SPEECH EMOTION RECOGNITION

A project report submitted in partial fulfillment of

the requirements for the award of the Degree of

Bachelor of Technology
in
Computer Science and Engineering

BY

Adithi Karre (18011A0501)

Pindiga Sushma Niveni (18011A0539)

Aishwarya Madishetty (18011A0564)

Under the guidance of

Dr. B. Padmaja Rani
Professor

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

DECLARATION BY THE CANDIDATES

We, **Adithi Karre (18011A0501), Pindiga Sushma Niveni (18011A0539) and Aishwarya Madishetty (18011A0564),** hereby certify that the project report entitled "**Speech Emotion Recognition**" carried out under the guidance of **Dr. B. Padmaja Rani**, is submitted. This is a record of bonafide work carried out by us and the results embodied in this project have not been reproduced/ copied from any source and have not been submitted to any other University or Institute for the award of any other degree or diploma.

Adithi Karre (18011A0501)

Pindiga Sushma Niveni (18011A0539)

Aishwarya Madishetty (18011A0564)

Department of Computer Scienceand Engineering,

JNTUH University College of Engineering,Hyderabad-85

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

CERTIFICATE BY THE SUPERVISOR

This is to certify that the project report entitled **"Speech Emotion Recognition"**, being submitted by **Adithi Karre (18011A0501), Pindiga Sushma Niveni (18011A0539) and Aishwarya Madishetty (18011A0564),** in partial fulfilment is a record of bonafide work carried out by them. The results are verified and found satisfactory.

Dr. B. Padmaja Rani

Professor

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

Date:

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

CERTIFICATE BY THE HEAD

This is to certify that the project report entitled "**Speech Emotion Recognition**", being submitted by **Adithi Karre (18011A0501), Pindiga Sushma Niveni (18011A0539) and Aishwarya Madishetty (18011A0564),** in partial fulfilment is a record of bonafide work carried out by them. The results are verified and found satisfactory.

Dr. D. Vasumathi,

Professor & Head of the Department,

Department of Computer Science and Engineering,

JNTUH University College of Engineering, Hyderabad-85

Date:

ACKNOWLEDGEMENT

I would like to express sincere thanks to our Supervisor **Dr. B. Padmaja Rani**, Professor of Computer Science and Engineering Department, JNTUH-CEH for her admirable guidance and inspiration both theoretically and practically and most importantly for the drive to complete the project successfully. Working under such an eminent guide was our privilege.

I owe a debt of gratitude to **Dr. D. Vasumathi**, Professor & Head of the department of Computer Science & Engineering, for her kind considerations and encouragement in carrying out this project successfully.

I am grateful to the Project Review Committee members and Department of Computer Science & Engineering who have helped in successfully completing this project by giving their valuable suggestions and support.

I express thanks to our parents for their love, care and moral support without which we would have not been able to complete this project. It has been a constant source of inspiration for all our academic endeavors. Last but not the least, we thank the Almighty for making us a part of the world.

Adithi Karre (18011A0501)

Pindiga Sushma Niveni (18011A0539)

Aishwarya Madishetty (18011A0564)

ABSTRACT

The idea behind this project is to build a machine learning model that could detect emotions from how we converse with each other.
Emotion detection has become one of the biggest marketing strategies, the mood of the consumer plays an important role. So, to detect the current emotion of the person and suggest to him the apt product or help him, accordingly, will increase the demand of the product or the company.

Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. The emotion recognition and the classifiers are used to differentiate emotions such as happiness, surprise, anger, neutral state, sadness, etc. The dataset for the speech emotion recognition system is the speech samples and the characteristics are extracted from these speech samples using the LIBROSA package. The classification performance is based on extracted characteristics. Finally, we can determine the emotion of the speech signal to build a model using an MLPClassifier. This will be able to recognize emotion from sound files. 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 be able to calculate the accuracy of our model.

Speech Emotion Recognition is a current research topic because of its wide range of applications, and it has become a challenge in the field of speech processing too. In this project, we will be carrying out a brief study on Speech Emotion Analysis along with Emotion Recognition. This project includes the study of different types of emotions, features to identify those emotions and various classifiers to classify them properly.

CONTENTS

1. INTRODUCTION

1.1 SPEECH EMOTION RECOGNITION

In this project we built a model to recognize emotion from speech using the librosa and Multi Layer Perception Classifier (MLPClassifier) and RAVDESS Dataset. This will be able to recognize emotion from sound files. We will load the data, extract features from it, then split the dataset into training and testing sets. Then, we initialized an MLPClassifier and trained the model. Finally, we calculated the accuracy of our model. Speech Emotion Recognition is tough because emotions are subjective and annotating audio is challenging.

1.2 PURPOSE

As human beings speech is amongst the most natural way 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. Emotion detection is a challenging task because emotions are subjective. There is no common consensus on how to measure or categorize them. We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings.

1.3 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, in voice-based virtual assistants or chatbots, in linguistic research, etc.

- 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.

  Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.
- 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.

1.4 Problem Definition

Emotion detection has become one of the biggest marketing strategies, in which mood of the consumer plays an important role. So, to detect the current emotion of the person and suggest him the apt product and help him/her, accordingly, will increase the demand of the product or the company. Humans have the natural ability to use all their available senses for maximum awareness of the received message. The emotional detection is natural for humans, but it is a very difficult task for machines. Detecting emotions is one of the most important marketing strategies in today's world. For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centres to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful.

2. LITERATURE SURVEY

2.1 SYSTEM REVIEW

This survey is done to comprehend the need and prerequisite of the general population, and as such, we went through different sites and applications and looked for the fundamental data. Based on these data, we made an audit that helped us get new thoughts and make different arrangements for our task. We reached the decision that there is a need of such application and felt that there is a decent extent of progress in this field too.

2.2 EXISTING CLASSIFICATION MODELS

- **Multivariate Linear Regression Classification**

Multivariate Linear Regression (MLR) is a simple and efficient computation of machine learning algorithms, and it can be used for both regression and classification problems. This is quite similar to the simple linear regression model but with multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine and complex computation due to the added variables.

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_n x_i^{(n)}$$

- **Support Vector Machine**

Support Vector Machines (SVM) is an optimal margin classifier in machine learning. It is also used extensively in many studies related to audio emotion recognition. It can have a very good classification performance compared to other classifiers especially for limited training data. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

- **Recurrent Neural Networks**

Recurrent Neural Networks (RNN) are suitable for learning time series data. While RNN models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, LSTM (Long Short Term Memory) RNNs were proposed by Hochreiter et al. It uses memory cells to store information so that it can exploit long range dependencies in the data.

2.3 DRAWBACKS AND FUTURE SCOPE

- You can try other different classifiers as well to predict the emotion behind the audio like CNN, SVM, etc.
- Predicting the Live Audio takes a lot of process and it is sometimes difficult to process as it is unlike the binary data with some csv file associated with it.
- In future, we can predict the random recorded audio as well.
- We can also embed our UI with ML Model, and we can build web based application with ML using Flask in future, this point can be a great scope.
- More advancements can be more variety of voices can be trained and dataset can be increased to deploy a more realistic model.

3. REQUIREMENT SPECIFICATION

3.1 FUNCTIONAL REQUIREMENTS

1. **Python**

   Python is the basis of the program that we wrote. It utilizes many of the python libraries.

2. **Libraries**

   - Pandas: for data manipulation and analysis
   - NumPy: for mathematical and logical operations
   - Scikit-learn: for data preprocessing, model selection, model evaluation and other utilities
   - Librosa: for analyzing audio and music
   - Soundfile: for reading and writing sound files

3. **Operating System**

   The operating system used for this project is WINDOWS 10.

3.2 NON-FUNCTIONAL REQUIREMENTS

All the other requirements which do not form a part of the above specification are categorized as Non-Functional needs. Comfortable network information measure may additionally be a non-functional requirement of a system.

Usability

It should be easy for the user, and easy to learn, operate through interaction with a system.

Scalability

Application performance should be consistent while processing different kinds of images.

Flexibility

Any hardware and software configuration application should work fine.

Efficiency

Tool should be able to translate any text efficiently.

4. IMPLEMENTATION

4.1 TECHNOLOGIES USED

- **Python**

Python is an interpreted, high-level, general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

- **Audio Processing**

Python has some great libraries for audio processing like Librosa and PyAudio. There are also built-in modules for some basic audio functionalities. It is a Python module to analyze audio signals in general but geared more towards music.

- **Machine Learning**

Machine learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", inorder to make predictions or decisions without being explicitly told.

4.2 STEP BY STEP IMPLEMENTATION

To build a model to recognize emotion from speech using the librosa and sklearn libraries and the RAVDESS dataset.

we will use the libraries librosa, soundfile, and sklearn (among others) to build a model using an MLPClassifier. This will be able to recognize emotion from sound files. 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.

1. Make the necessary imports

```
In [1]: import pandas as pd
        import numpy as np

In [2]: import sys
        import os

In [3]: import glob

In [6]: import librosa

In [7]: import soundfile

In [8]: import pickle

In [9]: from sklearn.model_selection import train_test_split
        from sklearn.neural_network import MLPClassifier
        from sklearn.metrics import accuracy_score
```

2. Define a function extract_feature to extract the mfcc, chroma, and mel features from a sound file. This function takes 4 parameters- the file name and three Boolean parameters for the three features:

   mfcc: Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound

   chroma: Pertains to the 12 different pitch classes

   mel: Mel Spectrogram Frequency

```
In [10]: def extract_feature(file_name, mfcc, chroma, mel):
             with soundfile.SoundFile(file_name) as sound_file:
                 X = sound_file.read(dtype="float32")
                 sample_rate=sound_file.samplerate
                 if chroma:
                     stft=np.abs(librosa.stft(X))
                 result=np.array([])
                 if mfcc:
                     mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
                     result=np.hstack((result, mfccs))
                 if chroma:
                     chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
                     result=np.hstack((result, chroma))
                 if mel:
                     mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
                     result=np.hstack((result, mel))
             return result
```

3. let's define a dictionary to hold numbers and the emotions available in the RAVDESS dataset, and a list to hold those we want- calm, happy, fearful, disgust.

```
In [12]: emotions={
           '01':'neutral',
           '02':'calm',
           '03':'happy',
           '04':'sad',
           '05':'angry',
           '06':'fearful',
           '07':'disgust',
           '08':'surprised'
         }
         observed_emotions=['calm', 'happy', 'fearful', 'disgust']
```

4. let's load the data with a function load_data() – this takes in the relative size of the test set as parameter. x and y are empty lists; we'll use the glob() function from the glob module to get all the pathnames for the sound files in our dataset.

```
In [35]: def load_data(test_size=0.2):
             x,y=[],[]
             for file in glob.glob(r"C:\Users\AISHWARYA\Desktop\speech-emotion-recognition-ravdess-data\Actor_*\\*.wav"):
                 file_name=os.path.basename(file)
                 emotion=emotions[file_name.split("-")[2]]
                 if emotion not in observed_emotions:
                     continue
                 feature=extract_feature(file, mfcc=True, chroma=True, mel=True)
                 x.append(feature)
                 y.append(emotion)
             return train_test_split(np.array(x), y, test_size=test_size, random_state=9)
```

Using our emotions dictionary, this number is turned into an emotion, and our function checks whether this emotion is in our list of observed_emotions; if not, it continues to the next file. It makes a call to extract_feature and stores what is returned in 'feature'. Then, it appends the feature to x and the emotion to y. So, the list x holds the features and y holds the emotions. We call the function

train_test_split with these, the test size, and a random state value, and return that.

5. Split the dataset into training and testing sets, Let's keep the test set 25% of everything and use the load_data function for this.

```
In [36]: x_train,x_test,y_train,y_test=load_data(test_size=0.25)
```

6. Observe the shape of the training and testing datasets

```
In [37]: print((x_train.shape[0], x_test.shape[0]))
         (576, 192)
```

7. And get the number of features extracted.

```
In [38]: print(f'Features extracted: {x_train.shape[1]}')
         Features extracted: 180
```

8. Now, let's initialize an MLPClassifier. This is a Multi-layer Perceptron Classifier; it optimizes the log-loss function using LBFGS or stochastic gradient descent. Unlike SVM or Naive Bayes, the MLPClassifier has an internal neural network for the purpose of classification. This is a feedforward ANN model.

```
In [39]: model=MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)
```

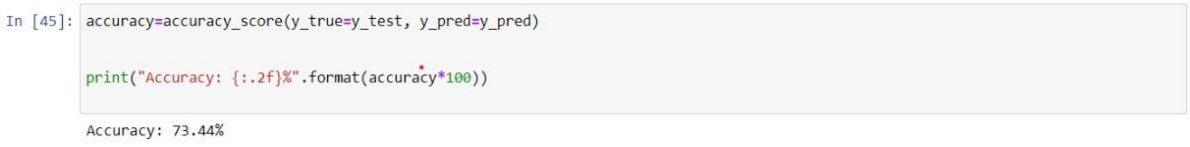
9. Fit/train the model.

```
In [40]: model.fit(x_train,y_train)
Out[40]: MLPClassifier(activation='relu', alpha=0.01, batch_size=256, beta_1=0.9,
                       beta_2=0.999, early_stopping=False, epsilon=1e-08,
                       hidden_layer_sizes=(300,), learning_rate='adaptive',
                       learning_rate_init=0.001, max_fun=15000, max_iter=500,
                       momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
                       power_t=0.5, random_state=None, shuffle=True, solver='adam',
                       tol=0.0001, validation_fraction=0.1, verbose=False,
                       warm_start=False)
```

10. Let's predict the values for the test set. This gives us y_pred (the predicted emotions for the features in the test set).

```
In [41]: y_pred=model.predict(x_test)
```

11. To calculate the accuracy of our model, we'll call up the accuracy_score() function we imported from sklearn. Finally, we'll round the accuracy to 2 decimal places and print it out.



5. RESULT

The Accuracy of 73.44% has been achieved.

```
In [45]: accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

         print("Accuracy: {:.2f}%".format(accuracy*100))

         Accuracy: 73.44%
```

6.CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

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, in voice-based virtual assistants or chatbots, in linguistic research, etc.

6.2 FUTURE WORK

- You can try other different classifiers as well to predict the emotion behind the audio like CNN, SVM, etc.
- Predicting the Live Audio takes a lot of process, and it is sometimes difficult to process as it is unlike the binary data with some csv file associated with it.
- In future, we can predict the random recorded audio as well.
- We can also embed our UI with ML Model, and we can build web-based application with ML using Flask in future, this point can be a great scope.
- More advancements can be more variety of voices can be trained and dataset can be increased to deploy a more realistic model.

7.REFERENCES

1. HEAD FIRST PYTHON A BRAIN FRIENDLY GUIDE, PAUL BARRY.
2. MACHINE LEARNING, TOM M. MITCHELL, VMCGRAW-HILL.
3. HUMAN SPEECH EMOTION RECOGNITION – Maheshwari Selvaraj, Dr.R.Bhuvana, S.Padmaja.
4. SPEECH EMOTION RECOGNITION USING NEURAL NETWORK AND MLP CLASSIFIER
   – Jerry Joy, Aparna Kannan, Shreya Ram, S.Rama.
5. Speech Emotion Recognition (SER) through Machine Learning (analyticsinsight.net)
6. Speech Emotion Recognition Using Deep Learning (dataiku.com)