Recognition of Spoken Words Using Mel Frequency Cepstral Coefficients & Dynamic Time Warping Algorithm

Project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

By

NIKHIL ELIAS (141056)

AMIT MEHTA (141074)

SHWET ANMOL (141075)

UNDER THE GUIDANCE OF

DR. MEENAKSHI SOOD

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

May-2018

TABLE OF CONTENTS

DECLERATION BY THE SCHOLAR

We hereby declare that the work reported in the B-Tech Thesis entitled "**Recognition of Spoken Words using Mel Frequency Cepstral Coefficients and Dynamic Time Warping Algorithm"** is an authentic record of my work carried out under the supervision of "**Dr. Meenakshi Sood"**. We have not submitted this work elsewhere for any other degree or diploma.

Nikhil Elias (141056)

Amit Mehta (141074)

Shwet Anmol (141075)

Department of Electronics and Communication Engineering,

Jaypee University of Information Technology, Waknaghat, India

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
(Established by H.P. State Legislative vide Act No. 14 of 2002)
P.O. Waknaghat, Teh. Kandaghat, Distt. Solan - 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: +91-01792-245362

CERTIFICATE

This is to certify that the work reported in the B.Tech project report entitled **"Recognition Of Spoken Words Using Mel Frequency Cepstral Coefficients And Dynamic Time Warping Algorithm** "which is being submitted by **Nikhil Elias (141056), Amit Mehta (141074)** and **Shwet Anmol (141075)** in fulfillment for the award of Bachelor of Technology in Electronics and Communication Engineering by the Jaypee University of Information Technology, is the record of candidate's own work carried out by him/her under my supervision. This work is original and has not been submitted partially or fully anywhere else for any other degree or diploma.

Dr. Meenakshi Sood

Department Coordinator
Assistant Professor (Senior Grade)
Department of Electronics & Communication Engineering
Jaypee University of Information Technology, Waknaghat,

ACKNOWLEDGEMENT

In performing our project, we had to take help and guidelines from some respected persons who deserves our greatest gratitude. We would like to show our gratitude to **Dr. Meenakshi Sood, Assistant Professor (Senior Grade)** for giving us the guidelines for project throughout numerous consultations. We would also like to extend our deepest gratitude to all those who have directly and indirectly guided us in completing the project and this report.

We also thank **Prof. Dr. Samir Dev Gupta, Head of Department, Electronics and Communication Engineering**, Jaypee University of Information Technology, Waknaghat for consent to include copyrighted pictures as a part of our report. We also thank **Mr. Kamlesh**, Technical and Laboratory, Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Waknaghat for providing us with all facilities, necessary components required to complete the project.

Nikhil Elias (141056)

Amit Mehta (141074)

Shwet Anmol (141075)

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

ABBREVIATIONS	FULL FORM
MFCC	Mel Frequency Cepstral Coefficients
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
Hz	Hertz
AWGN	Additive White Gaussian Noise
SNR	Signal To Noise Ratio
KHz	Kilo-Hertz
DCT	Discrete Cosine Transform
VQ	Vector Quantization
ms	Milli-seconds

LIST OF TABLES

ABSTRACT

Voice is an intriguing aspect of the human anatomy. Humans have the capability of producing sounds with frequency ranging from as low as zero Hertz to four thousand Hertz. But the interesting part lies in the fact that each individual voice is totally different from anyone else. Thus voice can also be used as a second fingerprint for recognition and other significant purposes. Also any software or application based upon speech recognition has an immense range of applications in fields of defence, security, healthcare and in automation too. To apply this hypothesis, one first has to extract some special hidden features from the human voice for recognition. These features are mainly stored in the frequency spectrum, namely the formants, pitch and frequency range.

Various algorithms such has LPC (Linear Predictive coding), MFCC (Mel Frequency Cepstral Coefficients) and HMM (Hidden Markov Models) are present that work over extracting these features. These algorithms obtain features by examining small pieces of the whole signal at a time, thus assuming that small piece of the signal to be stationary over the duration of the time. The signal is considered stationary in terms of the frequency in-variations .The recognition phase comes under the category of Feature Matching which is carried out using the DTW (Dynamic Time Warping) Algorithm. This algorithm helps in comparing any time varying signals with accurate results.

CHAPTER 1
INTRODUCTION

1.1 HISTORY

The first attempts to recognize spoken language goes as far as back to 1950s. Till this point, most of the attempts were at creation and recording of voice but none at interpretation of the same. A machine created by Bells Lab named, ‘AUDREY’, was the very first machine that could understand digits 0-9, with 90% accuracy. The accuracy varied from 70% to 80% when individuals other than the recorded spoke to the machine.

Until the 1990s, most of the successful recognition algorithms and machines were based upon template matching. The recorded set of voice signals would be translated into a set of numbers. These would then trigger only when similar sound was spoken into the machine. The main challenge faced under template matching was to record the sounds without background noise corrupting the signal.

The Hidden Markov Models were first used in IBM Tangora, released in mid 1980s designed by Albert Tangora. It could adjust itself to the speaker, required 20 minutes of training data and could recognize up to 20,000 words. But the input speech signals had to be slow, clear and noise free for better accuracy. The use of HMMs allowed greater flexibility in voice recognition machines.

In 1997 the first “Continuous Speech Recognizer”, named as Dragon, was released. One was no longer required to pause while entering words into machine. Capable of understanding 100 words per minute, the dragon software is still in use today and is favoured by numerous individuals in medical field for notation purposes.

In the 20th century, the research carried out in the field of machine learning paved the way to breakthrough for speech recognition. Google combined the latest technology with the power of cloud based computing to share data and to further improve the accuracy of machine learning algorithms.

Siri, “Apple’s” entry into the recognition market captured attention of the public. After Apple, Microsoft launched “Cortana” Amazon introduced “Alexa” both of which now competes to gain maximum market attention. As a result, wheels have been set in motion among the IT and Tech Giants to improve speech recognition platforms.

1.1 SPEECH RECOGNITION

Speech recognition is the ability of a machine or a particular programme to identify words and phrases in a given particular language. It the task of the program to convert the speech into a machine readable format so that further processing can be carried out on the recorded signal.

The recognition works on algorithms that depend on how modelling of language is to be carried so that proper matching of words can be carried out. Speech recognition refers to identifying the commands/utterances of the user and executing any function corresponding to the command. The objective is to make any computer or system understand the human language.

Speech recognition can be classified into two categories; isolated word and non-isolated word recognition. In former, the speaker only speaks a single word into the microphone. In the later type, no restrictions are imposed on the speaker over the number of words to be spoken. The speaker may pronounce a single word to a full sentence. It is the task of the system and the algorithm to grasp the required word for recognition from the sentence.

Isolated word recognition is relatively easier when compared to non-isolated word recognition. Non isolated recognition requires first identification of the required word and then it has to be separated from the complete sentence. The complete task requires complex procedure and imposes much difficulty. Also there is a risk of introducing noise into system if clipping is done. On the other hand isolated recognition requires no such use of complex procedure.

1.2 MFCC

Generally human speech conveys large magnitude of information such as the speaker's gender, emotions and present conditions. This information is referred to as features in terms of speech recognition. MFCC algorithm is one of the most used for converting this information into features unique to each speaker. This is known as Feature extraction

Feature extraction process is the characterization of the speaker specific information contained in any signal. It transforms the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies are suppressed. This speaker specific information is derived from the vocal tract. Pronunciation of various phones, accent, pitch, loudness, duration of each word and rhythm also contribute towards speaker specific information.

MFCC algorithm is based on the important idea of cepstrum. The application of algorithm results in a group of coefficients, each one having its own meaning.

1.3 DTW

DTW is a well-known technique to find an optimal alignment between two given time varying sequences under certain restrictions. One can say that using this algorithm the two sequences are non-linearly warped to each other. The task accomplished using DTW is known as Feature Matching. This algorithm calculates the least weighted path between the signals. DTW measures this distance on the basis of similarity between the stored vector in the database and the input voice signal from the user. The stated algorithm helps in compensation of time difference between any two spoken sequences. It is well known that any two sequences can never be exactly the same owing to different speaker rate. The non-linear matching used in this technique eliminates this problem effectively.

Speech recognition has ample of applications which uses both feature extraction as well as feature matching as discussed in the following sections. Practical applications for speaker identification consist of various kinds of security system where human voice can act as key for security purposes.

CHAPTER 2

LITERATURE SURVEY

1. **Hasina Shaikh, Dr. Luis C. Mesquita, Sufola Das Chagas Silva Araujo, "Recognition of Isolated Spoken Words and Numeric using MFCC and DTW" ,International Journal Engineering science and computing, April 2017,Volume 7, Issue No. 4**

- This paper focuses on recognizing certain words and decimal numeric zero to nine using MFCC and DTW algorithms for feature extraction and feature matching respectively.
- Implemented using C++ which provides good development flexibility and requires relatively short execution time.
- For the recognition of words, a database comprising of five speakers; and that for the numbers, a database of total six speakers, 3 males and 3 female
- A test voice sample of any of the stored word or numeric is again recorded and then the algorithm is applied to recognize the same. To cope with varying speaking speeds in speech recognition, Distance Time Warping, a method that measures similarity between two sequences which may vary in time or speed, is used.

2. **Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar, "Self-learning Vector Quantization for Pattern Discovery from Speech", Interspeech Brighton, 2009**

- In this paper a novel and computationally straightforward clustering algorithm was used for VQ of speech signals for a task of unsupervised pattern discovery from speech.
- The algorithm is computationally extremely feasible, and achieves comparable classification quality with the well-known k-means algorithm in the PD task.
- In addition to presenting the algorithm, general findings regarding the convergence of the clustering algorithm, and the ultimate quality of VQ codebooks are discussed.

3. **L indasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic time warping techniques", Journal of Computing, March 2010, Volume 2, Issue 3**

- This paper discuss several methods such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN) and etc to identify a straight forward and effective method for voice signal.
- The extraction and matching process is implemented right after the Pre Processing or filtering signal is performed. The non-parametric method of Mel Frequency Cepstral Coefficients (MFCCs) are utilize as extraction techniques.
- The nonlinear sequence alignment known as Dynamic Time Warping (DTW) introduced by Sakoe Chiba has been used as features matching techniques.
- This paper presents the viability of MFCC to extract features and DTW to compare the test patterns.

4. **Xuedong Huang and Kai-Fu Lee, "On Speaker-Independent, Speaker-Dependent and Speaker-Adpaptive Speech Recognition", IEEE transaction on Speech and audio processing, Volume 1, No. 2, April 1993**

- Speaker-independent speech recognition systems are desirable in many applications where speaker-specific data do not exist.
- In this paper, author have used the DARPA Resource Management task as his domain is to investigate the performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition.
- The error rate here was reduced substantially by incorporating sex-dependent semi-continuous hidden Markov model.
- This work demonstrate that speaker-adaptive systems outperform both speaker-independent and speaker-dependent systems and suggests that the most effective speech recognition system is the one that begins with speaker-independent training and continues to adapt to users**.**

CHAPTER 3

METHODOLOGY

The entire process of speech recognition involves use of two main algorithms, namely MFCC and DTW. The input voice signal from the users is first converted into digitalized signal using MATLAB software. Entire recording and storing of these signals is done using the same software. After recording the signal redundant part of the signal (the parts where the user does not speak) are clipped off. This is done only to take in important part of the signal.

Figure 3.1 describes the entire steps and process that were undertaken to complete the objective of recognizing speech signals.

The sub modules used in Feature Extraction system are Noise Filtering, Pre-Emphasis, Windowing, Energy calculations and Mel filter bank which is used for the extraction of the MFCC coefficients which is the major part in recognizing the spoken word and will be briefed in the further chapters.

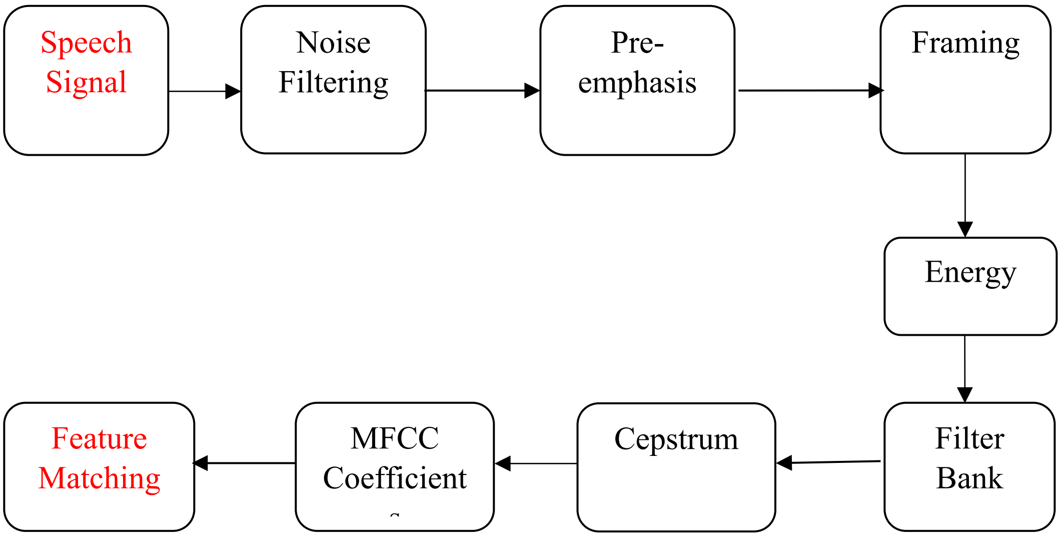

Figure 3.1: Flow chart of adopted Model

The end result of Feature Extraction is a feture Matrix. The feature matrix is then used as the basis for comparing and matching input signal with the stored signal in the database. DTW algorithm is the used for accomplishing the task of Feature Matching.

3.1 FEATURE EXTRACTION

Feature Extraction involves obtaining the hidden traits of a person’s voice. An individual has different way of pronunciation of alphabets or words, thus the frequency spectrum varies from one person to other. Also the features extracted have to be stored with reference to the human auditory perception.

Feature extraction is accomplished using the MFCC algorithm. MFCC is based upon Human listening behaviour. Human listening adopts a non linear behaviour. Up to 1000Hz, the graph can be approximated to be linear after which it becomes visibly non linear. The adopted MFCC algorithm takes this concept into consideration while calculating the features.

3.2 NOISE FILTERING

Noise filtering is the first step carried out before any other processing is conducted over the signal. The input voice signal is mainly corrupted by surrounding noises such as wind and breath. All these sources contribute towards corrupting the signal by introducing static. As such, the feature extracted contains less useful information. Thus it becomes imminent to employ a filter that can remove these noises without disturbing the original signal. To accomplish this objective Savitzky-Golay filter has been chosen. The noises introduced during recording degrade the SNR of the processed signal which further leads to degradation of the MFCC coefficients. Savitzky-Golay filter is used for creating an approximating function that tries to take into account of capturing important pattern in the data set or signal. This leads to increase in the SNR of the signal. One of the major advantages of using this filter is that it increases the SNR of the signal without introducing time delay in the signal.

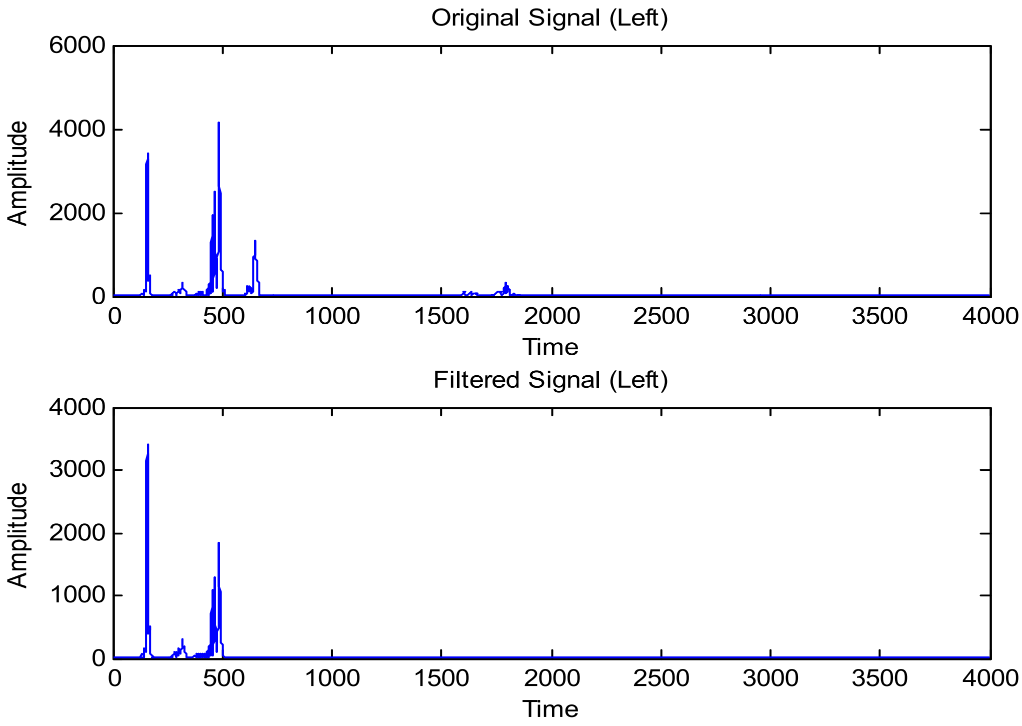

Figure 3.2: Filtering Using Savitzky Golay Filter

3.3 PRE-EMPHASIS

While considering the energy spectrum of speech signal, it was calculated that the energy content of the waves decreased rapidly as the frequency increased. The lower frequency spectrum contained majority of the energy. The higher frequency components provide substantial distinction between any two signals. Also the higher frequency component of the speech signal contains important information which helps in the recognition of the spoken commands. So Pre-emphasis is done to increase the energy of high frequency components of the input signal. Each value of the speech signal is raised using the equation

$$Y[n] = X[n] - \alpha * X[n-1] \qquad \text{---------------} (3.1)$$

Here the value of 'α' ranges from 0.8 to 1.0. The pre-emphasis filter is mainly a high pass filter with changes done in its levels.

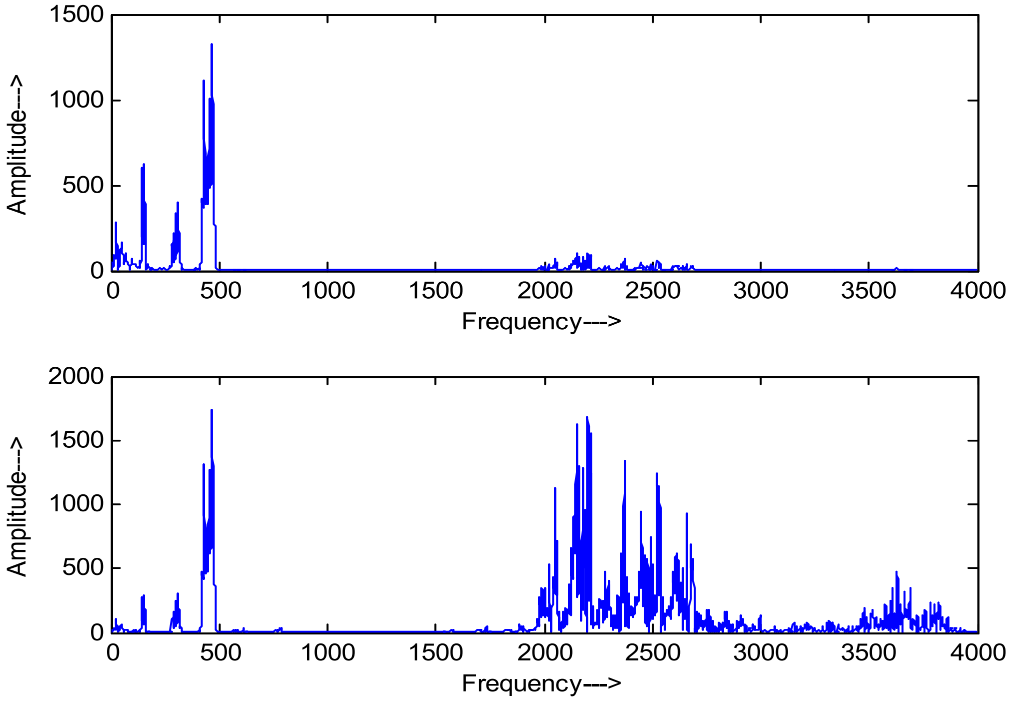

Figure 3.3: Pre-Emphasis of Alphabet ‘A’

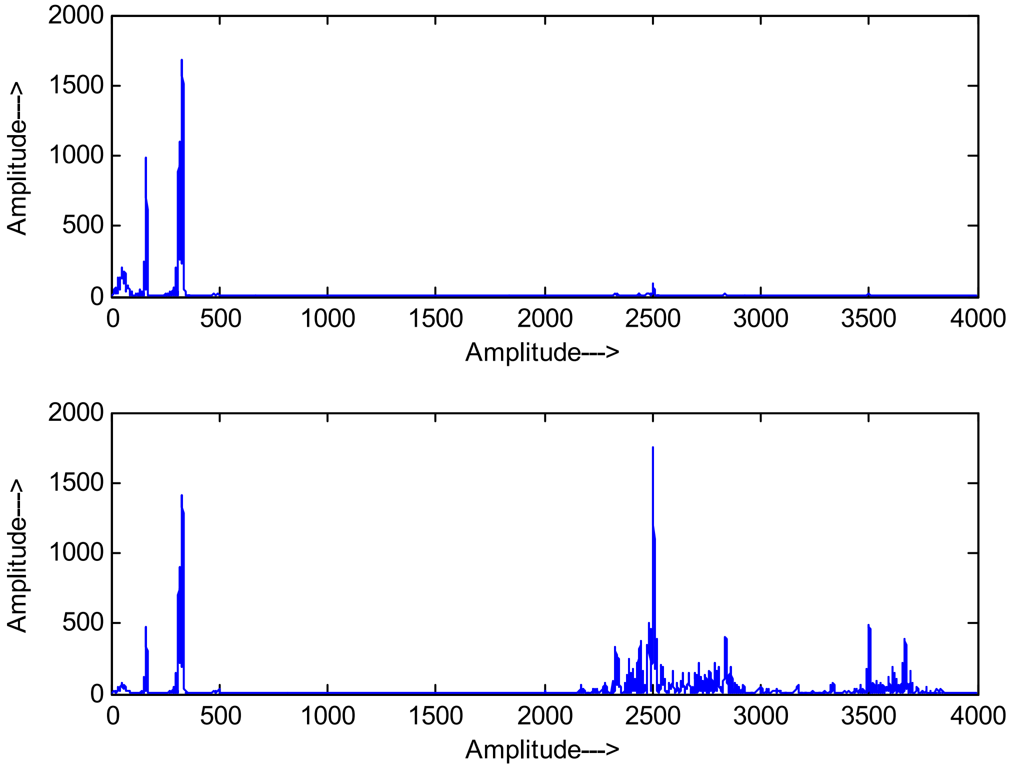

Figure 3.4: Pre-Emphasis of Alphabet ‘B’

Figure 3.2 and 3.3 shows the Pre-Emphasis of alphabets ‘a’ and ‘b’ in which we have taken the value of α to be 0.85. From the figure we can see that the high frequency components of the signal have a huge boost in their energy.

3.4 FRAMING & WINDOWING

Framing is the main part of the whole MFCC process .Since the speech signal is considered to be time varying it has to be approximated as a stationery signal. This is possible only by considering small parts of the signal one at time and calculating its coefficients. This is the whole idea behind carrying out framing.

The important part lies in calculating the framing size. Based upon the concept of STFT, size of the framing defines the amount of time and frequency information that will be contained in the individual frame. It is important that frame comprise out about 15ms-25ms of the whole signal. Limiting the size of the frame till the stated time range ensure that signal stored in the frame can be approximated to be stationery.

If the size of the frame is decreased further, the information extracted will be meaningless. On the similar lines if the size of the frames is increased beyond the upper range of about 25ms, the accuracy of the MFCC will decrease considerably as the signal will cease to be stationery. The size of framing also determines the size of the final feature matrix after carrying out the entire extraction process.

After performing framing over the signal another point that has to be considered is the overlapping between two frames. Overlapping ensures that some continuity remains between the frames. As a thumb rule, the size of overlapping is taken to be half of the framing size.

Windowing is the next step succeeding framing. Under this, a window signal is used upon each of the individual frame. Introducing such a windowing signal is important as it removes the Gibbs phenomenon, introduced by truncating the signal, which greatly corrupts the signal stored in each frame. The choice of a window function plays a crucial role in determining the quality of the results.

The parameters upon which to select the window function are mainly

1.) Maximum Side Lobe Level
2.) Half Main Lobe Width

3.) Side Lobe Fall Off Rate

Based upon the above stated parameters a particular window has to be selected among the various window function at our disposal

The following sub-sections discuss the distinct windows available for use in the MFCC algorithm. The response of each window is calculated and is then plotted. The mathematical equations governing each of the windows are also given along with their response.

3.4.1 TRIANGULAR WINDOW

The triangular window is described using the equation given below

$$w(n) = 1 - \frac{2|n - \frac{N-1}{2}|}{N-1} \qquad \text{---------------} \ (3.2)$$

The impulse response of the triangular window is shown in the figure 3.5.

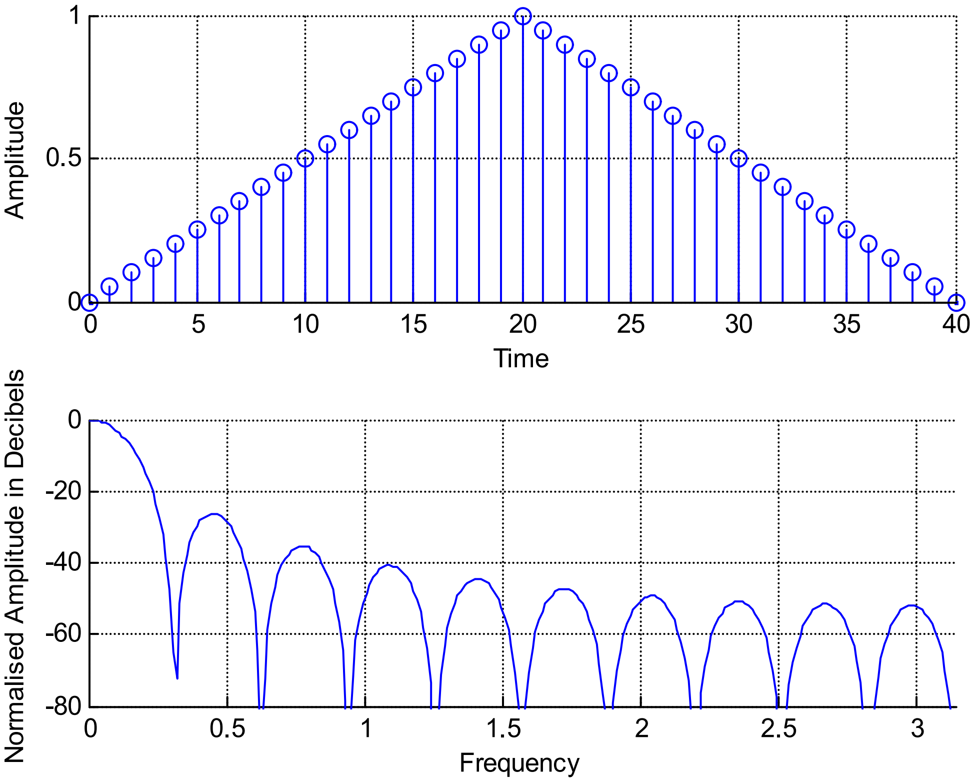

Figure 3.5: Response of Triangular Window

3.4.2 HAMMING WINDOW

The equation governing hamming window is given below

$$w(n) = 0.54 - 0.46 * \cos\,(\frac{2\pi n}{N-1}) \qquad \text{---------------} (3.3)$$

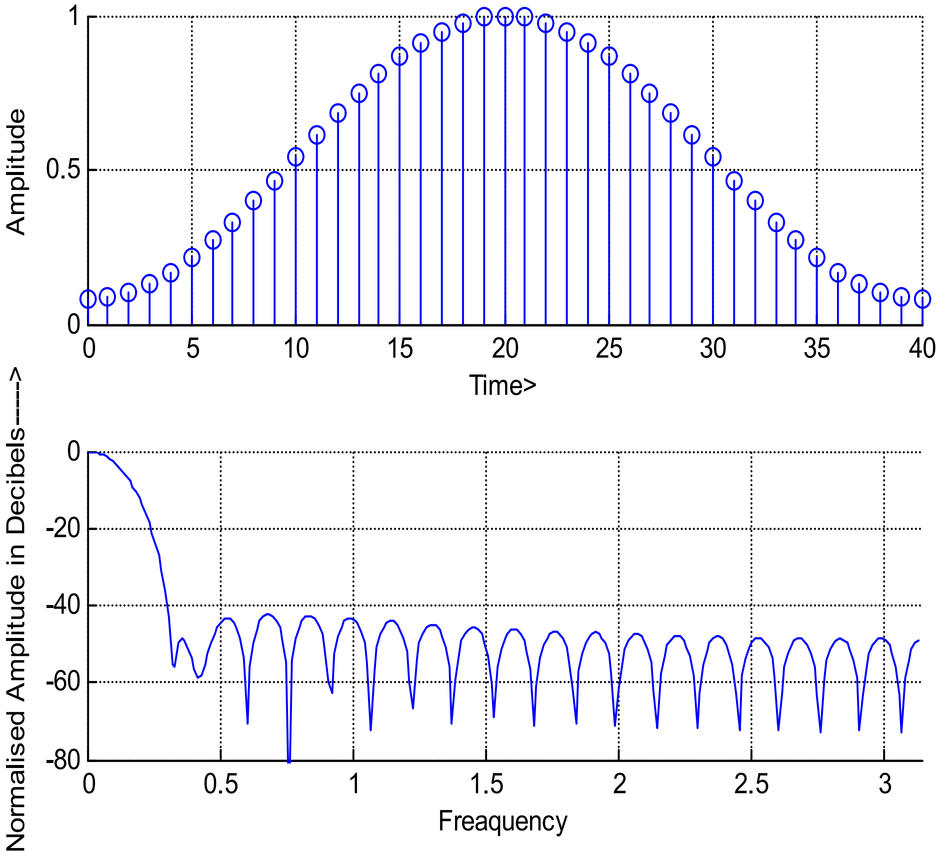

Figure 3.6: Response of Hamming Window

3.4.3 HANNING WINDOW

Mathematical equation of Hanning window is

$$w(n) = 0.5 * (1 - \cos\left(\frac{2\pi n}{N-1}\right)) \qquad \text{---------------} (3.4)$$

Figure 3.7 shows the response of the Hanning window.

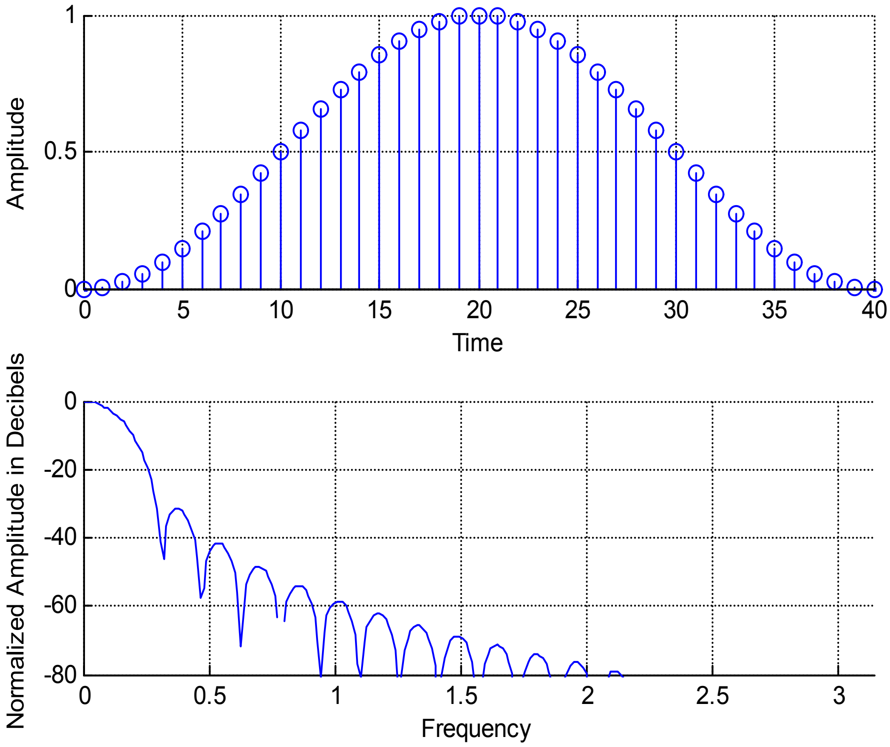

Figure 3.7: Response of Hanning Window

3.4.4 BLACKMANN WINDOW

Blackmann window is described by the equation

$$w(n) = 0.42 - 0.5 * cos\frac{2\pi n}{N-1} + 0.08 * \cos\frac{4\pi n}{N-1} \quad \text{---------------} \ (3.5)$$

The response of Blackmann Window is shown in the figure 3.8.

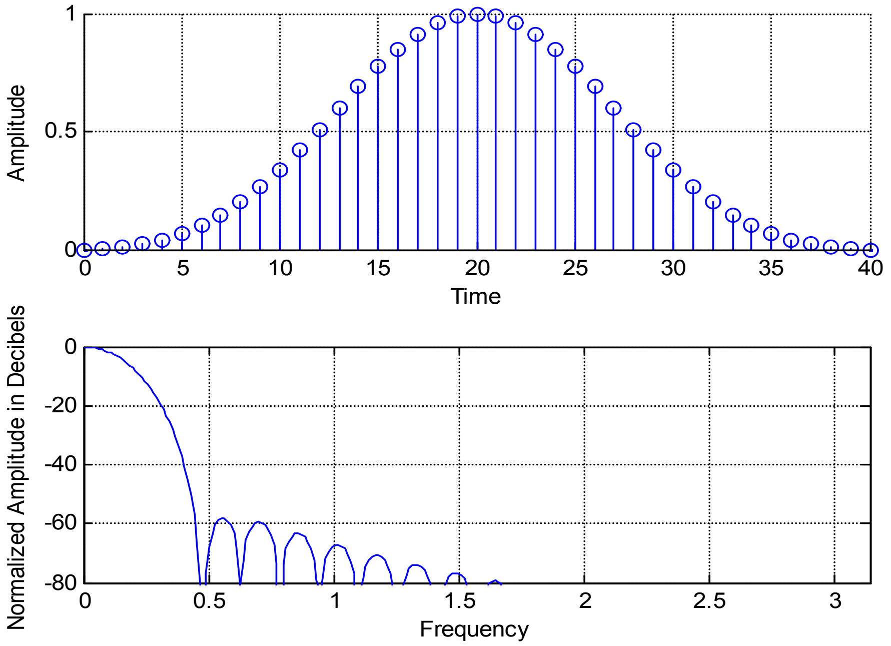

Figure 3.8: Response of Blackmann Window

Of These whole choices of windows available, hamming window was selected as it yielded better accuracy during comparisons.

3.5 FAST FOURIER TRANSFORM

Any information required for comparing two signals is stored in the frequency spectrum of the signal. Time domain representation of the signal yields no important information that can be used for recognition target.

The purpose of using the FFT is to convert speech signal in time domain to frequency domain which is carried out using the equation

$$X\,(k) = \sum_{n=0}^{N-1} X[n]e^{-j2kn\frac{\pi}{N}} \qquad \text{---------------} \; (3.6)$$

Where X (n) is the speech signal in time domain and ‘k’ varies from 0 to N.

By using FFT the complexity of the calculations required is greatly reduced. Further we calculate the energy of the above signal by first taking the absolute value of the above signal and then taking the square of the absolute signal.

3.6 MEL FILTER BANK

One of the main concepts that MFCC revolves around is to map the normal frequency spectrum of a signal to listening behaviour of a user. The auditory perception of a normal individual is linear up till 1 KHz which then becomes non linear on further increase in frequency.

Such an ideology is accomplished using a set of filters combined together in such a way that non linear mapping is achieved. This set of filters is known as Mel Filter Bank.

The filter bank comprises of triangular filters. Each filter calculates energy in a pre defined range of frequency.

The equation for defining filter bank is as follows

$$Hm(k) = 0 \text{ , if } 0 \leq k < fc(m-1) \text{ ;} \qquad \text{-------------} (3.7)$$

$$Hm(k) = \frac{k - fc(m-1)}{fc(m) - fc(m-1)} \text{ , if } fc(m-1) \leq k < fc(m) \text{ ;} \qquad \text{-------------} (3.8)$$

$$Hm(k) = \frac{k - fc(m+1)}{fc(m) - fc(m+1)} \text{ , if } fc(m) \leq k < fc(m+1) \text{ ;} \qquad \text{-------------} (3.9)$$

Where ‘m’ is the number of filters we want and f() is the list of M+2 Mel spaced frequencies.

Figure 3.9 shows the designed filter bank for the MFCC algorithm. The initial filters are spaced linearly followed by non liner spacing of the triangular filters. Each of the filters calculates the energy in its designated region. The outputs of the filters are the Mel coefficients which have been discussed in section 3.8.

The filters are mapped on the frequency axis using the Mel scale. The Mel scale follows a non linear behaviour mapping the entire frequency axis based upon human listening behaviour.

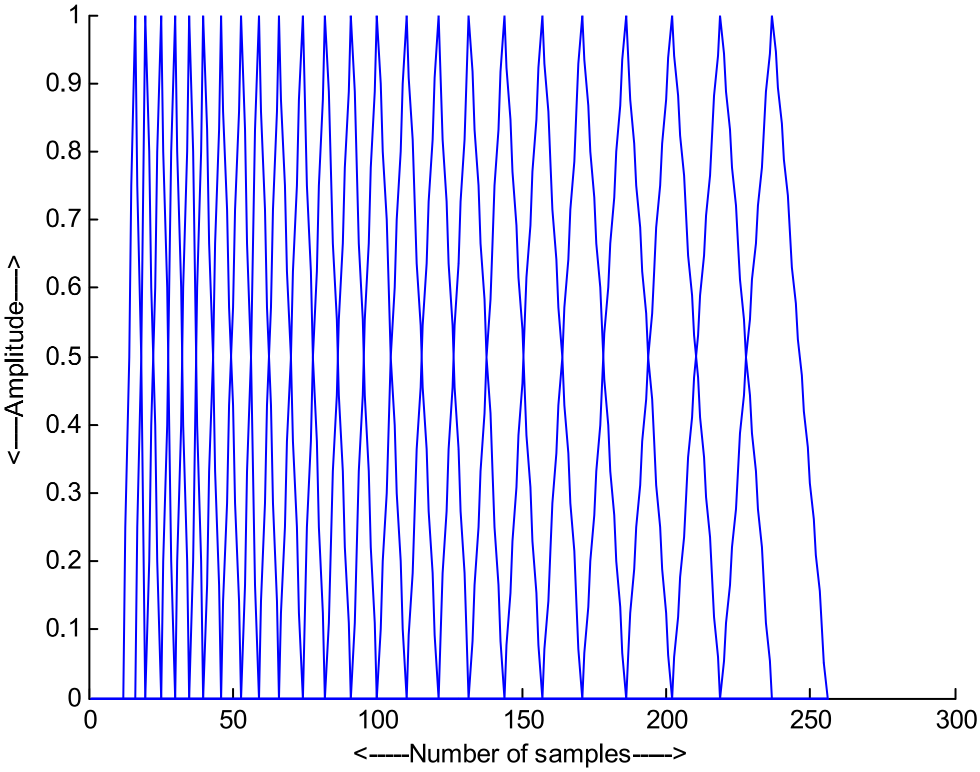

Figure 3.9: Filter bank at 8KHz

The equation for calculating Mel frequencies is

$$\text{Fmel} = 1125 * \log_e [1 + \frac{f(linear)}{700}] \qquad \text{---------------- (3.10)}$$

Where f (linear) is the linear frequencies and using above equation the range of Mel frequency is determined.

3.7 CEPSTRUM

Cepsturm analysis is a method that enables us to find out whether a given signal contains some periodic elements in it.

Calculating the cepstrum helps in multiple ways as it first compresses the dynamic range of the values thus reducing required storage space. In turn, the frequency estimates also becomes less sensitive to variations in input either due to power or noise.

Mathematically cepstrum is defined by equation 3.11.

$$c[n] = F^{-1}(\log|F\{x[n]\}|) \qquad \text{---------------} \ (3.11)$$

where F stands for Fourier transform and F^{-1}stands for inverse Fourier Transform. The log spectrum can be treated as a waveform and thus it is subjected to further Fourier analysis giving rise to cepstrum. The independent variable of cepstrum is nominally time but is interpreted as frequency since DCT is involved.

The calculation of cepstrum involves using only the magnitude and discarding the phase information. Also cepstrum coefficients are uncorrelated to a large extent thus their comparison becomes a lot easier.

3.8 MEL COEFFICIENTS

After performing the above described steps the final outcome is a set of numbers. These numbers are known as Mel Coefficients. For each frame we get a distinct set of Mel Coefficients which are then clubbed together to get a feature matrix.

In the algorithm implied during the complete Feature Extraction process, 37 Mel Coefficients are obtained. The First 12 coefficients are known as Cepstral coefficients. These coefficients mainly store the information found in the spectral domain of the signal.

The next 12 coefficients are called as delta coefficients. Delta coefficients describe the rate of change in the formants and the Cepstral coefficients. Double Delta coefficients comprise the next 12 numbers. These are obtained by calculating the slope of change in delta coefficients.

Apart from the above three categories of the coefficients, an energy function is also used. The energy function calculates the energy of the signal clipped out in the frame. Together these coefficients help in recognizing a signal.

The multiple set of coefficients from different frames are then clubbed together to form what is called as a Feature Matrix. The Feature Matrix is the end result of the whole MFCC algorithm. The matrix is then used as input for the DTW algorithm for Feature matching purposes.

CHAPTER 4
FEATURE MATCHING

4.1 INTRODUCTION

Feature Matching is the second phase of speech recognition. The term matching signifies that we have to compare and evaluate similarity between the stored signal in the database and the input utterance by the speaker. The similarity is calculated based upon the degree of co-relation between the features of the two signals being compared. The correlation depends on many factors during recording such as noise in surroundings, condition of speaker and the rate at which speaker pronounce the words. The input signal's features are extracted from the method described in the previous chapter and are compared with each set of signals stored in the database. Based upon the algorithm being utilised, each set of stored signal will result out in a comparison parameter which will be then used to find out the spoken word.

4.2 DYNAMIC TIME WARPING

One of the well-known techniques, DTW is used to calculate an optimal alignment between any two temporal sequences. This algorithm first calculates distance between each point in one signal to every other point in the second signal. This results out as matrix of distance between the signals. The algorithm then finds the least weighted path between from the matrix.

The whole process can also be seen as a function of non-linear warping between the signals. A non-linear alignment produces a more intuitive similarity measure allowing similar signals/shapes to match even if they are out of phase in the time axis. Normal distance calculation between any two points is known as Euclidean distance. It does not account for any time variations or phase difference between any two signals. As such it acts a poor measure of similarity for feature matching process. But since voice signal changes at a very fast rate and is never same exactly twice Euclidean distance will not suffice.

From the distance matrix calculated, there will be a fair chance that large number of paths exists giving the same weighted/cost path.

4.3 DTW CONSTRAINTS

The selection of path depends upon several restrictions which have been described in the following sections. All of these restrictions help in the reduction of search space under DTW algorithm.

4.3.1 MONOTONICITY

Monotonicity is the first restriction in DTW algorithm. It states that the least weighted path must not go back in time index. It signifies that the path must go in forward direction only and must exclude all the other previous point from its calculations.

This condition guarantees that same features are not repeated in the alignment path used.

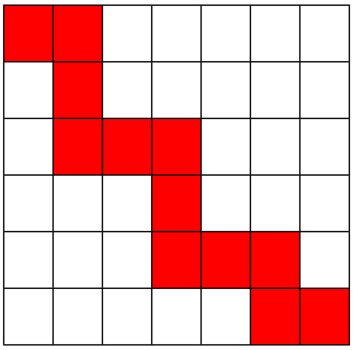

Figure 4.1: Distance matrix after DTW algorithm

The figure above shows a distance matrix between two signals. I and J are the indices of the two signals. Monotonicity condition states that J_{s-1}<J_s and I_{s-1}< I_s, where s is the indices of the two signals. In figure 3.1 the red path shows the least weighted path within the matrix.

4.3.2 CONTINUITY

Continuity restriction is the second restriction in the DTW algorithm. This condition states that the alignment path does not jump in the time index. In other terms it signifies that difference between the present point and the previous point's index must be only one.

In mathematical terms I_s-I_{s-1}=1 and J_s-J_{s-1}=1. Having this condition in the algorithm ensures that no important features are omitted from either of the two signals.

4.3.3 BOUNDARY CONDIITON

Boundary condition dictates the start and end points of the alignment path. The path selected must start from the top left corner of the matrix and must end at the bottom right corner. In mathematical terms, let N and M be the length of the two signals. Then boundary condition implies that $I_1=1$ and $I_k=N$, $J_1=1$ and $J_k=M$.

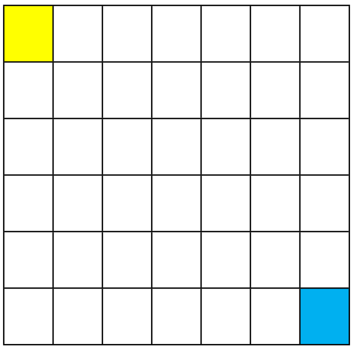

Figure 4.2: Start and End point of Algorithm

Figure 4.2 shows the start and end positions of the optimal path. The points highlighted in yellow are the start point and the blue is the end point.

Boundary condition ensures that the least weighted path selected does not give higher preference to either of the signal but considers them equally. Thus features form both the sequences are selected for proper comparison.

4.3.4 DIAGONAL CONSTRAINT

Diagonal constraint can also be seen as a window imposed upon the distance matrix. The window is in form of two diagonals running parallel to each other across the matrix. A good alignment path is unlikely to wander too far from the diagonals.

Figure 4.3 depicts the required diagonal constraint in DTW algorithm. The lines highlighted in green are the diagonals restricting the path selected as shown in red.

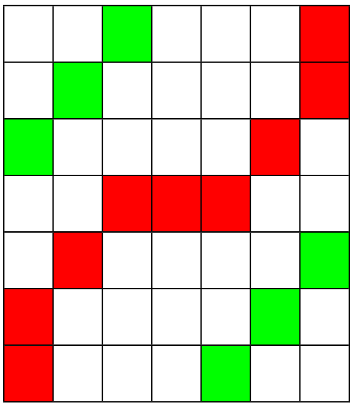

Figure 4.3: Diagonal Constraints in DTW.

The use of this restriction helps in ensuring that the path selected does not try to skip different features while traversing across the matrix and does not get struck at similar features. Thus diagonal constraint ensures that path travels across the whole matrix.

Mathematically diagonal constraint can be written as $|I_s – J_s|< r$, where r is the distance between the two diagonals.

4.3.5 SLOPE CONDITION

Slope condition restricts the slope of the alignment from being either too steep or too shallow. It states that number of continuous steps in either horizontal or vertical direction has to be limited. Doing so prevents that very short parts of either of the two sequences are not matched to very long parts of the other sequence.

It also states that after maximum steps in any of two directions are achieved, one must step in other direction for optimal path alignment.

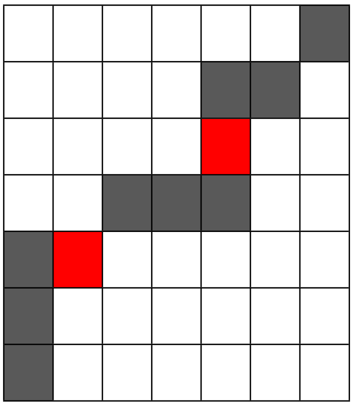

Figure 4.4: Slope Constraint in DTW

Figure 4.4 shows the slope constraint applied in DTW algorithm. The portion highlighted with grey is the optimal path. The red part shows the point where the slope is changed forcibly after maximum allowed slope in either direction is achieved.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 BASIC SETUP

The setup for executing the described algorithms is as follows

- Maximum Frequency selected for recorder……………………………… 4KHz
- Sampling Frequency.. 8KHz
- Scaling factor in Pre-emphasis filter…………………………………... 0.85
- Start frequency Of filter Bank…………………………………………. 200Hz
- End Frequency of filter Bank………………………………………….. 8KHz
- Window selected…………………………………………. Hamming Window
- Window Length……………………………………………………... 256
- Length of overlap in between window…………………………………. 100
- Size of FFT…………………………………………………………. 256
- Number of features extracted…………………………………………… 37
- Threshold selected for feature matching……………………………….. 3.5

5.2 RECOGNITION RESULTS

Table 5.1 shows the results of recognizing alphabets using the MFCC and DTW algorithms. Sampling Frequency was kept to be 8 KHz and comparison threshold was kept around 1.5.

ALPHABET	A	B	C	D	E
A	1.05	3.43	2.81	7.5	1.64
B	3.36	1.012	2.9	1.67	1.7
C	3.33	3.6	1.11	2.64	1.32
D	3.1	1.83	2.01	0.6	1.85
E	2.58	3.2	1.83	4.003	0.511

Table 5.1: Comparison Parameters of alphabets

Each of the alphabets was recorded 5 times. As such the initial database contained the coefficient matrices of 25 alphabets. During comparison, the spoken alphabet was compared with first five instances and then its average was calculated giving out the comparison parameter.

The degree of accuracy depends upon the surrounding noises while recording, sampling frequency and the size of each frame. Also the similarity between alphabets having similar phonemes decreases the comparison parameter which results in ambiguity among the results. The speaking rate of the speaker also plays an important role while comparison. Too much deviation from normal behaviour in speed while entering the signal will degrade the quality of results.

Table 5.2 shows the results of recognition of three different commands, left right and down.

COMMNADS	Left	Right	Down
Left	0.904	2.825	4.97
Right	7.884	2.137	4.87
Down	9.01	8.08	1.924
ACCURACY	80%	80%	100%

Table 5.2: Commands Recognition 256 point framing

Sampling frequency was kept 8 KHz and framing was of 256 points. In time domain the frame size accounts 32ms. The overlap size was taken as 100 points, 12.5ms in time domain.

Each of the command had three instances stored in the database. The comparison parameter came after calculating the average of the comparison output with each of the three stored signals. The entire comparison process was repeated five times and the accuracy came out as given in the table.

Table 5.3 depicts the results of recognition after the frame size was changed to 512 points from 256 points. Sampling frequency was kept 8 KHz (64ms) and overlap between frames was taken to be 100 points.

Upon comparison it is observed that the performance of the whole system degrades as size of frames is increased. The cardinal cause behind such reduction in accuracy is the change in time limit of each frame. Increase in frame size increases the time duration from 32ms to 64ms. Hence, the clipped out signal cannot be approximated as a stationery signal. The whole concept behind MFCC algorithm is that the signal has to follow stationery behaviour. In a large time duration of 64ms, the properties and frequency of the signal varies a by a large extent. As such MFCC algorithm efficiency will decrease.

COMMANDS	Left	Right	Down
Left	1.02	3.29	3.83
Right	4.42	1.84	5.44
Down	1.50	6.90	0.73
ACCURACY	80%	60%	80%

Table 5.3: Commands Recognition 512 point framing

For using the higher framing size, the solution is to use a higher sampling frequency. Using a higher sampling rate will allow larger frame size but redundancy will increase. Maximum frequency that an average human can produce is around the range of 3.5 KHz to 4 KHz. Thus following the nyquist criteria, best sampling frequency is 8 KHz.

Figure 5.1 and 5.2 visualizes the output of the DTW algorithm. The former figure shows the comparison between two similar alphabets while the latter is the comparison between different alphabets.

The red line shows the distance between two signals, vectors 1 and 2. A straight line from one corner to the other signifies that similarity between the two signals is high. If the line follows a zigzag pattern, the two signals being compared are different from each other.

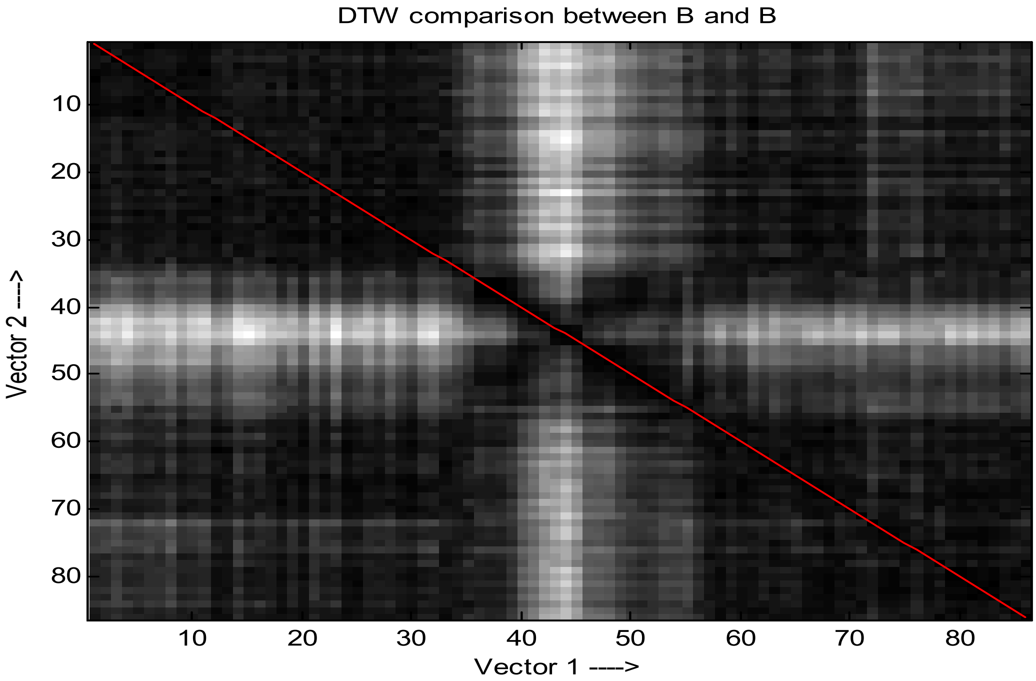

Figure 5.1: DTW Comparison Between B and B

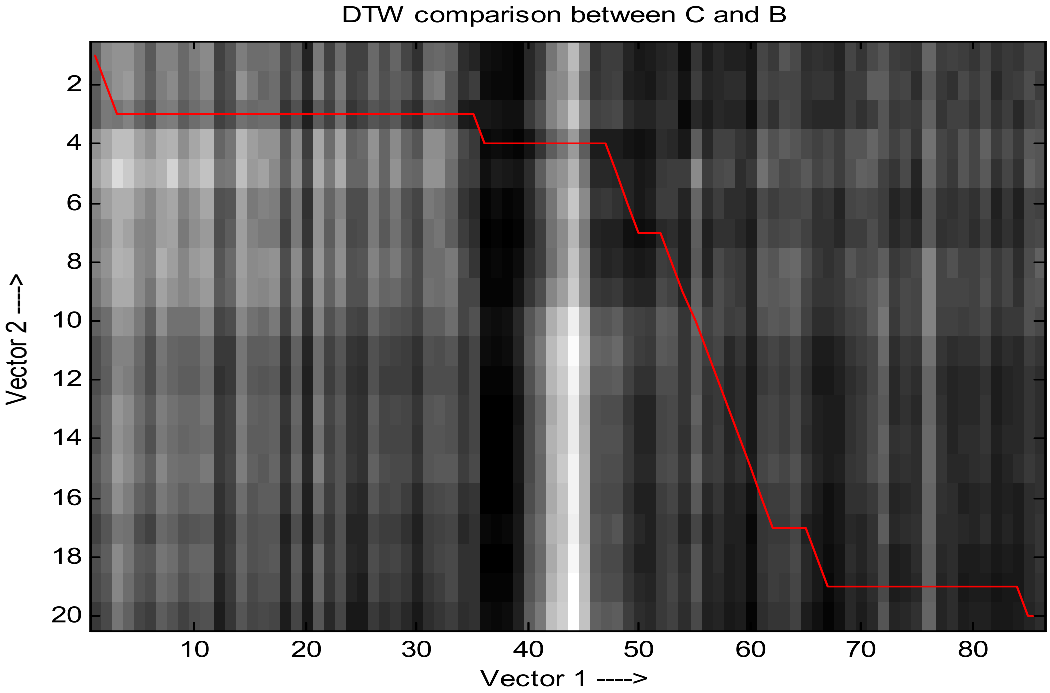

Figure 5.2: DTW Comparison Between C and B

Figure 5.1 shows that the alphabet b have been matched perfectly with itself but when alphabet b is matched with alphabet c the matching is not observed which can be observed from figure 5.2.

CHAPTER 6

REALISATION USING ROBOTIC MODEL

6.1 INTRODUCTION

Arduino Uno is used to construct a Robotic car in which four 300 rpm servo motors are operated using 9-12V DC battery. Also L293D motor driver Board is used for configuring the motor with the Arduino Uno. Bluetooth module HC-05 is used for the communication between the robot and the computer. The RC car is monitored using the commands spoken by the user which will be first recorded using a microphone and then that recorded signal is processed. Upon recognition of the spoken command the command is transferred to the RC car using HC-05 Bluetooth module which is then carried forward to the L293D motor driver module which carries out the respective command according to the Arduino code which is designed using the Arduino ISE design software and burned on to the Arduino Uno.

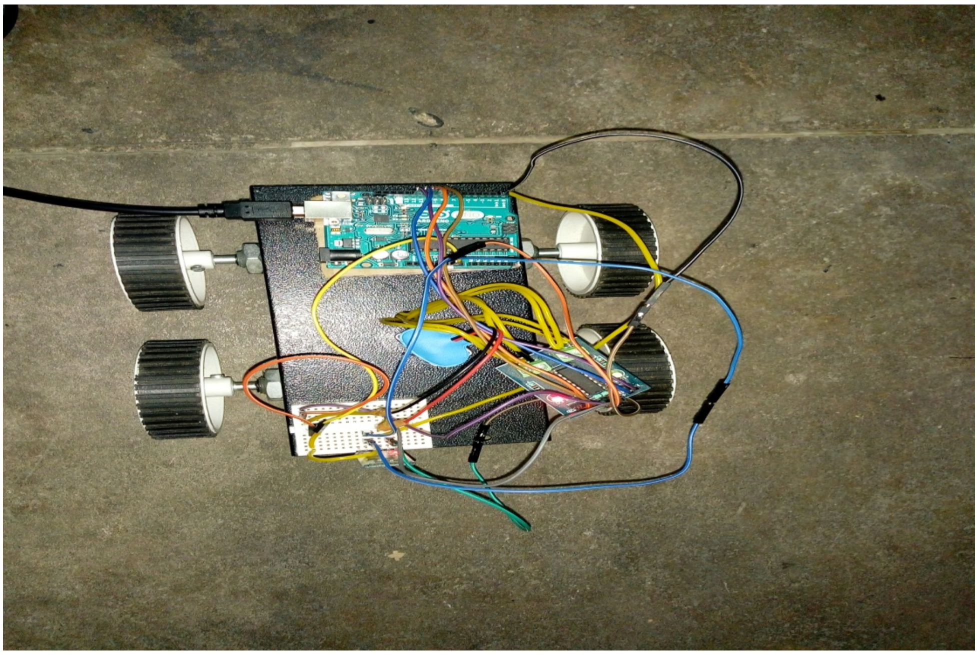

Figure 6.1: Robotic Model Controlled using Voice Signals

6.2 ARDUINO UNO

Arduino Uno is a broadly used microcontroller board based on ATmega328P microcontroller. This microcontroller board has various analog and digital input/output pins which can be used to link this board to additional boards and other circuitries. The microcontroller is powered either using a 9V DC battery or a USB cable.

6.2.1 HARDWARE SPECIFICATIONS

Arduino Uno has 14 digital input/output pins, 6 analog pins, a USB connection, a power jack and a reset button. The clock speed of the microcontroller is 16 MHz which is generated using a quartz crystal oscillator. The DC input/output pins can handle a maximum current of 20 mA and a DC current for 3.3V pin is 50 mA. This board has a flash memory of 32 KB of which 0.5 KB is used by bootloader. It also contains 2 KB of SRAM and 1 KB of EEPROM. The input voltage limits for this microcontroller is 6 to 20V but it is preferable to use it in between 7 to12V. This board provides facilities for serial communication with the computer or other microcontroller using digital pin 0(receiver) and 1(transmitter).

6.3 BLUETOOTH MODULE (HC-05)

HC-05 is a Bluetooth Serial Port Protocol module used for wireless serial connection setup. This module has 6 pins and their functions are as follows:

KEY / EN

It is used to Enable Bluetooth module in AT commands mode. If this pin is set high, then this module will work in command mode else it will work in data mode by default. The standard baud rate of HC-05 in command mode is 38400 bps and 9600bps in data mode.

VCC

This pin is connected to 3.3V.

GROUND

This pin is connected to ground.

TXD

The data which is received by the Bluetooth module is transmitted out serially on TXD pin.

RXD

Received data will be transmitted wirelessly by Bluetooth module.

STATE

It tells whether module is connected or not.

6.4 L293D MOTOR DRIVER MODULE

It is a module which is used to control the speed and direction of the motor simultaneously. This module is designed using a 16 pin L293D IC. It can provide bidirectional drive currents of upto 600 mA at voltages from 4.5 to 36V.

6.5 WORKING OF THE ROBOTIC MODEL

The movement of the robotic model is governed using the 300rpm dc motors which are controlled using L293D motor driver module which is powered using 9V dc battery. The Arduino board is powered using a power bank whose output is connected to the Bluetooth module using jumper wires. The Bluetooth module is set into the communication mode and the transmitter of the Bluetooth module is connected to the receiver of Arduino board which carries commands to be executed onto the Arduino board. For setting up the communication with the computer (MATLAB) initially the computer is connected with the HC-05 Bluetooth module then the communication is setup with the help of MATLAB by carrying out the following process:

- The first step is to figure out the name of the Bluetooth device using the command **instrhwinfo('bluetooth')** for checking out the available Bluetooth devices.
- Then we figure out the remote Id of the Bluetooth module using **ans.RemoteNames** command.
- Then we create an object for initiating the communication with the HC-05 bluetooth module.
- Further we use **fopen(Object name)** to initiate communication.
- Then the recognized commands are sent to the Adruino for execution in digitized format according to the code embedded into the Arduino.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

MFCC algorithm produces the best results when the size of frames and windows are limited to 15ms to 25 ms. The robotic model was successfully controlled using four commands, accuracy of which have been described in results and discussions. Hamming window proved to give better results as when compared when other windows were used.

MFCC algorithm used for implementing the concept of speech recognition has some setbacks despite of its good and accurate results. Primarily, MFCC depends on a database for recognizing speech signals. Any signal other than the ones stored in the database will not be recognized. Secondly, the algorithm depends upon the clarity of the signal being stored and as such is prone to noise interference. Requirement of ideal recording is a must for the algorithm to give better results. Lastly, the MFCC algorithm is a user dependent algorithm. As such, it can recognize commands of only those individuals whose voice signals have been stored in the database. Thus it cannot recognize voices of multiple users using the same database.

Speech processing has a huge potential in becoming an important factor for human machine interaction in the near future. A Speaker Independent speech recognition system can be proposed which uses MFCC and Artificial Neural Networks and the parameters of MFCC and Artificial Neural Networks can affect each other for better performance. A combination of Probabilistic neural network and Recurrent neural network followed by MFCC can be used for better accuracy.

REFERENCES

1. Hasina Shaikh, Dr. Luis C. Mesquita, Sufola Das Chagas Silva Araujo, “Recognition of Isolated Spoken Words and Numeric using MFCC and DTW”, International Journal Engineering science and computing, April 2017,Volume 7, Issue No. 4
2. Kiran R., Nivedha K., Pavithra Devi S., Subha T., “Voice and speech recognition in Tamil language”, Computing and Communications Technologies (ICCCT), 2017 2nd International Conference
3. Shivanker Dev Dhingra, Geeta Nijhawan , Poonam Pandit, “ISOLATED SPEECH RECOGNITION USING MFCC AND DTW”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8, August 2013
4. J. Kim, A. Lammert, P. Ghosh, and S. S. Narayanan, “Spatial and temporal alignment of multimodal human speech production data: real time imaging, flesh point tracking and audio,” in IEEE International Conference on Acoustics Speech and Signal Processing, 2013, pp. 3637–3641
5. K. Joshi, N. Kolhare, and V. M. Pandharipande, “Implementation of Speech Recognition System using DSP Processor ADSP2181,” *Int. J.*Electron. *Signals Syst.*, vol. 1, no. 3, 2012
6. Hamed Azami, Karim Mohammadi, Behzad Bozorgtabar, “An Improved Signal Segmentation Using Moving Average and Savitzky-Golay filter”, Journal of Signal & Information Processing, February 2012
7. Vikramjit Mitra, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein, “Articulatory information for noise robust speech recognition,” IEEE Transactinos on Audio, Speech and Language Processing, vol. 19, no. 7, 2011
8. LindasalwaMuda, MumtajBegam and I.Elamvazuthi, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic time warping techniques”, Journal of Computing, March 2010, Volume 2, Issue 3
9. Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar, “Self-learning Vector Quantization for Pattern Discovery from Speech”, Interspeech Brighton, 2009

10. D. Jurafsky and J. H. Martin, "Speech and Language Processing", 2nd ed. Pearson Education International, 2008
11. C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 6, pp. 623–635, 1980
12. E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in SIAM Conference on Data Mining, 2001
13. Xuedong Huang and Kai-Fu Lee, "On Speaker-Independent, Speaker-Dependent and Speaker-Adpaptive Speech Recognition", IEEE transaction on Speech and audio processing, Volume 1, No. 2, April 1993