MFCC Feature Extraction and Comparative Analysis of Indian Languages

Abhishek Kumar Singh
Indian Institute of Technology, Jodhpur
Jodhpur 342037, India
m23csa503@iitj.ac.in

Abstract

Spectrograms play a crucial role in speech analysis, offering a time-frequency representation of an audio signal. They illustrate how energy is distributed across frequencies over time, helping to analyze pitch, harmonics, and noise. However, spectrograms may contain frequencies outside the human hearing range. To address this, Mel-Frequency Cepstral Coefficients (MFCCs) are employed. These features, designed to replicate human auditory perception, provide valuable insights into vocal tract changes and formant variations. MFCCs are particularly effective for speech and language recognition as they emphasize perceptually significant information.

This study leverages MFCCs to analyze and classify audio samples from a dataset containing ten Indian languages. The research is divided into two primary tasks. In Task A, MFCC features are extracted and visualized for three languages—Hindi, Marathi, and Bengali—to compare spectral patterns and identify distinguishing characteristics. Statistical measures such as mean and variance are computed to quantify differences across languages.

In Task B, MFCCs are extracted from 1,000 samples per language to train a machine-learning model for language classification. A Random Forest Classifier is utilized for its interpretability and effectiveness in handling complex audio features. Preprocessing steps include min-max normalization and a 70-30 train-test split to ensure reliable evaluation.

The study also explores challenges in language identification using MFCCs, including speaker variability, regional accents, and background noise. Additionally, tonal languages may be underrepresented due to MFCCs discarding pitch information. The implementation and results are available at GitHub Repository

Contents

1 Question 2

1.1 Introduction

Speech signals are inherently complex, consisting of both source-related and filter-related components. The source component originates from the vibration of the vocal cords, producing variations in pitch and tone. Meanwhile, the filter component, determined by the vocal tract (comprising the pharynx, oral cavity, and nasal cavity), shapes the produced sound. To analyze and classify speech, it is essential to extract meaningful features that capture these characteristics while disregarding irrelevant variations.

Cepstral analysis decomposes an audio signal into its source and filter components, enabling the separation of pitch-related features from phonemes. One of the most widely used representations for speech analysis is the Mel-Frequency Cepstral Coefficients (MFCCs), which are indirectly derived from the cepstrum. These coefficients are designed to mimic human auditory perception by focusing on perceptually significant frequency ranges. Due to their ability to represent vocal tract characteristics effectively, MFCCs are widely applied in speech recognition, speaker verification, and language identification.

1.1.1 From Time Domain to Cepstrum

A speech signal $s(n)$ can be represented as a convolution of:

- **Excitation Signal** $e(n)$: A periodic signal generated by vocal cord vibrations, determining pitch.
- **Vocal Tract Response** $h(n)$: A filter that shapes the excitation signal, forming distinct phonemes.

Mathematically, this relationship is expressed as:

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

Applying the Discrete Fourier Transform (DFT) converts the convolution into a multiplication in the frequency domain:

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

where $S(k)$ is the spectrum of the speech signal, $E(k)$ is the spectrum of the excitation signal, and $H(k)$ represents the vocal tract response.

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

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

This transformation facilitates easier manipulation of source and filter contributions. Applying the Inverse DFT (IDFT) yields the cepstrum:

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

In the cepstral domain:

- Low quefrencies represent slow variations linked to vocal tract information.
- High quefrencies capture rapid variations associated with pitch.

Since speech recognition relies mainly on vocal tract shape, the low-quefrency region is retained while higher quefrencies are discarded.

1.1.2 From Cepstrum to MFCCs

While the cepstrum provides useful information, it does not align with human auditory perception, which is nonlinear. The Mel scale models this perception and is used to compute MFCCs through the following steps:

Step 1: Pre-emphasis - High frequencies are amplified using a filter:

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

where α is typically set to 0.97.

Step 2: Framing and Windowing - The signal is divided into short overlapping frames (e.g., 25 ms with 10 ms overlap). A Hamming window is applied:

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

Step 3: Fourier Transform and Power Spectrum - The Fast Fourier Transform (FFT) computes the magnitude spectrum, and the power spectrum is given by:

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

Step 4: Mel Filter Bank - The frequency axis is transformed to the Mel scale using:

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

Triangular filters are applied to measure energy in each Mel band:

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

where E_m represents the energy in the m-th Mel filter.

Step 5: Logarithm of Filter Outputs - The log operation mimics human perception:

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

Step 6: Discrete Cosine Transform (DCT) - The spectral data is compressed into MFCCs:

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

Only the first 12-13 MFCCs are retained, as they capture essential vocal tract features.

1.1.3 Importance of MFCCs

MFCCs are widely used due to:

- **Human-like Perception** - The Mel scale aligns with auditory perception.
- **Source-Filter Separation** - Vocal tract features are isolated from pitch.
- **Compact Representation** - The DCT reduces redundancy for efficient machine learning processing.
- **Sound Source Differentiation** - MFCCs capture the energy distribution unique to each source, aiding in classification.

1.2 Feature Extraction and Analysis

1.2.1 Dataset Collection

The dataset comprises speech samples from ten Indian languages, sourced from Kaggle, ensuring diverse linguistic representation.

1.2.2 MFCC Extraction

MFCCs were extracted using the `librosa` library:

- Audio files were loaded using `librosa.load()` while preserving the original sampling rate.
- 13 MFCC features were computed using `librosa.feature.mfcc()`.
- Extracted features were normalized for consistency.

1.2.3 MFCC Visualization and Comparison

MFCC spectrograms were generated for Hindi, Bengali, and Marathi. Comparative analysis highlighted distinctions in energy distribution, temporal variation, and phonetic structure.

1.3 Language Classification Using MFCC Features

1.3.1 Model Selection and Training

A Random Forest classifier was trained using MFCC features from 1000 samples per language. Data preprocessing included:

- Min-Max normalization of MFCC features.
- 70:30 train-test split.

The model achieved 84% accuracy, demonstrating MFCC effectiveness in language identification.

1.3.2 Challenges in Classification

Key challenges included:

- Speaker variability affecting MFCC consistency.
- Background noise introducing distortions.
- Regional accents influencing phonetic patterns.
- MFCC limitations in capturing tonal language characteristics.

References

[1] Rub´en 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-frequencycepstral 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.