MFCC analysis in Audio Engineering

Author: Thiksiga Ragulakaran (E/18/362)
Course: Audio Engineering
University: University of Peradeniya

1. Introduction

Mel Frequency Cepstral Coefficients (MFCCs) are widely used in audio signal processing, particularly for tasks such as speech and music analysis. MFCCs provide a concise representation of the audio signal's spectral envelope, capturing essential features while reducing the dimensionality of the data. This report explains the mathematical basis of MFCCs, their extraction process, and practical implementation using Python. Additionally, it covers how to interpret MFCC plots and their applications in various audio processing tasks.

2. Mathematical Background

The extraction of MFCCs involves several key steps:

1. **Pre-emphasis**: Enhances higher frequencies in the audio signal by applying a high-pass filter. This step compensates for the fact that human hearing is less sensitive to higher frequencies.

   $y(t)=x(t)-\alpha \cdot x(t-1)$

   where x(t) is the input signal, y(t) is the output signal, and α (typically 0.97) is the pre-emphasis coefficient.

2. **Framing**: Splits the signal into small overlapping frames to analyze short segments of the audio. Each frame typically lasts 20-40 milliseconds, allowing for time-localized analysis.
3. **Windowing**: Applies a window function (usually a Hamming window) to each frame to minimize spectral leakage during the Fourier Transform. The window function is defined as:

$$w(n) = 0.54 - 0.46.\cos(\frac{2\pi n}{N-1})$$

4. **Fast Fourier Transform (FFT)**: Converts the time-domain signal into the frequency domain to obtain the magnitude spectrum for each frame. The power spectrum is computed as:

$$P(f) = \frac{|X(f)|^2}{N}$$

   where X(f) is the FFT of the windowed frame, and N is the number of FFT points.

5. **Mel Filter Banks**: Applies a series of triangular filters spaced on the Mel scale to the power spectrum to approximate human auditory perception. The Mel scale is computed as:

$$m = 2595.\log_{10}\left(1 + \frac{f}{700}\right)$$

   where f is the frequency in Hertz.

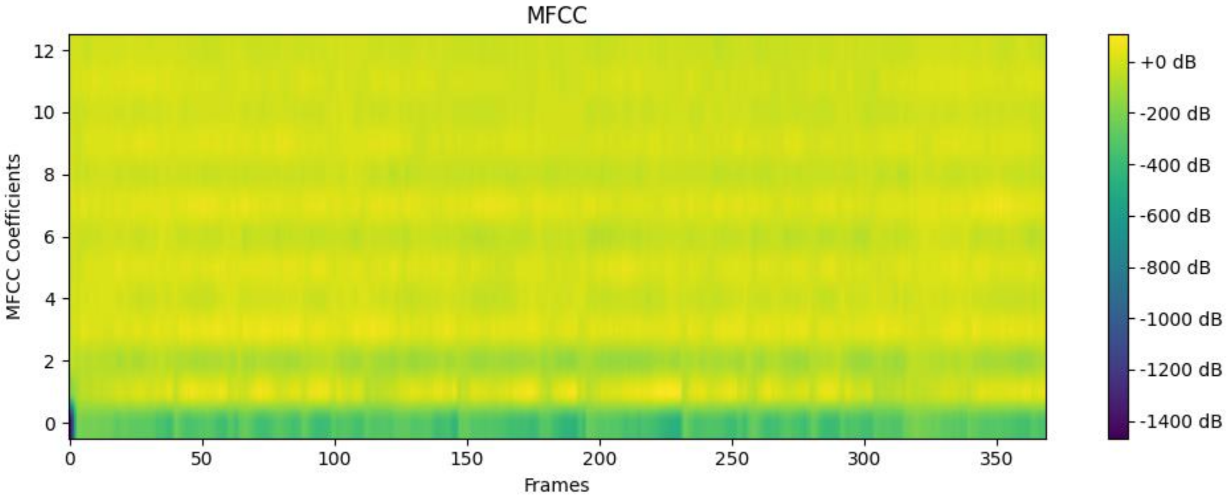
6. **Logarithm of Filter Bank Energies**: Takes the logarithm of each Mel-scaled filter bank output to mimic the human ear's sensitivity to sound intensity changes.
7. **Discrete Cosine Transform (DCT)**: Applies DCT to the log filter bank energies to produce the MFCCs, decorrelating the features and compressing them into a smaller set of coefficients. The DCT is defined as:

$$C(k) = \sum_{n=0}^{N-1} \left(logF(n)\right). \cos\left(\frac{\pi k(2n+1)}{2N}\right)$$

where C(k) is the k-th MFCC coefficient, and F(n) is the log filter bank energy.

3. Visualization and Interpretation

The MFCC plot visualizes the temporal evolution of the Mel Frequency Cepstral Coefficients over time. Each frame on the x-axis represents a short segment of the audio, while the y-axis shows the MFCC coefficients (typically 13). The color intensity indicates the magnitude of each coefficient in dB.



- **Patterns and Shapes**: Different patterns in the plot correspond to different sounds, such as phonemes in speech or notes in music.
- **Energy Distribution**: The color intensity helps identify which frequency ranges dominate the sound over time.

From the MFCC plot, it can be observed that:

1. **Coefficient 0 Corresponds to Vocals**:
    - At the beginning of the song, there is a noticeable intensity in the lower MFCC coefficients, particularly coefficient 0. This indicates that the energy is concentrated in the lower frequencies, which often corresponds to the presence of vocal sounds without much instrumental accompaniment.
    - Since coefficient 0 usually captures the overall loudness or energy of the signal, its prominence at the start suggests that the vocals are the dominant source of sound in this portion.
2. **Introduction of Background Instruments**:

   - As time progresses, the MFCC plot shows changes in the higher coefficients, indicating that additional frequency components are becoming more prominent. These changes reflect the introduction of background instruments that contribute more complex frequency content to the audio signal.
   - The smooth transitions and variations in the coefficients over time represent a gradual blending of vocals with the accompanying instrumental sounds, illustrating how both elements co-exist and evolve throughout the song.
3. **Dynamic Changes Over Time**:
   - The MFCC plot captures dynamic changes in the spectral content of the song. Variations in the color intensity suggest fluctuations in the energy levels of different frequency bands, reflecting the natural dynamics of musical compositions, such as changes in tempo, intensity, or the entrance of different instruments.

Conclusion:

The MFCC plot effectively shows how different audio components—like vocals and instruments—are distributed across time and frequency. The observation that coefficient 0 primarily represents vocals at the beginning of the song helps to understand the song's structure and composition, revealing how the vocals are initially dominant and gradually blend with instrumental sounds throughout the piece. This analysis can be further used in applications such as vocal isolation, music classification, and audio content analysis.