Mel-Frequency Cepstral Coefficients (MFCCs): Computation Steps

1. Pre-emphasis

- Apply a high-pass filter to boost high frequencies and balance the spectrum.

2. Framing

- Divide the continuous audio signal into short,overlapping frames (e.g.25 ms windows with a 10 ms hop).
- Each frame is assumed quasi-stationary,allowing spectral analysis.

3. Windowing

- Multiply each frame by a Hamming window to reduce edge discontinuities.

4. Fast Fourier Transform (FFT)

- Compute the N-point FFT for each windowed frame to obtain its magnitude spectrum.
- Result: a set of complex values; take absolute value to get magnitude.

5. Power Spectrum

- Square the magnitude spectrum and normalize by frame length to get power for each frequency bin.

6. Mel Filter Bank

- Map the linear frequency bins to the mel scale (perceptual scale where pitch ≈ mel frequency).
- Create a series of overlapping triangular filters spaced uniformly on the mel scale.Typical count: 26 filters.
- Apply each filter to the power spectrum,summing the energy under the triangle.

7. Logarithm

- Take the natural log of each filter-bank energy to mimic human loudness perception (log amplitude).

8. Discrete Cosine Transform (DCT)

- Compute the DCT of the log filter-bank vector to decorrelate and compress information.
- Keep the first M coefficients (e.g.M=13) as the MFCCs; discard the rest.

9. Liftering (Optional)

- Multiply the MFCC sequence by a window (e.g.sinusoidal lifter) to emphasize higher coefficients and de-emphasize lower ones.

10. Delta and Delta-Delta Features

- Compute the first and second derivatives (delta and acceleration) of the MFCC sequence over time.
- These capture dynamic changes in the spectral envelope.

Result: For each frame,you obtain a vector of M static MFCCs plus delta and delta-delta coefficients (commonly 39 total features). Summarizing these frame-level coefficients via mean and standard deviation yields a fixed-length representation for an entire audio clip. This MFCC pipeline transforms raw time-domain signals into compact,perceptually informed features ideal for speech analysis and classification tasks.