SPEECH EMOTION RECOGNITION

NAME:JOSIKA P
ROLLNO:2019202020
CLASS:MCA-R
GUIDE:MS.R.L.JASMINE

ABSTRACT:

This project aims at building and training speech and emotion recognition system by using machine learning and deep learning algorithm which uses CNN which is a type of artificial neural network which is widely used for image/object recognition.This project is to predict emotion based on the speech used to perform analytical research by applying different machine learning algorithms and neural networks with different architecture and compare their results for insights.This project uses data set RAVDESS for classification and uses 9 emotions neutral,calm,happy,sad,angry,fear,disgust,pleasant surprise and boredom.Feature Extraction is done using librosa library which comes under pre processing data.The next step is to build model using Convolution Neural Network then prediction is done to find accuracy with various algorithms to find the better one.

INTRODUCTION:

Speech is a rich, dense form of communication that can convey information effectively. It contains two types of information, namely linguistic and paralinguistic. The former refers to the verbal content, the underlying language code, while the latter refers to the implicit information such as body language, gestures, facial expressions, tone, pitch, emotion etc. Para linguistic characteristics can help understand the mental state of the person (emotion), gender, attitude, dialect]. Recorded speech has key features that can be leveraged to extract information, such as emotion, in a structured way. To get such information would be invaluable in facilitating more natural conversations between the virtual assistant and the user since emotion color everyday human interactions.This study focuses on identifying the best audio feature and model architecture for emotion recognition in speech. The experiments were carried out on "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)",Surrey Audio-Visual Expressed Emotion (SAVEE),Toronto Emotional Speech Set(TESS),Crowd Sourced Emotional Multimodal Actors Dataset(CREMA-D) dataset .In deep learning, a **convolutional neural network** (**CNN/ConvNet**) is a class of deep neural networks, most commonly applied to analyze visual imagery. Now when we think of a neural network we think about matrix multiplications but that is not the case with ConvNet. It uses a special technique called Convolution. Now in mathematics **convolution** is a mathematical operation on two functions that produces a third function that expresses how the shape of one is modified by the other. Convolution Neural Network (CNN) is an extension of DNN that operates on data that come in the form of several arrays, in particular images. Much as with signals represented as a single-dimensional array, the input has philters that are screwed over it and then packed together for a smaller dimension. This is done to collect local input data, and if replicated, a hierarchy of features is created.

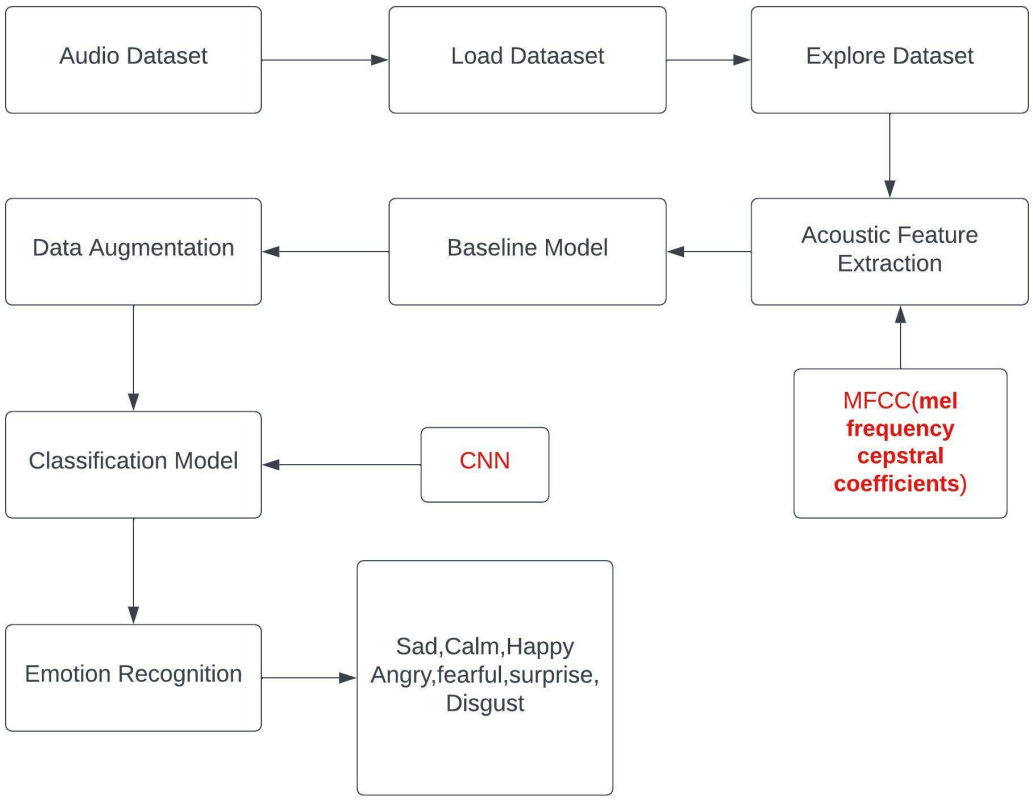
PROBLEM STATEMENT:

Recognition of emotions in audio signals has been a field of study in the past. Previous work in this area included use of various classifiers like SVM, Neural Networks, Bayes Classifier etc. The number of emotions classified varied from study to study, they play an important aspect in evaluating the accuracy of the different classifiers.Using Machine learning models to speech emotion recognition has less accuracy in order to overcome this issue we have been using a model from deep learning called convolution neural network(CNN).

OBJECTIVE:

- Using CNN to predict emotions (Neutral,Calm,Happy,Sad,Angry,Fear,Disgust,Pleasant ,Surprise,Boredom) based on voice.
- Bring better accuracy and comparing it with machine learning models like Hidden Markov Models (HMMs),SVM,Decision Tree and deep learning models like LSTM,RNN,MLP

ARCHITECTURE DIAGRAM :



ARCHITECTURE EXPLANATION:

AUDIO DATASET:

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

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.

Surrey Audio-Visual Expressed Emotion (SAVEE)

The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories.

Toronto Emotional Speech Set(TESS)

There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total.

The dataset is organised such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

Crowd Sourced Emotional Multimodal Actors Dataset(CREMA-D)

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

LOAD DATASET:

Each audio file contains a 7-part numerical identifier each denoting the modality, vocal channel, emotion, emotional intensity, statement, repetition and the actor respectively. The naming convention followed a pattern, wherein odd actors and even actors denoted male

and female sex respectively. We extracted all these information from the file names into metadata. The target variable is the emotion that the audio recording was classified as.

EXPLORE THE DATA

The wave plot is a graphical representation of a sound wave vibration overtime. Its in this wave that we need to find the key pattern that will help us distinguish the different emotions.We're going to plot one or two audio files here selected randomly, just to get a feel for the type of data we're dealing with. Eg. Does it contain lots of background noise. Is the emotions clear. The idea being that, if a human struggles to interpret the data, then its very likely the model isn't going to do a very good job either.

ACOUSTIC FEATURE EXTRACTION:

Audio features can be broadly classified into two categories, namely time-domain features and frequency-domain features. Time-domain features include the short-term energy of signal, zero crossing rate, maximum amplitude, minimum energy, entropy of energy. These features are very easy to extract and provide a simpler way to analyze audio signals. Under limited data, frequency domain features reveal deeper patterns in the audio signal, which can potentially help us identify the underlying emotion of the signal. Frequency-domain features include spectograms, Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral rolloff, spectral entropy and chroma coefficients

Mel-Frequency Cepstrum (MFC) Mel-Frequency Cepstrum is a representation of the short-term power spectrum of a sound by transforming the audio signal through a series of steps to mimic the human cochlea. The Mel scale is important because it better approximates human-based perception of sound as opposed to linear scales.Mel-Frequency Cepstral Coefficients (MFCC) are coefficients which capture the envelope of the short time power spectrum.

Mel-Spectograms A spectogram is a time vs frequency representation of an audio signal. Different emotions exhibit different patterns in the energy spectrum. Mel-spectogram is a representation of the audio signal on a Mel-scale. The logarithmic form of mel-spectogram helps understand emotions better because humans perceive sound in logarithmic scale.

BASELINE MODEL

Here we will build a baseline model for an emotion classifier.Baseline, mean its the simplest most parsimonious model. And view points will vary from one data scientist to another, but essentially its a model **NOT** meant to achieve full accuracy potential. It's just to qucikly test the framework and setup the blueprint for how we go about creating a workable emotion classifier.

DATA AUGMENTATION

To generate the copies or syntactic data for increment in dataset, augmentation has been applied through injection of noise, pitch change, time and speed change to syntactic data. The augmentation with injection of noise and pitch change is done in this work. Signal after adding white noise and pitch.

CLASSIFICATION MODEL:

Convolution layer A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i.e., convolution operation and activation function. Nonlinear activation function The outputs of a linear operation such as convolution are then passed through a nonlinear activation function.The most common nonlinear activation function used presently is the rectified linear unit (ReLU).

PREDICT EMOTION:

Hence the final step is to predict the emotions like angry,sad,happy,fearful,disgust based on the above cnn model.

LIST OF MODULES:

1. Audio Feature Extraction and Visualizations.
2. To train the model for accuracy calculation.
3. Implementation process of CNN model.
4. Classification of speech emotions.

BRIEF DESCRIPTION OF MODULES :

In our CNN model we have four important layers:

1. Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence.
2. Activation layer: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used corrected linear unit (ReLU) during our work.
3. Max Pooling layer: This layer enables options with the maximum value to the Dense layers. It helps to keep the variable length inputs to a fixed sized feature array.
4. Dense layer

➢ Audio Feature Extraction and Visualizations. (module01)

Characteristics extraction is required for classification and depiction. The audio signal is a 3D signal in which 3 axes indicate time, amplitude and frequency. We will use librosa to analyze and extract characteristics of any audio signal. (.load) function pulls an audio file and decrypts it into a 1D array which is of time series x, and SR is actually sampling rate of x. By default SR is 22 kHz. Here I will show one audio file display with the use of (IPython.display) function. Librosa.display is important to represent the audio files in various forms i.e. wave plot, spectrogram and colormap. Wave plots use loudness of the audio at a particular time. Spectrogram displays various frequencies for a particular time with its amplitude.

➢ To train the model for accuracy calculation. (module02)

Within this module we train the model for accuracy estimations. 1 st , import necessary modules. Then pull the dataset. We will receive the sampling rate value with librosa packages and mfcc function. Thereafter this value holds other variables. Now audio files and mfcc value hold a variable consequently it will add a list. Then zip the list and hold

two variables x & y. Then we have represented (x, y) shape values with the use of numpy package.

➢ **Implementation process of CNN model. (module03)**

Speech represented in the form of image with 3 layers. While using CNN, do consider, 1st and 2nd derivatives of speech image with time and frequency. CNN can predict, analyze the speech data, CNN can learn from speeches and identify words or utterances.

➢ **Classification of speech emotions. (module04)**

When testing we provide the audio input. Next, we run the audio in order to hear with ipython.disply packages. Thereafter plot the audio features with librosa.display.waveplot packages. Extract the Characteristics using librosa.load. It converts one data frame and display structured form. Further it compares loaded model by predict function batch size 32. Ultimately it displays the output from the audio file what sort of expression/emotion that audio file has.

REFERENCES:

[1] Dong Yu and Li Deng. AUTOMATIC SPEECH RECOGNITION. Springer, 2016.

[2] Samira Ebrahimi, Vincent Michalski, Kishore Konda, Goethe Roland Memisevic, Christopher Pal— Recurrent Neural Networks for Emotion Recognition in Video‖, Kahou École Polytechnique de Montréal, Canada ; Universität Frankfurt, Germany; Université de Montréal, Montréal, Canada; 2015.

[3] Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010.

[4] Demis Hassabis, Dharshan Kumaran, Christopher Summer eld, and Matthew Botvinick. Neuroscience inspired artificial intelligence.

[5] Marvin Minsky. The emotion machine: Commonsense thinking.

[6] artificial intelligence, and the future of the human mind. Simon and Schuster, 2007.

[7] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. Artificial intelligence: a modern approach, volume 2. Prentice hall Upper Saddle River, 2003.

[8] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition, volume 14. PTR Prentice Hall Englewood Clis, 1993.

[9] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. A maximum likelihood approach to continuous speech recognition. In Readings in speech recognition, pages 308{319. Elsevier, 1990.

[10] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. The Bell System Technical Journal, 62(4):1035{1074, 1983.