Speech Understanding (CSL7770)
Minor Examination Report
Shubh Goyal (B21CS073)

QUESTION 1 - Data Collection and Audio Analysis

The task was to collect 10 audio samples on our own voice, extract and analyze different characteristic features, and plot spectrogram for them.

Data Collection

Quotes and monologues from movies and web-series of varying emotions were used as text so as to cover as wide variety of voice variations as possible. A short summary of the tone of the speech samples in the collected dataset is as follows:

- 1, 5, 8 - sad and tired
- 2 - joke
- 3 - anger
- 4 - confident
- 6 - excited and serious announcement
- 7 - commanding
- 9 - announcing and asking to join
- 10 - paused and low tone

Feature Extraction

The features extracted for each of the audio file are - mean rms, maximum rms, mean amplitude, maximum amplitude, mean pitch, maximum pitch, minimum pitch, frequency and spectrogram plots. The plots of waveform, spectrogram, rms, amplitude and pitch envelope are present in the plots directory under respective audio files number's directory. The code for rms and amplitude calculation was written from scratch while `librosa` was used for calculating other features.

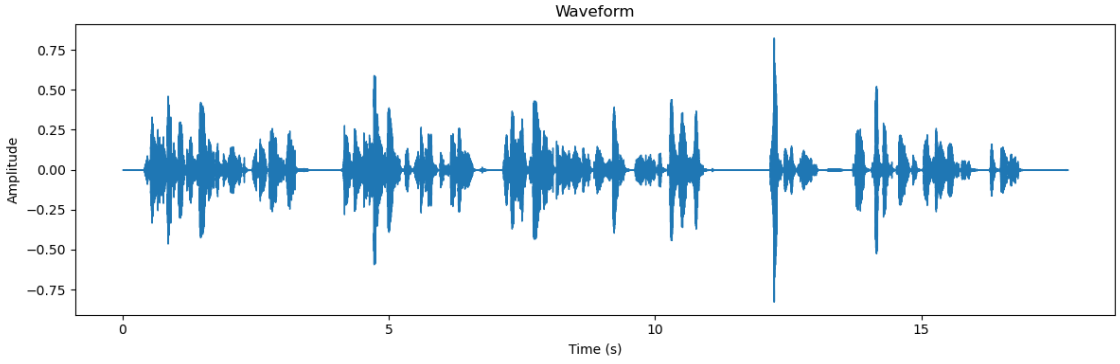

Fig. 1. Waveform - 4

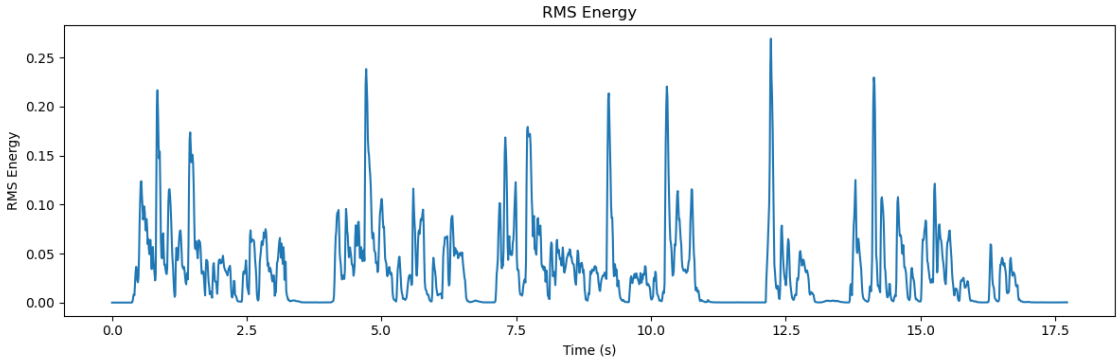

Fig. 2. RMS - 4

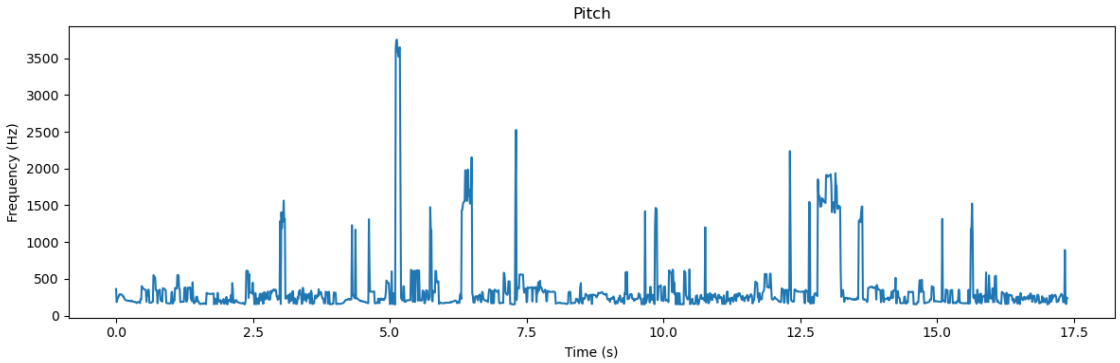

Fig. 3. Pitch - 4

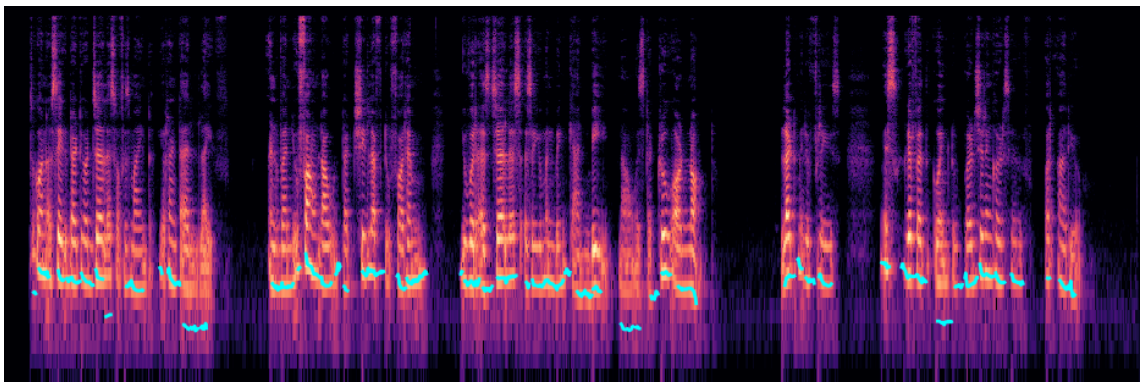

Fig. 4. Spectrogram - 4

QUESTION 2 - LEADER SPEECH ANALYSIS

The task involved analyzing the given historical speech recordings on the grounds of emotional tone and speaking styles. Zero Crossing Rate (ZCR), Short-Time Energy (STE) and Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from the audio samples and analyzed.

The librosa [1] Python package was used for implementation and the matplotlib [2] Python package was used for visualization.

AUDIO FEATURE EXTRACTION

To reduce noise and enhance the quality of recordings, `librosa.effects.preemphasis` was used. It performs high-pass filtering, which amplifies high-frequency components and reduces/removes low frequency components and makes speech samples less noisy.

To deal with varying size of speeches, mean of the feature values across the frames were taken where needed.

The code for calculating of **Zero-Crossing Rate (ZCR)** and **Short-Time Energy (STE)** was implemented from scratch. The formula for Zero-Crossing Rate and Short-Time Energy used as per code:

$$ZCR = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M-1}\frac{|\text{sgn}(x_n[m]) - \text{sgn}(x_n[m-1])|}{2}\right) \tag{1}$$

$$STE = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=0}^{M-1}x_n^2[m]\right) \tag{2}$$

- $x_n[m]$ represents the frame of the audio signal.
- M is the frame size.
- N is the total number of frames.

This features helps to identify variations in speech intensity across time.

For **Mel-Frequency Cepstral Coefficients (MFCCs),** the `librosa.feature.mfcc` function was used, extracting the first 13 coefficients as was instructed. To create feature representation using this, ***mean*** and ***standard deviation*** of these MFCCs were calculated, thus giving total of 26 features from MFCC.

To visualize and compare the extracted features across different speakers, the Zero-Crossing Rate, Short-Time Energy, and MFCCs were plotted, keeping the leaders (speakers) on the x-axis. This helped to make a direct comparision in speech patterns and energy variations between different individuals. The plots of ZCR (Fig. 6), STE(Fig. 5), and the mean(Fig. 7) and standard deviation(Fig. 8) of the first MFCC are included in this report. Additional plots for the remaining MFCC features can be found in the submitted ZIP file.

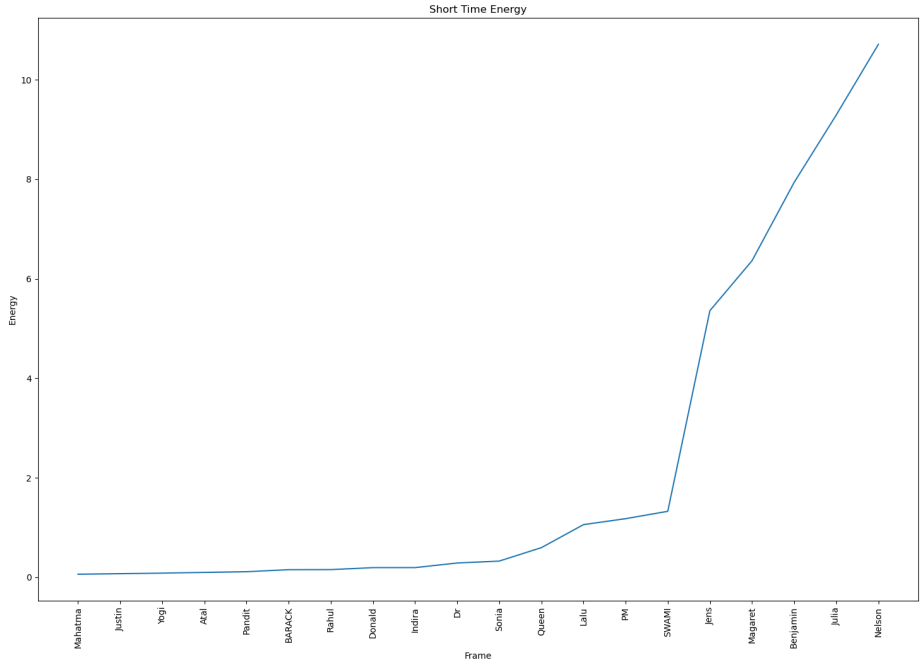

Fig. 5. Short Time Energy

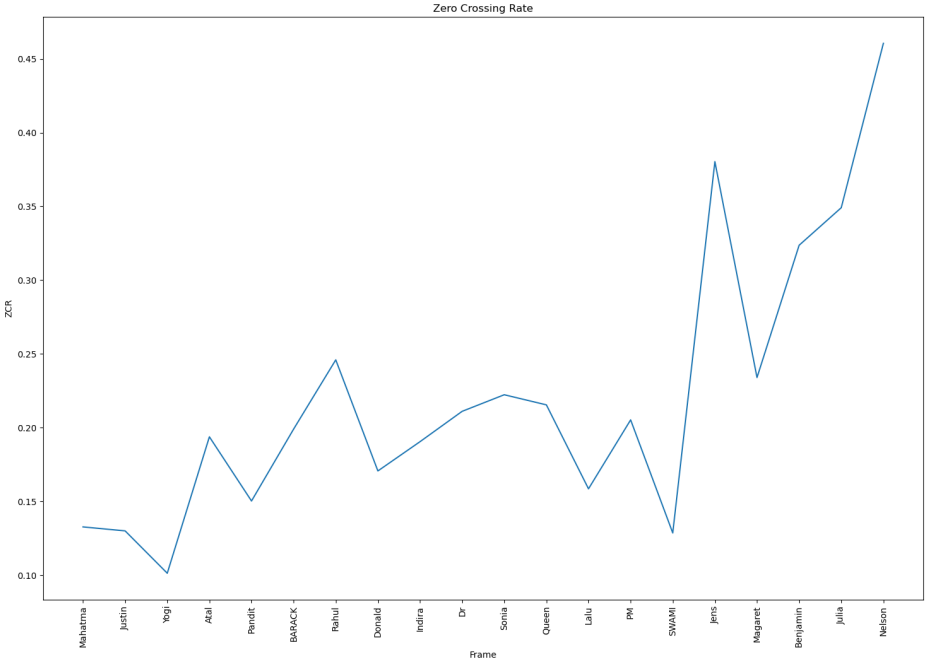

Fig. 6. Zero Crossing Rate

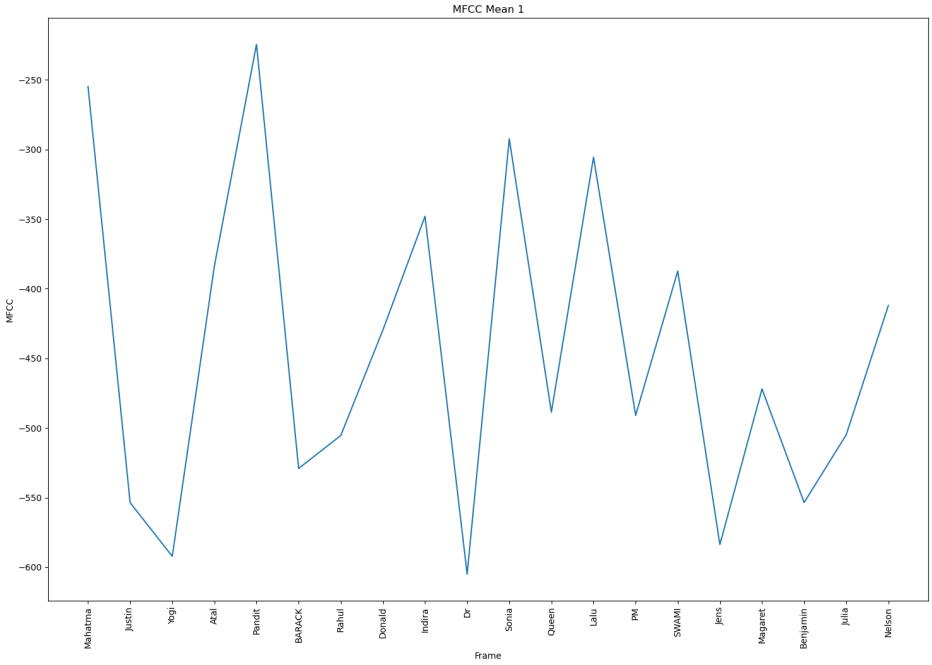

Fig. 7. MFCC 1 - Mean

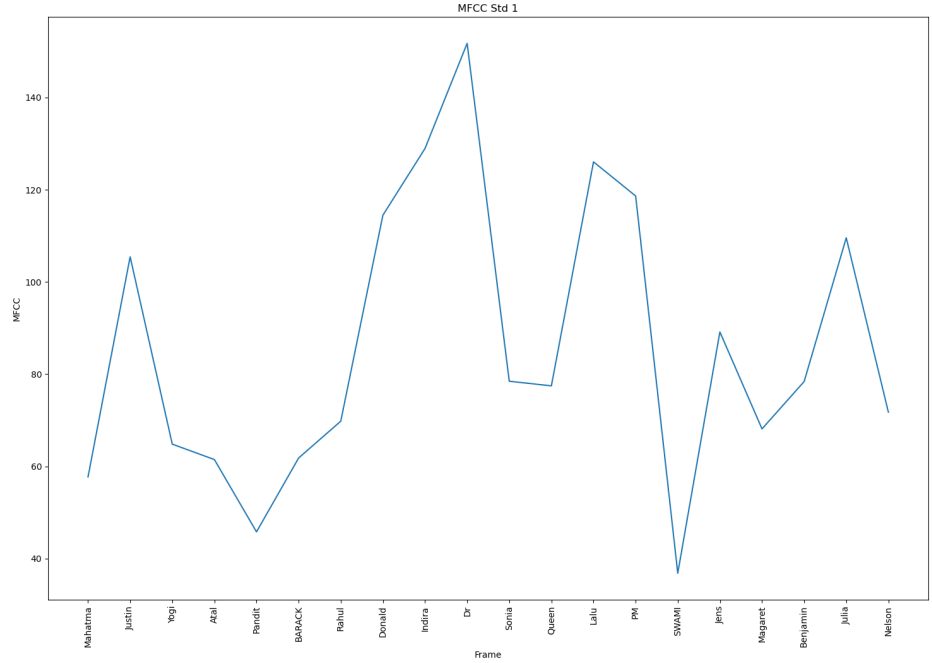

Fig. 8. MFCC 2 - Standard Deviation

COMPARATIVE ANALYSIS

The two speeches chosen for this task were from the given dataset itself. They are as follows:

- **Swami Vivekananda** - Calm and Formal
- **Nelson Mandela** - Passionate and Aggressive

The table (Table 1) below shows the extracted values of a few features from all.

Feature	Swami Vivekananda	Nelson Mandela
Zero Crossing Rate	0.1286	0.4606
Short-Term Energy	1.3267	10.7135
MFCC Mean 1	-383.6407	-224.4787
MFCC Mean 2	89.1737	30.3432

Table 1. Comparison of Speech Features between Swami Vivekananda and Nelson Mandela

Some of the observations on the points mentioned in the question are as follows:

- **Energy of the Signal.** Swami's speech has lower STE which indicates a calm and composed tone and a measured delivery. Contrary to this, Mandela's speech has a much higher STE which indicates more powerful delivery with energy bursts and an impactful manner of addressing.

- **Frequency Content.** As can be observed, ZCR is low in Swami's speech when compared to Mandel's, this suggests that Swami's speech had smooth pronunciations and less number of rapid phonemes transitions, while Mandela's speech had more number of as well as sharper transitions.
  Also, the MFCC values suggest that the spectral variation is high in Swami's speech thus showing more tone modulation while it is low in Mandela's speech due a powerful delivery.

- **General Shape of the Vocal Tract.** Lower ZCR in Swami's speech suggest stable vocal tract whereas high ZCR in Mandela's speech suggest dynamic vocal tract shifts.

LIMITATIONS AND IMPROVEMENTS

Limitations. Some of the possible limitations of using traditional speech features are as follows:

- ***Technical Limitations and Poor Audio Quality*** The early microphones had limited frequency capture ranges, which might cause them to lose some frequencies containing sensetive emotional information.
  Often, there is a lot noise and variation in historical recordings due to low quality of recording instruments and technology of the time. Thus recordings done at different times and places might vary too much even for the same person. Also, the features used are highly sensetive to noise and distortions in audio.

– ***Speaker Variability*** Everyone has a different style of speaking and different vocal characteristics which are inherent in them as a human. Such differences makes it difficult to understand if there is some emotional variation between the speeches of two people or is it just a natural difference. Someone's motivations tone maybe someone's normal way of speaking.

– ***Limited Context and Awareness*** The features used mainly capture physical properties related to speech, and may not be able to fully capture emotions in those speeches due to lack of linguistic context. Along with this, the features are often extracted from frames which are of limited size and then averaged out to give a global context. This makes it impossible to capture the emotional variation at specific points and the evolution of dynamics over extended time in the speech.

Possible Improvements. The following improvements could lead to better feature extraction and analysis:

– ***Deep Learning based Methods*** Deep Learning techniques such as CNN+LSTM architecture or Transformer architectures which provide advanced contextual awareness can be used. Moreover, such models trained on modern high quality data for feature extraction can be used on historical data using domain adaption and can be relied upon to make better feature spaces.

– ***Advanced Noise Reduction*** Using advanced noise reduction techniques involving learning methods can lead to better feature extraction and thus better analysis.

QUESTION 3 - VOWEL CLASSIFICATION AND FORMANT ANALYSIS

The task involved developing a Python-based system for classifying five different vowel sounds (/a/, /e/, /i/, /o/, /u/) using traditional speech processing techniques. The pipeline implemented extracted acoustic features, formant frequencies (F1, F2, F3) and fundamental frequency (F0), from the provided dataste of total 60 speech samples. These features were then used for classification and performance evaluation.

The librosa [1] Python package and scikit-learn [3] Python package were used for audio processing and machine learning algorithms, respectively. And matplotlib [2] Python Package was used for for visualization.

FEATURE EXTRACTION

As mentioned, acoustic features, the first three formant frequencies (F1, F2, F3) and the fundamental frequency (F0) were extracted from the audio samples to be used for training classifiers. The code was implemented from scratch and no direct library function was used. The method followed is as below:

- **Formant Frequency Extraction (F1, F2, F3):**
  - ***Linear Predictive Coding (LPC)*** was implemented. Reference taken from librosa [4].
  - Burg's algorithm was used to iteratively calculate reflection coefficients by minimizing prediction errors.
  - Then LPC polynomial's roots were computed, and only those within the unit circle with positive imaginary components were kept.
  - Finally, angles of these roots were converted into formant frequencies, and first three were selected.

- **Fundamental Frequency (F0) Extraction:**
  - ***Autocorrelation method*** was used to estimate this.
  - The speech signal was first divided into frames of size 1024 with a 50% overlap.
  - For each frame, autocorrelation was calculated and the lag corresponding to the highest peak (within human pitch range) was found.
  - Finally, the fundamental frequency was calculated as F0 = sampling rate / peak lag, and the average over all frames was taken as the final F0 of the speech sample.

The plots corresponding to each frequency distribution (Fig. 9) is attached. It shows the distribution of the extracted frequencies (F0, F1, F2, and F3) for different vowels across all the speech samples.

The plot for vowel space visualization (Fig. 10) using F1-F2 formants is also attached below.

CLASSIFICATION SYSTEM

The features extracted from the above method were stored in a CSV file along with the corresponding vowel labels.

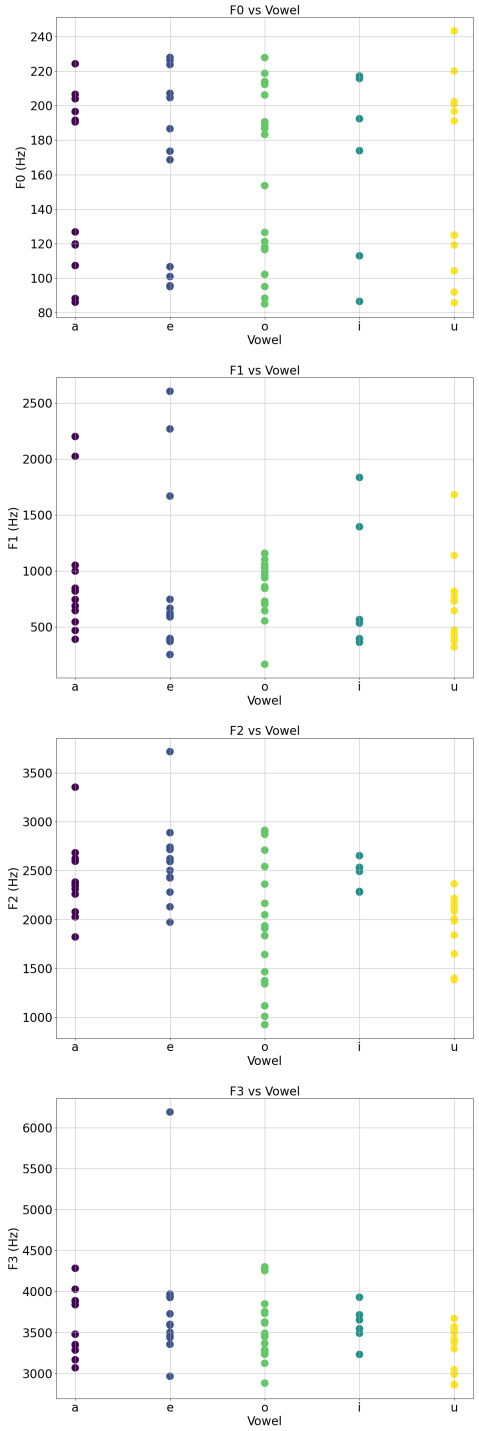

Fig. 9. Audio Features: Graphs of F0, F1, F2, and F3 frequencies for different speech samples representing vowels.

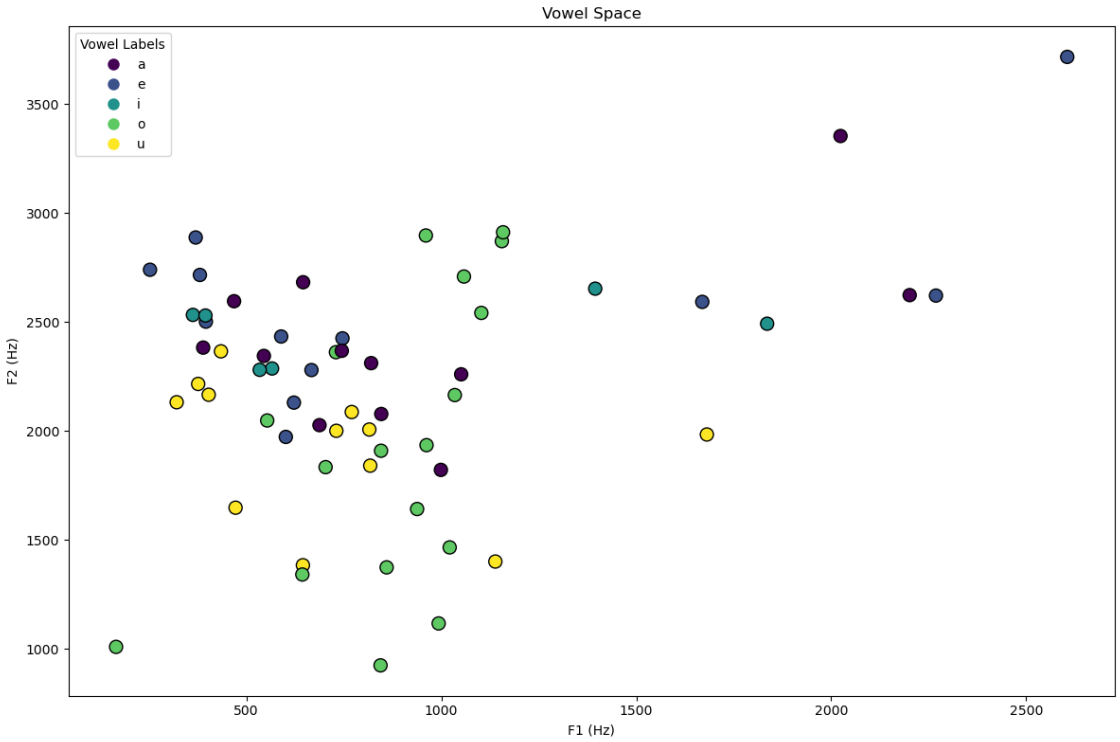

Fig. 10. Vowel Space (F1-F2 formants)

Then, as instructed, 80:20 train-test split was applied to the dataset. and the features were standardized using a *Standard Scaler* to normalize their distributions as the models are sensitive to order of the features.

Then, four different classification models were trained and tested on the dataset:

- K-Nearest Neighbors (KNN)
- Gaussian Mixture Model (GMM)
- Decision Tree Classifier (DTC)
- Support Vector Classifier (SVC)

The accuracy of each model was recorded, and confusion matrices were plotted to analyze their performance.

The confusion matrices for all models are presented in Fig. 15. The accuracies for all the models are reported in Table 2.

The individual confusion matrices can be found in the submitted zip.

Table 2. Training and Testing Accuracy of Different Models

Model	**Train Accuracy (%)**	**Test Accuracy (%)**
KNN	60.42	41.67
Decision Tree (DT)	100.00	41.67
Support Vector Machine (SVM)	60.42	50.00
Gaussian Mixture Model (GMM)	20.83	0.00

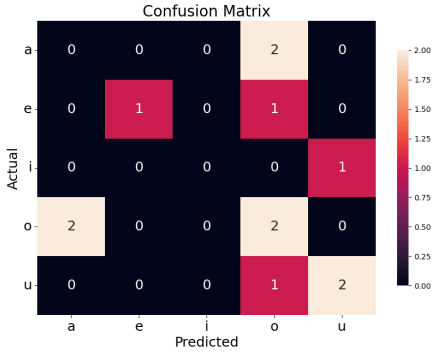

Fig. 11. Confusion Matrix - KNN

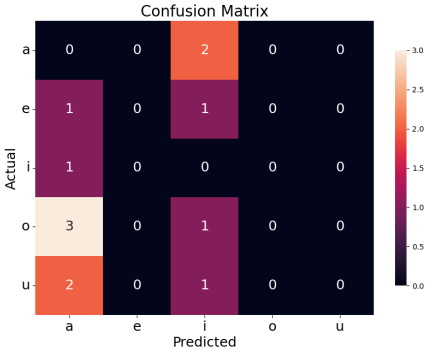

Fig. 12. Confusion Matrix - GMM

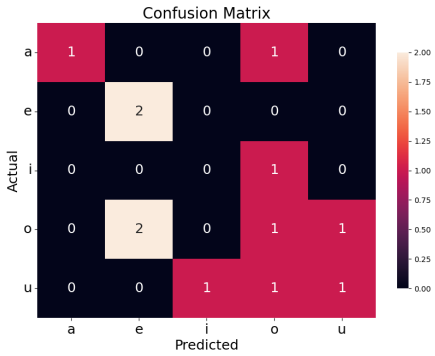

Fig. 13. Confusion Matrix - DTC

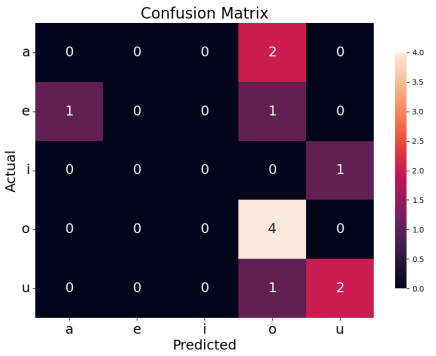

Fig. 14. Confusion Matrix - SVC

Fig. 15. Confusion Matrices for Each Classification Model

All the above was done using the scikit-learn Python package.

ANALYSIS AND REFLECTION

Comparison with Theoretical Expectations. In standard vowel acoustic practices, the first two formants (F1 and F2) are known to be important for classification. The plotted vowel space (Fig. 10) and the formant distribution plots (Fig. 9), depicts that the calculated formants are a bit higher than the expected formants for adult male and adult female as mentioned here [5]. This maybe due to some error in the implementation. Otherwise the frequency distribution of vowels for the formants seem to be good. Also, the vowel space of F1-F2 formant seem to be following the shape generally observed and as shown in the cited reference. The mix seen can be due to the combined plotting of female and male voice features.

Sources of Error and Confusion. Although the feature extraction techniques are well proven, the classification results are not so good and significant misclassification can be observed, especially in the GMM model, which achieved 0% accuracy on the test data. Possible sources of error can be:

- ***Inaccurate Formant estimation***: As mentioned above, although the formants are good, but they are not perfectly aligning with the expected distributions, thus they could have been a source of error. Also, the overlap which can be clearly observed would have been another source of error. This is evident in the confusion matrices attached, in which certain vowels are frequently misclassified as other vowels.
- ***Dataset limitations***: The dataset may have had inconsistent recordings, varying accents or noise, all of which affect the feature extraction process and thus the classification performance.
- ***Imbalanced classifier complexity***: The Decision Tree (DT) classifier overfitted on the train data (100% accuracy) but performed inaccurately on the test set. This indicates that some classifiers suffer from high variance.

Relation to Historical Speech Recognition Systems. The approach followed is a traditional feature extraction technique. Formant analysis and fundamental frequency estimation were among the earliest vowel classification techniques. Traditional approaches, started with handcrafting features like formants, pitch, and other properties to identify speech sounds. On the other hand, modern methods uses deep learning and embeddings, which are often not interpretable and require large amount of data. The technique used in the approach can be effective for low-data situations or when interpretable speech analysis tasks are needed, thus aligning with the traditional approaches..

Potential Improvements. To improve classification metrics, the following changes could be made:

- ***More Features***: Using other features such as Mel-Frequency Cepstral Coefficients (MFCCs), with F1, F2, and F3, could provide more robust vowel classification.
- ***Other Classifiers***: Instead of the used models, using ensemble methods (e.g., Random Forest, Gradient Boosting) to deal with high varianve or deep learning (e.g., CNNs for spectrogram analysis, only if we have enough data) may improve accuracy.

References

1. https://librosa.org/doc/latest/index.html.
2. https://matplotlib.org/.
3. https://scikit-learn.org/.
4. https://librosa.org/doc/main/_modules/librosa/core/audio.html#lpc.
5. Raymond D. Kent and Houri K. Vorperian. Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, 74:74–97, 2018.