ISOLATED WORD SPEECH RECOGNITION SYSTEM

Enrollment. No. - 071075

Name of Student - HARENDER KANWAR

Name of supervisor(s) - PROF. T.S LAMBA

May – 2011

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TALBLE OF CONTENTS

CERTIFICATE

This is to certify that the work entitled "**Isolated Word Speech Recognition System**" submitted by "**Harender Kanwar (071075)**" in partial fulfillment for the award of degree of B. Tech, of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor ……..

Name of Supervisor ……..

Designation ……..

Date ……..

ACKNOWLEDGEMENT

The final year's major project provide us an outlook towards the practical things in the world. I came to know about the work done in the speech recognition, work going in the field.

I express my heartfelt thanks to Prof. T.S LAMBA, who through his expert guidance, constructive criticism and valuable suggestions helped me throughout the course of this project. If it were not his motivation and encouragement, we would not have seen through this project in an honest course to the splendor of success.

Finally, we wish to convey our gratitude to all the faculty members of the Electronics department of our college for providing necessary help and encouragement in the course of completion of this project .

Signature of the student ……

Name of Student ……..

Date ……..

SUMMARY

Speech recognition has been an integral part of human life acting as one of the five senses of Human body, because of which application developed on the basis of speech recognition, has high degree of acceptance [1]. The various steps that will be followed and that will be dealt in speech recognition are **feature extraction**, **distance calculation**, **dynamic time wrapping**. Such an approach has been used which is both simple and efficient. After analyzing the steps above we will realize the process using small programs using **MATLAB** which is able to do isolated digits (0-9) recognition.

A Graphical User Interface (GUI) system has been built in MATLAB.

The speech signal is a slowly timed varying signal . When examined over a sufficiently short period of time (between 5 and 100 m sec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, *short-time spectral analysis* is the most common way to characterize the speech signal.

There are two main operations in a speech recognition system. First is the registration part in which we record a speech signal and process the features of speech signal and record in the codebook. Features from speech signal can be extracted using any of the techniques available i.e. linear predictive coding or mel-frequency cepstral coefficients and many more. Second is the verification

part in which the speech signal is recorded and its features are processed and those feature vectors are then compared with that of the feature vectors record in the codebook. This comparison is done with the help of Dynamic time warping technique (DTW). A graphical user interface has been made which will be calculating the distance between the recorded and stored speech signal feature vectors and the one with minimum distance will be declared the recognized word.

Signature of the student ……

Name of Student ……..

Date ……..

List of Figures

CHAPTER 1

Introduction

1.1 Production of Speech

While you are producing speech sounds, the air flow from your lungs first passes the glottis and then your throat and mouth or nasal cavity. The length of vocal tract is roughly 17.5 cms. Various body parts which are responsible for the speech production have been discussed below:-

Lips - supported by maxilla (upper jaw) and lower jaw (mandible)

– Speech function - varied movement, rounded, tensed, obstruct air flow.

Teeth

– speech function - anatomical obstacle for lips or Tongue.

Alveolar ridge (gum ridge of maxilla

– speech function - point of contact/constriction

Speech Mechanism hard palate - bony structure posterior to alveolar ridge

– speech function - point of contact; defines shape

of oral cavity

Soft palate/velum

– speech function - It separates the oral cavity (mouth) from the nasal cavity in order to produce the oral speech sounds.

Tongue

– speech function - direction of air flow: contacts other structures.

-changes size of oral cavity

Mandible (lower jaw)

– speech function - change size of oral cavity.

Oral cavity - from mouth opening to posterior wall of pharynx (posterior pharyngeal wall).

– speech function : channels airstream: contributes oral resonance

Nasal cavity - extends from nostrils (nares) to posterior speech Mechanism pharyngeal wall

– speech function : contributes nasal resonance

Pharynx - posterior portion of nasal cavity down through back of oral cavity to larynx

Vocal cords/folds - in lateral walls of larynx

– phonation It is the process by which the vocal cords produce certain sounds through quasi-periodic vibration.

Depending on which speech sound you articulate, the speech signal can be excited in three possible ways:

- **Voiced excitation**:- The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangle–shaped). All the vowels have voiced excitation.
- **Unvoiced excitation** :- The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.
- **Transient excitation** :- A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately.

With some speech sounds these three kinds of excitation occur in combination. The spectral shape of the speech signal is determined by the shape of the vocal tract (the pipe formed by your throat, tongue, teeth and lips). By changing the shape of the pipe , in addition opening and closing the air flow through your nose you change the spectral shape of the speech signal, thus articulating different speech sounds.

1.2 Technical Characteristics of the Speech Signal

- The bandwidth of the speech signal for speech recognition purposes will be considered 4 kHz and therefore will be having a sampling rate of 8khz and no. of bits per sample will be 16 bits per sample. Within a bandwidth of 4 kHz the speech signal contains all the information necessary to understand a human voice.
- The speech signal is periodic with a fundamental frequency between 80 Hz and 350 Hz. Using voiced excitation for the speech sound will result in a pulse train.
- There are peaks in the spectral distribution of energy , which are called the **formants**. Formants are the resonance peaks in a speech signal There is roughly a one formant for every 1 kHz of frequency. After passing the glottis, the vocal tract gives a characteristic spectral shape to the speech signal. If one simplifies the vocal tract to a straight pipe (the length is about 17cm). Depending on the shape of the vocal tract (the diameter of the pipe changes along the pipe), the frequency of the

formants (especially of the 1st and 2nd formant) change and therefore characterize the word being articulated.

- The envelope of the power spectrum of the signal shows a decrease with increasing frequency which can be 5 to 12 dB/octave. The spectrum for voiced segments has more energy at lower frequencies than higher frequencies. This is called spectral tilt. Spectral tilt is caused by the nature of the glottal pulse. Boosting high-frequency energy gives more info to acoustic Model Improves phone recognition performance.

1.3 Speech Signal representation

Speech signal can be represented in two ways:-

1. Amplitude vs. Time plot.
2. spectrograms.

1.3.1 Amplitude vs. Time plot

Fig 1.1 shows us the amplitude vs. time plot of the word ‘one’. The speech signal starts with noise and then after when one comes there is pulse train as voice sound i.e. vowel ‘o’ has been uttered. Then come ‘n’ which is unvoiced sound and that’s why it is much darker as compared to voiced sound. thereafter comes again a voice sound i.e. ‘e’ which again follows a pulse train. For a speech signal with a sampling rate of 8 KHz and bit resolution of 8 bits /sample have been used.

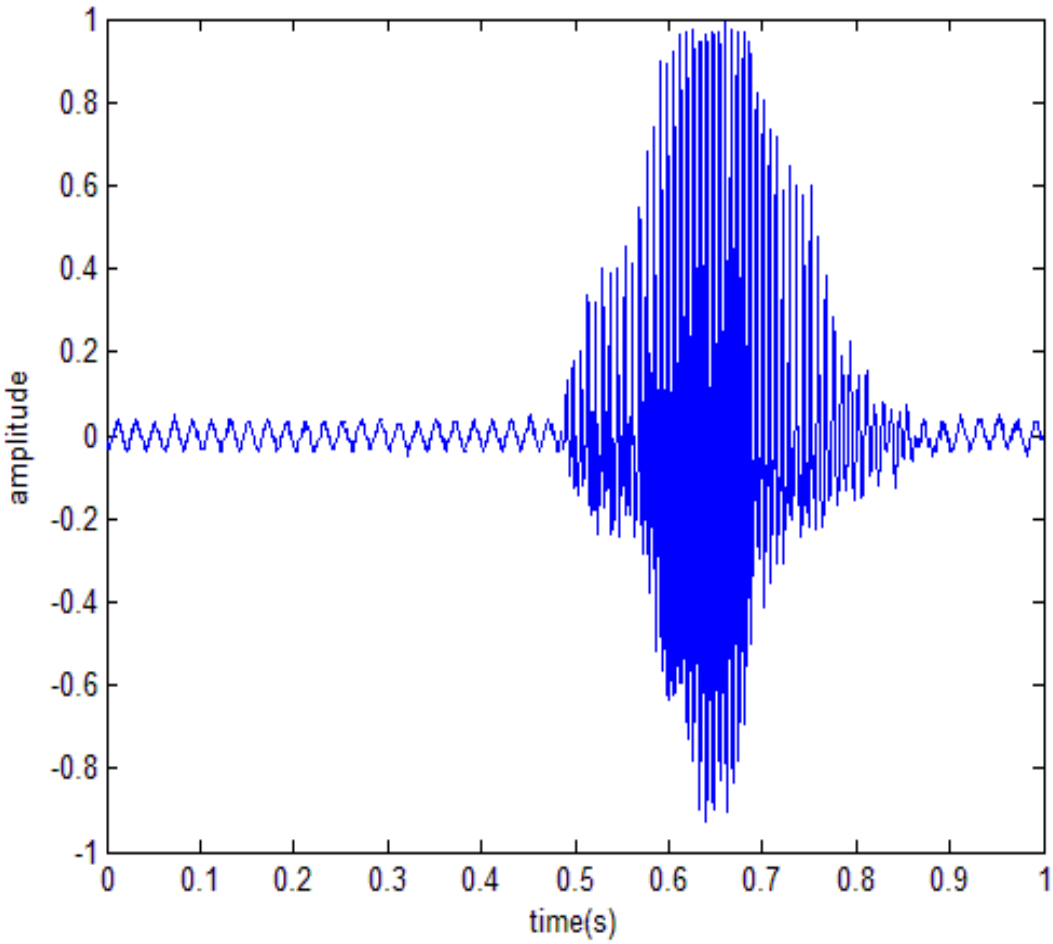
amplitude
time(s)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig 1.1 Amplitude vs. time plot of the word one .(fs = 8 khz and bit resolution = 8 bits/sample).

1.3.2 Spectrograms

A spectrogram is a three dimensional representation of the speech signal with time on x-axis and frequency on z axis and the colour intensity representing the magnitude of the short time Fourier transform. Speech signal is first windowed and thereafter for every windowed signal a short time Fourier transform will be calculated. We can not apply Fourier transform directly as speech is a slowly time varying signal and is periodic for only short time i.e. approximately 250~300 ms. The region with red colour is the region where the amplitude of the short time Fourier transform (**STFT**) is maximum and the blue color represents the region of very low or approximately zero magnitude of the **STFT**.

The resonance frequencies of the vocal tract show up as **"energy bands".**

Voiced intervals characterized by striated appearance .Un-Voiced intervals are more solidly filled in; i.e. the voiced sounds are lightly coloured and the unvoiced sounds are dark in colour. As can be seen from the fig 1.2 and 1.3 the utterance of word 'n ' is much darker in colour then the voiced sounds 'o' and 'e'.

There are two types of spectrograms, one is the narrowband spectrogram (fig1.2) and the other is wideband spectrogram(fig1.3).

NARROWBAND Spectrograms:- Narrowband spectrograms as shown in the fig1.2 has been implemented with the window size of 256 samples and **fft** of length 512.

WIDEBAND Spectrograms:- Wideband spectrograms as shown in the fig1. has been implemented with the window size of 128 samples and fft of length 256.

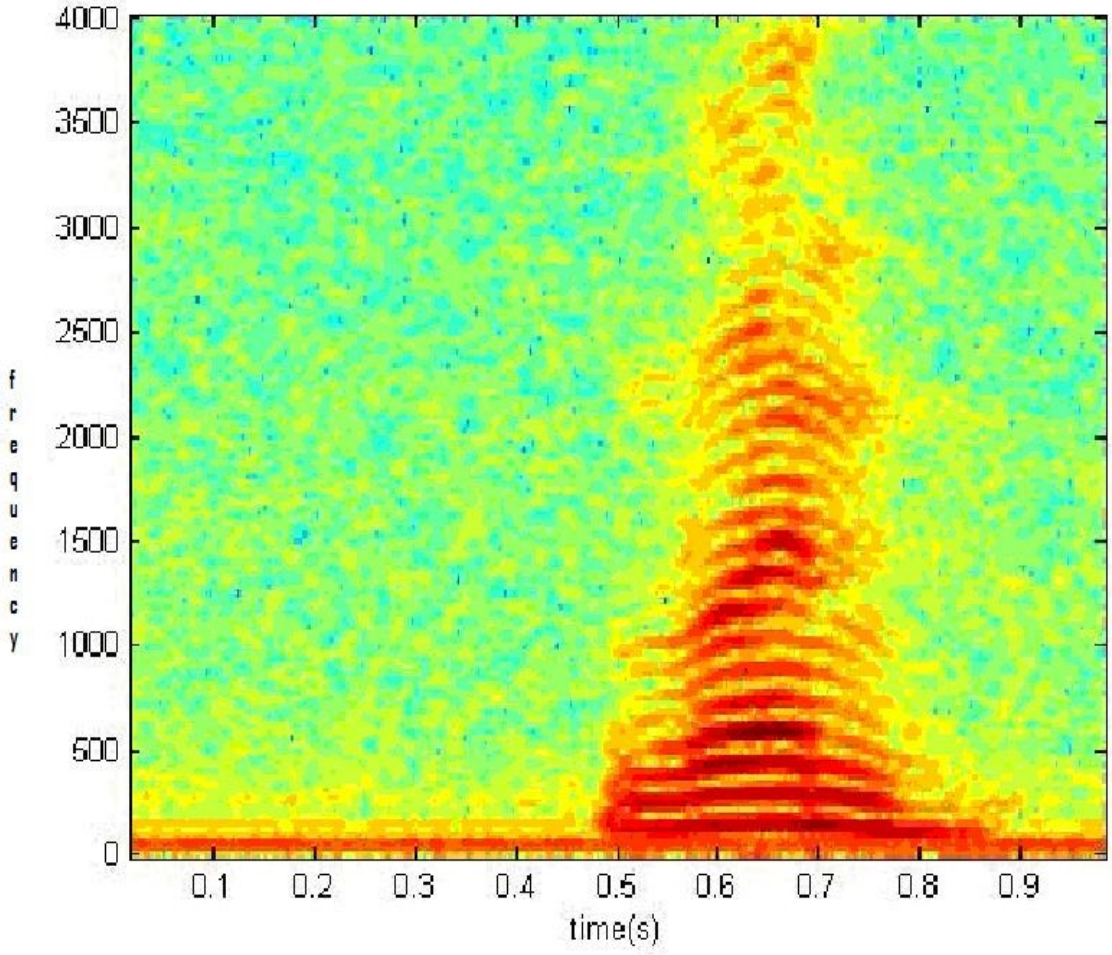

Fig1.2 Narowband representation of speech signal

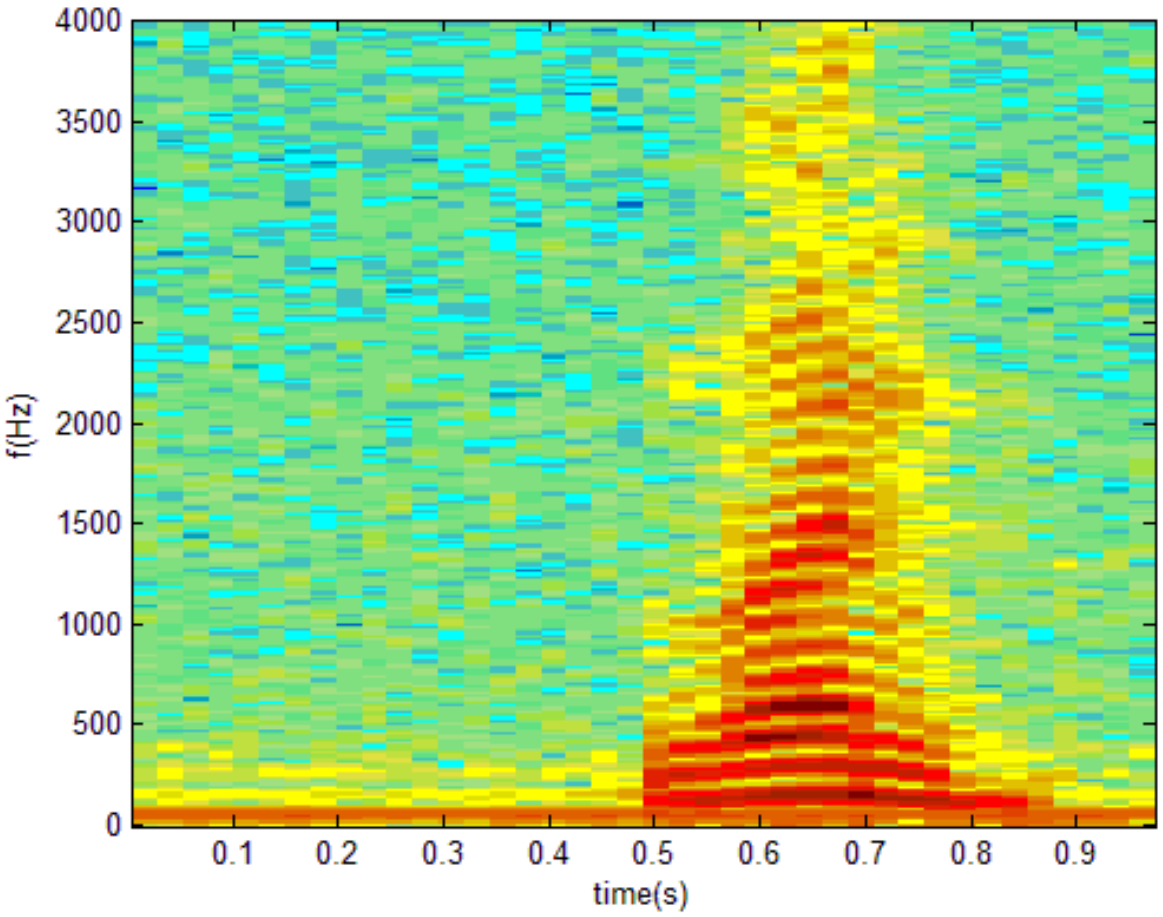

Fig 1.3 WIDEBAND spectrogram of speech signal.

1.4 Recording of Speech

Speech signal has been recorded with the windows sound recorder, the sound recorded had a sampling rate of 44.1 KHz and bit resolution of 16 bits per sample. Speech signal is said to have bandwidth in the range of 300-3000 Hz.

So a sampling frequency of 8 kHz is used and bit resolution is downsized to 8 bits /sample using **downsample.m.**

CHAPTER 2

Feature Extraction

One of the first decisions in any pattern recognition system is the choice of what features to use. In this part the details of extracting the features per frame of the speech signal have been discussed.

A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as **Linear Prediction Coding** (LPC), **Mel-Frequency Cepstrum Coefficients** (MFCC), and others. More recently, the majority of speech recognition systems have converged to the use of a cepstral vector derived from the filter bank that has been designed according to some model of auditory system. MFCC provides a representation corresponding to a smoothed short spectrum. MFCC has the reduced resolution at high frequencies that is indicative of auditory filter bank based methods. The resulting coefficients in MFCC are cepstral coefficients which are orthogonal. Both the methods have been studied in this project as a feature extraction method but because of the Mel Frequency Cepstral Coefficients (MFCC) is chosen because of the sensitivity of the low order cepstral coefficients to overall spectral slope and the sensitivity properties of the high-order cepstral coefficient [1]. Currently it is the most popular feature extraction method [1].

2.1 Linear predictive coding (LPC):-

The basic idea behind the LPC model is that a given speech sample at time *n ,s(n),* can be approximated as a linear combination of the past *p* speech

samples. The digitized speech sample s(n) is put through a low order digital system to spectrally flatten the signal . Framing of speech samples is done with **256** samples per frame and with an overlap of **128** samples per frame; after that a hamming window of has been applied to minimize the signal discontinuities at the beginning and end of each frame. Each frame of windowed signal is next auto correlated to give **LPC** coefficients for the order of **10**. The steps involved in calculating the LPC coefficients have been discussed in following sub-sections. Fig 2.1 shows the procedure to calculate the LPC coefficients and thereby converting the spectral coefficients to the cepstral coefficients later in the subsequent stage.

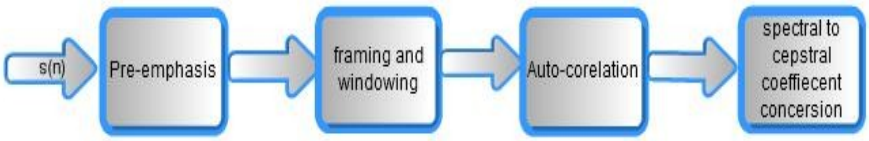

Fig 2.1 Flow process of LPC

2.1.1 Pre-emphasis

Speech has an overall tilt of 5 to 12 dB/octave, so a pre emphasis filter is

used. The transfer function of filter is given by H(z)

$$H(z) = 1 - 0.95z^{-1}$$

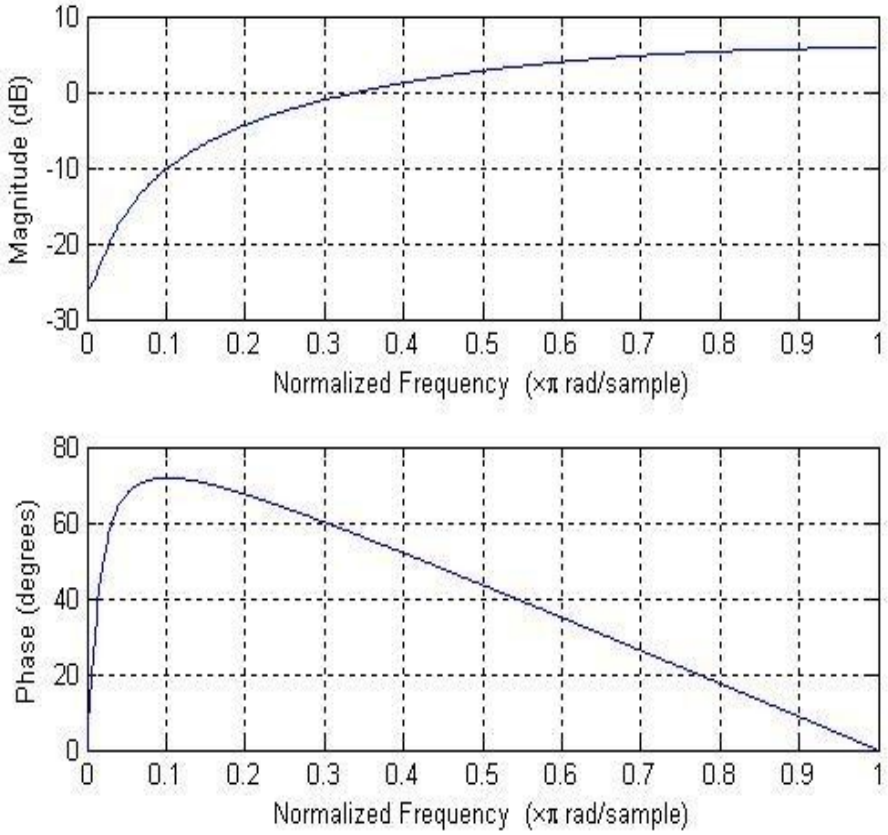

Fig.2.2 Magnitude and Phase response of the pre-emphasis filter.

Figure 2.2 shows the magnitude and phase response of the pre-emphasis filter. The speech signal has been passed through the pre-emphasis filter before the

feature extraction method has been applied. This step is common in both the LPC and MFCC feature extraction method.

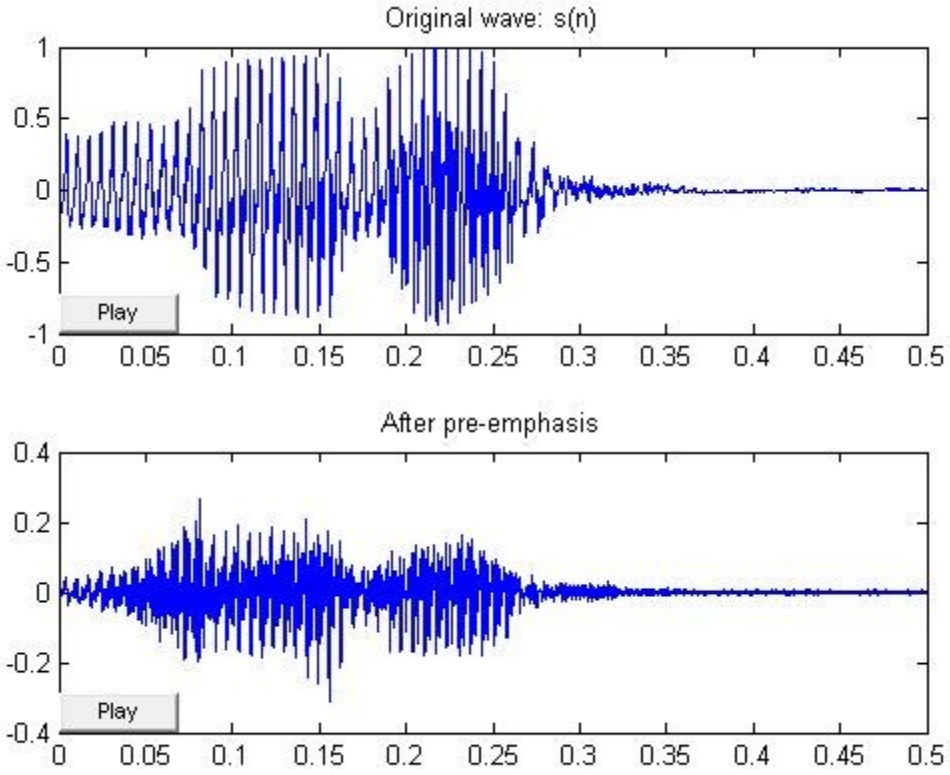

Fig 2.3 Speech signal before pre-emphasis and after emphasis in time domain.

Fig 2.3 shows us the speech signal before emphasis nd after emphasis. The speech after pre-emphasis sounds sharper with a smaller volume.

Fig 2.4.a shows us the FFT of the original speech signal and fig2.4.b shows us the FFT of the preemphasised speech signal. It is observed from these two figures that after applying preemphasis high frequency component of speech

signal have been preepmhasised in the range of 5-8 dB. The purpose of the pre emphasis filter is to spectrally flatten the signal.

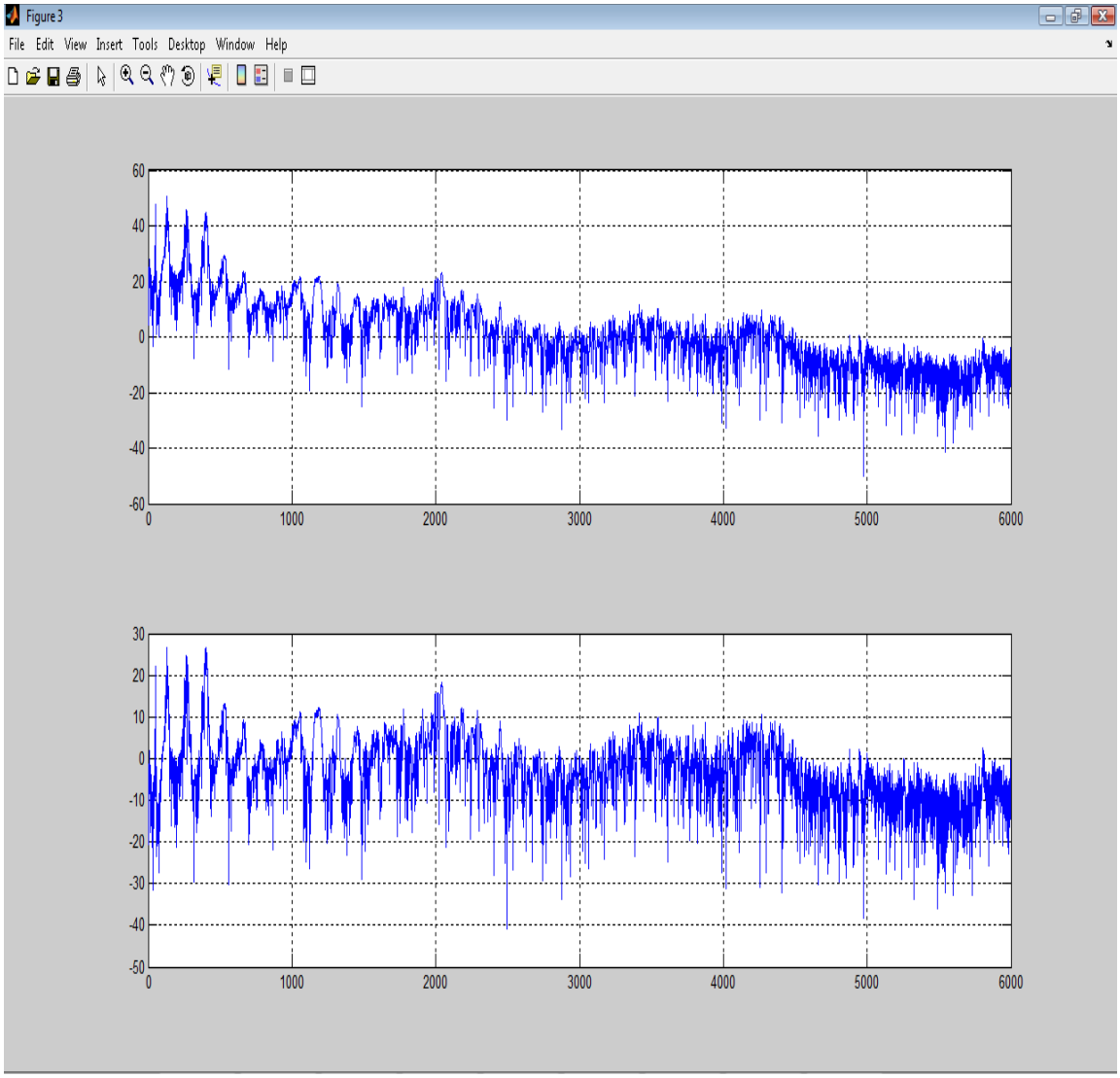

Fig 2.4 a) FFT of windowed speech signal.

b) FFT of preemphasized speech signal.

2.1.2 Framing

In this step the continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples. The second frame begins *M* samples after the first frame, and overlaps it by *N* - *M* samples and so on. This process continues until all the speech is accounted for within one or more frames. In this project framing has been done with **256** samples per frame and with an overlap of **128** samples per frame.

2.1.3 Windowing

Analysis of speech requires examination of small portions assumed to be pseudo stationary. Windowing yields a set of speech samples x(n) weighted by the shape of the window.

$$x(n) = s(n)w(n)$$

where,

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & n = 0,1,2.....N-1 \\ 0 & \text{otherwise.} \end{cases}$$

Generally, successive windows will overlap as *w*(n) tends to have a shape that will deemphasize samples near it's edges. This breaks the speech down into a sequence of Frames and these frames needs further processing which have been discussed in the subsequent subsections.

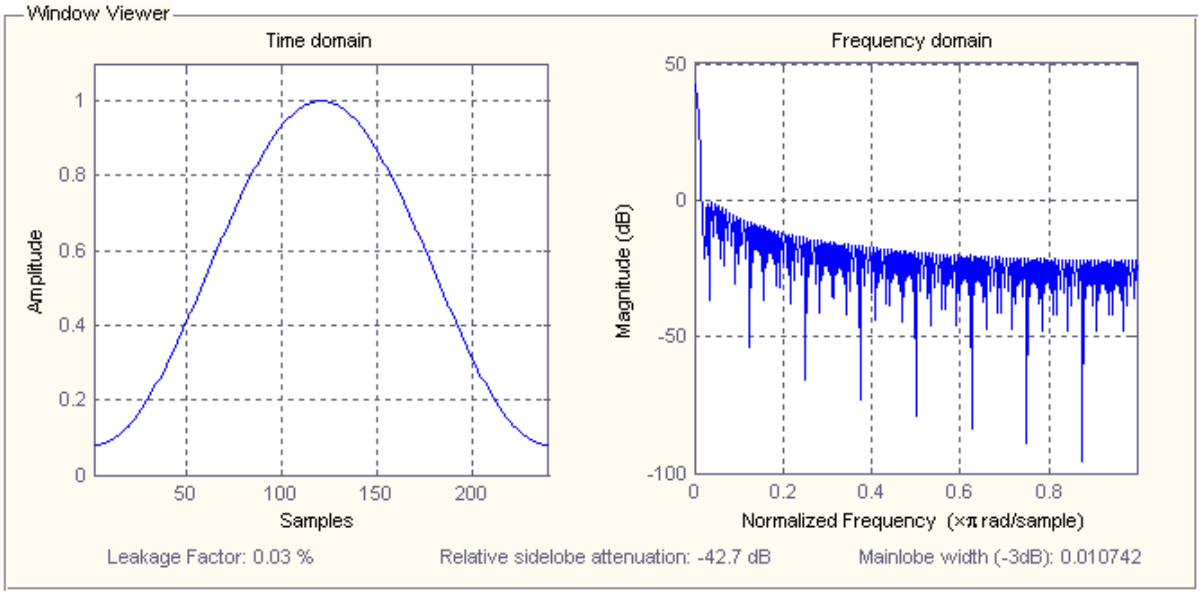

Fig 2.5 Hamming Window of length 256.

2.1.4 Autocorrelation[2]

LPC models the speech signal at time point n as an approximate linear combination of previous p samples:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p)$$

Where, a_1, a_2, …..,a_p are constant for each frame of speech. The basic idea is to predict the speech signal at a given time n based on the previous p samples.

We can make the approximation exact by including a "difference" or "residual" term, which is the excitation of the signal if the LPC coefficients are a filter:

LPC *can* be used to generate speech from either the error signal (residual) or a sequence of impulses as input:

$$\hat{s}(n) = e(n) + a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p)$$

where $\hat{s}$ is the generated speech, and $e(m)$ is the error signal or a sequence of impulses. However, we use LPC here as a *representation* of the signal. The values $a_1 \ldots a_p$ (where $p = 10$) *describe* the signal over the range of one window of data (256 samples). Each resonance (complex pole) in spectrum requires two LPC coefficients; each spectral slope factor (frequency=0 or Nyquist frequency) requires one LPC coefficient.
For 8 kHz speech, 4 formants $\Rightarrow$ LPC order of 9 or 10.

The autocorrelation of a speech signal s(n) can be given by:-

$$\phi(k) = \sum_{n=-\infty}^{\infty} s(n)s(n+k)$$

The autocorrelation of the segmented and windowed speech signal can be given by:-

$$R_n(k) = \sum_{n=0}^{N-1-k} \left[s(m+n)w(n) \right] \left[x(m+n+k)w(n+k) \right]$$

If we change $s(m)$ to s_m (signal x starting at sample m), then the equation becomes:

$$R_n(k) = \sum_{n=0}^{N-1-k} \left(s_m(n)w(n)\right)\left(s_m(n+k)w(n+k)\right)$$

and if we set $x_m(n) = s_m(n)\, w(n)$, so that x is the windowed signal of s where the window is zero for $n<0$ and $n>N-1$, then:

$$R_n(k) = \sum_{m=0}^{N-1-k} s_m(n) \cdot x_m(n+k) \qquad 0 \le k \le K$$

where K is the maximum autocorrelation index desired.

Note that $R_n(k) = R_n(-k)$, because when we sum over all values of m that have a non-zero y value (or just change the limits in the summation to $m=k$ to $N-1$), then

$$x_m(n) \cdot x_m(n+k) = x_m(n-k) \cdot x_m(n) = x_m(n) \cdot x_m(n-k)$$

In matrix form, equation looks like this

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \cdots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \cdots & R_n(p-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \cdots \\ \cdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \cdots \\ \cdots \\ R_n(p) \end{bmatrix}$$

Solve a Toeplitz (symmetric, diagonal elements equal) matrix for values of α Each resonance (complex pole) in spectrum requires two LPC coefficients;

each spectral slope factor requires one LPC coefficient. For 8 kHz speech, 4 formants $\Rightarrow$ LPC order of 9 or 10

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \qquad 1 \le i \le p$$

$$E^{(0)} = R(0)$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right] \Big/ E^{(i-1)} \qquad 1 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1$$

$$E^{(i)} = (1-k_i^2) E^{(i-1)}$$

$$\hat{a}_j = \alpha_j^{(p)}$$

The values of lpc coefficients($a_1, a_2, \ldots\ldots a_p$) can therefore be calculated for each framed and windowed speech signal. These feature further can be used for the speech recongition but these feature are not orthognal

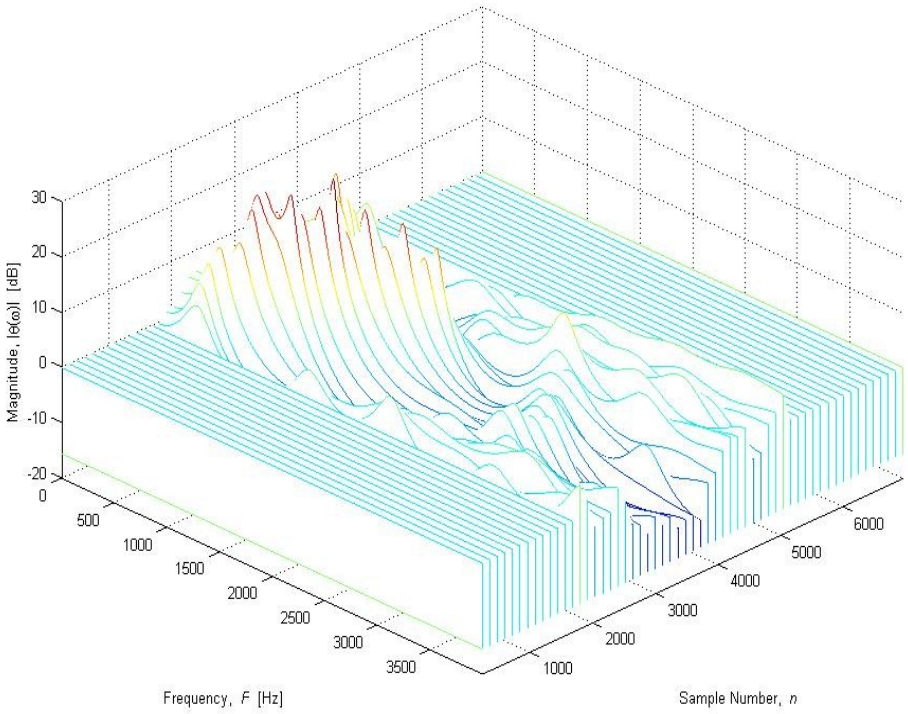

Fig 2.6 Vocal Tract filter resposnse of the word 'one'.

2.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

MFCC is one of the widely method that is used for the parametric representation of a speech audio signal. In MFCC speech signal is first preepmhasized then framing and windowing is done and thereafter Discrete Fourier Transform (DFT) is done for each windowed signal. After DFT is applied to each windowed signal, the signal is in the frequency domain and the signal is then applied to the mel filters. DFT analysis gives the amount of energy in the audio signal that is present within the frequency range for each bin. DFT analysis occurs in terms of number of equally spaced 'bins'. Each bin represents a particular frequency range DFT analysis gives the amount of energy in the audio signal that is present within the frequency range for each bin.

The result of applying mel-frequency filters is to reduce the amount of data. Instead of having a number of values same as number of bins produced by DFT, now have a number of values same as number of filters. In this project **512** point DFT has been used and **24** mel filters have been used. So **24** mel filter coefficients would be representing the speech signal, but for speech recognition purposes we only need **13** mel filter coefficients as they would be able to do our task so we will be neglecting the rest of the mel filter coefficients.

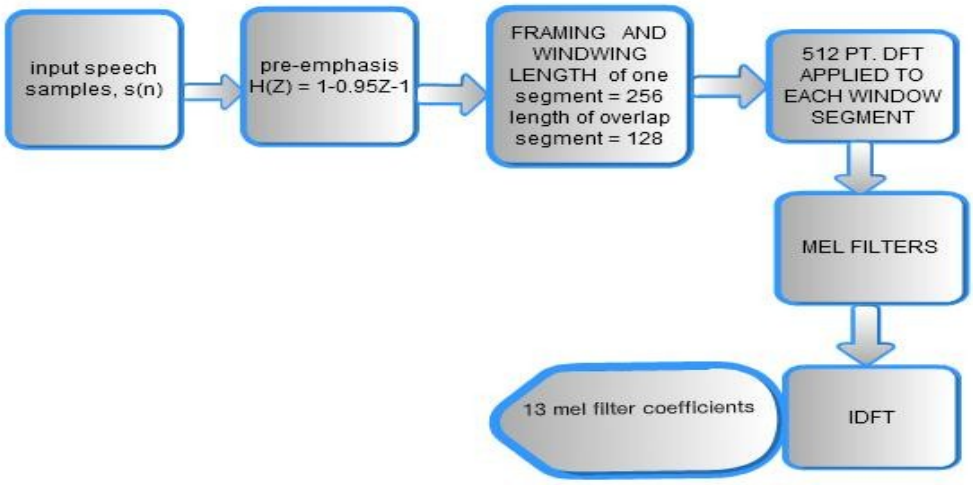

Fig 2.7 Flow diagram of MFCC process.

2.2.1 DFT

DFT applied to windowed segment

$$S(f) = \sum_{n=0}^{N-1} s(n) e^{-i2\pi kn/N} \qquad k = 0,1,\ldots\ldots..\text{N-1 (N 512)}$$

512 point **DFT** is applied to the each windowed segment and the signal from time domain has been converted into the frequency domain. **DFT** analysis occurs in terms of number of equally spaced 'bins'. Each bin represents a particular frequency range. **DFT** analysis gives the amount of

energy in the audio signal that is present within the frequency range for each bin.

2.2.2 MEL Filters:-

Linear frequency spacing below low frequencies below 1 KHz and a logarithmic spacing above high freq. above 1 KHz. Filters corresponding to these spacing seem to capture phonetically important characteristics of speech.

The human ear has high frequency resolution in low–frequency parts of the spectrum and low frequency resolution in the high–frequency parts of the spectrum. The coefficients of the power spectrum are now transformed to reflect the frequency resolution of the human ear. The human ear does not show a linear frequency resolution but builds several groups of frequencies and integrates the spectral energies within a given group.
Furthermore, the mid-frequency and bandwidth of these groups are non–linearly distributed. The non–linear warping of the frequency axis have been modeled by the so–called mel-scale. The frequency groups are assumed to be linearly distributed along the mel-scale. The so–called mel–frequency f_{mel} can be computed from the frequency f as follows

$$f_{mel}(f) = 2595 \log(1 + f / 700)$$

Fig 2.8 shows us the frequency to mel frequency curve. This basically plot the mel frequency with respect to the linear frequency using the above formula.

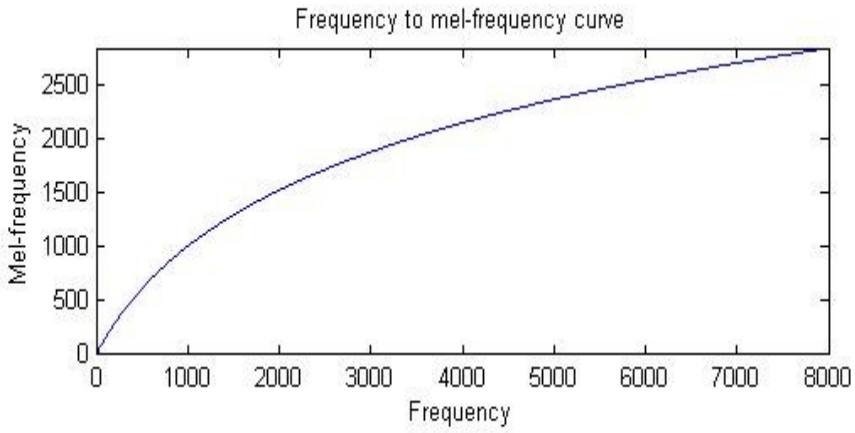

Fig 2.8 Frequency to mel- frequency curve.

Fig 2.9 shows us the mel-filters where total no. of filters used is 24. But we would be using only 13 mel spectral coefficients out of the total 24 that we would be getting for each framed and windowed signal. These 13 mel spectral coefficients are then inverted back into time-domain by using **IDFT** which has been discussed next.

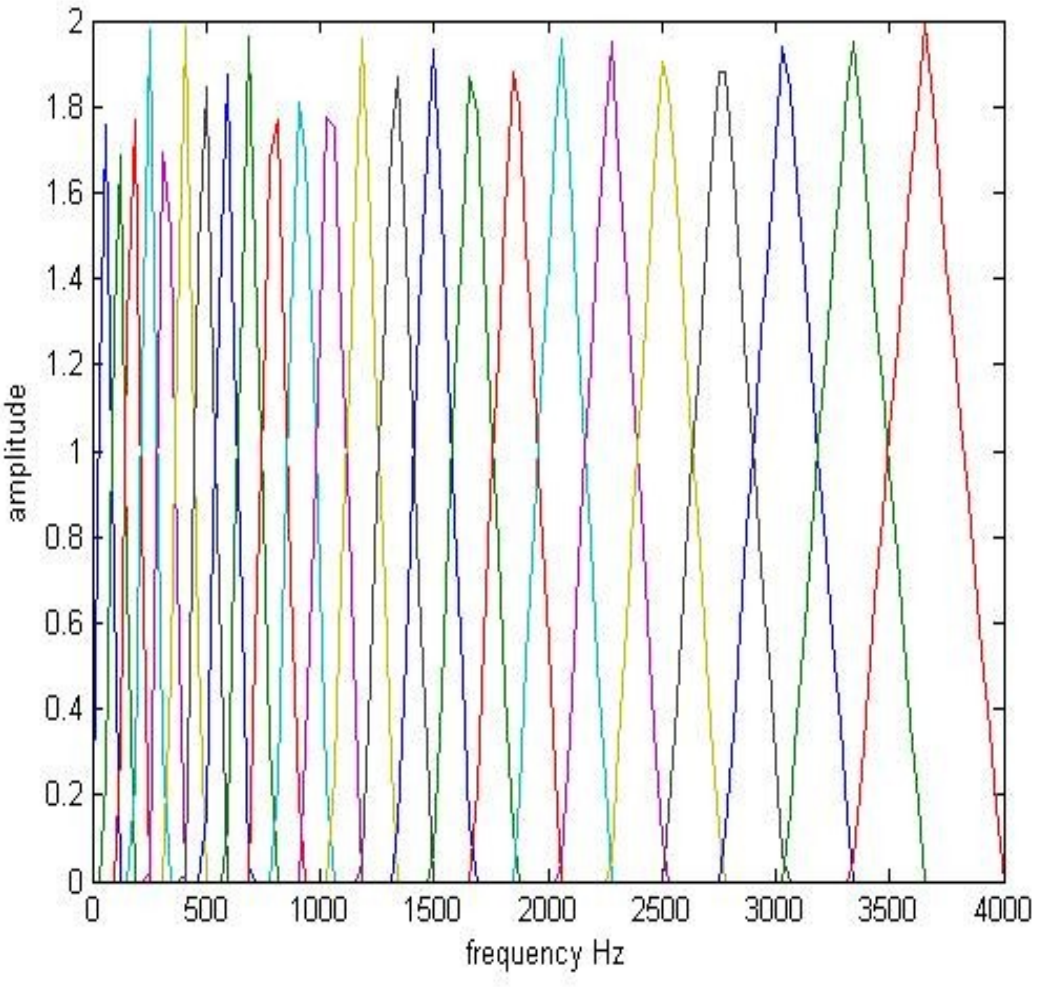

Figure 2.9 MEL Filters.(No. of filters = 24)

2.2.3 Inverse Discrete Fourier Transform (IDFT)

The speech signal in frequency domain $S(f)$ after passing through the mel filters we get $S'(f)$, we apply the inverse discrete Fourier transform .

$$s(n) = \frac{1}{N}\sum_{k=0}^{N-1} S'(f) e^{-i2\pi kn/N} \qquad \text{n = 0,1, … N-1 (N= 512).}$$

We get the 24 mel filter coefficients after all the process , but we will be using only first 13 mel filter coefficients for every windowed segment, which would be our features which we would be using for the speech recognition task.

CHAPTER 3

3.1 Dynamic time warping

In isolated word recognition systems the template of each word in the vocabulary has been stored as a time sequence of features i.e. MFCC. Recognition is comparing the acoustic pattern of the word to be recognized with the stored patterns and then choosing the word, which it matches best as the recognized word. This involves not only distance computation but also time-alignment of the input and reference patterns because a word spoken on different occasions, even by the same speaker, will exhibit both local and global variation in its time scale.[2]

The dynamic time warping uses the distance measure techniques to compare the distances between the known feature vectors in the codebook and the unknown feature vectors recorded during run time. In this project Euclidean distance has been used to compute the distances between the feature vectors.

3.2 Euclidean Distance

The distance measure between two feature vectors is calculated using the *Euclidean* distance metric. Therefore the local distance between feature vector X of signal 1 and feature vector Y of signal 2 is given by

$$d(x,y) = \sqrt{\sum_{i,j} (X_i - Y_j)^2}$$

X_i = known sequence in the codebook.

Y_j = unknown sequence to be compared with the known sequence in codebook.

3.3 DTW

Speech is a time-dependent process. Hence the utterances of the same word will have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed. "Time-time" matrix is used to visualize the alignment[2].

3.3.1 Algorithm[1]:-

- Compute the local distance i.e. the distance between frame 1 of the input and all the frames of each reference template. Call this cumulative distortion value for that element.
- Starting with frame 2 of input and beginning frame 1 of first reference template. Compute the local distance and add it to best cumulative distortion value for all possible predecessors.
- Continue this operation through each other column.
- Find the least global distance in last column for each reference template and declare it the distance associated with word
- Choose the word associated with best of global distance and declare it the winner.

Fig 3.1 shows the time-time alignment of the word “three” stored in the codebook and the word three spoken during the run time. From the left it starts with the comparison of the word ‘Three’ with the ‘One’ and goes down in serial order downwards till ‘Five’ and then upwards on right with ‘Six’ till ‘Zero’. X and Y axis represents the time in the multiple of 10 ms.

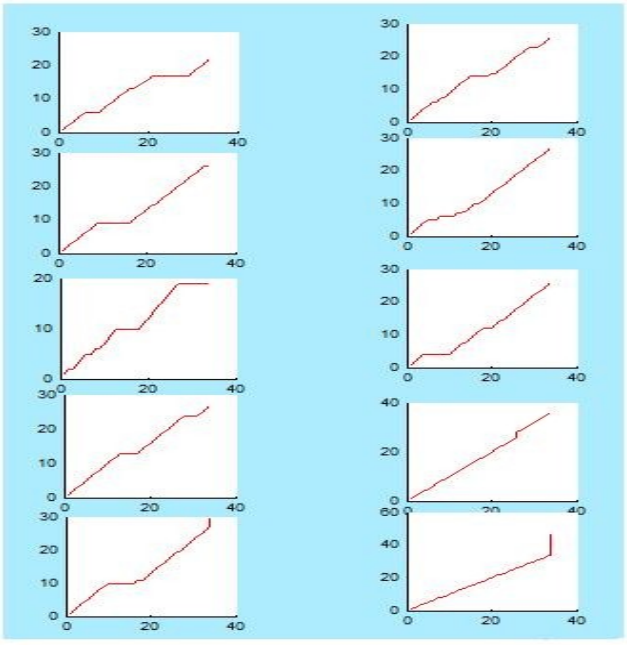

Fig 3.1 time-time alignment of word ‘five’.

CHAPTER 4GUI FOR SPEECH RECOGNITION

GUI in MATLAB has been made for the speech recognition task.

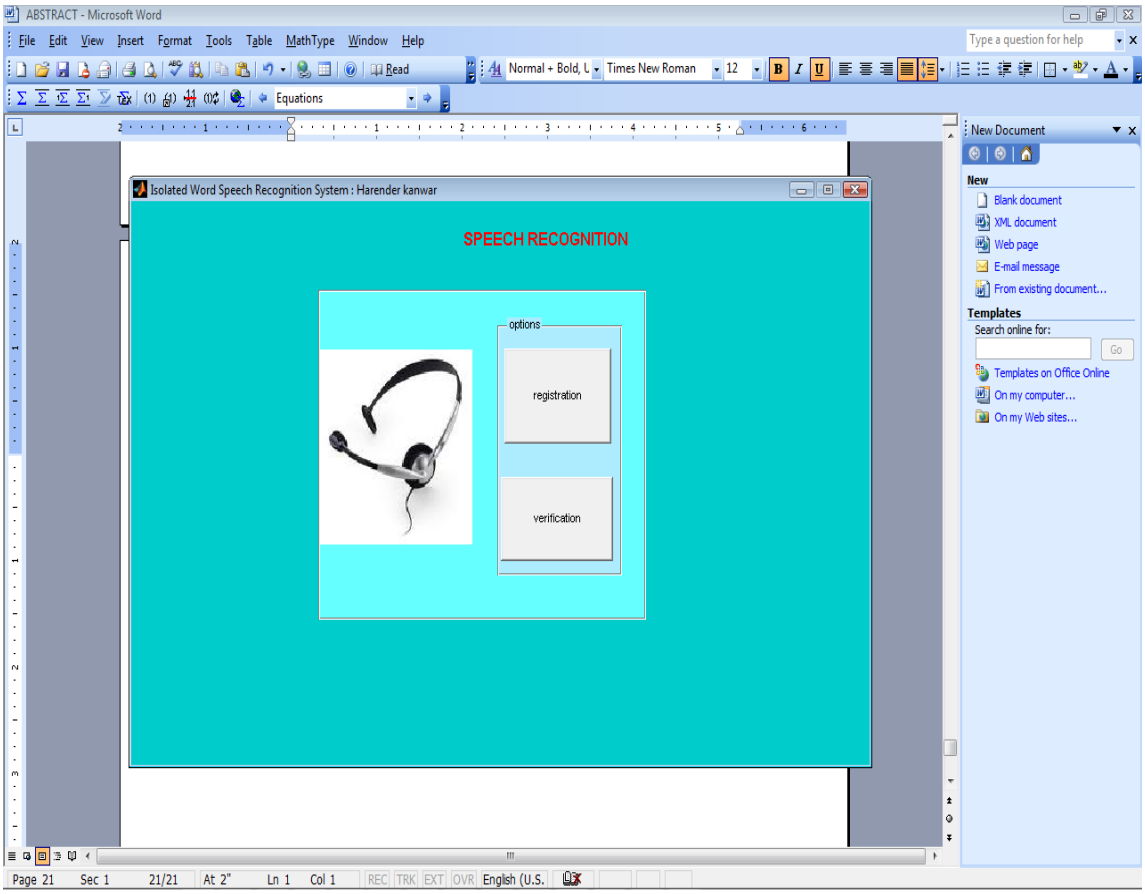

FIG 4.1 SPEECH RECOGNITION MAIN WINDOW

This is the main window of the GUI which has two options registration and the verification.

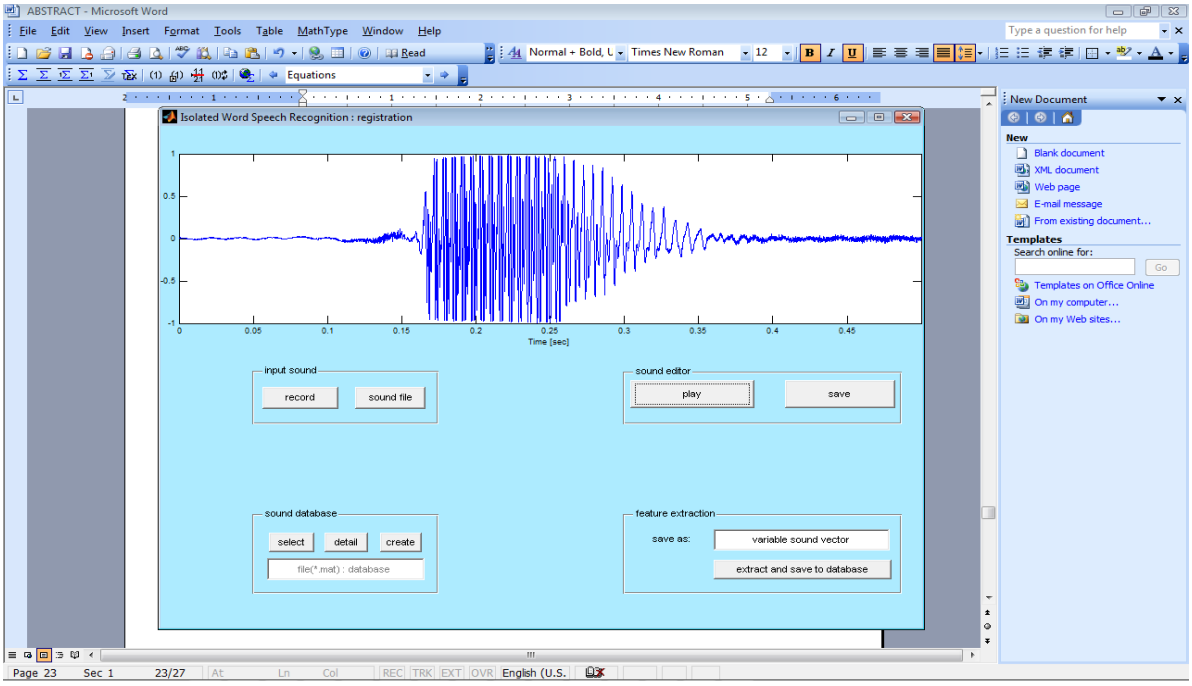

Fig 4.2 Registration Window

After clicking on the registration button in the main window this window will open up which has the options of recording sound and or loading the sound file from our pc. The recorded sound can also be saved with sampling rate of 8 KHz and bit resolution of 8 bits/sample. The MFCC features are extracted

from the speech recorded or loaded from the pc with the steps discussed in the chapter 2. You can also view the MFCC feature vectors by clicking on the detail button on the screen.

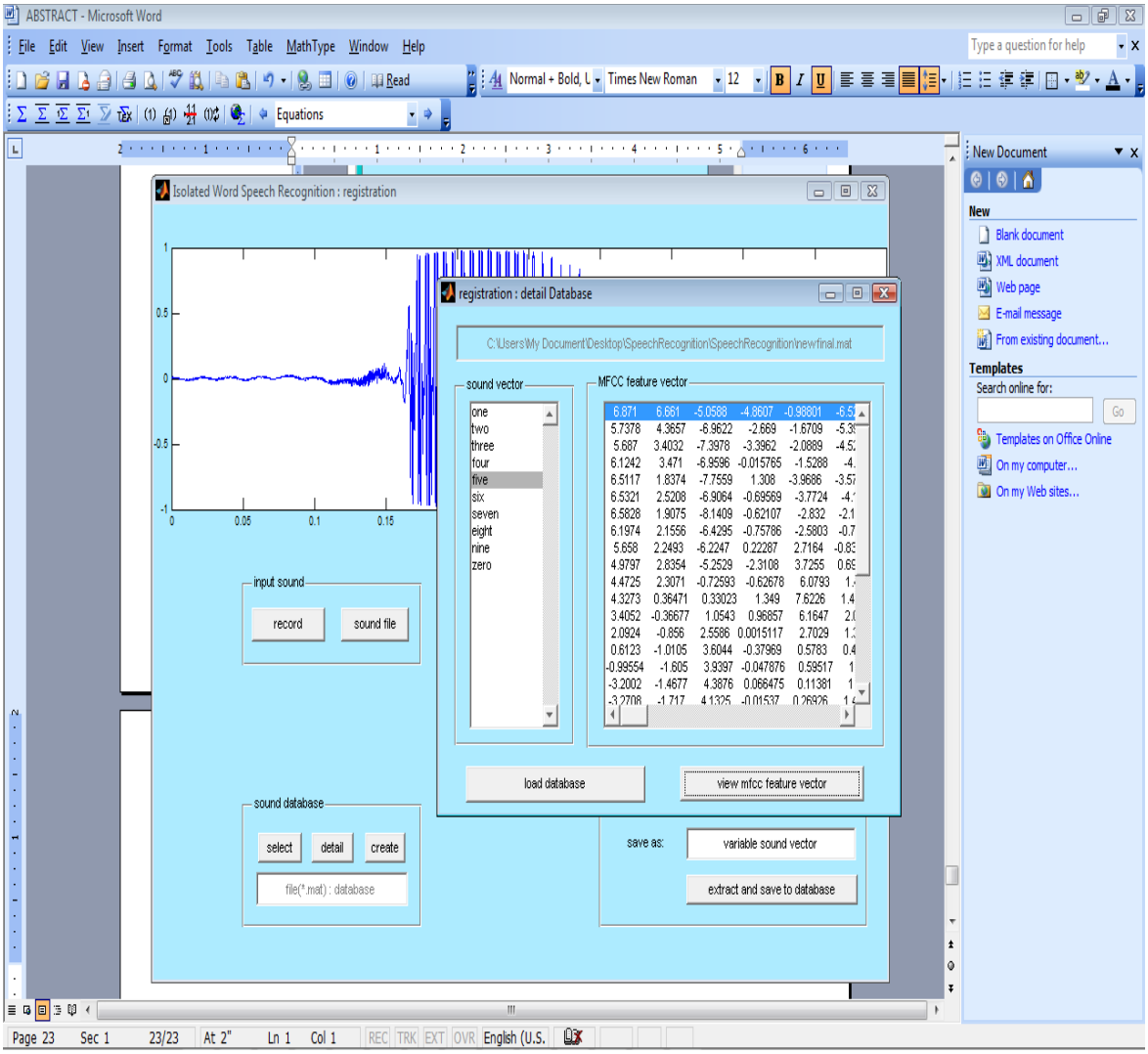

Fig4.3 MFCC vector for the recorded speech.

Above figure shows us the mel frequency cepstral coefficients that we got after applying the **MFCC** . We got the 13 mel cepstral coefficients for the speech signal recorded. We can view the cepstral coefficients of any word varying from zero to nine.

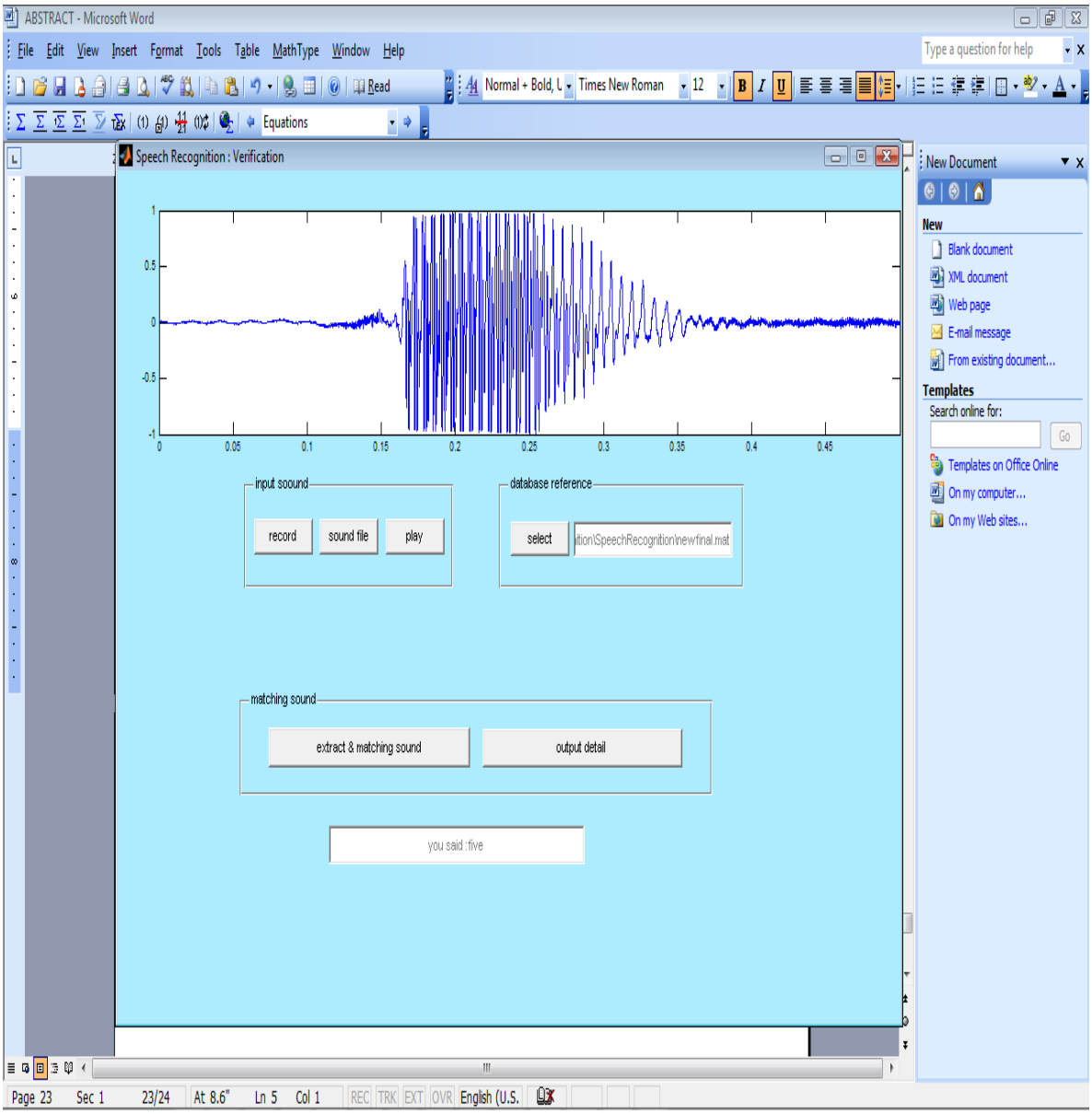

Fig 4.4 Verification Window

The speech can here be recorded or loaded from the pc .The extract and matching sound pushbutton on pressing will extract the features from the speech signal using MFCC and Dynamic Time Warping Technique (DTW)

and the global distances are calculated for the each word stored in the codebook and one with minimum distance is declared the winner.

–

APPENDIX

Downsample.m :- Function used to downsample the recorded speech signal from 44.1 KHz and bit resolution 16 bits/sample to 8 khz and 8 bits/sample.

Spectrogramproject.m :- Function used to plot the narrow band and wideband spectrograms.

Premphasise.m :- Function to preemphasise the speech signal.

FFToneside.m :- Function used to calculate the fft of speech signal of only real part.

Fftspeech.m:- Function to see the effect of preemphasizing the speech signal

LPCauto.m :- Function used to calculate the the lpc coefficients using auto-correlation.

LPC.M:- Function used to calculate the lpc coefficients and plot the vocal tract filter response.

MFCC.M:- Function used to calculate MFCC features from the speech signal.

Distecu.m:- Calculates the Euclidean distance between the feature vectors.

Dtw.m:- Implements the DTW.

DFT.m :- Function calculates the DFT of speech signal

IDFT.m :- Function calculates the IDFT of the spectral coefficients.

Mainspeechrecognition.fig:-Main window of the speech recognition GUI.

Mainspeechrecognition.m:- Function supporting the mainspeechrecognition.fig file

Registration.fig :- GUI window to extract the features and store them to the database.

Registration.m :- Function supporting the Registration.fig file.

Detaildatabase.m :- Shows the extracted features of speech.

Verification.fig:- GUI window to verify the uttered word.

Verification.m :- Function supporting the verification.fig file.

DISTANCETIME.m :- Function used to calculate the global distance and plot time alignment of stored signals vvs uttered word.

DISTANCETIME.fig :- Fig file used to support the distancetime.m

REFRENCES

[1]Lawrence and Rabiner and Biing- Hwang Juang "Fundamentals of Speech Recogntion"

[2] Gold, B. and N. Morgan, 2000. Speech and Audio Signal Processing, John Wiley and Sons, USA.