MFCC Feature Extraction and Comparative Analysis of Indian Languages

Neermita Bhattacharya

Indian Institute of Technology, Jodhpur, Jodhpur 342037, India
b22cs092@iitj.ac.in

Abstract

Spectrograms are an essential part of speech analysis. A normal spectrogram is a time-frequency representation of an audio signal. It shows how the signal's energy (or amplitude) is distributed over frequency as it changes over time. It is useful in analyzing pitch, various harmonics and noise. But, the spectrogram may contain frequencies that humans cannot perceive. Hence, additional features called the MFCCs (Mel-Frequency Cepstral Coefficients) are used. MFCCs are derived from audio signals to capture features that mimic how humans hear sounds. Utilizing MFCCs gives us better insights into the changes of the vocal tract and the formants of various utterances. They are particularly useful for tasks such as speech and language recognition because they prioritize perceptually relevant information. In this study, I use MFCC features to analyze and classify audio samples from the Audio Dataset with 10 Indian Languages. The work is divided into two main tasks.

In Task A, I extract MFCCs from the audio samples, generate and visualize MFCC spectrograms for at least three selected languages (Hindi, Marathi and Bengali), and compare the patterns to identify differences and similarities. This analysis helps to capture distinctive spectral characteristics that may differentiate between languages. Statistical analysis is also performed, including the calculation of mean and variance of the MFCCs, to quantify these differences.

In Task B, the function to extract MFFCs was used to extract MFFCs for 1000 samples of each language, to build a machine-learning model for language classification. I experimented with a Random Forest Classifier due to its interpretability and effectiveness in handling complex features. The dataset is preprocessed through normalization (min-max scaling) and a train-test split (70-30) to ensure robust model evaluation. By training and testing the classifier on MFCC features, I aim to assess how well these features distinguish between different Indian languages.

This study also discusses the potential challenges in using MFCCs for language identification. Variability in speakers, differences in regional accents, and background noise can affect the reliability of MFCC-based models. Additionally, since MFCCs discard pitch information, tonal differences between languages may be underrepresented. This is represented by the confusion matrix for the Punjabi and Gujarati samples. The findings aim to highlight the effectiveness of MFCCs in capturing language-specific characteristics and provide insights into the strengths and limitations of using MFCCs for speech-based language classification. The codes and reports are available at github.com/Neermita18/Speech-Understanding-PA2.

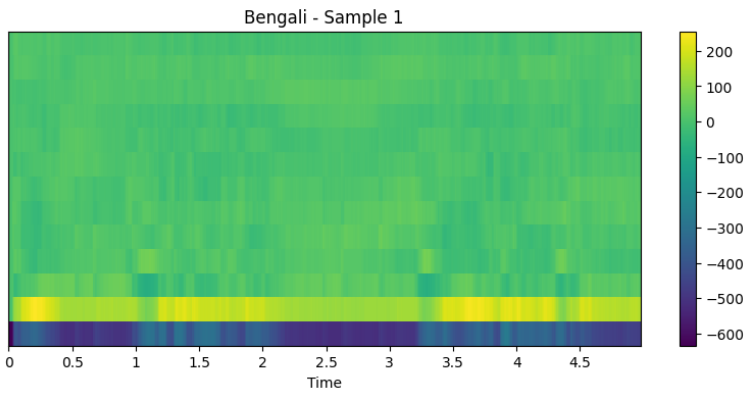


Figure 1: MFCC spectrogram of a Bengali audio sample

Contents

1 Question 2

1.1 Introduction

Speech signals are complex, containing both the source information (such as the vibration of the vocal cords) and the filter information (shaped by the vocal tract). Vibration of the vocal chords (larynx houses them) give rise to different pitches and tones, whereas the vocal tract (consisting of the pharynx, the oral cavity and the nasal cavity), act as a filter or resonator, shaping the sound produced. To analyze and classify speech, it is crucial to extract meaningful features that represent the characteristics of these signals while discarding irrelevant information. The cepstrum represents the audio signal as an addition of the laryngeal and vocal tract components. This can be useful to separate features such as pitch from phonemes. One of the most effective feature representations for speech analysis is the Mel-Frequency Cepstral Coefficients (MFCCs) which is indirectly obtained from the cepstrum. MFCCs are derived from the audio signal and are designed to mimic the human auditory system, focusing on perceptually relevant frequency ranges. These coefficients are widely used in applications such as speech recognition, language identification, and speaker verification due to their ability to capture the vocal tract's shape and structure.

1.1.1 From Time Domain to Cepstrum

A speech signal $s(n)$ is typically modeled as the convolution of two components:

- **Excitation Signal** $e(n)$: A periodic signal generated by the vibration of the vocal cords, which determines the pitch.
- **Vocal Tract Response** $h(n)$: A filter that shapes the excitation signal to produce different phonemes.

Mathematically, this can be represented as:

$$s(n) = e(n) * h(n) \tag{1}$$

Applying the **Discrete Fourier Transform (DFT)** to the speech signal transforms this convolution into a multiplication in the frequency domain:

$$S(k) = E(k) \cdot H(k) \tag{2}$$

Where:

- $S(k)$ is the speech signal's spectrum.
- $E(k)$ is the excitation signal's spectrum.
- $H(k)$ is the vocal tract's frequency response.

To separate these components, we compute the **log-magnitude spectrum**:

$$\log|S(k)| = \log|E(k)| + \log|H(k)| \tag{3}$$

Since addition is easier to manipulate than multiplication, this transformation allows us to decompose the excitation and filter contributions. By applying the **Inverse DFT (IDFT)** to the log-magnitude spectrum, we obtain the **cepstrum**:

$$c(n) = \mathcal{F}^{-1}\{\log|S(k)|\} \tag{4}$$

In the cepstrum, different features occupy distinct regions:

- **Low quefrencies** represent the slowly varying vocal tract information.
- **High quefrencies** capture the rapid variations associated with pitch.

Since speech recognition primarily relies on the shape of the vocal tract, we focus on the low-quefrency region while discarding higher quefrencies.

1.1.2 From Cepstrum to MFCCs

Although the cepstrum provides useful information, it does not align with how humans perceive sound. Human auditory perception is non-linear, meaning that we have finer frequency resolution at lower frequencies and coarser resolution at higher frequencies. The **Mel scale** models this perception and is used to extract the Mel-Frequency Cepstral Coefficients (MFCCs).

Step 1: Pre-emphasis To amplify higher frequencies and balance the spectral content, a pre-emphasis filter is applied:

$$y(n) = s(n) - \alpha s(n-1) \tag{5}$$

Here, α is a small constant, typically set to $\alpha = 0.97$.

Step 2: Framing and Windowing The continuous speech signal is divided into short overlapping frames (e.g., 25 ms with a 10 ms overlap) to capture local characteristics. A window function, such as the **Hamming window**, is applied to each frame:

$$x_w(n) = x(n) \cdot w(n) \tag{6}$$

Step 3: Fourier Transform and Power Spectrum For each frame, the Fast Fourier Transform (FFT) is computed to obtain the magnitude spectrum. The power spectrum is calculated as:

$$P(k) = |X(k)|^2 \tag{7}$$

Step 4: Mel Filter Bank The frequency axis is warped to the Mel scale using the following relation:

$$f_{\text{mel}} = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{8}$$

A set of triangular filters is applied to the power spectrum to measure the energy in each Mel band:

$$E_m = \sum_{k=1}^{K} P(k) \cdot H_m(k) \tag{9}$$

Where:

- E_m is the energy in the m-th Mel filter.
- $H_m(k)$ represents the filter weights.

Step 5: Logarithm of Filter Outputs Taking the logarithm of the Mel-filtered energies mimics human perception by emphasizing lower amplitudes:

$$\log(E_m) \tag{10}$$

Step 6: Discrete Cosine Transform (DCT) Applying the Discrete Cosine Transform compresses the Mel-scaled spectral data into a compact form:

$$\text{MFCC}(n) = \sum_{m=1}^{M} \log(E_m) \cdot \cos\left(\pi n \frac{m - 0.5}{M}\right) \tag{11}$$

Typically, only the first 12-13 MFCC coefficients are retained because they capture the essential characteristics of the vocal tract.

1.1.3 Importance of MFCCs

MFCCs are widely used in speech processing tasks due to the following advantages:

- **Human-like Perception:** The Mel scale aligns with how humans perceive sound frequencies.
- **Source-Filter Separation:** Isolates vocal tract features while discarding pitch-related information.
- **Compact Representation:** DCT reduces redundancy, leading to a low-dimensional but informative feature set suitable for machine learning models.
- **Distinguish Sources:** When extracting MFCCs from audio, we capture how much energy is present across different frequency bands. Since the distribution of harmonics is unique to each sound source, MFCCs can help distinguish between a violin, piano, or human speech.

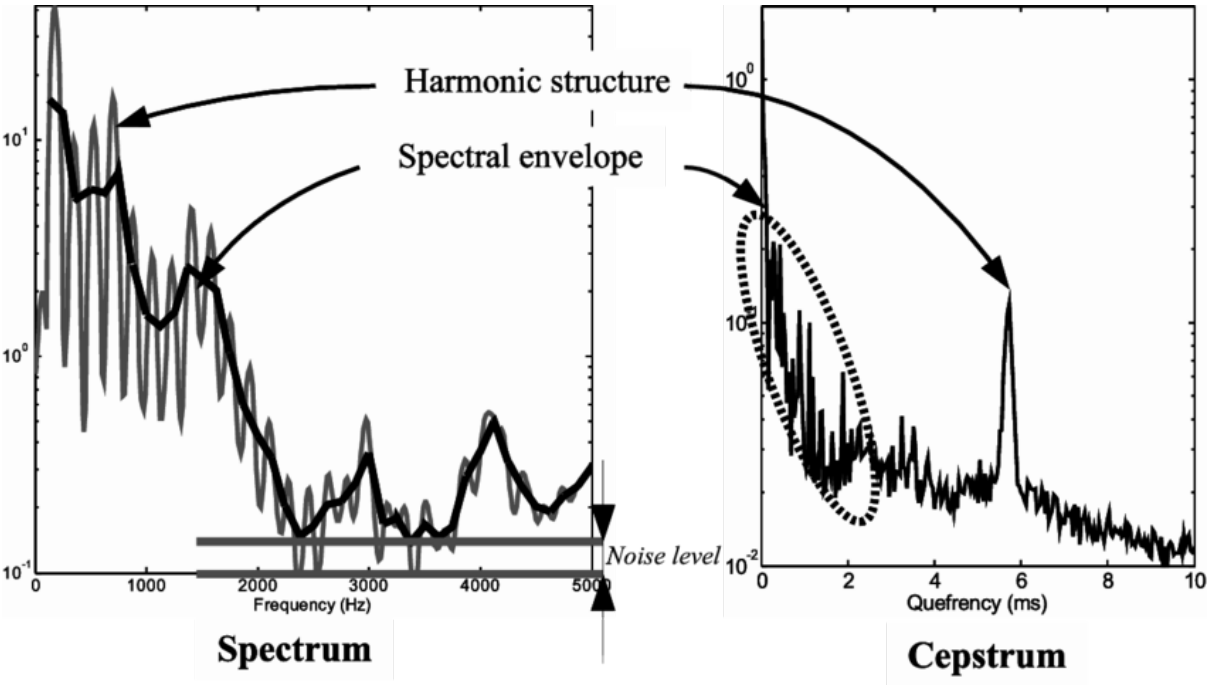


Figure 2: Example of a graphical interpretation of a cepstrum. Reference [1]

1.2 Task A: Feature Extraction and Analysis

1.2.1 Dataset Collection

The dataset used for this task is the **Audio Dataset with 10 Indian Languages** sourced from Kaggle. The dataset contains audio samples from various Indian languages, providing a diverse corpus for analyzing acoustic patterns.

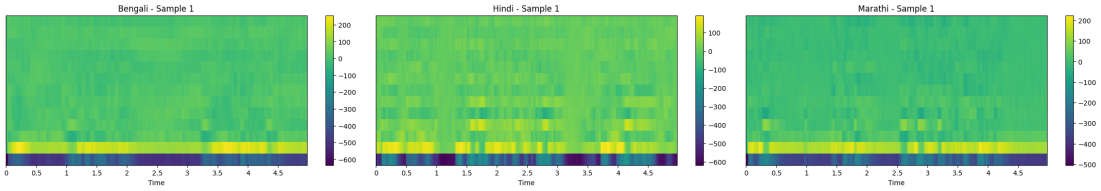
1.2.2 MFCC Extraction

A simple function was implemented to extract Mel-Frequency Cepstral Coefficients (MFCCs) from each audio sample using the `librosa` library. MFCC extraction follows these steps:

- Audio files were loaded using `librosa.load()` The sampling rate was kept as the original.
- 13 MFCC features were extracted with `librosa.feature.mfcc()`.
- The extracted MFCCs were normalized later to ensure consistency across samples for the model.

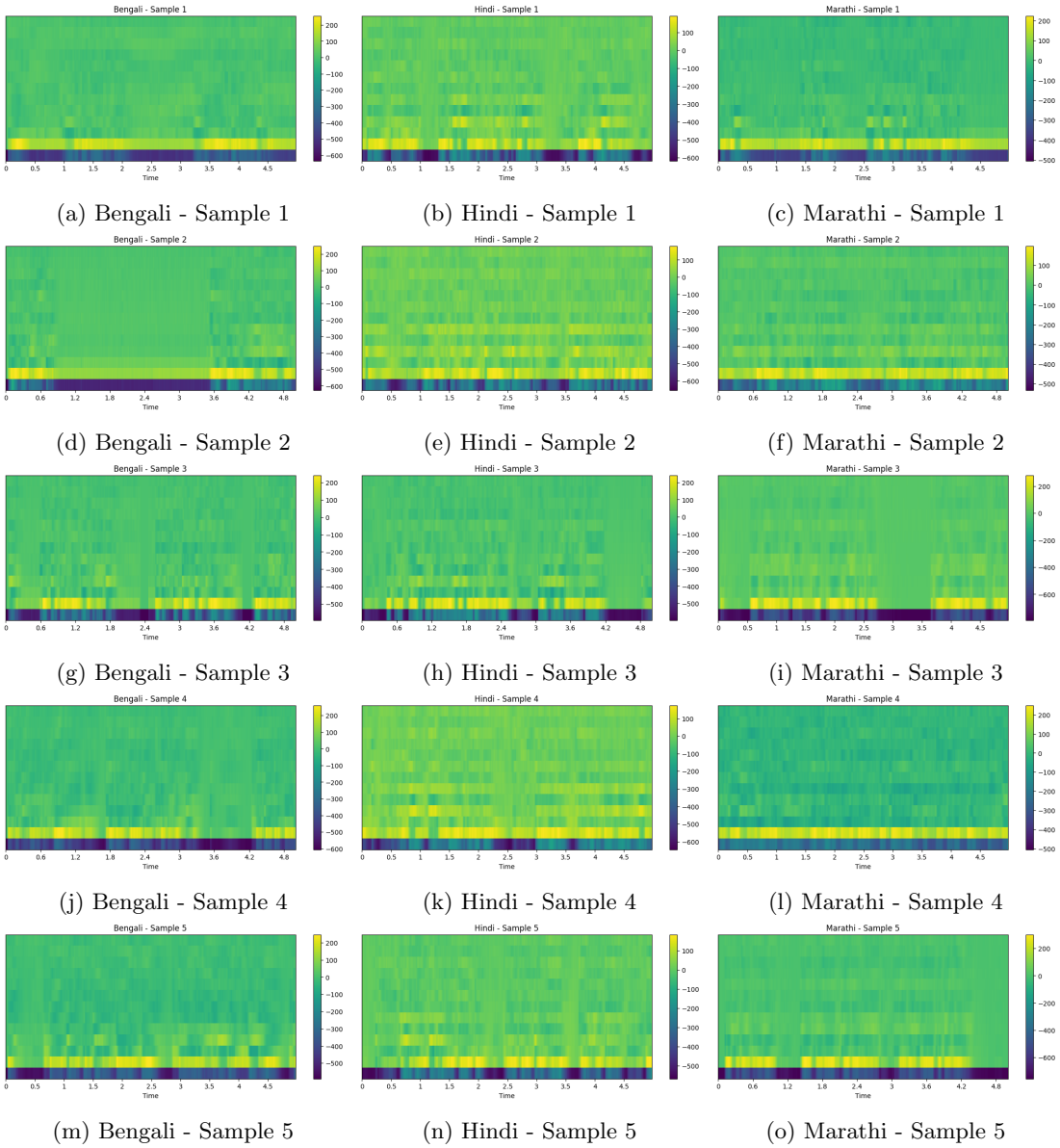
1.2.3 MFCC Visualization

MFCC spectrograms were generated to visualize the temporal patterns of the audio samples. The following figures present MFCC spectrograms for Bengali, Hindi and Marathi samples:



(a) Language 1: Bengali (b) Language 2: Hindi (c) Language 3: Marathi

Figure 3: MFCC Spectrograms for Three Languages



(a) Bengali - Sample 1 (b) Hindi - Sample 1 (c) Marathi - Sample 1

(d) Bengali - Sample 2 (e) Hindi - Sample 2 (f) Marathi - Sample 2

(g) Bengali - Sample 3 (h) Hindi - Sample 3 (i) Marathi - Sample 3

(j) Bengali - Sample 4 (k) Hindi - Sample 4 (l) Marathi - Sample 4

(m) Bengali - Sample 5 (n) Hindi - Sample 5 (o) Marathi - Sample 5

Figure 4: MFCC Spectrograms for Five Samples of Bengali, Hindi, and Marathi

1.2.4 Comparison of MFCC Spectrograms

The visual comparison of MFCC spectrograms across different languages reveals observable differences in spectral patterns. Certain languages exhibit denser formant structures, while others display more dispersed energy distributions.

- **Energy Distribution**: All three spectrograms show concentrated energy in the lower MFCC bands (near the bottom), indicating that the primary speech information lies in lower frequencies, which is common for human speech. Hindi shows stronger yellow bands, suggesting a more pronounced formant structure or higher energy in certain phonemes. Marathi and Bengali, on the other hand, shows relatively lower energy in these regions.

- **Temporal Variation**: Hindi exhibits more variation across time, with rapid shifts in intensity, indicating more dynamic acoustic patterns. This is visible in the more frequent alternating patterns of bright and dark regions in the Hindi spectrograms. Bengali samples consistently show strong energy in the lowest coefficients with relatively uniform patterns across time. Hindi samples display more variation in the middle coefficient ranges with distinctive horizontal striping patterns. Marathi samples appear to have a more balanced distribution of energy with specific temporal patterns

- **Language-Specific Differences**:
  Bengali: Shows strong energy bursts at the beginning, indicating prominent voiced segments or aspirated sounds, which are characteristic of Bengali phonetics.
  Hindi: Displays a more even distribution but with occasional high-energy spikes, which may reflect the diverse vowel-consonant transitions.
  Marathi: Exhibits a flatter structure with less variation, indicating a more uniform acoustic profile.
  Hindi spectrograms generally show more temporal variation (more yellow-green patterns across time) compared to Bengali and Marathi. Bengali appears to have the most concentrated energy in the lowest coefficients (bright yellow band at bottom). Marathi seems to have more moderate energy distribution with fewer extreme values.

1.2.5 Statistical Analysis

To quantify these differences, statistical measures such as the mean and variance of the MFCC coefficients were computed.

- The first MFCC coefficient (MFCC-0) shows the largest negative mean values for all three languages, with Bengali having the most negative value (-412.48), followed by Marathi (-393.03) and Hindi (-373.44). This suggests Bengali speech typically has lower overall energy or amplitude compared to the other languages.

- Hindi demonstrates the highest variance in MFCC-0 (14962.15), indicating greater variability in overall energy distribution compared to Bengali and Marathi. Intuitively, this means Hindi speakers likely use more dynamic volume changes within their speech.

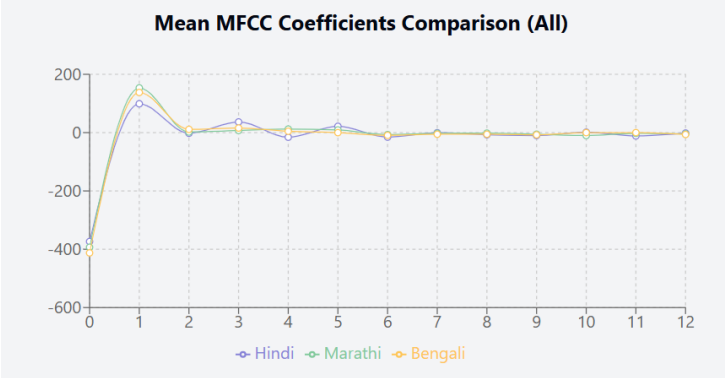
- Marathi exhibits distinctly higher values for MFCC-1 (153.31) compared to Bengali (137.79) and Hindi (98.97), suggesting unique spectral slope characteristics. This likely reflects Marathi's particular balance between low and high frequencies, giving it a distinct tonal quality.

- Bengali shows significantly higher positive mean value for MFCC-2 (11.14) compared to Marathi (2.18) and Hindi (-2.23), potentially indicating distinctive phonetic features. This could relate to Bengali's characteristic nasalization or vowel qualities that create unique resonance patterns.

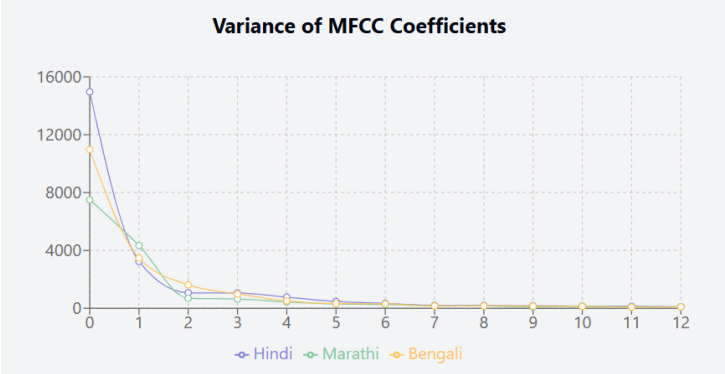
- Hindi displays a notably higher mean for MFCC-3 (36.72) compared to Bengali (15.74) and Marathi (7.45). This might reflect Hindi's distinctive retroflex consonants that create specific acoustic patterns not as prominent in the other languages.

- Marathi is the only language showing positive values for MFCC-4 (12.13) and MFCC-5 (9.21), while Hindi alternates between negative and positive values across these coefficients. This suggests Marathi has a more consistent articulation pattern in certain frequency ranges, possibly related to its unique consonant-vowel combinations.

- Bengali exhibits higher variance in MFCC-2 (1620.27) compared to Hindi (1071.18) and Marathi (692.36), suggesting greater variability in certain phonetic elements. This may reflect Bengali's wider range of vowel sounds and their modifications in different contexts.
- Hindi consistently shows higher variance across multiple coefficients, indicating more diverse acoustic patterns overall. This aligns with Hindi's large phonetic inventory and the significant regional variations in its pronunciation.
- The distinctive patterns in mean values across coefficients create unique "acoustic signatures" for each language that could be leveraged for automatic language identification. These patterns are like fingerprints that distinguish one language from another based on their characteristic sounds.
- Lower coefficients (0-6) show greater variance and more significant differences between languages compared to higher coefficients (7-12), suggesting that fundamental acoustic properties carry more language-specific information. This means the basic sound structures of these languages differ more than their finer acoustic details.



(a) MFCC means of all 3 languages



(b) MFCC variances of all 3 languages

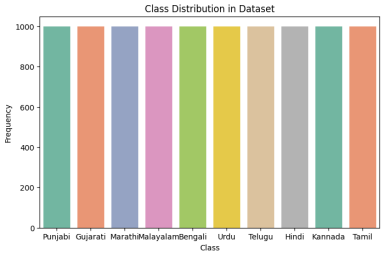
Figure 5: Mean and Variance of MFCC

1.3 Task B: Language Classification Using MFCC Features

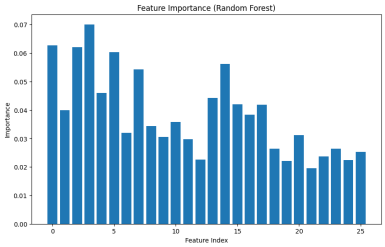
1.3.1 Model Selection and Preprocessing

A Random Forest classifier was selected to predict the language of each audio sample. 1000 samples for each languages was considered for the dataset. The model's hyperparameters are as follows:- n_estimators=100, random_state=42. Data preprocessing involved the following steps:

- The extracted MFCC features were normalized using the Min-Max Scaler.
- The dataset was split into training and testing sets using an 70:30 ratio.



(a) Distribution of all languages

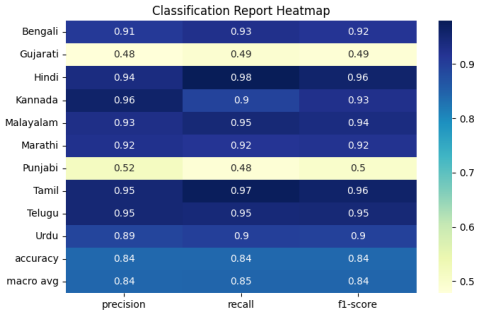


(b) Feature Importances

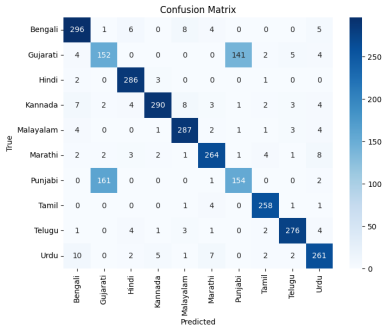
Figure 6: Distribution of Languages and Feature Importances

1.3.2 Model Training and Evaluation

The Random Forest classifier was trained on the extracted MFCC features. Performance was evaluated using accuracy, precision, recall, and the F1-score. The model achieved a good classification accuracy of 84%, indicating the discriminative power of MFCC features for language identification.



(a) Classification report



(b) Confusion Matrix

Figure 7: Model Outputs

1.3.3 Challenges in Language Classification Using MFCCs

Several challenges arise when using MFCC features for language classification. The model is unable to correctly distinguish between Punjabi and Gujarati samples. 141 Gujarati samples are misclassified as Punjabi, and 161 Punjabi samples are misclassified as Gujarati. Here, the possible causes are discussed:

- **Speaker Variability:** Differences in speaker pitch and articulation can affect MFCC consistency across samples. Regional accents within Punjab and Gujarat can vary. Additionally, bilingual speakers in these regions may introduce pronunciation patterns from neighboring languages.
- **Background Noise:** Noisy environments may introduce artifacts in the MFCC representations, reducing model accuracy.
- **Regional Accents:** Variations in pronunciation within the same language can overlap with patterns observed in other languages, making classification more difficult.
- **Variability in Tonality**: MFCCs don't take into account the pitch of audio samples. This is a major problem as can be seen by the accuracies in Punjabi and Gujarati samples. The model gets confused between these two languages mainly. Punjabi is a tonal language (having pitch-based phonemic differences), while Gujarati is not. These are probably not fully captured by MFCC, which focuses primarily on spectral envelope rather than pitch contours. Moreover, both languages share a significant amount of vocabulary from Sanskrit, which could cause the model to misidentify words.

References

[1] Rubén Fraile, Nicolas Saenz-Lechon, Juan godino llorente, V Osma-Ruiz, and Corinne Fredouille. Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 61:146–52, 02 2009. doi:10.1159/000219950.