Milestone 3 – White Paper

By Achraf Safsafi

Speech Emotion Recognition

Business Problem:

As we live in the age of artificial intelligence, we interact with computers in many aspects of our lives. Our communication with machines is no longer dependent on the keyboard or mouse, but our communication methods have become diverse. It can be via many means as image, touch, or voice. Since voices can be understood based on the words spoken and how they are said as well, in this research, I would focus on speech emotions detection. Recognizing the feelings behind the words would help to improve the audio-based interaction between us and computers. So the research question I will answer is as follows " Which machine learning model is best for speech emotion recognition?".

Background/History:

For machines to communicate well with humans and function effectively, many types of research have been conducted to recognize the human emotions associated with speech. However, recognizing human emotions using technology is a relatively recent area of research where the first seminal work on the topic was published in 1996 by Dellaert et al. However, the field has been steadily developing. Nowadays, we can find intelligent machines in all aspects of our lives like virtual assistants technologies Siri and Alexa.

Implementation Plan:

1) Problem understanding.
2) Data understanding.
    - Describing data.
    - Exploring data.
3) Data preparation:
    - Encoding.
    - Normalization.
    - Feature extraction.
4) Modeling
5) Model evaluation

Data Explanation:

In this project, the dataset is obtained from https://zenodo.org/record/1188976. The following table shows the basic details of the dataset.

Database Name	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
File Name	Audio_Song_Actors_01-24
File Type	zip file
File Size	225.5 MB
Actors	24 professional actors :12 female + 12 male
Accent	Neutral North American accent
Speech Emotions	neutral calm, happy, sad, angry, fearful, surprise, disgust, and surprised
Emotional intensity	normal / strong ❖ There is no strong intensity for the neutral emotion
Statement	"Kids are talking by the door" or “Dogs are sitting by the door"
Speech Files Number	60 trials per actor x 24 actors = 1440

The filename is created using the following constructor:

Modality-Vocal channel-Emotion-Emotional intensity-Statement -Repetition-Actor.wav

Audio File Name Identifiers	**Code**
Modality	03 = audio-only
Vocal channel	01 = speech
Emotion	Emotion : 01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised
Emotional intensity	01 = normal, 02 = strong
Statement	01 = "Kids are talking by the door" 02 = "Dogs are sitting by the door"
Repetition	01 = 1st repetition 02 = 2nd repetition
Actor	01 to 24 odd-numbered actors are male even-numbered actors are female

The dataset is balanced. Except for the neutral emotion, all other emotion labels are equal, as shown in Figure 1

After loading and exploring data, feature extraction was done using Mel-frequency cepstral coefficients (MFCCs), and then the labels were encoded (angry: 0, calm: 1, disgust: 2, fear: 3, happy: 4, neutral: 5, sad: 6, surprise: 7). Next, the extracted feature was normalized using the z-score method. Next, the dataset was split into training and testing sets with a 80:20 split.

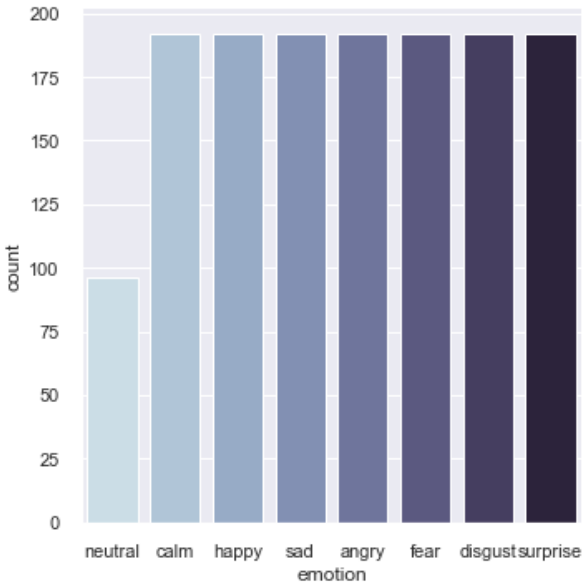


Figure 1:Class distribution

Methods:

After dataset preparation, traditional Machine learning classification algorithms were trained using stratified 10-fold cross-validation. The classifiers include Light Gradient Boosting Machine, MLP Classifier, Extra Trees Classifier, and Random Forest Classifier. Besides that, the deep learning technique of 1D Convolutional Neural Network (1D-CNN) was applied. Next, the performance of the models was compared based on their Accuracy. After, their performance was evaluated to select the final model that shows the best Accuracy.

Analysis:

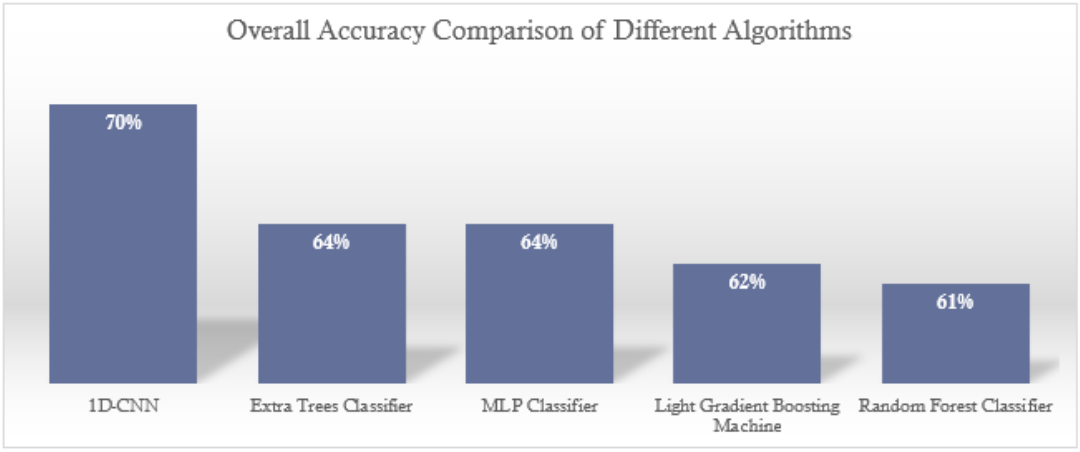


Figure 2:Performance results for all models

Figure 2 summarizes the average Accuracy reached by each individual algorithm. The 1D Convolutional Neural Network algorithm got a maximum Accuracy score of 70 %, followed by the Extra Trees Classifier algorithm and MLP Classifier algorithm with 64 %. The Light Gradient Boosting Machine algorithm achieved an average accuracy of 61 %.

Conclusion:

After different techniques, including four machine learning techniques and one deep learning method, were applied to the audio part of the RAVDESS dataset, the results show that 1D Convolutional Neural Network (1D-CNN) got better performance than the results of the machine learning approach.

limitations:

Emotion is a complex psychological state in which the social and the subjective overlap. Therefore, it is not easy to detect.

Recommendations:

Based on the results, the 1D Convolutional Neural Network model can reach an encouraging result in detecting emotions compared to the classical machine learning methods. So, it is recommended to use deep learning models instead of machine learning models to get better results.

Future Uses/Additional Applications:

Besides recognizing emotions, the model can also be used to detect other multi-class audio classification problems.

Challenges/Issues:

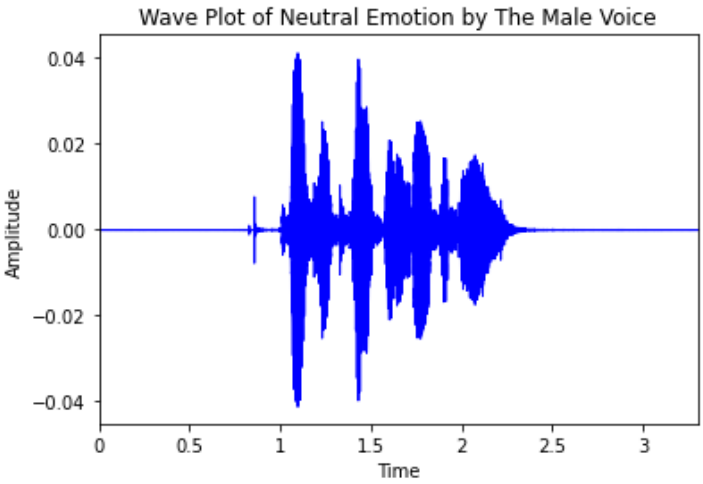
My biggest challenge in this project was preparing the audio dataset. This was the first time I worked on the audio file format.

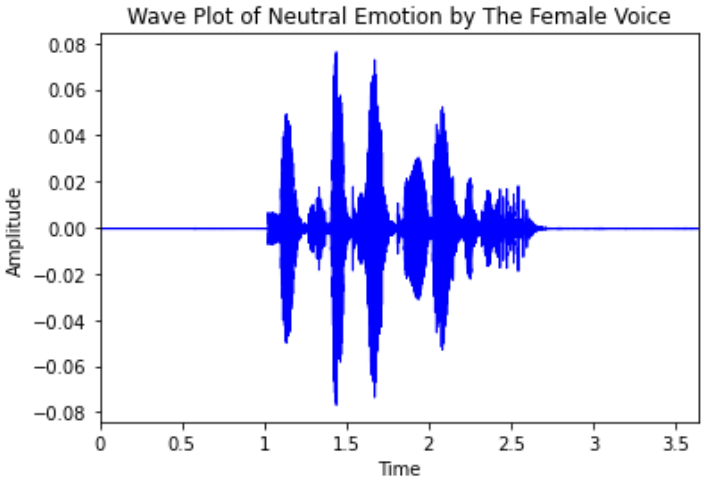
Ethical Considerations:

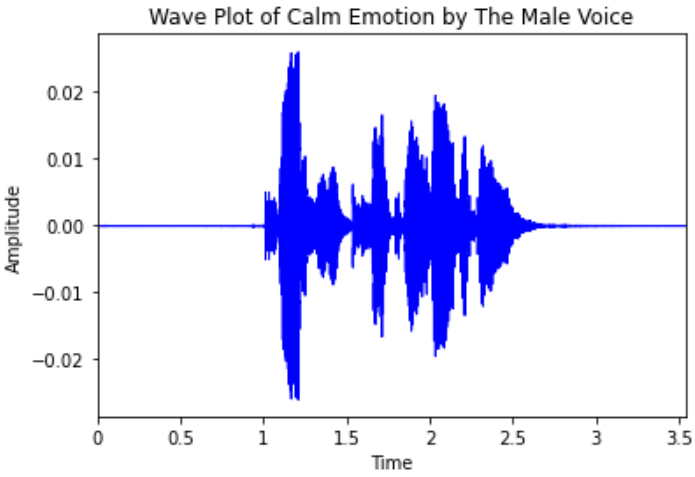
Suppose this method is used to benefit persons, such as getting to know their feelings to improve the services directed to them. In that case, there is nothing wrong with that. But I cannot entirely agree if it is used to judge whether persons are afraid or worried or otherwise, as during a job interview or during security monitoring. These techniques cannot be used against anyone. Human emotions are too complex to be determined carefully by a machine. What about people with anxiety disorders? Would they be more likely to be rejected in a job interview? or would they be investigated more than others just because the machine could reveal their feeling is nervous or fearful?!!

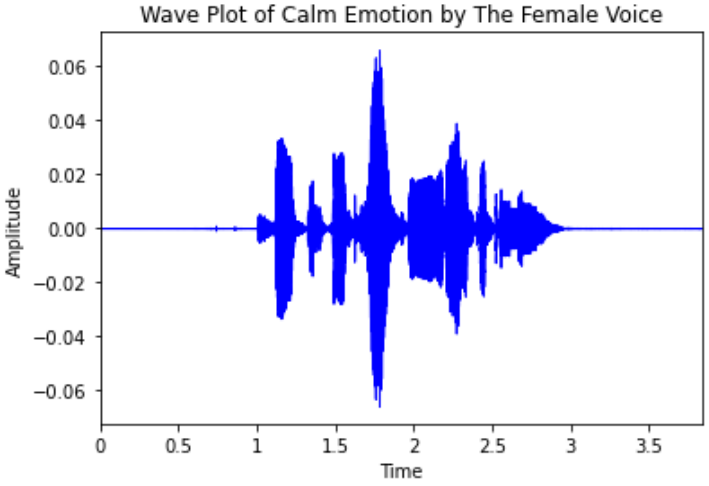
Appendix

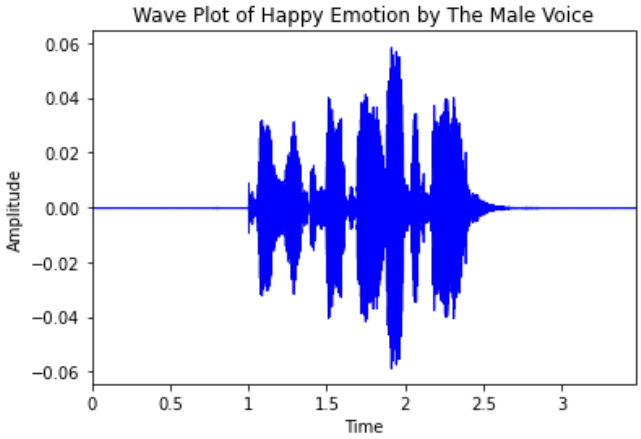
Wave Plots of Different Types of Emotions

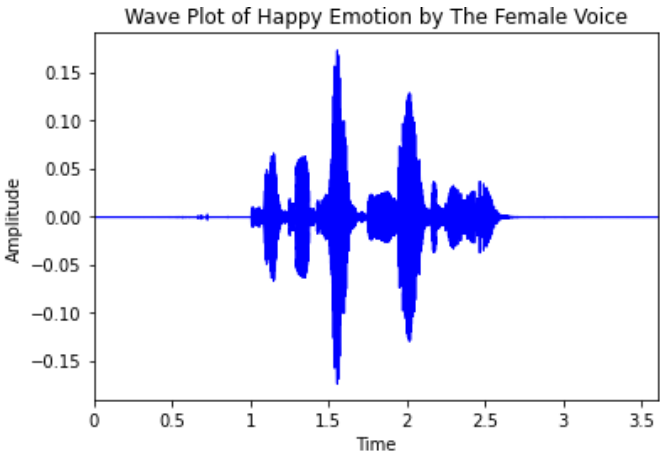


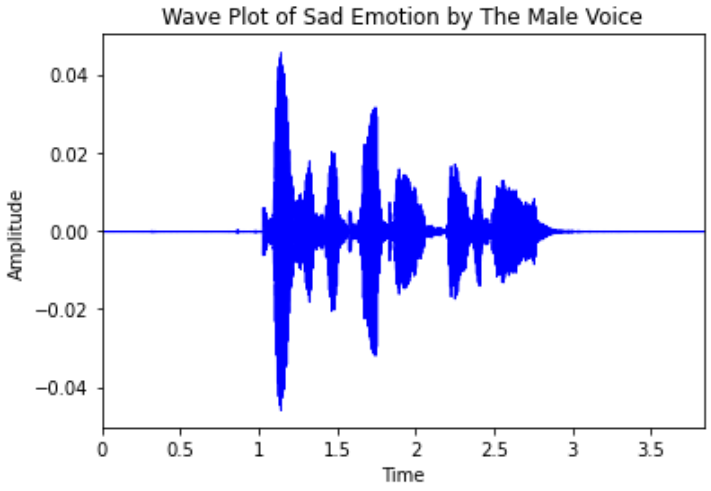


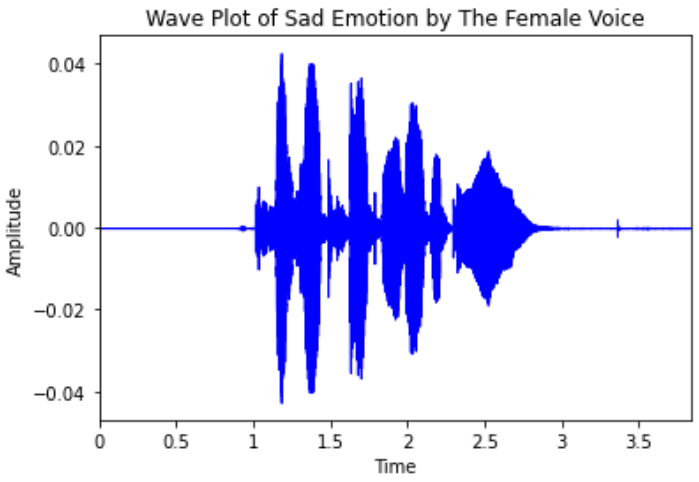


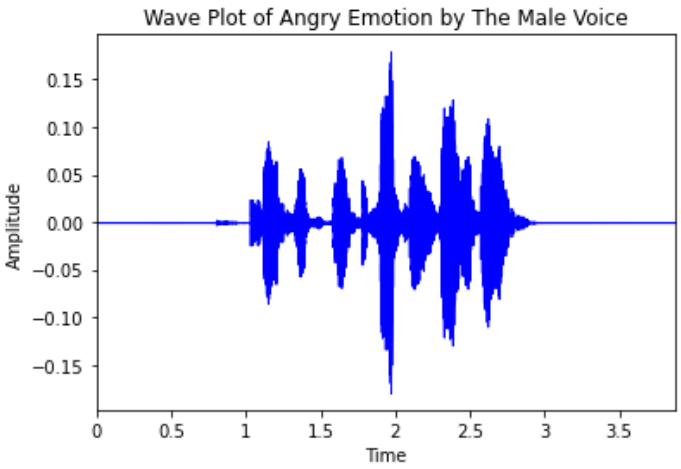


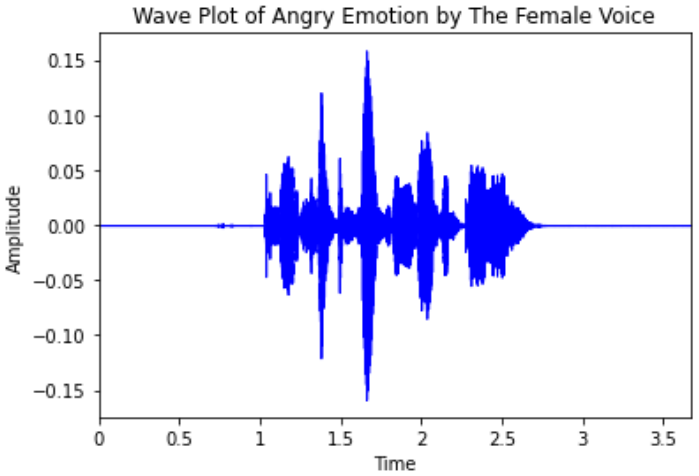


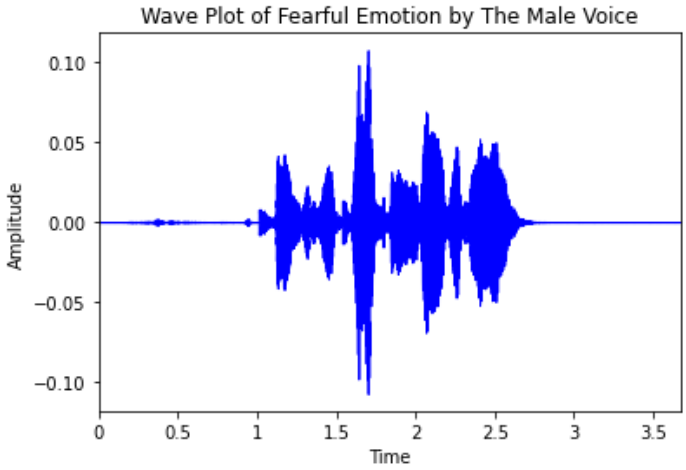


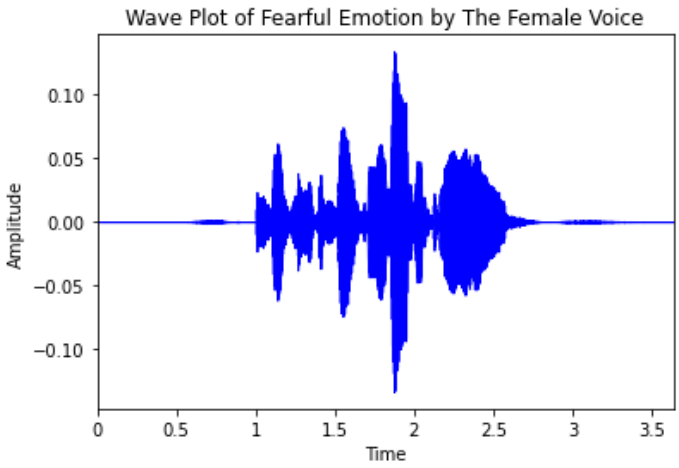

Wave Plot of Sad Emotion by The Male Voice
Amplitude
Time

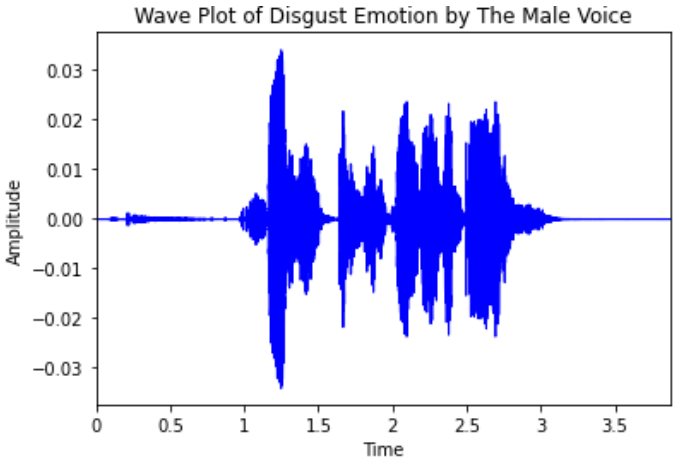

Wave Plot of Sad Emotion by The Female Voice
Amplitude
Time

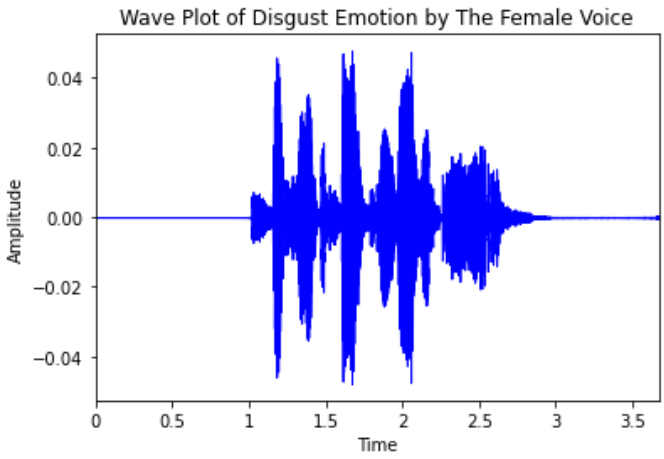

Wave Plot of Angry Emotion by The Male Voice
Amplitude
Time

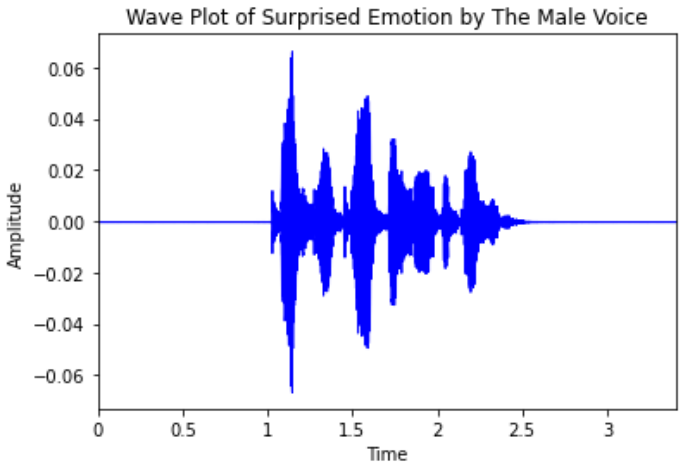

Wave Plot of Angry Emotion by The Female Voice
Amplitude
Time

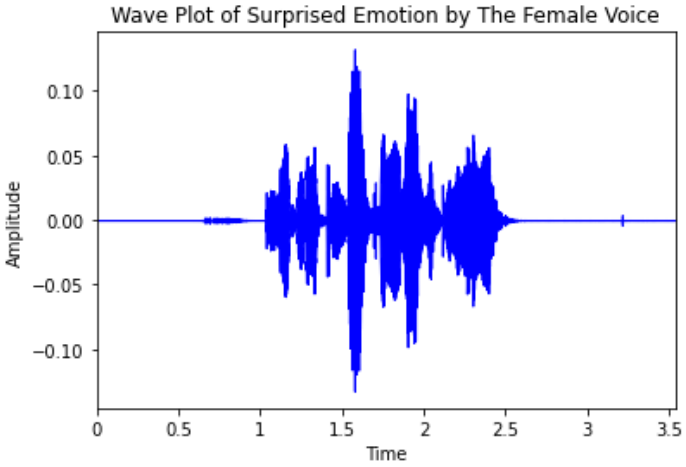

Wave Plot of Fearful Emotion by The Male Voice
Amplitude
Time


Wave Plot of Fearful Emotion by The Female Voice
Amplitude
Time

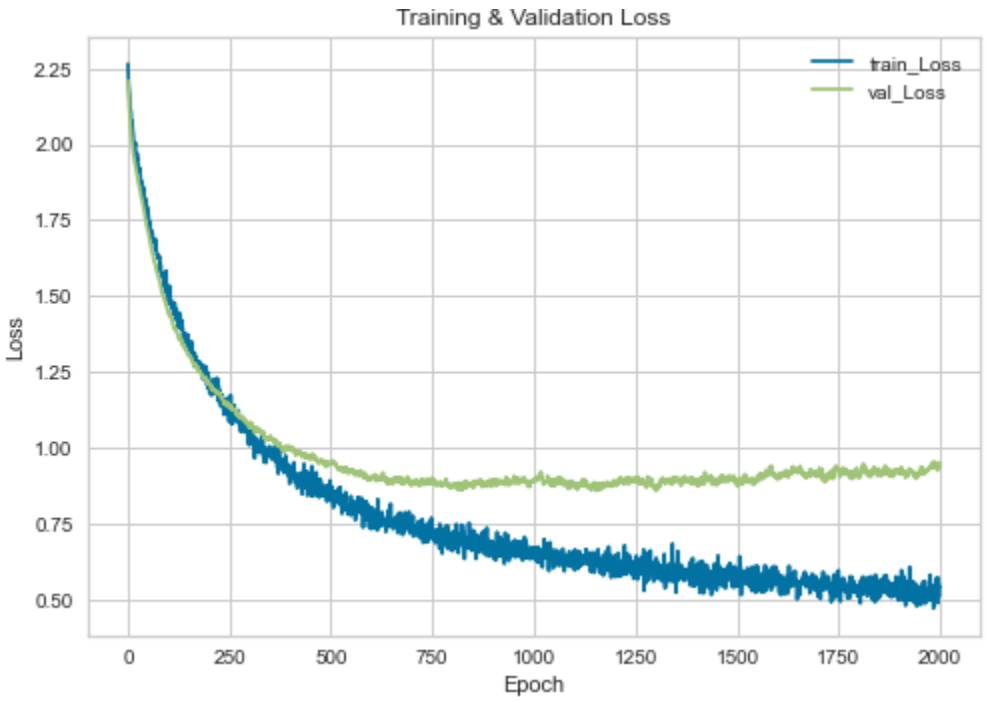

Wave Plot of Disgust Emotion by The Male Voice
Amplitude
Time


Wave Plot of Disgust Emotion by The Female Voice
Amplitude
Time


Wave Plot of Surprised Emotion by The Male Voice
Amplitude
Time


Wave Plot of Surprised Emotion by The Female Voice
Amplitude
Time

Plotting 1D-CNN Results



References:

Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

Emotion Recognition from Human Speech: Emphasizing on Relevant Feature Selection and Majority Voting Technique. (n.d.). Retrieved from https://arxiv.org/ftp/arxiv/papers/1807/1807.03909.pdf

Convolution Neural Network for Speech Emotion Recognition. (2020). Retrieved from https://ijcrt.org/papers/IJCRT2006076.pdf

Dellaert, Frank & Polzin, Thomas & Waibel, Alex. (1996). Recognizing Emotion In Speech. International Conference on Spoken Language Processing, ICSLP, Proceedings. 3.