Comparison of Speech Enhancement Algorithms in order to improve speech quality

Mahdi Alinaghizadeh Ardestani[1], *Zhila Mohammadi*2*

1. Faculty Member, electical Engineering, Technical and Vocational University (TVU), Tehran, Iran
2.P.H.D, Department of Basic Sciences, , Technical and Vocational University (TVU), Tehran, Iran

Abstract

Speech is one of the simplest sources of communication, yet it is a challenging phenomenon. Spoken communication includes two distinct aspects, which are the verbal component and the prosodic component. The former component combines two systems in which the first forms words from phonemes and the second constructs sentences from words. In other words, it helps to express emotions, to stress a particular word or the end of a sentence [1]. Both components share the same speech signal; however, they play peculiar roles in formulating natural language. Since we live in a natural environment where noise is inevitable and ubiquitous, speech signals can seldom be recorded in pure form and are generally contaminated by acoustic background noise. As a result, the speech signals have to be cleaned up with digital signal processing tools before they are stored, transmitted, or played out [1]. Noise reduction algorithms and systems for speech enhancement have received considerable interest in the past, primarily because the reduced speech intelligibility under noisy conditions is one of the major complaints in hearing impaired subjects. Recent years, noise reduction has been in great demand for an increasing number of audio applications, such as automatic speech recognition systems and cellular telephone.

Keywords: speech improvement, speech quality, noise removal

1. Introduction

Speech improvement is one of the most important topics in the field of telecommunications and signal processing. In telecommunications, it is possible that the speech signal is contaminated with environmental noise, and as a result, it affects the quality of communication due to the reduction of speech intelligibility. This article discusses the structure of speech and its conversion into a digital signal. The letters of a speech are divided into two groups of vowels and silent letters. It will also be discussed about the types of noises such as: white noise, pink noise and environmental noises that affect the speech signal. Speech processing includes different parts such as: coding, speaker recognition, synthesis (separation), compression and speech improvement. One of the most important parts of speech processing is improving speech and trying to improve the performance quality of speech communication systems related to it, under conditions that are affected by destructive factors. Based on the number of available microphones, speech enhancement is divided into two categories: single-channel methods (single microphone) and multi-channel methods (two-channel, microphone array, and distributed microphones). Among the important methods of removing noise from the speech signal and improving the quality of speech, we can use methods based on spectral subtraction, speech model, endowment methods, methods based on the signal subspace, Wiener filtering, linear estimation (linear prediction model), blind algorithms, He mentioned wavelet transform, neural networks, fuzzy networks, Kalman filtering .In the studies carried out on noise removal methods from single-channel speech signals, some of these methods are computationally long and some are complicated in terms of hardware. In the conducted investigations, it was found that the method of removing speech noise with spectral difference, especially in single-channel speech, has fewer calculations, is simpler and economically less expensive.

2 . noise and quality speech

It has been known for many years that a speaker will increase his/her vocal effort in the presence of a loud background noise. Informal observations confirm that people talk much louder in a noisy environment such as a subway, airplane, or cocktail party than in a quiet environment such as a library or doctor’s office. This effect, known as the Lombard reflex, was first described by Etienne Lombard in 1911 and has attracted a moderate degree of attention by researchers over the years. The observation that speakers increase their vocal effort in the presence of noise in the environment suggests that speakers monitor their vocal output rather carefully when speaking. Apparently, speakers attempt to maintain a constant level of intelligibility in the face of degradation of the message by the environmental noise source and the corresponding decrease in auditory side tone

at their ears. The most important means of communication between people is speech and voice, and it is necessary to understand and understand the subjects. A speech signal is divided into two parts, vowels and voiceless (consonant). In this article, we talk about noise and types of noise and its effect on the speech signal. Noise in practice is mostly colored and does not affect the speech signal uniformly over the entire spectrum. A lot of information is known about the effect of noise on the speech spectrum. Such knowledge can potentially help you develop better conversational enhancement algorithms. There are several important reasons for understanding how noise affects the speech spectrum. First, it helps us in better design of noise reduction algorithms and auditory understanding of speech despite noise. Second, it can help us to create better speech improvement algorithms and by choosing a suitable, relatively simple and cheap method, we can easily remove noise from the speech signal to improve speech quality. [1]A good and quality speech should have characteristics such as having the right speed, being clear, having low frequencies (bum) and high frequencies (sub) in order to convey its meaning. Also, speech is produced through sound of the human mouth and to some extent through the nose. In this process, teeth, small and large tongue, larynx and nose play an important role. Simply, we call any variable quantity in time or place that can be measured as a signal. Therefore, a speech signal can be analog (continuous) and digital (discrete). An analog speech signal is a signal whose value (amplitude) changes over time. A digital speech signal is a signal that is limited both in terms of occurrence time and value in a certain interval. [2]

Speech is caused by changes in acoustic pressure and can be measured directly in front of the mouth as a function of time. The speech signal is non-stationary and changes with the movement of the muscles of the oral cavity in the form of contraction and expansion. Speech can be divided into sound segments that share some sound characteristics with each other. Sounds are mainly divided into 2 categories:

A) Vowels, which create an unlimited flow of air in the oral cavity.
B) b) Voiceless ones, which limit the air flow in some places and are weaker than voiceless ones.

The mechanism of speech production in humans consists of the following parts : lungs, trachea, pharynx, tracheal cavity (throat), oral cavity (mouth), nasal cavity, soft palate, tongue, jaw, teeth. Lungs and trachea form the respiratory subsystem of the mechanism. These provide a source of energy for conversation and this is when air enters the trachea from the lungs. Speech production can be viewed as a filtering process where the sound source stimulates the filter of the oral cavity. The source is either rounds that cause voiced speech or non-rounds that cause voiceless speech. [3]. In general, any unwanted fluctuations and changes (unintentional) that appear on speech signals are called noise. In daily life, noise is an unwanted and loud sound that does not have any musical order. Noise is random and its distribution is usually considered Gaussian distribution (of course, this distribution is usually considered, but different distributions may be considered in different situations). The randomness of noise causes its average to be zero. So, its power of two

values are used to describe it. It is necessary to define the variability with time for noise. We call noise static (invariant with time) whose statistical characteristics do not change with time. For example, the variance or its effective value should not change with time [7]. Noise is characterized by time and frequency changes. The noises that have the most impact on speech include: white noise, facial noise, environmental noise which is explained as follows:

Spectral white noise, by definition, is noise whose density spectrum does not depend on frequency (is a constant value). Of course, this is an ideal definition, because if we take an integral from a constant number with respect to the frequency, the variance of the noise (or noise energy) is infinite. In most investigated systems, the noise is actually not white, but pink. In the sense that it has a cutoff frequency. This cutoff frequency limits the noise variance. All systems have white noise, but some application systems are contaminated by low-frequency noise in addition to that noise.

In electronics, the phenomenon known as 1/f noise, or pink noise due to its spectral characteristics, is a ubiquitous presence in most components. Also known as 1/f flicker noise, this distinctive type of electronic noise is characterized by a power spectral density inversely proportional to the frequency—hence its name. This phenomenon has profound implications for various electronic components and circuits, including radio frequency (RF) electronics oscillators, transistors, and more. Read on as we delve into 1/f flicker noise. Noises that are created by various devices such as microphones, loudspeakers, traffic (cars), restaurants, helicopters, and factories in a real environment, but are not usually present in laboratory environments, and are sounds other than the speaker's voice, are called environmental (background) noises. There is also conversation noise that can affect the speech signal. [7]

Improving the quality of the speech signal plays an important role in audio telecommunication systems. It is very important to improve the quality of noisy speech in communication systems. Speech quality means having a clear speech without noise that is understandable to the listener. The quality of speech can consist of two parts: one is observing and hearing speech anywhere and the other is the ability to understand speech. Good and quality speech may be important for the listener. When we naturally listen to a broadcast speech, carried over the air, this speech may not be of the same quality from different broadcast sources, such as the real human voice that can be heard anywhere. Another category is the ability to understand speech, where we can hear what is being said carefully. This ability to understand can be measured in terms of percentage [4]. Speech quality has been evaluated with respect to natural intelligibility and suitable for software used in some applications, for example machine reading [5]. On the other hand, high-speed speech clarity is one of the important and necessary features for blind people, which can be evaluated when dealing with multimedia applications, which can be evaluated at different levels, such as phonemes, words, or sentences.

3.Methods of removing noise from the speech signal

In order to have a quality speech, different methods of removing noise from the speech signal are used. In order to understand which of the methods of removing noise from the speech signal and improving the quality of speech is more appropriate and to be able to compare them, evaluation criteria such as: signal-to-noise ratio (SNR), perceptual evaluation of speech quality (PESQ), segmented signal-to-noise ratio (SNR) seg are used.

SNR describes the total noise present in the output edge detected in an image, in comparison to the noise in the original signal level. SNR is a quality metric and presents a rough calculation of the possibility of false switching; it serves as a mean to compare the relative performance of different situations . Signal-to-noise ratio (often abbreviated as SNR or S/N) is a measure used in science and engineering that compares the level of a desired signal to the level of background noise. The signal-to-noise ratio, the bandwidth, and the channel capacity of a communication channel are connected by the Shannon–Hartley theorem1. Signal-to-noise ratio is sometimes used informally to refer to the ratio of useful information to false or irrelevant data in a conversation or exchange.

The presence of noise in speech significantly reduces speech intelligibility. Speech impairment severely affects a person's ability to understand what the speaker is saying, whether they have good, normal hearing or impaired hearing. Noise reduction or speech improvement algorithms are used to reduce background noise and improve perceptual quality and the ability to understand speech. The complexity of implementing noise reduction algorithms is also important in some applications, especially applications related to portable devices such as mobile communications and digital hearing aids. In the past decades, many methods have been proposed and developed to reduce the environmental noise of the speech signal. Most of these methods are based on dedicated arrays and mostly require several sensors. The number of available audio channels is an important factor in the design of speech purification systems. Generally, the more the number of microphones, the better the cleaning. [3] Speech noise removal methods can be divided into two categories: single-channel (single-microphone) and multi-channel. The discussion in this paper is limited to single channel noise reduction methods, as these are the most common types of methods found in many applications.

3. 1. The method of removing noise from the speech signal as a single channel

In single-channel methods, only one microphone is available to receive the speech signal. In a single-channel system, the microphone should be placed as close as possible to the speaker. So, in this case, the only input for the speech improvement system is a noisy speech signal. In this method, by assuming that the noise is static during the active areas of speech, the statistical characteristics of the noise are obtained through its values in the inactive areas of speech (silence). These areas are identified using a speech signal detector. [3] In 1992, a speech noise removal method was presented by Ephraim, and the amplification systems in this class are called methods based on statistical models. These methods are usually used when there is no knowledge of the statistical characteristics of the speech or the interference signal. Conversation production models such as average or moving average (MA), automatic regression (ARMA) and autoregressive (AR)

are also used. This is combined with the estimation of the parameters of the speech model, and then an estimate of the amplified signal is used with separate synthesis using the parameters of the speech model or using a Wiener or Kalman filter [4]. The Wiener filter method is one of the most important and oldest speech improvement methods, which was first introduced in the field of speech improvement in 1949 by Wiener. The goal of the Wiener filter is to compute a statistical estimate of an unknown signal using a related signal as an input and filtering that known signal to produce the estimate as an output. For example, the known signal might consist of an unknown signal of interest that has been corrupted by additive noise. The Wiener filter can be used to filter out the noise from the corrupted signal to provide an estimate of the underlying signal of interest.

One of the important advantages of these filters is the use of different algorithms that can minimize the output error by changing the filter weights. Among the most important disadvantages of these algorithms is the high dependence of the convergence behavior of the algorithm on the density function of the power spectrum of the input signal. Also, this filter is only designed for constant noise with zero average.

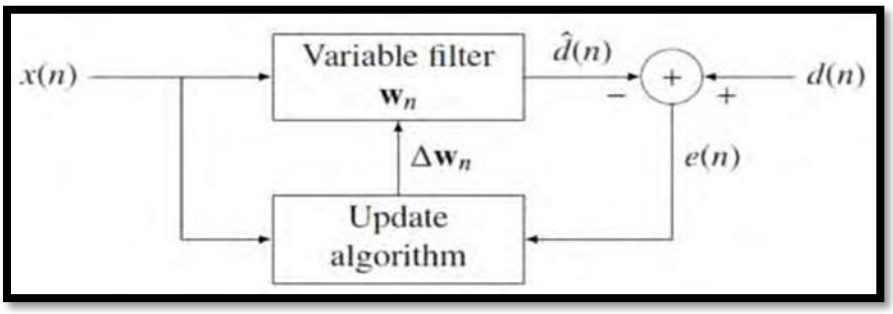

Figure 1: Block diagram of adaptive Wiener filter

According to the form X(n) (input signal), v(n) (noise signal), d(n) (desired signal), wn (filter coefficients or weights), n (number of repetitions), e(n) (error signal) and) d^ (n) are the estimation signal. In 2000, Deller et al. conducted studies in the field of speech noise removal and presented the short-term speech spectral domain (STSA) method. Short-term spectral domain (STSA) of speech has been successfully revealed in the preparation and development of different speech enhancement algorithms. The main idea is to use the short-term spectral domain of the noisy speech input and improve the estimate of the clean short-term spectral domain by removing the part attributed to additive noise. The input to the system is the noise-cut signal, Y(N).

The short-term Fourier transform (STFT) was proposed by Gabor in 1945. The short-term spectral shape transformation (STFT) of the signal with (OLA influence) has been used. The spectral range |Y(K)|of the noisy input signal Y(N) is changed using the correction factor. Usually, this correction factor can be the spectral amplitude of the estimated noise signal d(N), which is measured during periods of inactivity or silence in the conversational signal or obtained from the reference channel (two-channel method). The correction is obtained by reducing the spectral amplitude of the noise signal from the noisy speech input. Therefore, these methods are known as differential type algorithms.

A more detailed representation of the short-term spectrum amplitude amplification system of spectral difference is a well-known noise reduction method, which is based on the technique of short-term spectral amplitude estimation (STSA) and is shown in Figure (2) [6].

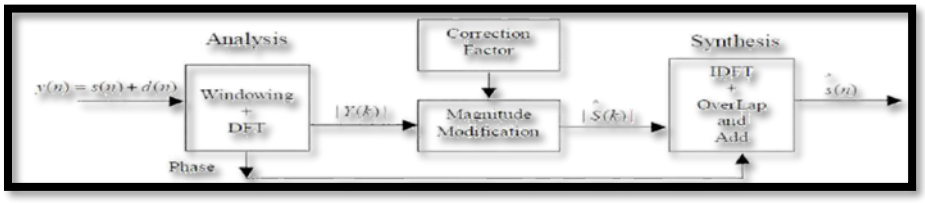

Figure 2: A more detailed representation of the short-term spectral difference spectrum amplification system. (STSA)

One of the important advantages of the short-term spectral domain method is that it works well in the preparation and development of speech enhancement algorithms. The disadvantage of using STFT transformation is that the length of the window remains constant at different frequencies, while many signals, such as speech signals, require more flexibility. In 1975, Boll founded another old and very famous speech improvement method, called spectral subtraction (difference), and in 1979, it was proposed with processing in the frequency domain. Spectral difference is used to recover the power spectrum of the signal contaminated with additive noise. This method subtracts the estimated noise power spectrum from the power spectrum of the noisy signal and produces the improved signal. In this method, the areas of speech and silence should be distinguished. It is also assumed that the noise remains stationary throughout the silence region and is uncorrelated with the original signal. In the spectral difference algorithm, first the initial spectrum of noise is estimated. This estimate is then updated using periods of silence where only noise is present. The advantages of the spectral difference method include less calculations, simple hardware, and low

cost. Among its disadvantages are the presence of residual noise called musical noise in the output signal and the roughness of the conversation due to the busy phase. [2]

In 1983, Ephraim conducted studies on noise removal methods to improve speech and introduced a statistical model. An example of the most common options for parametric modeling of the speech signal is the use of hidden Markov models (HMM). HMMs are Markov models whose process transitions between states in an unobservable path. However, output that is dependent on these states is visible. Basic theory behind HMM was published in a series of classical research papers. HMMs have been conventionally applied to problems that require the recovery of data sequences that are not immediately observable. These models have been used for speech recognition, gene prediction, part-of-speech tagging, activity recognition, etc.

In 2013, Liu and Yang conducted another research on removing speech noise to improve speech . They propose a noise reduction method based on the least-mean-square (LMS) adaptive filter of audio signals. It restores the desired audio signal by passing the noisy speech through a FIR filter whose coefficients are estimated by minimizing the mean square error (MSE) between the clean signals [8]. In many applications, LMS adaptive filtering algorithms are widely used, partly because they require less calculation and are simple to implement. It can also be delineated in the frequency domain, resulting in various derivative techniques [1]. For obtaining faster convergence, this paper will derive a normalized least square (NLMS) algorithm and the associated extended algorithm under Gaussian noise assumption. Simulation results indicate a higher quality of the processed speech signal than original observed signal.

It is possible to transmit a signal from one base to another for any purpose in different ways. One of these methods of time signal conversion was proposed by hohoho in 1995. The function is defined by the sum of the product of the said function, the scaled and shifted wavelet function in the whole time interval. [8]:

$$C(\text{scale},\ \text{position}) = \int_{-\infty}^{+\infty} (t)\ (\text{scale, position})dt$$

$$C(a, b) = \int_{-\infty}^{+\infty} (t)\ \frac{1}{\ }\ \left(\frac{t-b}{\ }\right)d(\ )$$

The multiplication of each of the coefficients in the corresponding scaled and shifted wavelet determines its contribution to the original signal. Shift means the movement of the wavelet along

the time axis and scale means the amount of expansion of the wavelet along the time axis. The decrease of the wavelet amplitude with the expansion of the scale in Figure (3) is to keep it constant. [8]

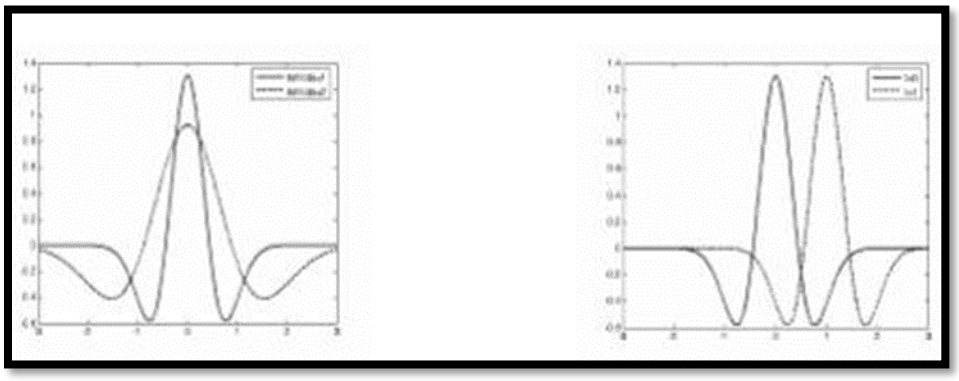

Figure 3: The right side of the shift and the left side of the scale of a wavelet

The reconstruction of the original signal is done through the following relationship:

$$C(a,b) = \int_{-\infty}^{+\infty} (t)\ \frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right) d(a)d(b)$$

$$(t) = \frac{1}{k} \int_{R} \int_{R}$$

In 1959, conducted activities in the field of noise removal using neural network. Artificial Neural Networks (ANNs) are finding increasing use in noise reduction problems and the main design goal of these Neural Networks (NNs) was to obtain a good approximation for some input output mapping. In addition to obtaining a conventional approximation, NNs are expected to generalize

from the given training data. The generalization is to use information that NN learned during training phase in order to synthesize, similar but not identical mapping [5].

The designed NNs are trained with input sequences that are assumed to be a composition of the desired signal plus an additive white Gaussian noise. The networks are expected to learn the noisy training data with the corresponding desired output and generalize the model. This research is an attempt to employ ANN for the enhancement of the measured corrupted signal and reduce the noise. The main contribution includes the following:

• The input training sequences to the designed NNs are assumed to be a composition of the desired signal plus an additive white Gaussian noise. This assumption speeds up the learning process and improves the approximation of the desired model [5].

• The development and comparison of NN architectures for use in noise reduction applications.

• A comparison of modeling performance using multi-layer and recurrent NNs.

• An examination of the relationship between training performance and training speed with the training algorithm used for a given NN architecture.

There are two main phases in the operation of ANN: learning and testing. Learning is the process of adapting or modifying the NN weights in response to the training input patterns being presented at the input layer. How weights adapt in response to a learning example is controlled by a training algorithm. Testing is the application mode where the network processes a tested input pattern presented at its input layer and creates a response at the output layer. Designing an ANN for a given application requires determining the NN architecture, the optimal size for the network (the total number of layers, the number of hidden units in the middle layers, and number of units in the input and output layers) in terms of accuracy on a test set, and the training algorithm used during the learning phase. Two types of neural networks are used to perform the required extraction of the knowledge from a noisy training set to achieve better signal enhancement [9].

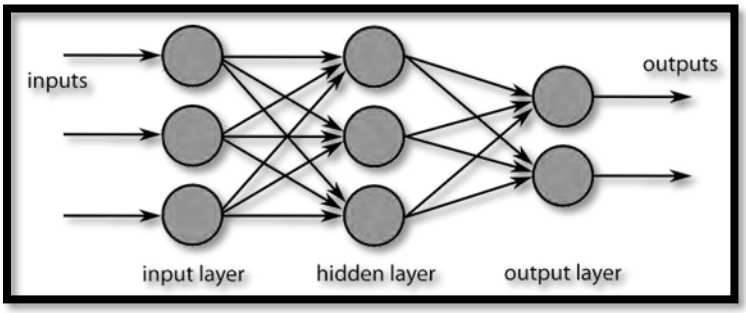

Figure 4: An artificial neural network

A feedforward neural network is one of the simplest types of artificial neural networks devised. In this network, the information moves in only one direction—forward—from the input nodes, through the hidden nodes (if any), and to the output nodes. There are no cycles or loops in the network.

The architecture of a feedforward neural network consists of three types of layers: the input layer, hidden layers, and the output layer. Each layer is made up of units known as neurons, and the layers are interconnected by weights [9].

- Input Layer: This layer consists of neurons that receive inputs and pass them on to the next layer. The number of neurons in the input layer is determined by the dimensions of the input data.
- Hidden Layers: These layers are not exposed to the input or output and can be considered as the computational engine of the neural network. Each hidden layer's neurons take the weighted sum of the outputs from the previous layer, apply an activation function, and pass the result to the next layer. The network can have zero or more hidden layers
- Output Layer: The final layer that produces the output for the given inputs. The number of neurons in the output layer depends on the number of possible outputs the network is designed to produce.

Each neuron in one layer is connected to every neuron in the next layer, making this a fully connected network. The strength of the connection between neurons is represented by weights, and learning in a neural network involves updating these weights based on the error of the output.

Applications of Feedforward Neural Networks

Feedforward neural networks are used in a variety of machine learning tasks including:

- Pattern recognition
- Classification tasks
- Regression analysis
- Image recognition
- Time series prediction

Despite their simplicity, feedforward neural networks can model complex relationships in data and have been the foundation for more complex neural network architectures.

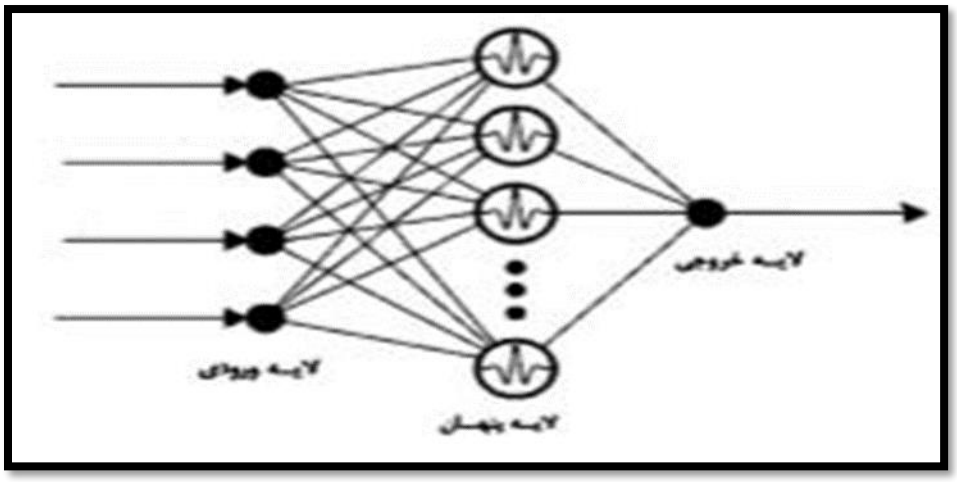

Figure 5: overview

4. Conclusion

Due to the destructive effect of various types of noise on the quality of speech and the distortion of the clean speech signal, speech recognition by humans in telephone and wireless communications faces many problems. Also, it is very difficult to identify noisy speech by computer systems in applications such as automatic speech and speaker identification. Therefore, one of the most important stages of speech signal processing is to remove noise from the noisy signal in order to improve the quality of speech in electronic systems. In this article, by reviewing the speech noise removal methods, Some of these methods, such as the hidden Markov model method, are computationally long and time-consuming. Therefore, among these methods, those that have the possibility of cheap implementation and require less computing power, and in addition, do not need multiple sensors and are able to remove noise on single-channel noise. Such as: spectral difference method because commercial speech processing systems such as digital

answering machines, mobile phones or even smart phones cannot have multiple inputs, so in this case they need multiple analog-to-digital models, which causes the system price to rise. In addition, the system with multiple inputs requires more memory and higher processing power, which directly leads to an increase in price.

References

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on acoustics, speech, and signal processing, 27(2):113–120, 1979.

[2] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? Journal ACM, 58(3):11:1–11:37, 2011.

[3] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. Speech Communication, 10(1):45 – 57, 1991.

[4] M.A. Ben-Messaoud, A. Bouzid, and N. Ellouze. Speech enhancement based on wavelet packet of an improved principal component analysis. Computer Speech & Language, 35:58 – 72, 2016.

[5] M.S. Khan, S.M. Naqvi, and J. Chambers. A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6566–6570, Canada, 2013. IEEE.

[6] A.A. Petrovsky, M. Parfieniuk, and A. Borowicz. Warped DFT based perceptual noise reduction system. In Audio Engineering Society Convention 116. Audio Engineering Society, 2004.

[7] Y. Lu and P.C. Loizou. A geometric approach to spectral subtraction. Speech communication, 50(6):453–466, 2008.

[8] S. Vihari, A. Sreenivasa, P. Soni, and D. Naik. Comparison of Speech Enhancement Algorithms. Procedia Computer Science, 89:666 – 676, 2016. Twelfth International Conference on Communication Networks, ICCN 2016, August 19–21, 2016, Bangalore, India Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19-21, 2016, Bangalore, India.

[9] V. Sunnydayal and T. Kishore-Kumar. Speech enhancement using sub-band wiener filter with pitch synchronous analysis. In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 20–25, Aug 2013.