ISSN: 2413-6999
Journal of Information, Communication, and Intelligence Systems (JICIS)
Volume 4, Issue 5, December 2018

Frame Blocking and Windowing Speech Signal

Oday Kamil Hamid

Abstract— The key objective of this research is frame blocking and windowing, a speech signal is a slowly time varying signal in the sense that, when examined over a short period of time (between 10 to 30 ms), its characteristics are short time stationary. This is not the case if we look at a speech signal under a longer time perspective (approximately time T › 0.5 s).in this case the signals characteristics are non-stationary, meaning that it changes to reflect the different sounds spoken by the talker. For this reason we use frame blocking and windowing to be able to use a speech signal and interpret its characteristics in proper manner. In this project speech signal is blocked into frames of N sample with adjacent frames being separated by M (M ‹ N) where N=256 sample correspond to (≈23 ms) and M(overlapping)=50% (128 sample)(11.37 ms)and signal is sampled at 11.25 ms, and then we use hamming window because it is the most widely used in speech processing.

The proposed speaker recognition systems are examined through theoretical analysis and computer simulation using Matlab version 6 programming language and sound forge 5 as a speech analyzer under Microsoft Windows 2007 operating system

I. Introduction

Speech recognition is a topic that is very useful in many application and environment in our daily life. generally speech recognizer is a machine which understand human and their spoken word in some way and can act thereafter it can be used, for example in a car environment to voice control non critical operations, such as dialing a phone number another possible scenario is on – board havigation, presenting the driving route to the driver applying voice control the traffic safety will be increased.

A different aspect of speech recognition is to facilitate for people with functional disability or other kinds of handicap to make their daily chores easier, voice control could be helpful .with their voice they could operate the light switch turn of/on the coffee machine or operate some other domestic appliances this leads to the discussion about intelligent homes where these operation can be made available for the common man as well as for handicapped [1].

Oday K.Hamid, Dept. of Computer Techniques Engineering, Dijlah University College, *(e-mail: oday.kamil@duc.edu.iq). Baghdad, Iraq.*

With this information presented so far one question comes naturally: now is speech recognition done? to get knowledge of how speech recognition problems can be approached today, a review of some research high lights will be presented the earliest attempt to device systems for automatic speech recognition by machine were made in the 1950's ,when various researchers tried to exploit the fundamental idea of acoustic –phonetics in 1952, at bell laboratories , davis biddulph ,and balashek built a system for isolated digit recognition for a single speaker the system relied heavily on measuring spectral resonances during the vowel region of each digit . in 1959 another attempt was made by forgie and forgie , constructed at MIT Lincoln laboratories ten vowel embedded in a/b/-vowel-/t format were recognized in speaker independent manner .in the 1970's speech recognition research achieved a number of significant mile stones ,first the area of isolated word or discrete utterance recognition became a viable and usable technology based on the fundamental studies by velichko andzagoruyko in Russia ,sakoe and chiba in japan and itakura, in united state .the Russian studies helped advance the use of pattern recognition ideas in speech recognition ,the Japanese research showed how dynamic program ming methods could be successfully applied and itakura's research showed now the idea of linear predicting coding (LPC).

The purpose with this research is getting a deeper theoretical and practical understanding of speech recognition .the work started by cutting the speech data signal into frames before analysis and the frame size is 10---30 ms and frames can be overlapped normally the over lapping region range from 0 to 50% of the frame size and then use the matlab to process the speech signal .in the future it could be possible to use this information to create chip that could be used as anew interface to humans .for example it would be desired to get rid of all remote controls in the home and just tell the TV,stereo or any desired device what to do with the voice[2].

II. Theory

Framing

Decompose the speech signal into a series of overlapping frames

– Traditional methods for spectral evaluation are reliable in the case of a stationary signal (i.e., a signal whose statistical characteristics are invariant with respect to time)

- Imply that the region is short enough for the behavior of (periodicity or noise-like appearance) the signal to be approximately constant
- In sense, the speech region has to be short enough so that it can reasonably be assumed to be stationary
- stationary in that region: i.e., the signal characteristics whether periodicity or noise-like appearance) are uniform in that region. Frame duration ranges are between 10 ~ 25 ms in the case of speech processing [3].

Frame blocking and Windowing

Due to the differences in phoneme's spectral features, changes in prosody, and random variations in the vocal tract, speech is a non-stationary signal. However, in a short time interval (generally from 10 to 20 ms) it is assumed that the speech signal is stationary, and therefore it is analyzed over these shot-time windows. So the frame blocking procedure consists essentially dividing the speech signal into short frames of *N* samples, which overlap by *M* samples, with adjacent frames.

In order to minimize spectral distortions when blocking the speech signal, each frame is multiplied with a Hamming window of the form

$$w(n) = 0.54 - 0.64\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

where *N* is the duration (in samples) of the speech frame. The output

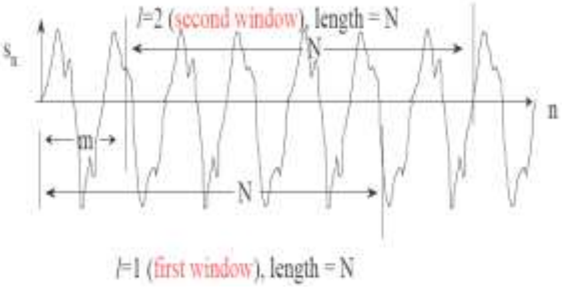
y (n) of the windowed signal becomes:

$$y(n) = x(n)w(n)$$

❧This windowing function acts as a low pass filter, enhancing the signal at the window center and smoothening it at the edges.

❧ To choose the frame size (N samples) and adjacent frames separated by m samples.

❧ i.e... A 11.5 KHz sampling signal, a 8ms window has N=256samples, (neighboring shift) m=128 sample[4].


l=2 (second window), length = N
N
s_n
n
m
N
l=1 (first window), length = N

Time frame and overlap

❧Since our ear cannot response to very fast change of speech data content, we normally cut the speech data into frames before analysis

❧Frame size is 10~30ms

❧Frames can be overlapped:

Normally the overlapping region ranges from 0 to 75% of the frame size.

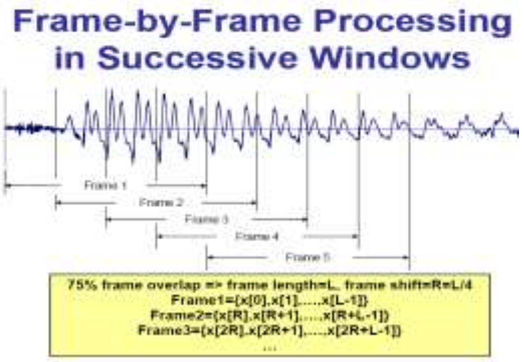

Frame-by-Frame Processing
in Successive Windows
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
75% frame overlap => frame length=L, frame shift=R=L/4
Frame1={x[0],x[1],....,x[L-1]}
Frame2={x[R],x[R+1],....,x[R+L-1]}
Frame3={x[2R],x[2R+1],....,x[2R+L-1]}
...

Figure (1) (sampled speech signal)

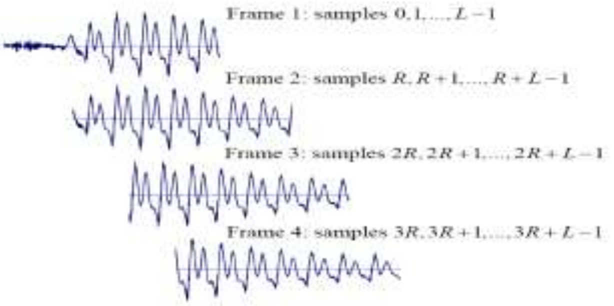
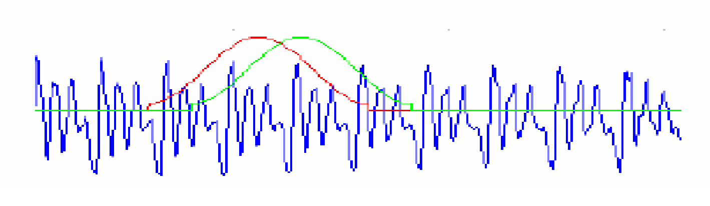

Frame 1: samples 0, 1, ..., L − 1
Frame 2: samples R, R + 1, ..., R + L − 1
Frame 3: samples 2R, 2R + 1, ..., 2R + L − 1
Frame 4: samples 3R, 3R + 1, ..., 3R + L − 1

Figure (2) (frame length of 256 samples and overlap of 128 samples)[5].

Frame shifting

It is normal to use overlapping windows to ensure better temporal continuity in the transform domain. An overlap of half the window size (or less) is typical.



- Frame rate: the number of frames computed per second, in general 33 to 100 frames per second in sort-term speech processing [6].

Framing and windowing–short-term processing

A frame-based analysis is essential for speech signals as shown in figure (3).

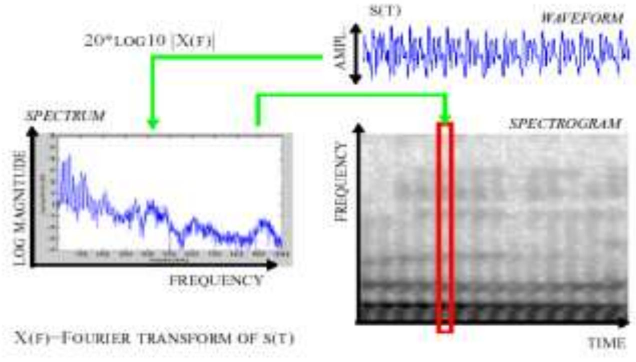
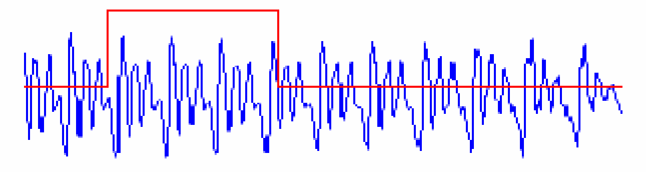


Figure (3) (frame analysis)

Windowing and the types of window

Since speech is non-stationary, we are interesting in short-term estimates of parameters such as the Fourier spectrum. This requires that a speech segment be chosen for analysis. We are effectively cross-multiplying the signal by a window function.



$$w(n) = \begin{cases} 1 & 0 \le n \le N \\ 0 & Otherwise \end{cases}$$

- Just extract the frame part of signal without further processing
- Whose frequency response has high side lobes.

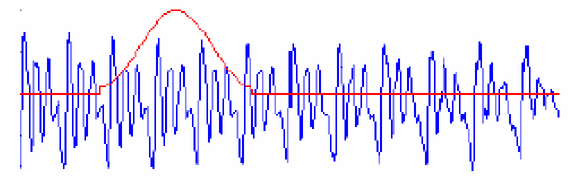
–Main lobe: spreads out in a wider frequency range the narrow band power of the signal, and thus reduces the local frequency resolution

–Side lobe: swaps energy from different and distant frequencies of xm[n], which is called leakage

However, it is desirable to use a tapered window such as:

Hamming window

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & \text{Otherwise} \end{cases}$$



Function of window

– rectangular window:

- *h*[*n*]=1, *0≤n≤L-1* and *0* otherwise

– Hamming window (raised cosine window):

- *h*[*n*]*=0.54-0.46 cos(2πn/(L-1)), 0≤n≤L-1* and *0* otherwise

– rectangular window gives *equal weight* to all *L* samples in the window (*n,...,n-L+1*)

– Hamming window gives *most weight* to middle samples and *tapers off* strongly at the beginning and the end of the window [7].

Windows in STFT

For Xn(e^{jw})to represent the short-time spectral properties of *X(n)* inside the window, Xn(e^{jw}) should be much narrower in frequency than significant spectral regions of Xn(e^{jw}) i.e., almost an impulse in frequency. Consider rectangular and hamming windows, where width of the main spectral lobe is inversely proportional to window length and side lobe levels are essentially independent of window length.

- Rectangular Window: flat window of length N samples; first zero in

frequency response occurs at FS/N, with side lobe levels of -14 dB or lower.

- Hamming Window: raised cosine window of length L samples; first zero in frequency response occurs at 2FS/N, with side lobe levels of -40 dB or lower as shown in figure (4) below:

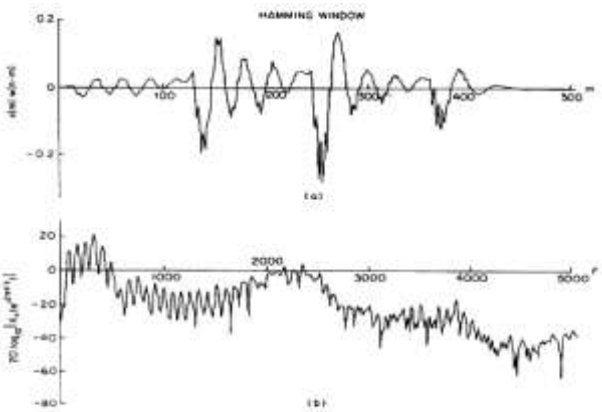


Figure (4) Frequency response

- 500sample windows (50 msec)
- can see periodicity in time and in frequency
- can see strong first formant (300-400 Hz), strong resonance at 2200 Hz, resonance at 3800 Hz as shown in figure(5)[8].:

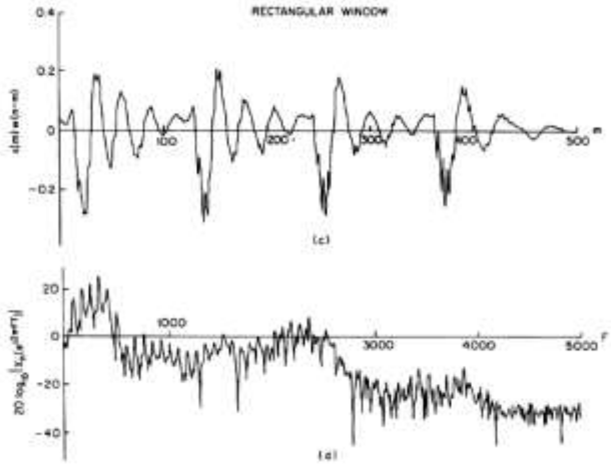


Figure (5) Frequency response

- to determine the sampling rate in time, we take a linear filtering view

1. $X_n(e^{j\omega})$ is the output of a filter with impulse response $w(n)$
2. $W(e^{j\omega})$ is a lowpass response with effective bandwidth of B Hertz

- thus the effective bandwidth of $X_n(e^{j\omega})$ is B Hertz => $X_n(e^{j\omega})$ has to be sampled at a rate of $2B$ samples/second to avoid aliasing

Example: Hamming Window

$$w(n) = 0.54 - 0.46\cos(2\pi n/(L-1)) \quad 0 \le n \le L-1$$
$$= 0 \quad \text{otherwise}$$

$\Rightarrow B \approx \frac{2F_s}{L}$(Hz); for $L = 100$, $F_s = 10,000$ Hz => $B = 200$ Hz => need rate of 400/sec (every 25 samples) for sampling rate in time

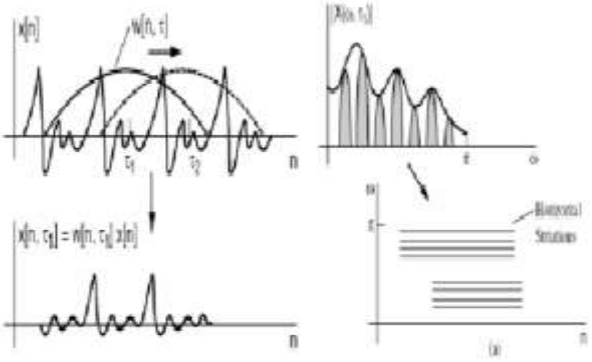


Figure (6) sampling theorem

Sampling rate in frequency

- since $X_n(e^{j\omega})$ is periodic in ω with period 2π, it is only necessary to sample over an interval of length 2π
- need to determine an appropriate finite set of frequencies, $\omega_k = 2\pi k/N$, $k = 0,1,\ldots,N-1$ at which $X_n(e^{j\omega})$ must be specified to exactly recover $x(n)$
- use the Fourier transform interpretation of $X_n(e^{j\omega})$

1. if the window $w(n)$ is time-limited, then the inverse transform of $X_n(e^{j\omega})$ is time-limited
2. the sampling theorem requres that we sample $X_n(e^{j\omega})$ in the frequency dimension at a rate of at least twice its ('symmetric') "time width"
3. since the inverse Fourier transform of $X_n(e^{j\omega})$ is the signal $x(m)w(n-m)$ and this signal is of duration L samples (the duration of $w(n)$), then according to the sampling theorem $X_n(e^{j\omega})$ must be sampled (in frequency) at the set of frequencies

$$\omega_k = \frac{2\pi k}{L}, \quad k = 0,1,\ldots,L-1 \quad \text{(where } L/2 \text{ is the effective width of the window)}$$

in order to exactly recover $x(n)$ from $X_n(e^{j\omega_k})$

Total" Sampling Rate of STFT

• the "total" sampling rate for the STFT is the product of the sampling rates in time and frequency, i.e.,

SR = SR(time) x SR(frequency)

= 2B x L samples/sec

B = frequency bandwidth of window (Hz)

L = time width of window (samples)

• for most windows of interest, B is a multiple of FS/L, i.e.,

B = C FS/L (Hz), C=1 for Rectangular Window

C=2 for Hamming Window

SR = 2C FS samples/second

• can define an 'over sampling rate' of

SR/ FS = 2C = over sampling rate of STFT as compared to conventional sampling representation of x(n)

for RW, 2C=2; for HW 2C=4 => range of over sampling is 2-4

this over sampling gives a very flexible representation of the speech signal

Short-term Energy

The long term definition of signal energy is as below:

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) = x^2(n-N+1) + \ldots + x^2(n)$$

There is little or no utility of this definition for time-varying signals, say speech.

For a short-term speech signal (the n-th frame speech after framing and windowing:

Xn(m)=x(m)w(n-m)   n-N+1<=m<=n

Where w (n) is the window, n is the sample that the analysis window is centered on, and N is the window size. Window jumps/slides across sequence of squared values, selecting interval for processing, as shown below in figure (7) [9].:

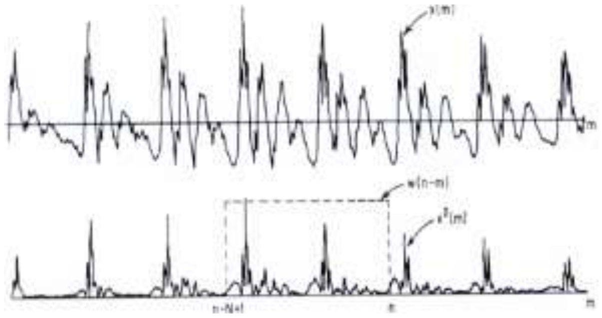


Figure (7) Original sample and windowed sample

III. Database

Any speech or speaker recognition system depend on the type of data input to this system, so some elements must be available in order to get on a good data as:

- High quality microphones used in recording for both training and testing sessions.
- Ideal recording must be used in rooms with little or no background noise or reverberation for both training and testing sessions.
- Collect a large database for many tries.
- Using modern programs in recording because it has a capability for cutting voices or synthesis and it gives a good representation of signal's shape.

database had been recorded by using high quality microphone to record voice(الخير) this word are recorded by using Sound Forge program figure (3.1) shows recorded signal displayed by this programs screen.

Data is sampled at 11.25 KHZ (sampling rate), with 16-bit sample value A/D, but unfortunately we did not have the chance of recording in suitable place for such a purpose and that deal of the collected database from each speaker considered as a little for suggested systems.

Preprocessing speech signal

The basic idea of the preprocessing is not to use the high dimensional redundant speech signal for the recognition procedure, but to describe it by a low dimensional set of features being typical mainly for the speaker's identity. The speech signal coming from microphone pass through these steps:

1.Reading the database (convert it from analog to digital signal $(x(n))$.

Frame blocking

Since speech is time-varying in that the vocal-tract configuration changes over time, an accurate set of predictor coefficients is adaptively determined over short time frames (typically 10ms to 30ms) during which time-invariance is assumed [31]. So for this reason the continuous speech signal is blocked into frames of N sample with adjacent frames being separated by $M(M<N)$. In this research $N=256$ sample correspond to (~23 msec) and *M* (over lapping) =50% (128 sample) (11.37 msec). The first frame consists of the first *N* samples; the second frame being *M* samples after the first frame, and overlaps it by *N-M* sample. Similarly, the third frame beings *2M* samples after the first frame (or *M* samples after the second frame) and overlaps it by *N-2M* samples. This process continues until all the speech is accounted for within one or more frames. We denote the l^{th} frame of speech by $x_l(n)$, and there are *L* frames within the speech signal, then $x_l(n)=x(Ml+n)$ n=0,1,…….., *N-1*   *l*=0,1,…..*L-1* that is, the first frame of speech, $x_0(n)$ encompasses speech samples $x(0),x(1),.......x(N-1)$,the second frame of speech $x_1(n)$,encompasses samples $x(M),x(M+1),.......x(M+N-1)$,and the L^{th} frame of Speech, $x_{L-1}(n)$,Encompasses speech samples

$x(M(L-1)),x(M(L-1)+1),............x(M(L-1)+N-1)$ **as shown in figure (8)**

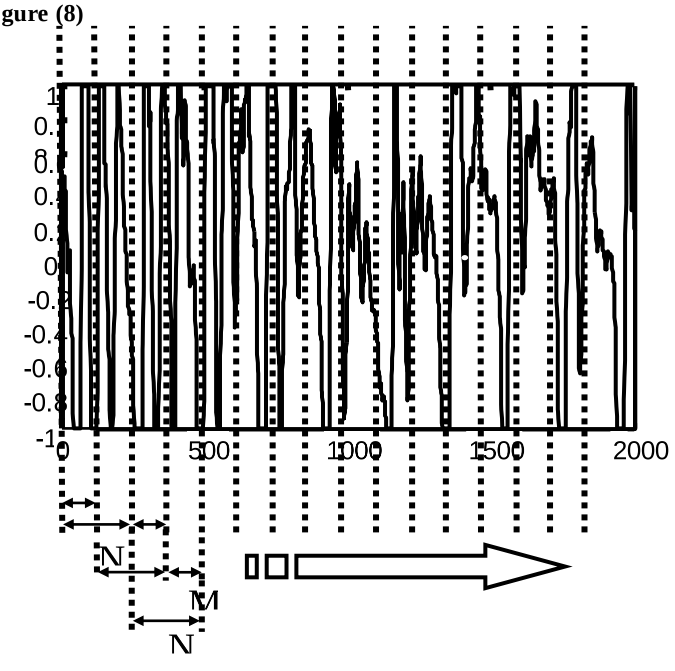


Figure (8): Blocking of speech into overlapping frames.

Windowing

The next step in the processing is to window each individual frame, the most widely used windows in speech processing are the rectangular window that weights all samples in the analysis frame equally, and Hamming window which is used to taper the segment because prediction residual must be kept to minimize at the beginning and end of the frame.

There is an important difference between rectangular window and hamming window that the bandwidth of hamming window is about twice the bandwidth of a rectangular window of the same length as shown in It is also clear that the hamming window gives much greater attenuation outside the pass band than the comparable rectangular window

–

❖ Hamming Window

Which it was used in this research, figure (9) and it's defined as:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N - \\ 0 & Otherwise \end{cases}$$

n=Length of window.

N=Number of sample [10].

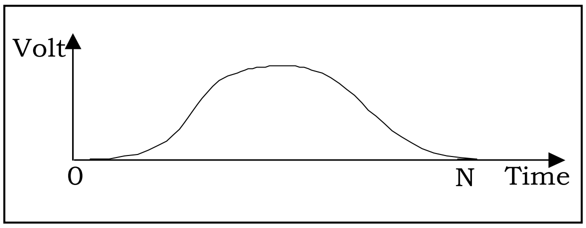


Figure (9): Hamming window

Windows are used in signal analysis so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame [6].

In this research a typical hamming window was used as in equation (3-2):

$$\tilde{x}_1(n) = x_1(n)w(n) \qquad 0 \le n \le N-1$$

N=256.

IV. Evaluation test for the proposed method

1- convert the analog signal to digital signal X(n) as shown in figure (10) below:

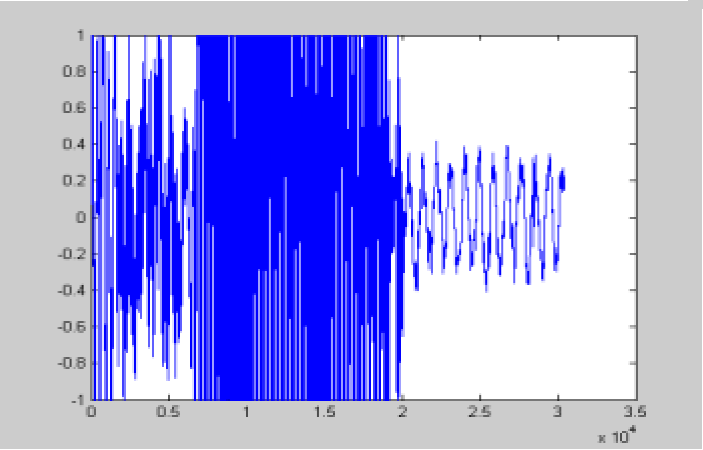


Figure (10) sampled speech signal

2- Then the original sampled speech signal was cut into frames each frame have 256 sample as discussed before, figure(11) shows frame number 25.

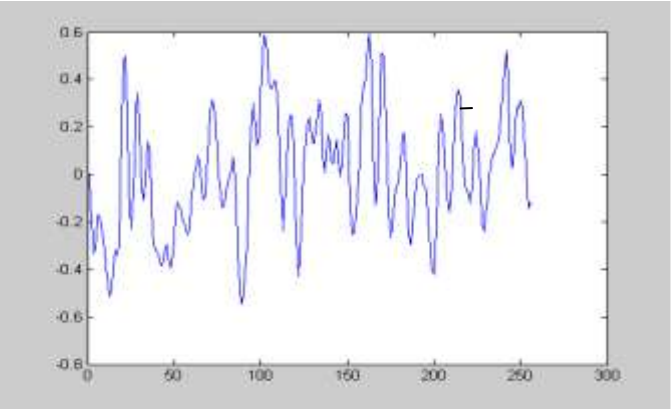


Figure (11) frame number 25

3- Applied hamming window for each frame as shown in figure (12).

Figure (13) shows both frame and windowed sampled and observe how the hamming window taper the beginning and end of the frame.

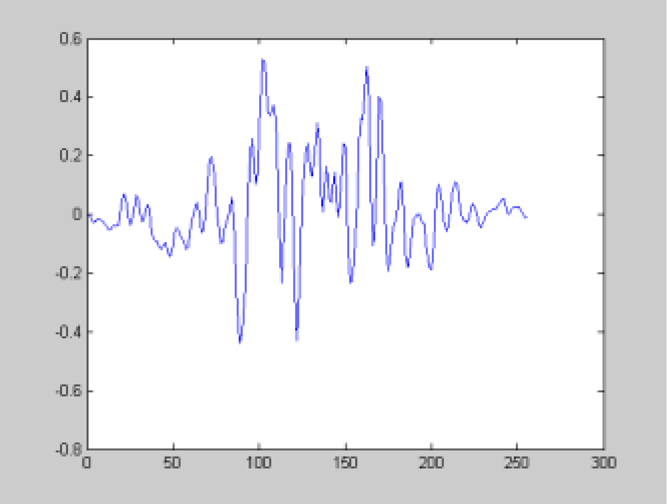


Figure (12) windowed frame

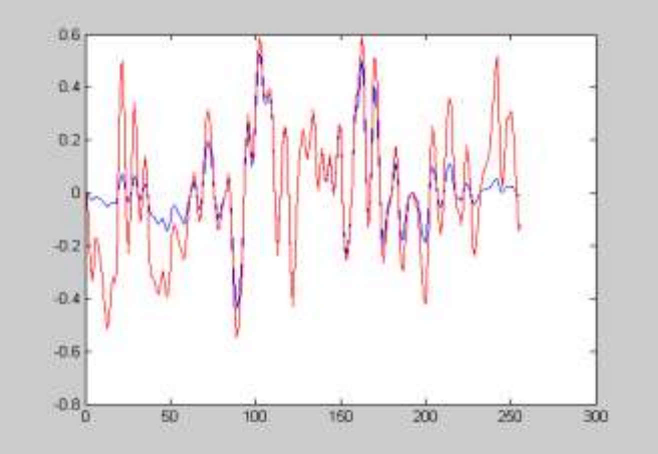


Figure (13) frame and windowed frame

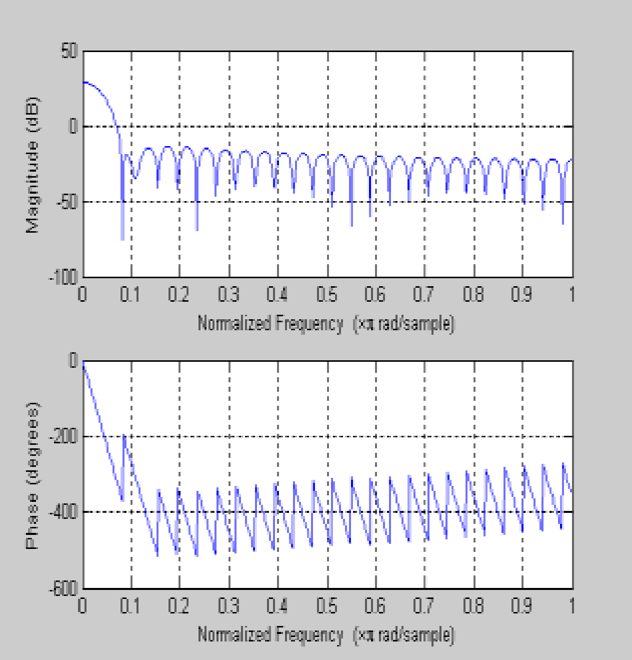


Figure (14): frequency response with N=51, for hamming window

The using hamming window instead of rectangular window is because this has to do with the assumption of periodicity made by the DFT, and become clearer in the frequency domain.

A window (on its own) tends to have an averaging effect. Thus it has a low pass spectral characteristic. Ideally, we want

• To preserve spectral detail

• To produce little spectral distortion The log magnitude spectrum of a rectangular window can be compared with that of a Hamming window:

• The Hamming has a wider main lobe, but much better attenuation of side lobes (typically 20-30 dB better than rectangular). For a designed window, wish that:

- A narrow bandwidth main lobe

- Large attenuation in the magnitudes of the side lobes

However, this is a trade-off!

Notice that:

1. A narrow main lobe will resolve the sharp details of speech signal (the frequency response of the framed signal) as the convolution proceeds in frequency domain

2. The attenuated side lobes prevents noise from other parts of the spectrum from corrupting the true spectrum at a given frequency

3. Band width of hamming window is about twice the band width of rectangular window of the same length as shown in figures (14) and (15).

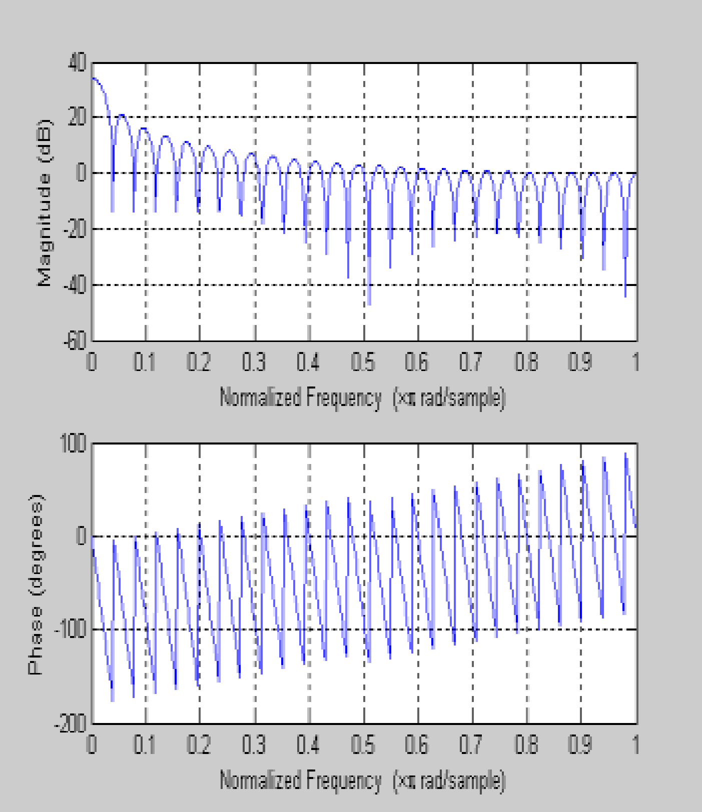


Figure (15): frequency response with N=51, for rectangular window

REFERENCES

1. J. P. Hosom, R. Cole and M. Fanty, *"Speech Recognition Using Neural Networks"*, Center for Spoken Language Understanding (*cslu)*
Oregon Graduate Institute of Science and Technology, July 6, 1999, "http://cslu.cse.ogi.edu/corpora/available/".
2. B. S. Atal, *"Automatic Recognition of Speakers From Their Voices"*, IEEE, Vol. 64, pp. 460-475, April1975.
3. B. S. Atal, *"Automatic Speaker Recognition Based Upon Pitch Contours"*, Journal of acoustic of America society (JAAS), Vol. 52, pp. 1687-1697, 1972.
4. L. R. Rabinar and B. H. Juang, *"Fundamentals of Speech Recognition"*, Prentice-Hell, New Jersey, 1993.
5. M. N. AL-Trfi, *"Speaker Recognition Based Upon Phonemes Using Wavelet Packet Transform"*, M.Sc. Thesis, College of Engineering, University of Baghdad, 2000.
6. N. Do. Minh, *"An Automatic Speaker Recognition System"*, Swiss Federal Institute of Technology, lausanne-Epel, "http://lcavwww. Epfl. Ch./~mindho/asr_project/. Html", January 2000.
7. R. D. Rodman, *"Speaker Recognition of Disguised Voices"*, "www.csc.Ncsu.edu/factly/rodman/", Speaker Recognition, Disguised Voices, 1997.and references there in
8. A. H. Al-Nakkash, *"A Novel Approach For Speakers Recognition Using Vector Quantization Technique"*, M.Sc. Thesis, University of Technology, Baghdad, 2001.
9. H. Fenglie and W. Bingxi, *"An Integrated System for Text-Independent Speaker Recognition Using Binary Neural Network Classifiers"*, Proceeding of ICASSP 2001.
10. K. G. Margaritis, *"Development of a Text-Dependent Speaker Identification System with the OGI Toolkit"*, 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece,
Proceedings, Companion Volume, pp. 525-530.