A Project Report on

Emotion Recognition Using MFCC & DTW

Submitted by

SIDDHARTH BANERJEE (ROLL NO. 7)

AKSHATA BHAT (ROLL NO. 10)

UMANG BHATT (ROLL NO. 11)

PANKAJ CHAUHAN (ROLL NO. 18)

in fulfillment of

MINI-PROJECT II

in

Electronics & Telecommunication Engineering

Under the Guidance of

Mr.Inderkumar Kochchar

Department of Electronics and Telecommunication Engineering

St. Francis Institute of Technology, Mumbai

University of Mumbai

(2016-2017)

ABSTRACT

Speech is a vocalized form of human communication. Emotions exert an incredibly powerful force on human behavior. Emotion plays an important role in a person's approach to a particular situation at that particular time. Unable to understand a person's emotion in a particular situation may cause a failure of communication. Thus recognizing the emotion becomes one of the important aspects.

Recently increasing attention has been directed to the study of emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance.

This project mainly aims to classify 5 emotions namely sad, happy, anger, surprise and neutral. The input signal is divided into various frames of 20ms and features are extracted from each frame using MFCC. Later on, DTW is used for classification of emotions.

CERTIFICATE

This is to certify that the project entitled **“Emotion Recognition Using MFCC & DTW”** is a bonafide work of **“Siddharth Banerjee(Roll No. 7), Akshata Bhat(Roll no. 10), Umang Bhatt(Roll no. 11) and Pankaj Chauhan(Roll no. 18)”**submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Electronics and Telecommunication Engineering**.

_____

Mr.Inderkumar Kochchar

Internal Guide

_____ _____

Internal Examiner External Examiner

ACKNOWLEDGEMENT

It is indeed a matter of great pleasure and proud privilege to be able to present this project on **“Emotion Recognition Using MFCC”**

The completion of the project work is a milestone in a student’s life and its execution is inevitable in the hands of the guide. We are highly indebted the project guide **Mr. Inderkumar Kochchar** for her invaluable guidance and appreciation for giving form and substance to this project. We would also like to thank **Mr.Santosh Chapaneri** for his guidance and support. It is due to his enduring efforts; patience and encouragement, which has given a sense of direction and meaning to this project and ultimately made it a success.

We would like to express our sincere thanks to the staff members for their co-operation.

We would also like thank the Lab assistants for their help and co-operation.

We would wish to thank the non - teaching staff and our friends who have helped us in some way or the other.

Contents

List of Figures

List of tables:

CHAPTER 1:

INTRODUCTION

Chapter 1: Introduction:

Emotion Recognition using MFCC and DTW is used to determine the emotion of the speaker from his voice. The speech signal given by the speaker is compared with the database which consists of various sentences which define the various emotions of the language.

1.1. Motivation:

Emotion Recognition is an important aspect in today's world wherein machine learning is a trending and growing technology. It involves processing of the input signal and the evaluation of the emotion by extracting the features of the input voice signal.

1.2. Scope of Project:

As mentioned earlier, Emotion Recognition is the basic step involved for machine learning and deep learning. We are interested in training the machines in such a way that they can provide service which is as good as that provided by humans and a high efficiency of performance.

1.3. Methodology:

In order to determine the emotion of the received speech signal, we have to obtain various features like energy, power spectral density (PSD) etc. of the signal and compare it with the recorded database. We are achieving this using MFCC. Also, in order to eliminate the difference between the database and the input signal, we use a method called Time Warping.

CHAPTER 2:

LITERATURE REVIEW

Chapter 2: Literature Review:

The overall steps involved in recognition of emotion from the speech are as shown below:

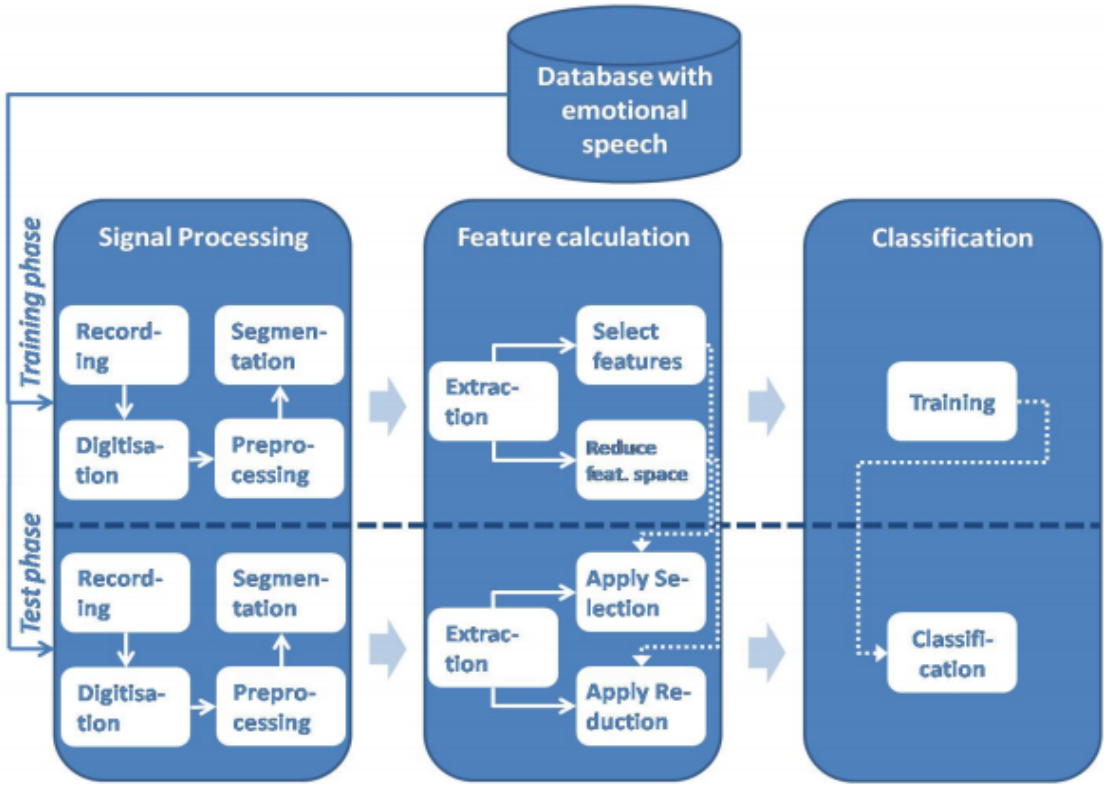


Figure 2.4: Overview of a speech emotion recognition system.

2.1. Signal Processing:

Signal processing is an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals. It uses mathematical, statistical, computational, heuristic, and linguistic representations, formalisms, and techniques for representation, modeling, analysis, synthesis, discovery, recovery, sensing, acquisition, extraction, learning, security, or forensics. **Examples:**

- Audio signal processing - for electrical signals representing sound, such as speech or music

- Digital signal processing - for the processing of digitized discrete-time sampled signals.
- Speech signal processing - for processing and interpreting spoken words.

2.2 .Voice Recognition Algorithms:

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word. The voice algorithms consist of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase.

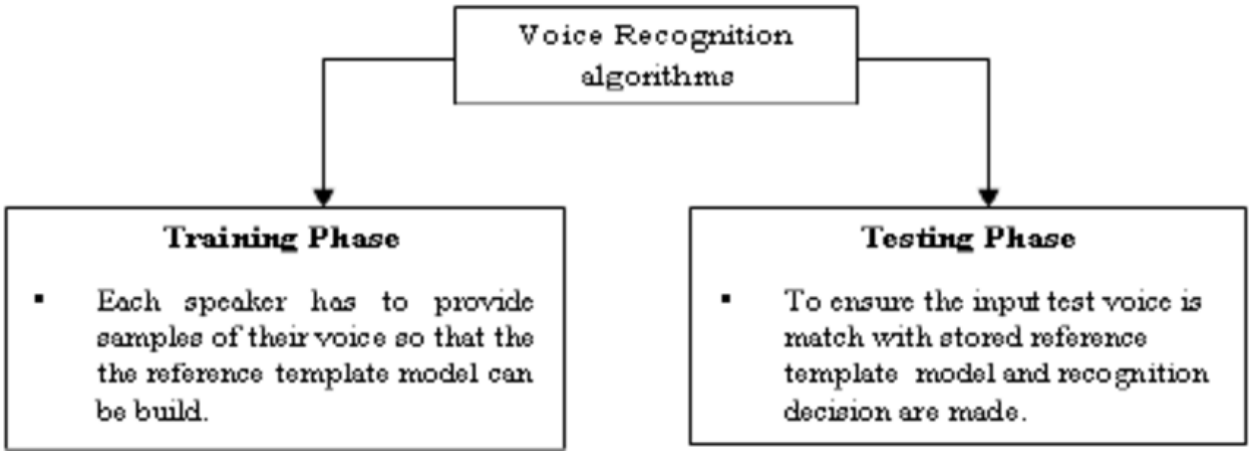


Fig 2: Voice Recognition Algorithms

2.3. Mel Frequency Cepstral Coefficients (MFCC Feature Extraction):

In sound processing, the **mel-frequency cepstrum** (**MFC**) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (**MFCCs**) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more

closely than the linearly-spaced frequency bands used in the normal cepstrum. There are 13 mel frequency cepstral coefficients. The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f/700) \qquad (1)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \qquad (2)$$

2.4. Dynamic Time Warping:

In time series analysis, **dynamic time warping** (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in speed. For instance, similarities in walking could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation. DTW has been applied to temporal sequences of video, audio and graphics data — indeed, any data which can be turned into a linear sequence can be analyzed with DTW. A well- known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. Also it is seen that it can be used in partial shape matching application.

In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" nonlinearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold.

In addition to a similarity measure between the two sequences, a so called "warping path" is produced, by warping according to this path the two signals may be aligned in time. The signal with an original set of points X (original), Y(original) is transformed to X(warped), Y(origina l).

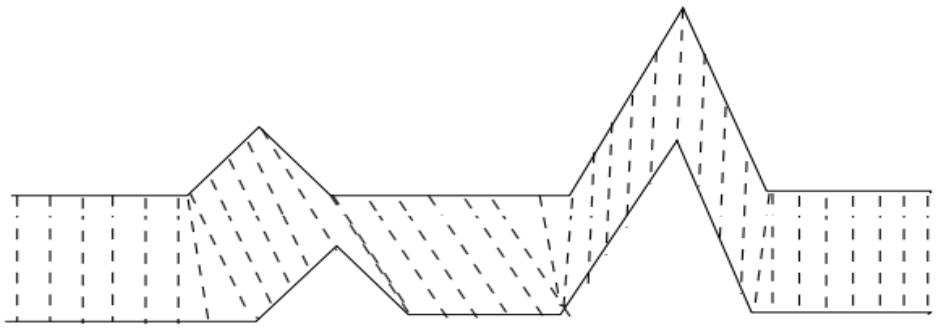


Fig 3.DTW

CHAPTER 3:

DESIGN METHODOLOGY, SOFTWARE AND HARDWARE SUPPORT

Chapter 3: Design Methodology, Software and Hardware Support.

3.1. Block diagram:

The basic block diagram of our project is as follows:

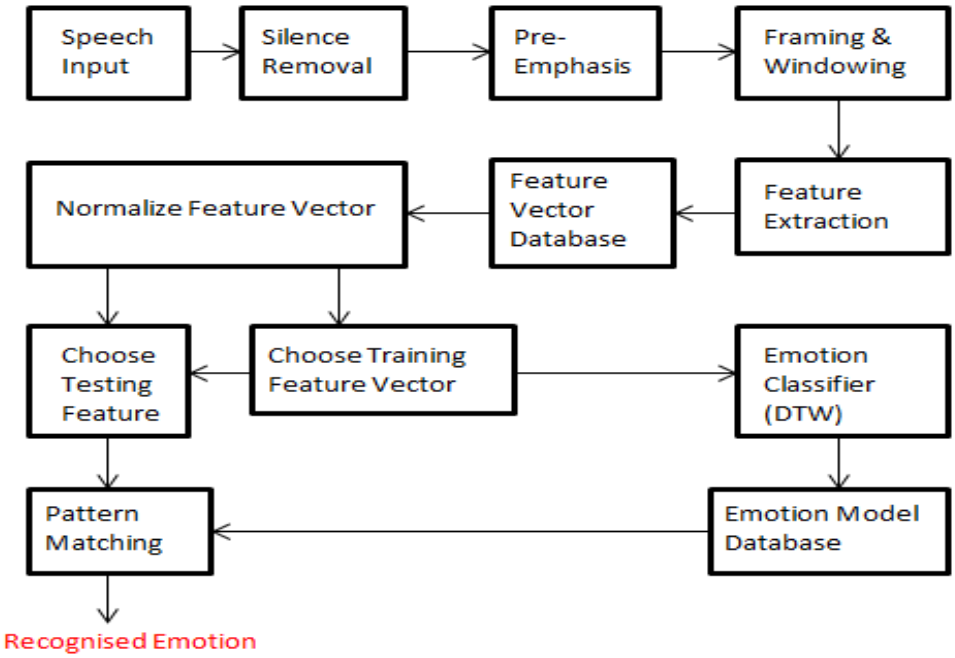


Fig4. Block diagram for MFCC Feature Extraction

Explanation:

The above block diagram consists of the following steps:

3.1.1. . Pre-Emphasis:

Pre–emphasis: This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

The speech signal s(n) is sent to a high-pass filter:

$$s_2(n) = s(n) - a*s(n-1)$$

where $s_2(n)$ is the output signal and the value of a is usually between 0.9 and 1.0. The z-transform of the filter is

$$H(z)=1-a*z^{-1}$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants. The next example demonstrates the effect of pre-emphasis.

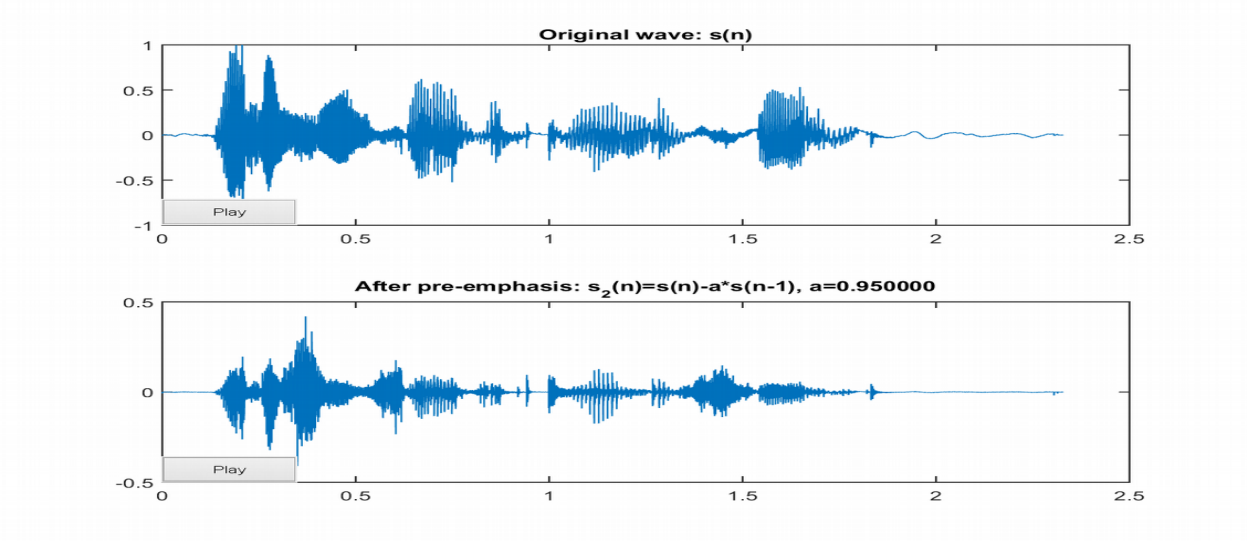


Fig5. Pre-emphasis

3.1.2. Framing:

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. 25ms is standard. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). The typical values used are M=100 and N=256.

Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is 320/16000 = 0.02 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second.

3.1.3. Windowing:

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

$$Y(n) = X(n) \times W(n) \quad \ldots\ldots \quad (3)$$

If the window is defined as W (n), 0 ≤ n ≤ N-1 where N = number of samples in each frame Y[n] = Output signal X (n) = input signal W (n) = Hamming window, then the result of windowing signal is shown below:

$$w(n, a) = (1 - a) - a\cos(2 * pi * n/(N-1)),\ 0 \leqq n \leqq N-1$$

Different values of a corresponds to different curves for the Hamming windows shown :

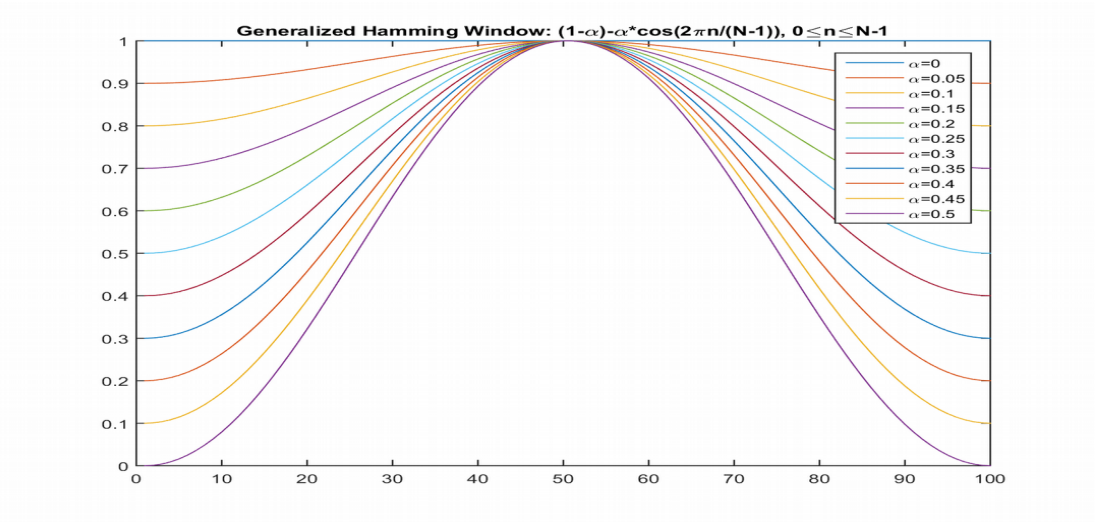


Fig.6. Hamming Window

3.1.4 Fast Fourier Transform:

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame.

When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies:

a. Multiply each frame by a Hamming window to increase its continuity at the first and last points.

b. Take a frame of a variable size such that it always contains a integer multiple number of the fundamental periods of the speech signal.

The second strategy encounters difficulty in practice since the identification of the fundamental period is not a trivial problem. Moreover, unvoiced sounds do not have a fundamental period at all. Consequently, we usually adopt the first strategy to multiply the frame by a Hamming window before performing FFT. The following example shows the effect of multiplying a Hamming window.

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain.

This statement supports the equation below:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \quad \ldots\ldots (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

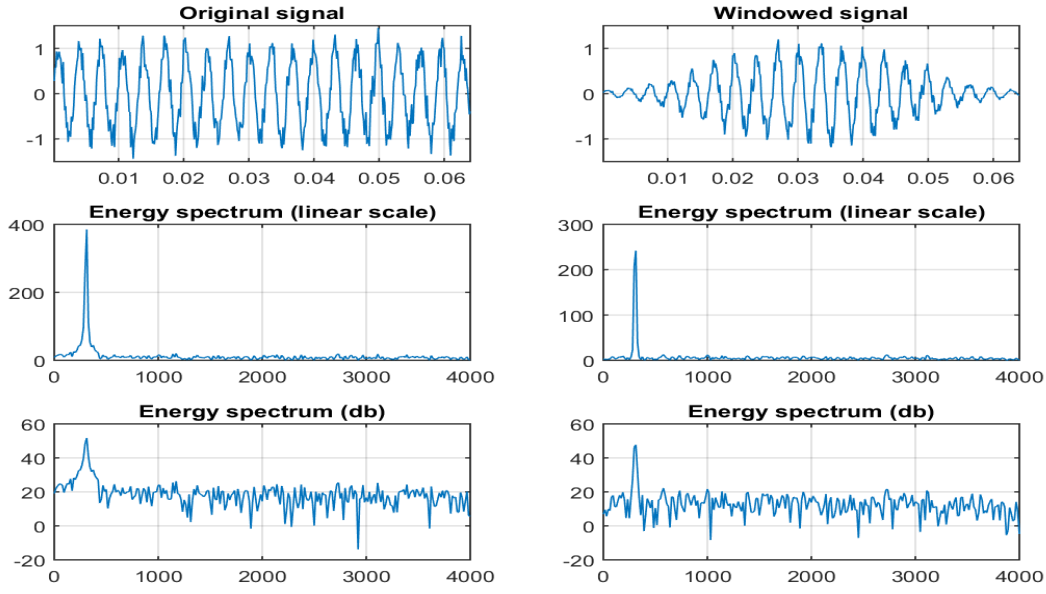


Fig7. Windowing

In the above example, the signal is a sinusoidal function plus some noise.

Without the use of a Hamming window, the discontinuity at the frame's first and last points will make the peak in the frequency response wider and less obvious. With the use of a Hamming, the peak is sharper and more distinct in the frequency response.

3.1.5 Mel Filter Bank Processing:

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. This is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate from step 2. Our filterbank comes in the form of 26 vectors of length 257 (assuming the FFT settings from step 2). Each vector is mostly zeros, but is non-zero for a certain section of the spectrum. To calculate filterbank energies we multiply each filterbank with the power spectrum, then add up the coefficients. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filterbank. For a detailed explanation of how to calculate the filter banks see below. Here is a plot to hopefully clear things up:

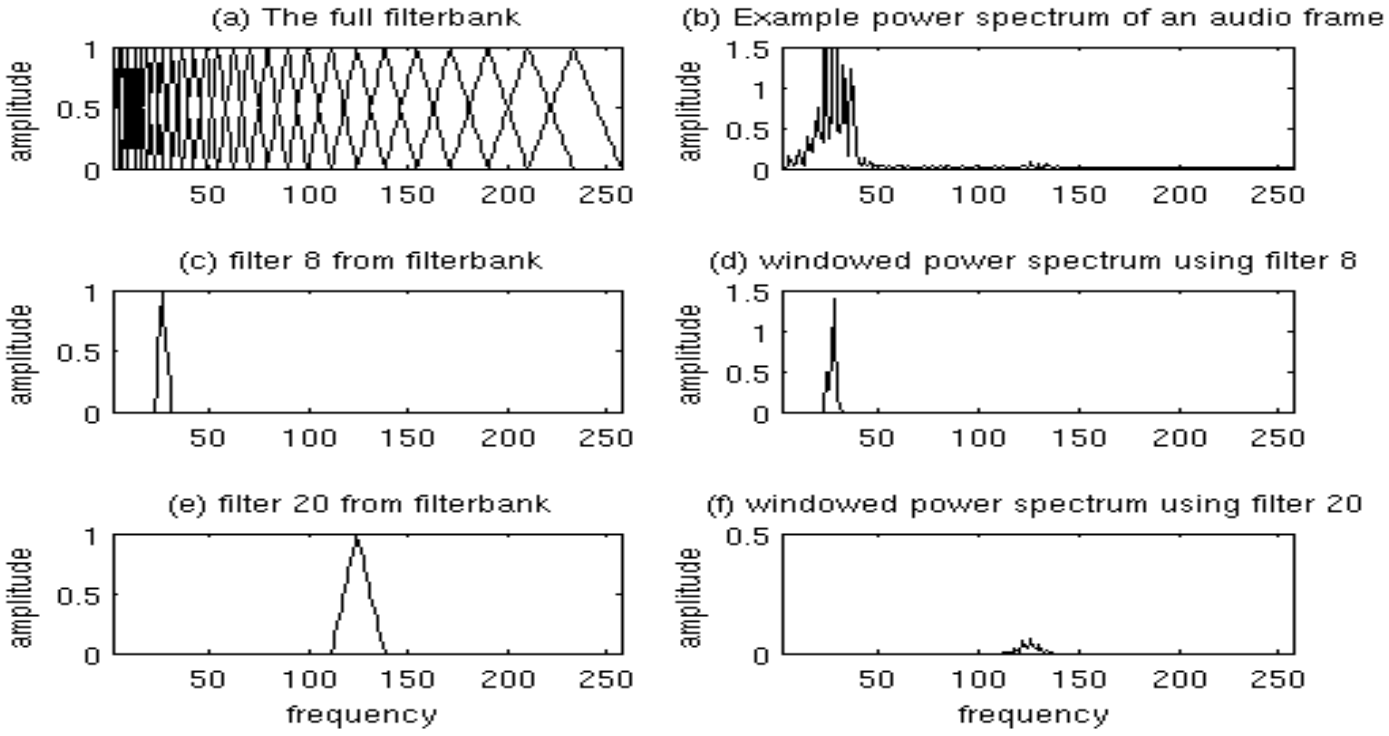


Fig.8. Calculation of power spectrum

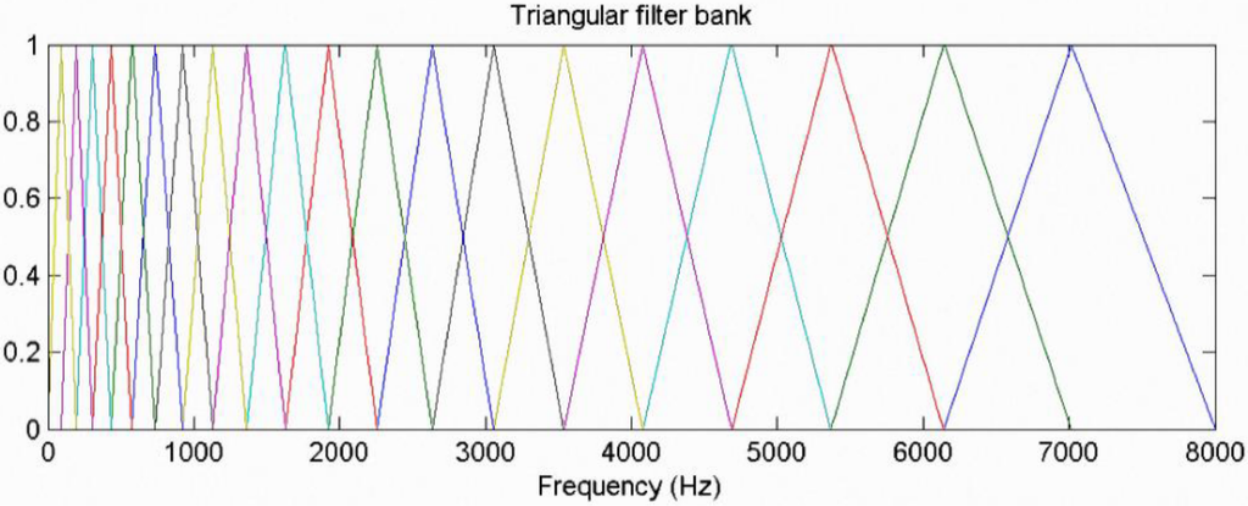


Fig8: Mel Filter Bank

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

F (Mel) = [2595 ∗ log 10 [1 + f /700]] ..... (5)

3.1.6 Discrete Cosine Transform:

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The set of coefficients (of MFCC) is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

To take the Discrete Fourier Transform of the frame, perform the following:

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \qquad 1 \leq k \leq K$$

Where ***h(n)*** is an ***N*** sample long analysis window (e.g. hamming window), and ***K*** is the length of the DFT. The periodogram-based power spectral estimate for the speech frame ***$s_i(n)$*** is given by:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

This is called the Periodogram estimate of the power spectrum. We take the absolute value of the complex Fourier transform, and square the result. We would generally perform a 512 point FFT and keep only the first 257 coefficients.

3.2. Dynamic Time Warping:

Dynamic time warping (DTW) is a time series alignment algorithm developed originally for speech recognition. It aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match (according to a suitable metrics) between the two sequences is found.

Consider two sequences of feature vectors:

$$\mathcal{A} = a_1, a_2, \ldots, a_i, \ldots, a_n$$
$$\mathcal{B} = b_1, b_2, \ldots, b_j, \ldots, b_m$$

The two sequences can be arranged on the sides of a grid, with one on the top and the other up the left hand side. Both sequences start on the bottom left of the grid.

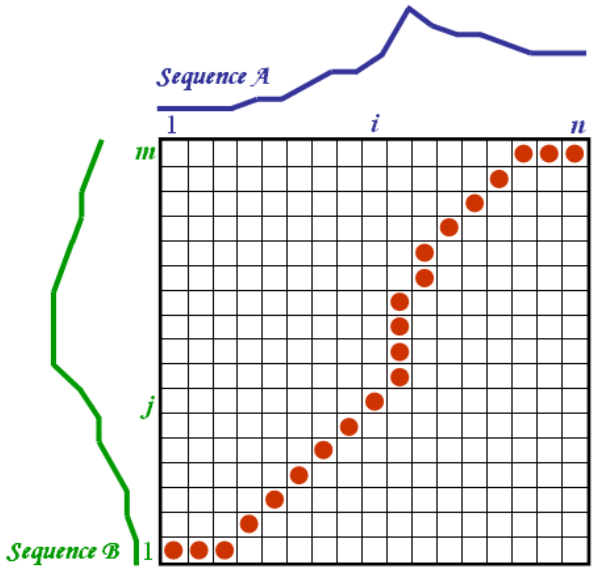


Fig9. DTW

3.3. Software used: MATLAB R2013b

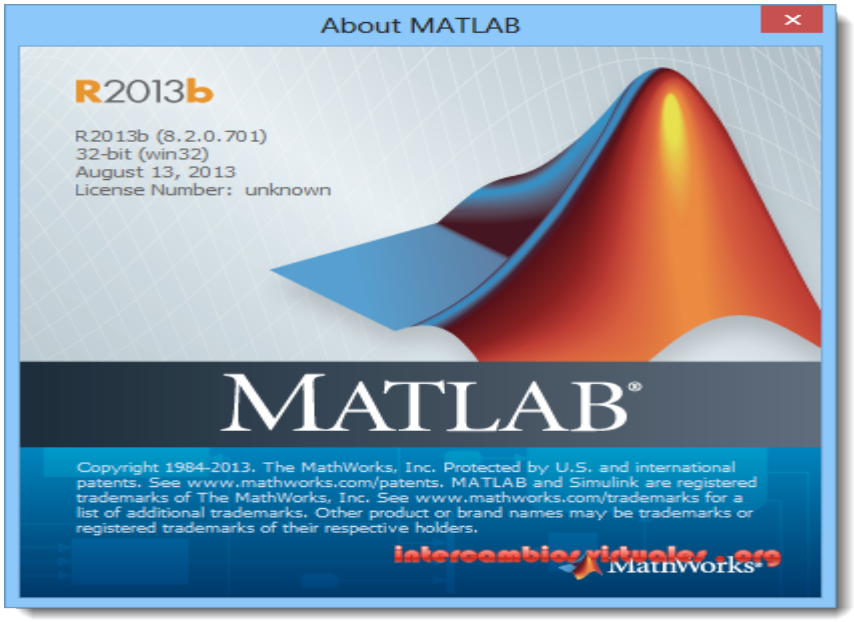


Fig10. MATLAB logo

We have performed the software simulation of our project using MATLAB R2013b.

Steps to be followed for simulation in MATLAB R2013b:

1. Open MATLAB R2013

2. Download Voicebox Toolbox for MATLAB

3. Read .wav files from the input given by user

4. Create download database for the Emotion Recognition from English Speech.

5. Write a program for calculation of the mel frequency cepstral coefficients of the recorded database as well as for the input signal.

6. Write a program which assigns the coefficients the sentences and their respective emotions.

7. Write a program to calculate the time warping of the mel frequency

cepstral coefficients of the input signal with that of the coefficients of the recorded database.

8. Depending on the result of these comparisons, display the emotion.

9. For determining the accuracy, write a program to calculate the confusion matrix.

CHAPTER 4:

Database Details

Corpus design:

We have used Surrey Audio-Visual Expresses Emotion (SAVEE) Database

The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, and KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. This is supported by the cross-cultural studies of Ekman [6] and studies of automatic emotion recognition tended to focus on recognizing these [12]. We added neutral to provide recordings of 7 emotion categories. The text material consisted of 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. The 3 common and 2 × 6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences. This resulted in a total of 120 utterances per speaker, for example:

Anger: Who authorized the unlimited expense account?
Disgust: Please take this dirty table cloth to the cleaners for me.
Fear: Call an ambulance for medical assistance.
Happiness: Those musicians harmonize marvellously.
Sadness: The prospect of cutting back spending is an unpleasant one for any governor.
Surprise: The carpet cleaners shampooed our oriental rug.
Neutral: The best way to learn is to solve extra problems.

The distribution includes a complete list of sentences.

CHAPTER 5:

SIMULATION

Chapter 5: Simulation

5.1. CODES:

1.MAIN.m: It is used to recognise the emotion of the speaker

2.SilenceRemoval.m: It removes silence from the speech signal

3.dtw.m: It is used to calculate the DTW distance between the test signal and reference signal

4.train.m: It is used to train the model.

5.melcepst.m: It is used to calculate the Mel Frequency Cepstral Coefficients (MFCC)

6.melbank.m: It is used to calculate the mel filter bank.

5.2. Results:

1. Training Phase:

We have used SAVEE database in which 4 speakers have recorded their voices. Of these 4 speakers we have selected one speaker's emotion database as the reference and we have reserved other 3 for the testing purpose.

Run **train.m** file

Command:

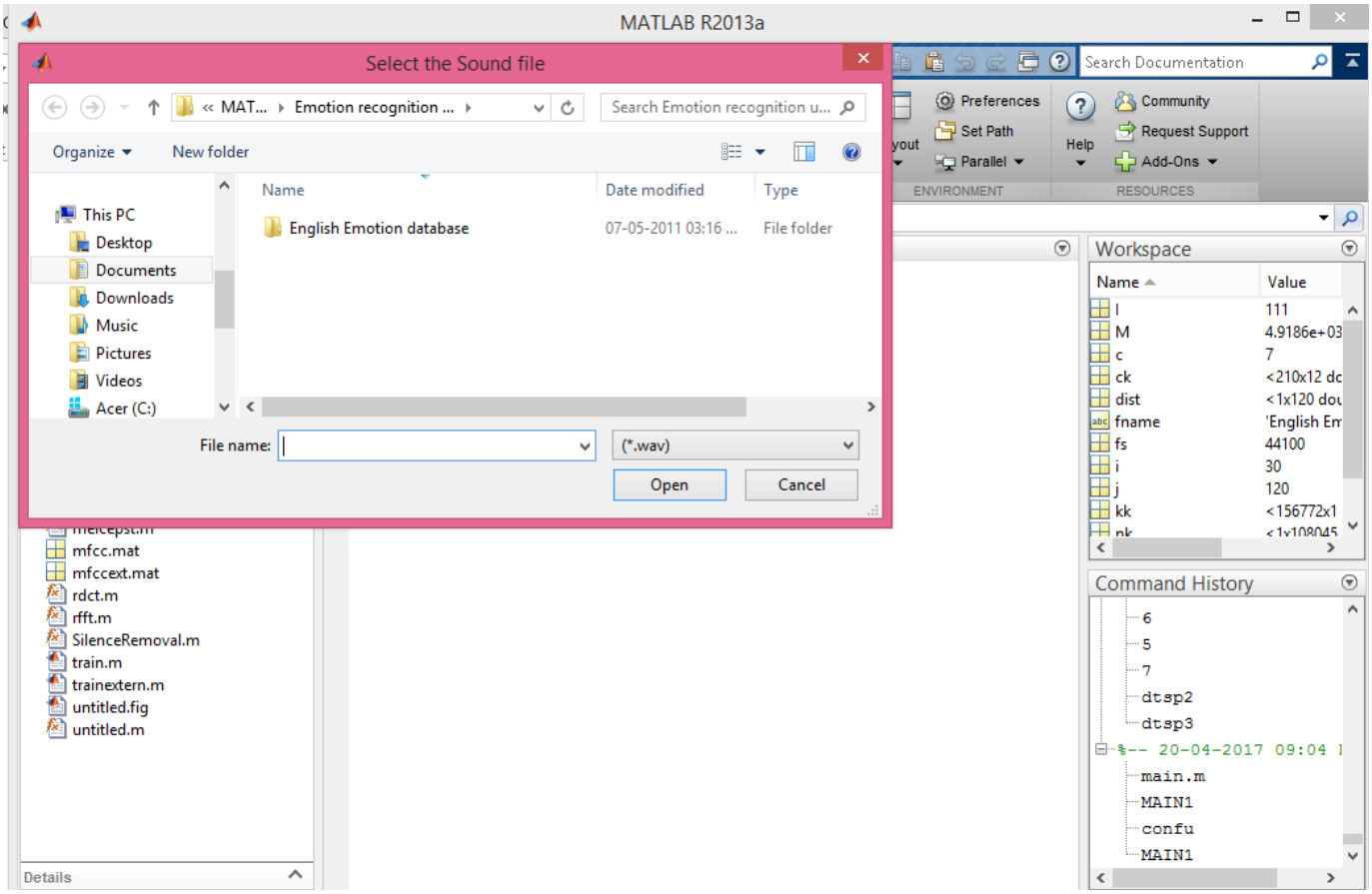
```
>>train   //It is used to train the model
```

2. Testing Phase:

For testing the emotion, we first run the **MAIN.m** file and select the desired emotion from the list of emotions.

Command:

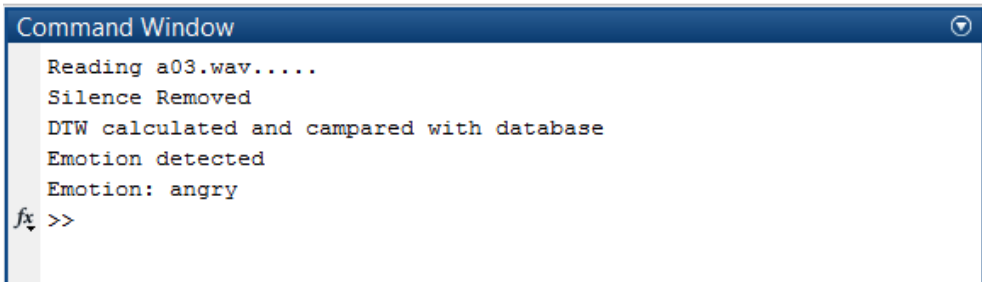
>> MAIN //It is used to test the emotion of the voice signal



Test signal: a03.wav

The selected sound file is of **‘angry’** emotion from the speaker ‘JK’

The detected emotion is angry .



Confusion Matrix Using DTW

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	53.33	-	13.33	3.33	10	-	20
Disgust	3.33	43.33	13.33	-	23.33	-	16.67
Fear	13.33	-	86.67	-	-	-	
Happy	-	-	13.33	80	-	-	6.67
Neutral	-	-	1.67	-	76.67	10	11.67
Sad	6.67	-	13.33	-	13.33	60	6.67
Surprise	-	-	20	-	-	-	80

The Average accuracy is 68.57%

CHAPTER 6:

CONCLUSION

Chapter 6: Conclusion

Conclusion/ Result Analysis:

We have implemented Emotion Recognition from voice signal for six different emotions namely happy, sad, disgust, fear, anger, surprise and neutral from the English database. The feature extraction technique that we have used is Mel Frequency Cepstral Coefficient (MFCC).

The accomplishment of this project requires various steps i.e. collecting the database, finding the features of the signals and training the system by computing a MFCC database and comparing the feature of the voice signal to be tested using DTW. After the comparison it detects and displays the emotion.

The accuracy of the obtained/recognized emotion is calculated using the concept of Confusion Matrix which assigns probability values to the events which should/should not or may occur, most likely, least likely etc. which in case of our project is getting **68.57%** as avg. accuracy. Thus, knowing the emotion of the user, the system can adapt to the user.

Future Scope:

In practical applications, MFCC and DTW are not widely used for accuracy as well as computational speed limitations. The overall human-system interaction maybe affected by this. This can be improved by using Hidden Markov Model (HMM) for further increasing the accuracy Also, we can extend this to various languages or even multilingual stage. Also, we can develop an android application for the same.

REFERENCES

References:

- Voice RecognitionAlgorithmsusingMelFrequencyCepstralCoefficient(MFCC) andDynamicTime Warping(DTW) Techniques, LindasalwaMuda, MumtajBegamandI. Elamvazuthi
- http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- http://wearables.cc.gatech.edu/paper_of_week/DTW_myths.pdf
- https://www.psb.ugent.be/cbd/papers/gentxwarper/DTWalgorithm.html
- http://www.phon.ox.ac.uk/jcoleman/old_SLP/Lecture_5/DTW_explanation.html
- http://www-db.deis.unibo.it/courses/SI-M/slides/04.5.TimeSeries2.pdf
- https://www.vocal.com/speech-recognition
- http://staffwww.dcs.shef.ac.uk/people/S.Wrigley/com326/sym.html
- https://lemonzi.files.wordpress.com/2013/01/dtw.pdf

- http://kahlan.eps.surrey.ac.uk/savee/
- http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
- Santosh Chapaneri , ”Spoken digits recognition using weighted MFCC and improved features for Dynamic time Warping” , *International Journal of Computer Applications,* vol.40,pp.975, Feb 2012.
- http://in.mathworks.com/matlabcentral/fileexchange/24583-mirtoolbox