Separate Speech Signal in Mobile Phone Calls from the Noise of Environment in Real-Time

Mohamad Al-Sadi
Syrian Virtual University (SVU)
Damascus, Syria
{ise34857@gmail.com}

ABSTRACT: *During voice calls through a mobile phone,there are a large number of unwanted sounds that transmitted with our speech signals, such as the noise of cars on the street, the machines in industrial places, group conversations, the sounds of people around us like children's screams and other sounds. These sounds are unwanted noise, because it affects the quality of the speech signal on the one hand. On the other hand, it provides information about where we are located now and what is environment around us. This is a private information,we do not want them be known by those we talk to. In this paper, we propose to isolate the speech signal that transmitted through mobile phone from the surrounding sounds, without affecting the quality of the signal and any delay time. We have implemented many algorithms. That take input as a sound signal with noise, and then the system estimates the random noise signal and deletes it from the input signal in real-time. That will improve the quality of the transmitted or recorded speech signal.*

Keywords: Voice Signal, Speech Signal, Transmitted Speech Signal, Recorded Speech Signal, Delay Time, Real-Time

Received: 29 August 2019, Revised 4 December 2019, Accepted 12 December 2019

DOI: 10.6025/stj/2020/9/1-16

1. Introduction

Natural language processing (NLP) systems intersect with modern digital communication systems in a field called "Speech Voice Signal". NLP is interested in processing this signal and studying its compounds and changes in both Time and Frequency domains in order to detect and recognize it, While Communication systems are concerned with how to transmit speech signals in different types of transport. However, the Speech Signal must be of a high quality before transmission so that the listener can understand the spoken speech. To achieve this quality, we must eliminate sources of noise that cause distortion of the signal.

Noise in the spoken voice signal is caused by several factors [1] [2]:

- **The noise around the Source of the useful Signal:** We consider any non-human voice is a noise for us.
- **Overlapping with Other Human Voices:** When we want to study the voice signal for a specific person, any different human voice or overlapping with it, we consider it a noise.
- **Voice Echo:** A phenomenon caused by the reflection of the sound wave and the arrival of later versions of it after periods of time

- often ms - from the arrival of the original sound wave.

• **Analog to Digital Converter (ADC):** Most modern communication systems are digital systems and based on ADC switches, where each sample of the signal is represented on a limited number of bits, which should be as small as possible without affecting the sound quality. This is also an important reason for the noise.

Today we use mobile phones everywhere and make voice calls over the Internet or the GSM network. All or some of the mentioned noise types may affect our calls. So why not think about using natural language processing to extract the useful speech signal during voice calls. We can dampen or cancel all kinds of noise that we have mention above, as we will explain in the next paragraph.

2. Research Goal

The purpose of this study is to provide an applied system for a type of artificial intelligence, which is the processing of sound signal and the extraction of useful information from it, this process similar to the mechanism of our human brain. To illustrate the idea of the research we can suggest the following example: If we walk with someone on a crowded street and talk to him, our hearing system picks up all the sound signals, such as car noise, engine noise and people's words. However, our brain can easily focus on the voice of the person we are talking to and delete all other sounds that are a noise to it. In a similar way, we will implement and test a system based on mathematical algorithms and adaptive filters to use it for remove noise from the speech signal. In our system, the input is a voice signal with noise and the system outputs is an improved signal without any type of noise. In real-time and without any delay. Figure1 illustrates the block diagram for the designed system.

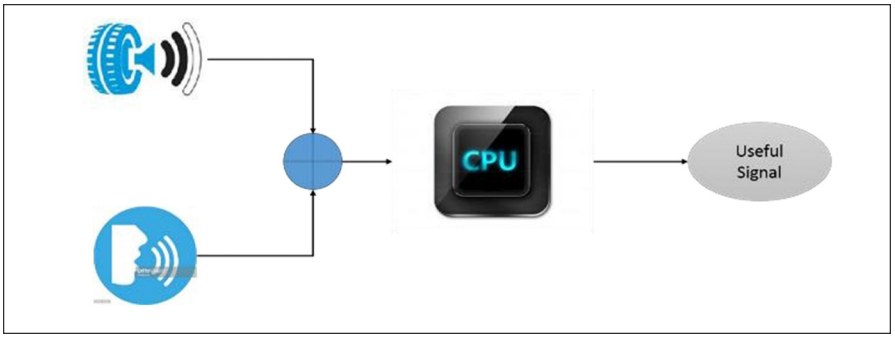

Figure 1. Block diagram for the designed system

As shown in Figure 1, the system's income is a composite signal of the useful speaker signal and a set of interconnected sounds, such as car sounds, human sounds, etc. The process begins by converting the analog signal to a digital signal A/D. then the digital signal enters a specialized algorithm for noise cancellation, which estimates the noise signal and deletes it, thus giving the output an enhancement and pure signal of noise. Figure 2 illustrates progressive stages of work in the proposed system.

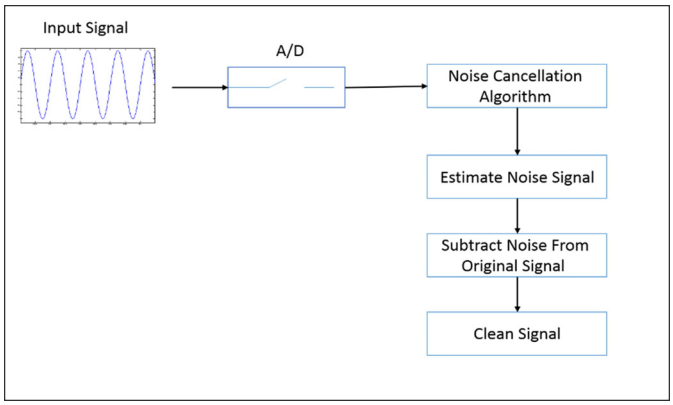

Figure 2. The stages of work in the proposed system to remove noise from the speech signal

We will test a set of algorithms and display the test results to demonstrate the effectiveness of the designed system to recognize speech signals and dampen the rest of other human and non-human sounds, which have low power compared to the main source signal, which negatively affects the quality of the transmitted signal and reduces SNR[1] rate. This system can be included in the operating systems of smartphones and computers, to become part of it.

3. Actual Challenges of the Proposed System

The proposed idea is not intuitive or easy to implement. In fact, there are many complexities in noise estimation and deletion issue. In this paragraph, we will present the noise that may exposed to the signal in the time and frequency level to highlight the importance of this idea and the challenges implicit its implementation. Figure 3 shows the clean signal at both the frequency and time levels, where we note that the spectrum of this signal contains only two compounds that correspond to the sine wave signal [7].

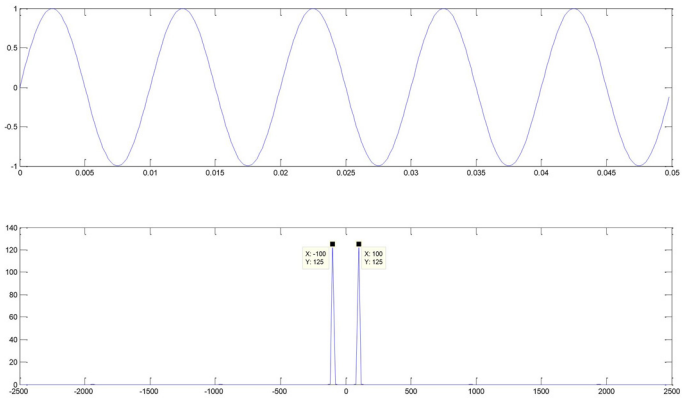

Figure 3. The sinusoidal signal without noise at both time and frequency levels

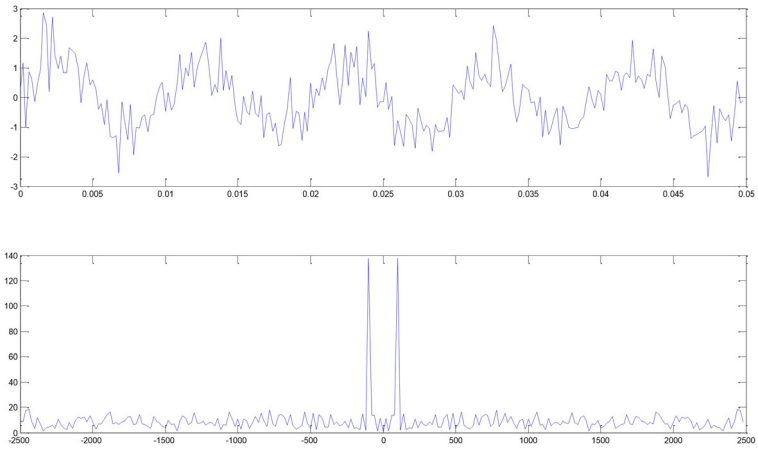

Figure 4. Signal with noise

The figure 4 illustrates the sinusoidal signal with noise; we note that the spectrum of this signal, which is supposed to contain only two compounds, contains many frequency compounds. The algorithms should estimate and delete these compounds [7].

This is a simple example, but in the case of speech signals, the frequency spectrum of the speech and noise signal is cross, here is the extreme difficulty of the issue.

---

[1] **SNR:** Signal to Noise Ratio

4. Origin of Speech Signals

The speech waveform is a sound pressure wave originating from controlled movements of anatomical structures making up the human speech production system. A simplified structural view is shown in Figure 5 Speech is basically generated as an acoustic wave that is radiated from the nostrils and the mouth when air is expelled from the lungs with the resulting flow of air perturbed by the constrictions inside the body. It is useful to interpret speech production in terms of acoustic filtering. The three main cavities of the speech production system are nasal, oral, and pharyngeal forming the main acoustic filter. The filter is excited by the air from the lungs and is loaded at its main output by a radiation impedance associated with the lips.

The vocal tract refers to the pharyngeal and oral cavities grouped together. The nasal tract begins at the velum and ends at the nostrils of the nose. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

The form and shape of the vocal and nasal tracts change continuously with time, creating an acoustic filter with time-varying frequency response. As air from the lungs travels through the tracts, the frequency spectrum is shaped by the frequency selectivity of these tracts. The resonance frequencies of the vocal tract tube are called formant frequencies or simply formants, which depend on the shape and dimensions of the vocal tract.

Inside the larynx is one of the most important components of the speech production system —the vocal cords. The location of the cords is at the height of the ‘‘Adam’s apple’’— the protrusion in the front of the neck for most adult males. Vocal cords are a pair of elastic bands of muscle and mucous membrane that open and close rapidly during speech production. The speed by which the cords open and close is unique for each individual and define the feature and personality of the particular voice.[1]

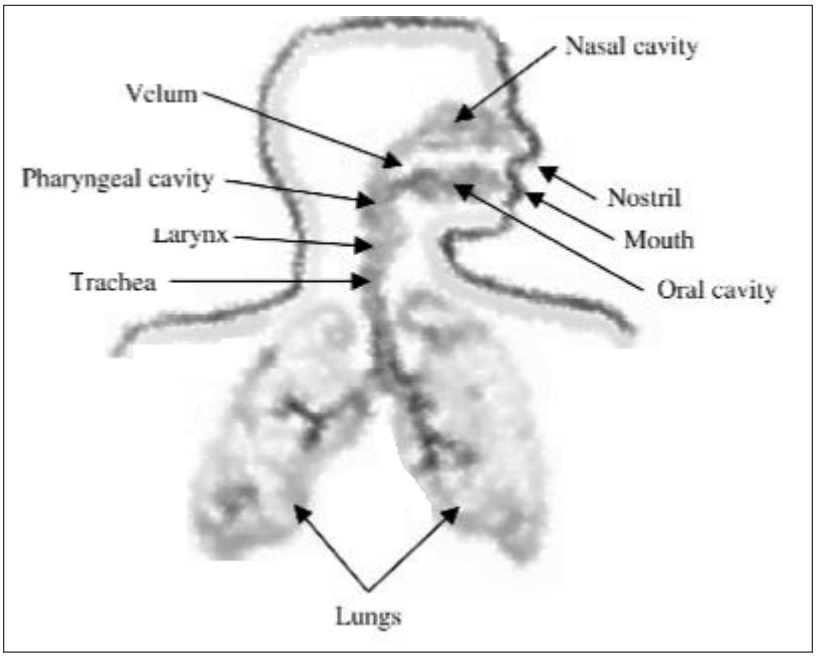

Figure 5. Diagram of the human speech production system

5. Speech Signal Processing Systems

There are two types of speech processing systems in the time domain: [1]:

• **Processing based on Frame:** In these systems, the speech signal spectrum is measured once in each frame, and the frame length is between 5 to 20 milliseconds. In order to extract the features vectors, we use one of the following analysis techniques: Fast Fourier transform, linear prediction coefficients or Mel-Cepstrum coefficients.

• **Processing depending on Segment or Landmark:** The spectrum is measured at certain points only, these points called segment or Landmark. The spectrum is measured between 6 to 10 times at each mark.

6. Related Works

In this paragraph, we present a simplified theoretical explanation about the most important algorithms and methods used so far in the field of deleting the negative effects on the sound signal. The focus will be on the issue of echo cancellation and the noise cancellation [3] [2].

6.1. Echo Cancellation Issue: [9]

Adaptive filters play an important role in echo deletion as they change their coefficients to fit the input signal changes and propagation environment. Figure 6 shows the general diagram of echo cancellation process:

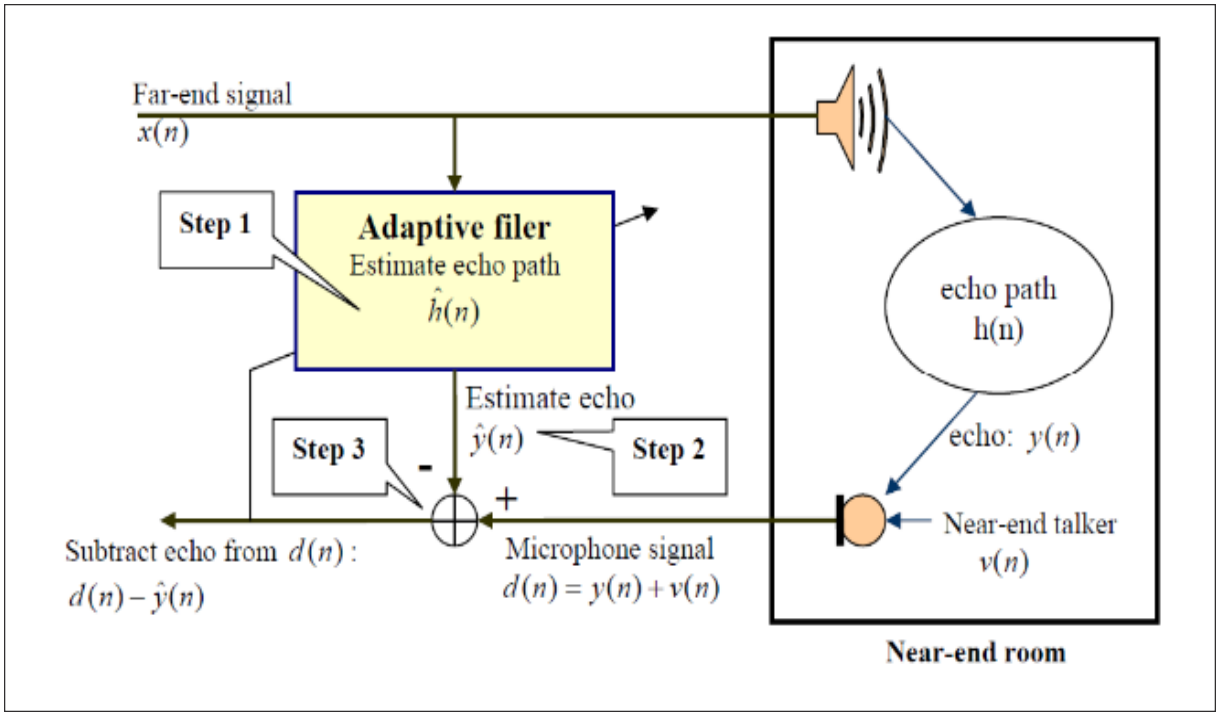

Figure 6. The general diagram of echo cancellation process

Some effective echo cancellation algorithms process signals through the following basic steps:

Step 1: Estimate the characteristics of echo path $h(n)$ of the room: $\hat{h}(n)$

Step 2: Create a replica of the echo signal: $\hat{y}(n)$

Step 3: Echo is then subtracted from microphone signal $d(n)$ (in- cludes near-end and echo signals) to obtain the desired signal as in relation (1).

$$Desired\ Signal = d(n) - \hat{y}(n) \ldots\ldots\ldots (1)$$

In the following paragraph, we will present a number of important and effective echo deletion algorithms.

6.1.1. Wiener filter: [7]

The work of the Weiner filter can be described as in Figure 7. We note that $x(n)$ represents the input signal of the adaptive filter $h(z)$, $d(n)$ represents the filter output signal, and $\hat{d}(n)$ the signal we want to estimate.

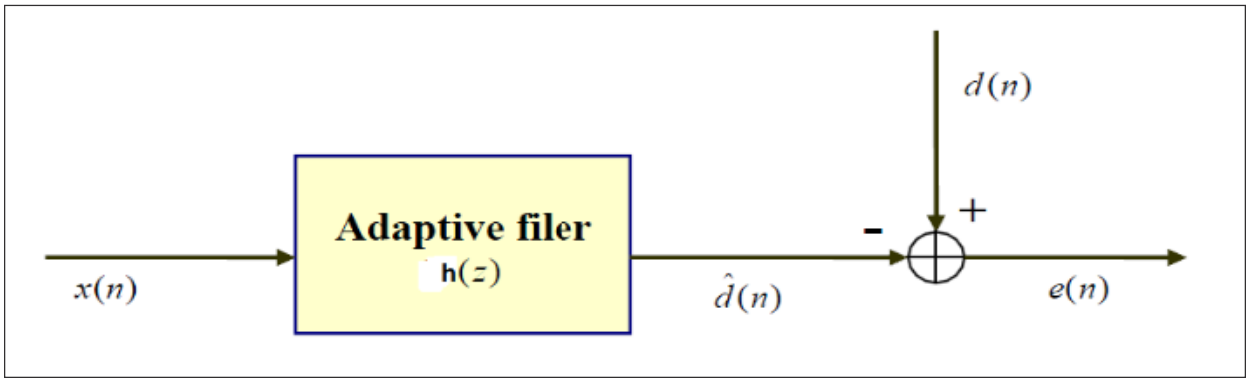

Figure 7. Illustrates the work of the Weiner filter

The constants required for the filter are the constants that give the smallest value to the MSE[2]. When calculating, we obtain the formula of equations (2), called Wiener-Hopf equations, which can be written in matrices:

$$\sum_x h_M = R_{dx} \tag{2}$$

Where Σ_x is a matrix whose elements are:

$$[\sum_x]_{lk} = R_x(l-k)$$
$$h_M = [h_0, \dots, h_{M-1}]^T$$
$$R_{dx} = [R_{dx}(0), \dots, R_{dx}(M-1)]^T$$

From the above we obtain the constants of the filter using the relationship (3):

$$h_m = \sum_x^{-1} R_{dx} \tag{3}$$

When these values are set for the constants of the filter, MMSE is given in relation (4):

$$\text{MMSE}_M = \min(\xi_M) = \sigma_d^2 - \sum_{k=0}^{M-1} h(k) R_x^*(k) \tag{4}$$

Where $\sigma_d^2 = E(|d(n)|^2)$.

SNR is defined in relation (5) as the ratio of the power input signal to the noise power.

$$\text{SNR} = \frac{E(|d(n)|^2)}{E(|w(n)|^2)} \tag{5}$$

Where $w(n)$ is the noise signal. To find out the relationship between the *SNR* and the *MSE*, we simulated 10 stable non-variable gausic signals with time (iid). At each *SNR* value, we calculate Wiener filter constants and *MSE*, we will obtain Figure 8. Where we find that the *MSE* value is decreasing with *SNR* value.

[2] MSE : Minimum Square Error

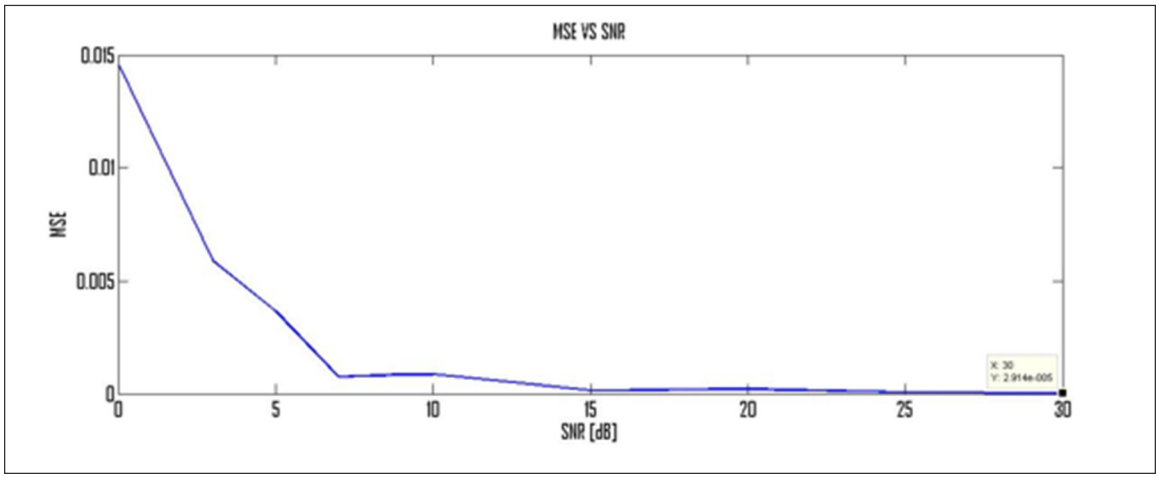

Figure 8. Illustrates the change in MSE value for SNR value

6.2. Noise Cancellation Issue: [10]

In this paragraph, we review the most famous algorithms used in the field of noise cancellation.

6.2.1. Linear Prediction

Linear prediction is the prediction of a future sample of a signal from a previous set of samples. This model is used in various applications such as audio and video codecs and noise reduction. The purpose of linear prediction is to model the mechanism of the production of bonding in the signal, which is widely used in speech processing applications such as speech coding at low encoding cost and improved speech signal quality. Assuming we have a signal have different amplitudes until the moment m .we want to guess $x(m)$ where we have p from previous samples

$$x(m-1) \ldots\ldots x(m-p)$$

$$\hat{x}\,(m) = \sum_{k=1}^{p} w_k x(m-k)$$

Where w_k is Prediction Coefficients and $w = R_{yy}^{-1}\, r_{yx}$. The Block diagram in Figure 9 can illustrate the linear prediction process.

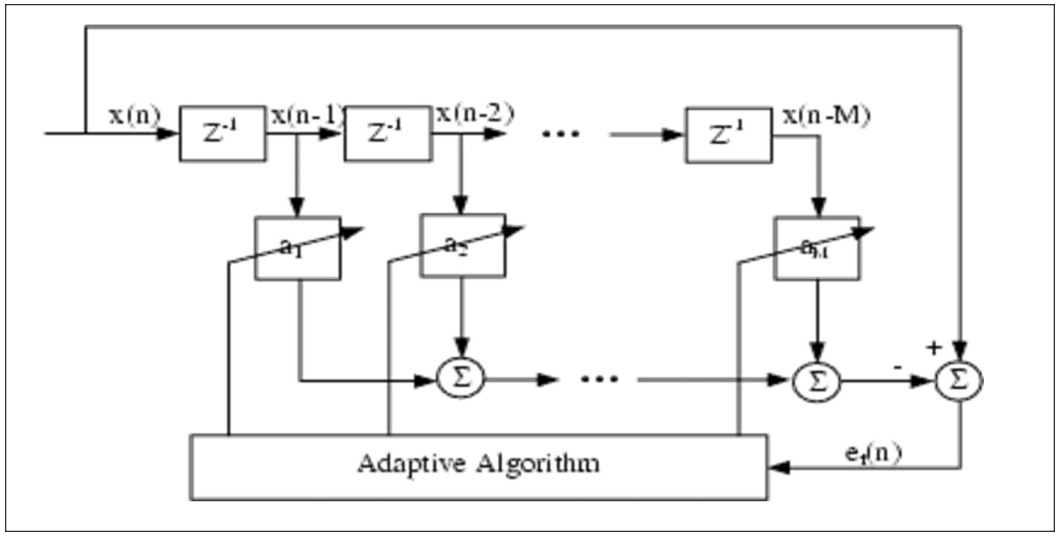

Figure 9. Block diagram for linear prediction process

7. The Algorithms used in Our Research

We have implemented and tested three algorithms for the noise deletion process. We programmed these algorithms using the MATLAB program because it provides a very large number of mathematical functions. We take the following steps:

• Record the audio clip containing the speech signal with the noise.

• The program converts this analog signal to a digital signal at a frequency of 8000 KHz and each sample is represented at 16 bits.
• So we have obtained the input signal as a column matrix.

• We enter this matrix into the function that represents the algorithm that we want to implement.

• Then we get the output signal and then read this signal through special window. We turn now to explaining the algorithms used in our research.

7.1. Spectral Subtraction algorithm: [11][13]

In figure 10 we shows how Spectral Subtraction algorithm worked.

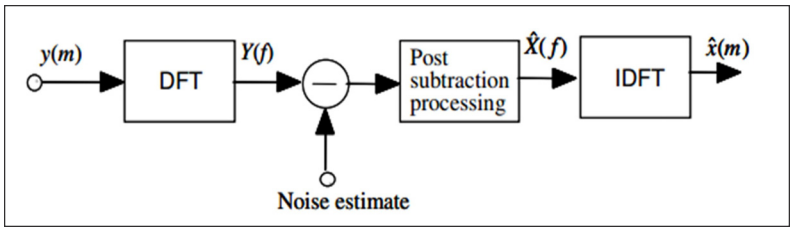

Figure 10. Block diagram for Spectral Subtraction algorithm

This algorithm is used to eliminate spectral distortion that can be observed through the energy spectrum or amplitude Fourier transform for signal. and is used to eliminate the collective noise that occurs on the signal. The frequency of the noise signal spectrum is estimated periodically and continuously to correct the mean noise Power. This operation is performed when the signal is absent.

This algorithm is based on estimate the median energy of the noise signal spectrum and subtracts it from speech signal spectrum. The spectral subtraction process can be illustrated by the equation (6):

$$\left|\hat{X}(f)\right|^b = |Y(f)|^b - \alpha\overline{|N(f)|^b} \tag{6}$$

Where the parameter alpha determines the amount of noise to be deleted, and parameter *b* specifies two cases:

• $b = 1$amplitude Subtract of spectrum

• $b = 2$ Subtract from power

The noise signal spectrum is estimated in the absence of the original signal, after dividing this signal into frames according to equation (7) (assume we have *k* farme):

$$\overline{|N(f)|^b} = \frac{1}{k}\sum_{i=0}^{k-1}|N_i(f)|^b \tag{7}$$

The frequency spectrum of the signal is calculated using a digital low pass filter according to equation (8):

$$\overline{|N_i(f)|^b} = p\overline{|N_{i-1}(f)|^b} + (1-p)|N_i(f)|^b \tag{8}$$

In order to restore the signal to the time domain, we multiply the signal spectrum estimated by the phase of the signal that have

noise, and then we perform reverse Fourier transform according to equation (9). This equation is based on the basic idea of the audible noise in the noisy signal resulting from the impact of the signal spectrum only.

$$\hat{x}(m) = \sum_{K=0}^{N-1} \left|\hat{X}(K)\right| e^{j\theta_Y(k)} e^{-j\frac{2\pi}{N}Km} \tag{9}$$

7.2. Wavelet Threshold Algorithm: [12] [14]
Wavelet is a limited continuity signal with zero rate, as opposed of a sine wave signal that extends theoretically from (–∞,+∞) and has a beginning and an end, Figure 11 shows the wavelet at the time domain.

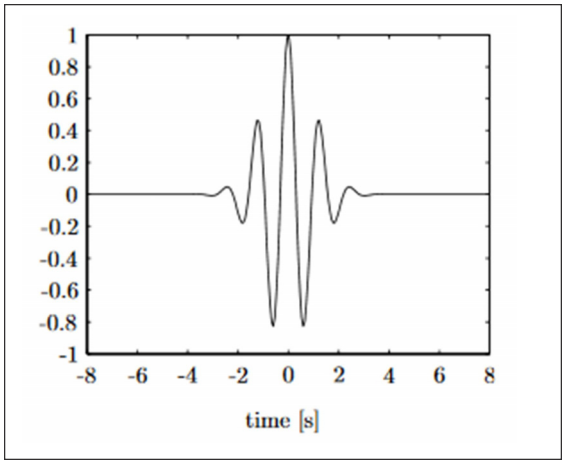

Figure 11. The Wavelet at time domain

Since the noise in the signal processing is a concern signal in one or all frequency bands, which is generally an undesirable signal, the less uniform noise that needs to apply the most sophisticated methods of signal purification.

The distinction between noise types is based on its properties in the domain of time and frequency, For example, white noise refers to a noise distribution in all frequency bands. In comparison to Gaussian noise and mono noise, Gauss noise characteristics indicate a probability density in the time domain. Mono noise has a fixed probability density in the specified period.

Sound filtering using wavelet transform is based on the basic idea that the signal energy is concentrated in speech signals while wavelet transform parameters for the noise signals are very small, allowing us to delete these parameters or replacing them with zeros, which gives the ability to delete noise and restore the pure signal.

Here we represent the speech signal that we want to purify as we did in the previous paragraph as a set of pure signals plus noise signals as in equation (10):

$$y(n) = x(n) + n(n) \tag{10}$$

Figure 12 illustrates how the wavelet algorithm works, but it should be noted that the general procedure for signal purification involves three basic steps:

• **Analysis:** In this steep, we select the wavelet and the level *N* then calculate the wavelet analysis for *x* signal in the level *N*.

Thresholding: In which we specify the detailed parameters of the threshold, for each level from 1 to *N*, the Thresholding process includes determining the type of threshold either soft or hard.

• **Reconstruction:** The re-construction of the wavelet is done by using the original detailed information in level *N* and adjusting the detailed information from 1 to *N* levels.

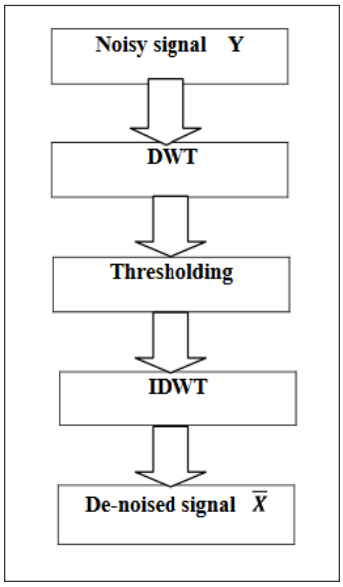

Figure 12. Flow chart for Wavelet Threshold process

7.3. Weiner Filter: [15] [16][17]

As explained above, the Weiner filter is used to delete the acoustic echo and we will prove in this paragraph that it can be used to eliminate noise from the speech signal and the purification process.

The Weiner filter is used to eliminate noise from speech signals in an adaptive manner. The power of this filter is capable of working without the need to estimate the signal spectrum or noise spectrum in advance. It is superior on traditional methods, which depend on damping the noise signal and maintaining the original signal after passing the noisy signal into a filter as shown in figure 13. These methods are called direct filtration. Figure 14 shows the structure of adaptive filters.

Figure 13. Traditional methods of noise deletion

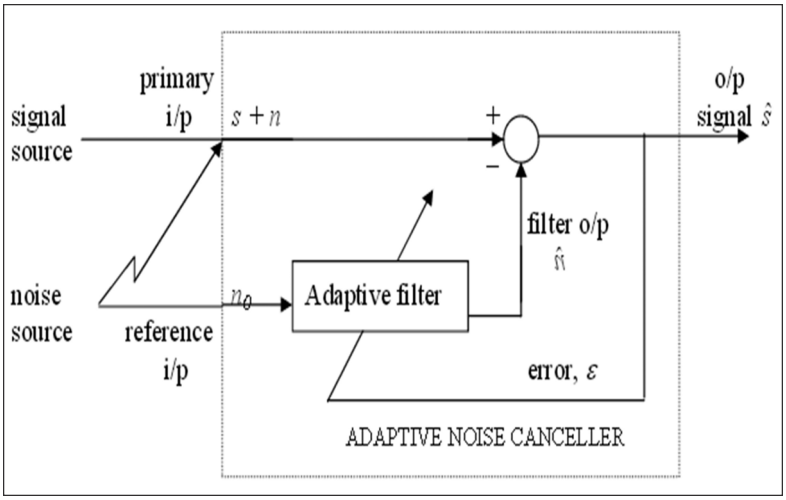

Figure 14. Structure of Adaptive Filter

Note that there are two types of input:

-Primary signal: which contains the primary signal plus the noise.

- Reference signal: which contains noise only and which are somehow interconnected with noise that added to the base signal.

Then the noise signal is estimated and deleted from the noisy signal in order to obtain the pure signal, where the estimated signal is given in relation (11):

$$\hat{s} = s + n - \hat{n} \Rightarrow$$
$$\hat{s}^2 = s^2 + (n - \hat{n})^2 + 2s(n - \hat{n}) \quad (11)$$

As noted in this paragraph, the noise deletion here is based on the idea of deleting the noise from the speech signal, but there must be some correlation between this noise in the reference signal and the noise in the primary signal. Thus, we will discuss two cases:

The First Case: Non-correlating input in the primary signal: Figure (15) shows noise signal non-Linked with the primary signal. Note that there are two noise signals, n and m. The reference channel only contain the noise $n*h(j)$ where h is the pulse response of the channel.

Both noise n and $n*h(j)$ have the same origin therefore, they are correlated in one way or another. The required response is:

$$s + m_0 + n \quad (12)$$

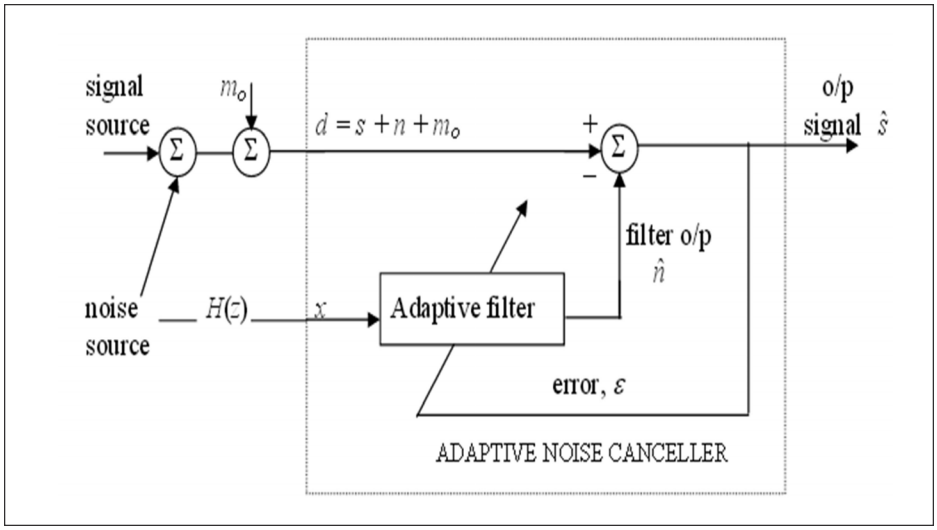

Figure 15. Non-correlated noise with the primary signal

Assume that the adaptive process is used for the MSE solution, then the adaptive filter is a Wiener filter and the transform function is as relation (13):

$$W^*(z) = \frac{\delta_{xd}(z)}{\delta_{xx}(z)} \quad (13)$$

The input spectrum for filters will be as relation (14) and (15):

$$\delta_{xx}(z) = \delta_{nn}(z)|H(z)|^2 \quad (14)$$

$$\delta_{xd}(z) = \delta_{nn}(z)H(z^{-1}) \quad (15)$$

Thus, the Weiner filter response is as in equation (16)

$$W^*(z) = \frac{\delta_{nn}(z)H(z^{-1})}{\delta_{nn}(z)|H(z)|^2} = \frac{1}{H(z)} \quad (16)$$

The Second Case: Non-correlating input in the reference signal: Figure (16) shows noise signal non-Linked with the reference signal. The noise on the reference signal representative by equation (17):

$$m_1 + n * h(j) \quad (17)$$

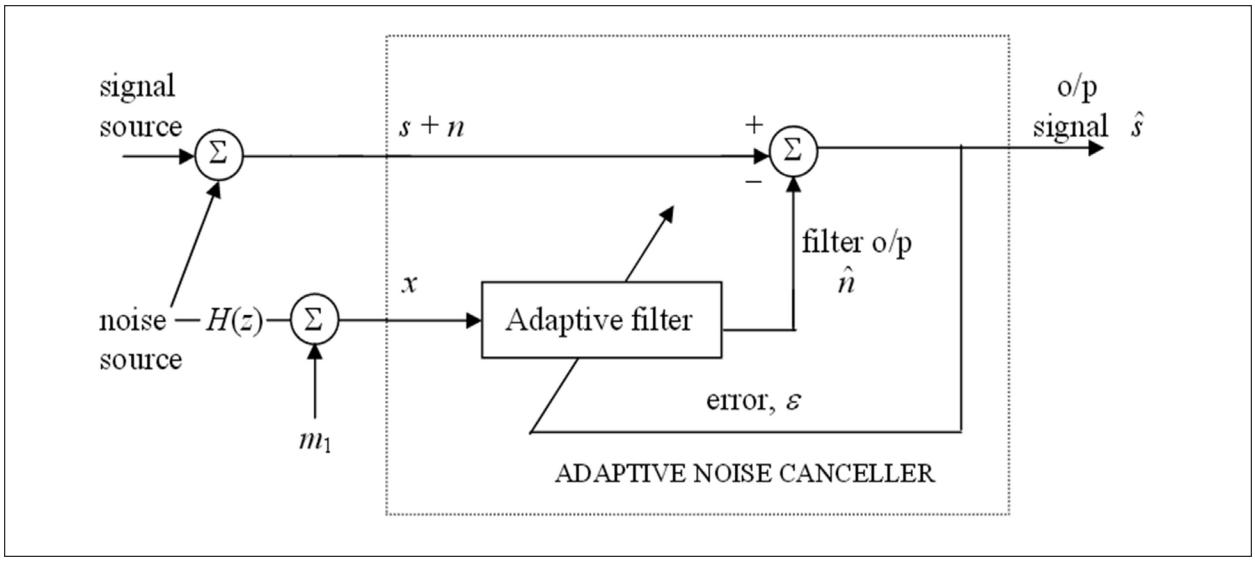

Figure 16. Non-correlated noise with the reference signal

Depending on the relationships 13, 14, 15 and 17, the filter's output in this case is as in Relationship (18):

$$W^*(z) = \frac{\delta_{nn}(z)H(z^{-1})}{\delta_{m1m1} + \delta_{nn}(z)|H(z)|^2} \quad (18)$$

8. Program Implementation

After presenting the theoretical study and determining the algorithms that we will work on, these algorithms explained in the paragraph 7. We implemented these algorithms within the Matlab program. Figure 17 shows the main form of the program that we have implemented.

This interface allows the user to choose how to delete the noise by checking the algorithm's check box. Then the user selects WAV file as input signal or records a voice clip with noise and determines the duration of the clip and number of frames per second that he wants. With the possibility of hearing this clip before deletion and after deletion. After pressing the "Start" button, the program will execute the specified algorithm and delete the noise with draw output signal on form. Figure 18 illustrates the program during work.

9. Experimental Results and Observations

In this paragraph, we present the practical results for our designed system. Where we will select a noisy input signal and pass it as parameter to the three algorithms in order to determine which ones are the best. Figure 19 shows the output of Spectral Subtraction algorithm. While Figure 20 shows the output of the Wavelet Threshold algorithm, finally Figure 21 shows the output of Weiner algorithm.

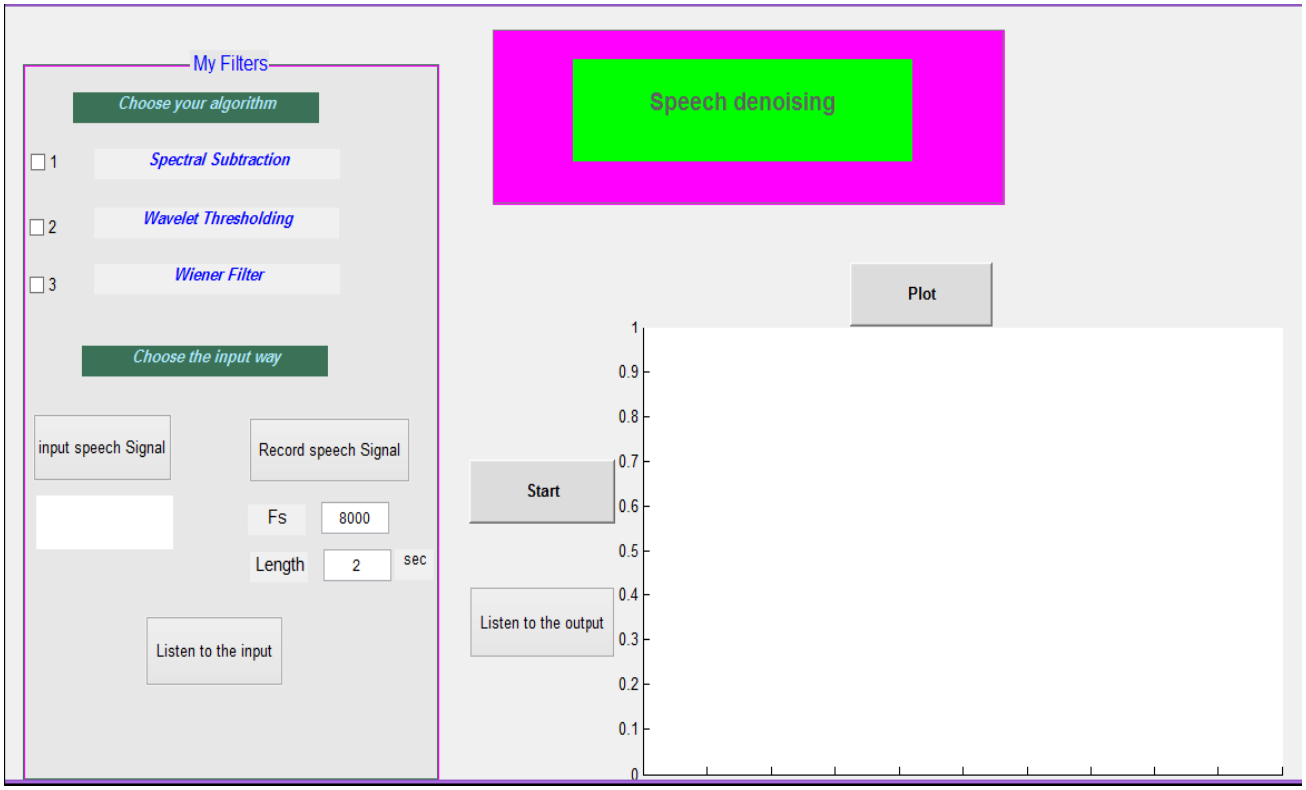

Figure 17. The program that we have implemented

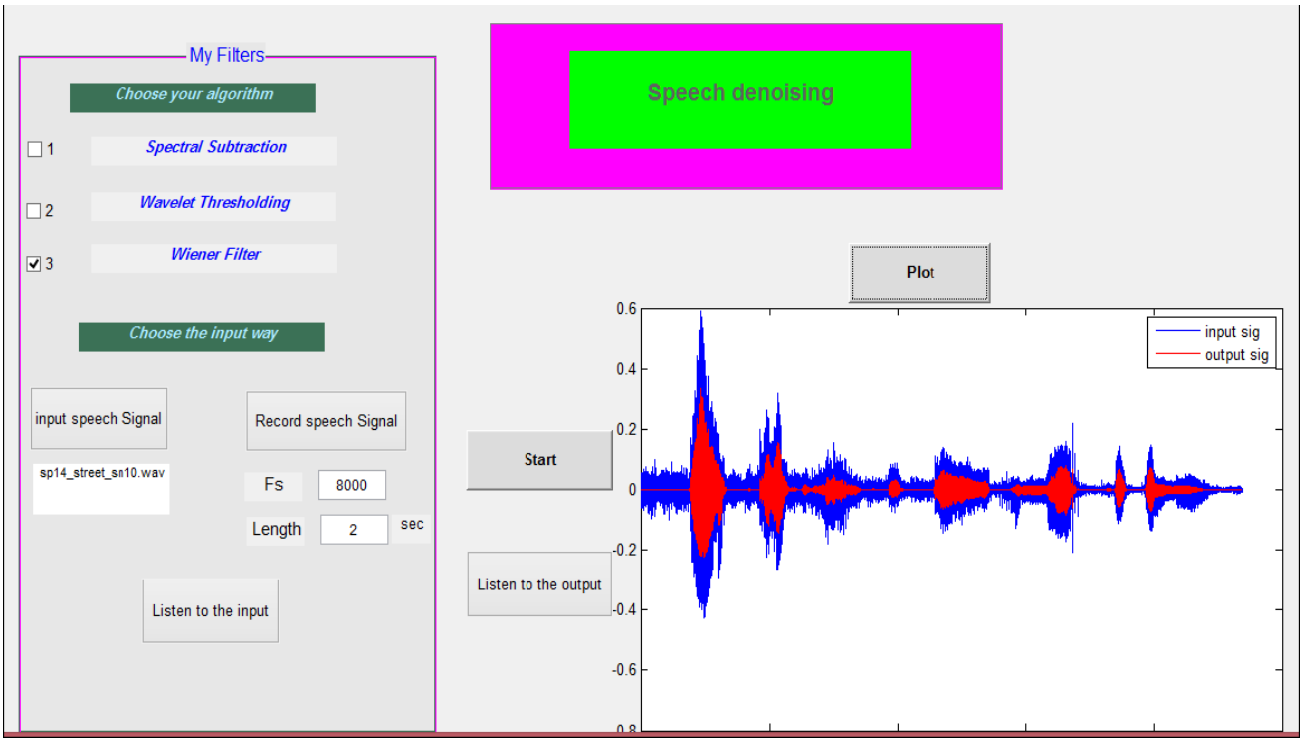

Figure 18. Our program during work

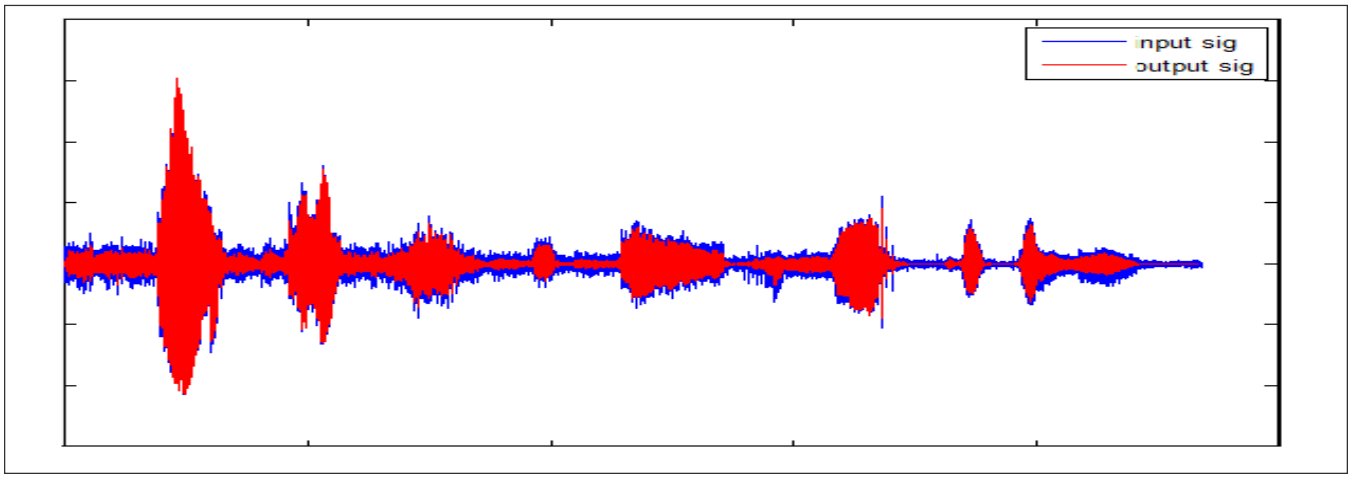

Figure 19. Output of the Spectral Subtraction algorithm

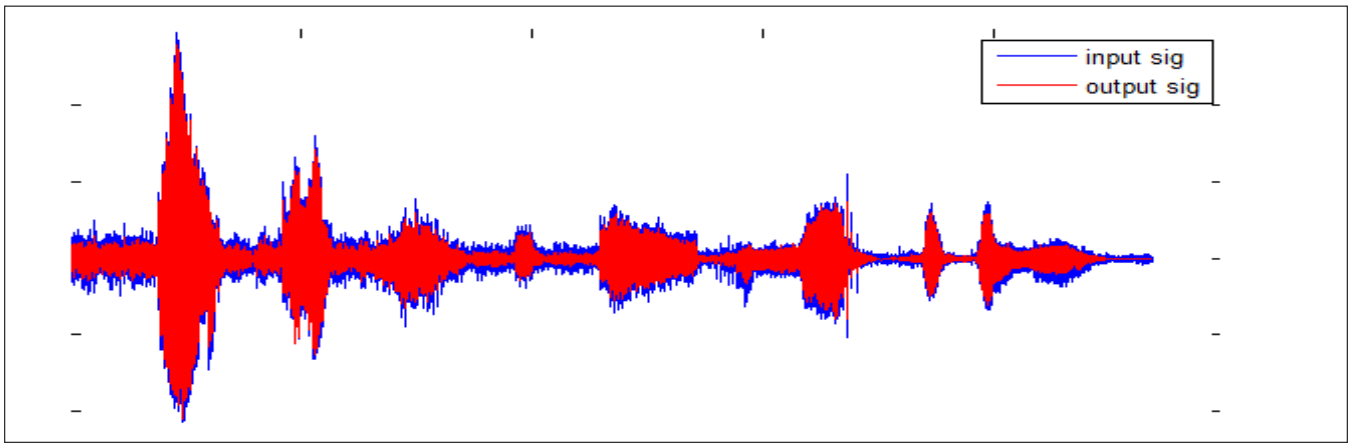

Figure 20. Output of the Wavelet Threshold algorithm

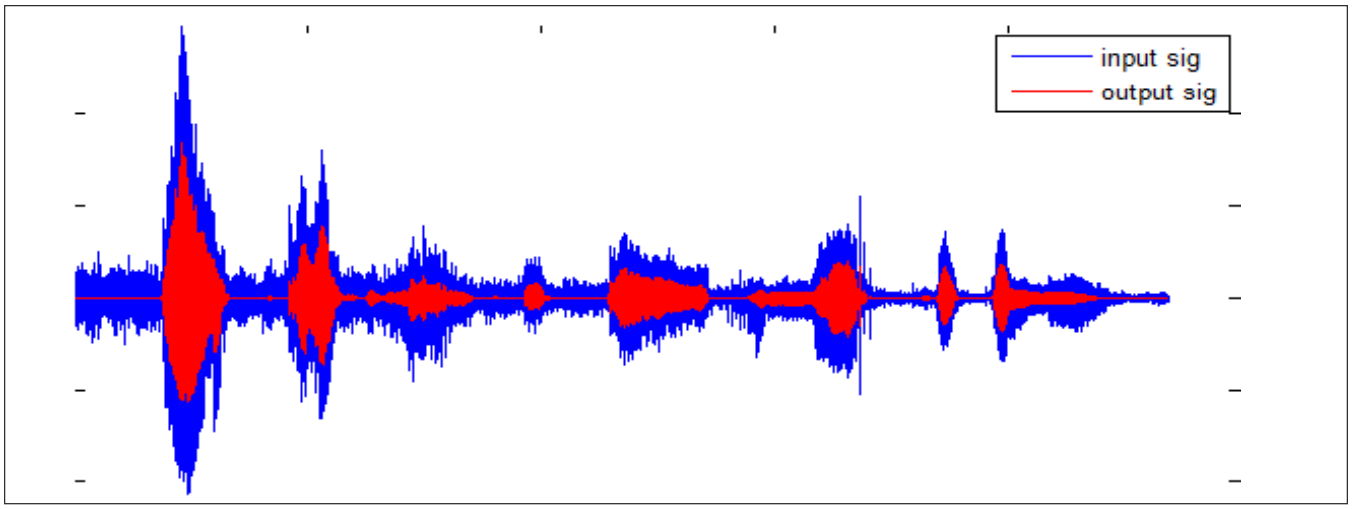

Figure 21. Output of the Weiner Filter algorithm

We Note that the best algorithms are the Wiener filter algorithm after enhancement it and used it to noise cancellation rather than echo. Followed by a Wavelet threshold algorithm and finally a Spectral Subtraction algorithm. The Wiener filter algorithm can be adopted in our proposed system to delete noise from around us during voice calls.

10. Conclusion

We must remember what was completed during this research, we designed voice-processing system capable of deleting the noise from the speech signal in the voice calls, which may have been exposed for different reasons, as we mention at the beginning of the search. This system has been implemented within the MATLAB R2013 a. The most important achievement was the implementation and testing of the following algorithms: Spectral Subtraction, Wavelet Threshold, and Wiener Filter. We found that the best of these algorithms is the Wiener filter algorithm followed by Wavelet threshold and then the Spectral Subtraction algorithm.

In the future, we aim to develop this system and apply it in real environments, to become a general software library that can be included in any operating system for mobile phones or personal computers.

References

[1] Chu, W. C.(2003). Speech Coding Algorithm. John Wiley & Sons, 2003.

[2] Vaseghi, Saeed V.. (2000) Advanced Signal Processing and Noise Reduction. John Wiley & Sons, 2000.

[3] Heigold, G. Zweig, G. Li, X. Nguyen, P. (2009). A flat direct model for speech recognition, *In: Proceedings* ICASSP, 2009.

[4] Zweig, G. Nguyen, P. (2009). Maximum mutual information multiphone units in direct modeling, *In*: *Proceedings Interspeech*, 2009.

[5] Zweig, G. (2003). Bayesian network structures and inference techniques for automatic speech recognition, *Computer Speech and Language*, 2003.

[6] Veth, J., Mauuary, L., Noe, B., Wet, F., Sienel, J., Boves, L., Jouvet, D. (2001). Feature vector selection to improve ASR robustness in noisy conditions. Eurospeech'01.

[7] Massachusetts Institute of Technology, MIT Open Course Ware. https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-011-introduction-to-communication-control-and-signal-processing-spring-2010/readings/MIT6_011S10_chap11.pdf

[8] Husn- Hsien Chang Jose, M. F. (2010). Moura Biomedical Signal Processing. Myer Kutz, in Biomedical Engineering and Design Handbook, 2nd Hill. 2010.

[9] Implementation of an Acoustic Echo Canceller Using Matlab by Srinivasaprasath Raghavendran College of Engineering University of South Florida October 15, 2003.

[10] Bai, L., Yin, Q. (2010). A modified NLMS algorithm for adaptive noise cancellation, *In*: *IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, ICASSP, , p.3726-3729, 14-19 March 2010

[11] Hymavathy, K. P., Janardhanan. (2013). Noise Filtering in Speech Using Frequency Response Masking Technique. *International Journal of Emerging Trends in Engineering and Development*, 2. P (2013)

[12] Aggarwal, R., Singh, J. K. Gupta, V. K., Rathore, S., Tiwari, M., Khare, A. (2011). Noise Reduction of Speech Signal Using Wavelet Transform with Modified Universal Threshold. *International Journal of Computer Applications*, 20, 15-19

[13] Verteletskaya, E.,Simak, B. (2010) Speech Distortion Minimized Noise Reduction Algorithm. *In: Proceedings of the World Congress on Engineering and Computer Science*, Volume 1, San Francisco, 20-22 October 2010.

[14] Steinbuch, M., van de Molengraft, M. J. G. (2005). Eindhoven University of Technology, Control Systems Technology Group Eindhoven, *Wavelet Theory and Applications, a literature study*, R. J. E. Merry, DCT 2005.53

[15] Jingdong Chen., Jacob Benesty., Yiteng (Arden) Huang., Simon Doclo. (2006). New Insights Into the Noise Reduction Wiener Filter, *Ieee Transactions on Audio, Speech, And Language Processing,* 14 (4), JULY 2006.

[16] Chandra Sekhar Yadav, G. V. P. Ananda Krishna, (2014). Study of different adaptive filter algorithms for noise cancellation in real time environment, *International Journal of Computer Applications* (0975-887), 96 (10), June 2014.

[17] Jasmeet Singh, Adaptive Noise Cancellation in Sinusoidal Signal using Wiener Filter, Thesis report, Thapar University, Patiala-147004, 15-7-2010.

Author Biographies

Mohamad Al-Sadi received his B.Sc. degree Information System Engineering from Syrian Virtual University (SVU), at the Department of Artificial Intelligence, Syria, in 2015, and Technical Diploma degree in Computer Science from The Technical Computer college (TCC), Syria, in 2010.Eng. AL-Sadi is currently working in the design and development of embedded systems and artificial intelligence. He can be reached at ise34857@gmail.com