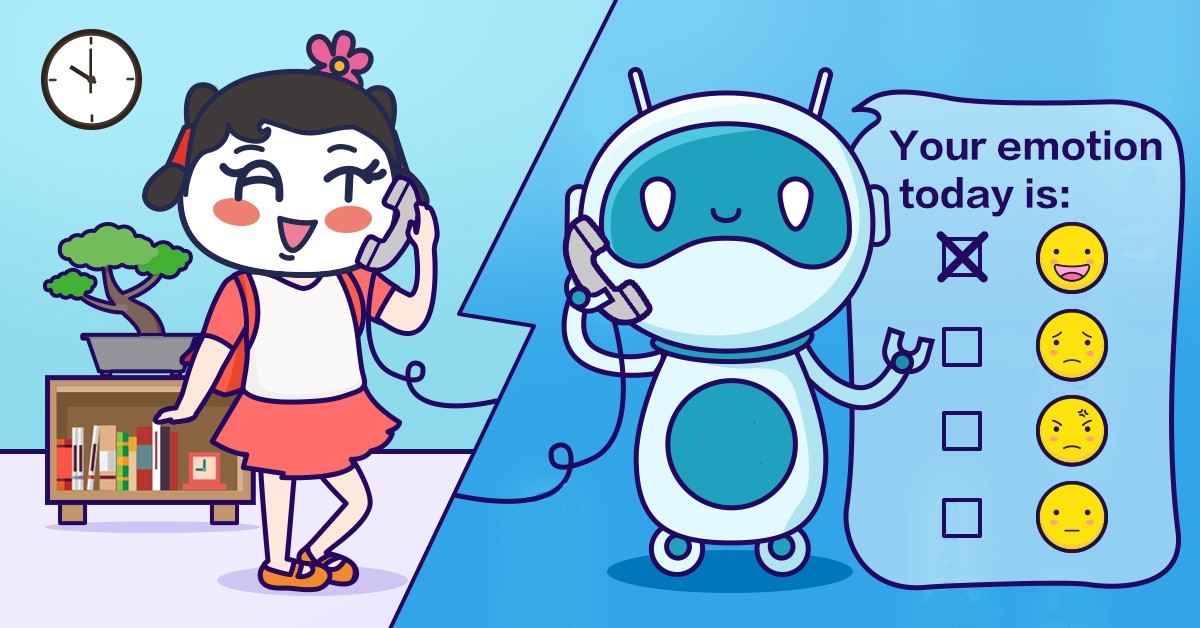


REPORT

Speech Emotion Recognition

Using Machine Learning

(SER)



VARSHITHA K

BACHELOR OF ENGINEERING IN COMPUTER SCIENCE

SIR M VISVESVARAYA INSTITUTE OF TECHNOLOGY

03-11-2020

Supervised by: AISHWARYA SAXSENA

INDEX

|  |  |  |
| --- | --- | --- |
| SL NO. | TOPIC | PAGE NO. |
| 1 | INTRODUCTION | 3 |
| 2 | OBJECTIVE | 3 |
| 3 | BACKGROUND | 3-5 |
| 4 | HARDWARE AND SOFTWARE REQUIREMENTS | 5-6 |
| 5 | CODE AND OUTPUTS | 6-12 |
| 6 | FUTURE SCOPE | 13 |
| 7 | CONCLUSION | 13 |
| 8 | BIBLIOGRAPHY AND REFERENCES | 13 |

INTRODUCTION

* Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions.
* It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others.
* Research has revealed the powerful role that emotion play in shaping human social interaction.
* Emotional displays convey considerable information about the mental state of an individual. This has opened up a new research field called speech emotion recognition, having basic goals to understand and retrieve desired emotions.

OBJECTIVE

* improve man-machine interface.
* Detecting of gender according to the datasets
* monitor the psycho physiological state of a person in lie detectors
* Analyse the representation capability of different feature extraction techniques based on acoustic and non-linear analysis with the aim of recognize emotions from speech.
* Evaluate the effect of different non-controlled acoustic conditions for the recognition of emotions from speech.
* Evaluate the performance of speech enhancement methods to improve the emotion recognition in non-controlled acoustic conditions

BACKGROUND

Librosa

Librosa is a python library for analyzing audio and music. It has a flatter package layout, standardized interfaces and names, backwards compatibility, modular functions, and readable code. It can by installed with pip

Jupyter Lab

Jupyter Lab is an open-source, web-based UI for Project Jupyter and it has all basic functionalities of the Jupyter Notebook, like notebooks, terminals, text editors, file browsers, rich outputs, and more. However, it also provides improved support for third party extensions.

* Categorization of emotions

1. Sensory modalities for emotion expression
2. Facial expressions
3. **Speech**
4. Physiological signals

* We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis.
* An SER system, based on different classifiers and different methods for features extraction, is developed. Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task.
* A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals.

Division of voice features

The important issue in acknowledgment of voice sentiments could be voice elicitation comprising that represents the enthusiastic voice substance, and in the meantime, it does not depend on lexical or speaker substance.

Various signal highlights of voice have examined in the acknowledgment of the voice signal, yet specialists could not detect the optimal features of voice for the events. Here, the work denotes features instances that have a place with each classification.

Features of voice characteristics

It is reliable that the enthusiastic articulation substance could detect emphatically with its characteristics of voice. Exploratory analysis tunes into human subjects show a robust connection among apparent feeling & voice quality. The work presents that several scientists deliberated the sound associated feelings parts that have been tried for connection characterize. The characteristics of voice through overall accounts represented in & out feelings. For instance, an individual feeling comes in the form of activities. This could be in averse to fundamental feelings that affect adversely or emphatically individual activities without controlling clinching. An extensive opportunity of formants parameters adds to an abstract influence of characteristics of the voice.

• The dimension of voice: energy, term & adequacy of the flag have occurred as robust voice level proportions.

• Voice pitch

• The phoneme, feature limits, word & expression

• Temporal frameworks

Data Sets

1. RAVDESS
2. TESS
3. SAVEE
4. CREMA

MFCC (Mel Frequency Cepstral Coefficients)

In the conventional analysis of time signals, any periodic component (for example, echoes) shows up as sharp peaks in the corresponding frequency spectrum (i.e. Fourier spectrum. This is obtained by applying a Fourier transform on the time signal). Any cepstrum feature is obtained by applying Fourier Transform on a spectrogram. The special characteristic of MFCC is that it 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.

MS (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.

HARDWARE REQUIREMENTS

|  |  |
| --- | --- |
| Hardware Tools | Minimum Requirements |
| Processor | I5 or above |
| Hard disk | 10GB |
| RAM | 4GB |
| Monitor | 15” Coloured |
| Mouse | Optical |
| Keyboard | 122 keys |

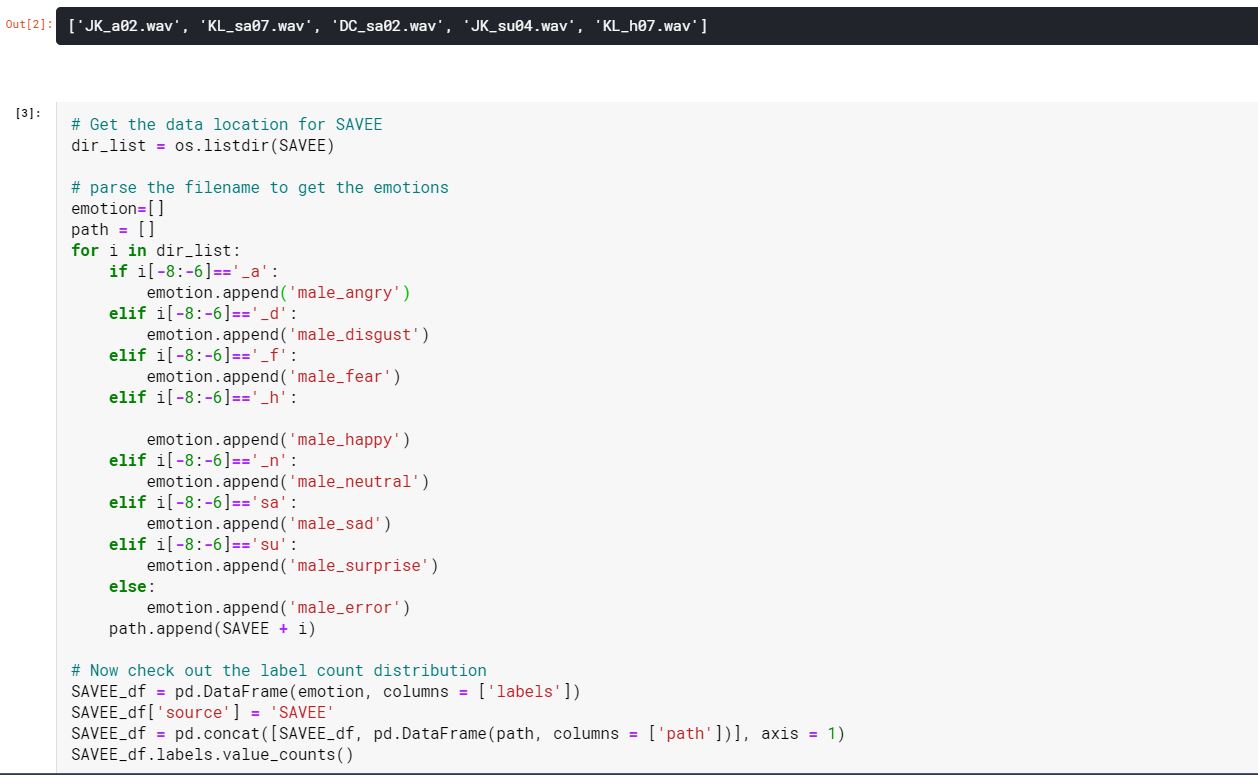
SOFTWARE REQUIREMENTS

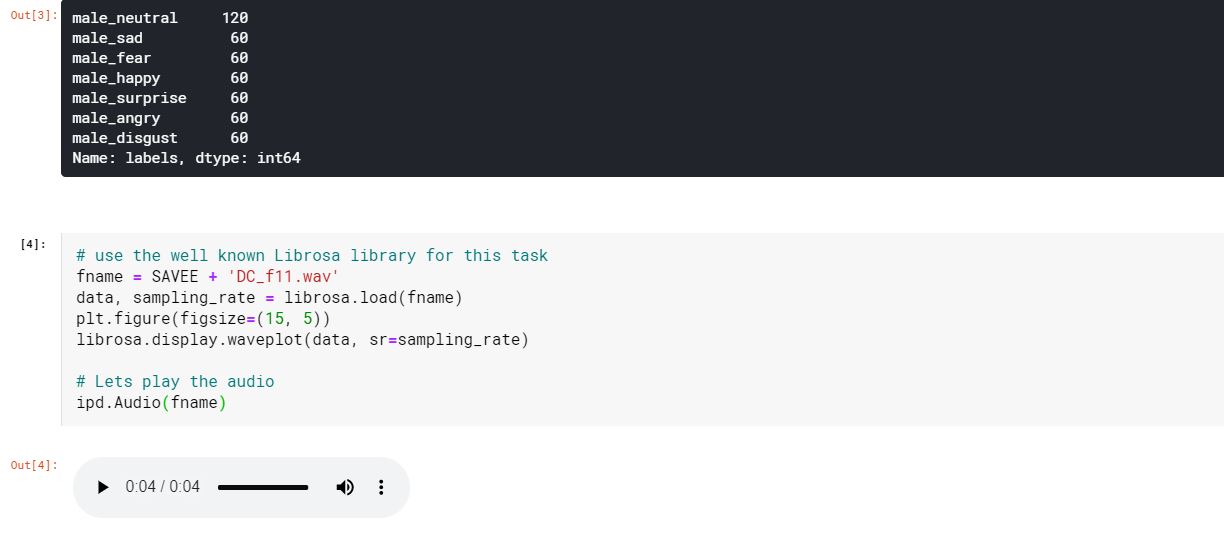
|  |  |
| --- | --- |
| Software Tools | Minimum Requirements |
| Platform | Windows, Linux or MacOS |
| Operating System | Windows, Linux or MacOS |
| Technology | Machine Learning-Python |
| Scripting Language | Python |
| IDE | Jupyter Notebook |

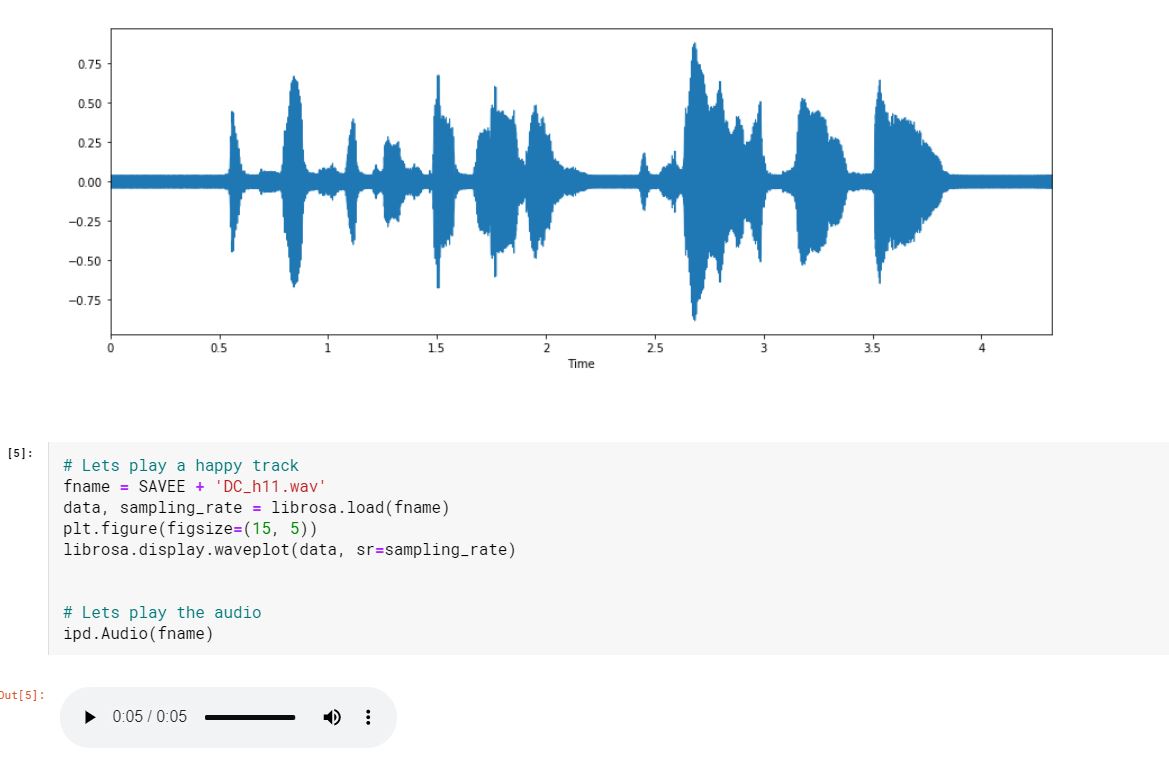
CODING AND OUTPUT SCREEN SHOTS

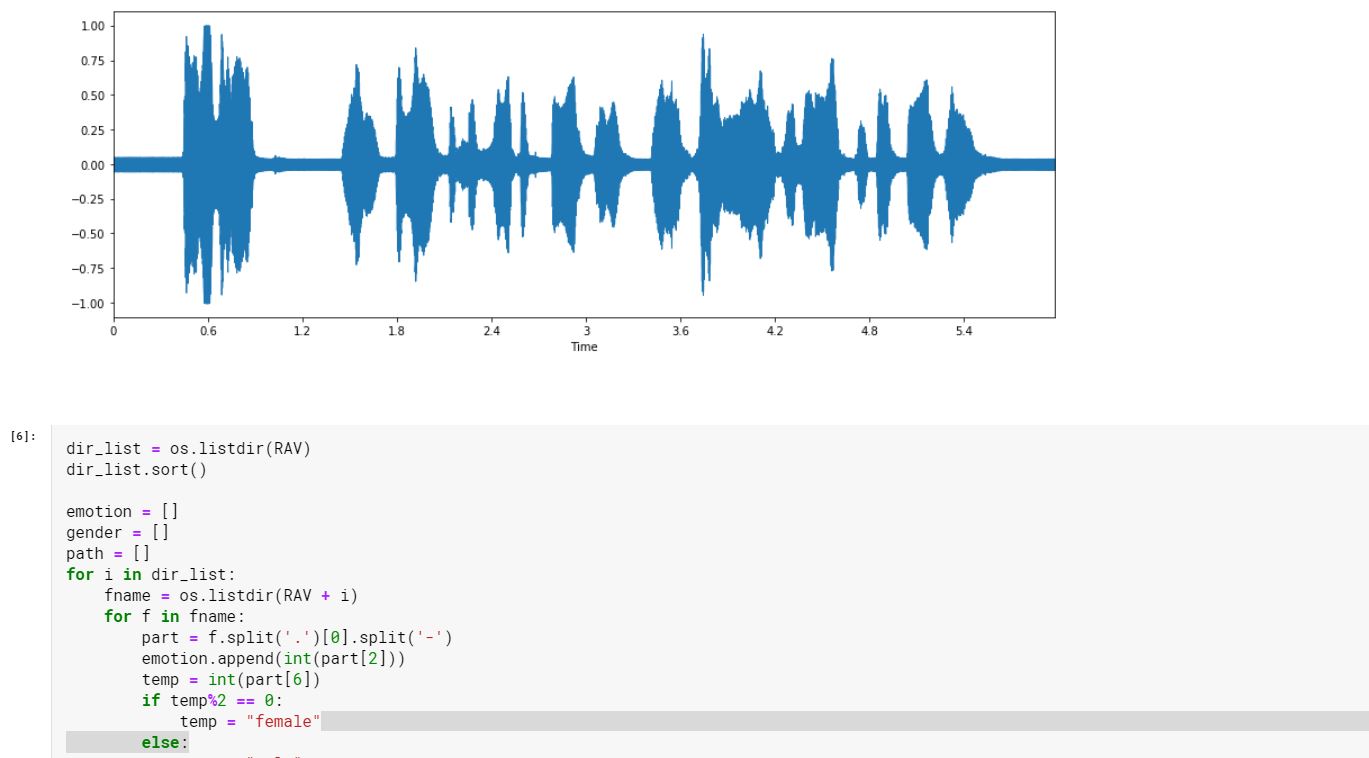
Speech Emotion Recogniton.ipynb

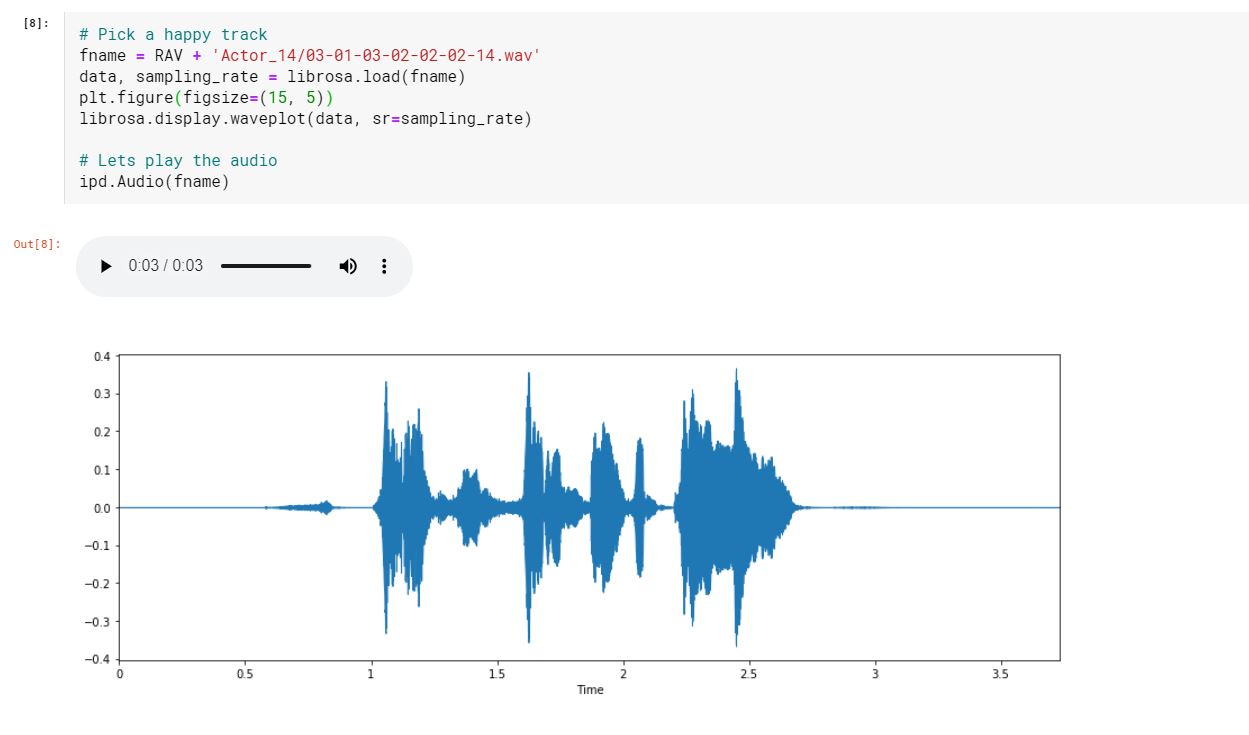
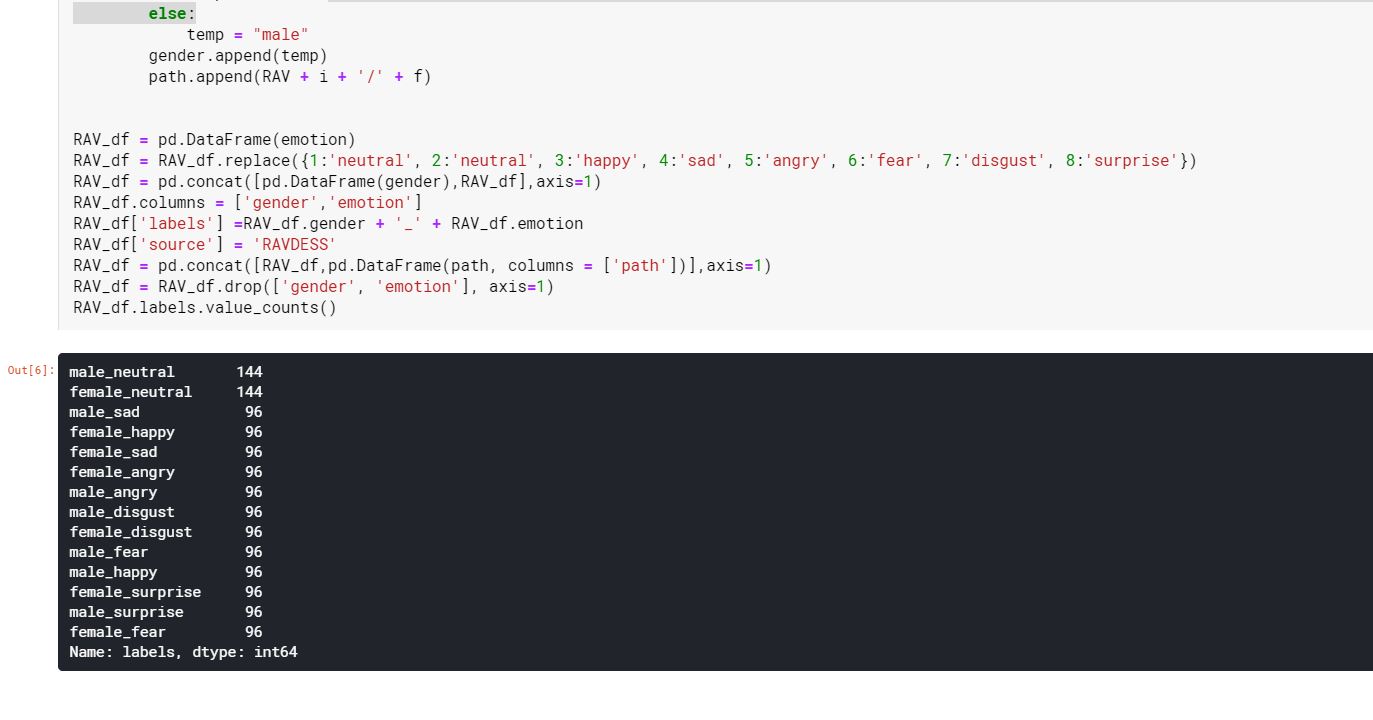


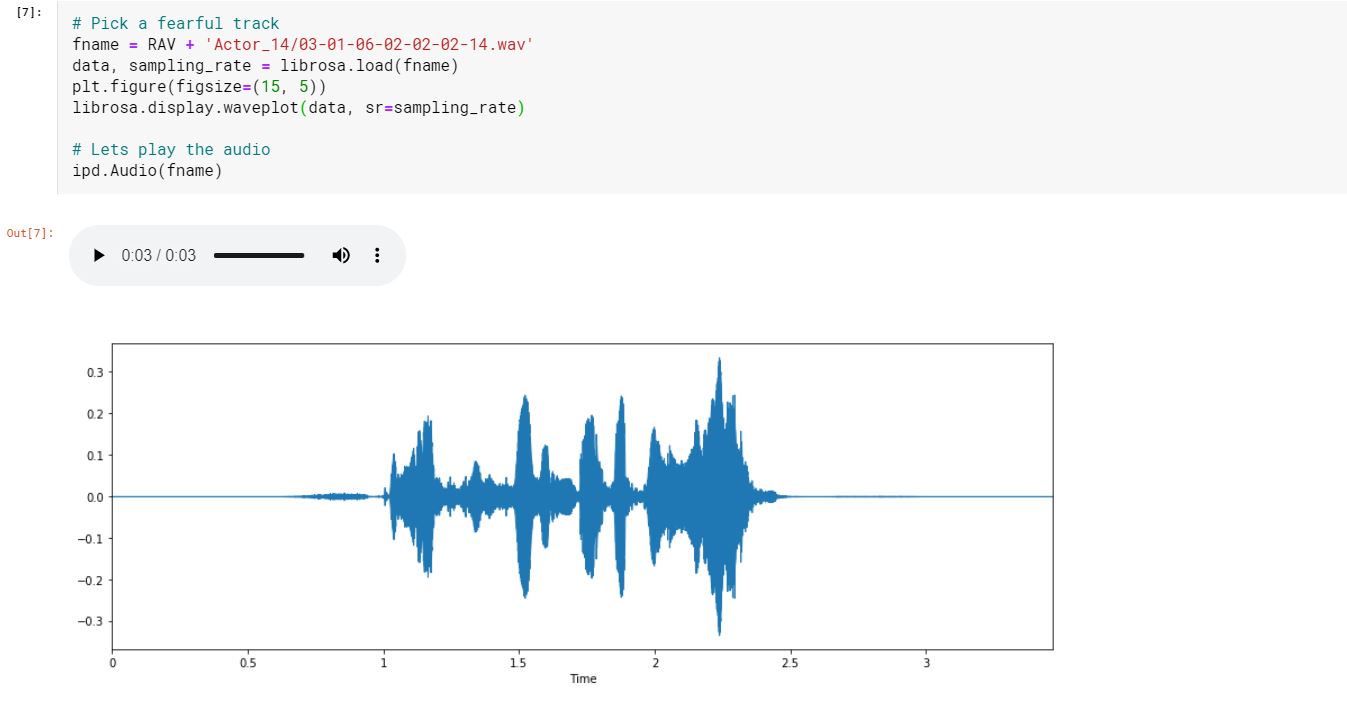


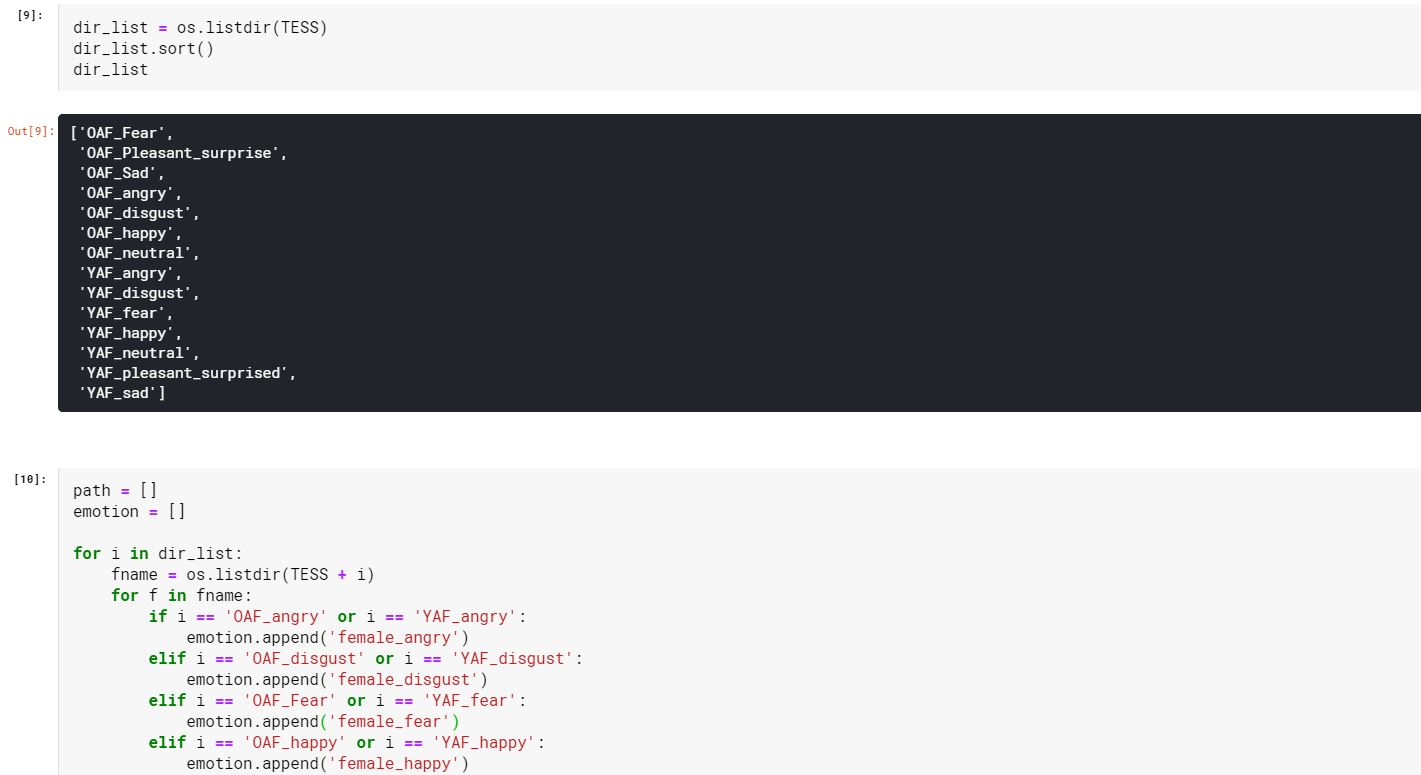


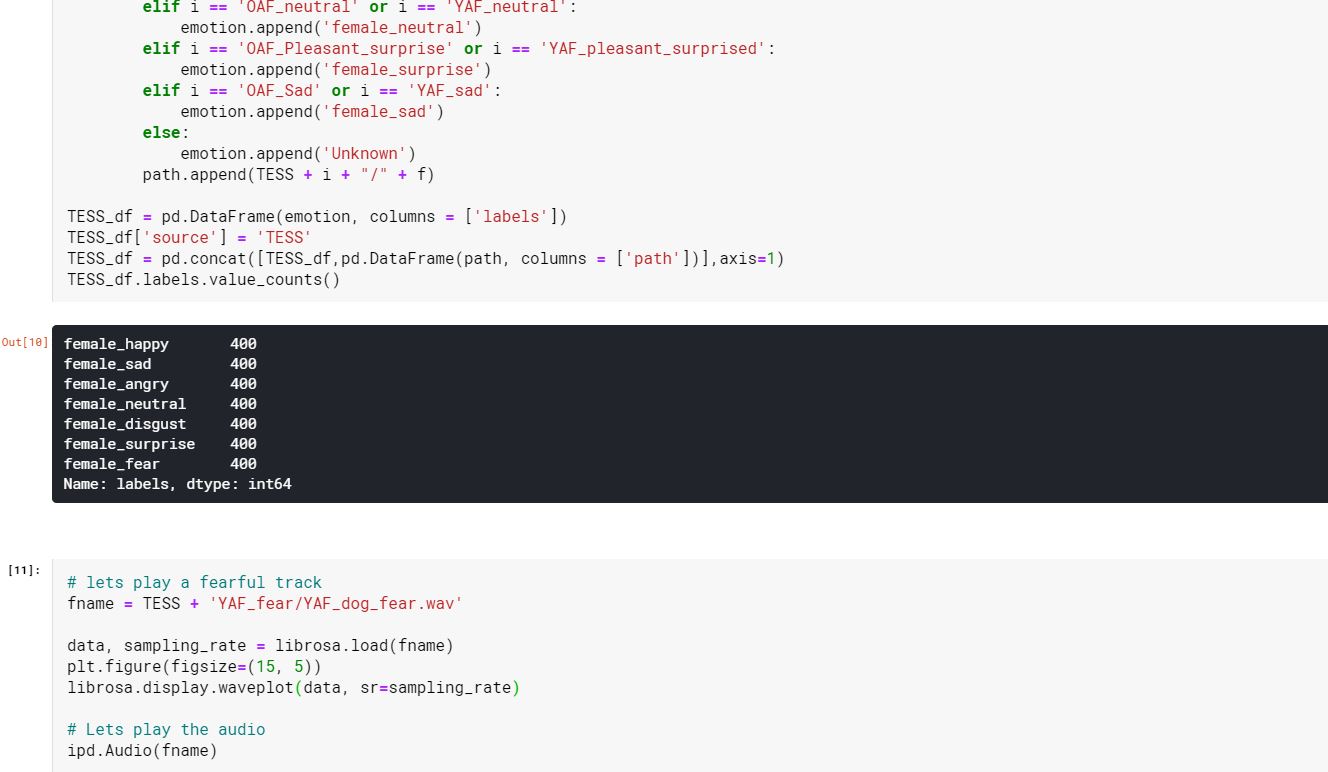


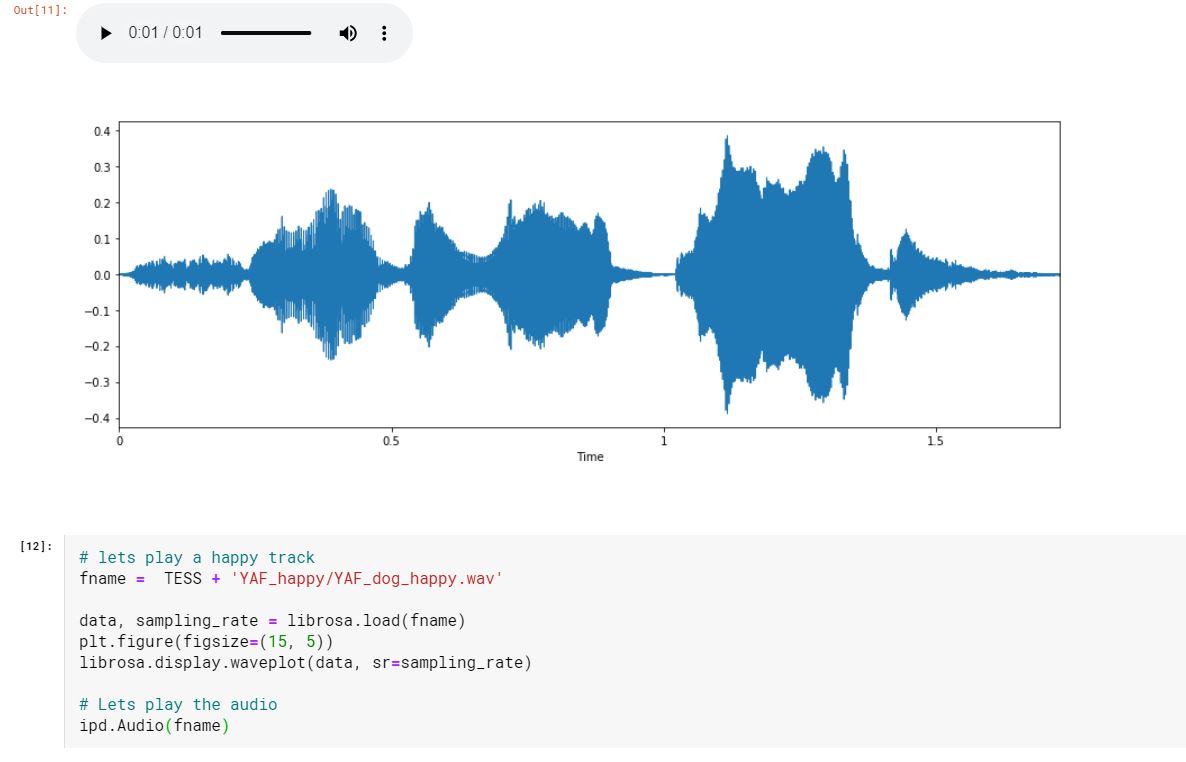


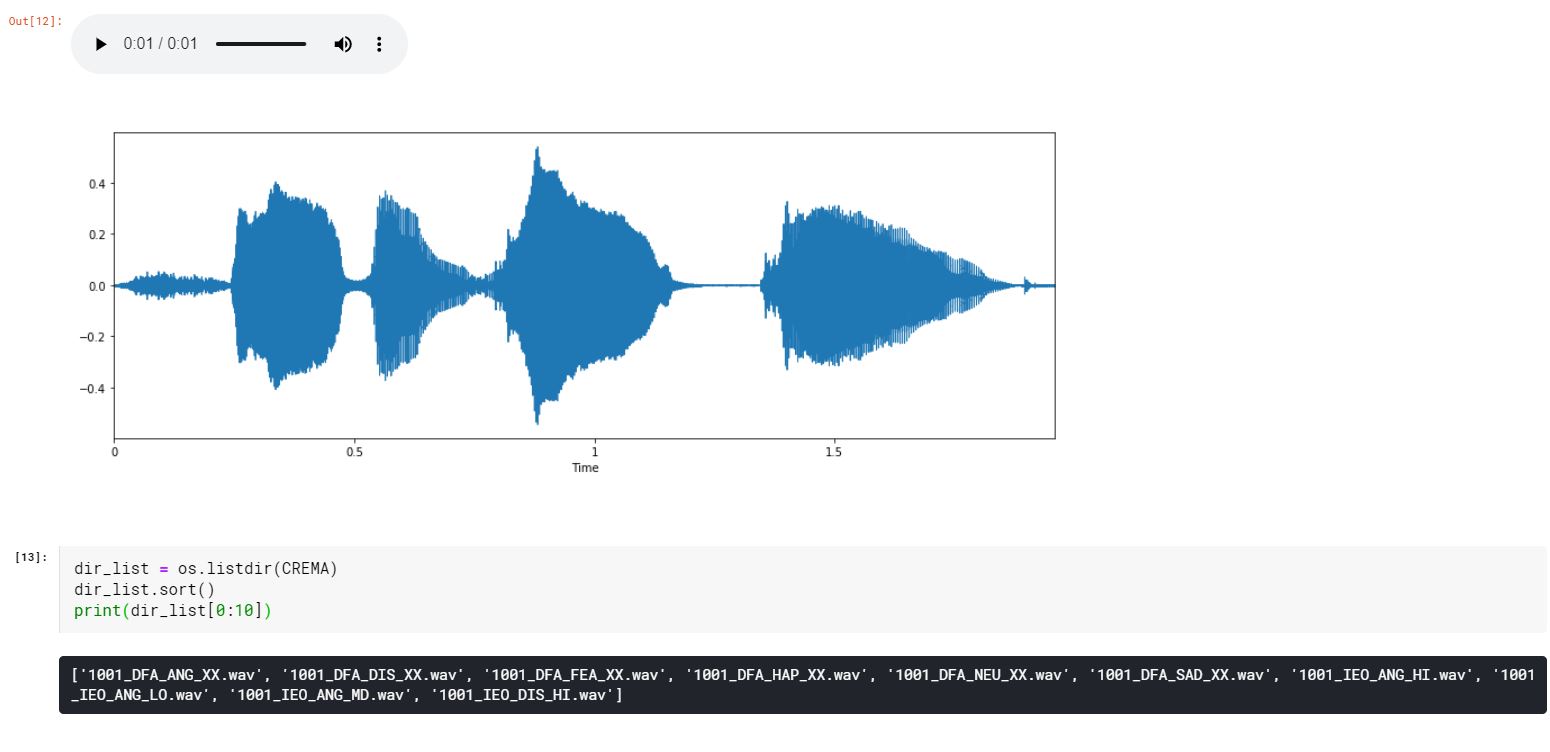


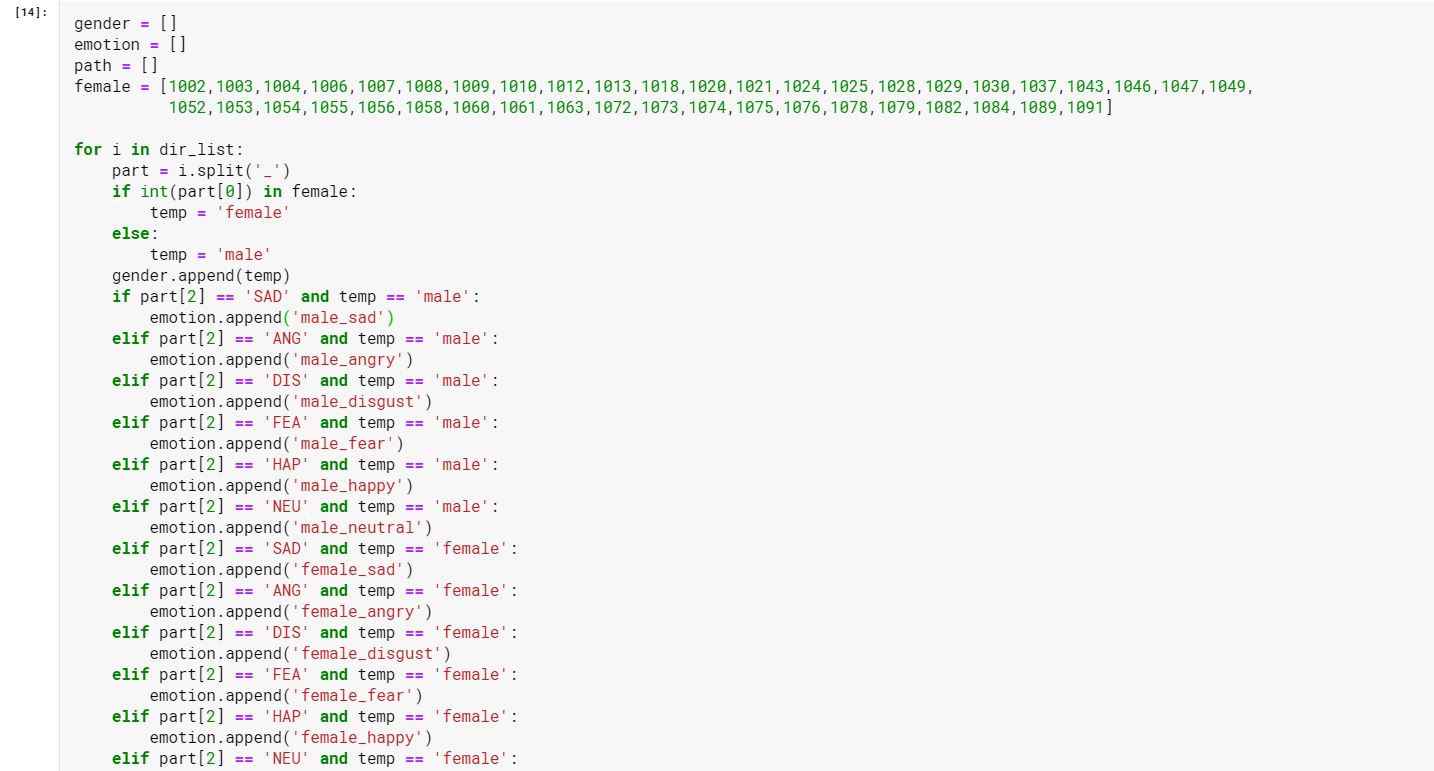




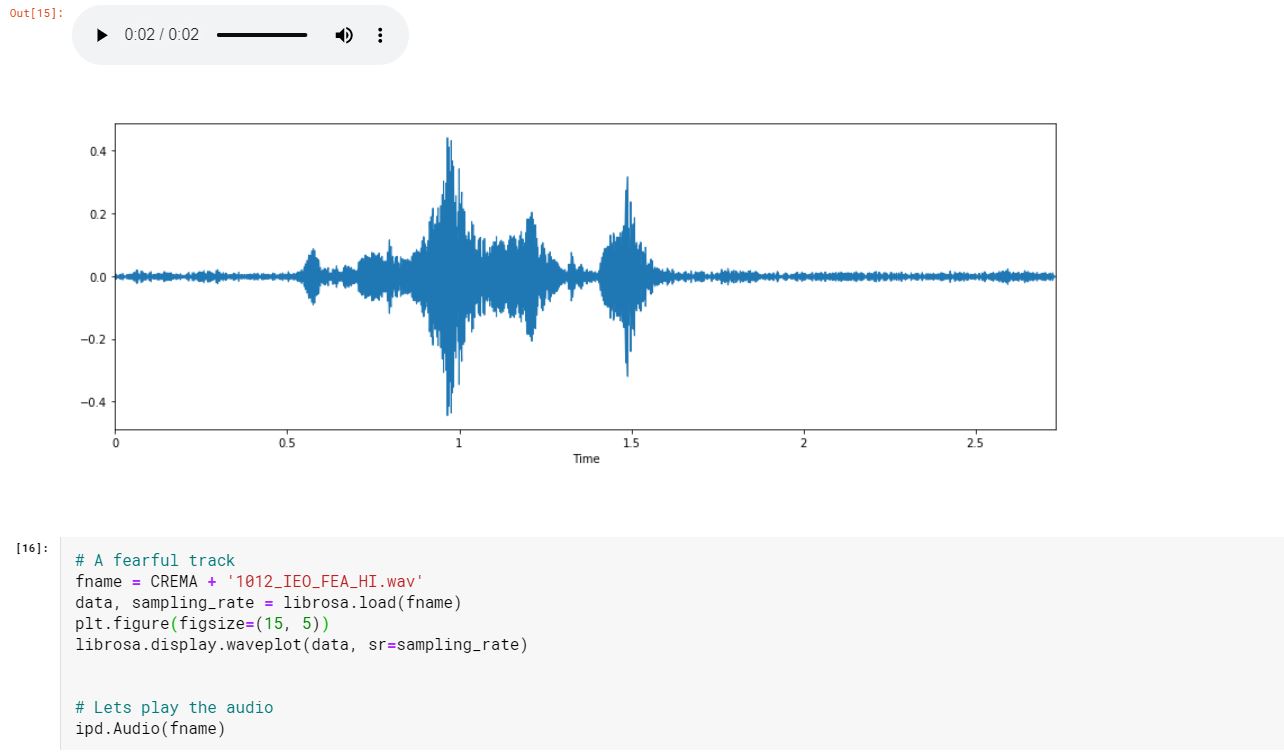


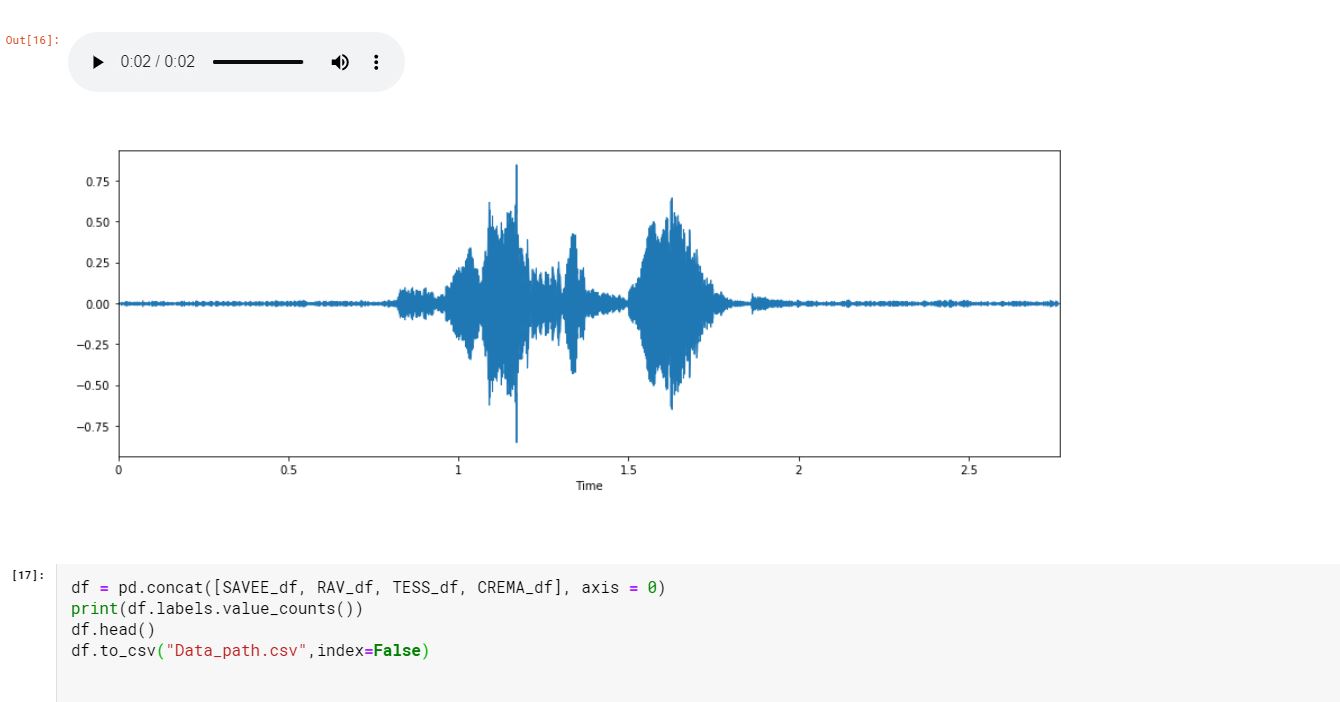






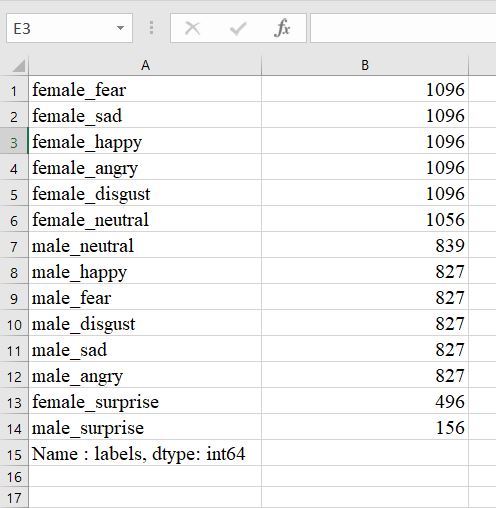








Output:



FUTURE SCOPE

An alternate approach that could be explored for this problem is splitting the classifying task into two distinct problems. A separate model could be used to classify gender and then separate models for each gender to classify emotion could be utilized. This could possibly lead to a performance improvement by segregating the task of emotion classification by gender.

It would be interesting to see how a human classifying the audio would measure up to these models, however, finding someone willing to listen to more than 2,400 audio clips may be a challenge in of itself because a person can only listen to “the children are talking by the door” or “the dogs are sitting by the door” so many times.

CONCLUSION

The use of three features (MFCC’s, Mel Spectrograms and chroma STFT) gave impressive accuracy in most of the models, reiterating the importance of feature selection. As with many data science projects, different features could be used and/or engineered. Tonnetz was originally used in modelling, however it led to decreased performance and was removed. Some other possible features to explore concerning audio would be MFCC Filterbanks or features extracted using the perceptual linear predictive (PLP) technique. These features could affect the performance of models in the emotion classification task.

REFERENCES AND BIBLIOGRAPHY

[1] Datasets (https://www.kaggle.com)

[2] Sklearn libraries (<https://scikit-learn.org/>)

[3] Scipy (<http://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfee.pdf>)  
[4] Speech Recognition Article (<https://towardsdatascience.com/ok-google-how-to-do-speech-recognition-f77b5d7cbe0b>)  
[5] Frequency Domain (<https://en.wikipedia.org/wiki/Frequency_domain>)  
[6] IEEE Paper: Machine Learning Based Emotion Recognition using Speech Signal (<https://www.ijeat.org/wp-content/uploads/papers/v9i1s5/A10681291S52019.pdf>)