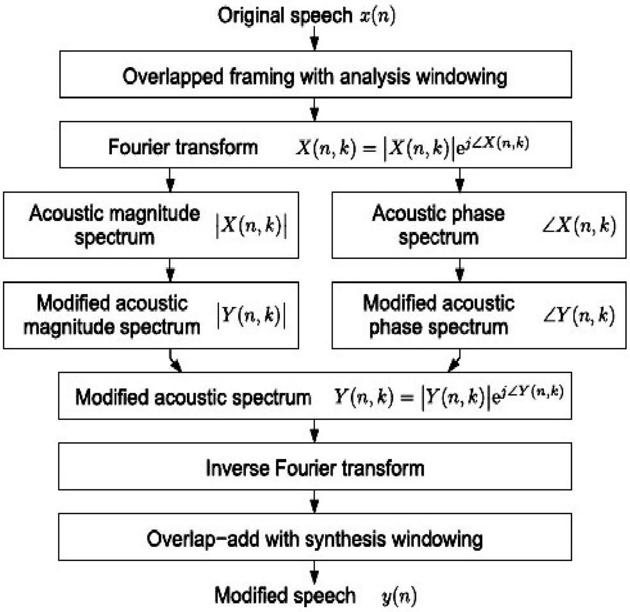
Original speech $x(n)$
Overlapped framing with analysis windowing
Fourier transform $X(n,k) = |X(n,k)|e^{j\angle X(n,k)}$
Acoustic magnitude spectrum $|X(n,k)|$
Acoustic phase spectrum $\angle X(n,k)$
Modified acoustic magnitude spectrum $|Y(n,k)|$
Modified acoustic phase spectrum $\angle Y(n,k)$
Modified acoustic spectrum $Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)}$
Inverse Fourier transform
Overlap-add with synthesis windowing
Modified speech $y(n)$

Fig 2: Block diagram of the acoustic AMS procedure.

C. Iterative reconstruction by using the proposed method

According to the combination of both components magnitude and phase spectra are important for the understanding signal reconstruction. In the proposed plan, it is used a combination of the magnitude spectrum information and phase spectrum. The goal that we follow is to reconstruct the original signal from the noisy input signal. For achieving this goal, we have used the AMS framework and iterative reconstruction of the STFT, in which by modifying the magnitude spectrum and then combine it with the phase spectrum, we reconstruct Part of the original signal in each iteration. Its block diagram is shown in Fig. 3 that in which with the arrival of the noisy signal, short-time analysis window at each iteration is applied the input signal and extract a part of the spectral information. Considering the certain amount of threshold and we apply on the magnitude spectrum. If the values of magnitude spectrum are less than the threshold value, it replaced with a zero value otherwise it will be preserved. According to formula (2) if $|X(n,k)|$ is the magnitude spectrum of the input signal is done this work as follows:

$$|Y(n,k)| = \text{Treshold}\,(\,|X(n,k)|\,) \quad (5)$$

Then we combined information the magnitude spectrum of modified and information the phase Spectrum of input signal so that, if $|Y(n,k)|$ the magnitude spectrum of

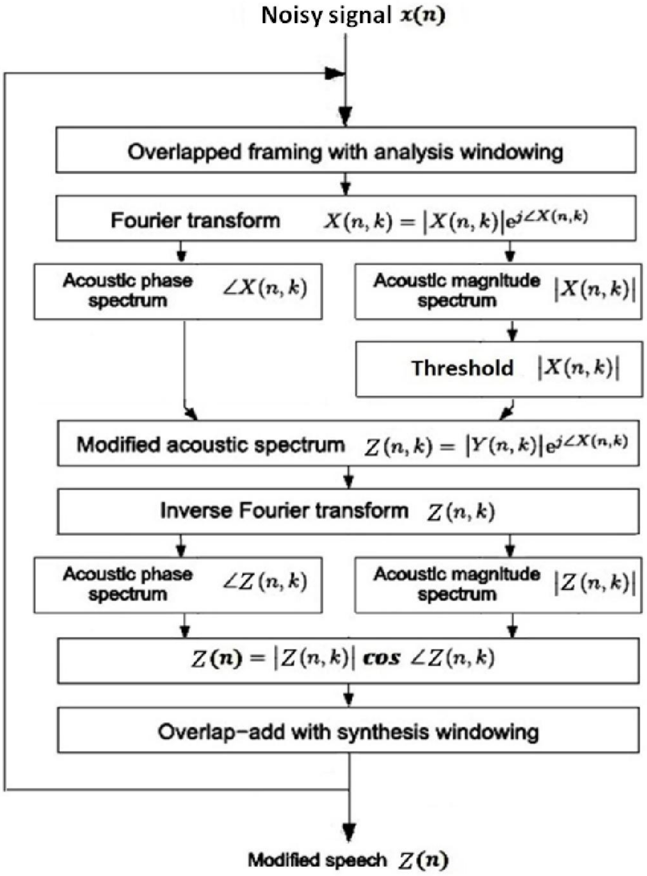
Noisy signal $x(n)$
Overlapped framing with analysis windowing
Fourier transform $X(n,k) = |X(n,k)|e^{j\angle X(n,k)}$
Acoustic phase spectrum $\angle X(n,k)$
Acoustic magnitude spectrum $|X(n,k)|$
Threshold $|X(n,k)|$
Modified acoustic spectrum $Z(n,k) = |Y(n,k)|e^{j\angle X(n,k)}$
Inverse Fourier transform $Z(n,k)$
Acoustic phase spectrum $\angle Z(n,k)$
Acoustic magnitude spectrum $|Z(n,k)|$
$Z(n) = |Z(n,k)| \cos \angle Z(n,k)$
Overlap–add with synthesis windowing
Modified speech $Z(n)$

Fig 3: Block diagram iterative reconstruction of the proposed method.

modified and $\angle X(n,k)$ information the phase Spectrum of input signal, we have:

$$Z(n,k) = |Y(n,k)|e^{j\angle X(n,k)} \tag{6}$$

Then we apply the inverse short-time Fourier transform. If $|Z(n,k)|$ the information of magnitude spectrum and $\angle Z(n,k)$ is information the Phase Spectrum of combined signal, for reconstruction Part of the input signal in the each iteration, we use the following formula:

$$Z(n) = |Z(n,k)|\cos \angle Z(n,k) \tag{7}$$

At the end of each iteration, we created a part of the speech input at each iteration, considering the overlapping area between the analysis windows, with the parts of speech in previous iterations, combined and going forward modification of the speech Spectrum. Operations of Iterative reconstruction the input speech with this method, will be repeated until the total of input signal is reconstructed.

III. Experiments

For evaluating the proposed method, we evaluated two experiments. In the first experiment, conducted the reconstruction of speech by changing the threshold value and in the second experiment, the reconstruction of input speech with different the numbers of iterations. In both experiments, this sentence, "how are you" was used as the experiment sentence.

A. Experiment 1: reconstruction of speech by changing the threshold value

In this experiment, we want to evaluate the noisy input signal when the threshold values is 0.5, 2.5 and 5. For this work, we used the frame with length of 30 ms and analysis window shifts with amount of 10 ms that can be reconstructed input signal with 300 iterations. Also for noisy input speech, we used formula the Gaussian random values In which generates an input signal array of random numbers of the standard normal distribution. To generate random numbers of a normal distribution with mean input signal and standard deviation, shift and stretch:

$$\text{noisy signal} = \text{input signal} + \text{standard deviation} * \text{randn(size(input signal))} \tag{8}$$

In this experiment, we used the standard deviation of 0.15. The comparison of mean square error (MSE) of the reconstructed signal by using a threshold and without the use of threshold, in different numbers of iterations of the input signal is shown in Fig. 4, that increasing amount of the threshold, is reduced the MSE. Course must be noted that increased of the threshold value to a certain amount because, it may lost many of the features of the original signal. For showing the results better, wave reconstructed of the signals with different values of the threshold, is shown in Fig. 5.

B. Experiment 2: reconstruction the input speech with different number of iterations

In this experiment we want to investigate the noisy input signal when the number of iterations is 600 and 200 times, was reconstructed respectively, with a shifts 5 and 15 ms. For this work, we used the frame with length of 30 ms and a threshold amount of 2.5. Also for enter the noise into speech input, by using formula 8 we added noise to the input speech with a standard deviation of 0.15. The comparison of MSE of the reconstructed signal by using a threshold and without using of threshold, in the number of iterations of the input signal is shown in Fig. 6. In Fig. 4(b) with 300 times iterations have been performed for the reconstruction of input signal. According to the Fig. 6, we observe that our method can significantly reduce the MSE of the noise signal. In the obtained results we observe that with increasing the number of iteration operations for reconstruction the input signal, the mean squared error is less. in Fig. 7, it is shown that wave of reconstruction of the signals with various iterations number by using threshold and without using of threshold that is more, the number of iterations operation for the reconstruction of input signal, reconstructed signal can be extracted the original signal from the input noisy signal with a higher ability, because the overlap area between the analysis windows is more. in Fig. 5(d), wave the reconstructed signal with 300 iterations is shown. in Fig. 5(a), wave the original signal is shown.

IV. Conclusion

In this paper, we propose a method for reconstruct the speech signal that by using it we will be able to extract the original speech signal from the speech signal with much noise. So that at each stage, the reconstruction operation iterations of the speech input, by taking the threshold value

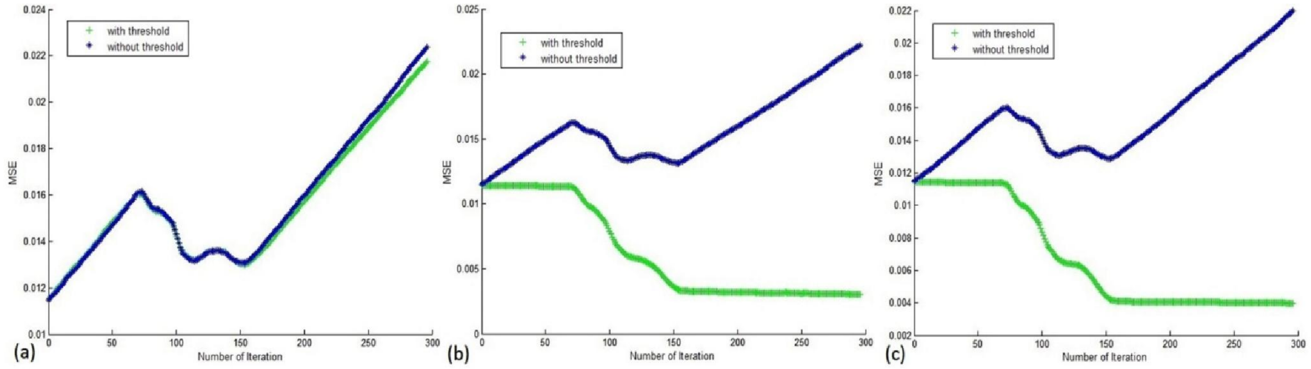

Fig 4: Reconstruction of the speech input with the threshold value a) 0.5, b) 2.5, c) 5.

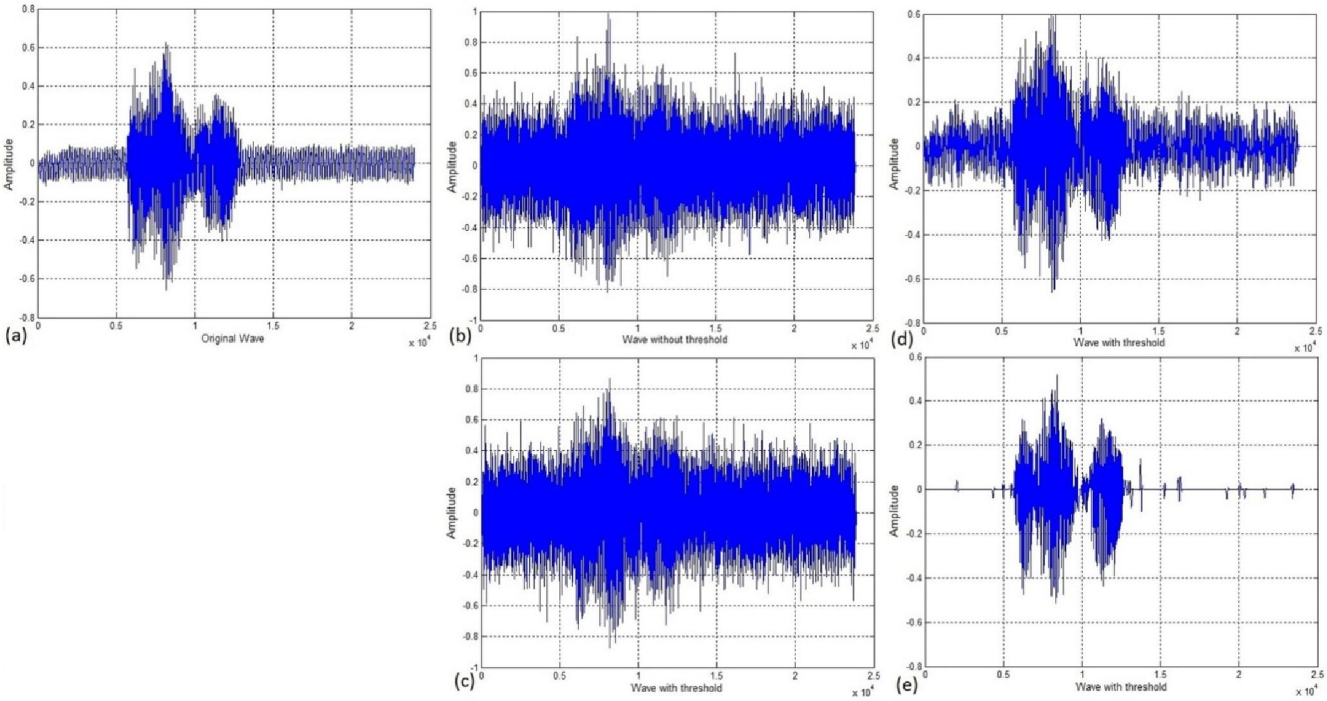

Fig 5: Comparison of wave a) the original signal, b) reconstruction of the without threshold, c) reconstruction of the with a threshold amount of 0.5, d) reconstruction of the with a threshold amount of 2.5, e) reconstruction of the with a threshold amount of 5.

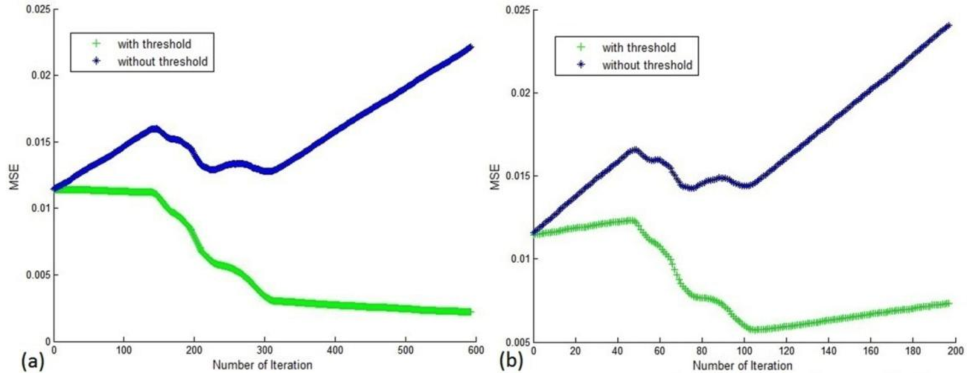

Fig 6: Reconstruction of the input speech with a) 600 iterations, b) 200 iterations.

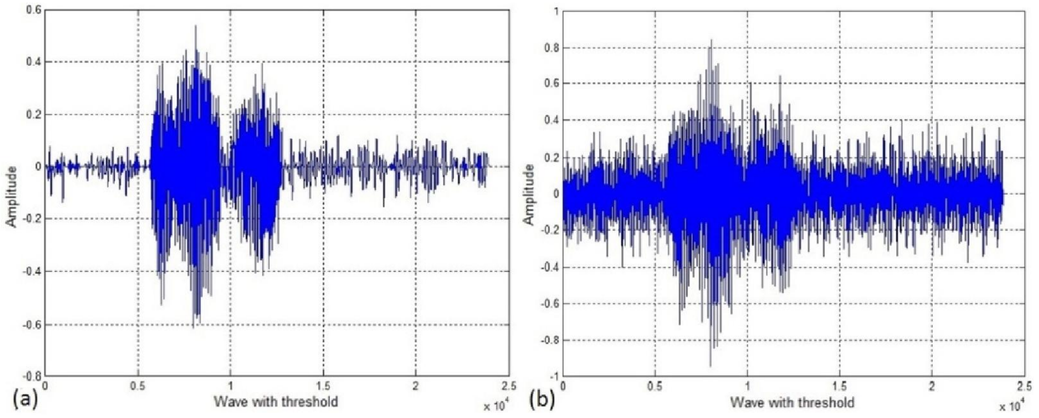

Fig 7: Compared the reconstructed wave with, a) 600 iterations and b) 200 iterations.

for the magnitude spectrum, have been removed smaller amounts of the threshold and for maintain the original signal features, we consider the phase spectrum then with combination of information the improved magnitude spectrum and phase spectrum information, we reconstruct a part of the signal in each iteration. The results showed that when the input speech signal has high values of noise ratio, it can be reconstructed the original speech signal very well. We also observed when we increase the threshold value; the original speech signal is reconstructed better. Other results showed that by using of our method, with more iteration numbers for reconstructing the input signal, the original signal is reconstructed better.

REFERENCES

[1] L. Liu, J. He, G. Palm, “Effects of phase on the perception of intervo-calic stop consonants,” Speech Communication, 1997, PP. 403-417.

[2] J. W. Picone, “Signal modeling techniques in speech recognition,” IEEE Trans, 1993, PP. 1215-1247.

[3] P. L. Van Hove, M. H. Hayes, J. S. Lim, A. V Oppenheim, “Signal reconstruction from signed Fourier transform magnitude,” IEEE Trans. Acoust. Speech Signal Processing ASSP-31 (5), 1983, pp. 1286–1293.

[4] L. D. Alsteris, K. Paliwal, “Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra,” Computer Speech and Language, 2007, PP.174–186.

[5] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, “Enhancement and Recognition of Reverberant and Noisy Speech by Extending Its Coherence,” arXiv:1509.00533 [cs, stat], Sep. 2015.

[6] K. Paliwal, B. Schwerin, K. Wojcicki, “Role of modulation magnitude and phase spectrum towards speech intelligibility,” proc. Speech Communication, 2011, PP. 327–339.

[7] D. W. Griffin, J. S. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoust. Speech Signal Processing ASSP-32 (2), 1984, PP. 236–243.

[8] S. H. Nawab, T. F. Quatieri, J. S. Lim, “Signal reconstruction from short-time Fourier transform magnitude,” IEEE Trans. Speech Signal Processing ASSP-31 (4), 1983, PP. 986–998.

[9] K. Paliwal, K. Wojcicki, B. Schwerin, “Single-channel speech enhan-cement using spectral subtraction in the short-time modulation domain,” Speech Comm. 52 (5), 2010b, PP. 450–475.

[10] P. Loizou, “Speech Enhancement: Theory and Practice,” Taylor and Francis, Boca Raton, FL, 2007.

[11] X. Huang, A. Acero, H. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development,” Prentice Hall, Upper Saddle River, New Jersey, 2001.

Saeed Karimi is master of science, computer hardware in Islamic azad university of dezful, iran and bachelor of science computer software in jahad university of ahvaz, iran. He has total 7 years teaching in university and published two isi papers.