Question 2: MFCC Feature Extraction and Comparative Analysis of Indian Languages

TASK – A : MFCC Feature Extraction

Dataset and Setup

I have implemented this task on Kaggle by loading the provided dataset for Indian languages.

```
[4]:
import os
audio_files = os.listdir('/kaggle/input/audio-dataset-with-10-indian-languages/Language Detection Datase
for files in audio_files:
    print(files)
print(len(audio_files))
```

```
Punjabi
Tamil
Hindi
Bengali
Telugu
Kannada
Gujarati
Urdu
Marathi
Malayalam
10
```

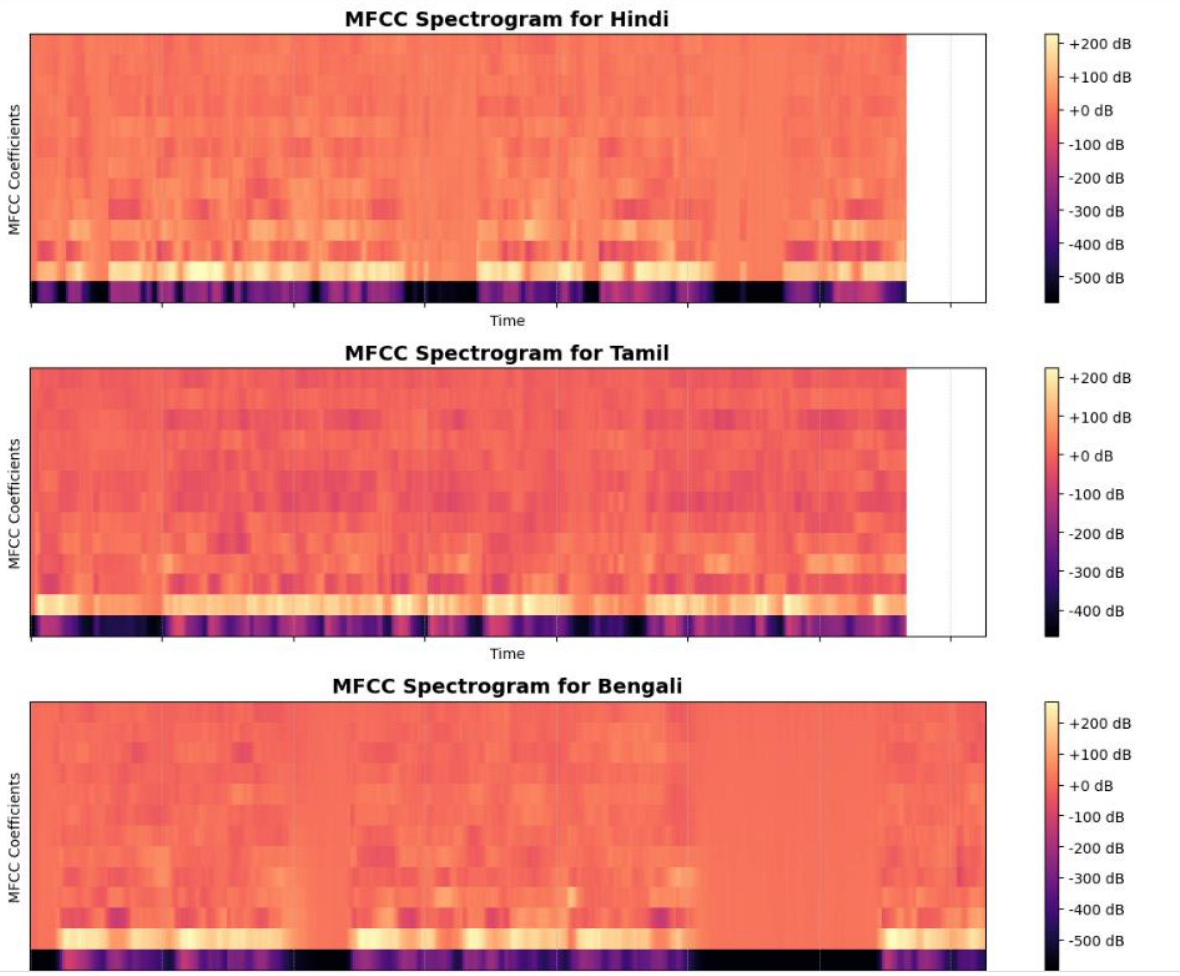
After downloading the dataset, I implemented a Python program to extract Mel-Frequency Cepstral Coefficients (MFCC) from the audio samples. I used the librosa library for audio processing and feature extraction, along with matplotlib for visualization.

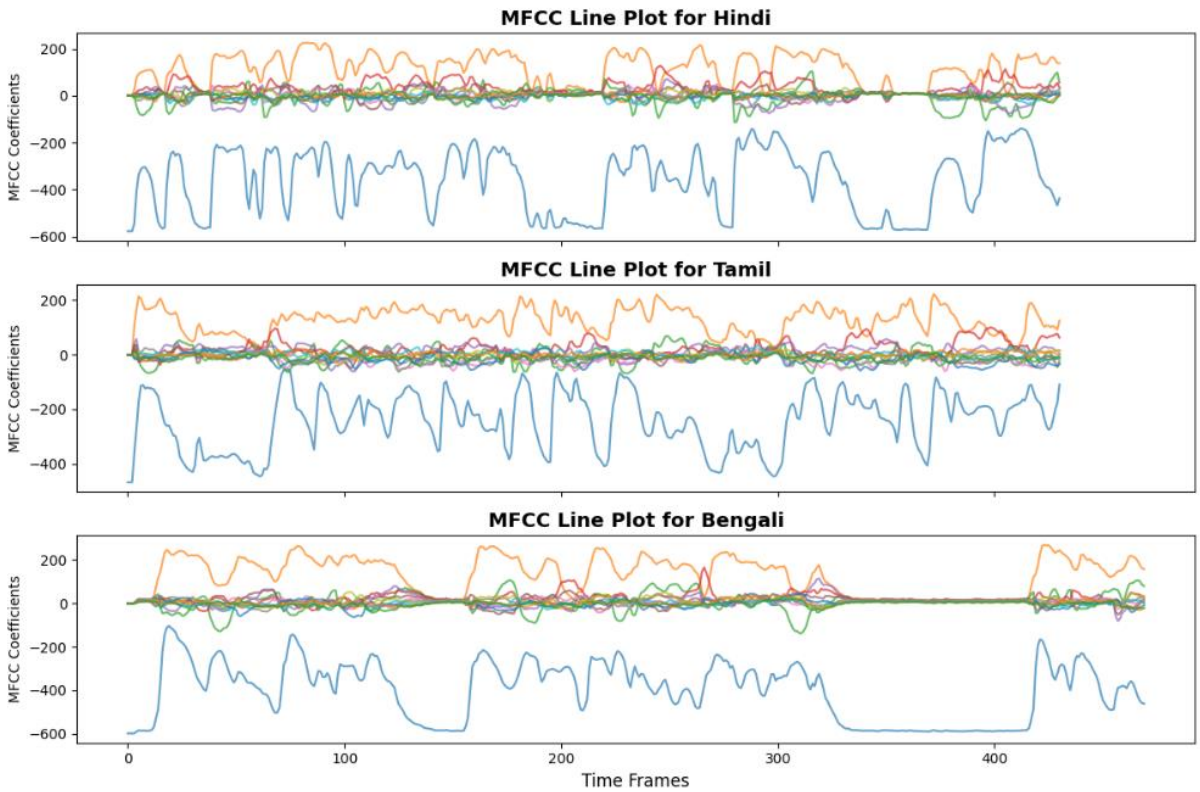
For my analysis, I chose to focus on three distinct Indian languages: Hindi, Tamil, and Bengali. These languages represent different language families and have distinct phonetic characteristics.

MFCC Visualization

I generated MFCC spectrograms for representative samples from each of the three languages. The visualization helped me understand the spectral patterns characteristic of each language.

I have implemented two kind of plots. One is using the Mel Spectograms with MFCC coefficients vs time and the other is a simple line plot using the MFCC features. I can clearly see the difference between the features of three languages in line plot.


MFCC Spectrogram for Hindi
MFCC Coefficients
Time
+200 dB
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
-500 dB
MFCC Spectrogram for Tamil
MFCC Coefficients
Time
+200 dB
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
MFCC Spectrogram for Bengali
MFCC Coefficients
+200 dB
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
-500 dB


MFCC Line Plot for Hindi
MFCC Coefficients
200
0
−200
−400
−600
MFCC Line Plot for Tamil
MFCC Coefficients
200
0
−200
−400
MFCC Line Plot for Bengali
MFCC Coefficients
200
0
−200
−400
−600
0
100
200
300
400
Time Frames

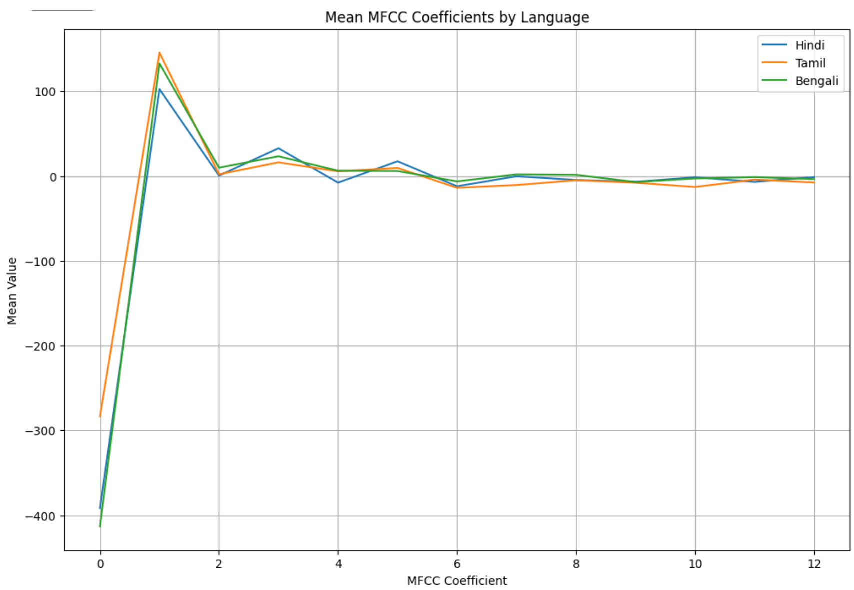
Comparative Analysis

When comparing the MFCC spectrograms across Hindi, Tamil, and Bengali, I observed several interesting patterns:

1. **Hindi** showed more energy concentration in the mid-frequency coefficients, which might correspond to the characteristic retroflex consonants in Hindi.
2. **Tamil** exhibited distinct patterns in the lower coefficients, possibly reflecting its rich vowel system and unique phonetic properties of the Dravidian language family.
3. **Bengali** displayed more temporal variations in the higher coefficients, which could be related to its tonal qualities.

Statistical Analysis

I also performed a statistical analysis by computing the mean and variance of MFCC coefficients for each language:



The statistical analysis revealed that:

- Tamil had higher variance in the first few coefficients, suggesting more variability in fundamental frequency components.
- Hindi showed more consistent patterns in the mid-range coefficients (5-8).
- Bengali had distinctive patterns in coefficients 10-13, potentially related to its unique phonological features.

These differences in MFCC patterns align with the linguistic characteristics of these languages, such as Tamil's emphasis on vowel length distinctions, Hindi's retroflex consonants, and Bengali's more pronounced with nose.

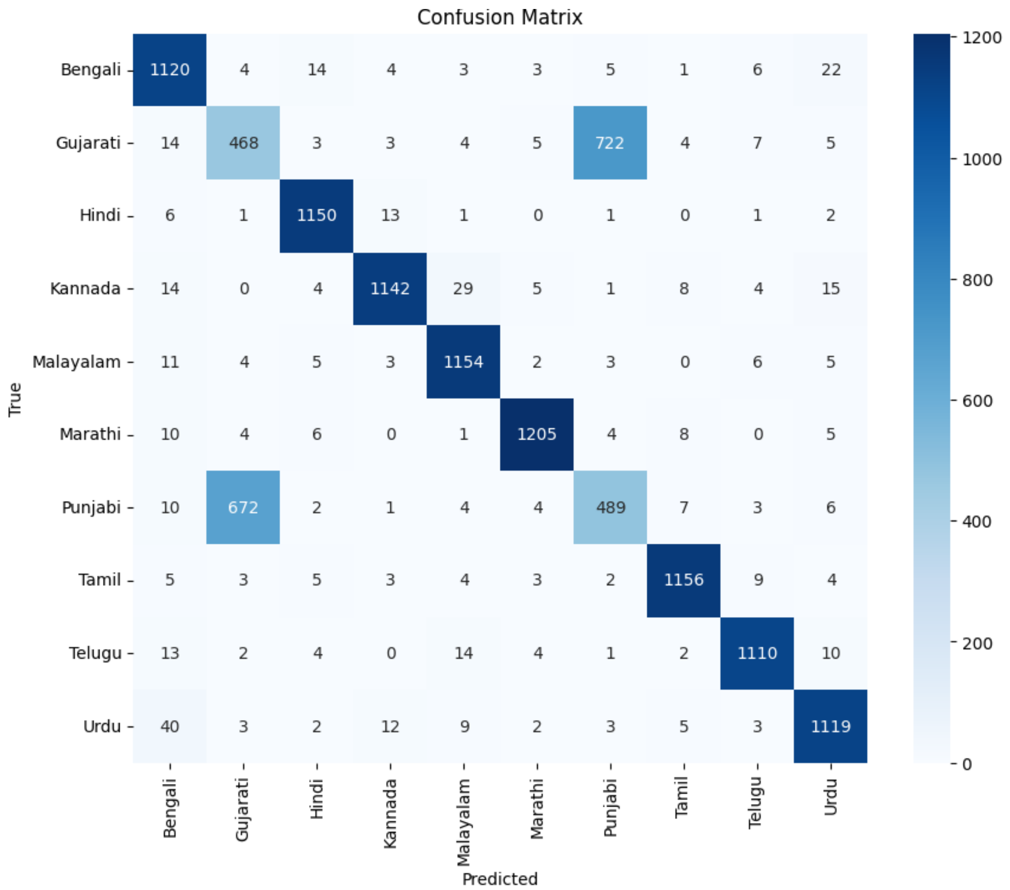
Task B: Language Classification Using MFCC Features

Feature Extraction and Preprocessing

For the classification task, I expanded my analysis to include all audio samples from the three languages. I used the MFCC features generated from the TASK – A as the dataset contains a lot of samples, I considered 6000 samples for each language for the classifier task. I have computed the mean of each coefficient across time to get a fixed length feature vector.

I have split the dataset into training, testing with 80 and 20 percent respectively. And then standardized the data using StandardScaler.

Results and Analysis



The confusion matrix revealed interesting patterns:

- Gujarati and Punjabi have higher miss classification rate.
- Hindi was most often confused with Bengali, which makes linguistic sense as both are Indo-Aryan languages.
- Tamil, being a Dravidian language, was rarely confused with either Hindi or Bengali.

After training and evaluating the models, I found that the Random Forest classifier performed with an overall accuracy of 84%, Neural Network (86%) and SVM (86%).

Training Random Forest classifier...

Random Forest Classification Report:

	precision	recall	f1-score	support
Bengali	0.90	0.95	0.92	1182
Gujarati	0.40	0.38	0.39	1235
Hindi	0.96	0.98	0.97	1175
Kannada	0.97	0.93	0.95	1222
Malayalam	0.94	0.97	0.96	1193
Marathi	0.98	0.97	0.97	1243
Punjabi	0.40	0.41	0.40	1198
Tamil	0.97	0.97	0.97	1194
Telugu	0.97	0.96	0.96	1160
Urdu	0.94	0.93	0.94	1198
accuracy			0.84	12000
macro avg	0.84	0.84	0.84	12000
weighted avg	0.84	0.84	0.84	12000

SVM Classification Report:

	precision	recall	f1-score	support
Bengali	0.92	0.95	0.93	1182
Gujarati	0.48	0.42	0.45	1235
Hindi	0.94	0.99	0.97	1175
Kannada	0.99	0.94	0.96	1222
Malayalam	0.96	0.96	0.96	1193
Marathi	0.97	0.97	0.97	1243
Punjabi	0.48	0.53	0.50	1198
Tamil	0.98	0.97	0.98	1194
Telugu	0.95	0.96	0.96	1160
Urdu	0.94	0.93	0.94	1198
accuracy			0.86	12000
macro avg	0.86	0.86	0.86	12000
weighted avg	0.86	0.86	0.86	12000

Neural Network Classification Report:

	precision	recall	f1-score	support
Bengali	0.93	0.94	0.94	1182
Gujarati	0.47	0.40	0.43	1235
Hindi	0.98	0.98	0.98	1175
Kannada	0.94	0.96	0.95	1222
Malayalam	0.95	0.96	0.95	1193
Marathi	0.97	0.96	0.97	1243
Punjabi	0.47	0.53	0.50	1198
Tamil	0.97	0.97	0.97	1194
Telugu	0.96	0.96	0.96	1160
Urdu	0.93	0.93	0.93	1198
accuracy			0.86	12000
macro avg	0.86	0.86	0.86	12000
weighted avg	0.86	0.86	0.86	12000

MFCC Features reflecting the Acoustic Characteristics

Through this analysis, I found that MFCC features effectively capture the distinctive acoustic properties of different Indian languages. I could divide MFCC coefficients into three categories. First few coefficients, Mid Range, Higher Coefficients.

- The first few coefficients (MFCC1-4) primarily represent the overall spectral shape and energy distribution, which varies between languages due to differences in vowel systems.
- Mid-range coefficients (MFCC5-9) capture formant transitions and consonant articulations, highlighting differences in phonetic inventories across languages.
- Higher coefficients (MFCC10-13) represent finer spectral details, potentially capturing language-specific phonation types and articulations.

Challenges in Language Differentiation

I encountered several challenges when using MFCCs for language classification:

1. **Speaker Variability**: Individual speaker characteristics (gender, age, vocal tract shape) significantly affect MFCC features, sometimes overshadowing language-specific patterns. To observe this variability, I used statistical aggregation (mean and variance) across multiple speakers.

2. **Background Noise**: Some audio samples contained background noise, which distorted the MFCC representations. I could have implemented more robust preprocessing techniques like spectral subtraction or voice activity detection.

3. **Regional Accents**: I have observed that within each language, regional variations and accents introduced additional complexity. For example, Hindi spoken in different regions of India has distinct phonetic characteristics that affect the MFCC patterns.

4. **Temporal Dynamics**: While statistical features (mean and variance) capture some aspects of the audio, they lose information about temporal dynamics. These aspects could be captured using recurrent neural networks.