Speech Understanding, Assignment 2

Akansha Gautam
M23CSA506
IIT Jodhpur
m23csa506@iitj.ac.in

1. Speech Enhancement

1.1. II: Speaker Verification Task

To perform the speaker verification task, I have used the **wav2vec2 xlsr** pre-trained speaker verification model from HugggingFace and evaluated this pre-trained model using the total number of 500 trial pairs of VoxCeleb1 dataset given in the question.

Metric	Pre-trained Model	Fine-tuned Model
EER (in %)	0.488	0.5
TAR@1% FAR	0.488	0.5
Speaker Identification Accuracy	0.488	0.5

Table 1. Comparison of speaker verification task for pre-trained and fine-tuned models

I have started this task by extracting the embeddings of audio files using the pre-trained wav2vec2 xlsr model. I have used the following metrics to evaluate the pre-trained model's performance:

- EER(in %)
- TAR@1%FAR
- Speaker Identification Accuracy

The pre-trained model performance was measured using the 500 trial pairs taken from the VoxCeleb1 dataset. The pre-trained model was then fine-tuned on the VoxCeleb2 dataset. Once fine-tuned, the performance of this model was measured on the same VoxCeleb1 data set. Table 1 shows the comparison of pre-trained and fine-tuned models performance on the speaker verification task. From this table, we can clearly see that the model has performed better when fine-tuned on the VoxCeleb2 dataset.

1.2. III: Perform speaker separation and speech enhancement on multi-speaker scenario dataset

I have created the multi-speaker scenario dataset of 1000 samples using first 50 identities (when sorted in ascending order) of the provided VoxCeleb2 dataset and use the next 50 identities (when sorted in ascending order).

2. MFCC Feature Extraction and Comparative Analysis of Indian Languages

2.1. Task A

2.1.1. Dataset

The dataset consists of audio samples of 10 different Indian languages, which can be found on Kaggle using the this link. In this dataset, each audio file is 5 seconds long and belongs to one of the following Indian language classes such as Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, and Urdu.

Figures 1, 2, and 3 show the audio waveforms for Bengali, Hindi, and Urdu languages respectively. Each waveform shows how the amplitude changes over time, highlighting the unique patterns in the speech of each language. By looking at these waveforms, we can get a sense of how speech varies between the languages.

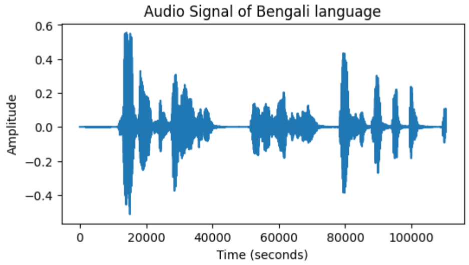


Figure 1. Audio signal of Bengali language

Figure 4 shows the distribution of audio samples for each Indian language. From this figure, we can see that the Urdu language contains more than 30,000 audio samples whereas Kannada contains roughly 20,000 audio samples.

2.1.2. Comparative analysis of MFCC spectrograms

Mel-Frequency Cepstral Coefficients (MFCCs) captures the essential characteristics present in audio-signal which are most discernible to the human ear. A positive MFCC value implies that the spectral energy lies in the low-frequency

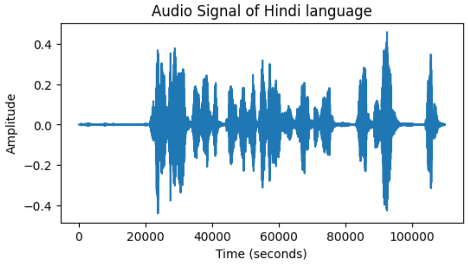


Figure 2. Audio signal of Hindi language

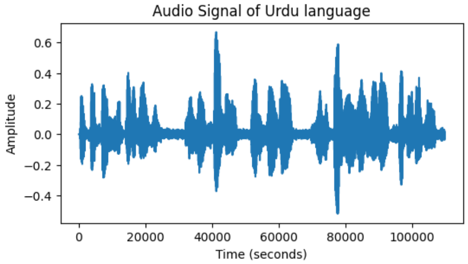


Figure 3. Audio signal of Urdu language

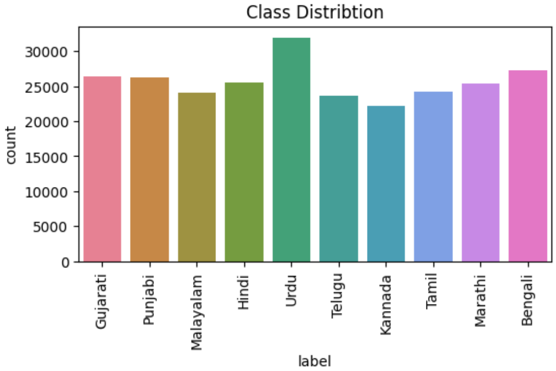


Figure 4. Class Distribution of audio samples

regions whereas the negative MFCC value implies that the spectral energy lies in the high-frequency regions.

To handle the potential issues like varying audio lengths and noise, I have first trimmed the audio signals to a duration of 5 seconds. Then, I have normalized the audio signal and applied a high-pass filter to reduce low-frequency noise. Figure 5 shows the MFCC spectrogram of audio samples taken for ten Indian languages: Gujarati, Punjabi, Malayalam, Hindi, Urdu, Telugu, Kannada, Tamil, Marathi, and Bengali. The x-axis represent the time, the y-axis shows the frequency and the colour intensity depicts the amplitude in each spectrogram. All the spectrograms are mostly in green and yellow colour demonstrating the presence of lower amplitude values across all the different Indian languages.

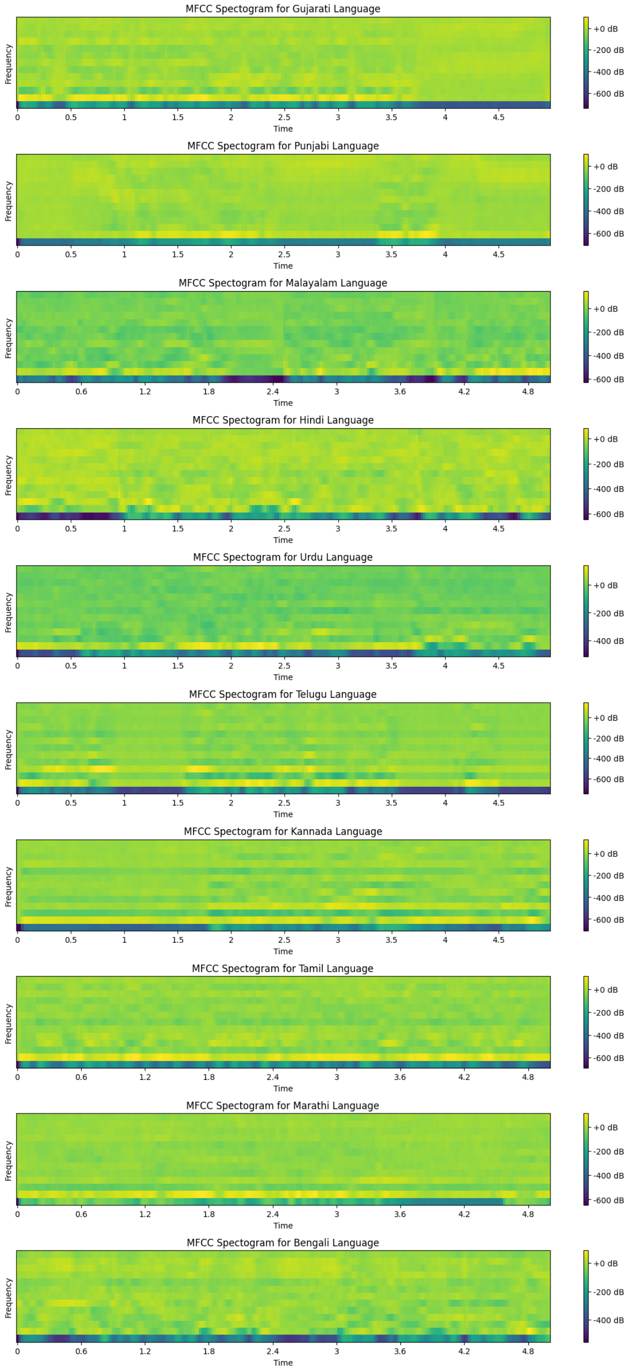


Figure 5. MFCC spectrogram of audio samples of each language

The MFCC spectrogram of Gujarati language is mostly green in colour, with some dark patches of blue color near the bottom. It signifies the presence of consistent low-frequency features which implies that this audio file likely has stronger low-frequency phonemes and less high-frequency variation.

The MFCC spectrogram of Hindi and Urdu language consists of smooth and consistent patterns over time with less frequency variations. However, Hindi has lower amplitude as compared to Urdu as hindi language spectogram lies more on the yellow side whereas urdu spectogram lies more on the greener side.

The MFCC spectrogram of Bengali language is mostly green with some dark patches at bottom however Marathi language spectrogram is lighter overall with more yellow-green tones. The amplitude (energy) is bengali language is higher than the marathi language. There are a lot of patches present in bengali language spectogram which possibly reflects the noisy features in comparison to marathi whose spectrogram is much smoother comparatively.

2.1.3. Statistical analysis of MFCC spectrograms

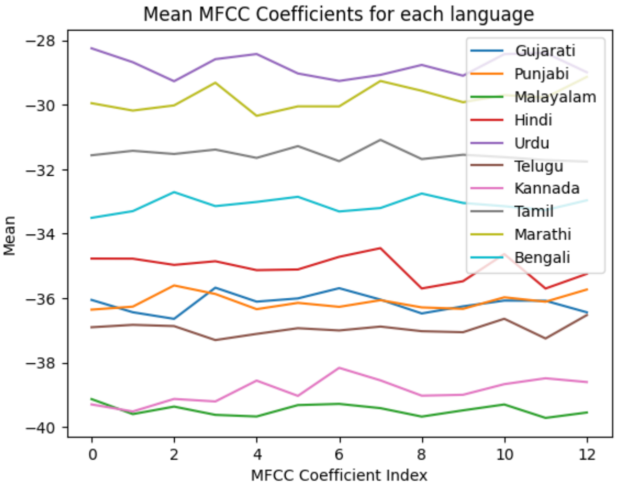


Figure 6. Mean MFCC spectrogram coefficients

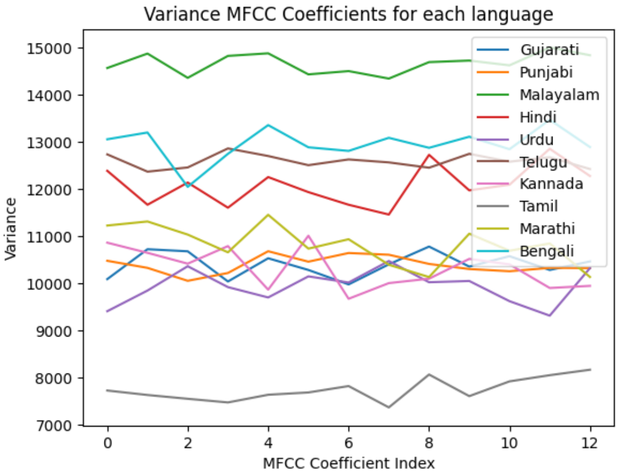
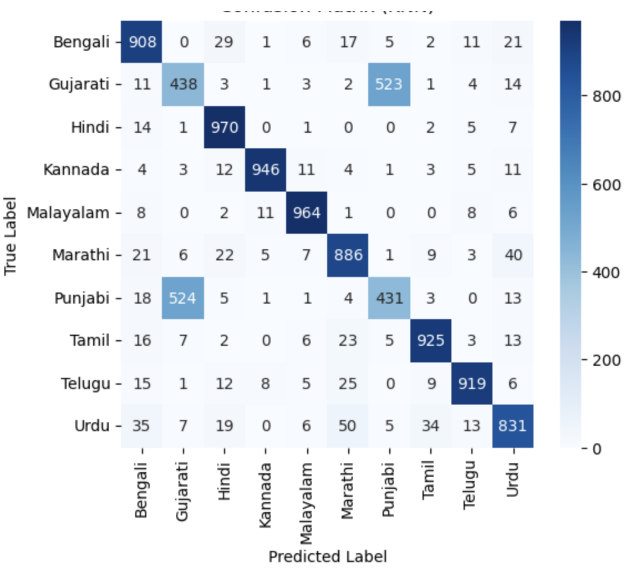


Figure 7. Variance MFCC spectrogram coefficients

I have extracted the first 13 MFCC coefficients to perform a statistical analysis (e.g., compute the mean and variance of MFCC coefficients) to quantify differences between languages. The figure 6 and figure 7 shows the Mean and Variance of MFCC spectrogram coefficients respectively. If we will compare the mean then, Urdu language has the highest mean whereas Malayalam language has the lowest mean. If we will look at the variance comparison, then, Malayalam language has the highest variance whereas Tamil has the lowest variance.

2.2. Task B



	Bengali	Gujarati	Hindi	Kannada	Malayalam	Marathi	Punjabi	Tamil	Telugu	Urdu
Bengali	908	0	29	1	6	17	5	2	11	21
Gujarati	11	438	3	1	3	2	523	1	4	14
Hindi	14	1	970	0	1	0	0	2	5	7
Kannada	4	3	12	946	11	4	1	3	5	11
Malayalam	8	0	2	11	964	1	0	0	8	6
Marathi	21	6	22	5	7	886	1	9	3	40
Punjabi	18	524	5	1	1	4	431	3	0	13
Tamil	16	7	2	0	6	23	5	925	3	13
Telugu	15	1	12	8	5	25	0	9	919	6
Urdu	35	7	19	0	6	50	5	34	13	831

Figure 8. Confusion matrix

I have taken the 5000 samples for each Indian language from the given dataset and then, the MFCC features have been extracted. Post feature extraction, the train and test dataset was created by splitting the features and labels into an 80-20 ratio with a random state of 45. The features dataset was then normalized using the StandardScaler function from the Scikit-learn library. The SVM Model was selected as a classifier to train the model on the training features and labels set. Then, the accuracy of the model's prediction was analysed using the test labels which came out to be 82% using the above extracted features.

Figure 8 shows the confusion matrix which determines how well the Support Vector Classifier predicted the Indian language labels. In this figure, the row represents the true labels, however, the column represents the predicted labels. We can see that the model is able to correctly classify the Indian language labels when trained using the first 13 MFCC coefficients.

2.2.1. Potential challenges in using MFCCs to differentiate between languages

Mel-Frequency Cepstral Coefficient (MFCC) features help us understand the frequency strength, audio variation and general shape and tone of the speech. However, there are a several potential challenges in using MFCC features to differentiate between languages as listed below:

- People speak in different tone and modularity with various accents which can confuse the model to classify languages when given MFCC features.
- Background noise can also interfere with the MFCC features and can degrade the model's performance.

References