Project report

# Speech Emotion Recognition

Jayesh B. Ahire Roshan D. Avhad

21111001 21111006

**Project Name:** Speech Emotion Recognition

**Name :** Jayesh Ahire

Roshan Avhad

**Roll No :** 21111001

21111006

**ACKNOWLEDGMENT**

The research work presented in this thesis has been performed at DEPARTMENT OF COMPUTER SCIENCE PUNE. While conducting this research I receive support from many people in one way or another without whose support this thesis would not have been completed in its present form I extend my unlimited thanks to all of them.

My first experience of the project has been successful and I would like to thank to the staff, my friends, and colleagues with gratitude to support me. I wish to acknowledge all of them however I wish to make special mention of the following.

The satisfaction that accompanies the successful completion and it asks would be incomplete without the mention of the people who made it possible and whose constant encouragement and guidance have been a source of inspiration throughout the course of this project.

I take this opportunity to express my sincere thanks to my guide respected Dr.

RITAMBARA KORPAL (Teacher) for their support and encouragement throughout the completion of this project.

Finally, I would like to thank all the teaching and non-teaching faculty members and lab staff of the department of electronics and communication engineering for their encouragement. I also extend our thanks to all those who helped us directly or indirectly in the completion of this project.

**INTRODUCTION**

Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents. While humans can efficiently perform this task as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research.

Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers. pilots, and many other human-machine communication users. Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human-like manner Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities. In some circumstances. humans could be replaced by computer-generated characters having the ability to conduct very natural and convincing conversations by appealing to human emotions. Machines need to understand emotions conveyed by speech. Only with this capability, an entirely meaningful dialogue based on mutual human-machine trust and understanding can be achieved.

Traditionally, machine learning (ML) involves the calculation of feature parameters from the raw data (e.g., speech, images, video, ECG, EEG). The features are used to train a model that learns

to produce the desired output labels. A common issue laced by this approach is the choice of features. In general, it is not known which features can lead to the most efficient clustering of data into different categories (or classes). Some insights can be gained by testing a large number of ditterent features. combining ditterent features into a common feature vector, or applying various feature selection techniques. The quality of the resulting hand-crafted features can have a significant effect on classification performance

An elegant solution bypassing the problem of an optimal feature selection has been given by the advent of deep neural networks (DNN) classifiers. The idea is to use an end-to-end network that there is no need to compute

hand-crafted features, nor to determine which parameters are optimal from the classification perspective. It is all done by the network itself. Namely, the network parameters (i.e., weights and bias values assigned to the network nodes) are optimized during the training procedure to act as features efficiently dividing the data into the desired categories. This otherwise very convenient solution comes at the price of much larger requirements for labeled data samples

compared to conventional classitication methods

**DATABASES USED FOR SER**

Speech emotional databases are used by many researchers in a variety of research activities 24].

The quality of the databases utilized and the performance achieved are the most important factors in the evaluation of emotion recognition. The methods available and objectives in the collection of speech databases vary depending on the motivation for speech systems development.

The categorization of databases can also be described as:

Simulated database: In these databases, the speech data has been recorded by well-trained and experienced performers [20], [21]. Among all databases, this one is considered the simplest way

to obtain the sneech-based dataset ot various emotions It is considered that almost 60% of speech databases are gathered by this technique

Induced database: This is another type of database in which the emotional set is collected by creating an artificial emotional situation [22], [23]. This is done without the knowledge of the performer or speaker. As compared to an actor-based database, this is a more naturalistic database. However, an issue of ethics may apply, because the speaker should know that they have

been recorded for research-based activities.

Natural database: While most rcalistic, these databases are hard to obtain due to the difficulty in recognition [24]. Natural emotional speech databases.

**TRADITIONAL TECHNUIQUES OF SER**

An emotion recognition system based on digitized speech is comprised of three fundamental components signal preprocessing feature extraction and classification 251. Acoustic preprocessing such as denoising as well as segmentation is carried out to determine meaningful units of this signal [26]. feature extraction is utilized to identify the rare event feature available in the signal. Lastly, the mapping of extracted feature vectors to relevant emotion is carried out by classifiers. In this section, a detailed discussion of speech signal processing, feature extraction, and classification is provided [27] Also, the differences between spontaneous and acted speech are discussed due to their relevance to the topic (281. (291. Figure 1 depicts a simplified system utilized for speech-based emotion recognition. In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, an feature selection. The required features are extracted from the pre-processed speech signal and the selection is made from.

**NEED FOR DEEP LEARNING TECHNIQUES FOR SER**

Speech processing usually functions in a straightforward manner on an audio signal 30. It Is considered significant and necessary for various speech-based applications such as SER, speech denoising and music classitication.

With recent advancements, SER has gained significance. However, it still requires accurate methodologies to mimic human-like behavior for interaction with human beings [311. As discussed earlier, an SER system is made up of various components that include feature selection and extraction, feature classification, acoustic modeling, recognition per unit, and most importantly language-based modeling. The traditional SER systems typically incorporate various classification models such as GMMs and HMMs. The GMMs are utilized for illustration of acoustic features of sound units, while, the HMMs are utilized for dealing with temporal variations occurrence in speech signals.

Deep learning methods are comprised of various nonlinear components that perform computation on a parallel basis [32]. However, these methods need to be structured with deeper layers of architecture to overcome the limitations of other techniques. Deep learning techniques such as Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN) and Auto Encoder (A) are considered a few of the fundamental deep learning techniques used for SER, that significantly improves the overall performance of the designed system

**`PROBLEM DEFINITION**

These methods required enormous engineering features and any variation in the features would need re-modeling the overall architecture of the technique. Nevertheless, recent development in deep learning applications and methods for Search Emotion Recognition can be varied also.

There are numerous literature and studies on the application of these algorithms to understand emotions and state of mind from human speech. Additionally, to deep learning, neural networks, and application of improvements of long short-term memory (LSTM) networks, generative adversarial models, and lots more, a wave in research on speech emotion recognition and its application now emerges. It is essential to understand its application and its role in emotion. For this reason, the objective of the current paper is to understand deep learning techniques for speech emotion recognition, from databases to models. After applying the deep learning and feature extraction methods further, it is very difficult to obtain high accuracy in the model because of the similarities between the different emotions like happy and surprising emotions have the same kind of frequency and tone. The length of the voice is also a problem because we all know that the human emotions do not remain the same throughout the sentence it keeps on changing so the system has to identify the parts of the data to understand the full emotion of the voice.

**OBJECTIVE**

This article looked at how you can use speech data in real-world applications, including automatic speech recognition (ASR) and speech emotion recognition (SER). We explored open-source Python packages to help start ASR and suggested project ideas. We also took a deeper dive into building a robust SER model using the TESS (TORONTO EMOTION SPEECHSET) dataset to train an LSTM model. This hands-on experience will equip you to start building projects and master the concepts of SER.

Scientists apply various audio processing techniques to capture this hidden layer of information that can amplify and extract tonal and acoustic features from speech Converting audio signals into numeric or vector format is not as straightforward as images. The transformation method will determine how much pivotal information is retained when we abandon the "audio" format.

If a particular data transformation cannot capture the softness and calmness, it would be challenging for the models to learn the emotion and classify the sample

Some methods to transform audio data into numeric include Mel Spectrograms that visualize audio signals based on their frequency components which can be plotted as an audio wave and fed to train a CNN as an image classifier. We can capture this using Mel-frequency cepstral coefficients (MCCs). Each of these data formats has its benefits and disadvantages based on the application.

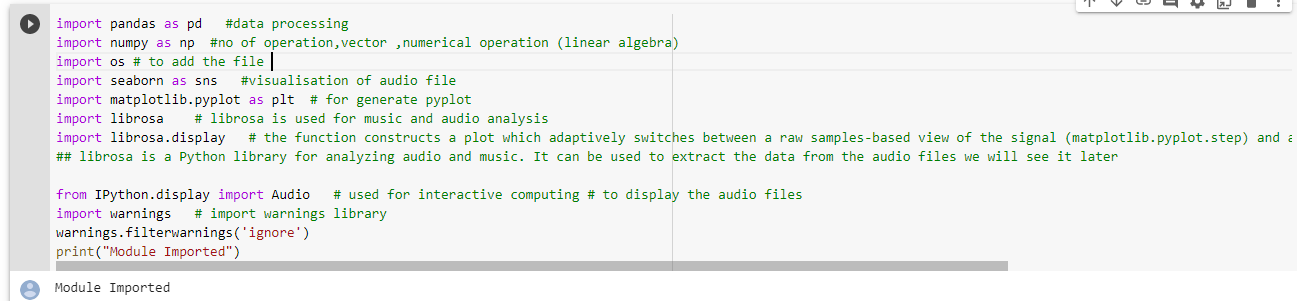
We will try to obtain the data from the MFCC and plot the data in a suitable array form that is used by the model for example we are using here the LSTM model of feature recognition we will use the numeric values given by the MFCC as input to the LSTM model and will try to recognize the emotion.

***Design Flow/Process :***

**LOADING THE DATASET:**

There are 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.

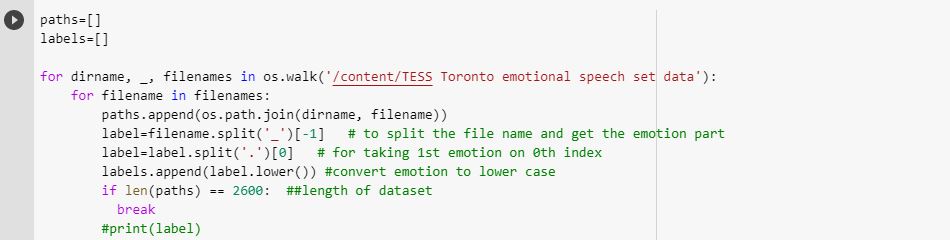
The dataset is organized such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is in WAV format.



First of all we have imported some libraries:

The first five i.e pandas, numpy, os, seaborn, and matplotlib are for visualization purpose. Now to import the audio files we have used the Librosa and Librosa.displav. if you want to play the audio means we have to import the python.display, import audio. To ignore the warnings we are using warnings. filterwarnings('ignore")

Now loading the dataset:



So all the paths are displayed using the syntax, for dirname. . filenames in os.walk('/kaggle/input')

So all the paths are displayed using the syntax, for dirname, , filenames in os.walk(/kaggle/input'):

# for filename in filenames:

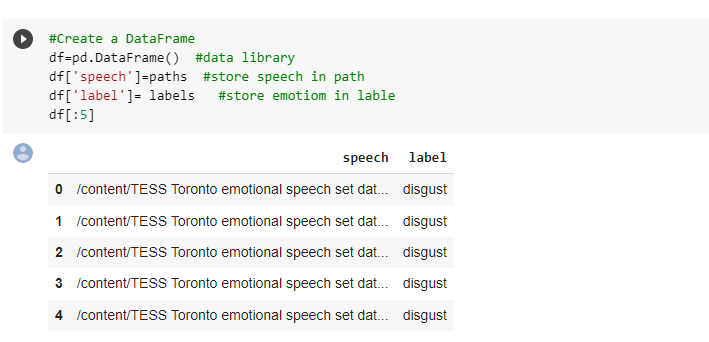
print(os.path.1o1n(dirname, filename))

Using paths and labels we have created two lists and by these lists we will create dataframe by adding everything into the list.

Using the paths we are going to load the audio

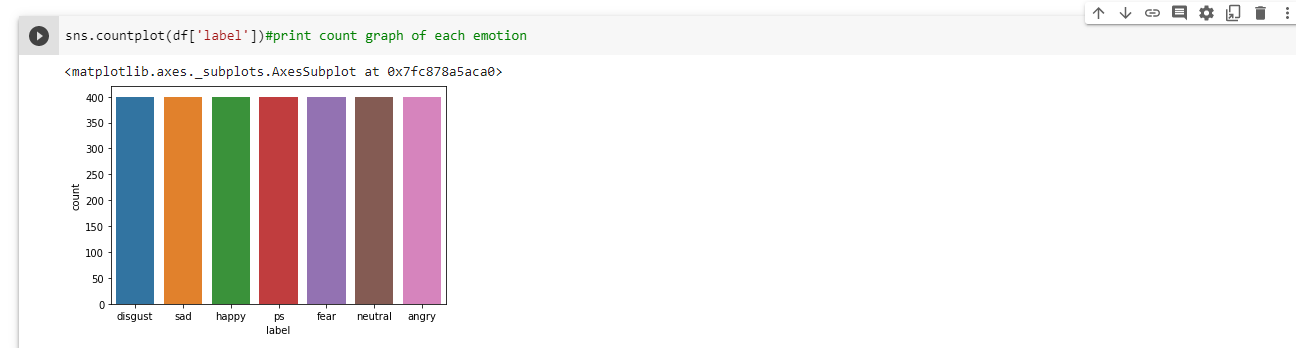
In order to get the label alone from the whole path of different file we are using the label column.

Label.append will move the labels to the label list.



Here we have created a data frame for paths and labels

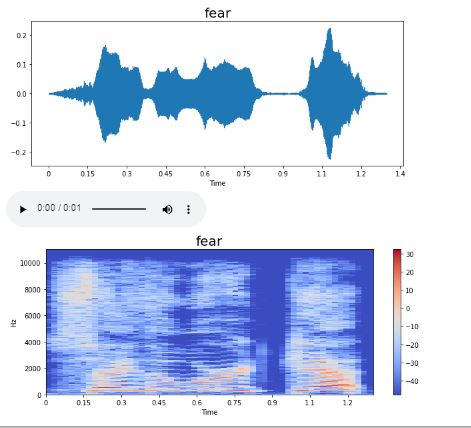
The total no samples use for each emotion ie fear, anger ,disgust, neutral, sad, ps, and happy are 400. Total 2800 samples



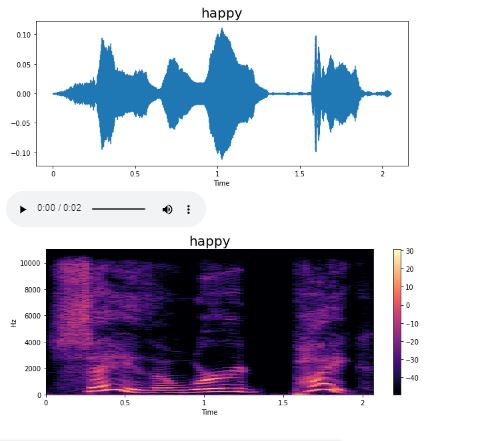
No of sample present of each dataset

Here are some images that are showing different waveforms and spectrograms of each emotion:

**Fear:**

****

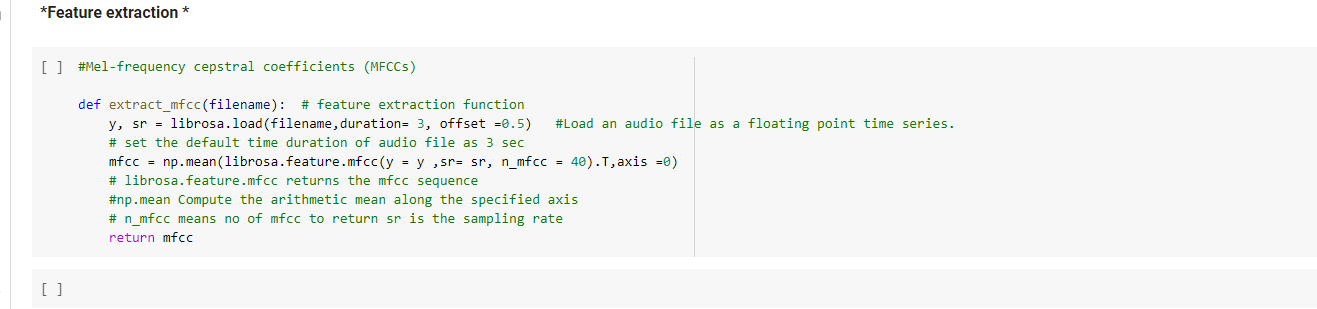
**Happy :**

****

**Extracting features Using MFCC :**

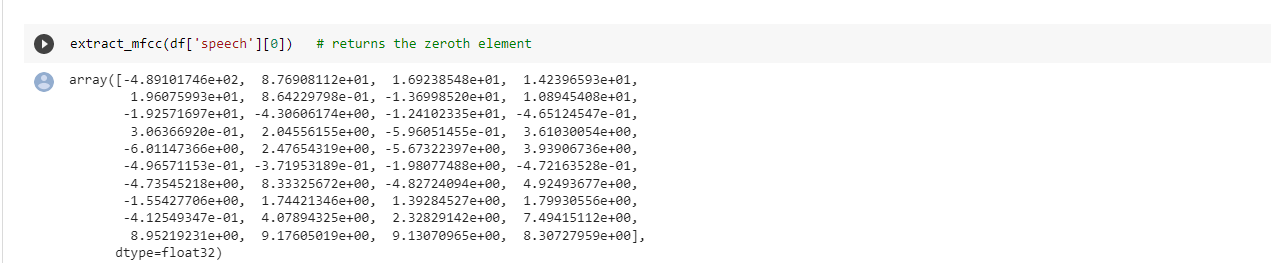
Feature extraction is the process of extracting various acoustic features of the speech which can directly affect the accuracy of the classification results, acoustic features include physical aspects of spoken language that can be recorded and analvzed these include waveform analvsis. FFT or LPC analvsis, voice onset time . format frequency. measurements and so on For the feature extraction purpose various methods are used such as MFCC, LPC, LPCC, LSF, PLP, DWT. We are using the MFCC method in our model for the process of feature extraction.

So Mel Frequency Cepstral Coefficient ( MCC) is basically one of the methods used to extract the various acoustic features of a speech from a raw data to perform various things on the extracted data which can be further used for the classification of the emotions of a speech



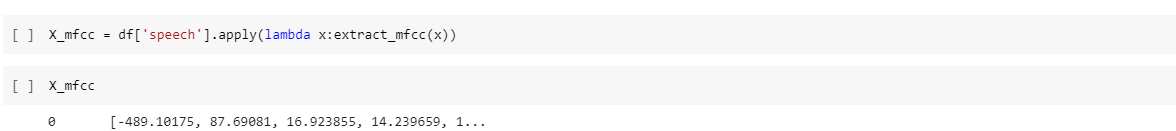
So we have created here a function named extract mfce and in which we have defined the duration of the leature extraction process as the various data have different lengths so we have set the duration to 3 sec and ofiset to 0.5 in the third step we have extracted the features of our data given in the filename.

For example for the first data the mfcc of the first file in the array file is looking something like this:



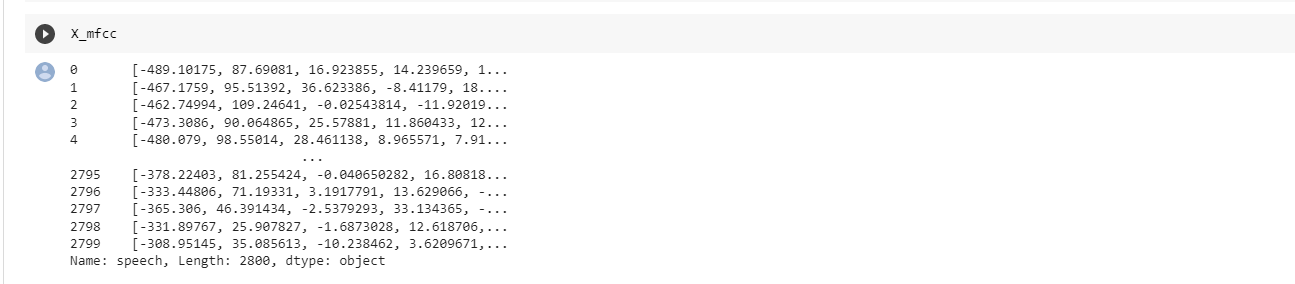
So we have got 40 values here and we will use these values for the input

Extracting features:



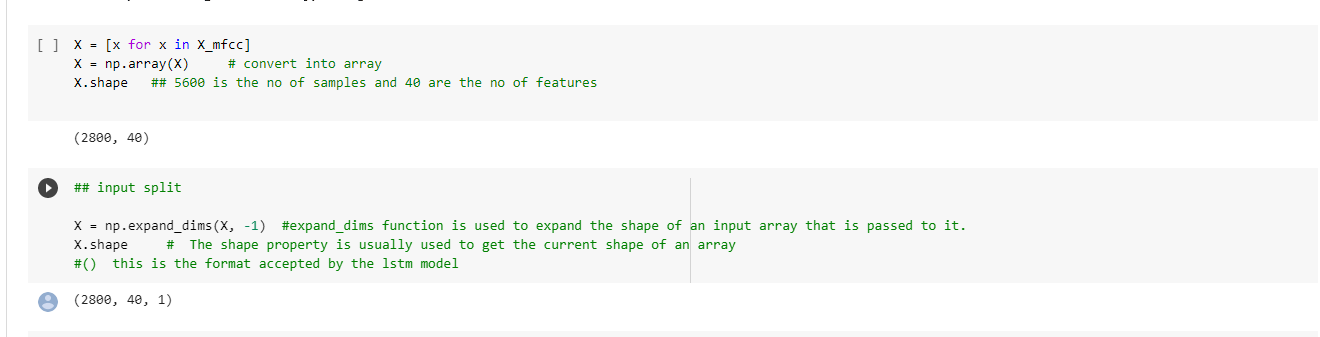
Like in the previous step we have used the mfec on only one sample here in this module we are going to apply the mic function on the whole sample that is our whole 2800 samples, and this is one of the maior steps in this whole process.

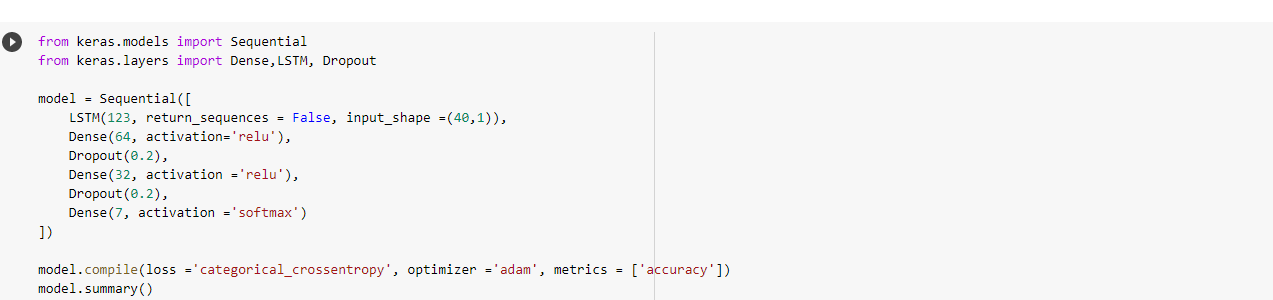
After this step we will get the data in some sequential form like this:



**Creating the LSTM Model**:

To convert the data into the 3d array form as used by the LSDM model we will use the following steps:

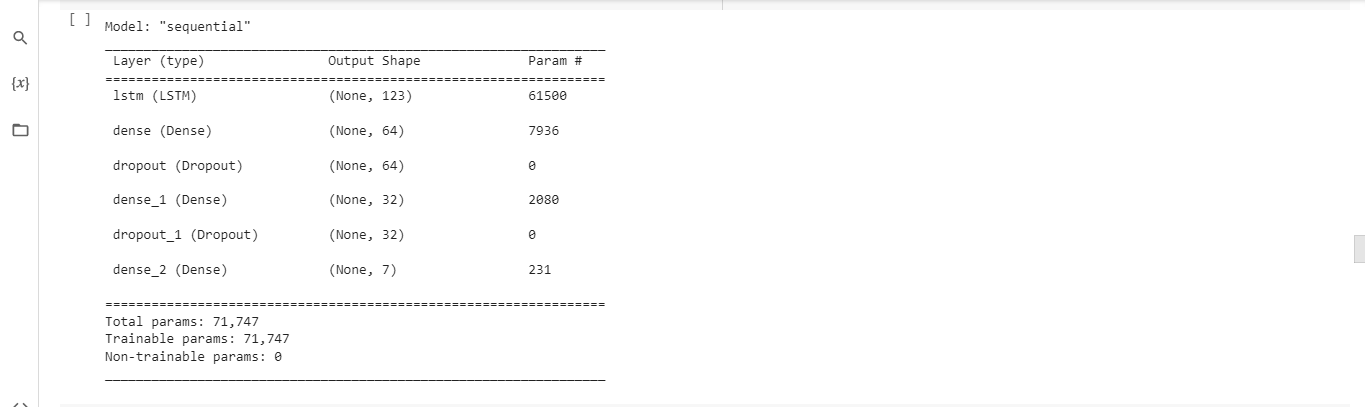




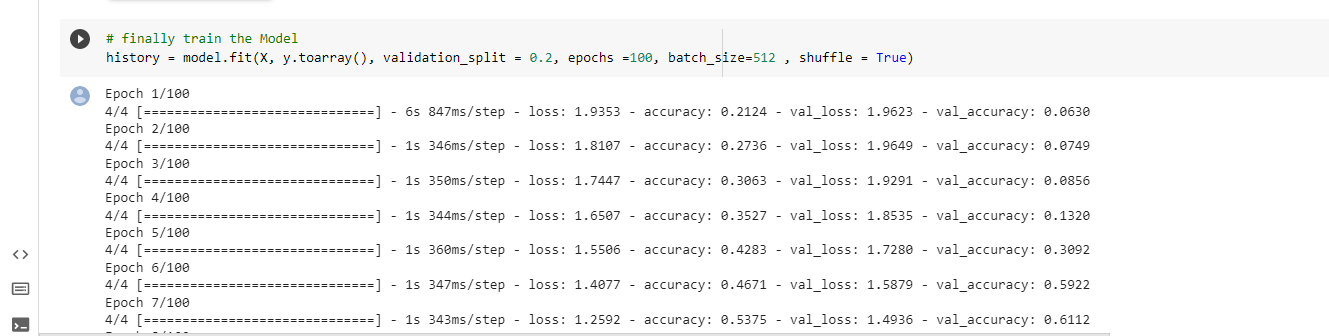
For creating the LSTM model we have to import the sequential, dense, LSTM and the dropout form keras models and keras.lavers respectively. Then specifying the various values required

Then specifying the various values required. In the next step compiling the model for loss and accuracy and then model. summary to view the model it will look something like this.

In the next step compiling the model for loss and accuracy and then model. summary to view the model it will look something like this.



**Training The Model.**

****

Validation split = it will do the splitting for us.

Epochs = the number of complete passes through the complete dataset.

Batch size= the number o training examples utilized in one literaion.

**Emotion Recognized.**

So at last by training and testing the model the emotion of the different samples of the voices are recognized by an overall accuracy o**f 71%.** The detailed result of the experiment are given below in the result analysis.