ASSIGNMENT

Module Code	SIP504
Module Name	Applied Machine Learning
Course	M. Tech. in Machine Learning and Intelligent Systems
Department	Computer Science and Engineering
Faculty	Faculty of Engineering And Technology

Name of the Student	:	Vishnu Prasad P
Reg. No	:	17ETCS075004
Batch	:	Full-Time 2017
Module Leader	:	Dr. Subarna Chatterjee

Ramaiah University of Applied Sciences

University House, Gnanagangothri Campus, New BEL Road,
M S R Nagar, Bangalore, Karnataka, INDIA - 560 054

Declaration Sheet				
Student Name	Vishnu Prasad P			
Reg. No	17ETCS075004			
Course	M. Tech. in Machine Learning and Intelligent Systems		Batch	Full-Time 2017
Module Code	SIP504			
Module Title	Applied Machine Learning			
Module Date	09 July 2018	to	11 August 2018	
Module Leader	Dr. Subarna Chatterjee			

Declaration

The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.

Signature of the student		Date	11 August 2018
Submission date stamp (by Examination & Assessment Section)			

Signature of the Module Leader and date	Signature of Reviewer and date

Firstly, The best way for the emotion classification is argued between the facial and the speech. The need of the emotion classification using speech is identified and the challenges that is on way for achieving it is also being discussed. A survey around all the models that is built for emotion recognition over speech is also discussed. Finally, the looking at the performance the stance of speech based emotion recognition is the best way to extract many emotion classes than that of the facial so the speech is choosen over image.

An Audio surveillance system is designed and built in this section. The surveillance system verifies the user i.e. it identifies the speaker in the vicinity.This is achieved from previously talking sample voice data from the user and identifying the features out of it and training the model to identify the same voice once more when received as an input.

A next advancement of capturing the emotion of the speaker is done in this stage. The steps involved are of 1.Emotional speech input ,2.Feature extraction and selection, 3.Training, 4.Classification, 5.Emotion recognition. All these are done and the classifier efficiency is noted to be as 85.71%.

Figure No.	Title of the figure	Page. No.

Speech processing methods are preferred over Image processing techniques for Emotion Recognition

1.1 Introduction and need for speech processing techniques in emotion recognition

Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues.

Speech emotion detection refers to analysing vocal behaviour as a marker of affect, with focus on the nonverbal aspects of speech. Its basic assumption is that there is a set of objectively measurable parameters in voice that reflect the affective state a person is currently expressing. This assumption is supported by the fact that most affective states involve physiological reactions which in turn modify the process by which voice is produced. For example, anger often produces changes in respiration and increases muscle tension, influencing the vibration of the vocal folds and vocal tract shape and affecting the acoustic characteristics of the speech [25]. So far, vocal emotion expression has received less attention than the facial equivalent, mirroring the relative emphasis by pioneers such as Charles Darwin.

In the past, emotions were considered to be hard to measure and were consequently not studied by computer scientists. Although the field has recently received an increase in contributions, it remains a new area of study with a number of potential applications. These include emotional hearing aids for people with autism; detection of an angry caller at an automated call centre to transfer to a human; or presentation style adjustment of a computerised e-learning tutor if the student is bored.

1.2 Parameters and challenges in speech detection for emotion analysis

- Firstly, discovering which features are indicative of emotion classes is a difficult task. The key challenge, in emotion detection and in pattern recognition in general, is to maximise the between-class variability whilst minimising the withinclass variability so that classes are well separated. However, features indicating different emotional states may be overlapping, and there may be multiple ways of expressing the same emotional state. One

strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others, creating a compact emotion code that can be used for classification. This avoids making difficult a priori assumptions about which features may be relevant.

- Secondly, previous studies indicate that several emotions can occur simultaneously [14]. For example, co-occurring emotions could include being happy at the same time as being tired, or feeling touched, surprised and excited when hearing good news. This requires a classifier that can infer multiple temporally co-occurring emotions.
- Thirdly, real-time classification will require choosing and implementing efficient algorithms and data structures.
- Fourth, in persuasive communication, special attention is required to what non-verbal clues the speaker conveys. Untrained speakers often come across as bland, lifeless and colourless. Precisely measuring and analysing the voice is a difficult task.

Despite there existing some working systems, implementations are still seen as challenging and are generally expected to be imperfect and imprecise.

1.3 Critical analysis of techniques in speech recognition for emotion analysis

Table-1-3-1 presents the current available methods, which targeted the speech emotion recognition systems, and these are evaluated with its classifier, features set, and recognition rate and on different dataset levels.

Table 1-3-1:Analysis different methods of Speech emotion Recognition

Ref	Types of classifier	Types of features	Recognition Rate	Type of Dataset	Methods
H. Cao(2015)	SVM	Prosodic and spectral features	44.40%	Berlin & LDC & FAU Aibo dataset	Ranking SVM

T. L. New (2003)	HMM	Log frequency power coefficients (LFPC), MFCC	Average and best result 78% and 96% respectively	Two private speech dataset	Discrete HMM and LFPC to characterize speech signal
J.-H. Yeh(2011)	k-NN	Jitter, shimmer, formants, LPC, LPCC, MFCC, LFPC, PLP, and Rasta-PLP, SFS and SBS	Best 86%	Chinese emotional speech corpus We invited 18 males and 16 females	Segment based method by employing k-NN, SFS (sequential forward selection), SBS (sequential backward selection)
C.-C. Lee (2009)	Bayesian Logistic Regression, SVM	large-margin feature	70.1% & 65.1% for two and five class	AIBO dataset	Hierarchical structure for binary decision tree

1.4 Stance taken with justification

Various speeches emotional recognition systems reviewed and discussed based on different approaches. Here the performance is also compared in terms of classifier, features, recognition rate, and datasets. Well-design classifiers have obtain high classification accuracies between different types of emotions. In this study HMM with adopting short time LFPC as a feature proves a good accuracy on different levels in the chart. Basharirad(2017) The majority of the current datasets are not capable for evaluation of speech emotion recognition. In most of them, it is hard even for human to specify different emotion of certain collected utterances; e.g. the human recognition accuracy was 80% for Berlin

1.5 Conclusion

From the above statements it is clearly evident that the facial expression does not exhibit as many classes of the emotions that is exhibited in voice/speech of a person.

Develop and simulate an algorithm for audio surveillance in PYTHON.

2.1 Discuss and Compare recent algorithms for audio surveillance

This section intents to provide a representative picture of what has been developed so far in the area of audio surveillance. The emphasis of previous approaches is mainly placed on the classifier, the feature extraction process, the training data, and the number of classes.

Table 2.1-1:Algorithms of Audio Survillance (Ntalampiras, 2015)

Paper	Atypical sound classes	Model adaptation	Classifier	Features	Database
Ntalampiras et al. (2009)	Scream, gunshot, and explosion	MAP adaptation of GMMs	GMM	MFCC, MPEG-7, CB-TEO, Intonation	Large audio corpora from professional sound effects
Valenzise et al. (2007)	Scream and gunshot	-	GMM	Temporal, spectral, cepstral, correlation	Movie soundtracks, Internet, and people shouts
Radhakrishnan & Divakaran (2005)	Banging and nonneutral speech	-	GMM	MFCC	Elevator recordings
Clavel et al. (2005)	Gunshot	-	GMM	MFCC, spectral moments	CDs for the national French public radio
Rouas et al. (2006)	Shout	Adaptive threshold for sound activity detection	GMM, SVM	Energy, MFCC	Recorded during four scenarios
Vacher et al. (2004)	Scream and glass break	-	GMM	Wavelet based cepstral coefficients	Laboratory recordings and RCWP

Atrey et al. (2006)	Shout	-	GMM	ZCR, LPC, LPCC, LFCC	Recorded in offi ce corridor
Ito et al. (2009)	Glass clash, scream, fi re cracker	Adaptive threshold for abnormal sound event detection	GMM	MFCC, Power	Recorded under laboratory conditions

2.2 Review various feature extraction and classification techniques for Speech Recognition.

2.2.1 feature extraction techniques for Speech Recognition

Feature extraction is the main part of the speech recognition system. It is considered as the heart of the system. The work of this is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction compresses the magnitude of the input signal (vector) without causing any harm to the power of speech signal.

Table 2.2-1:Comparison of different extraction techniques

Technique	Characteristics	Advantages	Disadvantages
LINEAR PREDICTIVE CODING	*Provides autoregression based speech features. *Is a formant estimation technique . *A static technique. *The residual sound is very close to the vocal tract input signal.[7]	*Is a reliable, accurate and robust technique for providing parameters which describe the timevarying linear system which represent the vocal tract. * Computation speed of LPC is good and provides with accurate parameters of speech. * Useful for encoding speech at low bit rate.	Is not able to distinguish the words with similar vowel sounds . * Cannot represent speech because of the assumption that signals are stationary and hence is not able to analyze the local events accurately. *LPC generates residual error as output that means some amount of important speech gets left in the residue resulting in poor speech

			quality.
MEL – FREQUENCY CEPSTRUM (MFCC)	* Used for speech processing tasks. *Mimics the human auditory system * Mel frequency scale: linear frequency spacing below 1000Hz & a log spacing above 1000Hz.	* The recognition accuracy is high. That means the performance rate of MFCC is high. * MFCC captures main characteristics of phones in speech. * Low Complexity.	*In background noise MFCC does not give accurate results. * The filter bandwidth is not an independent design parameter * Performance might be affected by the number of filters.
RelAtive SpecTrAl (RASTA Filtering)	* Is a band pass filtering technique. * Designed to lessen impact of noise as well as enhance speech. That is, it is a technique which is widely used for the speech signals that have background noise or simply noisy speech.	*Removes the slow varying environmental variations as well as the fast variations in artefacts. * This technique does not depend on the choice of microphone or the position of the microphone to the mouth, hence it is robust. * Captures frequencies with low modulations	*This technique causes a minor deprivation in performance for the clean information but it also slashes the error in half for the filtered case. RASTA combined with PLP gives a better performance ratio.

		that correspond to speech.	
Probabilistic Linear Discriminate Analysis (PLDA)	*Based on i-vector extraction. The ivector is one which is full of information and is a low dimensional vector having fixed length. * This technique uses the state dependent variables of HMM. * PLDA is formulated by a generative model.	* Is a flexible acoustic model which makes use of variable number of interrelated input frames without any need of covariance modelling. * High recognition accuracy	* The Gaussian assumption which are on the class conditional distributions. This is just an assumption and is not true actually. * The generative model is also a disadvantage. The objective was to fit the date which takes class discrimination into account.

2.2.2 Classification techniques for Speech Recognition.

Classifier compares the obtained features with stored features. Based upon this comparison classifier recognizes the particular speaker.

Table 2.2-2:Comparison of different Classification techniques

Technique	Characteristics	Advantages	Disadvantages
Guassian Mixture Model	Unsupervised	Needs less trainning and test data	Compromise between DTW and HMM

Dynamic Time Warping (DTW)	Unsupervised	Requires less storage space,beneficial for variable length	Cross-channel issue
Hidden Markov Model (HMM)	Unsupervised	Rail system outputs,efficient performance	Computationally more complex,more storage space
Vector Quantization(VQ)	Unsupervised	Computationally less complex	Real time encoding is complex
Support Vector Machine (SVM)	Supervised	Simple operation	binary SVM has limitations in speaker recognition

2.3 Capture real time audio samples and design and simulate a model for audio analysis using MATLAB.

The speech surveillance system that is being designed here can also be called as a Speaker Authorization/Recogniser. This system can be implemented at places of restricted access and the need for the identity of the person present at a highly protected areas fig shows the pictorial representation of the application.

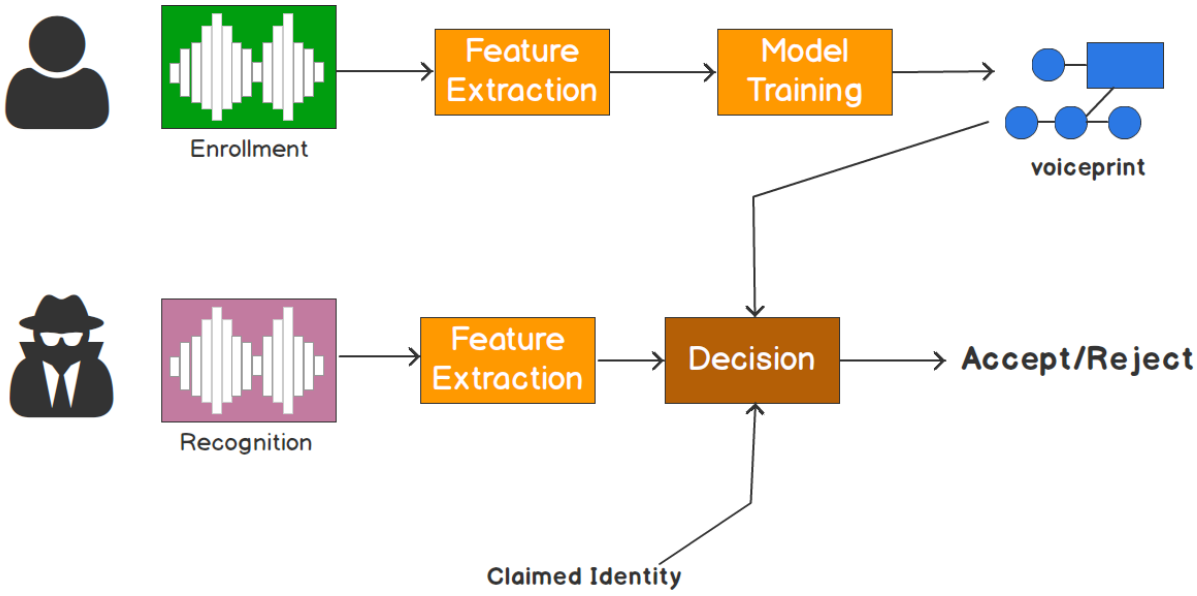


Figure 2.3-1:Survillance system Block Diagram

Speech Signals used for the system

The audio samples used for the identification of the speaker are as described in tabel. All the voices are computer generated and each of them vary from one another.

Table 2.3-1:Input files Description

File name	Voice	Uttered word
S1.wav	Female (Computer Generated)	Zero
S2.wav	Female (Computer Generated)	Zero
S3.wav	Female (Computer Generated)	Zero
S4.wav	Female (Computer Generated)	Zero
S5.wav	Female (Computer Generated)	Zero
S6.wav	Female (Computer Generated)	Zero
S7.wav	Female (Computer Generated)	Zero
S8.wav	Female (Computer Generated)	Zero

Feature Extraction

Choosing which features to extract from speech is the most significant part of speaker recognition. Some popular features are: MFCCs, LPCs, Zero-Crossing Rates etc. In this work, I have concentrated on MFCCs and LPCs. Here is a brief overview of these features.

Mel-Frequency Cepstral Coefficients

Human hearing is not linear but logarithmic in nature. This implies that our ear acts as a filter. 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. The relationship between frequency in Hz and frequency in Mel scale is given by:

$$m = 1125\ln\left(1 + \frac{f}{700}\right)$$

To calculate MFCCs, the steps are as follows.

The speech signal is divided into frames of 25ms with an overlap of 10ms. Each frame is multiplied with a Hamming window.

1. The periodogram of each frame of speech is calculated by first doing an FFT of 512 samples on individual frames, then taking the power spectrum as:

$$P(k) = \frac{1}{N}|S(k)|^2$$

Where P(k) refers to power spectral estimate and S(k) refers to Fourier coefficients for the kth frame of speech and N is the length of the analysis window. The last 257 samples of the periodogram are preserved since it is an even function.

2. The entire frequency range is divided into 'n' Mel filter banks, which is also the number of coefficients we want. 'For 'n' = 12, the filter bank is shown in Figure 3 - a number of overlapping triangular filters with increasing bandwidth as the frequency increases.

3. To calculate filter bank energies we multiply each filter bank with the power spectrum, and add up the coefficients. Once this is performed we are left with 'n' numbers that give us an indication of how much energy was in each filter bank.
4. We take the logarithm of these 'n' energies and compute its Discrete Cosine Transform to get the final MFCCs.

Linear Prediction Coefficients

LPCs are another popular feature for speaker recognition. To understand LPCs, we must first

understand the Autoregressive model of speech. Speech can be modelled as a p^{th} order AR process, where each sample is given by:

$$x(n) = -\sum_{k=1}^{p} a_x x(n-k) + u(n)$$

Each sample at the n^{th} instant depends on 'p' previous samples, added with a Gaussian noise u(n). This model comes from the assumption that a speech signal is produced by a buzzer at the end of the tube (voiced sounds), with occasional added hissing and popping sounds.

LPC coefficients are given by $\alpha_.$ To estimate the coefficients, we use the Yule-Walker equations. It uses the autocorrelation function R_X. Autocorrelation at lag l is given by:

$$R(i) = \sum_{n=1}^{N} x(n)x(n-l)$$

While calculating ACF in Python, the Box-Jenkins method is used which scales the correlation at each lag by the sample variance so that the autocorrelation at lag 0 is unity.

The final form of the Yule-Walker equations is:

$$\sum_{k=1}^{p} a_k R(l-k) = -R(l)$$

The solution for a is given by:

$$a = R^{-1}r$$

In this case, I have normalised the LPC coefficients estimated so that they lie between [-1,1]. This was seen to give more accurate results. We first divide speech into frames of 25ms with 10ms overlap, then calculate 'p' LPCs for each frame.

Feature Matching

The most popular feature matching algorithms for speaker recognition are Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Vector Quantization (VQ). Here, I have used Vector Quantization.

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that

space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all codewords is called a codebook.

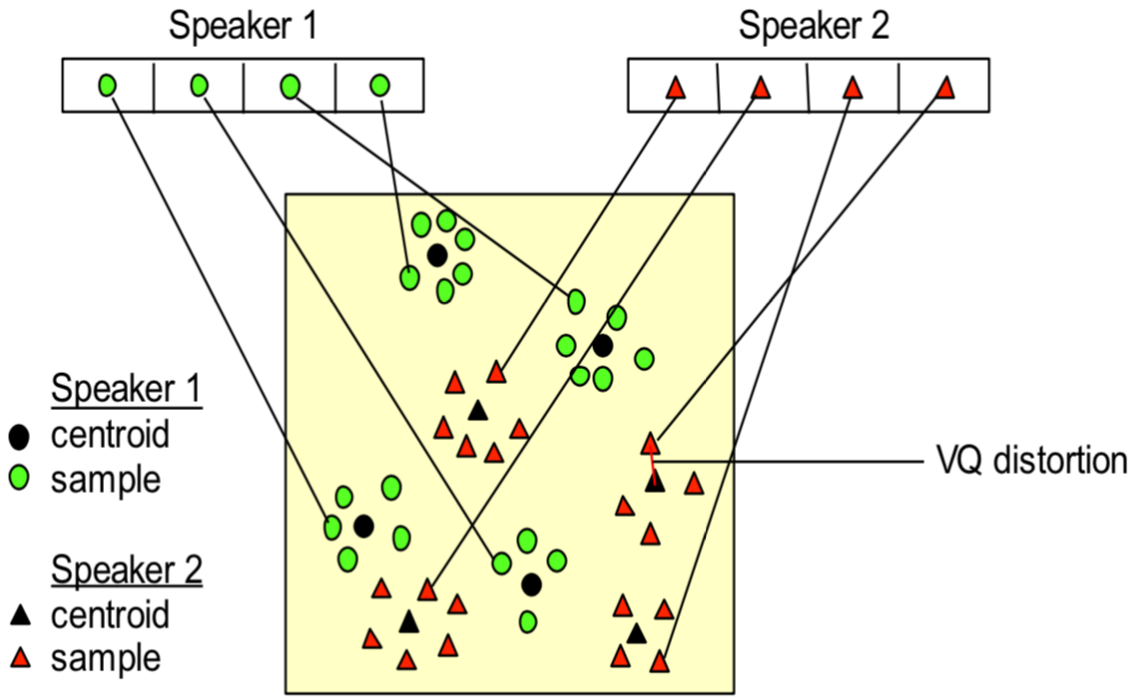


Figure 2.3-2:Visualizarion of vector quantization

LBG Algorithm

The LBG algorithm [Linde, Buzo and Gray], is used for clustering a set of L training vectors into a set of M codebook vectors. The algorithm is formally implemented by the following recursive procedure:

Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).

Double the size of the codebook by splitting each current codebook yn according to the

$$y_n^+ = y_n(1+\varepsilon)$$
$$y_n^- = y_n(1-\varepsilon)$$

where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).

Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

Iteration 1: repeat steps 3 and 4 until vector distortion for current iteration falls below a fraction of the pervious iteration's distortion. This is to ensure that the process has converged.

Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed. Intuitively, the LBG algorithm designs an M-vector codebook in stages. It starts first by

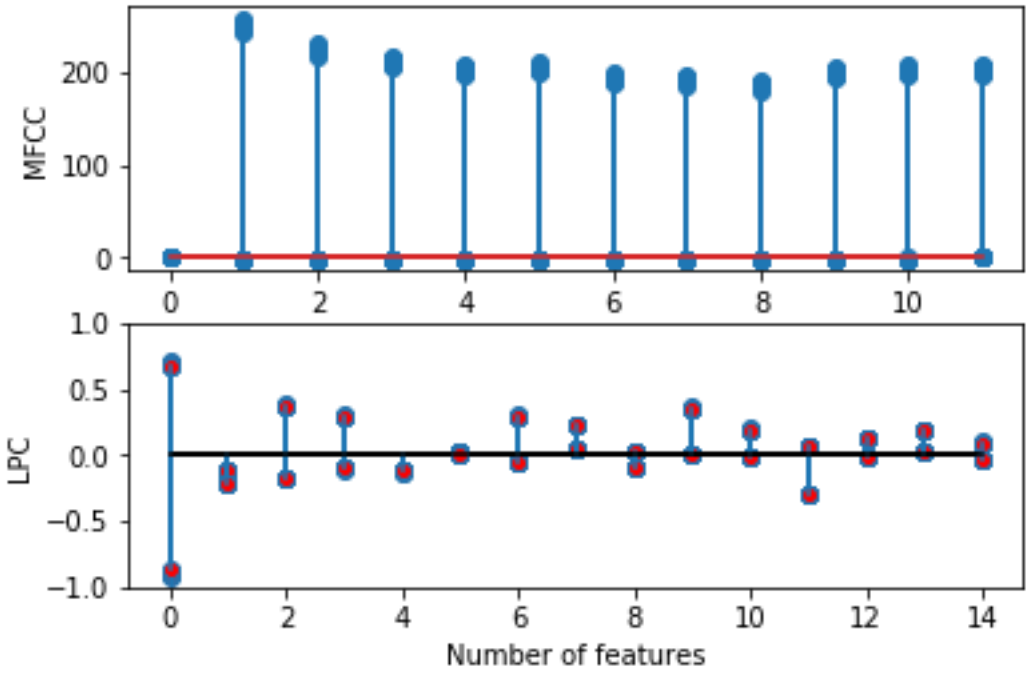
designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained.

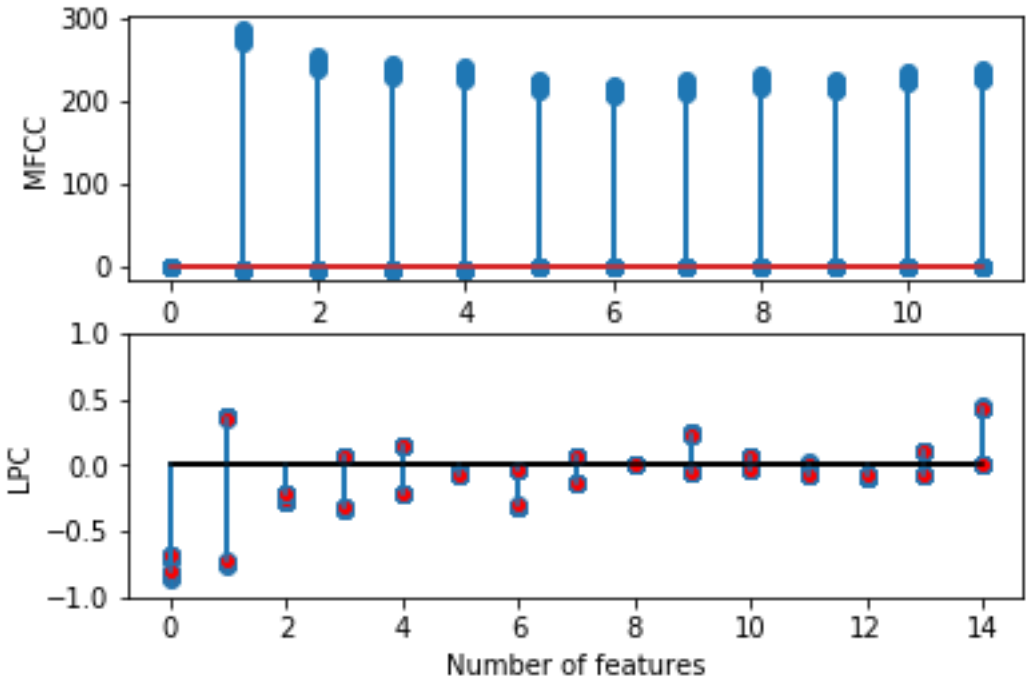
2.4 Design an audio surveillance system using MATLAB.

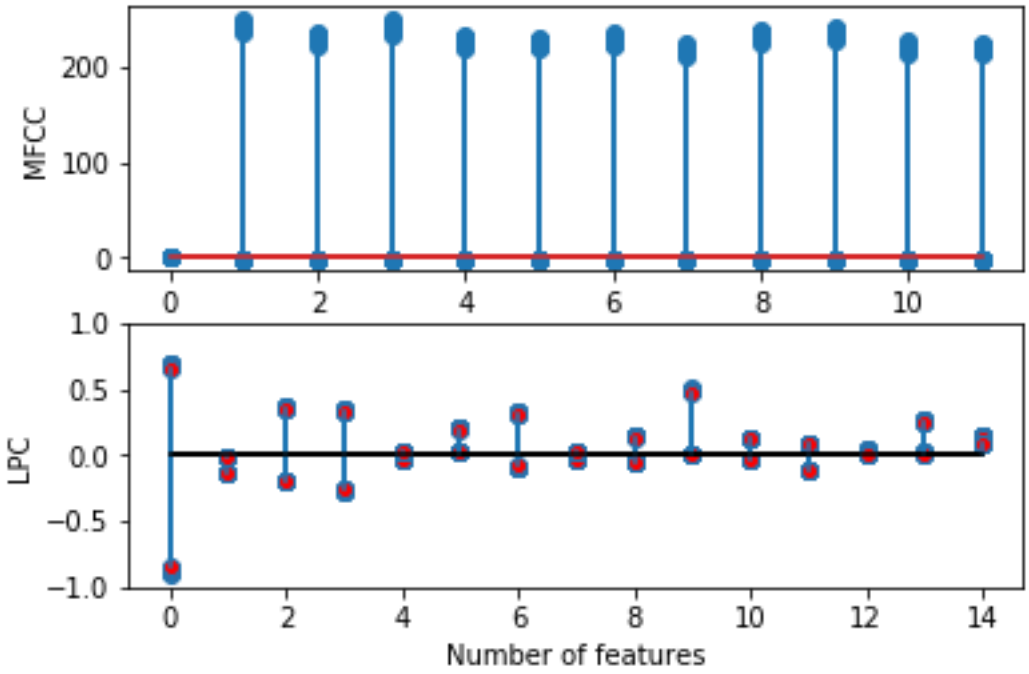
Feature Training

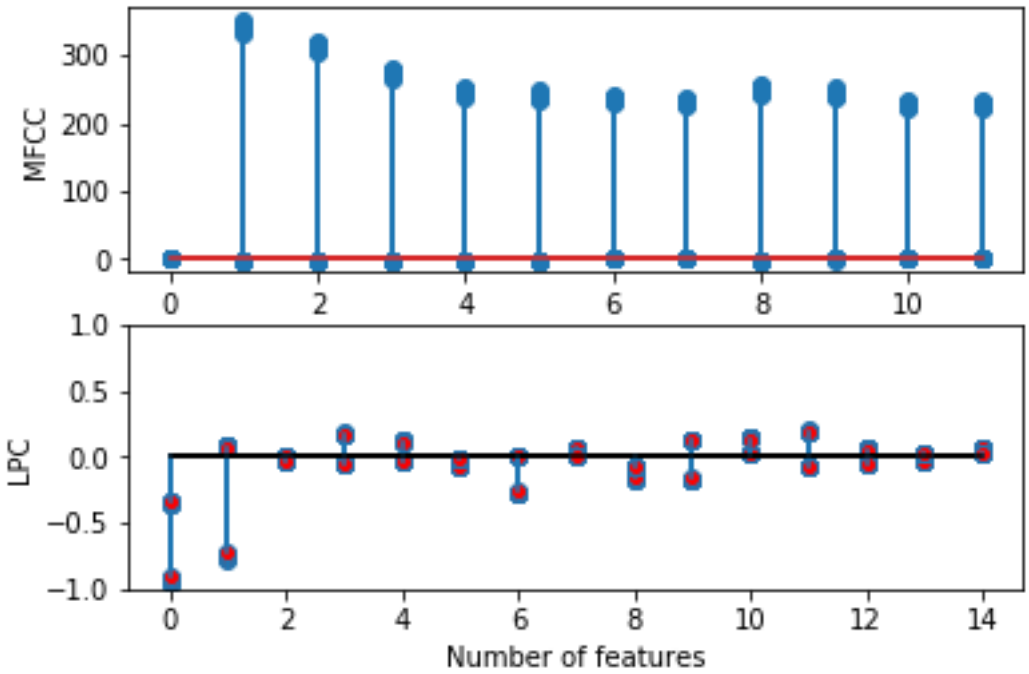
The main algorithms needed for speaker recognition have been implemented. Now, everything needs to be brought together to train our dataset and derive codebooks for each speaker using VQ. The number of speakers is nSpeaker = 8. As mentioned before, speech recordings of 8 female speakers uttering the word 'zero' has been taken for training and testing. Each codebook should have 16 codewords, hence nCentroid = 16 (it is highly recommended to keep this number a power of 2).

Codebooks for both MFCC features and LPC features are plotted for all 8 speakers on the figure.


MFCC
200
100
0
0
2
4
6
8
10
LPC
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10
12
14
Number of features


MFCC
300
200
100
0
0
2
4
6
8
10
LPC
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10
12
14
Number of features


MFCC
200
100
0
0
2
4
6
8
10
LPC
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10
12
14
Number of features


MFCC
300
200
100
0
0
2
4
6
8
10
LPC
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10
12
14
Number of features

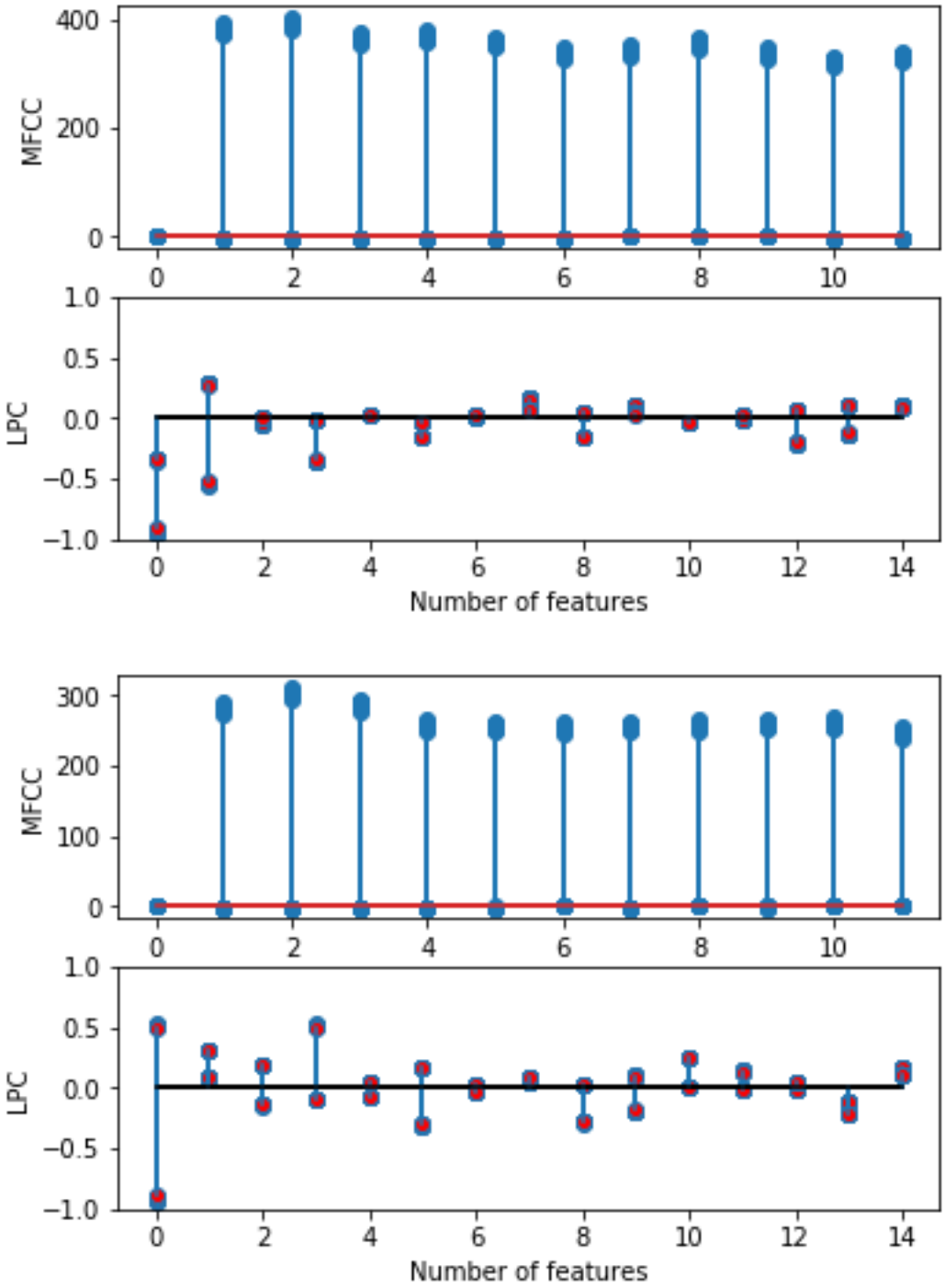
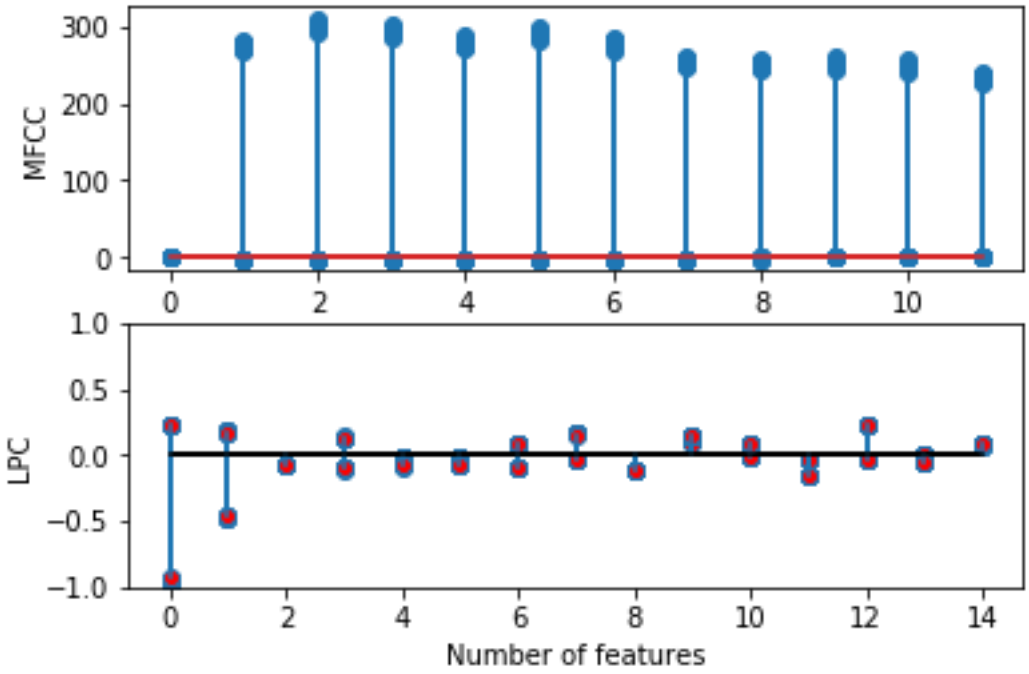
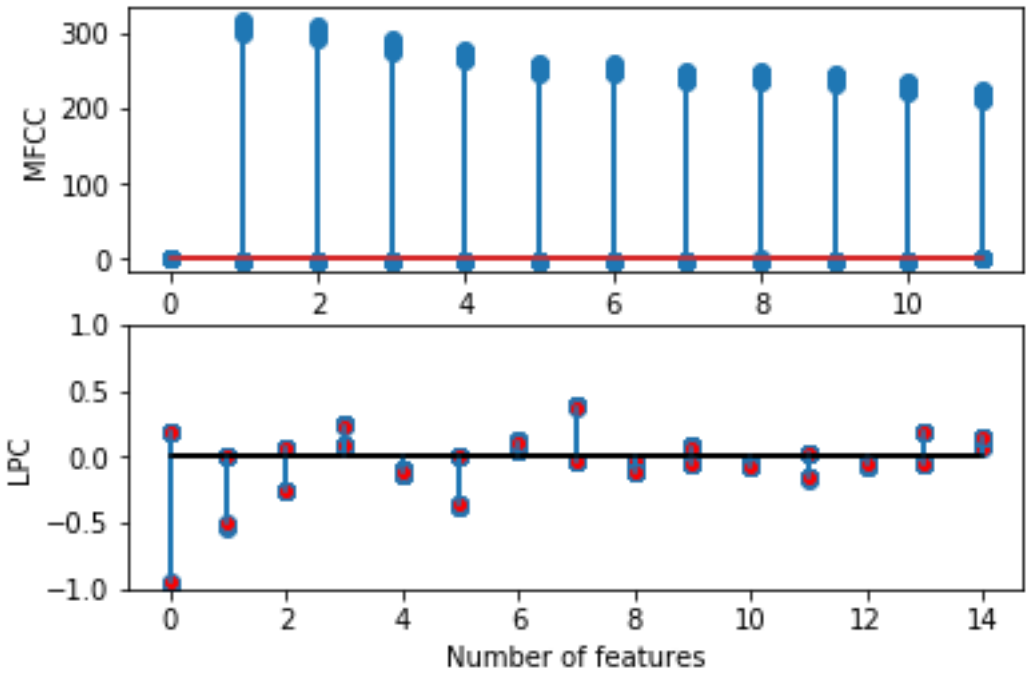


Figure 2.4-1:Number of features extracted and their respective values for the two feature extraction technique


MFCC
LPC
Number of features


MFCC
LPC
Number of features

2.5 Test and validate the developed algorithm on benchmark audios.

It is finally time to test our speaker recognition algorithm. the speaker recognition is done by comparing their feature vector to the codebooks of all trained speakers and computing the minimum distance between them. The results are yielding an accuracy of 37.5% with MFCC and 50% with LPC. Reasons for this low accuracy can be the fact that there wasn't enough data to train on. Other complex classification algorithms such as ANNs and SVMs should yield better results. I observed that training with12 MFCC features and LPC coefficients of order 15 gives the best results. There are other parameters that can be varied, such as number of codewords in a codebook and FFT size while computing MFCCs. It is possible that a different combination of these will give higher accuracy.

The following table gives the identification results for each of the 8 speakers with MFCC and LPC coefficients and Vector Quantization with LBG algorithm for classification.

Table 2.5-1:Verification of the speaker recognition system

True Speaker	Identified as (MFCC)	Identified as(LPC)
S1	S1	S5
S2	S8	S2
S3	S5	S1
S4	S6	S4
S5	S5	S5
S6	S3	S1
S7	S8	S8
S8	S8	S8
	Accuracy=37.5%	Accuracy = 50%

3.1 Analysis and comparison of existing algorithms for emotion detection

The traditional K-nearest neighbor algorithm and fuzzy K nearest neighbor algorithm are adopted for pattern recognition. And the two methods are compared and the respective characteristics are analyzed through the experimental results. In the use of the traditional K nearest neighbor and fuzzy K nearest neighbor algorithm, the K has to be selected The recognition results are shown in Table 3.1-1.

Table 3.1-1:Comparison between two models of ERS (Xia(2015))

Identification method	Anger	Happy	Neutral	Sadness	The average recognition rate
KNN	83.29	82.36	78.53	79.62	80.95
FKNN	85.35	83.59	80.14	81.47	82.63

The four kind of emotion recognition rate were also having varying degrees of increase, this is mainly because of more neighbors, which reduces the risk of miscarriage of justice, but also increases the amount of computation. At the same time, it can be seen that because the sample size is small, the traditional K nearest neighbor algorithm and fuzzy K nearest neighbor algorithm recognition rate is relatively high. For the recognition of four emotions anger, sad, happy, neutral, recognition effect for anger is the best, mainly reason is that anger emotion characteristics is more obvious compared to the other three emotion. In expressing the emotion of anger, the speaker's speed is often faster, and the tone is high, for the neutral and sad recognition rate is relatively low, mainly because the sadness and neutral state, some physiological characteristics are similar, easily confused by mistake. From the comparison of two algorithms we can also find that fuzzy K nearest neighbor algorithm calculates the Euclidean distance to take full account of the proportion between various parameters compared with the traditional K nearest neighbor algorithm, which makes the recognition effect, is more prominent, and the recognition rate has improved in a way.

3.2 Choice of an appropriate emotion detection algorithm and its justification for the speech files given

This system consists of 5 steps, namely

1.Emotional speech input

2.Feature extraction and selection,

3.Training,

4.Classification,

5.Emotion recognition

There are so many prosody and spectral features in voice samples which contains the emotional information. Change in this feature leads to abrupt change in the emotions. With the different emotional state, corresponding changes occurs in pitch, energy, speak rate and spectrum. Here a set of 14 potential features are extracted.

Pitch: It is the fundamental frequency of the sound. The relative highness and lowness of tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal card. Pitch is the main acoustic correlate of tone and intonation.

Standard deviation: Indicates the variation exists from the mean or expected value. Lower the value of standard deviation indicates that data point tend to be nearer to the mean. Higher the value of standard deviation indicates data points are spread out over large range of values. **Energy intensity:** Represents the loudness of the audio signal. It is correlated with the amplitude.

Energy entropy: It represents the abrupt change in the energy level of the audio signal. **Autocorrelation:** Cross-correlation of the signal with itself. It is similarity between observations as time lag between them.

Shimmer: A frequent back and forth changes in amplitude of voice sample. It provides an evaluation of variability of peak to peak amplitude within analyzed speech sample. Also represents the relative period to period variability of peak to peak amplitude.

Jitter: Jitter is the deviation from true periodicity of a presumed periodic signal. In speech, it is defined as the varying pitch in the voice, which causes a rough sound. It describes varying loudness in voice.

Harmonics to noise ratio: HNR represents the degree of acoustic periodicity, it is also called as Harmonicity object. It is expressed in DB, HNR 0db means there is equal energy in harmonic and noise.

Noise to harmonic ratio: Evaluation of noise present in the analyzed audio signal. It is average ration of inharmonic component to the harmonic component in the audio signal. **Short time energy:** Short time energy provides the convenient representation that reflects the amplitude variations. It also provides the basis for distinguishing voiced and unvoiced speech.

Zero crossing rates: ZCR is defined as rate of change of sign along the signal, i.e. it represents changes from positive to negative or back in a speech signal. ZCR is high for the unvoiced signals and low for voiced signals.

Spectral flux: Spectral flux measures how quickly power spectrum changes in present frame with respect to previous frame and it measures the changes in the power spectrum.

Spectral centroid: Spectral centroid indicates the centre of mass in spectrum of the audio signal. It is the weighted mean frequency. Spectral centroid is good predictor of brightness of sound.

Spectral roll off: spectral roll off is defined as the percentile of power spectrum. Usually N is 85%-95%. This measure is useful in distinguishing voiced speech from unvoiced speech.

Justification for feature extraction: MFCC (Mel frequency cepstral coefficient): Mel frequency cepstral coefficient is parametric representation known as voice quality feature, widely used in the area of speech emotion recognition. It provides better rate of recognition in both speech and emotion recognitions.

The feature extraction is done as same as explained in **section**

These extracted 14 potential features are analysed and data base is created in .sav files for 7 emotional categories. MFCC feature is selected in recognizing the emotions in SVM classification, all other features increase the accuracy of the emotion recognition.

Justification for Classification: The highly optimised SVMs produce higher cross validation accuracies than other algorithms. SVM classifier is binary decision algorithm and classification is mainly dependent on the MFFC feature.

Training: In classification, Training set used to learn the model that can classify data samples in to known classes.

Testing: In order to assess the model accuracy test the model using unseen test data

3.3 Development of a software reference model of the chosen algorithm using PYTHON

- Firstly, the train and test data are split and are placed in different folders named train_sounds and new_test_sounds.

```
X, sample_rate = librosa.load(file_name)
```

- All the features of the input files present in train_sounds is extracted using a function.

```
def extract_feature(X):
stft = np.abs(librosa.stft(X))
mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
chroma = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
mel = np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T, axis=0)
contrast = np.mean(librosa.feature.spectral_contrast(S=stft, sr=sample_rate).T, axis=0)
tonnetz = np.mean(librosa.feature.tonnetz(y=librosa.effects.harmonic(X),
sr=sample_rate).T, axis=0)
return mfccs, chroma, mel, contrast, tonnetz
```

- The extracted features are made into vectors and are stacked and labelled.

```
def parse_audio_files(path):
features = np.empty((0, 193))
for fn in glob.glob(path):
try:
mfccs, chroma, mel, contrast, tonnetz = extract_feature(fn)
except Exception as e:
print("Error encountered while parsing file: ", fn)
continue
ext_features = np.hstack([mfccs, chroma, mel, contrast, tonnetz])
features = np.vstack([features, ext_features])
target_files.append(fn)
return np.array(features)
```

- The SVM classifier is loaded and is given the features to fit itself(Training process).

```
model = pickle.load(open(filename, 'rb'))
prediction = model.predict(ts_features)
```

- Finally the each of the speech file from the floder new_test_sounds are read one by one and the features of them are extracted and the classifier is made to predict the output.

```
for i, val in enumerate(prediction):
print("Input File: ", target_files[i], "|", " Predicted Emotion Is:", classes[int(val)])
```

3.4 Testing and validation of the reference model with the given speech files

Emotion recognition by means of SVM is implemented and allowed to test the accuracy of the system.

The provided Data set is been tested over the model and the result of the software programme are as follows.

Input File: ./new_test_sounds/YAF_young_angry.wav | Predicted Emotion Is: angry

Input File: ./new_test_sounds/YAF_young_disgust.wav | Predicted Emotion Is: disgust

Input File: ./new_test_sounds/YAF_young_fear.wav | Predicted Emotion Is: fear

Input File: ./new_test_sounds/YAF_young_happy.wav | Predicted Emotion Is: happy

Input File: ./new_test_sounds/YAF_young_neutral.wav | Predicted Emotion Is: neutral

Input File: ./new_test_sounds/YAF_young_ps(Pleasant Surprise).wav | Predicted Emotion Is: surprise

Input File: ./new_test_sounds/YAF_young_sad.wav | Predicted Emotion Is: disgust

Table 3.4-1:Result validation of the Emotion Recognition system

Audio file	Emotion	Predicted Emotion
YAF_young_angry.wav	Angry	Angry
YAF_young_disgust.wav	Disgust	Disgust
YAF_young_fear.wav	Fear	Fear
YAF_young_happy.wav	Happy	Happy
YAF_young_neutral.wav	Neutral	Neutral

YAF_young_ps(Pleasant Surprise).wav	Surprise	Surprise
YAF_young_sad.wav	Sad	disgust
		Accuracy=85.71%

3.5 Conclusion and justification.

Various speeches emotional recognition systems based approaches are analysed. We also compare its performance in terms of classifier and features. Well-design classifiers obtain high classification accuracies between different types of emotions. this project shows that building a fast and efficient speech emotion detector is a challenging but achievable goal. The key design principles behind the successful implementation of a large real-time system included choosing efficient data structures and algorithms, and employing suitable software engineering tools. In sum, this section employs an understanding of a wide area of Computer Science to demonstrate that highly accurate speech emotion detection is possible, and that it can be done in realtime.

Avetisyan, H., Bruna, O. and Holub, J. (2016). Overview of existing algorithms for emotion classification. Uncertainties in evaluations of accuracies. *Journal of Physics: Conference Series*, 772, p.012039.

Ntalampiras, S. (2015). Audio Pattern Recognition of Baby Crying Sound Events. *Journal of the Audio Engineering Society*, 63(5), pp.358-369.

Ntalampiras, S., Potamitis, I. and Fakotakis, N. (2009). An Adaptive Framework for Acoustic Monitoring of Potential Hazards. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, pp.1-15.

Pfister, T. (2010). *Emotion Detection from Speech*. Ph.D. Gonville & Caius College.

Xia, S., Wang, J., Wang, R. and Zhao, L. (2015). An Improved Algorithm of Speech Emotion Recognition. *International Journal of u- and e- Service, Science and Technology*, 8(12), pp.217-226.