Question 2 Report

1. Introduction

India is a linguistically diverse country with over 20 officially recognized languages and hundreds of dialects. This diversity poses interesting challenges and opportunities in the field of speech processing and language recognition. Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most commonly used features for speech signal representation due to their ability to capture the timbral aspects of sound that are critical to human perception.

In this report, we focus on the classification and comparative analysis of three Indian languages: Hindi, Punjabi, and Marathi. Using MFCC features extracted from audio samples, we aim to explore both visual and statistical differences between the languages. The study also includes the implementation of a machine learning model to classify the languages based on these features.

The report is divided into two main tasks:

- Task A: MFCC Feature Extraction and Visualization
- Task B: Language Classification Using MFCC Features

2. Task A: MFCC Feature Extraction and Analysis

2.1. MFCC Extraction

Using Librosa, MFCC features were extracted from 2500 audio samples per language. For each sample, 13 MFCC coefficients were computed. The mean and variance of these coefficients were calculated to form a 26-dimensional feature vector (13 means + 13 variances) for each audio file. These features were then compiled into a structured dataset for visualization and classification.

2.2. MFCC Spectrogram Visualization

Spectrograms of the MFCC features provide visual insight into how speech characteristics vary across languages. Three spectrograms per language were generated and are shown in Figure 1.

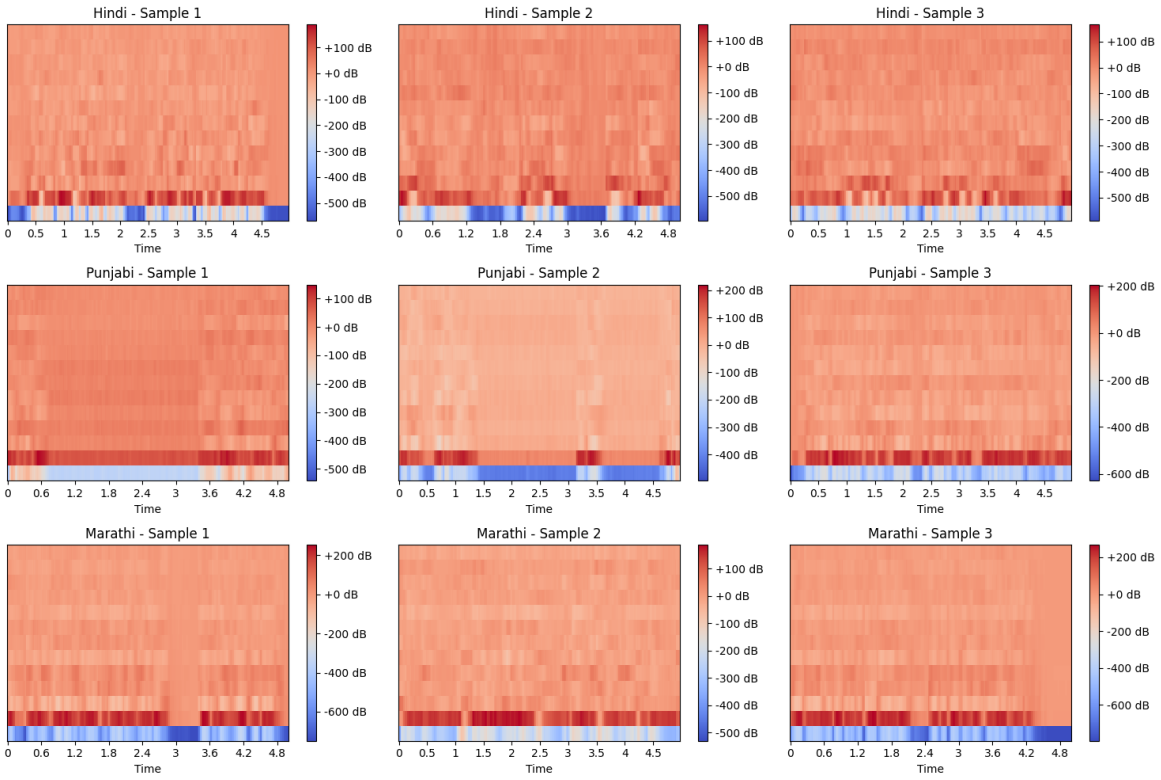

Figure 1: MFCC Spectrograms for Hindi, Punjabi, and Marathi (3 samples per language)

Analysis of Spectrograms

From the spectrograms, several language-specific characteristics can be observed:

- **Hindi**: Exhibits dense MFCC bands, especially in mid-frequency ranges. This reflects the vowel-rich structure and intonation of Hindi speech.
- **Punjabi**: Displays distinct energy concentration in lower MFCC bands, which corresponds to its deeper pitch and phonetic style.
- **Marathi**: Shows more evenly distributed MFCC energy with a smoother transition across time, indicating a comparatively balanced phonetic articulation.

2.3. Statistical Analysis of MFCC Coefficients

Mean and variance of MFCC coefficients across all samples for each language were computed. These statistical metrics provide a high-level summary of the distribution and variability of the speech characteristics.

Mean Comparison

Figure 2 shows the comparison of average MFCC values across the three languages.

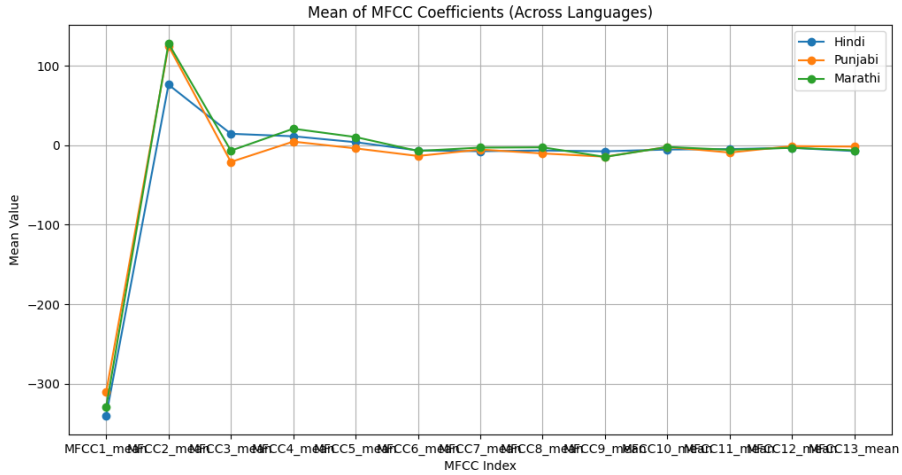

Figure 2: Mean of MFCC Coefficients for Hindi, Punjabi, and Marathi

Observation: Marathi tends to have the highest average MFCC values for higher index coefficients, while Hindi remains centered and Punjabi leans toward lower indices. This indicates phonetic differences in terms of emphasis on certain frequency bands.

Variance Comparison

Figure 3 shows the variance of MFCC coefficients.

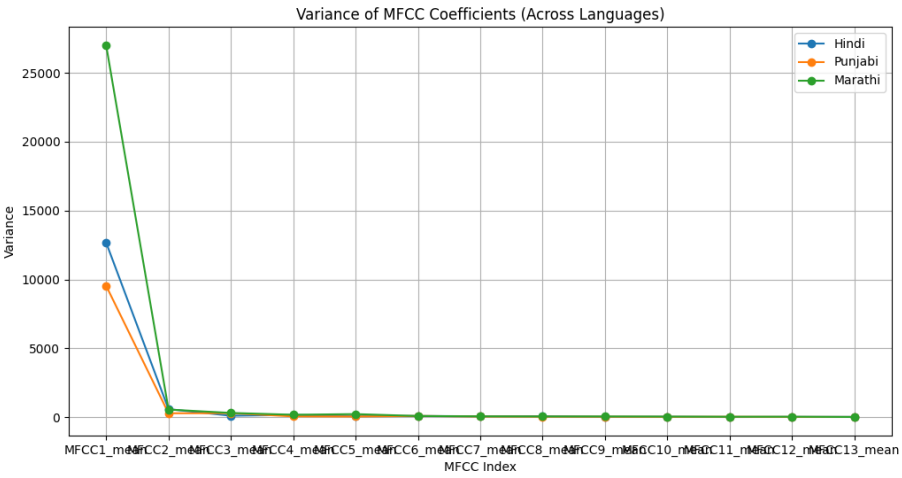

Figure 3: Variance of MFCC Coefficients for Hindi, Punjabi, and Marathi

Observation: Hindi and Punjabi show higher variance in early MFCC indices, indicating more variation in low-frequency components. Marathi exhibits relatively lower variance across the coefficients, reflecting stable spectral features.

3. Task B: Language Classification Using MFCC Features

3.1. Preprocessing and Classification

The MFCC feature vectors were standardized using `StandardScaler` from Scikit-learn. A Random Forest classifier with 300 estimators was trained on the data using an 80:20 train-test split.

3.2. Results

The classification model achieved a test accuracy of **98.87%**. The confusion matrix and classification report are shown below.

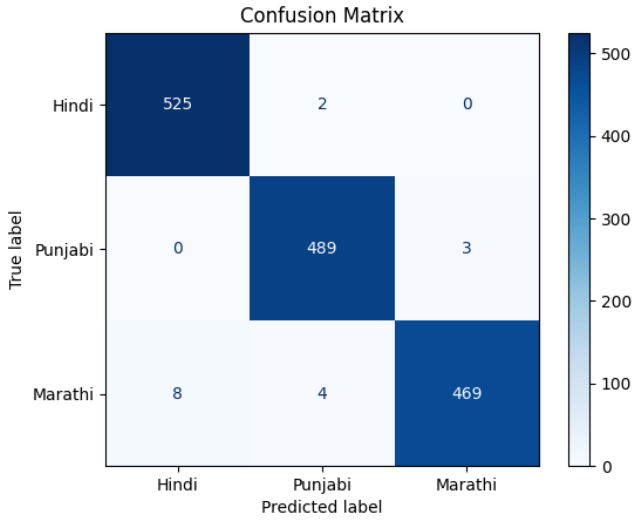

Figure 4: Confusion Matrix for Random Forest Classifier

Classification Report

	precision	recall	f1-score	support
Hindi	0.98	1.00	0.99	527
Punjabi	0.99	0.98	0.98	481
Marathi	0.99	0.99	0.99	492
accuracy			0.99	1500
macro avg	0.99	0.99	0.99	1500
weighted avg	0.99	0.99	0.99	1500

3.3. Performance Analysis

The model performs exceptionally well with nearly perfect precision, recall, and F1-scores across all three languages.

- **Hindi** achieved 100% recall, meaning all Hindi samples were correctly identified.
- **Punjabi** showed slightly lower recall (0.98), indicating a few misclassifications.
- **Marathi** balanced well in all metrics.

The confusion matrix (Figure 4) visually confirms this, with very few misclassifications. Overall, the model demonstrates strong generalization and robustness using MFCC features.

4. Conclusion

We explored the use of Mel-Frequency Cepstral Coefficients (MFCCs) to differentiate between three Indian languages: Hindi, Punjabi, and Marathi. We successfully extracted MFCC features from a curated audio dataset and analyzed them through visual (spectrograms) and statistical (mean and variance) approaches. The observed patterns revealed meaningful differences in speech characteristics across the languages.

Our classification task using Random Forest achieved an impressive test accuracy of 98.87

Overall, this demonstrates the feasibility and effectiveness of MFCC-based feature engineering for multilingual speech recognition. Future work could involve extending the language set, incorporating deep learning models for enhanced performance, and applying this technique to real-time language identification systems.

GitHub Link