Speech Emotion Recognition using MLP Classifier

Roshan Wagh[1], Yash Gade[2] , Abhishek Wagh[3] , Amon Bansod[4]
Department of computer Engineering
K.K.Wagh Institute of Engineering Education and Research

Abstract :
As human beings speech is natural way to express ourselves. Humans depend so much on it. Emotions play a important role in communication . Detection and analysis of emotion is very important in today's digital world.Emotion detection is a challenging task. There is not a general agreement on how to measure or categorize them. Speech Emotion Recognition process and classify speech signals to detect emotions embedded in them. Speech Emotion Recognition system can be used in various areas.The application area are like interactive voice based-assistant , caller agent conversation analysis,security and other fields. This System attempts to detect emotions in audio file passed by analysing the acoustic features. System uses MLP Classifier to classify the emotions from the given wave signal. RAVDESS dataset will be used .The features to be extracted from the audio input provided will be attracted by these five parameters which are as follows, MFCC, Contrast, Mel Spectrograph
Frequency, Chroma and Tonnetz.

Keywords : MLP Classifier,MFCC ,Chroma, Ravdess Dataset, Tonnetz, Neural Networks,
Contrast ,Mel,etc.

Introduction :

Speech Emotion Recognition is one among the booming analysis topics within the computer science world. Emotion is also a medium by that one expresses however a private feels and one's state of mind. Emotions play an important role in sensitive job areas, like that of a medical, a Military Commander and plenty of others wherever one must maintain their emotions. Predicting emotions
may be a strong task as each individual has a different tone and intonation of speech. The different sorts of emotions are anger ,happy, neutral, disgust, sad, surprised. The goal of the system is to classify these emotions from a given speech sample with most applicable technique . System goes to use MLP Classifier for predicting emotions. two classifiers i.e., Support Vector Machine (SVM) and Multi Layer Perceptron Classifier (MLP Classifier) are been campared. Support Vector Machine is efficient in predicting emotions for sound input with no discrepancy, in presence of noisy input it deviates from its pre- diction. Support Vector Machine only classifies employing a single plane . SVM restricts the prediction. The comparison shows that the system using the Support Vector Machine i.e., SVM features a more computational time, even although having a decent accuracy.

Motivation : Emotions are staple items for humans which are impacting their everyday routine and activities like decision-making ,communication and learning they are expressed through speech ,facial expressions, gestures and other actions. SER's
basic use is to research various aspects of speech from human voice. There are various features in human voice like pitch , tone, etc which affect emotions. This assumption is supported by the very fact that majority affective states involve physiological reactions which in turn the method by which voice is produced. For Example ,When a person is happy ,His body feels calmness ,when he is angry, muscle tension increases affecting acoustics of speech. In the past, Emotion Recognition was a difficult tasks many scientist did not take a initiative to try and do research thereon, Although the sphere has recently received an increase in contributions, it is new emerging field with multiple applications. These include user friendly application for emotion recognition via user sound; Emotion Detection at automated call center ; or changing kind of teaching by tutor if students are becoming bored. A replacement application of emotion detection is in security field to acknowledge inner feelings of criminals through his voice. Precise measurement and analysis of the voice may be a difficult task and was completely subjective with in the past Now many latest methods are been applied which makes this field one in every of the interesting field of study.

Literature Survey :

1 . A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM:

In this system, the quality of feature extraction directly affected the accuracy of speech emotion recognition. With within the process of feature extraction, it always took the whole emotion sentence as units for feature extracting, and extraction contents were several aspects of emotion speech, which were several acoustic characteristics of time, fundamental construction, fundamental frequency construction, and formant construction. Then contrast emotion speech with no emotion sentence from these aspects, acquiring the law of emotional signal distribution, then classify emotion speech . Deep neural network (DNN) has unprecedented success with within the field of speech recognition and image recognition ; however, so far no research on deep neural network has been applied to speech emotion processing. It is found that the deep belief network of DNN in speech emotion process in contains a enormous advantage.

Therefore the proposed a way to the understand to comprehend to realize the emotional features automatically extracted from the sentence. It used DBN to coach a 5-layer-deep network to extract speech emotion features. The speech emotion recognition rate reached 86.5 percent,which was 7 percent over the initial method.

2. Human Emotion Recognition From Audio And Video Signals:

In the past decade a plenty of research has gone into Automatic Speech Emotion Recognition(SER). The primary objective of SER is to boost man-machine interface. It can also be used to monitor the psycho physiological state of a someone in lie detectors. In recent time,speech emotion recognitional so find it s applications in medicine and forensics. During this this study 7 emotions are recognized using pitch and prosody features. Majority of the speech features utilized during this work are in time domain. Support Vector Machine (SVM) classifier has been used for classifying the emotions. Berlin emotional database is chosen for the task. An honest recognition rate of 81 percent was obtained.

3.Speech Emotion Recognition Using SVM:

This study examines the implications of reduced speech bandwidth . A step by step description of a real-time speech emotion recognition employing a pre-trained image classification network AlexNet is given. The results showed that the baseline approach achieved a mean accuracy.of 82percent when trained on the Berlin Emotional Speech (EMO-DB) data with seven categorical emotions. Reduction of the frequency from the baseline 16–8 kHz led to a decrease of SER accuracy by about 3.3 percent. The companding procedure on its own reduced the common accuracy by 3.8 percent, and also the combined effect of companding and band reduction decreased the accuracy by about 7percent compared to the baseline results. The SER was implemented in real-time with emotional labels generated every 1.033–1.026 s. Real-time implementation timelines are presented.

4. Speech Emotion Recognition using MLP Classifier:

Speech Emotion Recognition, abbreviated as SER, is that the act of attempting to recognize human emotion and so the associated affective states from speech. With within the Speech Emotion Recognition System (SER), the audio files are given because the input. The information sets travels through a variety of blocks of processes which makes it executable to assist for the analysis of the speech parameters. The data is preprocessed to alter it to the acceptable format and also the respective features from the audio files are extracted using various steps such as framing, windowing, etc. This process helps in breaking down the audio files into the numerical values which represents the frequency, time, amplitude orany other such parameters which can help within the analysis of the audio files. After the extraction of the desired features from the audio files, the model is trained. We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. For the training, we store the numerical values of emotion and the irrespective features correspondingly in different arrays. The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. The results obtained during this study demonstrate that speech recognition is feasible, Which MLPs may be used for any task concerning recognizing of speech and demonstrating the accuracy of every emotion present within the speech

Proposed Methodology :

The underlying emotion in speech is reflected in voice through tone and pitch .System aims to classify various types of emotions like anger, disgust ,happy, sad, neutal, etc. System uses neural network for that purpose.MLP Classifier will be used to perform this task.The dataset used is RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset).

A. Database Description RAVDESS dataset has recordings of 24 actors, 12 male actors and 12 female actors, the actors are numbered from 01 to 24. The emotions contained in the dataset are as sad, happy, neutral, angry, disgust, surprised, fearful and calm expressions.

B. Neural Network and Multi-Layered Perceptron classifier

A MLP is a network which is made up of perceptron. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and there can be multiple hidden layer between input and output layers based on our requirement.

Multilayer perceptron is applied for supervised learning problems. The MLP is used for the classification. The MLP is trained on given dataset.. MLP learns the correlation between the set of inputs and outputs from training phase. The MLP adjusts model parameters like weights and biases to minimize the error. Multi-Layer Perceptron Classifier Multi-layer Perceptron Classifier (MLP Classifier) relies Neural Network to perform classification Task . MLP Classifier implements a MLP algorithm and trains the Neural Network using Back-propagation. Building the MLP Classifier involves the subsequent steps.

1. Initialize the MLP Classifier by initiating the specific parameters.

2. Neural Network are fed with data for training purpose.

3. Output is predicted on basis of trained neural network.

4. Calculate the accuracy of the predictions.

Begin itemize

• Feature Extraction

Tone and pitch are the emotions frequently reflected by voice . The objective of feature extraction is to reveal applicable feature from discourse signals as for feelings. Five feature are extracted from the discourse signals given as information. The five features are, MFCC, Contrast, Mels Spectrograph Frequency, Chroma and Tonnetz.

• MFCC

Mel Frequency Cepstral Coefficients(MFCC) is utilised to recover the sound from the given wav audio file by utilising distinct hop length and HTK-styles mel frequencies.

Formula : Mel (freq)= 2595 * log 10 (1+freq/700)

• Mel

The Mel scale is employed to relate evident repeat, or pitch, of an unadulterated tone to its real recurrence. Individuals are incredibly improved at perceiving little changes in pitch at low frequencies as compared to high frequencies.

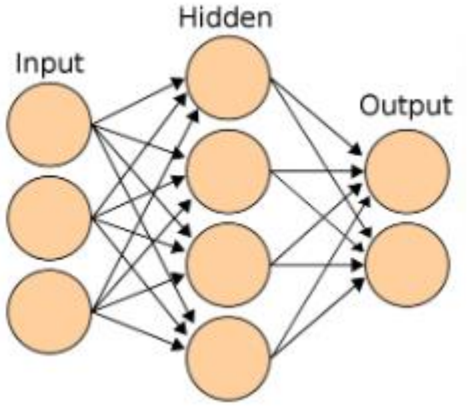

Fig1 . MLP classifier

Implementation :

OVERVIEW OF PROJECT MODULES

1)Routing Module : App routing is used to map the specific URL with the associated function which is intended to perform some specific task. It is used to access a particularpage. For eg. This System : @app.route('/conclusion', methods = ['POST'])

It will help to route system to the conclusion page

2)Feature Extracting Module : The goal of Feature Extraction is to reduce the number of features in a data set by creating new features from the existing ones . As the data set contains the audio file, useful features need to be extracted from it. Features being extracted here are MFCC, Chroma, and MEL.

3)Graph Plotting Module : Different plots like spectogram and Amplitude plot have been used. The python library 'Matplotlib' is used to plot the respective graphs.

4)Model Saving Module : After building the model successfully, the 'Pickle' library is used to save the model. It helps to get the model into action whenever required .

TOOLS AND TECHNOLOGY USED

Tools:- Jupyter Notebook, VS Code.

Technologies:- Machine Learning ,Python, HTML, CSS, JavaScript, Bootstrap.

ALGORITHM DETAILS

step **1** : Start

step **2** : Receive the input from user. The file uploaded must be in (.wav) format

step **3** : Save the file to a particular location.

step **4** : Make use of different plots to show the behaviour of audio file
step **5** : Pass the same file to Extract Features, it will extract the features like MFCC,
CHROMA, MEL.
step **6** : Pass these features to the model built using MLP Classifier .
step **7** : It will predict the emotion .
step **8** : END

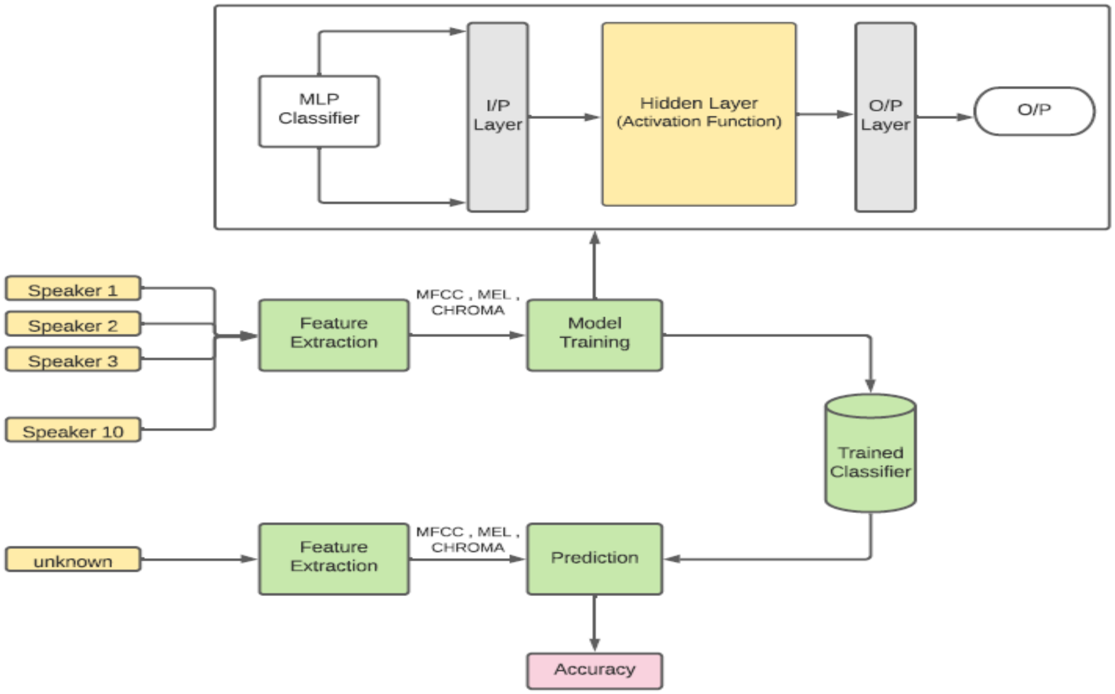

Fig2. Final Design

Training workflow :

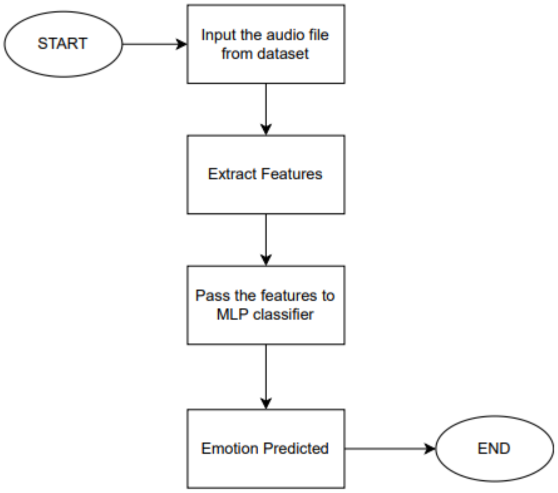
START
Input the audio file from dataset
Extract Features
Pass the features to MLP classifier
Emotion Predicted
END

Fig.Training workflow

Testing workflow :

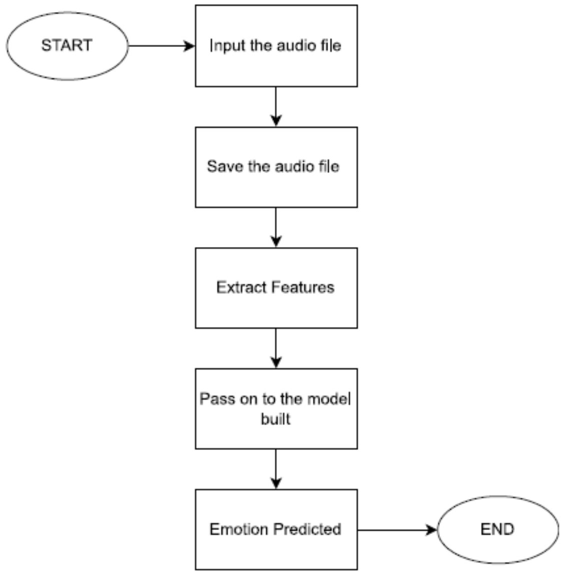

Fig .Testing Workflow

Result Analysis :

This system shows how Machine Learning can be used to obtain the underlying emotion from speech audio data and obtain various insights about human voice. The system can be employed in a variety of setups like Call Centre: for complaints or marketing, voice-based virtual assistants or chatbots, education sector for assisting teachers to identify the emotions of students pertaining to the content been taught etc.

Sample GUI :

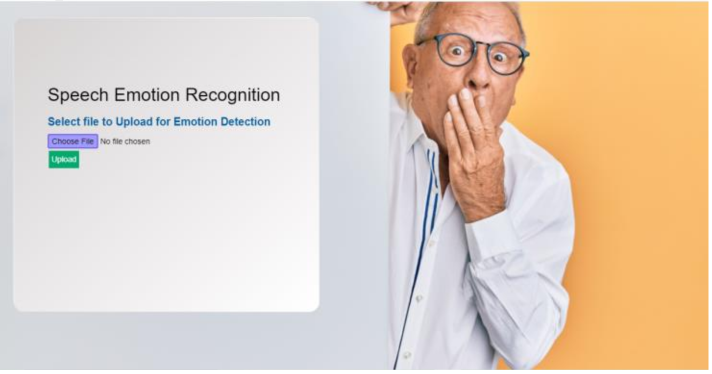
Speech Emotion Recognition
Select file to Upload for Emotion Detection
Choose File
No file chosen
Upload

Fig3 .sample GUI

Efficiency Issues :

- MLP requires tuning of a number of hyper parameters such as the number of
neurons hidden, hidden layers, and iterations.

- Poor robustness leads to noise signals as if there is at least one frequency band
which is skewed. The noise signal changes all MFCC's

Future Work :

1 . The system manages to achieve an 80 - 85 percent accuracy. So the next task will be to improve the accuracy of the system; resulting in better prediction.

2 . In the current scenario, the system can predict four emotions, which motivates to improvise the model to predict more than four emotions.

Conclusion :

The motive to build this system is to use Neural Network and imply the concept of Emotion Recognition from the input-speech audio file. The basic task is to use MLP classifier to perform classification of different emotions form the input. RAVDESS Dataset has been used io implement the test cases of the system. In this system, extraction
of different features of speech takes place. They are MFCC, Contrast, MEL Spectrograph Frequency, Chroma and Tonnetz. In this System; Initially the input audio file is loaded. Next, filters are applied to remove noise and unwanted signals. Further, different MFCC features are extracted. Hereafter, training of Neural Network and simulation takes place to result as the final output. As per the researches, MLP Classifier performs with high accuracy as well as efficiency. It is cost effective
as compared to other methods like SVM, CNN, Binary Classification, etc. The system shows how Machine Learning can be used to obtain the underlying emotion from speech audio data and obtain various insights about human voice. This
system can be used in Call Centre for complaints or also for marketing purpose , in voice-based virtual assistants or chat bots, in education sector for assisting teachers to identify the emotions of the students about content been taught etc.

Application :

1) Used by chatbots, companies/industries for recognizing emotion of customer.
2) Simulated online learning environment

References :

1]. Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition. IJTEE, October 2020.
2]. Jianfeng Zhao, Xia Mao Deep features to Recognise Speech Emotion using Merged Deep Convolution neural network(CNN). IET Signal Process., 2018
3]. H.K. Palo, Mihir Narayana Mohanty Application of different features for recognizing features via MLP network. Springer India 2018, Computational Vision and Robotics, Advancement in Intelligent Systems and Computing

4]. Ayush Kumar Shah ,Mansi Kattel,Araju Nepal. Chroma Feature Extraction using Fourier Transform. January 2019
5]. Sabur Ajibola and Nahrul Khair Commonly Used Speech Feature Extraction Algorithms.DOI: 10.5772/intechopen.80419.
6]. Davis, P.(1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech and for Signal Processing, Vol. 28 No. 4, pp. 357-366 .
7].X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001
8]. Kerkeni, et al.: Automatic speech emotion recognition using machine learning. In: Social Media and Machine Learning. IntechOpen (2019)
9]. Mirsanddi, S., Barsoum, E., Zhangi, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–223.
IEEE, 15 Mar 2017
10] M. S. Hussain and G. Muhammad, “Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data”, Information Fusion, vol. 49, pp. 69-78, September 2019.