IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Emotion Recognition using Speech Signals

S. Harsha Vardhan[1], M. P. Rahul[2], P. Kavyasri[3], A. Sraavani[4],
B. Tech Students, Department of Information Technology[1,2,3,4]
Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *Communication is the key to express one's thoughts and ideas. Among all the forms of communication, the most preferred form of communication is speech. In the era of Intelligent systems, the computer human interaction plays a major role in the functionality. This python project uses the libraries present to analyse the audio files. This project gives an overview of the deep learning techniques which are based on feature extraction and model creation which recognizes the emotion of the speaker. This project was made more efficient by the usage of data, efficient methods for feature extraction and classification. The existing models are capable of analysing only the three emotions namely Happy, Angry and Neutral, using this project we can identify Eight different emotions namely Sad, Happy, Calm, Angry, Fearful, Disgust, Surprised and Neutral. The emotion detection through speech improves the functionality of the Intelligent Systems.*

Keywords: Sound File, NumPy, Librosa, Sklearn, MFCC, MEL, CHROME, MLP Classifier, Confusion Matrix

I. INTRODUCTION

Speech Emotion Recognition is a fascinating and challenging task of human-computer interaction. It is the most emerging technology that Artificial Intelligence (AI) has developed from computer scientists to the common man. In this intelligent system functionality, human-computer interaction plays a dominant role. Speech is an efficient way of communication. Humans can alter conversations based on their emotions. Speech emotion detection is the classification problem solved using many machine learning techniques. This project gives an idea about deep learning techniques based on feature extraction, model creation, and prediction.

Emotion recognition is the part of speech recognition that is gaining more popularity and the need for it increased enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.These types of projects can be used for classifying calls in a call center according to emotions, and can be used as a performance parameter for conversational analysis by identifying the unsatisfied customer, customer satisfaction, and helping companies improve their services. It can also be used in-car board systems to prevent accidents by considering the mental state of the driver provided by the system.

II. EXISTING SYSTEMS

The Speech detection system is implemented using Machine Learning(ML). The model will learn from the data provided to it and the predictions and results produced are guided by the data. In this, feature engineering refers to data representation and data quality issues. Feature Engineering is followed by an Algorithmic Based Model development. Here, an ML Algorithm is used to make the model learn about the data and train itself to respond to any new data. The last step in all the ML Models is to evaluate the functioning and efficiency. Developers keep on analyzing different algorithms to achieve efficiency.

III. PROPOSED SYSTEM

In this project python libraries like Librosa, Sound File, NumPy, PyAudio and Scikit-Learn build a model using an MLP Classifier. This can recognize emotion from audio signals by extracting the features using MFCC, Chroma, MEL. MFCC has been used for classifying the speech data into various categories of emotions employing Artificial Neural Networks, this provides the advantage of classifying various emotions in variable length audio signals in a real time environment. This method establishes balance between the computational volume and performance accuracy. The accuracy of the System is analysed using the confusion matrix.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

IV. TECHNOLOGIES USED

4.1 Google Colab

Colaboratory (colab) is a product of Google research. Colab notebooks allow you to combine executable code and text in a single document along with images and datasets. It allows anybody to write and execute python code through the browser. It is well suited to Machine learning, Data analysis and education[1].

4.2 Python Libraries

There are several libraries and packages used in this project for audio analysis. Librosa is a python package that provides the building blocks to create the audio information retrieval system and feature extraction using various signal processing techniques[11]. Sound File is the module used for reading and writing audio files[10]. NumPy is the library used for working on arrays[12]. Scikit-Learn (Sklearn) is the most useful and robust library for Machine Learning and Statistical Modelling. PyAudio is the Cross-Platform audio I/O library[10].

4.3 Data

In this project the RAVDESS dataset is used. This dataset has recordings of 24 actors, 12 male and 12 female in North American Accent. All the emotional expressions come out with two levels of intensity: Normal and Strong, except for the 'neutral' emotion, a total of 2,900+ audio files are used in this project. Here 25% of the total data is taken for testing.

V. FLOW DIAGRAM

The following Fig.1 shows the flow diagram that gives an idea about the process carried out in the project. The data set collected from Kaggle is Imported into the program and the features are extracted. The Feature Extraction is followed by Model building. An MLP classifier model is built using the audio samples in the dataset. Here all the features are classified according to their features. Using the Model the emotion of the audio signal in the testing set is predicted and the accuracy of the model is evaluated.

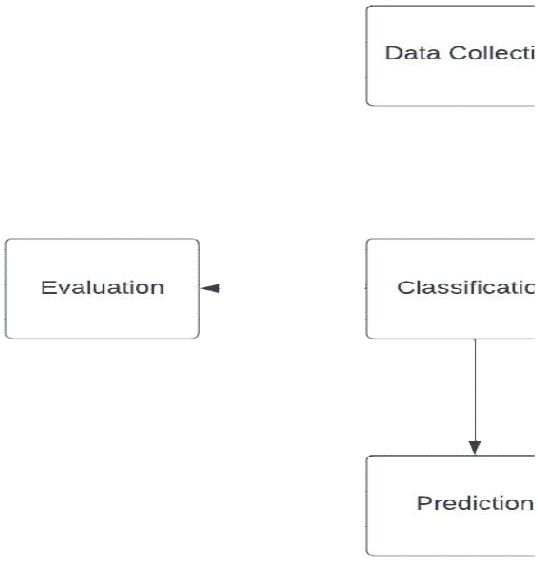

Figure 1: Flow Diagram

VI. MODULES

In this project there are five main modules. They are

- Data Collection
- Feature Extraction
- Splitting

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

- Classification
- Evaluation
- Prediction

6.1 Data Collection

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is taken from the Kaggle[2]. This is a gender balanced dataset with the audios of 24 Professional Actors in North American Accent.

6.2 Feature Extraction

The features that are supported by MFCC, MEL, Chroma are extracted from the audio inputs and a classification model is generated.

- **MFCC** - The features extracted by Mel Frequency cepstral Coefficient (MFCC) feature extraction technique includes a small set of features which concisely describe the overall shape of a Spectral Envelope[5].
- **Chroma** – Chroma features are that they capture Harmonic and Melodic characteristics of audio[7].
- **MEL** – Mel- Spectrogram is one of the efficient methods for audio processing and 8KHz sampling is used for each sample[6].

6.3 Splitting

The data set splitting is the common process in any machine learning program. The dataset is divided into two parts one is the training set that is used for training and model building, and the other one is the testing set that is used to examine the functionality of the system. In this project 25% of the total data is considered as test data. The train_test_split method is used to split the data into train and test sets[9].

6.4 Classification

Speech Emotion Recognition is a classification problem that can be solved using various machine learning techniques. This project makes use of machine learning techniques like Multi Layer Perceptron (MLP) Classifier is used to categorize the data into respective groups which are non-linearly separated. MLP Classifier is a Feed forward Artificial Neural Network model that maps input dataset to a set of outputs[3].

6.5 Evaluation

- **Accuracy:** Accuracy of the system is measured using a confusion matrix[4,8]. It can be used for binary as well as multi classification problems. It measures the quality of prediction from a classification model by looking at how many predictions are true and how many predictions are false.
- **Precision:** Out of all the positive predictions, what percentage is Truly positive. Its value lies between 0 and 1.
- **Recall:** Out of all positives, what percentage are predicted positive. It is the same as True positive rate.
- **F1 Score:** It is the harmonic mean of precision and recall. It takes both false positive and false negative into consideration. It performs well on an Imbalanced Dataset.
- **Support:** It is defined as the Number of actual occurrences of the class in a specified set of data.

6.6 Prediction

As the final step of Speech Emotion Recognition, a single audio signal is given as the input for testing the model. The features from the audio signal are extracted and reshaped. With the help of the MLP Classifier model the emotions are identified. This project can predict Eight Different Emotions Namely Happy, Sad, Angry, Surprised, Disgust, Fearful, Clam and Neutral[3].

VII. RESULTS

Fig 2, Fig 3 Shows the splitting phase, the division of training and testing samples of the dataset along with the number of features and accuracy of the generated MLP Classifier model is retrieved.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

In this project, 89 samples can be considered for training and 30 samples can be used for testing. For all these 119 samples contain 180 features.

Fig.2: Model Building

Fig.3: Data Splitting

Fig 4,5 shows the Evaluation phase, a classification report along with a confusion matrix are returned as the output. This report contains Actual Labels and Prediction Labels. The Actual labels are the eight emotions (angry, disgust, fearful, happy, neutral, sad, surprised) and accuracy, Macro avg, and Weighted avg. The Predicted Labels are Precision, Recall, F1 Score and support.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Fig.4: Accuracy

Fig.5 : Classification Report & Confusion Matrix

The Fig 6 prediction phase a sample audio file from the test sample is given as the input. The features are extracted, reshaped and emotion is predicted

Fig.6: Emotion Prediction

Finally the Accuracy of the model is efficient when compared to the existing models.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Table 1 : Comparison among Different Techniques

MODEL	TIME TAKING	ACCURACY
CNN	more time	86%
MLP	less time	83%
SVM	less time	82%

Table 1 shows the comparison among the three different techniques used for speech emotion recognition. The classification using Convolutional Neural Network (CNN) is more accurate among all the techniques but it takes longer time for execution. The Support Vector Machine (SVM) algorithm is slightly more than Multi Layer Perceptron (MLP) Classifier but it can only classify Three emotions namely (Angry, Happy and Neutral) . So we choose MLP Classifier as it can classify the eight emotions available in the RAVDESS dataset and less time taking than CNN.

VIII. CONCLUSION

This project showed how we can leverage Machine Learning to obtain the underlying emotion from speech audio data and some insights of human expression of emotion through voice. This system can be employed in a variety of setups like call centers for complaints or marketing, in voice based Virtual assistants or chatbot, etc. In this project MLP classifier is used for emotion detection. The accuracy of this algorithm is 83% when training samples are 89 and testing samples are 30. Even though CNN showed 86% of accuracy, it takes more time to run the algorithm. Where as MLP classifer have less accuracy compared to CNN but the execution time is very less.

IX. SCOPE OF FUTURE WORK

In the era of Artificial intelligence there is a chance of Robot intervention into Human life , to make the robots react according to the human feelings. This Speech Emotion Recognition System may be helps a lot. Training the model with different languages and accents can help in enabling the system to identify the emotion more accurately.

REFERENCES

[1]. https://colab.research.google.com/drive/1pXYDKsBVenXQhiFLrII4tv4Wp46LqMl4#scroll To=5I6CDcGoCzG6
[2]. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
[3]. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.htm l
[4]. https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classificati on/#:~:text=Confusion%20Matrix%20is%20used%20to,number%20of%20classes%20 or%20 outputs.
[5]. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/
[6]. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
[7]. https://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/
[8]. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance
[9]. -measures/
[10]. https://www.sharpsightlabs.com/blog/scikit-train_test_split/
[11]. https://pysoundfile.readthedocs.io/en/latest/
[12]. https://www.javatpoint.com/librosa-library-in-python
[13]. https://www.w3schools.com/python/numpy/numpy_intro.asp
[14]. https://www.sciencedirect.com/science/article/pii/S1746809418302337
[15]. 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.
[16]. 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.
[17]. T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models,"

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.