EMOTION ANALYSIS USING SPEECH RECOGNITION

A MINI PROJECT-I REPORT

Submitted by

YOGAPRIYA.P (1805161)
SAMAYANJALI.M (1805121)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

SRI RAMAKRISHNA ENGINEERING COLLEGE

[Educational Service: SNR Sons Charitable Trust]
[Autonomous Institution, Accredited by NAAC with 'A' Grade]
[Approved by AICTE and Permanently Affiliated to Anna University, Chennai]
[ISO 9001:2015 Certified and All Eligible Programmes Accredited by NBA]
Vattamalaipalayam, N.G.G.O. Colony Post,

COIMBATORE – 641 022

ANNA UNIVERSITY CHENNAI 600 025

APRIL 2020

ANNA UNIVERSITY CHENNAI 600 025

BONAFIDE CERTIFICATE

16IT258 – MINI PROJECT- I

Certified that this Mini project-I Report " **Emotion Analysis using Speech Recognition"** is the bonafide work of **"P.Yogapriya, M.Samayanjali"** who carried out the project work under my supervision.

SIGNATURE
Dr.M.SenthamilSelvi
HEAD OF THE DEPARTMENT
Professor
Department of Information Technology
Sri Ramakrishna Engineering College
Vattamalaipalayam
Coimbatore-641022

SIGNATURE
Dr.J.Anitha
PROJECT GUIDE
Associate Professor
Department of Information Technology
Sri Ramakrishna Engineering College
Vattamalaipalayam
Coimbatore-641022

Submitted for University Examination viva voce held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We thank the Almighty for his blessings on us to complete this project work successfully.

With profound sense of gratitude we sincerely thank the **Managing Trustee Thiru D.Lakshmi Narayanaswamy** and **Joint Managing Trustee R.Sundar** providing us the necessary infrastructure required for the completion of our project.

With profound sense of gratitude, we sincerely thank the **Head of the Institution, Dr.N.R.Alamelu,** for her kind patronage, which helped in pursuing the project successfully.

With immense pleasure, we express our hearty thanks to **Head of the Department, Dr.M.SenthamilSelvi ,** for her encouragement towards the completion of this project.

With profound sense of gratitude, we sincerely thank the Academic Coordinator **Dr.K.Deepa, Professor**, for her kind patronage and encouragement towards the completion of the project.

We also thank our project co-ordinator **Dr. M.Kalaiarasu , Associate Professor,** and our project guide, **Dr.J.Anitha, Associate Professor**, Department of Information Technology for their encouragement, valuable guidance and keen interest towards the completion of the project

We convey our thanks to all the teaching and non-teaching staff members of the IT Department of our college who rendered their cooperation by all means for completion of this project.

ABSTRACT

Speech is a complex signal consisting of various information, such as information about the message to be communicated, speaker, language, region, emotions etc. Speech Processing is one of the important branches of digital signal processing and finds applications in Human computer interfaces, Telecommunication, Assistive technologies, Audio mining, Security and so on. Speech emotion recognition is important to have a natural interaction between human being and machine. In speech emotion recognition, emotional state of a speaker is extracted from his or her speech. The acoustic characteristic of the speech signal is Feature. Feature extraction is the process that extracts a small amount of data from the speech signal that can later be used to represent each speaker. Speaker emotions are recognized using the data extracted from the speaker voice signal.

Mel Frequency Cepstral Coefficient (MFCC) technique is used to recognize emotion of a speaker from their voice. The designed system was validated for Happy, sad and anger emotions and the efficiency was found to be about 83%. Emotion Analysis is an important area of work to improve the interaction between human and machine. Moreover, Deep Learning technique with neural network extended the success ratio of machine in respect of emotion analysis. It works with Deep Learning technique has performed with different kinds of input of human behavior like audio-visual inputs, facial expressions, body postures, EEG signal. Still many aspects in this area to work on to improve and make a particular app will detect and classify emotions more accurately. In this project, the relevant significant works, their techniques, and the effectiveness of the methods and the scope of the improvement the results are explored. The result of which the person can be aware and maintain his emotions at the correct level.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

SYMBOL	DESCRIPTION	PAGE
≤	Lesser than or equal to	10
Π	Pi	10
∑	Summation	11

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
API	Application Programming Interface
MHL	Microsoft HoloLens
MFCCs	Mel-Frequency Cepstral Coefficients
HMIs	Human Machine Interactions
SER	Speech Emotion Recognition

CHAPTER 1

INTRODUCTION

Emotion plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to our rational and intelligent decisions. This important aspect of human interaction needs to be considered in the design of human– machine interfaces (HMIs). To build interfaces that are more in tune with the user's needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication. In these years, literature about automatic emotion recognition is growing dramatically due to the development of techniques in computer vision, speech analysis and machine learning. However, automatic recognition on emotions occurring on natural communication setting is a largely unexplored and challenging problem. In addition to the features corresponding to the speaker and/or the speech, the sound signals do have some features that represents the emotional state of the speaker.

Speech signals can be analysed to detect heart rate, which will naturally enable a remote diagnosis of a patient by only using audio data.A tremendous research is being done on Speech Emotion Recognition (SER) in the recent years with its main motto to improve human machine interaction. In this work, the effect of cepstral coefficients in the detection of emotions is performed. Also, a comparative analysis of cepstum, Mel-frequency Cepstral Coefficients (MFCC) and synthetically enlarged MFCC coefficients on emotion classification is done. The proposed method has facilitated a considerable reduction in the misclassification efficiency which outperforms the algorithm where the feature vector included only synthetically enlarged MFCC coefficients

CHAPTER 2

2.1 LITERATURE SURVEY

Various researches have been done for Emotion Analysis using voice. This research is done prior to taking up the project and understanding the various methods that were used previously.

2.1.1 Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition"

The author introduces speech emotion recognition by use of continuous hidden Markov models. Two methods are propagated and compared throughout the paper. Within the first method a global statistics framework of an utterance is classified by Gaussian mixture models using derived features of the raw pitch and energy contour of the speech signal. A second method introduces increased temporal complexity applying continuous hidden Markov models considering several states using low-level instantaneous features instead of global statistics. The paper addresses the design of working recognition engines and results achieved with respect to the alluded alternatives.

Advantages

- It is Text-independent

Disadvantages

- Significant increase in computational complexity
- The need of proper initialization for the model parameters before training.

2.2.2 Surabhi Agrawal, ShabdaDongaonkar, "Emotion recognition from speech using Gaussian Mixture Model and vector quantization"

There is a demand to evaluate the effectiveness of anchor models applied to the multiclass drawback of Emotion recognition from speech. Within the anchor models system, associate in nursing emotion category is characterized by its line of similarity relative to different emotion categories. Generative models like Gaussian Mixture Models (GMMs) are typically used as front-end systems to get feature vectors want to train complicated back-end systems like Support Vector Machine (SVMs) to enhance the classification performance.

There is a tendency to be employing a hybrid approach for recognizing emotion from speech that may be a combination of Vector quantization (VQ) and mathematician Mixture Models GMM. A quick review of labor applied within the space of recognition victimization VQ-GMM hybrid approach is mentioned here.

Advantages

- Unsupervised clustering
- Low computational burden

Disadvantages

- It is Text-independent
- Does not take temporal evolutions into accoun

2.2.3 V. Anoop, P.V. Rao, S.Aruna,"An Effective Speech Emotion Recognition Using Artificial Neural Networks"

The speech signal has emerged as the quickest and the most normal mode of communication between the human beings. As an indispensable method of the human emotional behaviour comprehension, the speech emotion recognition (SER) has evinced zooming significance in the human-centred signal processing. In the current document, the speech signal is identified and distinguished by way of four distinct phases such as the feature extraction; modified cuckoo search-based generating finest weight, feature analysis and the artificial neural network (ANN) classification.

The speech signal categorization, in turn, is performed in the working platform of MATLAB, and outcomes are assessed to ascertain the efficiency in the performance of the system.

Advantages

- Less parameters
- Has very high potential given more hidden layers

Disadvantages

- Network must be retrained when a new emotion is added to the system

2.2 PROPOSED SYSTEM

The system proposed is to be implement the working principle in wearables like smart watches, rings and neck bands. Or it can also be implemented to alert the person by notifying through his phone message or mail. So that the person can be aware of his own emotions and able to control it. The working methodology is that , the audio input which is in the form of real time voices is fed into the system and the emotion can be detected using the MFCC algorithm and notify according to the output. The analysis can be done using an app which is to be created in the smart watches , mobile phones or whatever the device that can be used. Instead of notifying , the app can also suggest some more relaxing techniques for the person based on the emotion detected. For example, If the emotion detected is angry , so that the app can suggest some relaxing techniques or some funny videos to make the person to back into normal. And the other feature can be if the emotion detected is neutral , there need not to be notified. So we can conclude that the system is built for the mental health of a person in such a way that the person can be notified and get alert whenever he is getting emotionally unstable. So the system can built in such a way that the voice signals should be received every 30 seconds or 1 minute, alert the person accordingly.

CHAPTER 3

PROJECT DESCRIPTION

3.1 INTRODUCTION:

The basic Emotion Recognition system from speech comprises of following steps as shown:

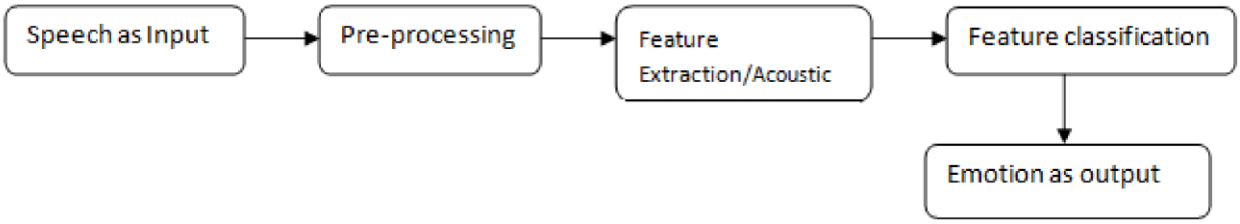


Fig 3.1.1 Speech Emotion Recognition System

The speech samples are taken as input in first phase , if no standard database used then those samples are pre-processed for removing the noise from samples . The main purpose of doing this is to obtain high frequency characteristics of signals.

Feature Extraction and classification:

Another important task in emotion recognition system is Feature Selection and Feature Extraction. Mainly they are classified into two Prosodic Features and Spectral Features. The prosodic features include pitch, Intonation, Duration, The mean Standard deviation, minimum, maximum, range and variance of pitch. The features are selected various feature selection algorithms and they are processed. Spectral features are extracted using MFCC with average rate of recognition of RAVDASS dataset and extracting emotions like sadness, happiness, neutral and Anger.

Feature Selection :

The intuition behind using the acoustic and prosodic features is to summarize the intentional variations observed in humans.

The acoustic features are ,

1. Maximum & Minimum counter ascent energy.
2. Mean and Median values of energy.
3. Mean and Median of energy decline in values.
4. Maximum of pitch frequency
5. Mean and Median of pitch frequency.
6. Maximum duration of pitch in terms of frequency.
7. Mean and Median of first format
8. Rate of change in formats.
9. Speed in voice frames.

3.2 BLOCK DIAGRAM:

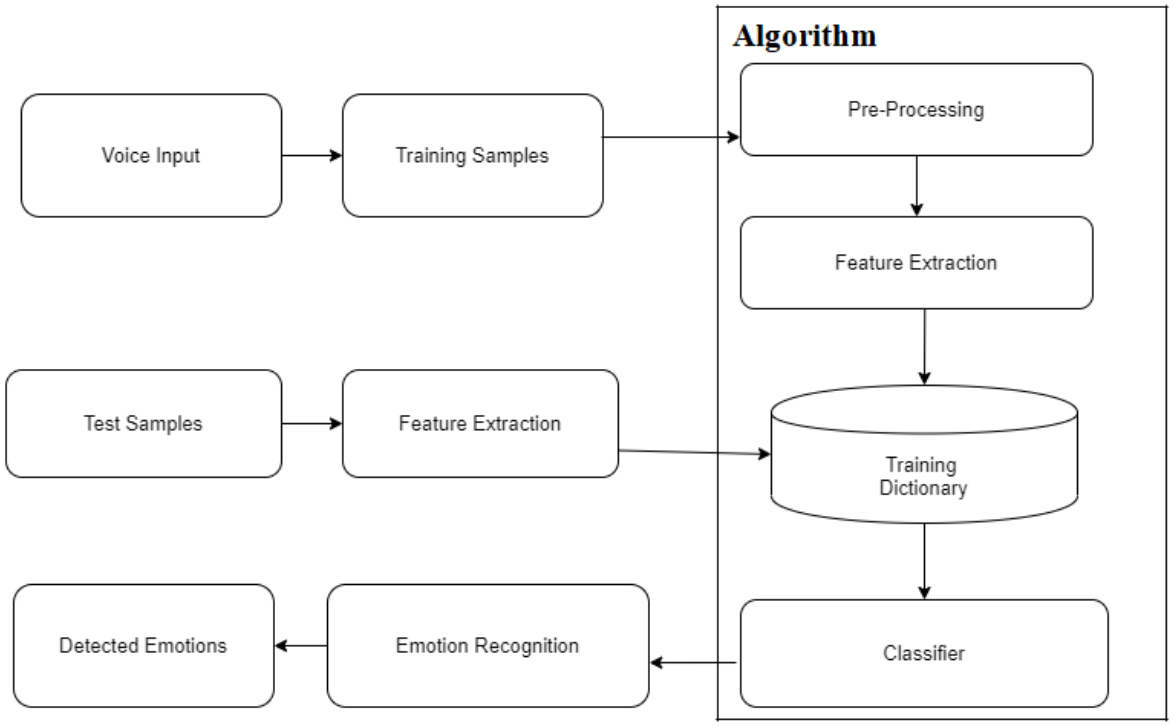


Fig 3.2.1 Block Diagram

The voice input is given as the training sample and then the training samples is given into the algorithm in which it classifies the emotion using MLP classifier. The test samples which is the real time audio is given as the input and the algorithm detects the emotion and the detected emotion is given as the output.

3.3 WORKING PRINCIPLE:

The Mel-frequency Cepstral Coefficients (MFCC) are widely used in audio classification experiments due to its good performance. It extracts and represents features of speech signal. To calculate these coefficients the cosine transform of real logarithm of the short-term spectrum of energy must be done. Then it is performed in melfrequency scale. Further, after pre-emphasizing the speech segments are windowed based on reduction of leakage effect. The concept of windowing is based on multiplying the signal frames by window function w[n]. All the recorded data was labelled into three categories/classes of emotions; neutral, anger and joy. Instead of including all of the different human emotions, we have used only a limited number of emotions as it can clearly reveal the fact that there are certain distinguishing emotion related elements in human speech. So, the output of the data mining is an accuracy percentage with standard deviations from thirty repetitions for each classifier.

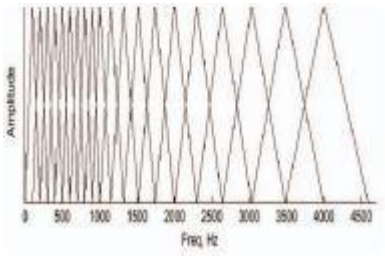


Fig 3.3.1 Filter Bank in Mel Frequency Scale

3.3.1 PSEUDOCODE OF THE ALGORITHM

Algorithm : Mel Frequency Cepstral Coefficient

Input : Signal (Audio signal)

Output : MFCC (MFCC of the given audio signal)

Function : MFCC (parameters)

Initialize the parameters;

Split into frames audio signals;

Apply **Hamming Windowing** to the splitted frames;

Get Spectrum by applying **Fast Fourier Transform** to all frames;

Determine matrix for a Mel-spaced filter bank;

Transform spectrum to **Mel spectrum**;

Obtain **MFCC Vector** for each frame by applying **Discrete Cosine Transform;**

End function

3.3.2 FEATURE EXTRACTION USING MFCC

The acoustic characteristic of the speech signal is Feature. A small amount of data from the speech signal is extracted to analyze the signal without disturbing its acoustic properties. This extracted signal is used for training and testing phases. It comprises of the following steps:

A. Frame Blocking

Processing of speech signals is done in short time intervals called frames with sizes generally between 20 and 40 ms. Overlapping of frames is done to smoothen the transitions between frames by a predefined size. The first frame consists of the first N = 256 (typical value) samples. The second frame begins at M = 100 (typical value) samples after the first frame, and overlap it by N −M samples and so on. This process continues until the entire speech signal is accounted

B. Windowing

After frame blocking, the signal is pass through a windowing function to minimize discontinuities and spectral distortion at the extremes of each frame. The result of windowing a signal x(n) is given by [2],

$$\mathbf{Y(n) = x(n)\, w(n),\ 0 \le n \le N-1 \ldots\ldots..(1)}$$,where w(n) is the window function, $0 \le n \le N-1$, and N is the number of samples in each frame. In this project hamming window is used, which has the form:

$$\mathbf{W(n)=0.54 - 0.46 \cos 2\pi n / N - 1\ ,\ 0 \le n \le N-1 \ldots\ldots(2)}$$

C. Fast Fourier Transform

FFT is performed on the windowing signal to convert it into frequency domain. The discrete Fourier Transform (DFT) over a discrete signal x(n) of N samples, converts each frame of N samples into the frequency domain from its time domain. The FFT is the fast algorithm to implement DFT, which is defined as

$$Xk = \sum x(n)\, e^{-j2\pi kn N}, \quad k = 0, 1, 2, 3, \ldots, N-1 \quad \ldots\ldots\ldots (3)$$

The result after this step is often referred to as spectrum.

D. Mel Frequency Warping

The Mel-frequency scale is linearly spaced for frequencies below 1000 Hz and logarithmically spaced above 1000 Hz. A pitch of 1000 Hz tone, which is equivalent to 1000 Mels when it is above the perceptual hearing threshold level by 40 dB. The following approximate formula can be used to compute the mels for a given frequency f in Hz [10]:

$$Mel(f) = 2595 * \log 10 (1 + f/100) \quad \ldots\ldots\ldots (4)$$

E. Mel Frequency Cepstrum Coefficient (MFCC)

MFCCs are coefficients that represent audio based on perception with their frequency bands logarithmically positioned and mimics the human vocal response.

DECISION MAKING

A. Mean of MFCC

Mean for a data set is termed as arithmetic mean. In this paper, mean of MFCCs is taken to reduce the huge set of values which are obtained from MFCC. When 'x' = {x1, x2, x3, . . . , xn} represents MFCC values and total number of MFCCs is n then the arithmetic mean is taken as,

$$\underline{x = x1 + x2 + x3 + \cdots + xn}$$

$$n$$

B. Standard Deviation

The standard deviation (σ) is a measure of variation or dispersion of a set of data from their mean. A smaller standard deviation indicates that the data to be very close to their mean and a high standard deviation indicates that the data are spread out from their mean and among themselves [8]. In this paper standard deviation of 60 speakers for different emotions, namely sad, happy and anger, are experimentally computed and range of standard deviations for these emotions are optimized as follows : **Sad: 0.1700 to 0.2199; Happy: 0.2200 to 0.2899; Angry: 0.2900 to 0.3999.**

3.4 MERITS AND DEMERITS

3.4.1 Merits

- Negative emotions like anger, sadness and any other emotion leads to the psychological diseases and disorders can be easily eliminated.
- Working is absolutely safe and does not involve any costly equipments.
- Other than being a wearable device, the entire system can be implemented in hospitals and in some health care centers.
- This is Ready-To-Use system, and does not require any special trainings.
- The emotion can be detected within a minute and does not consume more amount of time.

3.4.2 Demerits

- The detection cannot be more accurate when the voice is interrupted with additional noises.
- The system will consume large amount of time in its worst case.
- Sometimes the system finds it difficult to understand the words used by the speaker and detect his/her emotions.

3.5 APPLICATIONS:

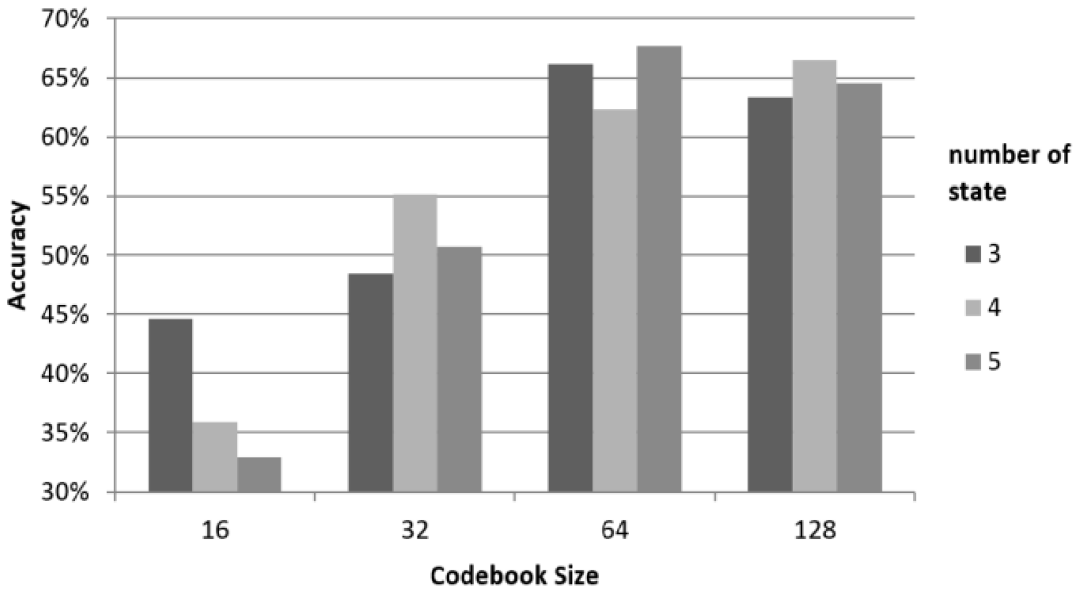
- This new technology develops smart watches which can be applied for Hospitals.
- The user interface provides whole information to promote the smart watch for patients.
- Addition, the particular app can maintain heart rate, blood pressure process so that the better treatment experience for patients can be created.
- This system with MFCCs can be implemented in field of medicine like hospital, health care.
- MFCCs are often seen as a prerequisite for the CNN.
- Further extension of this pattern can be done for other MFCCs applications such as to analyze voice, detect emotion or control individual target through sound waves.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 PERFORMANCE EVALUATION

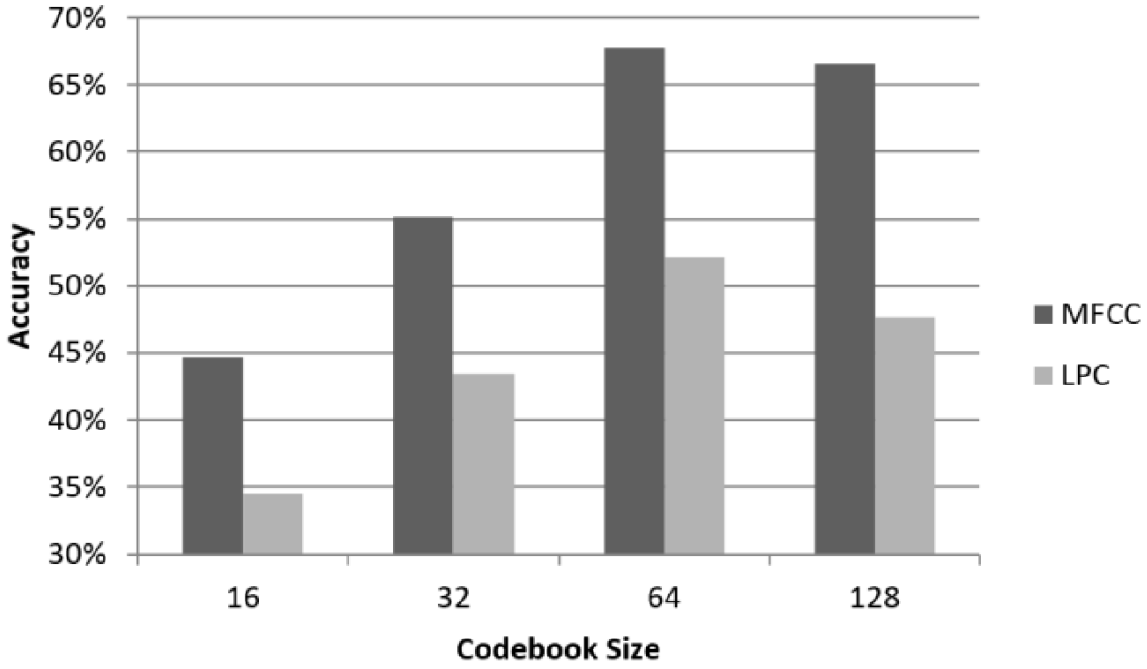
The database has been generated with different people with different voices. The audio signal which is given as input is received through the microphone, and then the audio signals are pre-processed. And the feature extraction using MFCC is done and the Mel-Frequency Cepstrum Coefficient is obtained. Codebook is a representation of feature vector of all voice signals that have passed the clustering process. The following figure is a result of accuracy to a number of state and the size of codebook.



LPC (Linear Predictive Coding) is a method used mostly in speech processing for representing the spectral envelope of digital signal of speech in compressed form, using the information of a linear predictive model. It is one of

the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides highly accurate estimate of speech parameters. LPC is the most widely used method in speech coding and speech synthesis.

Comparison between MFCC and LPC in recognize same dateset with same system is given in the below figure.



Based on the above figure, it can be concluded that MFCC is better feature extraction method in this case. Whatever codebook size used, MFCC has always better accuracy.

4.2 DATASET DESCRIPTION

The dataset used here is audio files. There are totally around 740 audio files by different with different emotions. The dataset given is in the wav form. RAVDESS dataset is used. The audio files are of 1 or 2 seconds duration. The sample image of the dataset is given below. The entire datasets can be accessed using, Speech Emotion Recognition\data .

Fig.4.2.1 Datasets

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

The project explored the idea of detecting the emotional state of a person by speech processing techniques. The study on words and letters under different emotional situations proved that the emotional state can alter the speech signal. It was observed that there are distinguishable features in a speech segment that characterizes each emotion state. After performing the classification tests on a dataset prepared from different subjects, it was observed that it is better considering data from an individual subject rather than a group of people. The development of a software based agent for emotion detection and heart rate analysis can greatly improve telemedicine based systems. Various papers were reviewed and tabulated with the features as corpus names, languages, size in terms of length of the sentence collected as male, female and children voices. Most common emotions searched and extracted are Happiness, Sadness, Anger, Neutral .The extraction rate depends on the classifier used. With new level of contributions it can be concluded that, for creation of new paradigm in Machine learning and Deep Learning with increasing efforts in education, medicine, psychology and agriculture.

APPENDICES

APPENDIX 1

SOFTWARE AND MODULES DESCRIPTION

- SPYDER IDE
- LIBRARIES
  - librosa==0.6.3
  - soundfile==0.9.0
  - pyaudio==0.2.11
  - numpy
  - sklearn

SPYDER IDE

Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the python language. Spyder uses QT for its GUI, and is designated to use either of the PyQT or Pyside python bindings. Spyder integrates with a number of prominent packages in the scientific Python stack, including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy, and Cython as well as other open source software.

The following are some of the more salient features of Spyder:

- It analyzes code to provide automatic code completion, horizontal/vertical splitting, and a go-to-definition.
- It is specifically designed for data scientists; hence, it integrates well with data science libraries like NumPy.
- It allows you to run the IPython console.

- It includes a powerful debugger.
- It contains an integrated documentation browser.

LIBRARIES

1.Librosa==0.6.3

LibROSA is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. To build librosa from source, say python setup.py build . Then, to install librosa, say python setup.py install. To install Librosa Library in Python

1. Using PyPI(Python Package Index) Open the command prompt on your system and write any one of them. pip install librosa sudo pip install librosa pip install -u librosa.

2. Conda Install. If you use conda/Anaconda environments, librosa can be installed from the conda-forge channel.

2. soundfile==0.9.0

SoundFile is an audio library based on libsndfile, CFFI and NumPy. SoundFile can read and write sound files. File reading/writing is supported through libsndfile, which is a free, cross-platform, open-source (LGPL) library for reading and writing many different sampled sound file formats that runs on many platforms including Windows, OS X, and Unix. In a modern Python, you can use **pip install soundfile** to download and install the latest release of SoundFile and its dependencies. Data can be written to the file using **soundfile.write()** or read from the file using **soundfile.read()**. SoundFile can open all file formats that libsndfile supports, for example WAV, FLAC, OGG and MAT files.

3.PyAudio==0.2.11

PyAudio provides Python bindings for PortAudio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to play and record audio on a variety of platforms. To use PyAudio, first instantiate PyAudio using **pyaudio.PyAudio()**, which sets up the portaudio system. To record or play audio, open a stream on the desired device with the desired audio parameters using **pyaudio.PyAudio.open()**.This sets up **pyaudio.stream** to play or record audio. Play audio by writing audio data to the stream using **pyaudio.Stream.write()**, or read audio data from the stream using **pyaudio.Stream.read()**.

4.Numpy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with **Python**. The ndarray (**NumPy** Array) is a multidimensional array **used** to store values of same datatype. These arrays are indexed just like Sequences, starts with zero.

5.Sklearn

Sklearn features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and **is** designed to interoperate with the **Python** numerical and scientific libraries NumPy and SciPy. The **sklearn** library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction\

APPENDIX 2

SOURCE CODE

```
import pyaudio
import os
import wave
import pickle
from sys import byteorder
from array import array
from struct import pack
from sklearn import MLPClassifier
from utils import extract_feature


THRESHOLD = 500
CHUNK_SIZE = 1024
FORMAT = pyaudio.paInt16
RATE = 16000

SILENCE = 30

def is_silent(snd_data):
    "Returns 'True' if below the 'silent' threshold"
    return max(snd_data) < THRESHOLD

def normalize(snd_data):
```

```
    "Average the volume out"
    MAXIMUM = 16384
    times = float(MAXIMUM)/max(abs(i) for i in snd_data)

    r = array('h')
    for i in snd_data:
        r.append(int(i*times))
    return r

def trim(snd_data):
    "Trim the blank spots at the start and end"
    def _trim(snd_data):
        snd_started = False
        r = array('h')

        for i in snd_data:
            if not snd_started and abs(i)>THRESHOLD:
                snd_started = True
                r.append(i)

            elif snd_started:
                r.append(i)
        return r

    # Trim to the left
    snd_data = _trim(snd_data)
```

```
    # Trim to the right
    snd_data.reverse()
    snd_data = _trim(snd_data)
    snd_data.reverse()
    return snd_data

def add_silence(snd_data, seconds):
    "Add silence to the start and end of 'snd_data' of length 'seconds' (float)"
    r = array('h', [0 for i in range(int(seconds*RATE))])
    r.extend(snd_data)
    r.extend([0 for i in range(int(seconds*RATE))])
    return r

def record():
    """
    Record a word or words from the microphone and
    return the data as an array of signed shorts.
    Normalizes the audio, trims silence from the
    start and end, and pads with 0.5 seconds of
    blank sound to make sure VLC  can play
    it without getting chopped off.
    """
    p = pyaudio.PyAudio()
    stream = p.open(format=FORMAT, channels=1, rate=RATE,
        input=True, output=True,
```

```
    frames_per_buffer=CHUNK_SIZE)

num_silent = 0
snd_started = False

r = array('h')

while 1:
    # little endian, signed short
    snd_data = array('h', stream.read(CHUNK_SIZE))
    if byteorder == 'big':
        snd_data.byteswap()
    r.extend(snd_data)

    silent = is_silent(snd_data)

    if silent and snd_started:
        num_silent += 1
    elif not silent and not snd_started:
        snd_started = True

    if snd_started and num_silent > SILENCE:
        break

sample_width = p.get_sample_size(FORMAT)
stream.stop_stream()
```

```
    stream.close()
    p.terminate()

    r = normalize(r)
    r = trim(r)
    r = add_silence(r, 0.5)
    return sample_width, r

def record_to_file(path):
    "Records from the microphone and outputs the resulting data to 'path'"
    sample_width, data = record()
    data = pack('<' + ('h'*len(data)), *data)

    wf = wave.open(path, 'wb')
    wf.setnchannels(1)
    wf.setsampwidth(sample_width)
    wf.setframerate(RATE)
    wf.writeframes(data)
    wf.close()



if __name__ == "__main__":
    # load the saved model (after training)
    model = pickle.load(open("result/mlp_classifier.model", "rb"))
    print("Please talk")
```

```
filename = "test.wav"
# record the file (start talking)
record_to_file(filename)
# extract features and reshape it
features = extract_feature(filename, mfcc=True, chroma=True, mel=True).reshape(1, -1)
# predict
result = model.predict(features)[0]
# show the result !
print("result:", result)
```

APPENDIX 3

SCREEN SHOTS

```
In [1]: runfile('C:/Users/user/Desktop/Mini Project/
Speech Emotion Recognition/train.py', wdir='C:/Users/
user/Desktop/Mini Project/Speech Emotion
Recognition')
[+] Number of training samples: 246
[+] Number of testing samples: 62
[+] Number of features: 180
[*] Training the model...
C:\Users\user\Music\anacondaa\envs\PythonCPU\lib
\site-packages\sklearn\neural_network
\_multilayer_perceptron.py:352: UserWarning: Got
`batch_size` less than 1 or larger than sample size.
It is going to be clipped
  warnings.warn("Got `batch_size` less than 1 or
larger than "
Accuracy: 82.26%
```

Fig A3.1 Trained Model Accuracy

```
Python 3.7.6 (default, Jan  8 2020, 20:23:39) [MSC v.
1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more
information.

IPython 7.12.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/user/Desktop/Mini Project/
Speech Emotion Recognition/test.py', wdir='C:/Users/
user/Desktop/Mini Project/Speech Emotion
Recognition')
Please talk
```

Fig A3.2 Test Model Output

```
In [6]: runfile('D:/Emotional Analysis/test.py', wdir='D:/Emotional Analysis')
Reloaded modules: utils
Please talk
result: neutral
```

Fig A3.3 Emotion detected as Neutral

```
In [4]: runfile('D:/Emotional Analysis/test.py', wdir='D:/Emotional Analysis')
Reloaded modules: utils
Please talk
result: happy
```

Fig A3.4 Emotion detected as Happy

```
In [3]: runfile('D:/Emotional Analysis/test.py', wdir='D:/Emotional Analysis')
Reloaded modules: utils
Please talk
result: sad
```

Fig A3.5 Emotion detected as Sad

```
In [5]: runfile('D:/Emotional Analysis/test.py', wdir='D:/Emotional Analysis')
Reloaded modules: utils
Please talk
result: angry
```

Fig A3.6 Emotion detected as Angry

REFERENCES

1. R. P. a. T. P. Alexander Schmitt, *"Advances in Speech Recognition,"* Springer, pp. 191-2001.
2. A. Joshi, "*Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm*," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 387-392, 2013.
3. D. S. L. N. Akshay, S. Utane, "*Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine*," International Journal of Scientific & Engineering Research, no. 5, pp. 1439-1443, 2013.
4. M Grimm and K. Kroschel, "*Emotion Estimation in Speech Using a 3D Emotion Space Concept[M]*", Robust Speech Recognition and Understanding. InTech, pp. 381-385, 2007.
5. A Harimi, A Shahzadi and A. Ahmadyfard, "*Recognition of emotion using non-linear dynamics of speech*", International Symposium on Telecommunications. IEEE, pp. 446-451, 2015.
6. M E Ayadi, M S Kamel and F. Karray, "*Survey on speech emotion recognition: Features classification schemes and databases[J]*", Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.
7. J. Yuan, L. Shen and F. Chen, "*The acoustic realization of anger fear joy and sadness in Chinese*", Proceedings of ICSLP, pp. 2025-2028, 2002.
8. C. H. Wu and W. B. Liang, "*Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels*", IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10-21, 2011.
9. D. Ververidis and C. Kotropoulos, "*Emotional speech recognition: Resources features and methods*", Speech Communication, vol. 48, no. 9, pp. 1162-1181, 2006.

10. H. K. Palo, M. N. Mohanty and M. Chandra, "*Computational Vision and Robotics*", Advances in Intelligent Systems and Computing, vol. 332, pp. 63-70, 2015.
11. H. A. Murthy and B. Yegnanarayana, "*Formant extraction from group delay function*", Speech Communication, vol. 10, no. 3, pp. 209-221, 1991.
12. Q. Jin, C. Li and S. Chen, "*Speech emotion recognition with acoustic and lexical features"*.. vol. 1, pp. 4749-4753, 2015.