

SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of science Degree in Computer Science

By

SOWNDARYA.S.S (REGNO:38290085)

VIJAYALAKSHMI.S (REGNO:38290102)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**

APRIL 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SOWNDARYA.S.S(REG.NO:38290085),VIJAYALAKSHMI.S(REG.NO:38290102)**, who carried out the project entitled “**SPEECH EMOTION RECOGNITION USING MACHINE LEARNING**” under our supervision from November 2020 to **APRIL 2021**.

Internal Guide

Dr.S PRINCE MARY, M.E., Ph.D.,

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **SOWNDARYA.S.S (REG.NO:38290085)** hereby declare that the Project Report entitled " **SPEECH EMOTION RECOGNITION USING MACHINE LEARNING**" done by me under the guidance of **Dr. S. Prince Mary, M.E., Ph.D.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of science Degree in Computer Science

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA, M.E., Ph.D.**, Dean, School of Computing and **Dr. S. VIGNESHWARI, M.E., Ph.D.**, and **Dr. L. LAKSHMANAN, M.E., Ph.D.**, Head of the Department, Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S. Prince Mary, M.E., Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways For the Completion of the project.

ABSTRACT

Emotion is a complex state of human being that depicts the physical, physiological or mental condition of a person. According to the human sciences, emotion is a mental process of neural mechanisms and disorders. Digital processing of speech signal is very important for high-speed and precise automatic voice recognition technology. Nowadays it is being used for health care, telephony military and people with disabilities therefore the digital signal processes such as Feature Extraction and Feature Matching are the latest issues for study of voice signal. In order to extract valuable information from the speech signal, make decisions on the process, and obtain results, the data needs to be manipulated and analyzed. Basic method used for extracting the features of the voice signal is to find the Mel frequency cepstral coefficients. Mel-frequency cepstral coefficients (MFCCs) is the coefficients that collectively represent 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. After calculating feature, neural networks are used to model the speech recognition. Based on the speech model the system decides whether or not the uttered speech matches what was prompted to utter.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	v
	LIST OF FIGURES	8
	LIST OF ABBREVIATIONS	9
1	INTRODUCTION	10
	1.1 INTRODUCTION	10
	1.2 OVERVIEW OF PROJECT	11
2	LITERATURE SURVEY	13
3	AIM AND SCOPE OF PRESENT INVESTIGATION	17
	3.1 AIM OF THE PROJECT	17
	3.2 OBJECTIVE OF THE PROJECT	17
	3.3 SYSTEM REQUIREMENTS	17
	3.3.1 SOFTWARE REQUIRED	17
	3.3.2 PACKAGES USED	17
	3.4 OVERVIEW OF PLATFORM	18
	3.4.1 ANACONDA PROMPT	18
	3.4.2 JUPYTER NOTEBOOK	18
	3.4.3 NUMPY	18
	3.4.4 MATPLOTLIB	19
	3.4.5 PANDAS	19
	3.4.6 PATHLIB	19
	3.4.7 SCIKIT LEARN (SKLEARN)	19
	3.4.8 OS	19
	3.4.9 SCIPY	20
	3.4.10 PLOTLY	20
	3.4.11 LIBROSA	20
4	EXPERIMENTAL OR MATERIAL ANALYSIS	21

	4.1 DESIGN METHODOLOGY	21
	4.1.1 EXSISTING SYSTEM	21
	4.1.2 PROPOSED SYSTEM	21
	4.2 MODULE DESCRIPTION	22
	4.2.1 EMOTION RECOGNITION	22
	4.2.2 FEATURE EXTRACTION	23
	4.2.2.1 MLP NETWORK	23
	4.2.2.2 MFCC NETWORK	23
	4.2.2.3 MEL SPECTROGRAM	24
	4.3 ARCHITECTURE	24
	4.3.1 ARCHITECTURE EXPLANATION	25
5	RESULT AND PERFORMANCE ANALYSIS	26
	5.1 VISUALISATION	26
	5.1.1 WAVE AND SPECTROGRAM	26
	5.1.1 MFCC	27
	5.2 SILENCE REMOVAL	29
	5.3 RESAMPLING	30
6	CONCLUSION AND FUTURE ENHANCEMENT	32
	6.1 CONCLUSION	32
	6.2 FUTURE ENHANCEMENT	32
	REFERENCE	33
	APPENDIX	34
	SOURCE CODE	

LIST OF FIGURES

FIGURE NO:	FIGURE NAME	PAGE NO
1.1	Categorization of Emotions	11
2.1	Comparision of Speech Emotion Recognition in Terms of Recognition rate	15
4.2.1	Speech emotion recognition	22
4.3.1	Architecture diagram	24
5.1.1	Raw wave form	26
5.1.2	Spectrogram	27
5.1.2.1	Mel spectrogram	28
5.2.1	Silence removal	29
5.2.1	Raw wave form after removal of sound	30
5.2.3	Spectrogram after removal of sound	30
5.3.1	FFT of recording sample	31

LIST OF ABBREVIATIONS

SER	Speech Emotion Recognition
MLP	Multi Layer Perceptron
MSE	Mean Square Root
ML	Machine Learning
NumPY	Numerical python
FFT	Fast Fourier Transform
ASR	Automatic Speech Recognition
AI	Artificial Intelligence
VAD	voice Activity Detection
SVM	support vector machine
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
RBF	Radial Basis Function
OS	Operating system
MFCC	Mel-Frequency Cepstral Coefficients
HZ	Hertz
STE	Short Term Energy

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Emotion is a complex state of human being that depicts the physical, physiological or mental condition of a person. According to the human sciences, emotion is a mental process of neural mechanisms and disorders. The expression of emotion may vary in linguistics as per the meaning of sounds. Numerous research activities have been carried out by researchers during the last few decades in understanding human emotion and for its effective recognition. It may help the society in many ways. The speech emotion recognition (SER) system can serve as a tool for assessing the efficiency of a call center executive dealing with potential customers. Intelligent on-line tutors can take clues from the emotional state of the students in delivering fruitful lectures by finding new methods of teaching or motivation. Designing personal robots for entertainment or education purpose is another such significant application. The recognition system can alert the security organization by identifying the stress from a phone conversation of security guards on duty. Interactive movies, and bio-medical engineering, etc. are few other vivid application domains that will be benefited from this type of study.

A categorization of emotions into three broad areas such as primary, secondary and tertiary is shown in Fig. 1.1. However, automatic SER is a complex research area because of the dynamism involved due to a varied

degree of psychological perception to the speech emotion, which is further categorized into dimensional (continuous) or categorical (discrete) emotions. The complexity in research is further aggravated due to unavailability of universal agreement on the basic taxonomy of emotion.

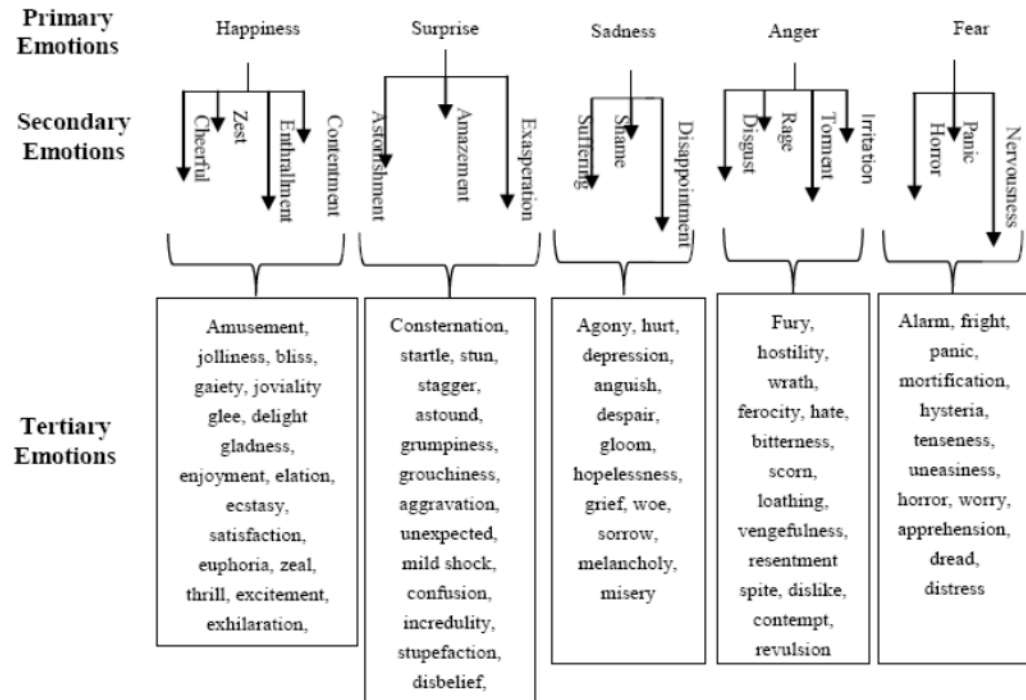


Fig.1.1 Categorization of Emotions

1.2 OVERVIEW OF PROJECT

The idea behind creating this project is to build a machine learning model that could detect emotions from the speech we have with each other all the time. Nowadays personalization is something that is needed in all the things we experience everyday. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each file was rated 10 times on emotional validity, intensity, and genuineness. This process about loading the dataset in python which involves extracting audio features, such as obtaining different features such as power, pitch

and vocal tract configuration from the speech signal, librosa file is used to do that. The set is true for 25% of the entries, as it is hard to predict the exact emotions. A classifier will be defined which is the MLP (Multi-Layer Perceptron). The model is trained for many entries and it does require a lot of training. finally the emotion of the audio file is determined.

CHAPTER 2

LITERATURE SURVEY

The researchers have used a number of different emotions for classification task based on the application area. It is hard to find concrete evidence on the number of emotions to be included for better classification accuracy. However, the recognition system based on large numbers of emotional states is likely to be complex. This has made most of the researchers to focus on the classification of emotions ranging from three to six classes. For our purpose, a maximum of five numbers of emotional states has been considered.

Due to different shape and size of the vocal tract, the SER system may not perform well as among different age groups. Children, in particular, have underdeveloped vocal tract and vocal folds. This leads to higher formants, non-linear formant variations, speaking rate, fundamental frequency and acoustic correlations for children than adults. This has motivated the authors to analyze children speech emotion in the current scenario as the work progresses.

Efficient feature extraction is arguably the long-standing challenge for emotion identification of the speech signal. It is necessary to derive new feature extraction by way of modification, reduction, combinational framework, and in conjunction to eliminate ambiguous emotion management. Hence, the majority of our work is focused in this area. The proposed techniques are tested for possible enhancement in recognition accuracy and a reduction in the mean square error (MSE). A comparison of the proposed and the state-of-art features in this area has been made using different databases and classification methods to validate our work.

Wu et al. proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for convey important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) are adopted. Berlin and Vera am Mittag (VAM) are employed to evaluate MSFs. In experimental result, the MSFs display capable performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC). When MSFs utilized augment prosodic features, there is a considerable improvement in performance of recognition. Furthermore overall recognition rate of 91.6% is achieved for classification].

Albornoz et al. investigate a new spectral feature in order to determine emotions and to characterize groups. In this study based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75%

Generally, analysis on speech emotion recognition aims to increase the recognition rate as well as accuracy. the current available methods are targeted the speech emotion recognition systems, and these are evaluated

with its classifier, features set, and recognition rate and on different dataset levels. the SVM has many motivating properties such as easy to fulfilment because of mathematical basis compare to simple classifier such as quadratic discriminant analysis (QDC) . Normally, SVM is employed as alone and also with combination of other classifier such as ANN and RBF to reduce the dimensional

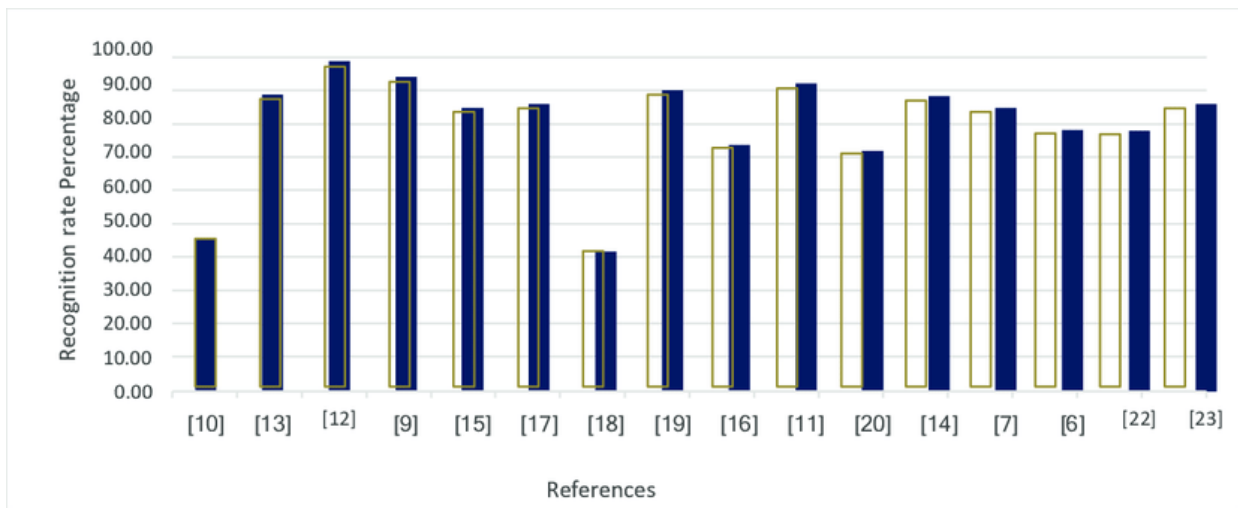


FIG. 2.1 Comparision of Speech Emotion Recognition in Terms of Recognition rate

The above fig.2.1 depicts the recognition rate percentage of the emotion using various classifiers,

[10]→ SVM classifier

[13]→SVM & ANN

[12] →HMM

[9] →SVM & RBF

[15] →Decision tree & random forest Linguistic

[17] → GMM, SVM, MLP and MDT(Meta decision tree)
Prosodic information and semantic labels (SL,)

[18] → k-NN & linear discriminate Fundamental frequency (F0)

[19] → Bayesian learning framework Energy,

[16] → HMM, GMM, MLP and hierarchical model Mean of the log-spectrum (MLS).

[11] → SVM Zero crossing rate, [20] Bayesian Logistic Regression,

[14] → k-NN Jitter, shimmer, formants, LPC, LPCC, MFCC, LFPC, PLP, and Rasta-PLP, SFS and SBS

[7] → LS-SVR Short-time Energy , Pitch , Short-time zero crossing rate

[6] → G GMVAR [22] Binary classifier & QDC Prosodic [23] k-NN Acoustic features

CHAPTER 3

AIM AND SCOPE OF PRESENT INVESTIGATION

3.1 AIM OF THE PROJECT

Our aim is to build a machine learning model using an MLPClassifier with the help of libraries such as librosa, soundfile, and sklearn (among others). This will be able to recognize emotion from sound files. Data will be loaded, features will be extracted from the data, then the data will be splitted into training and testing sets. Then, MLPClassifier is initialized to train the model. Finally, the emotion of the audio file is determined.

3.2 OBJECTIVE OF THE PROJECT

Emotion recognition from speech is one of the major challenges in the field of human computer interaction. The formulation of powerful emotion recognition systems are thus beneficial and the objective of a good emotion recognition system is to be able to mimic human perception in the way that humans are able to detect emotions such as anger, sadness, and happiness.

3.3 SYSTEM REQUIREMENTS

3.3.1 SOFTWARE REQUIRED

- Anaconda prompt
- Jupyter notebook

3.3.2 PACKAGES USED

- Numpy

- Matplotlib
- Pandas
- Pathlib
- Scikit learn(sklearn)
- Os
- Scipy
- Plotly
- Librosa

3.4 OVERVIEW OF PLATFORM

3.4.1 ANACONDA PROMPT

Anaconda command prompt is just like command prompt, Anaconda Prompt is a command line shell (a program where you type in commands instead of using a mouse). The black screen and text that makes up the Anaconda Prompt doesn't look like much, but it is really helpful for problem solvers using Python.

3.4.2 JUPYTER NOTEBOOK

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document.

3.4.3 Numpy

NumPy is a Python library used for working with arrays. NumPy is

a Python library that provides a simple yet powerful data structure: the n-dimensional array. This is the foundation on which almost all the power of Python's data science toolkit is built.

3.4.4 MATPLOTLIB

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3.4.5 PANDAS

Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. The Pandas library provides a really fast and efficient way to manage and explore data.

3.4.6 PATHLIB

Pathlib is an object oriented interface to the filesystem and provides a more intuitive method to interact with the filesystem in a platform agnostic and pythonic manner. The pathlib is a Python module which provides an object API for working with files and directories. The pathlib is a standard module. Path is the core object to work with files.

3.4.7 SCIKIT LEAR (SKLEARN)

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine **learning** and statistical modeling including classification, regression, clustering and dimensionality reduction.

3.4.8 OS

The OS module in Python provides functions for interacting with the operating system. This module provides a portable way of using operating system-dependent functionality

3.4.9 SCIPY

SciPy is a scientific computation library that uses NumP underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing.

3.4.10 PLOTLY

Plotly is a collaborative, web-based graphing and analytics platform .Plotly allows users to import, copy and paste, or stream data to be analyzed and visualized.The Plotly Python graphing library is a scientific graphing library.

3.4.11 LIBROSA

librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It is the starting point towards working with audio data at scale for a wide range of applications such as detecting voice from a person to finding personal characteristics from an audio. Librosa can be defined as a package which is structured as collection of submodules which further contains other functions.

CHAPTER 4

EXPERIMENTAL OR MATERIAL METHODS

4.1 DESIGN METHODOLOGY

4.1.1 Existing System:

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better.. The data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data. feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address the several data representation and data quality issues. the core of an ML project where an algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to. evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms. Comparison results help to choose the appropriate ML algorithm most relevant to the problem

4.1.2 PROPOSED SYSTEM

Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterise speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognise a particular emotion accurately. The North American English language open source dataset was divided into training and testing manually. Speaker vocal tract information, represented by Mel-frequency cepstral coefficients (MFCC), was extracted from the audio samples in training dataset. Pitch, Short Term

Energy(STE), and MFCC coefficients of audio samples in emotions anger, happiness, and sadness were obtained. These extracted feature vectors were sent to the classifier model. The test dataset will undergo the extraction procedure following which the classifier would make a decision regarding the underlying emotion in the test audio.

4.2 MODULE DESCRIPTION

4.2.1 EMOTION RECOGNITION

Three major modules that shape the success of a SER system are a genuine database, reliable features that correlate the acoustic variety of speech emotions, and effective classifiers. The basic modules of a SER system are shown in Fig. 4.2.1

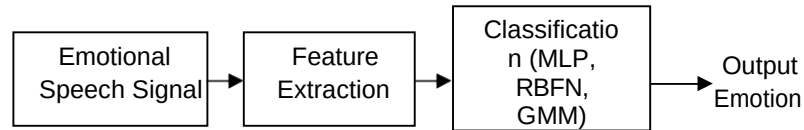


FIG. 4.2.1 Speech Emotion Recognition

The emotional speech databases used for research purposes have been generalized into three broad categories. These categories are natural, induced and simulated. Features extracted from natural and spontaneous emotion can faithfully represent the states with greater ecological validity. These are collected from a real life scenario during TV news or interviews, interactive games or talk shows.

However, bad and uncontrollable recording conditions, limited sources, the possibility of overlapped or mixed emotions, ethical, legal and copyright complications are some of the inherent issues that make the data procurement cumbersome. Induced emotional utterances are collected by putting the subjects under certain stress or situation, applying film or slide induced emotions and sometimes simulating the subjects with psychoactive drugs, card games, computers, robots or electronic gadgets. Professionals mimic these emotions as they are capable of expressing the desired emotion more reliably with much more control. The utterances are better emotionally portrayed than induced emotions. These emotions fall under the discrete emotional description. Although this method of collection is favored by psychologists, it cannot guarantee the availability of desired emotional states. Since the human being tends to react differently, even under similar situations, these emotions are often exaggerated or biased. The recording environment is often monolog and not dialog based. The text materials are read rather than spoken. Finally, there is every possibility of the actor's cultural background being inflicted during simulation of the desired emotions.

4.2.2 FEATURE EXTRACTION

4.2.2.1 MLP NETWORK

The multilayer perceptron comes under supervised learning category as it needs the desired output during learning. The aim is to develop a model so as to map the input into the output based on historical data during training. The constructed model is then utilized to provide an output with unknown data during testing. It works on the principle of back-propagation algorithm. The basic MLP structure comprises of three subsidiary sections as input, hidden and output layer. The network operates in two phases, i.e. forward pass (phase-I) and the backward pass (phase-II). When the desired input is presented to the network, it is propagated from the input to the output layer via the hidden layer during the forward pass. Consequently, a set of output is generated and are termed as the actual response of the network. The synaptic weights and biases are all remained unchanged during this phase.

4.2.2.2 MFCC NETWORK

Speech recognition mainly focuses on training the system to recognize an individual's unique voice characteristics. The most popular feature extraction technique is the Mel Frequency Cepstral Coefficients called MFCC as it is less complex in implementation and more effective and robust under various conditions. MFCC is designed using the knowledge of human auditory system. It is a standard method for feature extraction in speech recognition. Steps involved in MFCC are Pre-emphasis, Framing, Windowing, FFT, Mel filter bank, computing DCT.

4.2.2.3 MEL SPECTROGRAM

A spectrogram is a visual way of representing the signal strength, or "loudness", of a signal over time at various frequencies present in a particular waveform. Not only can one see whether there is more or less energy at, for example, 2 Hz vs 10 Hz, but one can also see how energy levels vary over time. A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.

4.3 ARCHITECTURE

The architecture diagram is as follow,

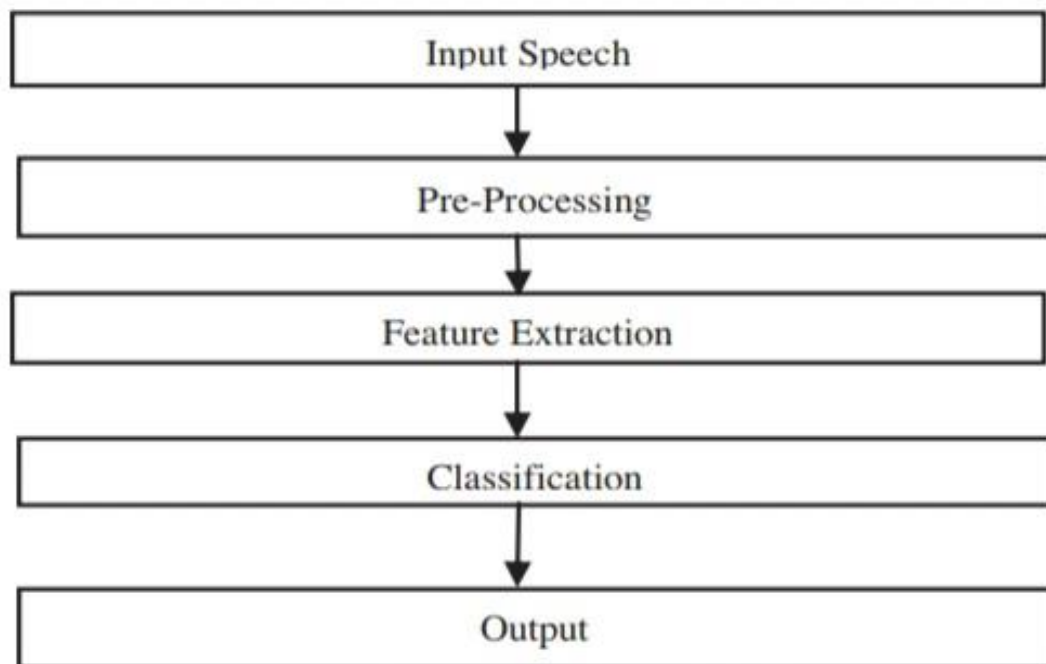


FIG.4.3.1 ARCHITURE DIAGRAM

4.3.1 ARCHITECTURE EXPLANATION

The proposed human emotion recognition system is of five components: input speech signal, pre-processing, feature extraction and selection, classification and finally emotions recognition. The Architecture of Emotional speech recognition system is as shown in figure 4.3.1. The Accuracy of the Emotional speech recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations . The emotion recognizer system identifies the emotion state of the input speech signal and displays the corresponding emotions of that particular speech. The speech processor, deals with the emotion features selection, speech preprocessing, and extraction algorithm. Determining emotion features is a vital issue in the emotion recognizer design. The emotion recognition result is strongly depending on the emotional features and which provided to various types of classifier that have been used to represent the emotion We want to extract Feature vectors from the input speech. We used the (MFCC) algorithm to extract the features. Extraction of MFCC we transform the input waveform into a sequence of acoustic feature vectors, each vector representing the information in a small time window of the signal. Now that we have a digitized, quantized representation of the waveform, now we are prepared to extract MFCC features.

CHAPTER 5

RESULTS AND PERFORMANCE ANALYSIS

5.1 VISUALISATION

There are two theories of a human hearing - place (frequency-based) and temporal. In speech recognition, I see two main tendencies - to input spectrogram (frequencies), and more sophisticated features MFCC - Mel-Frequency Cepstral Coefficient.

5.1.1 WAVE AND SPECTROGRAM

Define a function that calculates spectrogram. The logarithm of spectrogram values are taken. It will make the plot much more clear, moreover, it is strictly connected to the way people hear. Assure that there are no 0 values as input to logarithm.

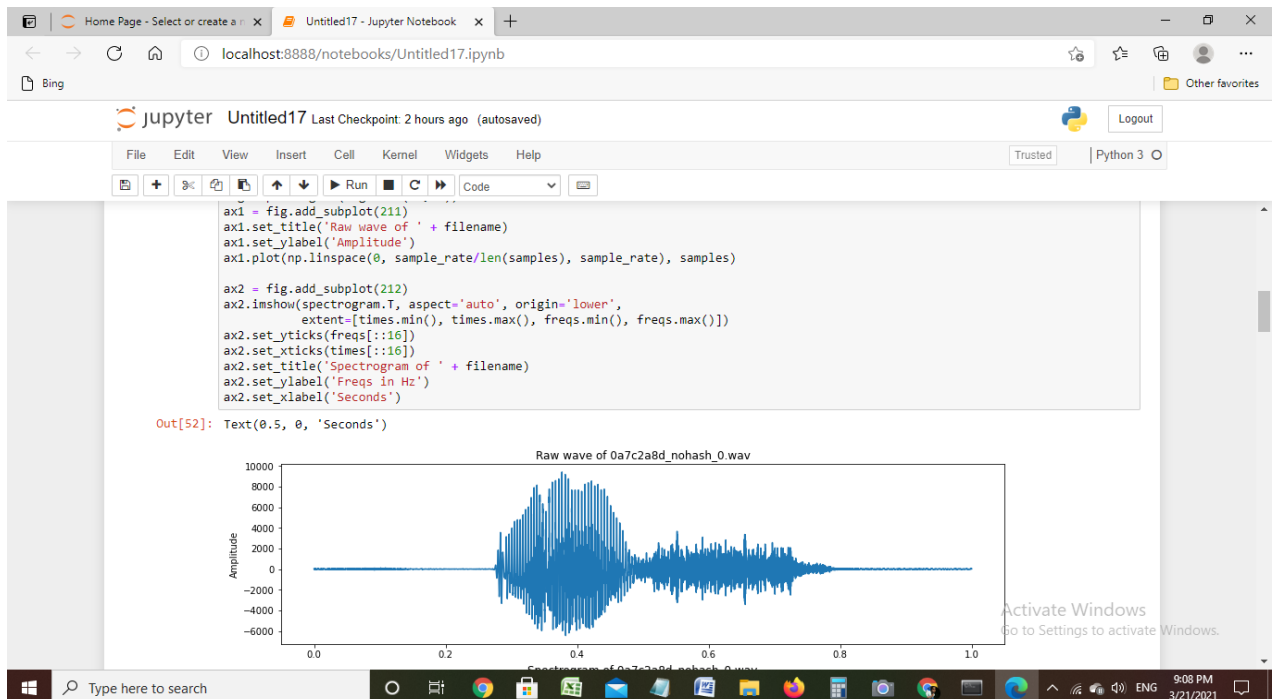


FIG.5.1.1 RAW WAVE FORM

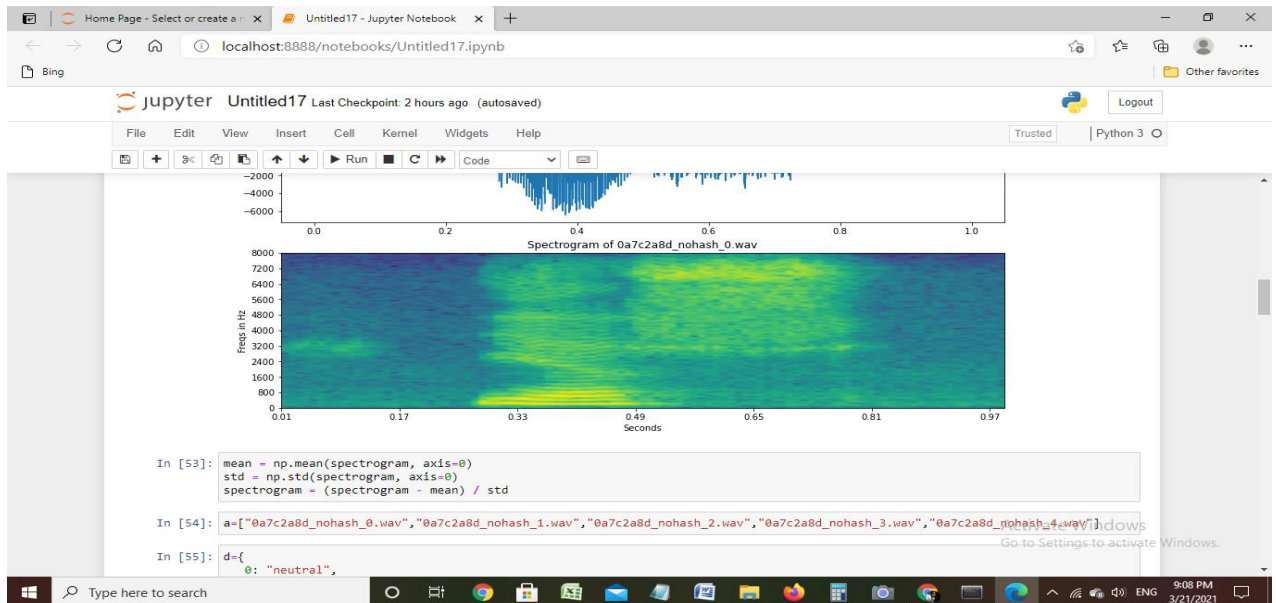


FIG.5.1.2 SPECTROGRAM

5.1.2 MFCC

An audio signal is constantly changing, so to simplify things to assume that on short time scales the audio signal doesn't change much. This is why the signal is framed into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.

The next step is to calculate the power spectrum of each frame. This is motivated by the human cochlea (an organ in the ear) which vibrates at different spots depending on the frequency of the incoming sounds. Depending on the location in the cochlea that vibrates (which wobbles small hairs), different nerves fire informing the brain that certain frequencies are present. Our periodogram estimate performs a similar job for us, identifying which frequencies are present in the frame.

The periodogram spectral estimate still contains a lot of information not required for Automatic Speech Recognition (ASR). In particular the cochlea can not discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. For this reason we take clumps of periodogram bins and sum them up to get an idea of how much energy exists in various frequency regions. This is performed by our Mel filterbank: the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. As the frequencies get higher our filters get wider as we become less concerned about variations. We are only interested in

roughly how much energy occurs at each spot. The Mel scale tells us exactly how to space our filterbanks and how wide to make them. See below for how to calculate the spacing.

Once we have the filterbank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the perceived volume of a sound we need to put 8 times as much energy into it. This means that large variations in energy may not sound all that different if the sound is loud to begin with. This compression operation makes our features match more closely what humans actually hear. The logarithm allows us to use cepstral mean subtraction, which is a channel normalisation technique.

The final step is to compute the DCT of the log filterbank energies. There are 2 main reasons this is performed. Because our filterbanks are all overlapping, the filterbank energies are quite correlated with each other. The DCT decorrelates the energies which means diagonal covariance matrices can be used to model the features in e.g. a HMM classifier. But notice that only 12 of the 26 DCT coefficients are kept. This is because the higher DCT coefficients represent fast changes in the filterbank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them.

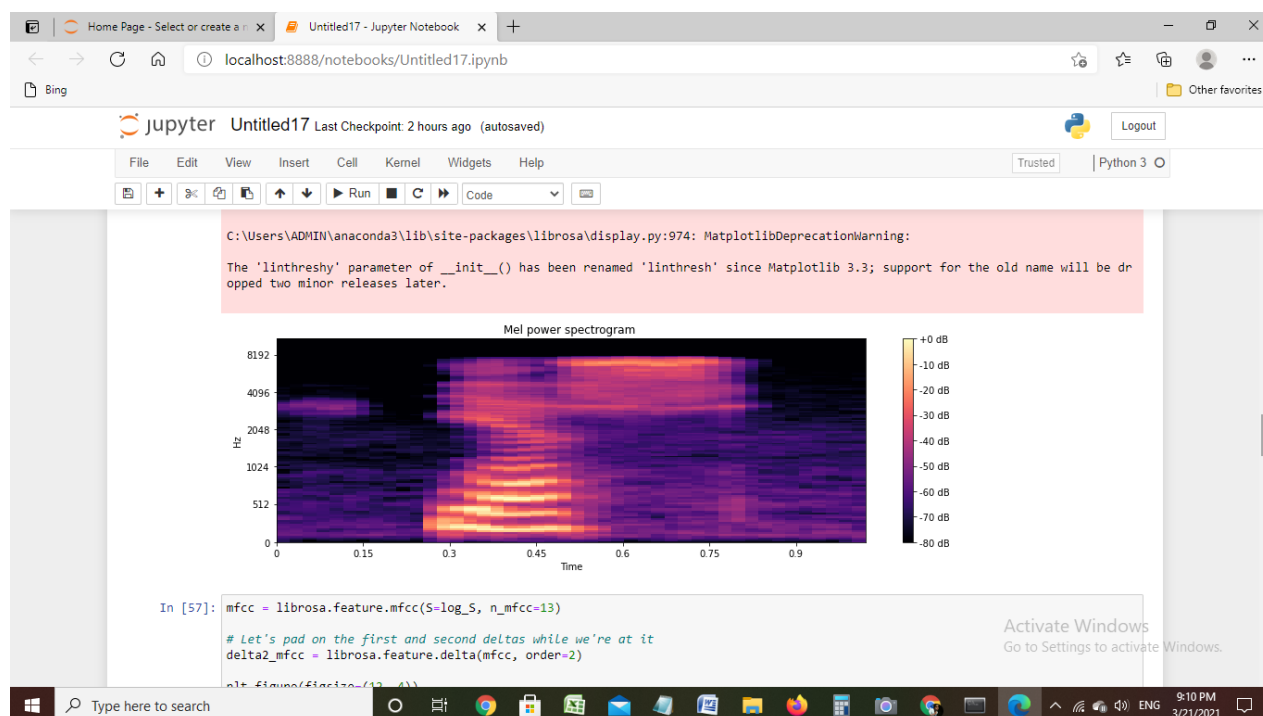


FIG.5.1.2.1 MEL POWER SPECTROGRAM

5.2 SILENCE REMOVAL

Here consider some VAD (Voice Activity Detection) that will be really useful. Although the words are short, there is a lot of silence in them. A decent VAD can reduce training size a lot, accelerating training speed significantly. cut a bit of the file from the beginning and from the end. and listen to it again.

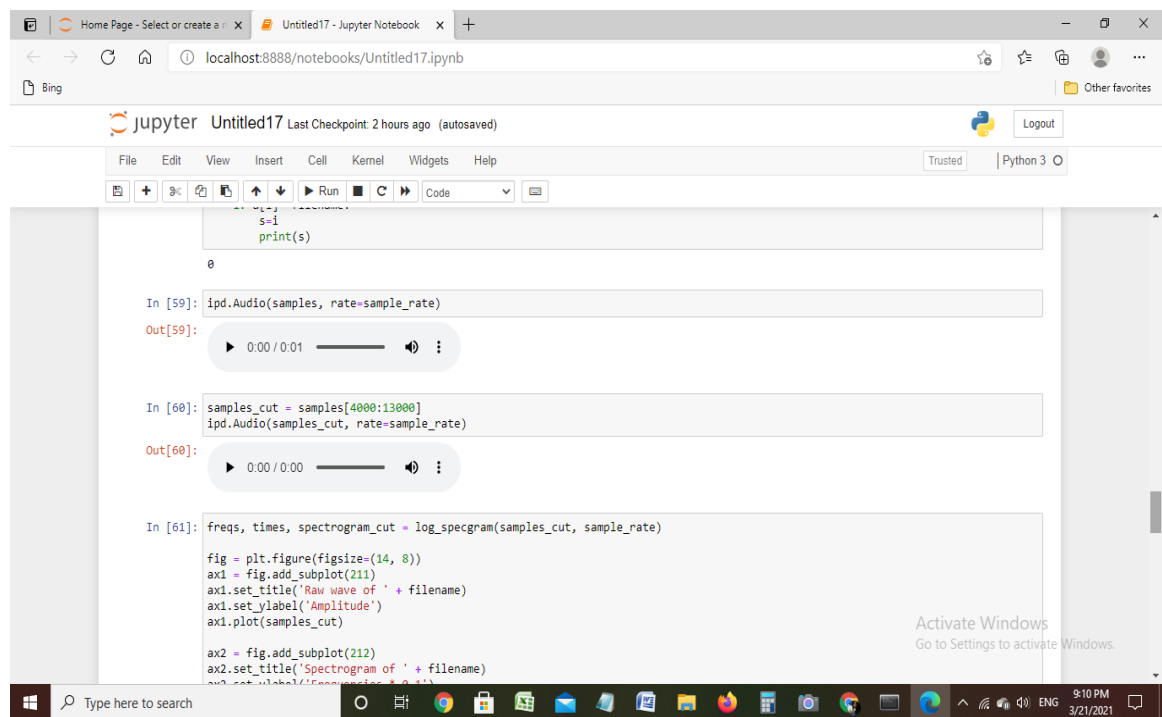


FIG.5.2.1 SILENCE REMOVAL

The entire word can be heard. It is impossible to cut all the files manually and do this basing on the simple plot. plot it again, together with guessed alignment of 'y' 'e' 's' graphemes.

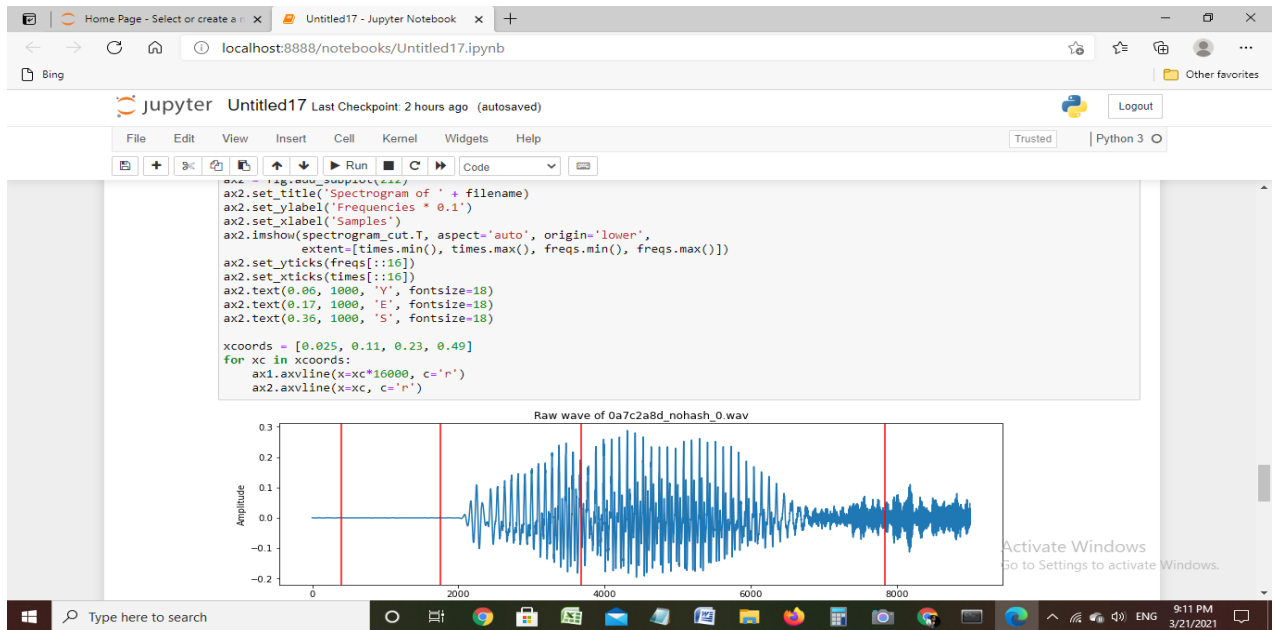


FIG.5.2.2 RAW WAVE FORM AFTER REMOVAL OF SILENCE

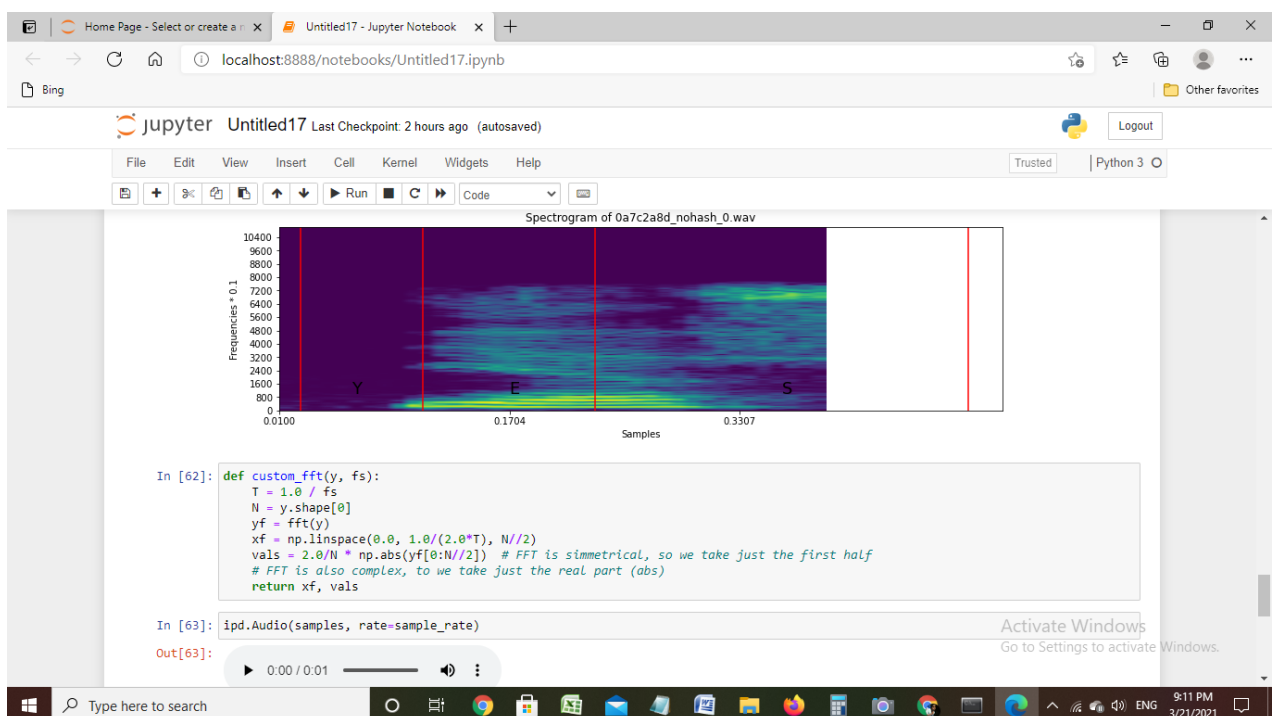


FIG.5.2.3 SPECTROGRAM AFTER REMOVAL OF SILENCE

5.3 RESAMPLING – DIMENSIONALITY REDUCTION

Another way to reduce the dimensionality of data is to resample recordings. the recording don't sound very natural, because they are sampled with 16k frequency. However, the most speech related frequencies are presented in smaller band. That's why people can still understand another person talking to the telephone, where GSM signal is

sampled to 8000 Hz. Summarizing, the dataset can be resampled to 8k. We will discard some information that shouldn't be important, and we'll reduce size of the data.

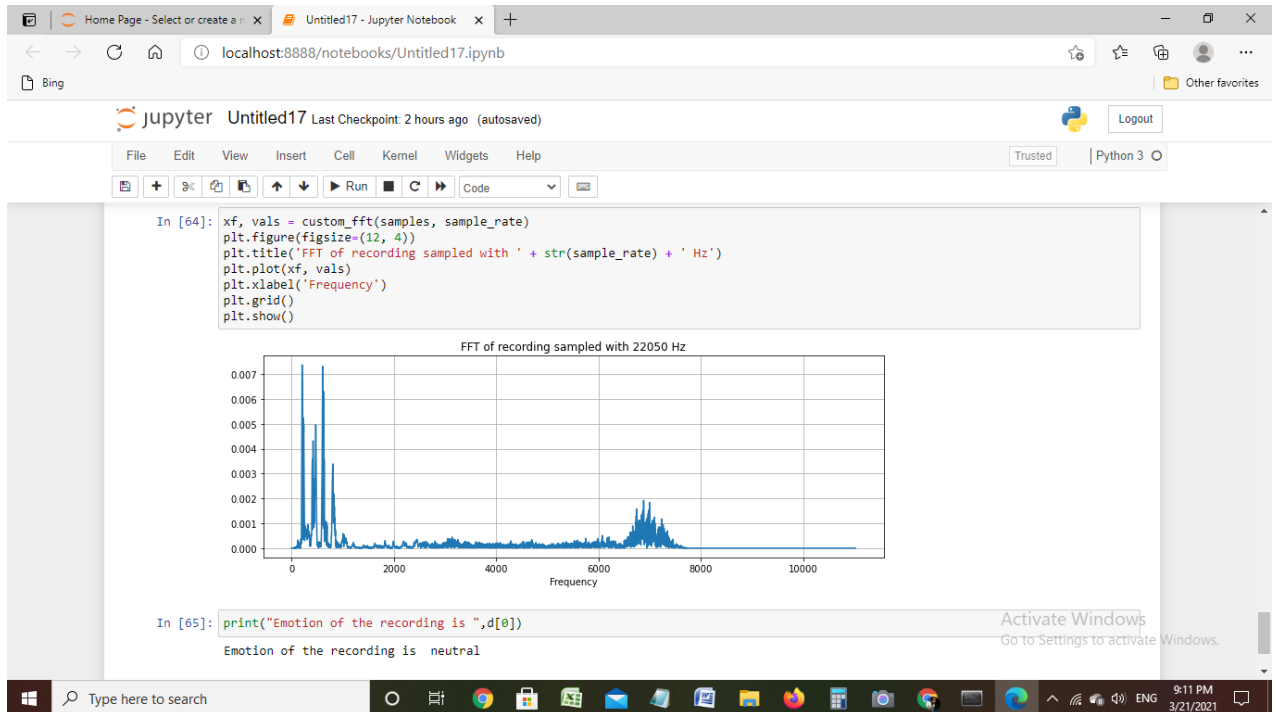


FIG.5.3.1 FFT OF RECORDING SAMPLE

Based on the pitch, tone, frequency of the audio file that is given as the input, the emotion of the recording is determined.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated devices work based on voice commands from the user. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system. we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

6.2 FUTURE ENHANCEMENT

For future advancements, the proposed project can be further modeled in terms of efficiency, accuracy, and usability. Additional to the emotions, the model can be extended to recognize feelings such as depression and mood changes. Such systems can be used by therapists to monitor the mood swings of the patients. A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem of emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. A sentiment detection using vocabulary, can be integrated with speech emotion detection to identify a possible sarcasm. Therefore, in the future, there would emerge many applications of a speech-based emotion recognition system.

REFERENCES

- [1]. Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4), 290-296.
- [2]. Harár, P., Burget, R., & Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In *Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on* (pp. 137-140). IEEE.
- [3]. Rawat, A., & Mishra, P. K. (2015). Emotion recognition through speech using neural network. *Int. J*, 5, 422-428.
- [4]. Sun, W., Zhao, H., & Jin, Z. (2017). An efficient unconstrained facial expression recognition algorithm based on Stack Binarized Autoencoders and Binarized Neural Networks. *Neuro computing*, 267, 385-395
- [5] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 186–202, Jan. 2015.
- [6] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
- [7] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.
- [8] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, May 2009.

APPENDIX

SOURCE CODE

```
import os
from os.path import isdir, join
from pathlib import Path
import pandas as pd

# Math
import numpy as np
from scipy.fftpack import fft
from scipy import signal
from scipy.io import wavfile
import librosa

from sklearn.decomposition import PCA

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns
import IPython.display as ipd
import librosa.display

import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import pandas as pd

%matplotlib inline
x=input()
train_audio_path = 'C:\\Users\\ADMIN\\one audio file\\'
filename = x
sample_rate, samples = wavfile.read(str(train_audio_path) + filename)
def log_specgram(audio, sample_rate, window_size=20,
                 step_size=10, eps=1e-10):
```

```

nperseg = int(round(window_size * sample_rate / 1e3))
noverlap = int(round(step_size * sample_rate / 1e3))
freqs, times, spec = signal.spectrogram(audio,
                                         fs=sample_rate,
                                         window='hann',
                                         nperseg=nperseg,
                                         noverlap=noverlap,
                                         detrend=False)

return freqs, times, np.log(spec.T.astype(np.float32) + eps)
freqs, times, spectrogram = log_specgram(samples, sample_rate)

fig = plt.figure(figsize=(14, 8))
ax1 = fig.add_subplot(211)
ax1.set_title('Raw wave of ' + filename)
ax1.set_ylabel('Amplitude')
ax1.plot(np.linspace(0, sample_rate/len(samples), sample_rate), samples)

ax2 = fig.add_subplot(212)
ax2.imshow(spectrogram.T, aspect='auto', origin='lower',
           extent=[times.min(), times.max(), freqs.min(), freqs.max()])
ax2.set_yticks(freqs[::16])
ax2.set_xticks(times[::16])
ax2.set_title('Spectrogram of ' + filename)
ax2.set_ylabel('Freqs in Hz')
ax2.set_xlabel('Seconds')
mean = np.mean(spectrogram, axis=0)
std = np.std(spectrogram, axis=0)
spectrogram = (spectrogram - mean) / std
a=["0a7c2a8d_nohash_0.wav","0a7c2a8d_nohash_1.wav","0a7c2a8d_nohash_2.wav","0a7c2a8d_nohash_3.wav","0a7c2a8d_nohash_4.wav"]
d={
    0: "neutral",
    1: "calm",
    2: "happy",
    3: "sad",
    4: "angry",
    5: "fearful",
    6: "disgust",

```

```

7: "surprised"
}
# From this tutorial
# https://github.com/librosa/librosa/blob/master/examples/LibROSA%20demo.ipynb
samples, sample_rate = librosa.load(str(train_audio_path)+filename)
S = librosa.feature.melspectrogram(samples, sr=sample_rate, n_mels=128)

# Convert to log scale (dB). We'll use the peak power (max) as reference.
log_S = librosa.power_to_db(S, ref=np.max)

plt.figure(figsize=(12, 4))
librosa.display.specshow(log_S, sr=sample_rate, x_axis='time', y_axis='mel')
plt.title('Mel power spectrogram ')
plt.colorbar(format='%+02.0f dB')
plt.tight_layout()
mfcc = librosa.feature.mfcc(S=log_S, n_mfcc=13)

# Let's pad on the first and second deltas while we're at it
delta2_mfcc = librosa.feature.delta(mfcc, order=2)

plt.figure(figsize=(12, 4))
librosa.display.specshow(delta2_mfcc)
plt.ylabel('MFCC coeffs')
plt.xlabel('Time')
plt.title('MFCC')
plt.colorbar()
plt.tight_layout()
s=0
for i in range(len(a)):
    if a[i]==filename:
        s=i
        print(s)
ipd.Audio(samples, rate=sample_rate)
samples_cut = samples[4000:13000]
ipd.Audio(samples_cut, rate=sample_rate)
freqs, times, spectrogram_cut = log_specgram(samples_cut, sample_rate)

fig = plt.figure(figsize=(14, 8))

```

```

ax1 = fig.add_subplot(211)
ax1.set_title('Raw wave of ' + filename)
ax1.set_ylabel('Amplitude')
ax1.plot(samples_cut)

ax2 = fig.add_subplot(212)
ax2.set_title('Spectrogram of ' + filename)
ax2.set_ylabel('Frequencies * 0.1')
ax2.set_xlabel('Samples')
ax2.imshow(spectrogram_cut.T, aspect='auto', origin='lower',
           extent=[times.min(), times.max(), freqs.min(), freqs.max()])
ax2.set_yticks(freqs[::16])
ax2.set_xticks(times[::16])
ax2.text(0.06, 1000, 'Y', fontsize=18)
ax2.text(0.17, 1000, 'E', fontsize=18)
ax2.text(0.36, 1000, 'S', fontsize=18)

xcoords = [0.025, 0.11, 0.23, 0.49]
for xc in xcoords:
    ax1.axvline(x=xc*16000, c='r')
    ax2.axvline(x=xc, c='r')
def custom_fft(y, fs):
    T = 1.0 / fs
    N = y.shape[0]
    yf = fft(y)
    xf = np.linspace(0.0, 1.0/(2.0*T), N//2)
    vals = 2.0/N * np.abs(yf[0:N//2]) # FFT is simmetrical, so we take just the first half
    # FFT is also complex, to we take just the real part (abs)
    return xf, vals
xf, vals = custom_fft(samples, sample_rate)
plt.figure(figsize=(12, 4))
plt.title('FFT of recording sampled with ' + str(sample_rate) + ' Hz')
plt.plot(xf, vals)
plt.xlabel('Frequency')
plt.grid()
plt.show()
print("Emotion of the recording is ",d[0])

```