VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi , Karnataka, INDIA

A Project Report
on

Speech Emotion Recognition

Submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering
in
Computer Science and Engineering

Submitted By

CHANDANA R	**1GA17CS041**
NISHMITHA P	**1GA17CS099**
SAMSKRUTHI K JOSHI	**1GA17CS133**

Under the Guidance of
Ms. VANISHREE M L
Assistant Professor

Department of Computer Science and Engineering
Accredited by NBA(2019-2022)
GLOBAL ACADEMY OF TECHNOLOGY
Rajarajeshwarinagar, Bengaluru - 560 098

2020 – 2021

GLOBAL ACADEMY OF TECHNOLOGY

Department of Computer Science and Engineering

Accredited by NBA(2019-2022)

CERTIFICATE

Certified that the Project Entitled **"Speech Emotion Recognition"** carried out by **CHANDANA R**, bearing **USN 1GA17CS041, NISHMITHA P**,bearing **USN 1GA17CS099, SAMSKRUTHI K JOSHI**, bearing **USN 1GA17CS133,** bonafide students of Global Academy of Technology, is in partial fulfillment for the award of the **BACHELOR OF ENGINEERING** in **Computer Science and Engineering** from Visvesvaraya Technological University, Belagavi during the year 2020-2021. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the report submitted to the department. The Partial Project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Ms. Vanishree M L Assistant Professor Dept. of CSE GAT, Bengaluru.	Dr. Srikanta Murthy K Professor & HOD Dept. of CSE GAT, Bengaluru.	Dr. Rana Pratap Reddy Principal GAT, Bengaluru.

GLOBAL ACADEMY OF TECHNOLOGY

Rajarajeshwarinagar, Bengaluru – 560 098

DECLARATION

We, **CHANDANA R**, bearing **USN 1GA17CS041, NISHMITHA P**,bearing **USN 1GA17CS099, SAMSKRUTHI K JOSHI**, bearing **USN 1GA17CS133**, students of Seventh Semester B.E, Department of Computer Science and Engineering, Global Academy of Technology, Rajarajeshwarinagar Bengaluru, declare that the Project Work entitled "**Speech Emotion Recognition"** has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree in **Bachelor of Engineering** in **Computer Science and Engineering** from **Visvesvaraya Technological University, Belagavi** during the academic year **2020-2021**.

1. CHANDANA R	1GA17CS041
2. NISHMITHA P	1GA17CS099
3. SAMSKRUTHI K JOSHI	1GA17CS133

Place: Bengaluru

Date:

ABSTRACT

As human beings speech is amongst the most natural ways to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication.

There is a wide diversity and non-agreement about the basic emotion or emotion-related states on one hand and about where the emotion-related information lies in the speech signal on the other side.motion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them.

ACKNOWLEDGEMENT

The satisfaction and the euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. The constant guidance of these persons and encouragement provide, crowned our efforts with success and glory. Although it is not possible to thank all the members who helped for the completion of the phase - 1 of the project individually, we take this opportunity to express our gratitude to one and all.

We are grateful to management and our institute **GLOBAL ACADEMY OF TECHNOLOGY** with its very ideals and inspiration for having provided us with the facilities, which made this, phase - 1 project a success.

We express our sincere gratitude to **Dr. N. Rana Pratap Reddy,** Principal, Global Academy of Technology for the support and encouragement.

We wish to place on record our grateful thanks to **Dr. Srikanta Murthy K**, HOD, Department of CSE , Global Academy of Technology, for the constant encouragement provided to us.

We are indebted with a deep sense of gratitude for the constant inspiration, encouragement, timely guidance and valid suggestions given to us by our guide **Ms. Vanishree M L, Associate Professor**, Department of CSE, Global Academy of Technology.

We are thankful to all the staff members of the department for providing relevant information and helped in different capacities in carrying out this phase -1 project.

Last, but not least, we owe our debts to our parents, friends and also those who directly or indirectly have helped us to make the phase - 1 project work a success.

CHANDANA R	**1GA17CS041**
NISHMITHA P	**1GA17CS099**
SAMSKRUTHI K JOSHI	**1GA17CS133**

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

SRS	Software Requirement Specification
DFD	Data Flow Diagram
MFCC	Mel Frequency Cepstrum Coefficient
MLP	Multi Layer Perceptron
ML	Machine Learning
RAVDESS	Ryerson Audio Video database of Emotional Speech and Song
DL	Deep Learning
SVM	Support Vector Machine
DBN	Deep Belief Networks
VUI	Voice User Interface
LSA	Latent Semantic Analysis
SER	Speech Emotion Recognition
LDA	Latent Dirichlet Allocation

CHAPTER 1

INTRODUCTION

In machine learning, computers apply statistical learning techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions. Machine learning brings together computer science and statistics to harness predictive power. There are two types of machine learning approaches:

- Unsupervised
- Supervised

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables. In supervised learning, we are given a dataset and already know what our correct output should look like, having the idea that there is a relationship between the input and the output. We are using React to create a user-friendly GUI. React is an open source web development framework created by Facebook. MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schema.

1.1 Definitions

- Machine Learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference.
- As stated by Tom M. Mitchell, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".
- Deep Learning (DL) is an AI function that mimics the workings of the human brain in processing data for use in detecting objects, recognizing speech, translating languages and making decisions.

1.2 Project Outline

- To implement a real-time classification algorithm for inferring emotions from the verbal features of speech.
- To successfully extract a set of features and use them to train and detect emotions from speech.
- To compare the performance of various kinds of models and choose the best algorithm for the dataset.
- To create a user-friendly web page where users input the audio file and test it against the model.

CHAPTER 2

REVIEW OF LITERATURE

2.1 System Study

1. Latent Semantic Analysis (LSA), 2008
    - This paper presents a system that used several variations of Latent Semantic Analysis and evaluated several knowledge-based and corpus-based methods for the automatic identification of six emotions in text when no affective words exist.
    - However their approach achieved a low accuracy because it is not context sensitive and lacks the semantic analysis of the sentence.
2. Latent Dirichlet Allocation (LDA)methodology, 2011
    - This paper proposed a framework that depends heavily on the pre-processing of the input data (Czech Newspaper Headlines) and labeling it using a classifier.
    - They achieved an average accuracy of 80% for 1000 Czech news headlines using SVM with 10-fold cross validation. However their method was not tested on English dataset. Also it is not context sensitive as it only considers emotional keywords as features.
3. Speech Emotion Recognition Based on Deep Belief Network and SVM, 2014
    - Deep Belief Networks(DBN's) is used to extract emotional characteristic parameter from emotional speech signal automatically.
    - Combined deep belief network and support vector machine (SVM) is the proposed classifier model.
4. New Fuzzy Cognitive Map Learning Algorithm for Speech Emotion Recognition,2017
    - Different acoustic features are extracted and fused to overcome the deficiency of one another in classifying certain emotions.
    - GA and swarm intelligence optimization algorithms are used.Most of these methods require domain experts who can specify in advance the initial weight matrix of an FCM.

2.2 Proposed Work

- In this project, we plan to use the libraries librosa, soundfile, and sklearn (among others) to build a model using an MLPClassifier (Multi-Layer Perceptron Classifier). Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function by training on a dataset. Given a set of features and a target , it can learn a non-linear function approximator for either classification or regression. This will be able to recognize emotion from sound files.
- The Dataset we plan to use is the RAVDESS (Ryerson Audio Video database of Emotional Speech and Song) dataset to recognise eight emotions namely,
    1. Neutral
    2. Happy
    3. Calm
    4. Sad
    5. Angry
    6. Fear
    7. Disgust
    8. Surprised
- We will load the data, extract features from it, then split the dataset into training and testing sets. Then, we'll initialize an MLPClassifier and train the model. Finally, we'll calculate the accuracy of our model.

2.3 Scope of the Project

The importance of Speech emotion recognition is getting popular with improving user experience and the engagement of Voice User Interfaces (VUIs). Developing emotion recognition system that is based on speech has practical application benefits. However, these benefits are somewhat negated by the real-world background noise impairing speech-based emotion recognition performance when the system is employed in practical applications

CHAPTER 3

SYSTEM REQUIREMENT SPECIFICATION

3.1 Functional Requirements

The functional requirements for a system describe what the system should do. These requirements depend on the type of software being developed; the general approach taken by the organization when writing requirements. The functional system requirements describe the system function in detail, its inputs and output, exceptions and so on.

Functional requirements are as follows:

- The RAVDESS Dataset will be fed to the Jupyter Notebook environment.
- Pre-processing of data and analysis is done.
- Various machine learning algorithms are implemented on the dataset.
- The algorithm with the highest accuracy is chosen.
- A user-friendly web application is developed using React.

3.2 Non-Functional Requirements

Non-Functional requirements, as the name suggests, are requirements that are not directly concerned with the specific functions delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy. Alternatively, they may define constraints on the system such as capabilities of I/O devices and the data representation used in system interfaces.

The non-functional requirements are as follows:

- Usability
- Performance
- Portability
- Reliability
- Supportability

3.3 Hardware Requirements

- Processor – Intel core i5 or AMD Ryzen 5 and above.
- Memory – 4 GB Ram and above
- 100 GB Hard Disk Drive
- 64-bit Operating System
- Mouse or any other pointing device
- Keyboard
- Display Device

3.4 Software Requirements

- Operating system : Windows 7/8/10
- IDE : Jupyter Notebook version 6 or above
- Programming Language : Python
- Libraries and tools:

  1. Librosa: Librosa is a Python library for analysing audio and music.

  2. NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices.

  3. Soundfile: SoundFile is an audio library based on libsndfile, CFFI and NumPy.

  4. Scikit-Learn: Scikit-learn is a free software machine learning library for the Python programming language.

  5. PyAudio: PyAudio provides bindings for PortAudio, the cross-platform audio I/O library.

CHAPTER 4

SYSTEM DESIGN

System design is the process of defining the architecture, components, modules, interface and data for a system to satisfy specified requirements. System design could see it has the application of systems theory to product development. There is some overlap with disciplines of system analysis, system architecture, system engineering.

4.1 Design Overview

The primary objective of SER is to improve man-machine interface. It can also be used to monitor the psycho physiological state of a person in lie detectors. The goal of SER is to build a model to recognize emotion from speech using the librosa and sklearn libraries and the RAVDESS dataset. It aims to mimic the human perception mechanisms.

4.2 System Architecture

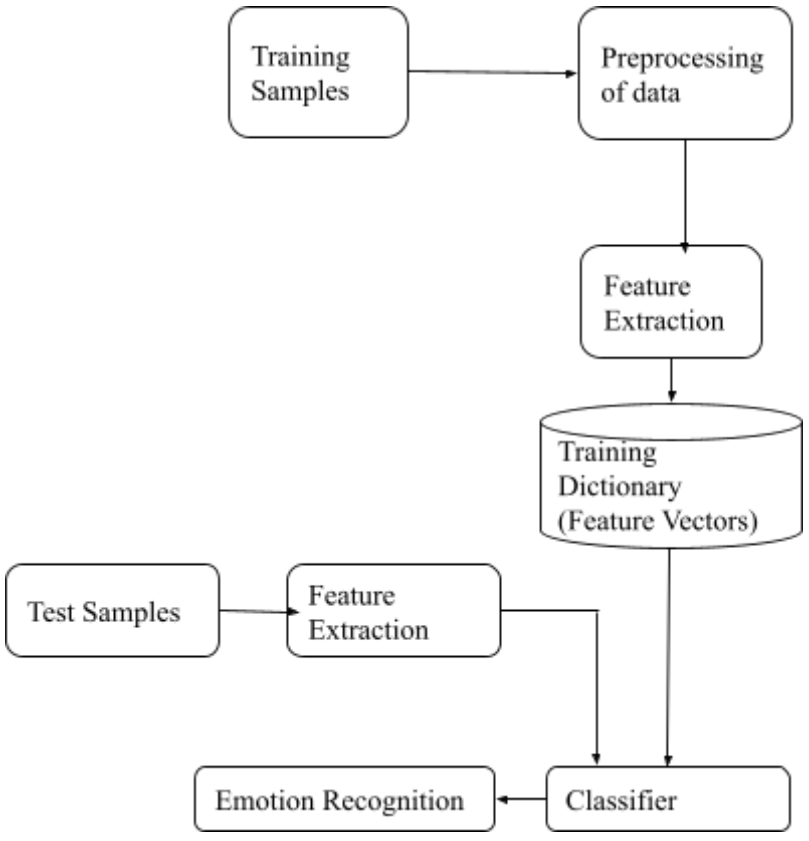

Fig 4.2 System Architecture

4.3 Data Flow Diagram

The Data Flow Diagram provides a visual representation of the flow information within a system. By drawing a Data Flow Diagram, you tell the information provided to someone who takes part in the system processes, the information needed in order to complete the process and the information needed to be stored and accessed.

4.3.1 Data Flow Diagram – Level 0

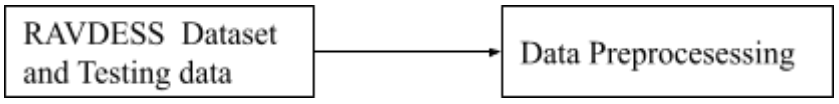

Fig. 4.3.1 DFD Level 0

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

4.3.2 Data Flow Diagram – Level 1

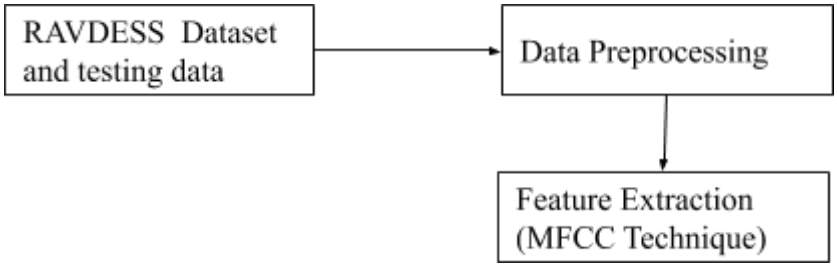

Fig. 4.3.2 DFD Level 1

After Preprocessing, the dataset is divided into a training set and test set. MFCC Technique is applied on the dataset to extract the three features, MFCC (Mel-frequency cepstral coefficients), chroma and Mel spectrogram as speech features.

4.3.3 Data Flow Diagram – Level 2

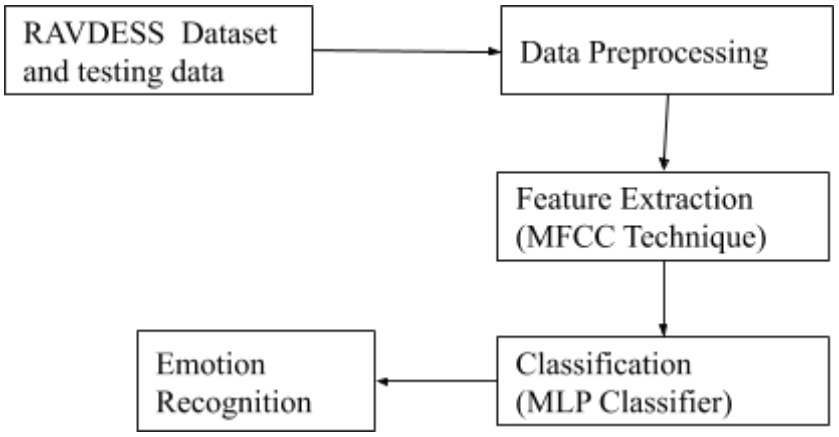

Fig. 4.3.3 DFD Level 2

The extracted features are fed to MLP Classifier (Multilayer Perceptron Classifier). It implements the Multilayer Perceptron algorithm that trains using back propagation. It optimizes the log-loss function using stochastic gradient descent, the MLPClassifier has an internal neural network for the purpose of classification.

4.4 USE CASE DIAGRAM

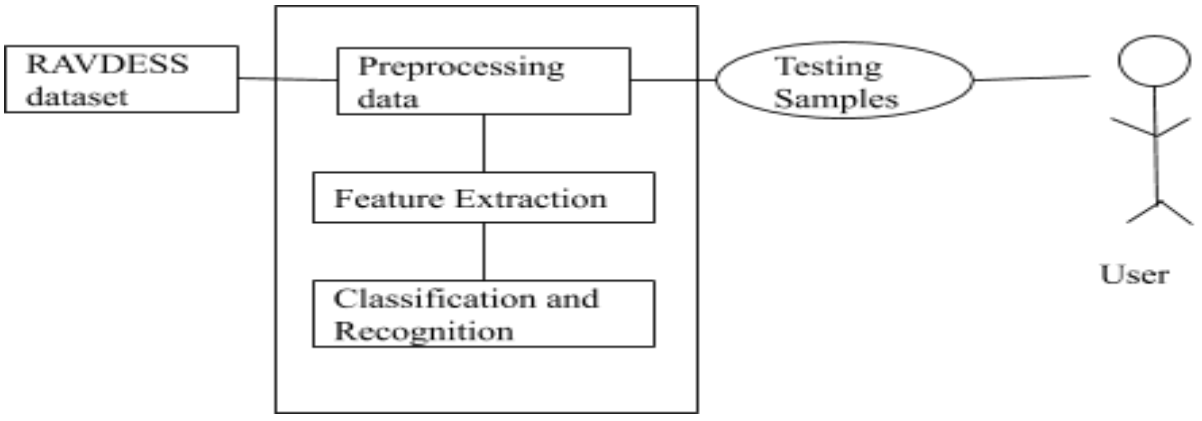

Fig 4.4 Use Case Diagram

- The dataset is fed into the execution environment.
- The next process is to preprocess the data and extract the features using MFCC Technique.

- The next step is to implement a classification algorithm to classify the feature and recognise the emotion.
- The output for the user input is displayed to the user.

4.5 SEQUENCE DIAGRAM

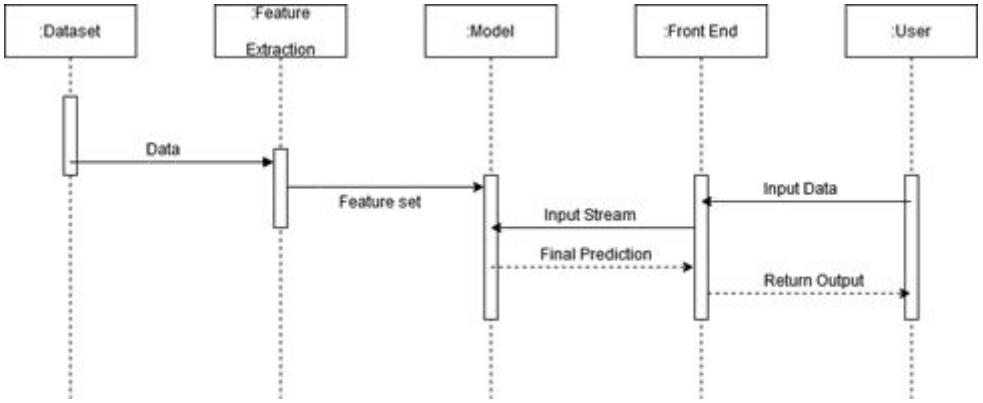

Fig 4.5 Sequence Diagram

4.6 COLLABORATION DIAGRAM

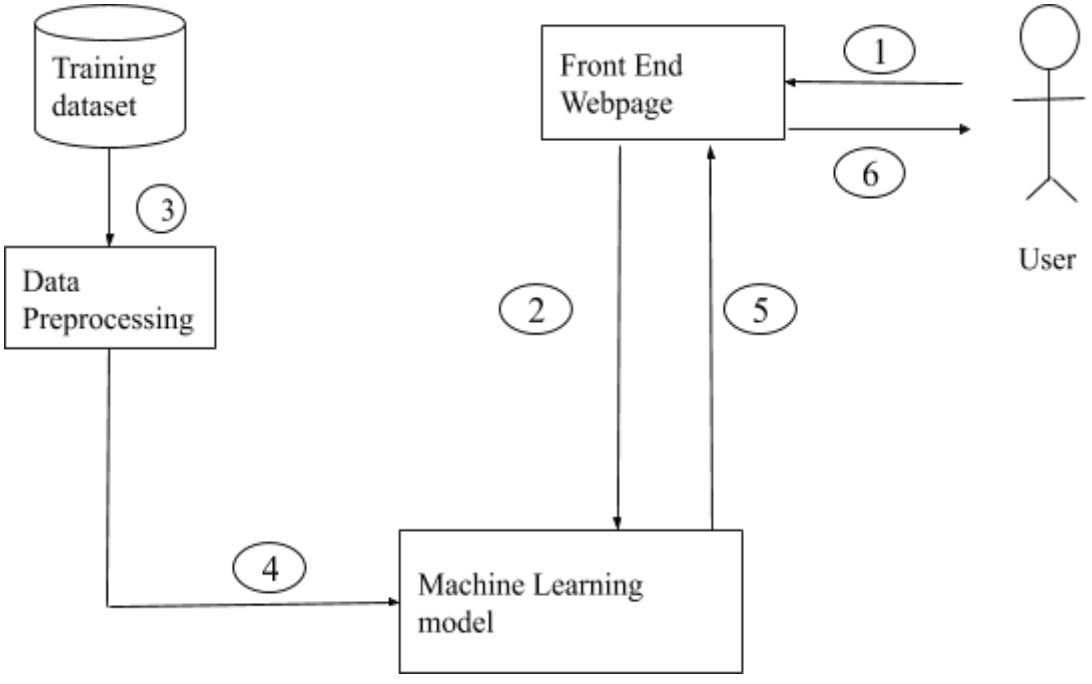

Fig 4.6 Collaboration Diagram

4.7 ACTIVITY DIAGRAM

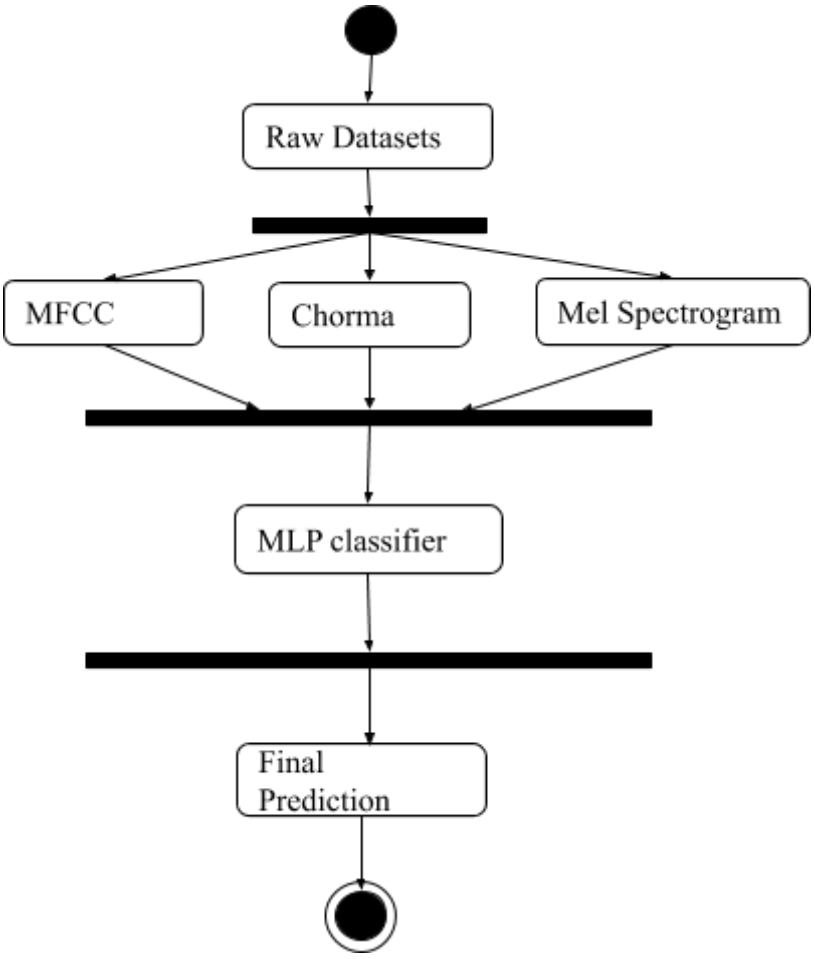

Fig 4.7 Activity Diagram

4.8 MODULES

Module 1: Selection and loading the data

Module 2: Preprocessing and splitting the dataset

Module 3: Feature Extraction

Module 4: Classification and Testing the model against user input

4.8.1 Module 1

Module Name: Selection and Loading the data

Functionality: This function will load the data from the interface and convert it to MFCC format and store it in a dataframe.

Input: RAVDESS Dataset

4.8.2 Module 2

Module Name: Preprocessing and splitting the dataset

Functionality: Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which increases the accuracy and efficiency of the machine learning model. Later, the dataset is divided into a training set and test set. Training set is used to train a model and define its optimal parameters. RAVDESS dataset is the training set we intend to use. A test set is needed for an evaluation of the trained model and its capability for generalization.

Input: The converted MFCC formats of the data are included here.

Used: 'train_test_split' package form sklearn model is imported.

Output: Training set and test set.

4.8.3 Module 3

Module Name: Feature Extraction

Functionality: This function uses MFCC(Mel-frequency cepstral coefficients) technique to extract features from the split dataset. MFCC coefficients, chroma and Mel spectrogram are the three features extracted here. The features extracted for each data are stored in a dictionary.

Input: The preprocessed dataset.

Used: MFCC algorithm.

Output: Feature Vectors(dictionary).

4.8.4 Module 4

Module Name: Classification and Testing the model against user input.

Functionality: Here the MLP Classifier optimizes the log-loss function using stochastic gradient descent. Unlike SVM or Naive Bayes, the MLPClassifier has an internal neural network for the purpose of classification. And hence, predicts the label(emotions) for each

data.

Used: MLP Classifier implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation.

Input: The features vectors extracted.

Output: Classified emotion on the basis of the extracted features.

CHAPTER 5

CONCLUSION

In this work, we tried multiple approaches for recognising emotion through speech. And 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 is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. We classify eight emotions from speech, happy, sad, neutral, anger, fear, disgust, surprise and calm. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, linguistic research, medical uses like therapy, counselling etc.

This simple system with the classifiers is easy to understand and implement because the utilization from a small group of features would work remarkably well on real-world data, making it possible to develop a real-time system where fast decisions in accordance with the emotional feedback provided from humans are taken. As a real application, it could be considered a real-time system that can serve like a motor of emotional knowledge in order to understand the autistic children, to accurately describe their internal state and show the real content of their emotions. This type of emotional devices working with emotional feedback will have the potential to reveal more about emotional state and the early detection of crisis, balanced lifestyle including and regulated stress level.

BIBLIOGRAPHY

[1] Marinrez DP, Alfonseca E, Rodriguez P, Gliozzo A, Strapparava C, Magnini B (2005) About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. Revista Signos 38: 325–43.

[2] Hernandez, S., Garden, K.L., Sallis, P.J.: A signal denoising method for text meaning vectors. In: Proceedings of the Fifth Asia Modelling Symposium .

[3] Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng, "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM", *Mathematical Problems in Engineering*, vol. 2014, Article ID 749604, 7 pages, 2014.

[4] Wei Zhang, Xueying Zhang, Ying Sun, "A New Fuzzy Cognitive Map Learning Algorithm for Speech Emotion Recognition", *Mathematical Problems in Engineering*, vol. 2017, Article ID 4127401, 12 pages, 2017.

[5] "RAVDESS dataset." [Online]. Available: https://zenodo.org/record/1188976.