J. Basic. Appl. Sci. Res., 3(1s)50-56, 2013
© 2013, TextRoad Publication
ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

The Improvement of Speech by Using Reconstructed Speech

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

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

Received: June 10 2013
Accepted: July 9 2013

ABSTRACT

In recent years, speech recognition has become one of the most important areas of research. The current system has been used to extract features from Mel-frequency cestrum coefficient (MFCC) if the signal is degraded by noise, it cannot create a system with high ability of recognition. In this paper we present an approach in which the speech recognition systems will be capable to do the recognition operations with higher capability. To achieve this goal we create the speech recognition by combining MFCC and analysis–modification–synthesis (AMS) methods and multiply it in input noise signal. Three experiment were investigated, in the first experiment, sub-bands weighted and non-weighted are studied, In a second experiment, the input signal was compared with the sum sub-bands weighted in the third trial, we compared the all of the band input signal with the whole band input signal, and the reconstructed weight signal was multiplied. The results showed that we can improve the reconstructed weighted signal by multiplying it in the input signal and also the ratio of the sum of the weighted sub-bands to the duration of the multiplication of all of the input signal bands by the reconstructed weighted signal had a lower mean squared error (MSE) value.

KEYWORDS: Phase spectrum, Magnitude spectrum, Sub-band frequency, Speech reconstruction, Speech recognition.

1. INTRODUCTION

It is well known that current automatic speech recognition (ASR) systems don't work as well as humans and they need to be exactly process. Noise-robust speech recognition has become an important area of research in recent years. In current speech recognition systems, the MFCCs are used as recognition features. When the speech signal is corrupted by narrow-band noise, the entire MFCC feature vector gets corrupted and it is not possible to exploit the frequency-selective property of the noise signal to make the recognition system robust. many methods have been proposed for improving speech, including the proposed the whole band noise signal divided into sub-bands and earn independent weighted coefficients of each sub-bands and the sub-bands has multiplied and by discrete cosine transform (DCT) that they made up the cepstral vectors [1]. the weighting procedure could not only degrades noise but it could enhances clean speech signal. In automatic speech recognition (ASR), the speech is processed frame-wise using a temporal window duration of 20–40 ms . The short-time Fourier transform (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 qqqspeech intelligibility [2] 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 [3]. 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. With respect to these variables the phase spectrum, other researchers were able to derive the frequency and time of phase spectrum, the signal to reconstruct [4]. Other results were obtained Using a combination of information is 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 [5]. In this paper, we show that the combined magnitude spectrum input signal and signal phase spectrum that is most likely to be adapt with the input signal, we can reconstructed the speech signal that causes speech to improve. To achieve these goals, a dual AMS framework such as proposed in [6] is used, we recommend. First, it introduces ,we then modify it until it can be desired speech signal in order to rebuild.

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

2. Reconstruction of the weighted speech signal

For reconstruction the proposed signal, first AMS and MFCC framework are described, then the combination of these two frameworks, signal can be reconstruction to multiplying the input speech signal and causes it to improve.

2.1. 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

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

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

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.2 In speech processing, an acoustic frame duration of 20–40 ms is typically used [3,7,8], 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 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. 1.

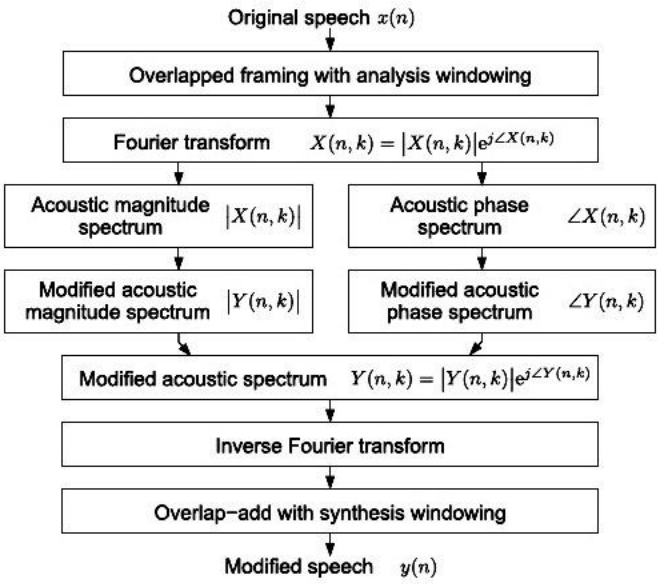

Fig. 1. Block diagram of the acoustic AMS procedure.

2.2. 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. 2.

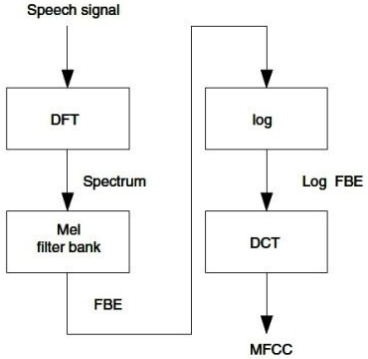

Fig. 2. Standard feature extraction procedure for MFCC.

2.3. Reconstruction Weight signal

According to the combination of both components phase spectrum and magnitude spectrum are important to reconstruct the signal are understandable. In the proposed plan, the purpose of a speech signal is that it has the most

adaption with the noise signal until multiplying the input signal to improve it. for this task, we change and develop the AMS framework, as firstly because of preserving the shape of the signal, we consider magnitude spectrum of the signal. according to formula (2), if $|X(n,k)|$ is characteristics the input signal magnitude spectrum, It is intended and if $\angle X(n,k)$ is characteristics the input signal phase spectrum, we leave it to the side and then to reconstruct the weight signal, we obtain the phase spectrum of the MFCC. as the phase spectrum of speech signals trained that it has the most likely to be adapt with the input signal that we get. In this method first, for stored speechs features extraction, speech signals is divided by DCT into spectrums. In the next step first, we take the logarithm of the amplitude spectrum of the input then a filter bank that is distribution and is based on Mel standard Imposed on the spectrum and the outputs of each filters is calculated then this outputs make up the vector $f = \{f_1, f_2, \dots f_D\}$ Where D is vector magnitude. In the next step, of this vector, we take logarithm and using the DCT is converted to a cepstral vector:

$$C = \text{DCT}\left(\hat{f}_i\right), \qquad 1 \le i \le D \tag{5}$$

If we assume $(C_1, C_2, \dots C_n)$ are stored cepstrals signals and n is the number of given training signals, which has the maximum adaptability with the input noise signal is obtained as follows:

$$weight = \max\left(\hat{C}_i\right), \qquad 1 \le i \le n \tag{6}$$

if the short-time Fourier transform is as follows:

$$weight(n,k) = |weight(n,k)|e^{j\angle weight(n,k)}, \tag{7}$$

where $|weight(n,k)|$ is magnitude spectrum the signal it aside and if $\angle weight(n,k)$ is signal phase spectrum, we consider it as phase spectrum the reconstructed signal and the magnitude spectrum $|X(n,k)|$ are extracted from the input Speech signal. Finally, for weight signal reconstruct, we use the following formula:

$$weight = |X(n,k)|cos\ \angle weight(n,k), \tag{8}$$

in Fig. 3, diagram the weight signal reconstruction, is shown.

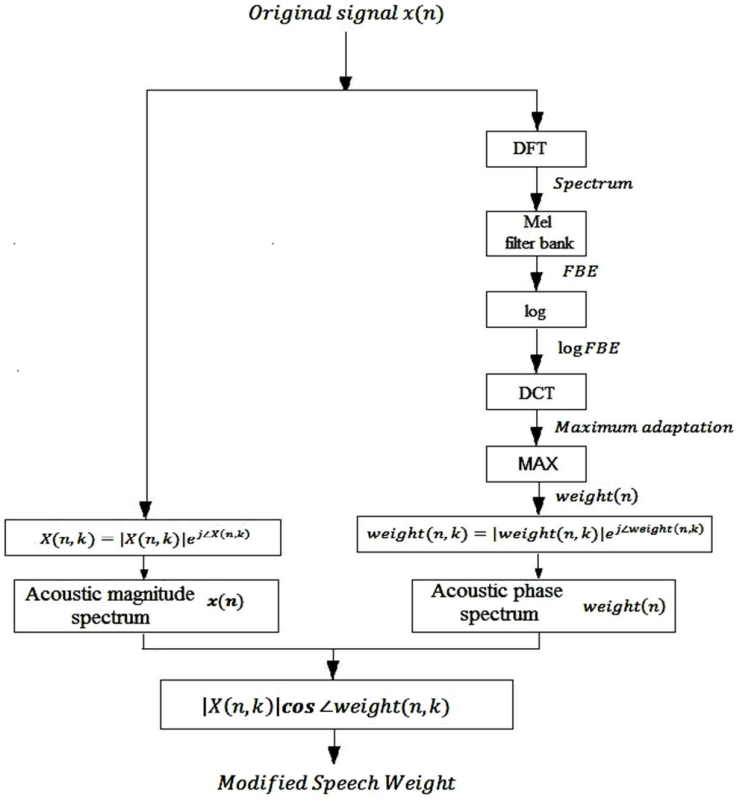

Fig. 3. Block diagram of weight signal Reconstruction.

3. Experiments

To check the experiments, first by using of a ordinary microphone, we considered some sentences as training data in the recognition system. This sentences are shown in Fig. 4, sentences may have environment noises or microphone noises. In the experiments, we examine the different aspects of the input signal with the noise starts from 0.001 and then we increase the amount of noise 0.002 at each stage Independently than previous stage.

#	speech	wave
1	امتحانت چطور بود	
2	آخر هفته هستی	
3	سلام علیکم	
4	بنام خدا	
5	امروز خوب بود	
6	کی دانشگاه میری	
7	حال شما خوبه	
8	ورزش می کنی امروز	
9	ماشین خریدی یا نه	
10	اهل کجا هستید	

Fig. 4. trained Sentences.

3.1. Experiment 1: Comparison sub-bands of weight and not-weight the input noisy signal.

In this experiment, the whole band input signal into sub-bands based on low frequencies 1-3, 3-5, 5-7, 7-9, 9,11 and 11-14 hz, are divided. Selecting of this lower frequencies was based on previous work [10-13]. then the reconstructed speech signal multiplied each sub-bands. in all experiments, sentence number 5 was considered as the experimental input speech and by entering the noise in this speech, we evaluated it. we tested the amounts of the different noises in the range of 0.001 to 0.5 and by increase the noise 0.002 at each stage. we saw that at each stage from increased noises, of trained signals phase spectrums, phase spectrum were intended that the speech signal had the more adaptation with the input signal and used from magnitude spectrum the input noise signal for weight signal reconstruction. In this experiments we observed that multiplying the weight of each frequency sub-bans, of first noise input into input signal, we faced with reducing the MSE, only when frequency sub-band was 1 to 3 hz, first MSE was increased by multiplying the weight signal, but with increasing the noise ,MSE decreases, which is shown in Fig. 5. So we considered the signal waveform before and after the increased and decreased MSE with increasing noise 0.01 and 0.1, that is shown in Fig. 6.

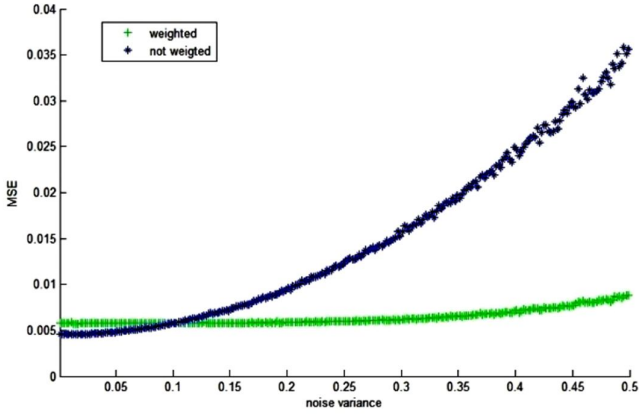

Fig. 5. weighted and non-weighted Sub-band frequency sub-band 1-3 hz.

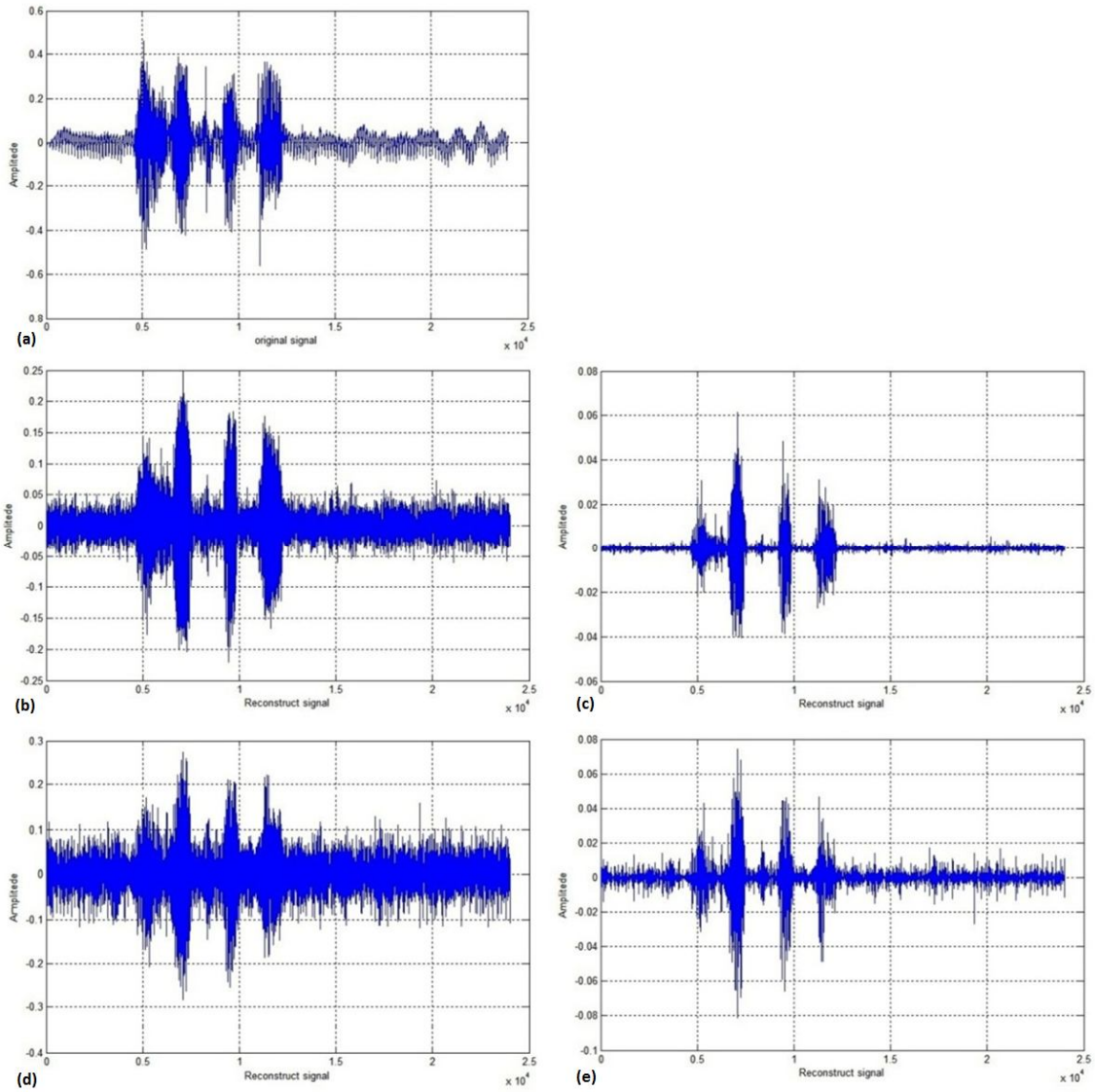

Fig. 6. Weighted signal wave (c, e) and not-weighted (b, d) frequency sub-band 1-3 Hz, (a) waveform of the original signal, (b and c), 0.05 adding noise to the input signal (d and e.) 0.1 adding noise to the input signal.

3.1.1.Results

According to the obtained shapes and diagrams also MSE values by entering amount of different noise to the input signal, we observe that with noise increasing, MSE increases for each sub-band but when the reconstructed weight signal are multiplied to each sub-bands and the amount of noise to the input signal increases, MSE is reduced. Also, we observe that by multiplying weight signal in the each sub-band, signal wave amplitude decreases. at frequency sub-band of 1 to 3 hz up a certain amount of noise in the input signal, by multiplying the weight signal MSE is increased, but with increased noise, MSE is reduced. Also, the frequency sub-band is 11-14 hz, MSE more increases than other frequency sub-bands.

3-2- Experiment 2: comparison input noisy signal with sum weighted sub-bands

In this experiment, all weighted sub-bands has added together at experiment 1, we compare with total full-band input signal, amounts the noises difference in the range of 0.001 to 0.5 and the noise increase 0.002 at each stage, was tested. at each stage of the assessment, compared the weighted sub-bands sum and full-band input signal with the original signal and MSE is calculated that shown in Fig. 7.

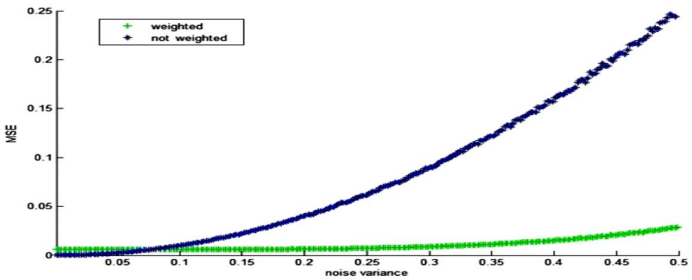

Fig. 7. weighted Sub-bands sum and input signal

The figure above shows at first weighted sub-bands sum than the input signal, are faced with a slightly higher MSE but gradually the noise increases, MSE the input signal than weighted sub-bands sum, increases. Note was that at this experiment we had dealt with when the whole band input signal was divided by the sub-bans and then were collected sub-bands without multiplied at reconstructed weighted signals, faced with MSE. In Fig. 8, MSE difference whole of the band input signal and whole band gathered obtained from sub-bands not-weighted is shown.

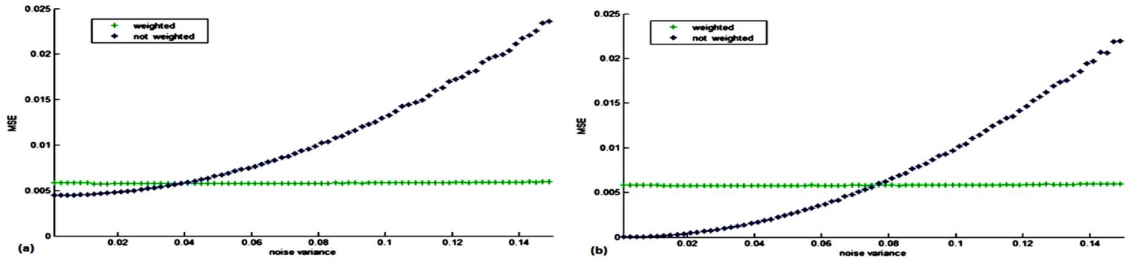

Fig. 8. compare the weighted sub-bands sum with (a) the sum sub-bands without the weight and (b) total band input signal.

In continuation this experiment we want created signal wave at both cases whole band input signal and the weighted sub-bands sum for noise added to the input signal with amount 0.01 and 0.1 the gain is shown in Fig. 9. Assessments showed when amount the input noise was 0.01 has received its phase spectrum of phase spectrum trained speech number 2 in Fig. 4 and when amount the input noise was 0.1, has received its phase spectrum of phase spectrum trained speech number 1 in Fig. 4. wave original signal at Fig. 6(a) is shown.

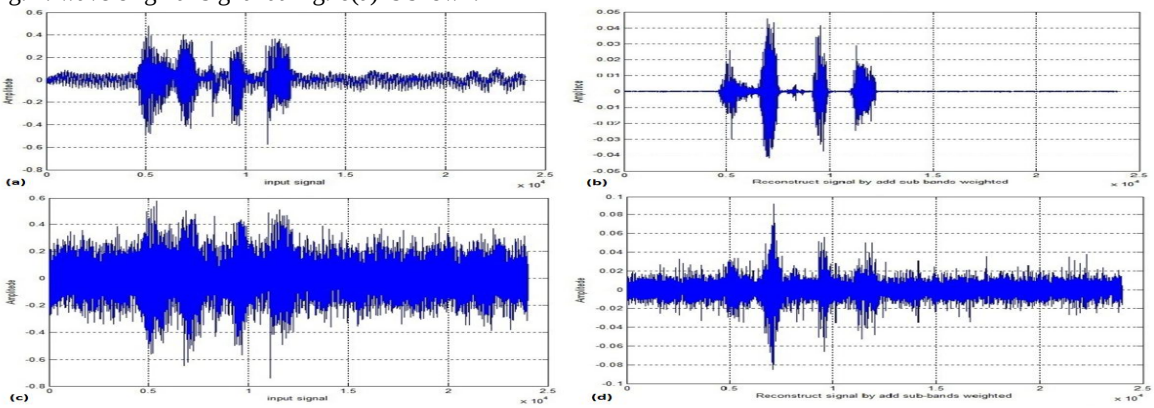

Fig. 9. Compare wave weighted sub-bands sum (b and d), and the whole band input signal (a and c), by 0.01 adding noise to the original signal (a and b) and by 0.1 adding noise to the original signal (c and d).

3.2.1. Results

at these experiments, we observed that the increased noise in the input signal, weighted sub-bands sum than the whole input signal, the MSE is less but at first face with a slightly higher MSE but many extra signals disappear and the signal obtained from the input signal is more clean. We also observed when the whole signal are obtained of not-weighted sub-bands sum than the when whole signal, without division into sub-bands is intended, MSE is a more. In fact, decomposed signal into sub-bands and then reconstruct the signal from sub-bands cause some of speech features is lost or changed which MSE is increasing.

3.3. Experiment 3: Comparison whole input noisy signal by multiplying weight at whole the input signal.

At this experiment, we have compared the input signal when the input signal was multiplied with weight signal. We also tested the range of the entered noises that was between 0.001 to 0.5 and the noise increase 0.002 at each stage. At each stage of the evaluation, MSE was calculated for both cases and it is shown in Fig. 10.

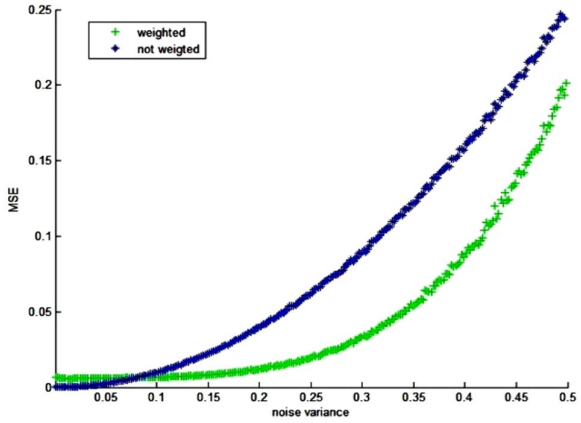

Fig. 10. Compare whole the input signal and whole the weighted input signal.

Above figure shows the increase in noise to 0.08 MSE input signal has a less MSE than the weighted input signal. However, we observe that with increasing noise, MSE input signal is increased and by weight signal, the input signal is reduced MSE.

3.3-1. Results

The shapes and results obtained of this experiment we observed signal when the weight signal is multiplied in the input signal, MSE is reduced, but again at first MSE is increase. we also observed at higher noise values, when the weight signal is multiplied in the input signal than in previous experiments, MSE is more. thus is better the input signal is divided into sub-bands and by multiplying weight signal at each sub-band and then their collecting causes to reduce the MSE at high noise.

4-CONCLUSION

In this paper, suggested a method for weight signal reconstruction, that it could be used to improve the noisy speech signal. The results showed that by multiplying the reconstructed weight signal with input signal, we can reduce the MSE. also other results showed that by dividing the whole band into sub-bands and then multiplying the reconstructed weight signal with each sub-bands and collecting weighted sub-bands, we can reduce the input signal MSE compared with when whole input signal was multiplied with the reconstructed weight signal. We used of Phase spectrum the trained signal for reconstruction of the weight signal, that has the most adaptation with the input signal and the magnitude spectrum input signal. Our method compared to previous methods show that with training limited sentences, we can improve the input signal.

REFERENCES

1. Zhu, D., Nakamura, S., Paliwal, K. and R.Wang, Maximum likelihood sub-band adaptation for robust speech recognition. Speech Communication, 2005. P. 243–264.
2. Liu, L., He, J. and G.Palm, Effects of phase on the perception of intervocalic stop consonants. Speech Communication, 1997. p. 403-417.
3. Picone, J.W, Signal modeling techniques in speech recognition. IEEE, 1993. P. 1215-1247.
4. 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.
5. Paliwal, Kuldip., Schwerin, Belinda. and Kamil.Wojcicki, Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Communication, 2011. P. 327–339.
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.
10. Shien, W. and et al, A Precluding But Not Ensuring Role of Entrained Low-Frequency Oscillations for Auditory Perception. The Journal of Neuroscience, 2012. P. 12268 –12276.
11. Kerlin, Jess R., Shahin ,Antoine J. and Lee.Miller, Attentional Gain Control of Ongoing Cortical Speech Representations in a Cocktail Party. The Journal of Neuroscience, 2010. P. 620–628.
12. Avendano, Carlos., van Vuuren, Sarel. and Hynek.Hermansky, Data Based Filter Design for RASTA-like Channel Normalization in ASR. ICSLP'96. Philadelphia, 1996. P. 2087-2090.
13. Jesen,ole. and et al, oscillations in the alpha band (9-12hz) increase with memory load during retention in a short-term memory task. cerebral cortex, 2002. P. 877-882.