J. Basic. Appl. Sci. Res., 3(1s)104-112, 2013
© 2013, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Iterative Reconstruction of High Ability Speech by Using of Reconstructed Combined Speech

Saeed Karimi[1], Mehdi Sadeghzadeh[2], Javad Mirabedini[3]

[1]Msc, Department of Computer, Islamic Azad University of Dezful Branch
[2],3Assistant Professor, Department of Computer, Islamic Azad University of Dezful Branch, Iran

Received: June 10 2013
Accepted: July 10 2013

ABSTRACT

Attention to the importance of speech reconstruction in various applications, this subject, has become one of the important area of research in recent years. The current system of speech reconstruction from the short time fourier transform (STFT) is used. For this work, are performed through the phase spectrum or magnitude spectrum or both of them and are used for feature extraction from mel freqency cepstrum coefficients (MFCC), if the signal is degraded with noise, it can not do recognition operation well. In this paper we present an approach to create the high ability Speech signals. to reach this goal, we create a speech signal the use from combines of the MFCC method, iterative reconstruction STFT method and analysis–modification–synthesis (AMS) method and at every stage of the repetition of reconstruction the input speech, reconstructed part the input speech multiplied by the adapt part with it in the reconstruction weight signal. We examined three tests, in the first experiment, we changed the length of the input speech frames, in the second experiment, number of iteration the signal reconstruction evaluated by shifting change of the analysis window, and in the third experiment, with an increase in noise into the input speech, we evaluated its. The results show that, our method in the different overlapping areas and frames with different size, can provide with high ability speech signal.
KEYWORDS: Iterative reconstruction, Speech recognition, Phase spectrum, Magnitude spectrum.

1. 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]. Van Hove et al. (1983) have determined that such signals can be uniquely specified by the signed-magnitude spectrum (magnitude spectrum with one bit of phase spectrum information). 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 [3]. 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 [4]. In this paper, we show that with a reconstruction of the input speech part and multiplying it in the adapt of its section at the reconstructed weight signal in each iteration, with a different frames and overlapping areas of the adjacent analysis windows, Speech reconstruction to improve. To achieve these goals, we recommend the framework by use of a combination Iterative reconstruction the STFT as proposed in [5] is used and a dual AMS framework such as proposed in [6] is used and MFCC. First, it introduces them, then modify it, until we can desired speech signal in order to rebuild.

This paper is organized as follows. in section 2, we described the weight signal reconstruction method, in section 3, some tests and results of any reviews and finally, conclusions are given in section 4.

2. The proposed iterative reconstruction of speech signal

For reconstruction this signal, first we briefly described the iterative reconstruction framework the STFT then AMS and MFCC, then we reconstruction signal with a combination of these three framework, until the created each segment of the input speech at each step of the iterative reconstruction multiplied with the adapt part with it in the reconstruction weight signal.

2.1. Reconstruction within the STFT framework

This method was employed for time-scale modification of speech. It is referred to as simultaneous extrapolation'. In this method, the known spectral information of all short-time sections are used simultaneously to determine the unknown signal (i.e., the whole signal is analyzed and synthesized in every iteration). This framework is illustrated in Fig. 1.

Corresponding author: Saeed Karimi, Department of Computer, Islamic Azad University of Dezful Branch, Azadegan Blv, Dezful, phone; +98-0641-6260601-9, Iran . Email address: karemisaeed@yahoo.com

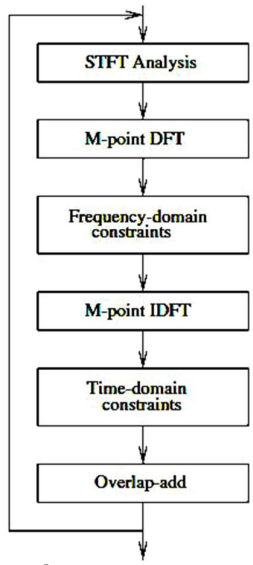

Fig. 1. STFT-based iterative reconstruction framework.

2.2. 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 [2,7,8], 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 [9]. Here, the modified Hanning window [5] 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.

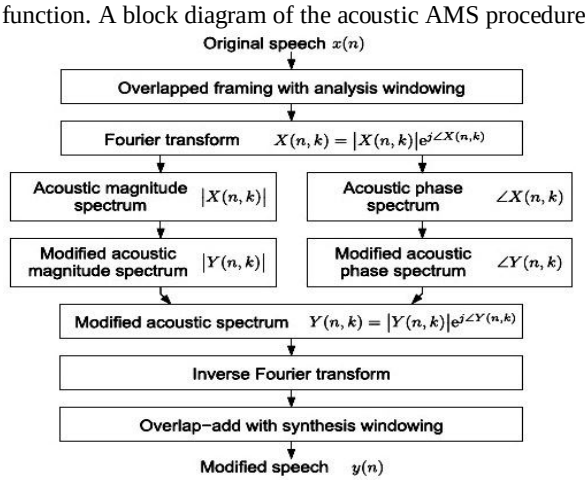

Fig. 2. Block diagram of the acoustic AMS procedure.

2.3. General MFCC procedure

In a general feature extraction procedure for the MFCC, the speech signal is converted to spectra via discrete fourier transformation (DFT), the spectra are passed through a Mel-frequency filter bank to get Mel-frequency FBEs, a logarithm is applied, and finally the MFCC is obtained from the log FBEs via a DCT. The procedure is shown in Fig. 3.

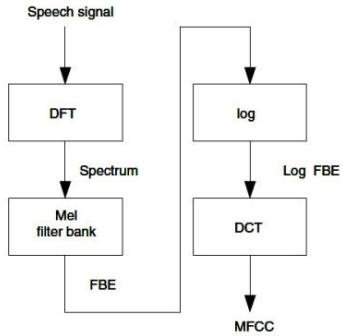

Fig. 3. Standard feature extraction procedure for MFCC.

2.4. The proposed iterative reconstruction of speech signal

By attention to the combination of both the components magnitude and phase spectra are important for the understanding signal reconstruction. In the purpose of the proposed plan, reconstruction has higher ability signal. For reach to this goal, we used from framework AMS, MFCC and Iterative reconstruction the STFT that we create the speech signal which is greatest similarity with the input noisy speech, until by multiplying it, at input signal at each iteration, it cause improvement of its reconstruction. in each of iteration our iterative reconstruction is based on the magnitude spectra components. Its diagram is shown in Fig. 4. where the input signal is applied in two directions, it imposed on both the short time analysis window and in each iteration, it is extracted spectral information from of both of the input. The first path is for weight signal reconstruction. According to the formula 3 if $|Y(n,k)|$ magnitude spectrum information and $\angle Y(n,k)$ information are obtained from the phase spectrum are input signal, we can created the speech signal from combinations of the magnitude spectrum:

$$Y(n) = |Y(n,k)| cos\, \angle Y(n,k)\,, \tag{5}$$

and other input, is weight signal reconstruction into case which is for earning the speech signal which is the greatest adaptation with the input noisy speech signal through the DCT is divided into the spectrums. In the next step, first, the logarithm of the input spectra amplitude then a filter bank which its distributions is based on mel criterion, Imposed on the spectrum and is calculated the outputs of each filters Then this outputs create $f = \{f_1, f_2, \dots f_D\}$ vector; where D is size of the vector. In the next step, The logarithm of the This vector and is used DCT transform into a cepstrum vector:

$$C = \mathrm{DCT}\left(\hat{f}_i\right), \qquad 1 \le i \le D \tag{6}$$

If we assume $(C_1, C_2, \dots C_n)$ are stored the cepstrum signals and n is the number of trained signals, the speech signal which is greatest similarity with the input noisy speech, obtained as follows:

$$weight = \max\left(\hat{C}_i\right), \qquad 1 \le i \le n \tag{7}$$

If the short-time Fourier transform is as follows:

$$weight(n,k) = |weight(n,k)| e^{j\angle weight(n,k)}\,, \tag{8}$$

where if $|weight(n,k)|$ is magnitude spectrum the adapted signal, its pretermit and if $\angle weight(n,k)$ is phase spectrum the adapted signal, consider it as phase spectrum the weight signal, and created magnitude spectrum the weight signal with attention to formula 3, In which $|Y(n,k)|$ is information of magnitude spectrum the input signal. for this reason we used of magnitude spectrum the input signal until it is preserved the signal form, finally for weight signal reconstruction used of the following formula:

$$weight(n) = |Y(n,k)| cos\, \angle weight(n,k)\,, \tag{9}$$

then, according to the formula 5 and 9, the input signal and reconstructed weight signal are multiplied:

$$Y(n) = Y(n).weight(n)\,, \tag{10}$$

with this method, the input speech iterative reconstruction would be repeated, until the whole input signal is reconstructed.

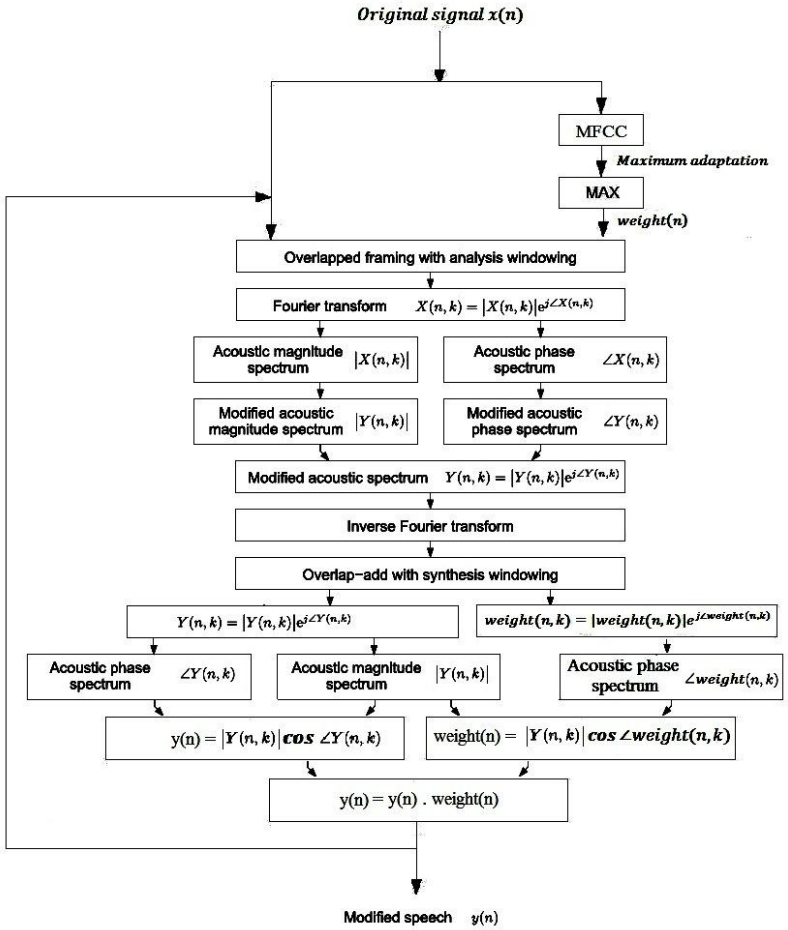

Fig. 4. Block diagram of the proposed Iterative reconstruction of speech signal.

3. Experiments

For evaluating the proposed method, there were conducted different tests. In these experiments, evaluated by changing the length of the input speech frames and number of steps iteration the reconstruction operation also with enter the amount of noise to the input speech. For experiments reviews, first, using a conventional microphone, sentences as training data, was considered the recognition system, sentences may be environment noise or microphone noise. That the experiments have been followed.

3.1. Experiment 1: change of frames length of input speech

In this experiment, we want to observe when duration of the frames is more or less, what impact it has on the input signal reconstruction and compare it with suggested method. On this basis, the input speech are divided into frames with different lengths. In this experiment of windows the henning short time used with time shifts 80 ms. after iteration 300 the reconstruction operation and reconstruction the input signal, results obtained are shown in Fig. 5. Where the input speech is divided into frames with length 160, 240, 320 and 640 ms. In all cases, when we had used the proposed method, it was reconstruction better the input signal. We also observe, in Fig. 5(b) that was 240 the frame length and the input signal is divided into 100 frames, had less mean squared error (MSE) than the other modes.

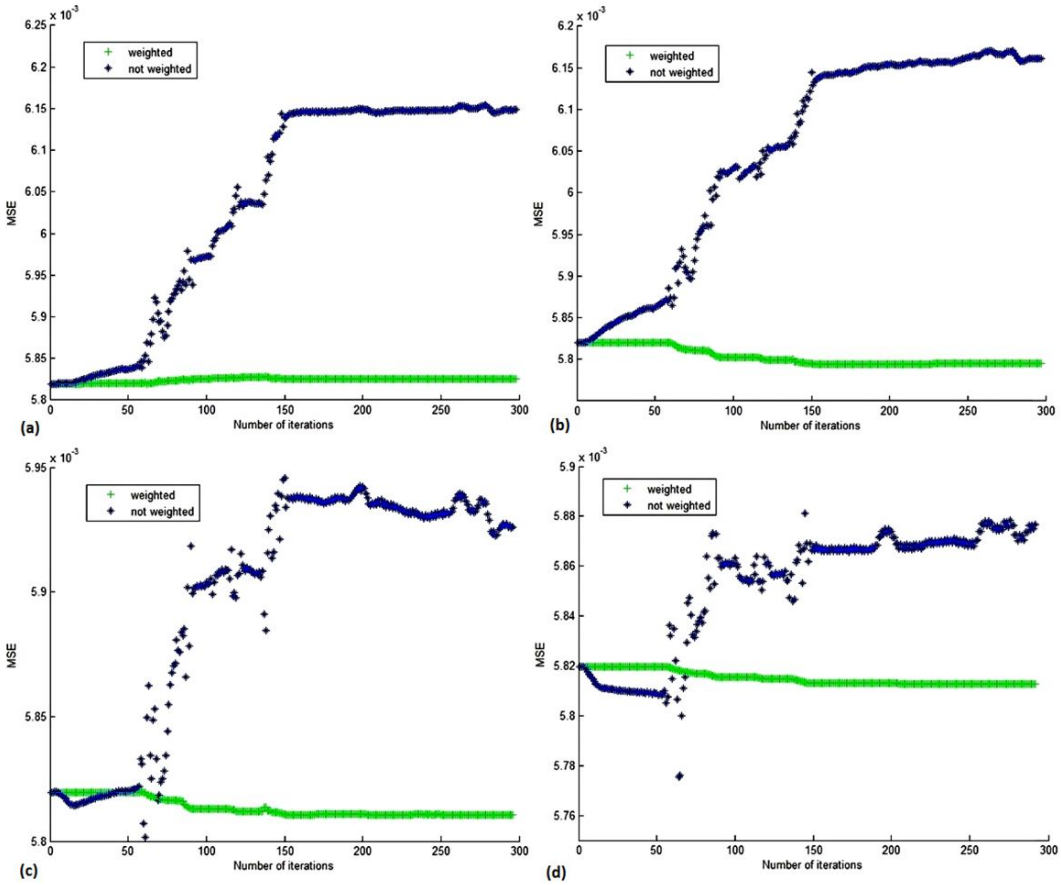

Fig.5. reconstruction of the input speech (a) frames of length 160, (b) frames of length 240, (c) frames of length 320 and (d) frames of length 640 ms .

To show the obtained results, in Fig. 6, reconstructed speech wave are shown as frames of length 640. By attention to generated waves, can be observed when reconstruction is based on magnitude spectrum compounds, when great the frame during by using our method, the reconstruction speech is better. For better understanding obtained results the generated spectra when frame length is 640, as shown in Fig. 7.

3.1.1. Results

This experiment showed that the using of the suggested method, it can be used when we use of the frames with the different duration, improve the reconstructed speech. Also from obtained results of the comparison frames with length different, we understand that, when the reconstruction is by using magnitude spectrum of compounds we can also reconstructed the speech by using of large frames, which it was not like this case in the past.

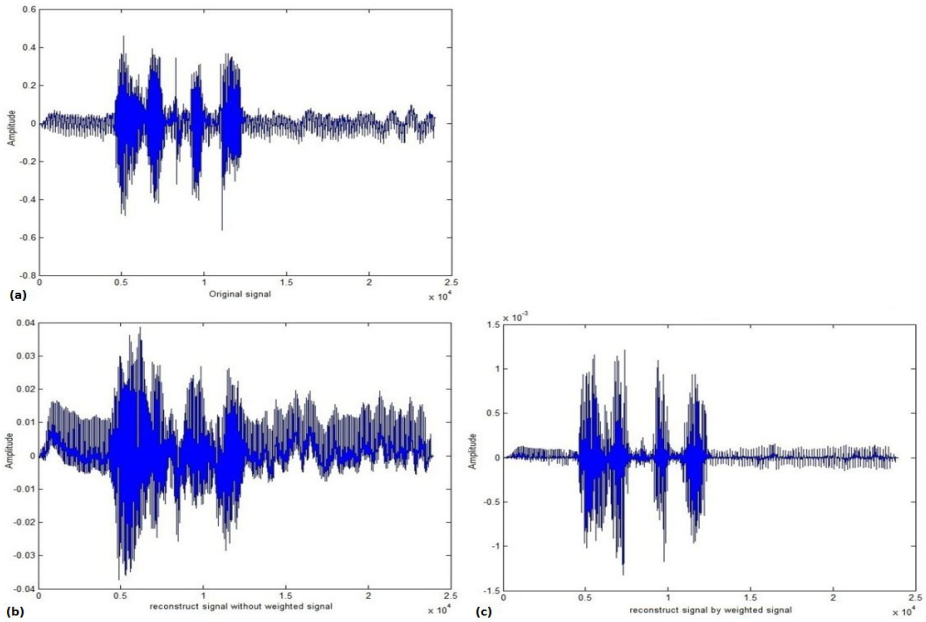

Fig. 6. comparison of the input signal reconstruction with frame length of 640 ms (a) wave the original signal, (b) reconstruction without weight, (c) reconstruction by weight.

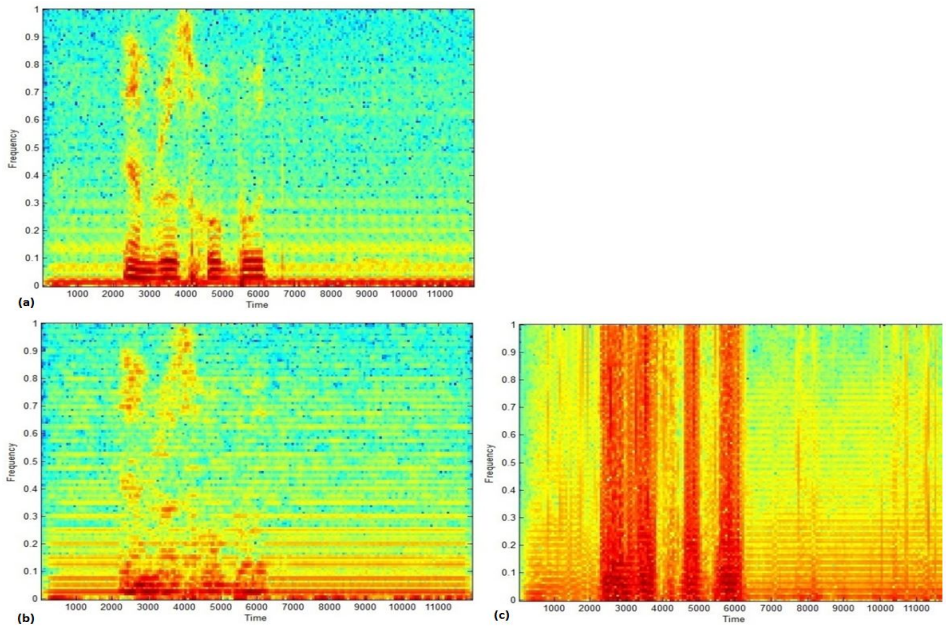

Fig. 7. spectrum comparing the input signal reconstruction with frame length of 640 ms (a) the original signal wave, (b) without weight reconstruction, (c) weight reconstruction.

3.2. Experiment 2: Increase or decrease the number of iteration of the reconstruction Operations

In this experiment, we want to observe when the number of the iterations of the reconstruction operations is more or less, what impact it has on reconstruction the input signal and we compare it with suggested method. On this basis, we used of fixed frame with length of 240 ms which the input speech was divided into 100 frames and also we use of the henning windows and increasing or decreasing the number of the iteration of the reconstruction operations, we used of shifts different time the analysis windows which are shown in Fig. 8. The input speech has been reconstructed with shifts short time 30, 60, 80 and 120 ms . When we used of shifts 30 ms the time window, the number of the iteration reconstruction stages the input signals was to 800. When we used of shifts 60 ms the time window, the number of the iteration reconstruction stages the input signals was to 400. When we used of shifts 80 ms the time window, the number of the iteration reconstruction stages the input signals was to 300 and when we used of shifts 120 ms the time window, the number of the iteration reconstruction stages the input signals was to 200.

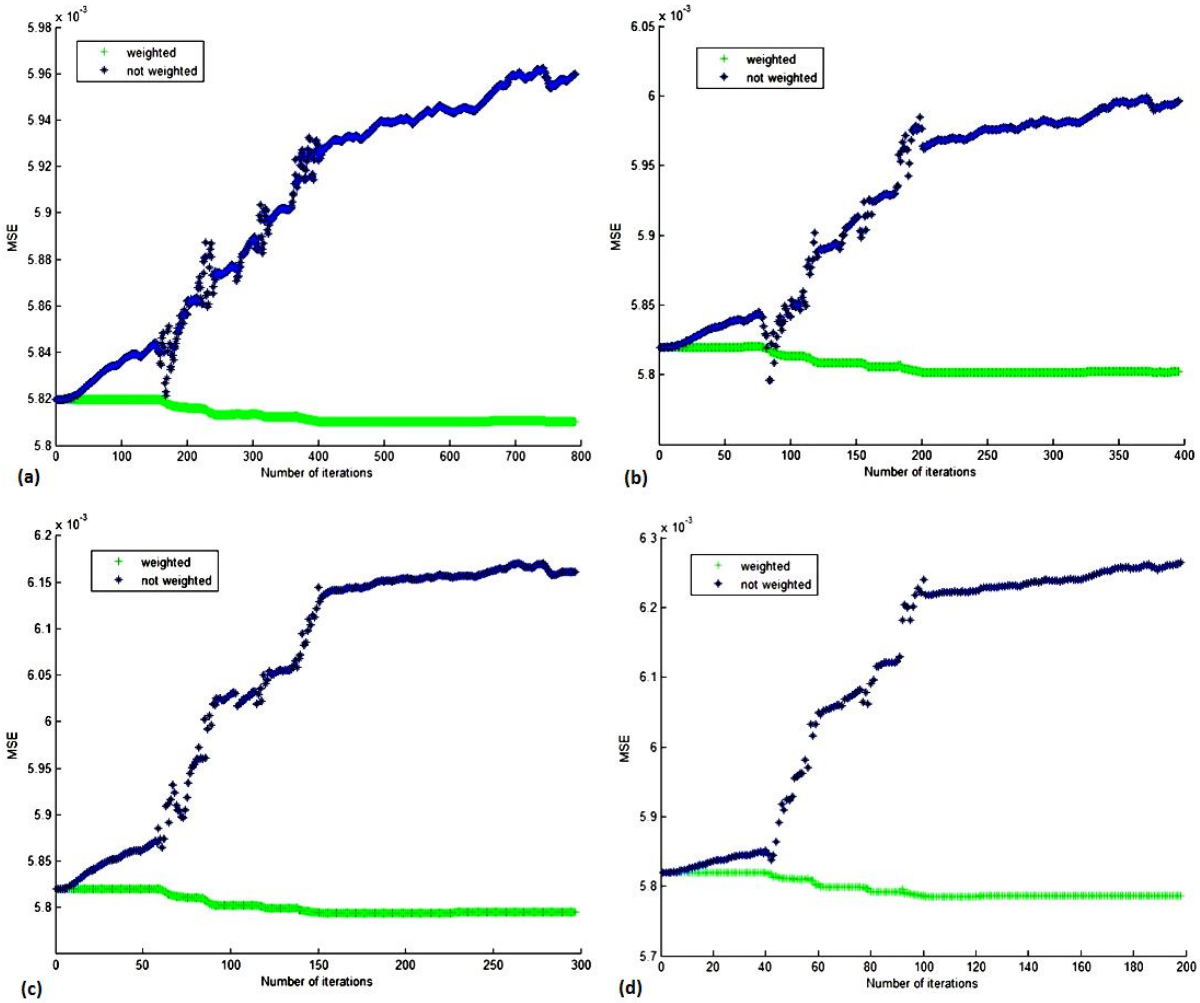

Fig. 8. reconstruction of the input speech with the iteration count(a) 800 With shift of 30 ms, (b) 400 With shift of 60 ms, (c) 300 With shift of 80 ms and (d) 200 With shift of 120 ms.

We can observe, when the shifts amounts are great , although areas of overlapping the short time windows of adjacent are low, our method can better the signal reconstruction whereas in the previous method if areas of overlapping the adjacent short time windows are high, better could reconstruct the speech signal. In our method also improvement in speech with areas of low overlapping the windows of adjacent and with areas of high overlapping.

3.2.1. Results

This experiment shows when we used of the proposed method, in more or less the shifts window of time short, it improve the speech. Also we observed when we used of windows with low overlapping areas, the signal could better reconstruction whereas in previous methods, the signal with high overlapping areas, better was reconstructed.

3.3. Experiment 3: The iterative reconstruction with an increase in noise into the input speech

In this experiment, we want to compare the reconstruction of input noisy signals with different amounts of noise in weighted and non-weighted. for this task, amounts the noises difference in the range of 0 to 0.15 and the noise increase 0.002 at each stage into the input speech. Also we compare the reconstruction with the input signal that does not use of the reconstruction operation and are affected by these values of noise. It is shown in Fig. 9. Increases MSE of input signal with much noise and until amount of noise 0.08 has lower MSE than iterative reconstruction framework, using a combination of magnitude spectrum, MSE of iterative reconstruction method with increasing noise using a combination of magnitude spectrum is less than the input signal. by using of our method, we see that the MSE decreases In both time increasing and decrement of MSE, iterative reconstruction framework by using of combinations the magnitude spectrum. at each stage of the increase in noise, input noisy signal compared with trained speech signals and the phase spectrum speech that is best adaptation with the input noise speech and magnitude spectrum the input noisy signal, uses for reconstruction the suggested signal.

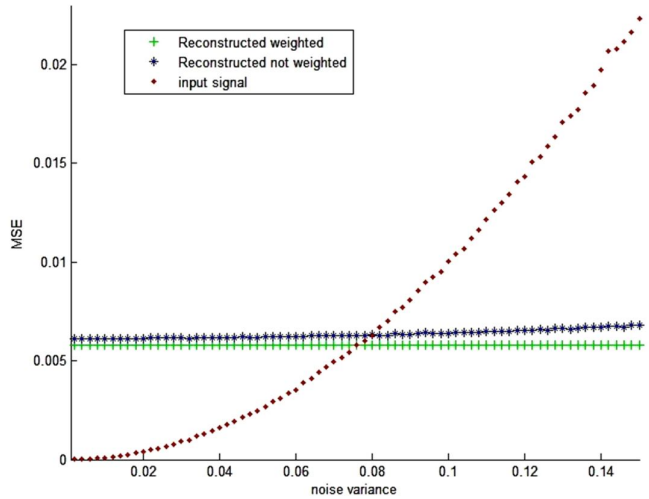

Fig. 9. Comparison of the input signal, the signal reconstructed the weighted and not-weighted by increasing of the noise amounts.

For understanding the results better, the wave of the reconstructed signal we have compared when amount of noise 0.15 is added to the input signal that are shown in Fig. 10. wave of the original signal is shown in Fig. 6(a).

3.3.1. Results

According to the amounts and the results achieved in this experiment, we see that when the input speech is affected by much noise it can be improve by using iterative reconstruction and by using our method, this work will be higher capability.

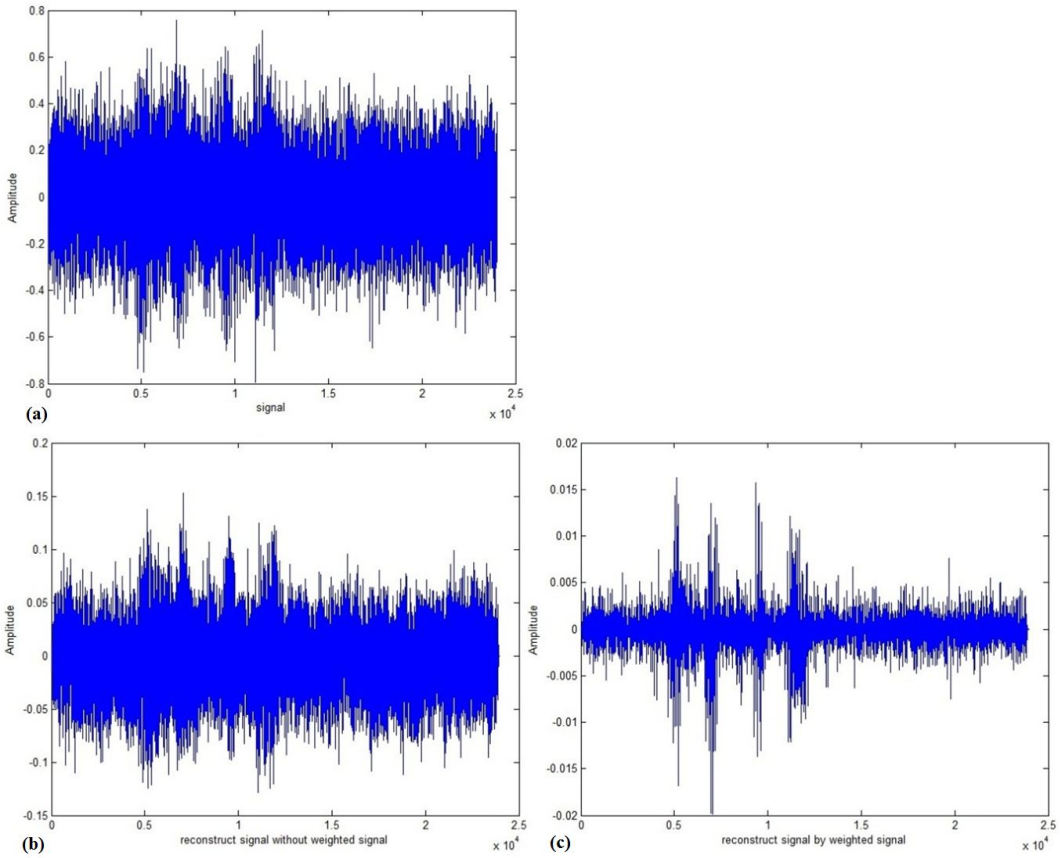

Fig. 10 - Comparison of the wave by adding noise 0.15 (a) input signal, (b) not-weight reconstructed signal, and (c) weighted reconstructed signal.

4. Conclusion

In this paper, we propose a method for reconstruction of speech signal that by using, can be reconstructed the noisy speech with higher capability. thereby which the improvement of reconstructed signal at each iteration by using multiplying it in the proposed reconstructed signal, it can be better reconstructed the speech signal. The results showed that when is reconstruction by using a combinations of magnitude spectrum and the frames are large can be with the higher ability than the previous method is reconstructed the speech signal, also we observed when the overlap areas between adjacent analysis windows, are small, than previous methods can be reconstructed speech signal with higher quality and showed other experiments that by using our method when increases noise amount of input signal, can be much decreased MSE the input signal.

REFERENCES

1. Liu, L., He, J. and G.Palm, Effects of phase on the perception of intervocalic stop consonants. Speech Communication, 1997. p. 403-417.
2. Picone, J.W, Signal modeling techniques in speech recognition. IEEE, 1993. P. 1215-1247.
3. Alsteris, Leigh D. and Kuldip.Paliwal, Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra. Computer Speech and Language, 2007. P. 174–186.
4. Paliwal, Kuldip., Schwerin, Belinda. and Kamil.Wojcicki, Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Communication, 2011. P. 327–339.
5. Griffin, D.W. and J.S.Lim, Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Processing ASSP-32 (2), 1984. P. 236–243.
6. Paliwal, K., Wojcicki, K. and B.Schwerin, Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm, 2010b. p. 450–475.
7. Loizou, P., Speech Enhancement: Theory and Practice. Taylor and Francis, Boca Raton, FL, 2007.
8. Huang, X., Acero, A. and H.Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, Upper Saddle River, New Jersey, 2001.
9. Quatieri, T., Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall, Upper Saddle River, NJ, 2002.