Vocular – The Speech Emotion Analyzer

MINOR PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

Tanishqa
Enrollment No: **44115603117**

Akshita Jain
Enrollment No: **44715603117**

Lakshay Singhal
Enrollment No: **44615603117**

Vikas
Enrollment No: **45515603117**

Guided by

Ms. Anjani Gupta
Assistant Professor of IT Department

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Dr. AKHILESH DAS GUPTA INSTITUTE OF TECHNOLOGY & MANAGEMENT
(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)
NEW DELHI – 110053

NOVEMBER 2020

CANDIDATE'S DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Minor Project Report entitled **"VOCULAR – SPEECH EMOTION ANALYSER"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** and submitted in the **Department of Information Technology** of **Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi)** is an authentic record of our own work carried out during a period from **October 2020 to November 2020** under the guidance of **Ms. Anjani Gupta, Assistant Professor of IT Department.**

The matter presented in the B. Tech Minor Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

Tanishqa
Enrollment No: 44115603117

Akshita Jain
Enrollment No: 44715603117

Lakshay Singhal
Enrollment No: 44615603117

Vikas
Enrollment No: 45515603117

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. They are permitted to appear in the External Minor Project Examination

(Ms. Anjani Gupta)
Assistant Professor of IT Department

Prof. (Dr.) Prashant Singh
Head, IT

The B. Tech Minor Project Viva-Voce Examination of **Tanishqa (Enrollment No: 44115603117),** has been held on 20th November 2020.

Prof. (Dr.) Preety Dhaka
Project Coordinator

(Signature of External Examiner)

ABSTRACT

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion. Various changes in the autonomic nervous system can indirectly alter a person's speech, and affective technologies can leverage this information to recognize emotion. For example, speech produced in a state of fear, anger, or joy becomes fast, loud, and precisely enunciated, with a higher and wider range in pitch, whereas emotions such as tiredness, boredom, or sadness tend to generate slow, low-pitched, and slurred speech. SER is tough because emotions are subjective and annotating audio is challenging.

This project is a web based application that would comprise of voice recording, recognition and analysis and thereby concluding the emotions that a person has using machine learning algorithms. In other words, the application does sentiment analysis on a voice that can be recorded in real time via the recorder provided in the app.

The type of sentiments that can be analysed and recognized from the voice of the user are happy, calm, sad, angry, fearful, neutral, disgust, surprised.

As a result of the emotion that is recognized, the application suggests some things in which a person can get involved in. The application can suggest some books, videos, quotes, events happening around from the web application's database and can redirect the user so that he/she can view the details of the same.

For example, the app can suggest some inspirational books, quotes for a person whose sentiment was detected sad and some videos, events that are happening around to further boost the mood of a person detected happy.

The application comes with two options that either a user can use voice based classification or can use a text/chat based to detect the sentiment. In the voice based classification, the sentiment is analysed on the basis of the pitch, intensity, loudness and other aspects of the sound. Whereas in text based classification, sentiment is analysed on the basis of the type of words used together to make a sentence and this recognition of sentiment using words and sentences is done using natural language processing.

Emotional speech processing technologies recognize the user's emotional state using computational analysis of speech features. Vocal parameters and prosodic features such as pitch variables and speech rate can be analyzed through pattern recognition techniques.

Speech analysis is an effective method of identifying affective state. These systems tend to outperform average human accuracy (approximately 60%) but are less accurate than systems which employ other modalities for emotion detection, such as physiological states or facial expressions. However, since many speech characteristics are independent of semantics or culture, this technique is considered to be a promising route for further research.

ACKNOWLEDGEMENT

We express our deep gratitude to **Ms. Anjani Gupta, Assistant Professor**, Department of Information Technology for his valuable guidance and suggestion throughout my project work. We are thankful to **Prof.** (**Dr.) Preety Dhaka,** Project Coordinator for their valuable guidance.

We would like to extend my sincere thanks to **Prof**. (**Dr.) Prashant Singh, Head of Department,** for his time to time suggestions to complete my project work. I am also thankful to **Prof. (Dr.) Sanjay Kumar, Director** for providing me the facilities to carry out my project work.

Tanishqa
Enrollment No: 44115603117

Akshita Jain
Enrollment No: 44715603117

Lakshay Singhal
Enrollment No: 44615603117

Vikas
Enrollment No: 45515603117

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATION

SER	Speech emotion recognition
AI	Artificial intelligence
CNN	Convolutional Neural Network
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
FFT	Fast Fourier Transform
MFCC	Mel Frequency Cepstral Coefficients
HNR	Harmonics to noise ratio
ANN	Artificial Neural Network

CHAPTER 1: INTRODUCTION

Speech is simply the most common method for communicating as people. It is just common at that point only natural then to extend out this correspondence medium to PC applications. We characterize speech emotion recognition (SER) as an assortment of systems that procedure and classify speech signals to detect the embedded emotions. In simple words, It is the act of attempting to recognize human emotion and affective states from speech. This is the system that will significantly take a shot at the way that voice frequently reflects hidden feelings through tone and pitch. SER is tough because emotions are subjective and annotating audio is challenging. By using this system, we can identify the human emotion like sad, cheerful, calm, angry, happy, fearful, regret, etc. by their speech or voice or we can say using some audio.

A web based application that would comprise of voice recording, recognition and analysis and thereby concluding the emotions that a person has using machine learning algorithms. In other words, the application does sentiment analysis on a voice that can be recorded in real time via the recorder provided in the app.

The type of sentiments that can be analysed and recognized from the voice of the user are happy, calm, sad, angry, fearful, neutral, disgust, surprised.

As a result of the emotion that is recognized, the application suggests some things in which a person can get involved in. The application can suggest some books, videos, quotes, events happening around from the web application's database and can redirect the user so that he/she can view the details of the same.

If you ever noticed, call centres employees never talk in the same manner, their way of pitching/talking to the customers changes with customers. Now, this does happen with common people too, but how is this relevant to call centres? Here is your answer, the employees recognize customers' emotions from speech, so they can improve their service and convert more people. In this way, they are using speech emotion recognition.

Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities. In some circumstances, humans could be replaced by computer-generated characters having the ability to conduct very natural and convincing conversations by appealing to human emotions. Machines need to understand emotions conveyed by speech. Only with this capability, an entirely meaningful dialogue based on mutual human-machine trust and understanding can be achieved.

CHAPTER 2 : CONCEPTS AND DEVELOPMENTAL COURSE

2.1 Concepts and Approaches

2.1.1 What is Machine Learning?

Machine learning is an application of artificial **intelligence** (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves.
Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

2.1.2 Machine Learning Methodologies

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed. Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labelled by humans, and unsupervised learning which provides the algorithm with no labelled data in order to allow it to find structure within its input data. Let’s explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labelled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data. For example, with supervised learning, an algorithm may be fed data with images of sharks labelled as fish and images of oceans labelled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabelled shark images as fish and unlabelled ocean images as water. A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

Unsupervised Learning

In unsupervised learning, data is unlabelled, so the learning algorithm is left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly valuable.
The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.
Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases. Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

CNN (Convolution Neural Networks)

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.
The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex.

2.2 Methodologies and Developmental course

2.2.1 Data gathering or Data Elicitation

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

As a result, we get a coherent set of data called as a **dataset**. In this project, the dataset into a flat file format that is a .data and .item.

ML depends heavily on data, without data, it is impossible for an "AI" to learn. It is the most crucial aspect that makes algorithm training possible… No matter how great your AI team is or the size of your data set, if your data set is not good enough, your entire AI project will fail. In every AI projects, classifying and labelling data sets takes most of our time, especially data sets accurate enough to reflect a realistic vision of the market/world.

Dataset

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):

Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS. Full dataset of speech and song, audio and video (24.8 GB) available from Zenodo.

Files

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

File naming convention

Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav).

Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").

- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

2.2.2 Setting Up a Python Programming Environment

Python is a flexible and versatile programming language suitable for many use cases, with strengths in scripting, automation, data analysis, machine learning, and back-end development. First published in 1991 the Python development team was inspired by the British comedy group Monty Python to make a programming language that was fun to use. Python 3 is the most current version of the language and is considered to be the future of Python.

Step 1 — Installing Python 3 Many operating systems come with Python 3 already installed. You can check to see whether you have Python 3 installed by opening up a terminal window and typing the following:

python3 –V

You'll receive output in the terminal window that will let you know the version number. While this number may vary, the output will be similar to this:

Output

Python 3.7.2

If you received alternate output, you can navigate in a web browser to python.org in order to download Python 3 and install it to your machine by following the instructions. Once you are able to type the python3 -V command above and receive output that states your computer's Python version number, you are ready to continue.

Step 2 — Installing pip To manage software packages for Python, let's install pip, a tool that will install and manage programming packages we may want to use in our development projects. If you have downloaded Python from python.org, you should have pip already installed. If you are on an Ubuntu or Debian server or computer, you can download pip by typing the following:

sudo apt install -y python3-pip

Now that you have pip installed, you can download Python packages with the following command:

pip3 install package_name

Here, package_name can refer to any Python package or library, such as Django for web development or NumPy for scientific computing. So if you would like to install NumPy, you can do so with the command pip3 install numpy. There are a few more packages and development tools to install to ensure that we have a robust set-up for our programming environment:

sudo apt install build-essential libssl-dev libffi-dev python3-dev

Once Python is set up, and pip and other tools are installed, we can set up a virtual environment for our development projects.

2.2.3 Data cleaning

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

2.2.4. Pre-processing

Pre-processing is the very first step after collecting data that will be used to train the classifier in a SER system. Some of these pre-processing techniques are used for feature extraction, while others are used to normalize the features so that variations of speakers and recordings would not affect the recognition process.

Framing

Signal framing, also known as speech segmentation, is the process of partitioning continuous speech signals into fixed length segments to overcome several challenges in SER.

Emotion can change in the course of speech since the signals are non-stationary. However, speech remains invariant for a sufficiently short period, such as 20 to 30 ms. By framing the speech signal, this quasi-stationary state can be approximated, and local features can be obtained. Consequently, fixed size frames are suitable for classifiers, such as Artificial Neural Networks, while retaining the emotion information in speech.

Windowing

After framing the speech signal, the next phase is generally applying a window function to frames. The windowing function is used to reduce the effects of leakages that occurs during Fast Fourier Transform (FFT) of data caused by discontinuities at the edge of the signals.

Normalization

Feature normalization is an important step which is used to reduce speaker and recording variability without losing the discriminative strength of the features. By using feature normalization, the generalization ability of features are increased.

Noise reduction

In real life, the noise present in the environment is captured along with the speech signal. This affects the recognition rate, hence some noise reduction techniques must be used to eliminate or reduce the noise.

Feature selection and dimension reduction

Feature selection and dimension reduction are important steps in emotion recognition. There is a need to use a feature selection algorithm because there are many features and there is no certain set of features to model the emotions. Otherwise, with so many features, the classifiers are faced with the curse of dimensionality, increased training time and over-fitting that highly affect the prediction rate.

Feature selection is the process of choosing a relevant and useful subset of the given set of features. The unneeded, redundant or irrelevant attributes are identified and removed to provide a more accurate predictive model.

2.2.5. Features

Features are an important aspect of speech emotion recognition. Carefully crafted set of features that successfully characterize each emotion increases the recognition rate. Various features have been used for SER systems; however, there is no generally accepted set of features for precise and distinctive classification. The existing studies have all been experimental so far.

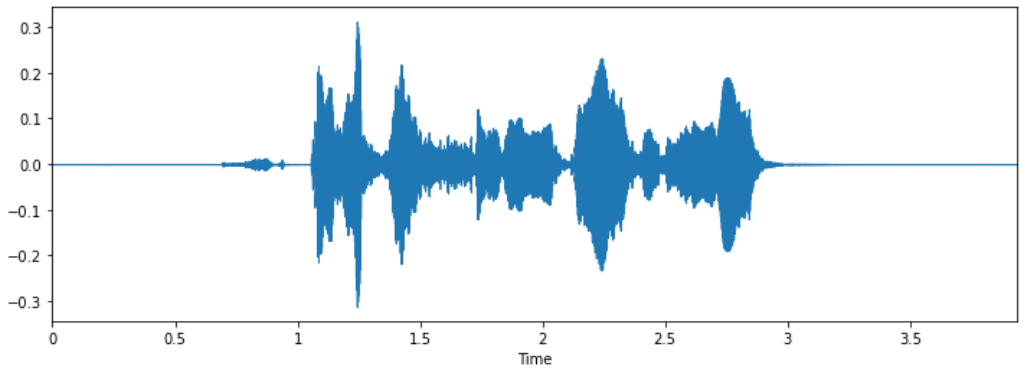

Figure 2.1: Features waveform of audio file

Speech is a continuous signal of varying length that carries both information and emotion. Therefore, global or local features can be extracted depending on the required approach. Global features, also called long-term or supra-segmental features, represent the gross statistics such as mean, minimum and maximum values, and standard deviation. Local features, also known as short-term or segmental features, represent the temporal dynamics, where the purpose is to approximate a stationary state.
These local and global features of SER systems are analyzed in the following four categories.

- Prosodic Features
- Spectral Features
- Voice Quality Features

Prosodic and spectral features are used more commonly in SER systems. Some of the features are listed under different categories by various studies depending on their approach. TEO features are specifically designed for recognizing stress and anger. These features are detailed individually; however, in practice, they are commonly combined to obtain better results.

Prosodic features

Prosodic features are those that can be perceived by humans, such as intonation and rhythm. A typical example is rising the intonation in a sentence that is meant as a question: “You are coming tonight?,” where in this case, the intonation rises on the word “tonight,” hinting that this is meant as a question. They are also known as para-linguistic features as they deal with the elements of speech that are properties of large units as in syllables, words, phrases, and sentences. Since they are extracted from these large units, they are long-term features. Prosodic

features have been discovered to convey the most distinctive properties of emotional content for speech emotion recognition

Spectral features

When sound is produced by a person, it is filtered by the shape of the vocal tract. The sound that comes out is determined by this shape. An accurately simulated shape may result in an accurate representation of the vocal tract and the sound produced. Characteristics of the vocal tract are well represented in the frequency domain. Spectral features are obtained by transforming the time domain signal into the frequency domain signal using the Fourier transform. They are extracted from speech segments of length 20 to 30 milliseconds that is partitioned by a windowing method.
Mel Frequency Cepstral Coefficients (MFCC) feature represents the short term power spectrum of the speech signal. To obtain MFCC, utterances are divided into segments, then each segment is converted into the frequency domain using short time discrete Fourier transform.

Voice quality features

Voice quality is determined by the physical properties of the vocal tract. Involuntary changes may produce a speech signal that might differentiate emotions using properties such as the jitter, shimmer, and harmonics to noise ratio (HNR). There is a strong correlation between voice quality and emotional content of the speech.

2.2.6 Building web framework using Flask web framework

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

Dependencies

These distributions will be installed automatically when installing Flask.

- Werkzeug implements WSGI, the standard Python interface between applications and servers.
- Jinja is a template language that renders the pages your application serves.
- MarkupSafe comes with Jinja. It escapes untrusted input when rendering templates to avoid injection attacks.

Install Flask

Python 3 comes bundled with the **venv** module to create virtual environments. Within the activated environment, use the following command to install Flask:

```
$ pip install Flask
```

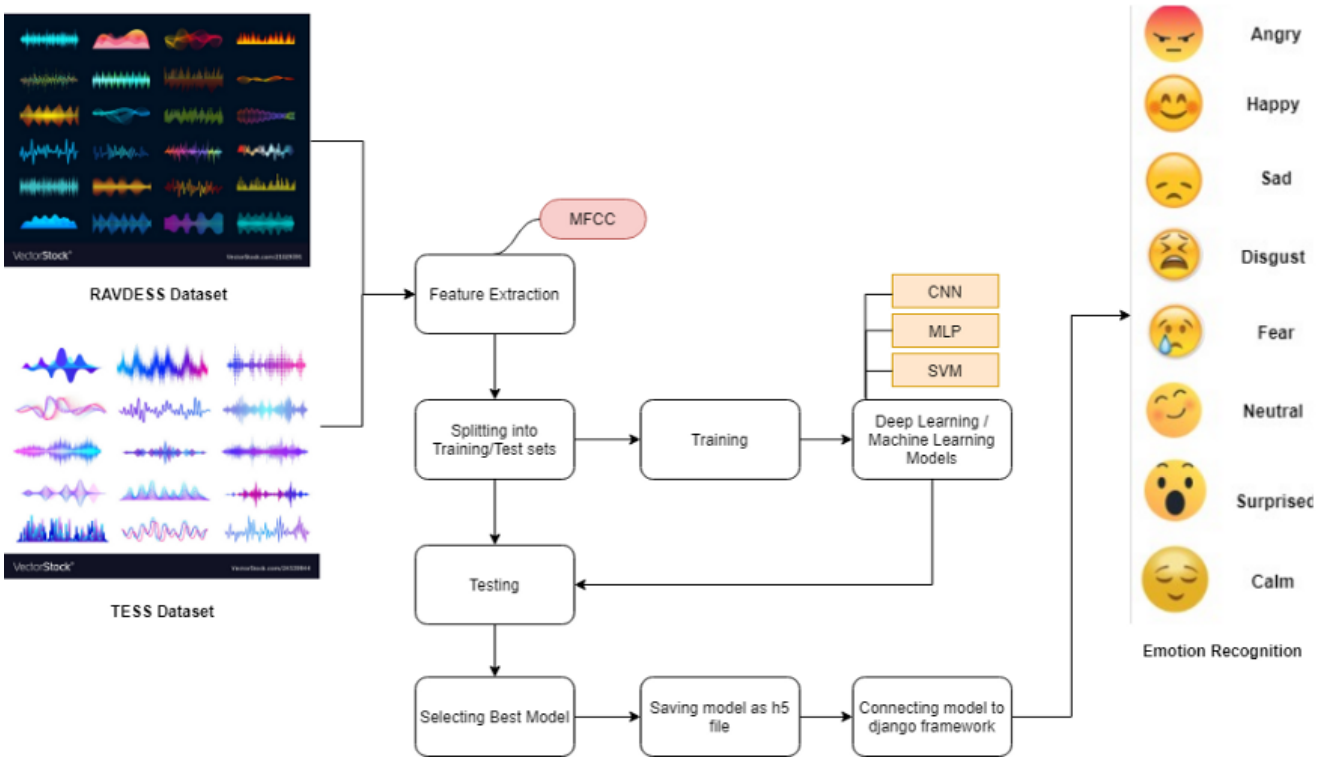

Figure 2.2: Developmental course of SER

CHAPTER 3 : TOOLS AND LIBRARIES

SUBLIME TEXT

Figure 3.1: Sublime Text logo

Sublime Text is a shareware cross-platform source code editor with a Python application programming interface. It natively supports many programming languages and markup languages, and functions can be added by users with plugins, typically community-built and maintained under free-software licenses.

FLASK

Figure 3.2 : Flask logo

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

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. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

NUMPY

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

PANDAS

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

LIBROSA

Figure 3.3 : Librosa logo

librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

CHAPTER 4 : RESULT ANALYSIS AND CONCLUSION

4.1 Working project's snapshots

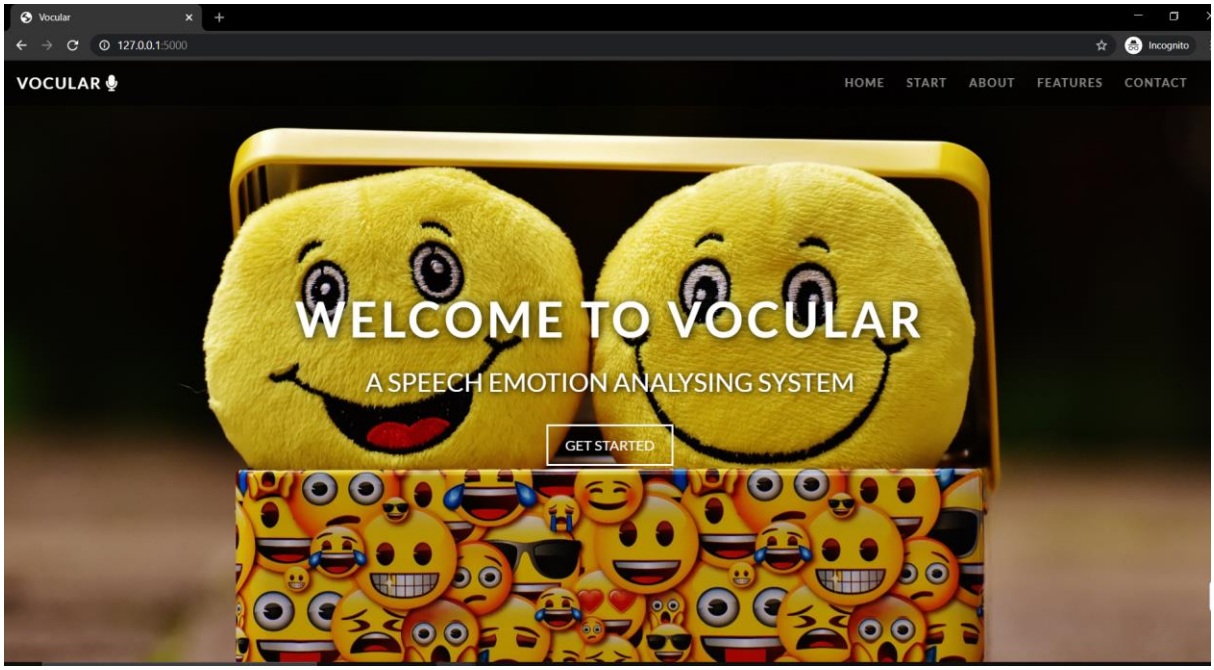

Figure 4.1 : Project image 1

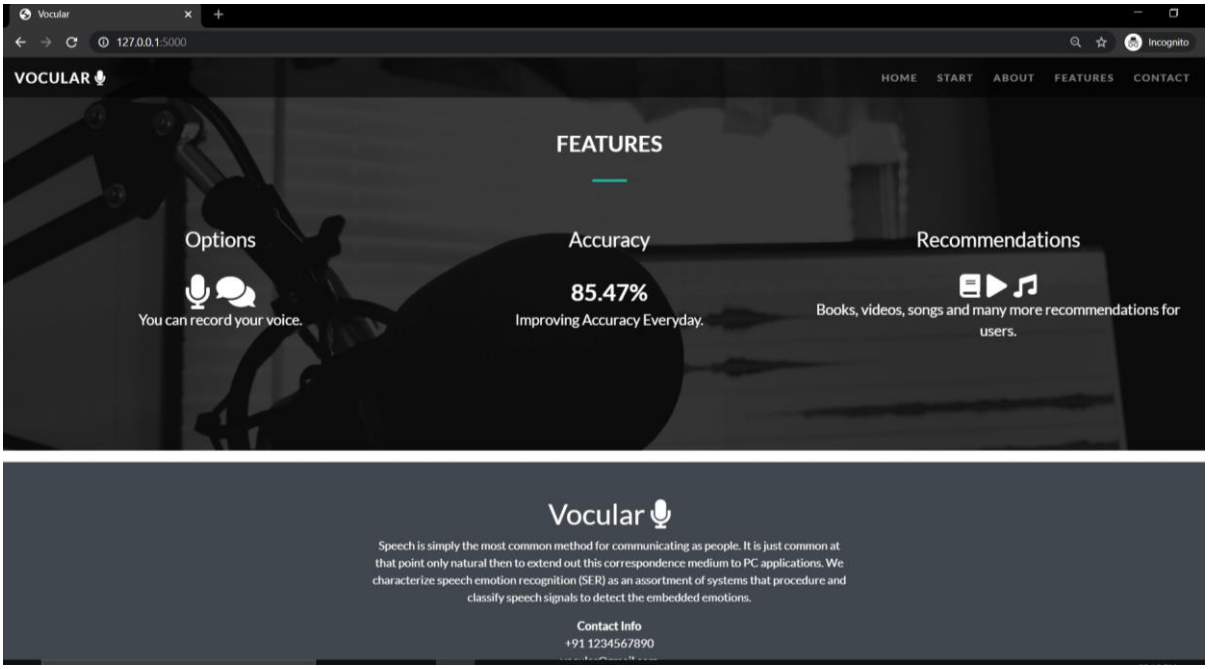

Figure 4.2 : Project image 2

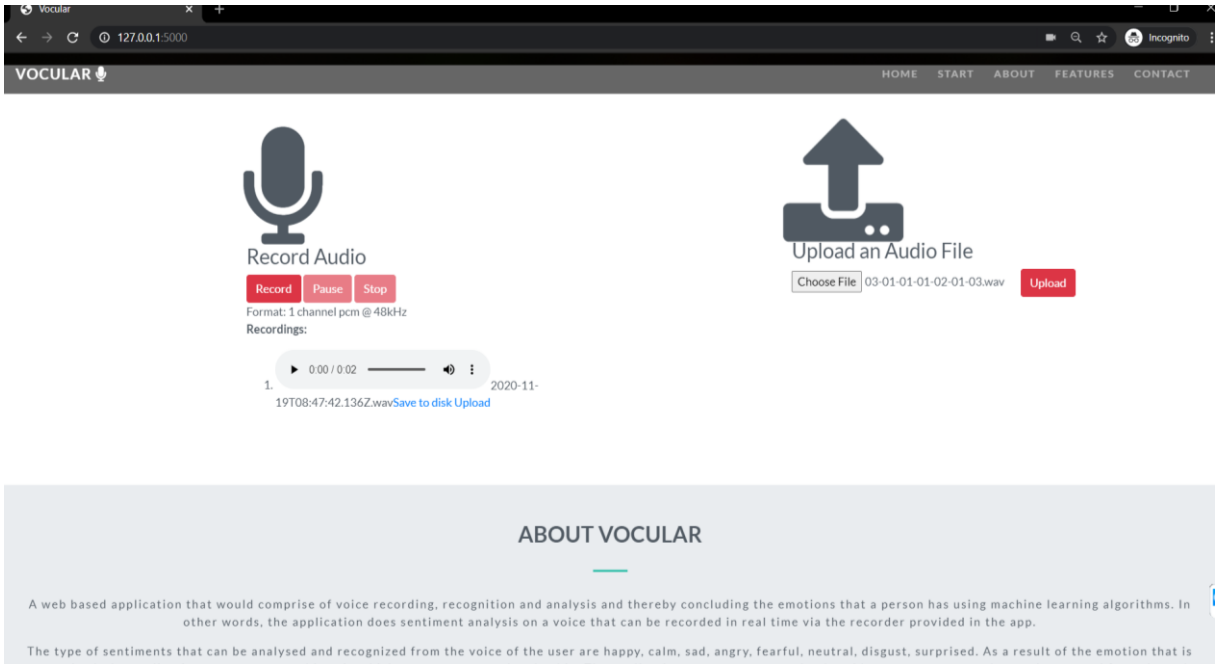

Figure 4.3 : Project image (Recording/ uploading of voice)

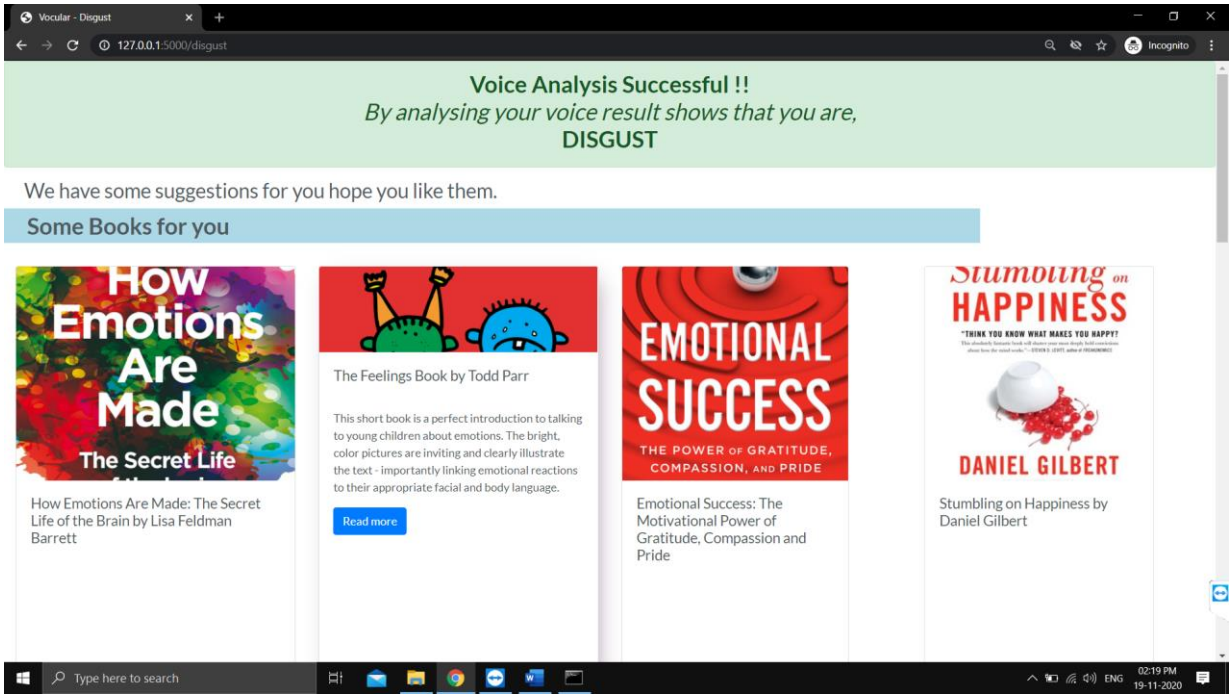

Figure 4.4 : Project image (Prediction of the emotion and suggestions on the basis of the emotion)

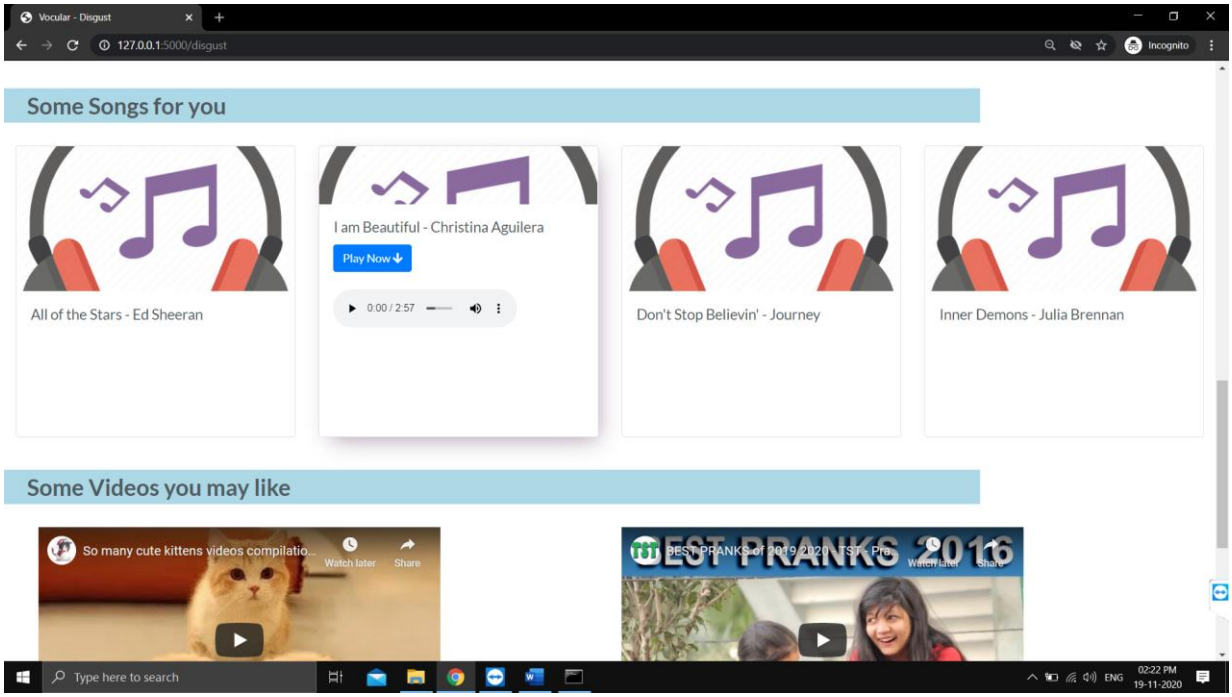

Figure 4.5 : Project image (Suggestions on the basis of the emotion)

4.2 Conclusion

We have identified and detailed the parts that make up a speech emotion recognition system. These systems require training data provided by speech databases that are created using either acted, elicited, or natural sources. The signals are then pre-processed to make them fit for feature extraction. SER systems most commonly use prosodic and spectral features since they support a wider range of emotion and yield better results. The results can further be improved by adding features from other modalities, such as the ones that depend on visual or linguistic features.

In this project, we have used CNN (Convolution Neural Networks) with training data = 67% and test data = 33%. The accuracy from our model is 85.47%.

Once all the features are extracted, SER systems have a wide range of classification algorithms to choose from. While most use classical approaches, there are an increasing number of studies that incorporate recent advances, such as Convolutional or Recurrent Neural Networks.

All of these pre-processing and feature extraction are done to detect the emotion in the speech signal, yet emotions are still an open problem in psychology. There are several models that define them. SER systems use manual labelling for their training data, which, as mentioned earlier, is not always exactly correct.

Although there are systems and realizations of real-time emotion recognition, SER systems are not yet part of our every day life, unlike speech recognition systems that are now easily accessible even with mobile devices. To reach this goal, SER systems need more powerful

hardware so that processing can be done faster; more correctly labelled data so that the training is more accurate; and more powerful algorithms so that the recognition rates increase. We believe that the research will continue towards solutions that apply deep learning algorithms, and since they require more data and more powerful processors, and these advances are likely to follow.

We believe that, as SER systems become more part of our daily lives, there will be more data available to learn from, which will improve their performance, even when at times humans can fail. The subtle differences which may not be registered by humans can be picked up by these networks that will improve the areas where emotion recognition is applicable, such as human computer interaction, healthcare, and alike.

4.3 Future Scope

Through this project, 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
- Voice-based virtual assistants or chatbots
- Linguistic research

A few possible steps that can be implemented to make the models more robust and accurate are the following:

- An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
- Figuring out a way to clear random silence from the audio clip.
- Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
- Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

REFERENCES

[1] Margaret Lech1, “Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network” Front. Comput. Sci., 26 May 2020

[2] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Available : https://zenodo.org/record/1188976#.X6mkqmgzZPY

[3] Kun Han, Dong Yu, Ivan Tashev,“Speech Emotion Recognition Using Deep Neural Network and Extreme Machine learning” The Ohio State University, Columbus, 43210, OH, USA Microsoft Research, One Microsoft Way, Redmond, 98052, WA, USA