**Speech Emotion Recognition**

**Kartik Pandey,Aniket Nichat,**

**Rohit Thawali,Sagar Khekale**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Emotion recognition is a rapidly growing research domain in recent years. Unlike humans, machines cannot perceive and show emotions. But human-computer interaction can be improved by automated emotions recognition, thereby reducing the need for human intervention

In this paper, the basic seven emotions (Angry, Happy, Fear, Neutral, surprise, sad, and disgust ) are analyzed from emotional speech signals.

In this Project, we used some techniques to develop this project such as KNN, MLP classifier, and CNN and also use some feature extraction and data augmentation

To increase the accuracy of our model.

***Keywords:* KNN,MLP, CNN, feature extraction, data augmentation**

**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 occupations consist mainly of presenting, most careers require and thrive 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 vocal presentations tone, disfluencies, and stutters, in particular, remain one of the most common and prominent factors of someone’s demonstration. 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**

### In today’s digital era, speech signals have become a mode of communication between humans and machines which is possible by various technological advancements. Speech recognition techniques with methodologies signal processing techniques made leads to Speech-to text (STT) technology which is used mobile phones as a mode of communication.

### Speech Recognition is the fastest growing research topic in which attempts to recognize speech signals. This leads to Speech Emotion Recognition (SER) growing research topic in which lots of advancements can lead to advancements in various fields like automatic translation systems, machine to human interaction, used in synthesizing speech from the text so on. In contrast, the paper focus to survey and review various speech extraction features, emotional speech databases, classifier algorithms.

There are many voice products has been developed like Amazon Alex, Google Home, Apple Home Pod which functions mainly on voice-based commands. It is evident that Voice will be the better medium for communicating to the machines

## **3. Dataset**

### The RAVDESS is a validated 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, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at 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 were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater 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

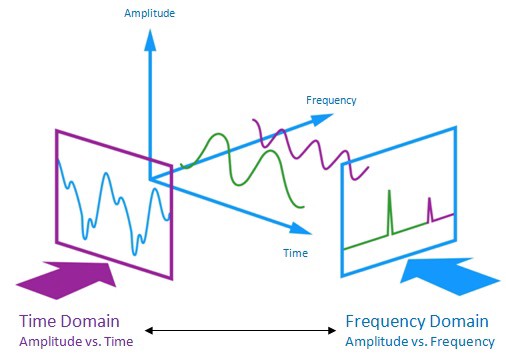
## **4. Data Augmentation**

* Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set.
* To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed.
* The objective is to make our model invariant to those perturbations and enhace its ability to generalize. In order to this to work adding the perturbations must conserve the same label as the original training sample

# **5. Feature Extraction**

## Extraction of features is a very important part of analyzing and finding relations between different things. As we already know that the data provided by audio cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction is used.

## The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency.



Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively lesser data rate for subsequent processing and analysis. This is usually called the front end signal-processing.

**6. Steps involved:**

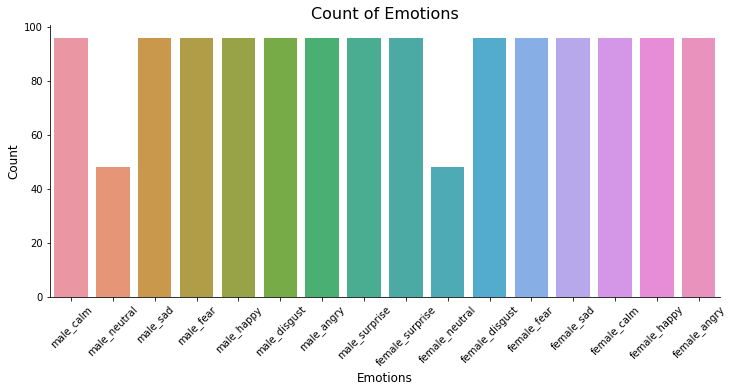
* **Exploratory Data Analysis**

The combined data set from the original 5 sources are thoroughly analyzed with respect to the following aspects

* Emotion distribution by gender
* Variation in energy across emotions
* Variation of relative pace and power across emotions

We checked the distribution of labels for emotions and gender and found that while the data is balanced for six emotions viz**.** neutral, happy, sad, angry, fear and disgust, the number of labels was slightly less for surprise and negligible for boredom.

While the slightly fewer instances of surprise can be overlooked on account of it being a rarer emotion, the imbalance against boredom was rectified later by clubbing sadness and boredom together due to them being similar acoustically. It’s also worth noting that boredom could have been combined with neutral emotion but since both sadness and boredomare negative emotions, it made more sense to combine them.



* **Feature Extraction**
* MFCC (Mel Frequency Cepstral Coefficients)- MFCC is taken on a Mel scale which is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.
* Mel Spectrogram- A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.
* **Fitting different models**

For modelling we tried various classification algorithms like:

1. **KNN(K-nearest Neighbour)**
2. **MLP Classifier**
3. **CNN**

**7.1. Algorithms:**

1. **KNN(K-nearest Neighbour):**

KNN belongs to the supervised learning domain and is majorly used in pattern recognition, and data mining. The K-nearest neighbor scheme requires a training set and desired classification for each item

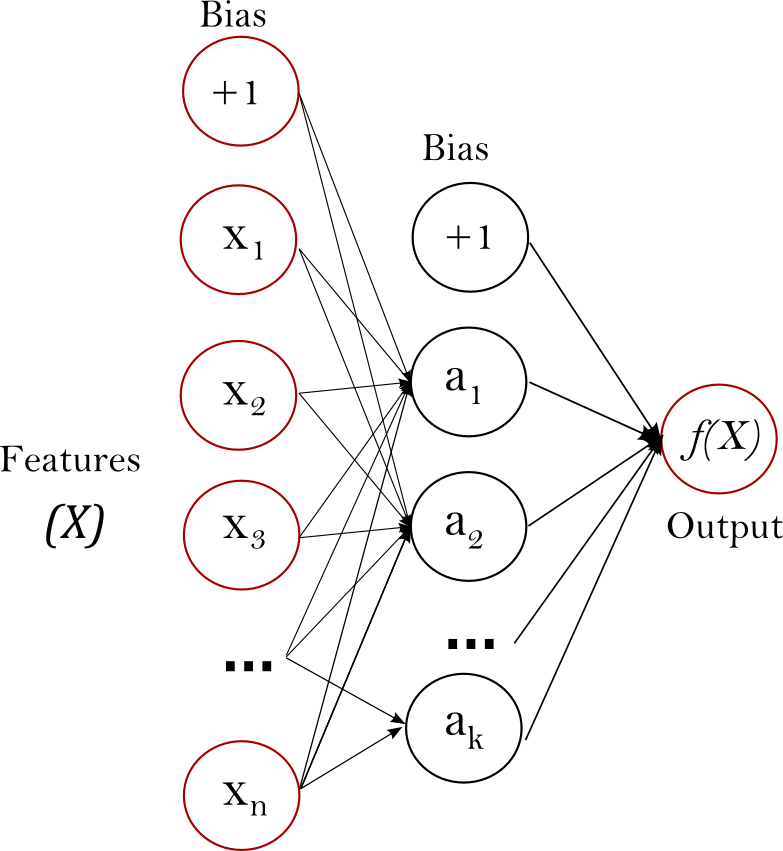
When we need to make a classification for the new data item, its distance to each data in the training set is to be made. Only the k closest entries in the training set are considered. The new data item is then placed in the class that holds the most number of items for this set of k closest data items.



.

1. **MLP Classifier:**

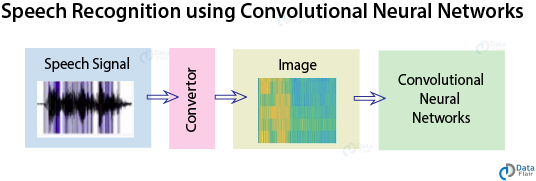
Multilayer perceptron (MLP) classifier is a supervised classification technique that uses backpropagation for training. It is one of the feed-forward artificial neural networks (ANN) classes. It consists of more than one perceptron. It consists of one output layer, one input layer, and in between these 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.

****

:

1. **CONVOLUTION 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.



**3.1 Convolution Layer-**

Convolutional layers in CNNs use convolution instead of multiplication to compute the output. As a result, the neurons in the convolutional layers are not connected to all the neurons in their preceding layers. This architecture is inspired by the fact that neurons of the visual cortex have a local receptive field. That is, the neurons are specialized to respond to the stimuli limited to a specific location and structure. As a result, using convolution introduces sparse connectivity and parameter sharing to CNNs, which decreases the number of parameters in deep neural networks drastically. The figure demonstrates the convolution of a kernel, which is a 2 × 2 matrices, with a one-channel 3 × 3 images. The output is a volume of 2 × 2 × 1. Generally, the size of the output is (nh −f + 1) × (nw −f + 1) ×nf, where nh is the height of the input, nw is the width of the input, and nf is the number of kernels. The depth of the kernel is determined by the depth of the input.

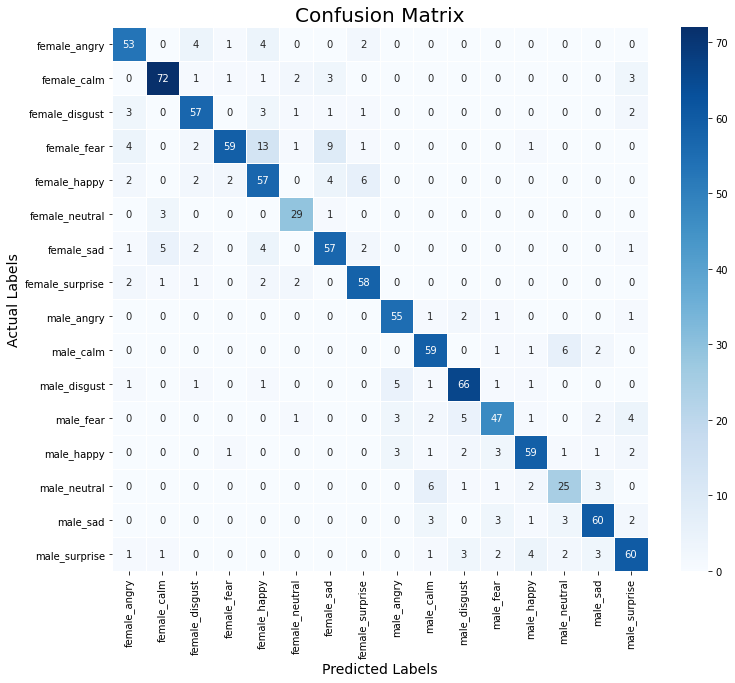
. 

**7.2. Model performance:**

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 at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.



**8. Conclusion:**

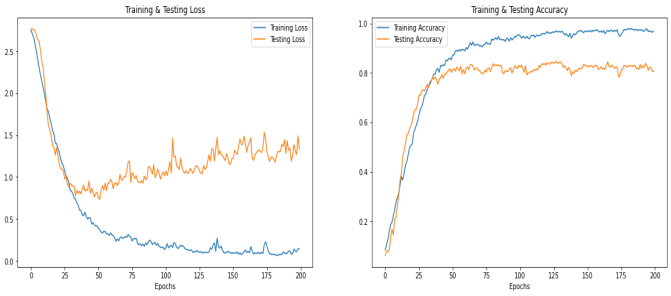
That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA, Data Augmentation, Feature Extraction, and then model building.

In all of these models, our accuracy revolves in the range of 96 to 97%.

So the accuracy of our best model is 95%

Which best accuracy using speech emotion dataset **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. In this project, we use some beautiful techniques like MLP classifier, CNN, KNN every model gave us good accuracy but we take only one model that names model3.h5 file.**



**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/