Features Extraction from Speech for Emotion Recognition

Er. Pragati Pal[1] Ms. Dalal Fatema[2] Ms. Neha Shaikh[3] Mr. Ozair Shaikh[4]
[1,2,3,4]Affiliated to Mumbai University

Abstract— Emotions form a basis of analyzing what an individual is feeling or experiencing at any given point of time. Emotion recognition will thus help us evaluate the temperament of an individual. In this paper emotion recognition is carried out by extracting certain necessary features from speech. Thus speech is forming the basis of emotion recognition. Emotion recognition is divided into two parts: feature extraction and pattern matching. Here feature extraction is carried out using Mel Frequency Cepstral Coefficients (MFCC) and pattern matching is done using Hidden Markov Model (HMM).

Key words: Emotion recognitio, feature extraction, feature matching, Mel Frequency Cepstral Coefficient(MFCC), Hidden Markov Model (HMM).

I. INTRODUCTION

Speech is the most common and widely used mode of communication. It holds a lot of features and attributes that form the basis of analysis in the much techno-world today. The speech signal is comprised of 3 parts: voiced part, unvoiced part and silenced part. An Automatic Speech Recognition (ASR) system helps in human computer interaction bridging the communication gap between humans and machines. Analyzing the voice signal directly increases complexity because the voice signal carries a lot of information. Thus digital processing of the voice signal coupled with voice recognition algorithms are features of a voice recognition technology. Digital signal processes such as Feature extraction and Feature matching are carried out to represent the voice input. Voice recognition algorithms are made up of two phases: training phase and testing phase. Emotion recognition thus helps us to recognize and evaluate the internal expressions of an individual from a speech database.

II. SPEECH EMOTION RECOGNITION SYSTEM

A speech emotion recognition system consists of emotional speech as input, pre-processing, feature extraction and selection, classifier and the recognized emotion as the output [1].

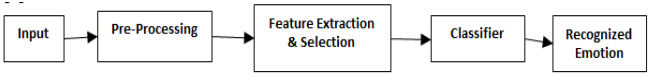

Fig. 1: Structure of A Speech Emotion Recognition System

A. Pre-Processing

In pre-processing analysis of the speech signal is done before we extract the necessary features. We perform the following operations on our input speech signal: End point detection and silence removal.

Knowing the end points of the speech signal plays an important role in further extraction of the features as we get to know the exact end of the signal and hence we can apply further techniques accordingly.

Silence removal on the other hand gets us only the voiced part of the speech and thereby removes the unvoiced part. This is one important process which reduces the computation and gets us the exact voiced signal containing features which are needed to be extracted, as the silence part contains no features

B. Feature Extraction and Selection

Voiced signal contains all the necessary features to be extracted wherein pitch plays the major role. Emotion detection from speech is thereby in a way dependant on the pitch of the signal. We use MFCC to extract the necessary features, as it is very much near to the human perception for audio.

C. Classifier

Classifier compares the features extracted from the testing data with the training data. It then returns us the perfect match from the training data. We are using HMM as the classifier as it is an appropriate tool to compare any two signals.

III. MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

MFCC feature extraction method has been used. The spectral information of a sound signal is represented in a very compact form using MFCC. We require 12-13 coefficients from each frame to accurately identify the emotion. [2]. The algorithm following MFCC feature extraction is as follows:

1) Pre-emphasis.
2) Framing.
3) Hamming windowing.
4) FFT.
5) Mel filter bank.
6) DCT.
7) Delta energy and delta spectrum.

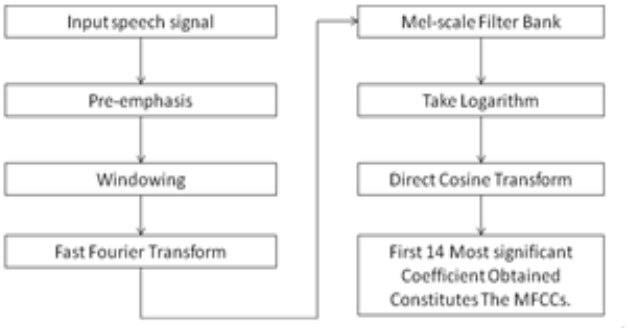

Fig. 2: MFCC Block Diagram

A. Pre-Emphasis

The high frequency region is more prone to noise, and contains more information compared to the low frequency region. We therefore use pre emphasis to get rid of the noise present in the high frequency region and thereby emphasise the high frequency region. The signal energy at high frequencies are increased considerably to an appropriate level. It involves passing the speech signal through a FIR filter. The filter equation used is as follows: [3]

$$Y[n]=X[n]-0.95X[n] \quad (1.1)$$

B. Framing

Audio signal tends to change continuously (with respect to the spectral characteristics). Though there is no significant change observed in short time intervals. Hence we divide the signal into frames of 20-40ms. We frame the signals with 50% overlapping.

C. Windowing

On each block of frames we apply Hamming window in order to remove the sudden discontinuities that occur at the ends in a way by tapering the ends. Therefore we apply overlapping windows so that we can accommodate all the information correctly. Hamming window is represented as: [3]

$$Y(n)=X(n)*W(n) \quad (3.1)$$

$$w(n)=0.54-0.46\cos(2*pi*n/(N-1)) \quad 0 \le n \le N-1 \quad (3.2)$$

where,
N = number of samples we need
Y(n) = output signal
X(n) = input signal
W(n) = hamming window

D. Fast Fourier Transform (FFT)

To convert a frame of N samples from time domain to frequency domain we make use of FFT. As we need the log magnitude spectrum for the determination of MFCC, FFT is used. 1024 point FFT is used for better frequency resolution. FFT is given as: [3]

$$Y(w) = FFT[h(t)*X(t)]= H(w)*X(w) \quad (4.1)$$

E. Mel Filter Bank Processing

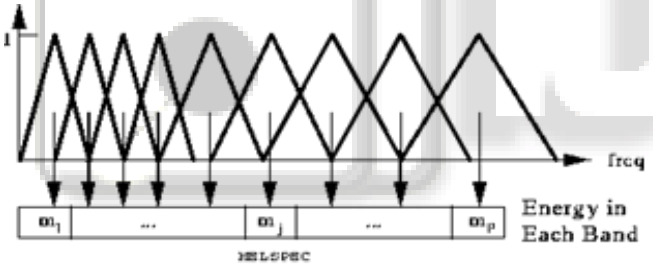

Fig. 3: Mel Filter Bank

Our input voice signal does not follow the linear scale. Thus we perform Mel Filter Bank processing in which the frequency scale is converted into Mel scale. A set of triangular filters with 50% overlapping are used to find a weighted sum of filter spectral components. The spectrum from each filter is added to get one coefficient each. Thus we consider the first 13 coefficients as our features because it gives better recognition accuracy [2]. The following formula is used to convert the frequencies to Mel scale: [1]

$$f(mel) = 2595*\log10(1+f/700) \quad (5.1)$$

F. Discrete Cosine Transform (DCT)

DCT gives us the final MFCC coefficients. It converts the log Mel spectrum into time domain, thereby giving us the final MFCC coefficients [3]. The magnitude spectrum of fourier transform is chosen over the phase spectrum because our ear is familiarized with the magnitude spectrum. These set of coefficients are known as acoustic vectors. Therefore, by transforming each input speech signal into a sequence of acoustic vectors we generate reference templates. [2]

G. Delta Energy And Delta Spectrum

The voice signal and the frames changes such as the slope of a formant at its transitions. Thus there is a need to relate features to change in cepstral features overtime [6]. 13 delta or velocity feature (12 cepstral features plus one energy feature), and 13 double delta or acceleration features are added. Each of the 13 delta features represents the change between frames while each of the 13 double delta features represents the change between frames in the corresponding delta features. The total of 39 MFCC feature are calculated for every frame that constitute the feature vector. To calculate the delta coefficients the following formula is used: [2]

$$d(t) = (c(t+1)-c(t-1))/2 \quad (6.1)$$

IV. HIDDEN MARKOV MODEL (HMM)

Emotions are classified from the given speech signal on the basis of the different features extracted. These features are given to the classifier. HMM is used in speech recognition because a speech signal can be viewed as a stationary signal. Thus speech can be thought of a Markov model having a chance of probability [7]. HMM is used because of its simplicity and computational feasibility. HMM consists of first order Markov chain whose states are hidden but the outcome is visible. They describe the sequence of events. Due to the presence of state transition matrix in HMM the temporal dynamics of speech features can be trapped [1].

V. RESULTS

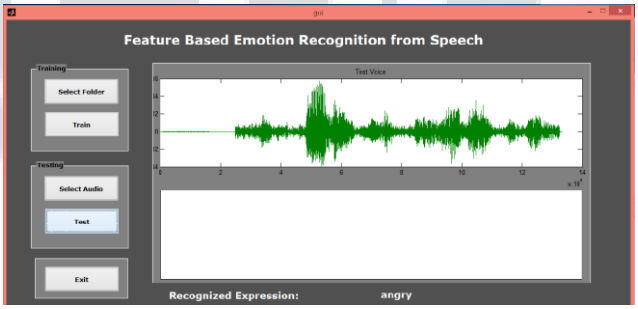

Fig. 4: Recognized Emotion Is Angry

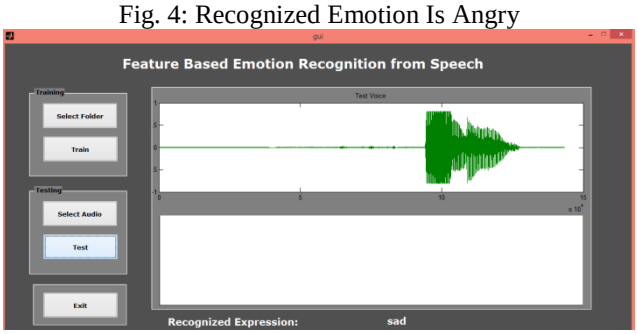

Fig. 5: Recognized Emotion is Sad

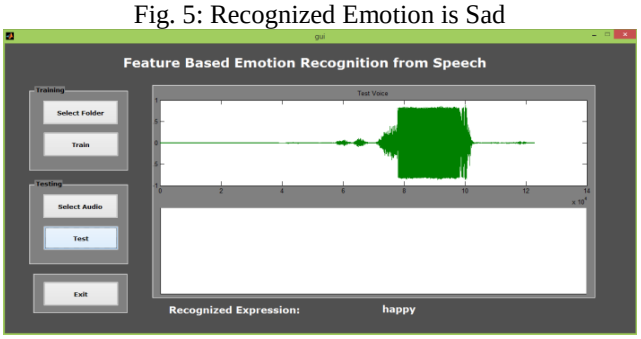

Fig. 5: Recognized Emotion is Happy

A. Confusion Matrix for All the Emotions

Emotion	Happy	Angry	Sad
Happy	-388.4991	-719.7140	-579.0544
Angry	-517.5993	-409.6627	-526.8452
Sad	-486.3848	-576.5781	-339.6184

Table 1:

VI. CONCLUSION

We use MFCC to extract features from the speech, which on a comparative study proves out to be a better feature extraction technique. Focusing on the pitch of the speech signal has made it speaker independent as male, female and kids have a specific fundamental frequency of speech. Using HMM as the classifier makes it easy to detect the closest possible match, and thereby displays the emotion of that matched signal. We are not interested in the signal with which it matches, instead we focus on the class of emotion that signal belongs. The maximum probability match is given as the output. On comparing various research papers we get to know that HMM and MFCC together works out as a good emotion recognition system for speech, and that is even being tested in this project.

REFERENCES

[1] Ashish B. Imgale, Dr.D.S.Chaudhari, International Journal of Advanced Engineering Research and Studies, Vol. 1, Issue 3, April-June 2012, E-ISSN2249-8974.
[2] N. Murali Krishna, P.V.Lakshmi, Y.Srinivas, J.Sirisha Devi, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1 September 2011, ISSN (Online): 1694-0814.
[3] Lindasalwa Muda, Mumtaj Begum, I.Elamvazuthi, Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617
[4] Pratik K. Kurzekar, atnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P.Shrishrimal, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, December 2014, ISSN: 2319-8753.
[5] Ibrahim Patel, Dr.Y.Srinivas Rao, Signal & Image Processing: An International Journal (SIPIJ), Vol. 1, Issue 2, December 2010.
[6] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366.
[7] Iosif Mpros, Todor Gangchev, Mihalis Siafarikas, Nikos Fakotakis, Journal of Computer Science, Vol. 3, Issue 8:608-616, 2007, ISSN 1549-3636.

