NLP/ASR

Unit 5: Advanced Topics in ASR/NLP

5.2.2

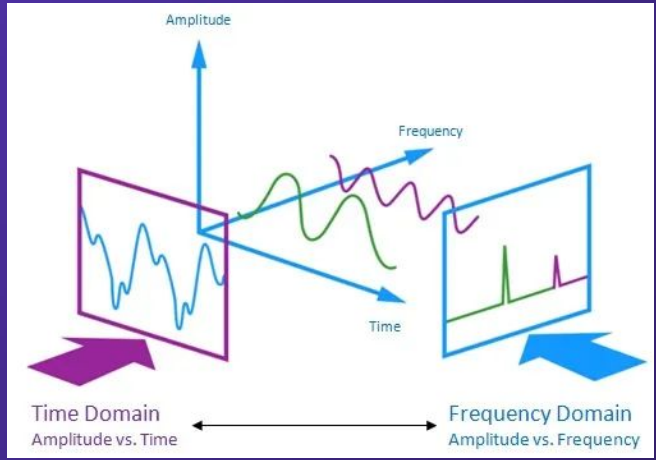
Advanced Audio Processing

MFCC, Filter banks, and Feature extraction methods

TIL-AI
TODAY I LEARNED·AI

Importance of Feature Extraction

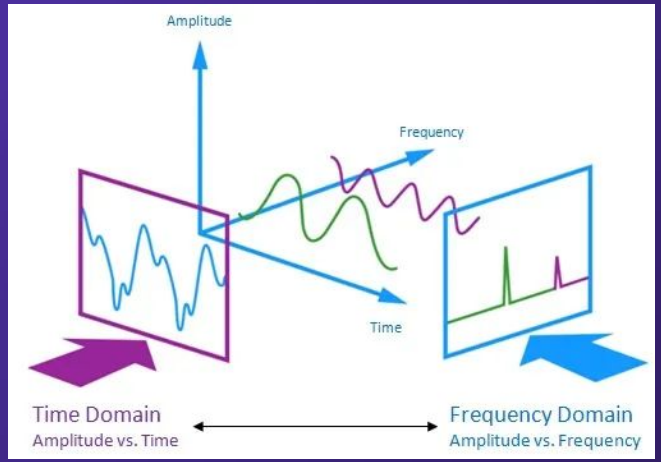
- Raw audio signals are complex and contain a mix of relevant and irrelevant information
- Feature extraction transforms raw audio into compact, meaningful representations that highlight speech patterns
- Effective features make it easier for ASR models to distinguish between different sounds and words
- The quality of feature extraction directly impacts the overall accuracy of an ASR system



TIL-AI
TODAY I LEARNED-AI

From Time to Frequency Domain

- ASR systems rely heavily on transforming speech signals from time domain to frequency domain to reveal the underlying frequencies that make up the speech
- Time domain: Speech is represented as a waveform where the amplitude (intensity) of the signal is plotted over time
- Discrete Fourier Transform (DFT) is used to decompose a time-domain signal into individual frequencies and amplitudes

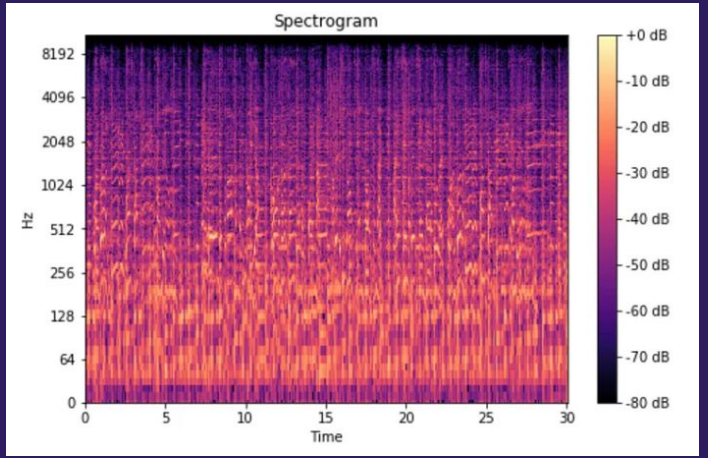


- The output of the DFT tells us how much energy is present at each frequency in the original speech signal

TIL-AI
TODAY I LEARNED-AI

Spectrograms

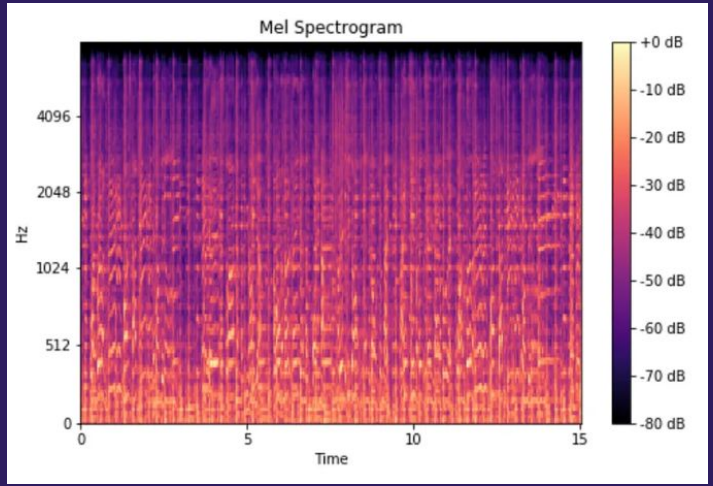
- Visual representation of the spectrum of frequencies in a sound as they vary with time
- Shows how different frequencies appear, disappear, or change intensity over time in an audio signal



TIL-AI
TODAY I LEARNED AI

Mel-Spectrograms

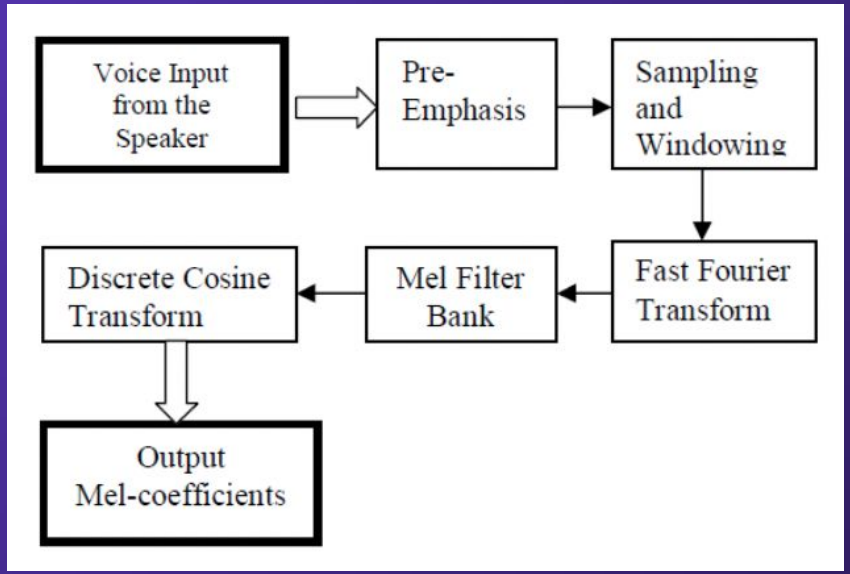
- Type of spectrogram where the frequency scale is converted to the Mel scale
- Mel scale more closely approximates human auditory system's response than the linear frequency scale; making it more effective for audio-related tasks in human speech and music



TIL-AI
TODAY I LEARNED AI

Mel-Frequency Cepstral Coefficients (MFCCs)

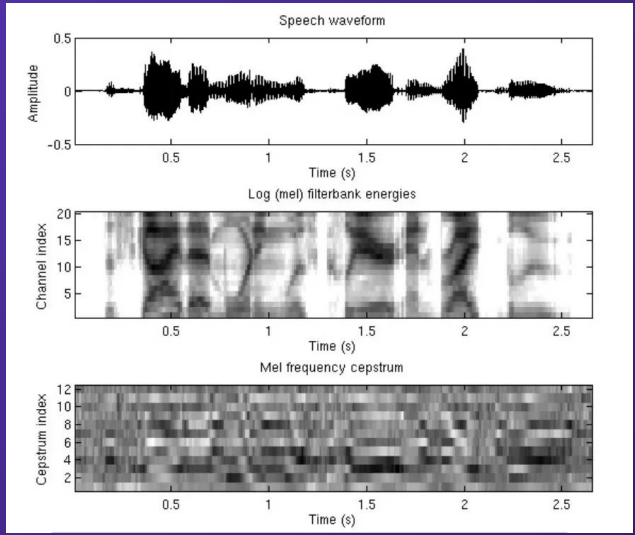
- MFCCs are widely used in ASR for their efficacy in capturing the essential characteristics of spoken language
- MFCCs are designed to closely mimic the way the human ear processes sound frequencies
- The Mel scale is a perceptual scale, meaning frequencies humans perceive as equally spaced are not linearly spaced on the Hertz scale



TIL-AI
TODAY I LEARNED-AI

Filter Banks

- Filter banks are collections of bandpass filters that split the audio spectrum into distinct frequency bands
- Each filter focuses on a specific frequency range
- Mel-scale filter banks are a specialized type designed for speech recognition



TIL-AI
TODAY I LEARNED-AI

Other Feature Extraction Techniques

- Linear Predictive Coding (LPC): LPC models the speech signal by predicting a sample based on a linear combination of past samples

- Perceptual Linear Predictive (PLP): An improvement on LPC, PLP incorporates aspects of human auditory perception (like critical band analysis and equal-loudness pre-emphasis) for more robust speech representation