Speech Emotion Recognition

Soma Pavan Kumar
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Speech Emotion recognition is a speedly growing research domain in recent years.
In this paper, the basic seven emotions (Angry, Happy, Fear, Neutral, surprise, sad, and disgust) are analysed from emotional speech signals.
Not at all like people, machines cannot see and appear feelings. But human-computer interaction can be progressed by mechanized feelings acknowledgment, in this manner decreasing the require for human intervention
In this project, we used some techniques to develop this project such as GRU and LSTM and also use some feature extraction and data augmentation
To increase the accuracy of our model.

Keywords: GRU, LSTM, CNN, Melspectrogram

1.Problem Statement

Verbal Communication is valuable and sought after in workplace and classroom environments alike. There is no denying the notion that Indians lack verbal communication and consequently lag in the workplace or classroom environments. This happens despite them having strong technical competencies. Clear and comprehensive speech is the vital backbone of strong communication and presentation skills. Where some work consists mainly of presenting, most careers require and prosper from the ability to communicate effectively.

Research has shown that verbal communication remains one of the most employable skills in both the perception of employers and new graduates. Of the possible improvements to speech, pause, and stutter, in particular, remain one of the most common and prominent factors of someone's show. Millions of people are affected by stuttering and other speech disfluencies, with the majority of the world having experienced mild stutters while communicating under stressful conditions.

Research shows that mild disfluencies can be cured without medical help, just practicing speech regularly and constructive feedbacks are effective ways to improve. We, Data Scientists recognize this problem and say hello.

2. Introduction

Speech is the most natural form of communication between humans and computers speech is a complex signal.

It contains information regarding message speaker language and emotion. so, emotion makes speech more attractive more effective more expressive. Speech emotion recognition means understanding the emotional state of a human by extracting or detecting the feature extracted by his/her voice.
There are some universal emotions including Neutral, anger, joy, sadness in which any intelligent system with limited computer resources can be trained to recognize or synthesize as needed.

3. Dataset

The RAVDESS is a standard multimodal database of emotional speech and song. The database is gender-balanced consisting of 24 professional actors, vocalizing lexically matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, frightened, surprised, and disgusting expressions, and the song contains calm, happy, sad, angry, and frightened emotions. Each expression is formed on two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and voice, face-only, and voice-only formats.

The set of 7356 recordings was rated 10 times each on emotional validity, intensity, and authenticity. The Ratings were provided by 247 individuals that were characteristic of untrained research participants in North America. The next set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intruder reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. All recordings are made freely available under a Creative Commons license and can be downloaded at https://www.kaggle.com/uwrfkaggler/ravdess-emotional-song-audio

4. Data Augmentation

Data augmentation is that the method by which we make unused engineered information tests by including little annoyances to our introductory preparing set.

To create syntactic data for sound, we are going to apply commotion infusion, moving time, changing pitch, and speed.

Data augmentation is a technique that it reduces the overfitting of model and act as a regularizer

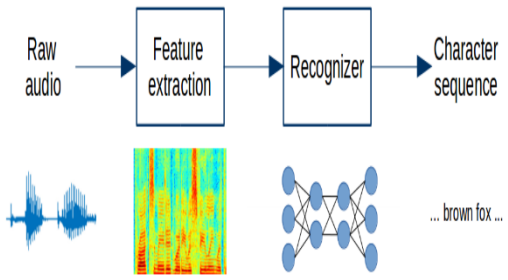
5. Feature Extraction

In Feature Extraction, we extract features and processing the info.

Broadly highlight extraction procedures are classified as worldly examination and unearthly examination techniques.

In temporal analysis, the speech waveform itself is employed for analysis.

In spectral analysis spectral representation of the speech, a sign is employed for analysis.



There are few techniques that we are using in this project MFCC and Mel spectrogram this is the feature extraction technique.

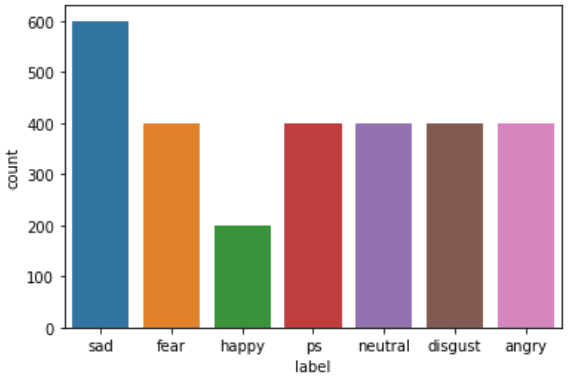
6. Steps involved:

- **Exploratory Data Analysis**

  The First Thing First EDA using this we get intuition about how the data has structed.

  After saw the data set first we analysis how many emotion are present in the dataset so after that we got most common seven emotion

  It helped us understand which features behave in which ways in relation to the target variable.



- **Feature Extraction**

- MFCC (Mel Frequency Cepstral Coefficients)- MFCC can be used to extract the distinctive properties of human voice, and this MFCC also represents the short-term power spectrum of human voice. MFCC is used to produce the coefficients that describe the frequency Cepstral; these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale frequency.

- Mel Spectrogram- A Fast Fourier Transform is computed on overlapping windowed segments of the signal and that we get what's called the spectrogram A spectrogram may be a visual way of speaking to the flag quality, or “loudness”, of a flag over time at different frequencies display in a specific waveform.

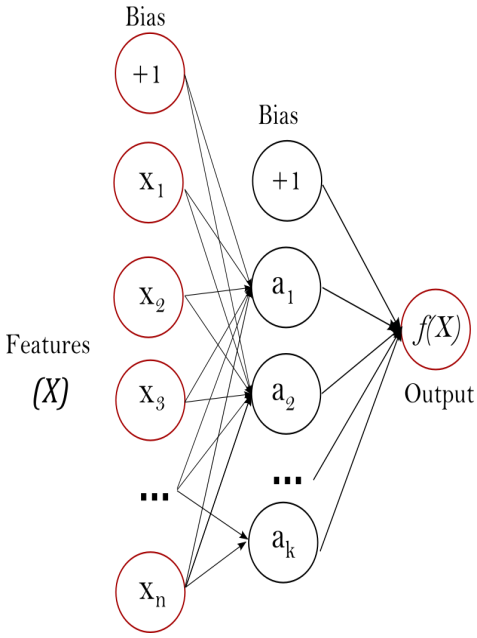
- **<u>Fitting different models</u>**
  For modeling, we tried various classification algorithms like:

1. **GRU**
2. **LSTM**

7.1. Algorithms:

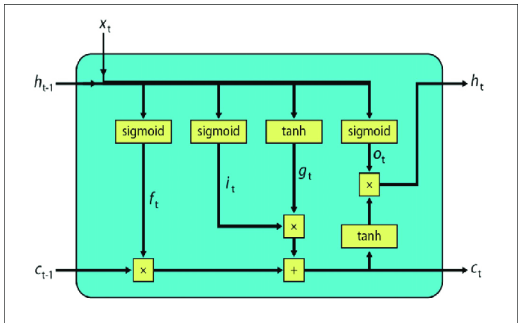
1. MLP (Multi Layer Perceptron):

Multilayer perceptron (MLP) classifier is a supervised classification technique that uses backpropagation for training. It is one of the feed-forward artificial neural networks (ANN) classes. It consists of more than one perceptron. It consists of one output layer, one input layer, and in between these input and output layers, there may be an arbitrary number of hidden layers based on the user's choice. That means it should contain at least three layers input layer, hidden layer, output layer. Expect the input layer; every layer is a neuron that uses a nonlinear activation function. Its nonlinear activation function, multiple layers distinguish this from a single layer feed-forward neural network. Since it has nonlinear activation, it can be able to distinguish the data that is not linearly separable.



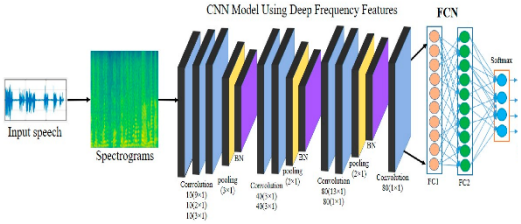
2. LSTM (Long Short-Term Memory):

Long Short-Term Memory systems are the special mode of RNN that have the capabilities to memorize long-term conditions and rectify choice to work in a profound assortment of problems. These are specially planned to manage with long-term conditions issues, by default behaviour, they can keep in mind data for a colossal span to time. Like the chain-like structure of rehashing modules of RNN, LSTM incorporates a distinctive structure of the rehashing modules, it has the set of four neural organize layer that connected with each other in a special way.



3. CNN (Convolutional Neural Network):

Convolutional neural networks (CNNs) are one of the most popular deep learning models that have manifested remarkable success in the research areas such as 14 object recognition, face recognition, handwriting recognition, speech recognition, and natural language processing. The term convolution comes from the fact that convolution—the mathematical operation—is employed in these networks. Generally, CNNs have three fundamental building blocks: the convolutional layer, the pooling layer, and the fully connected layer. Following, we describe these building blocks along with some basic concepts such as SoftMax unit, rectified linear unit, and dropout.

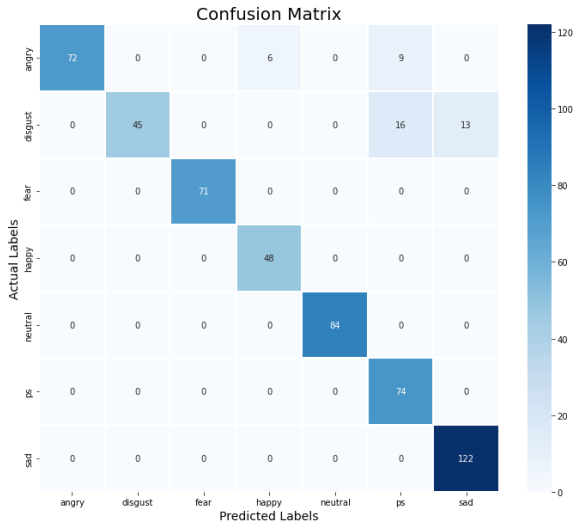


7.2. Model performance:

A model can be evaluated by various metrics such as:

1. **Confusion Matrix-**

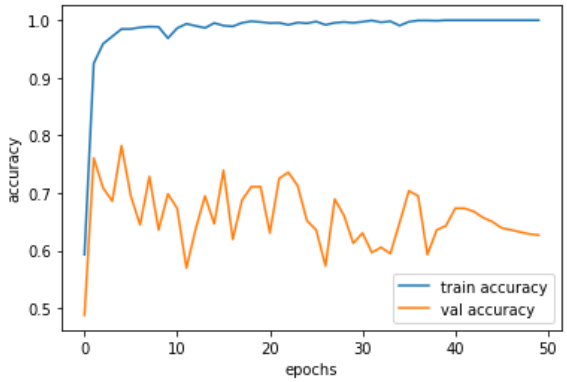
The confusion matrix is a table that summarizes how successful the classification model is in estimating examples related to different classes. One axis of the confusion matrix is the label predicted by model , and the other axis is the actual label.

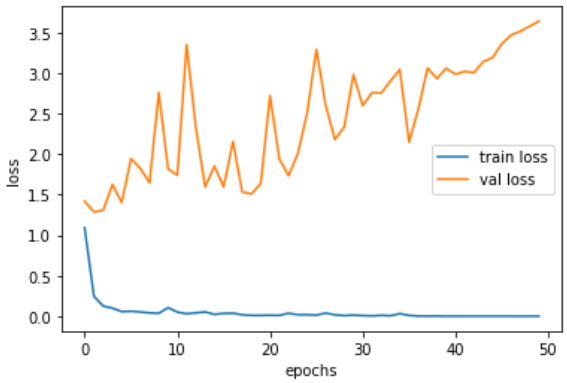


8. Conclusion:

In this Project, we utilize a few excellent procedures like LSTM, GRU. After utilizing all demonstrate LSTM gave a good accuracy. After that we also use Data

Augmentation could be a strategy utilized to extend the sum of information by including marginally adjusted duplicates of as of now existing information or recently made engineered information from existing information. It acts as a regularizer and makes a difference decrease overfitting when preparing a machine learning demonstrate. So this extend makes a difference to anticipate feeling utilizing discourse.





References-

- Dong Yu and Li Deng. AUTOMATIC SPEECH RECOGNITION. Springer, 2016.
- 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.
- Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010.
- https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/