***MFCC***

Mel-frequency cepstral coefficients

They provide a compact and effective representation of the spectral charateristics of audio signals making them versaltile for a wide range of audio processing tasks like

a.Speech Recognition

b.Speal Recognition

c.Audio Classification

d,Keyword Spotting

e.Emotion Recognition

f.Noise Reduction

g.Speaker Diarization

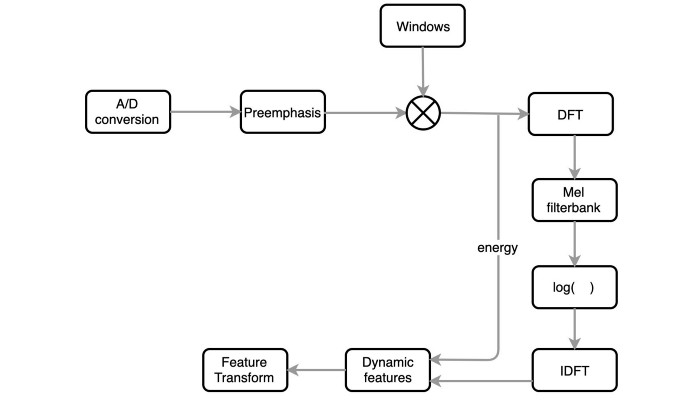
h.Voice Acitivity Detection(VAD)

Their application continues to expand with advancements in machine learning and signal processing techniques.

Basic Working-

They are derived by first transforming the raw audio signal into a frequency domain using a technique like the Discrete Fourier Transform (DFT), and then applying the mel-scale to approximate the human auditory perception of sound frequency. Finally, cepstral coefficients are computed from the mel-scaled spectrum.

They are useful as they are able to distinguish important audio features from noise, due to this they are used for a wide application.

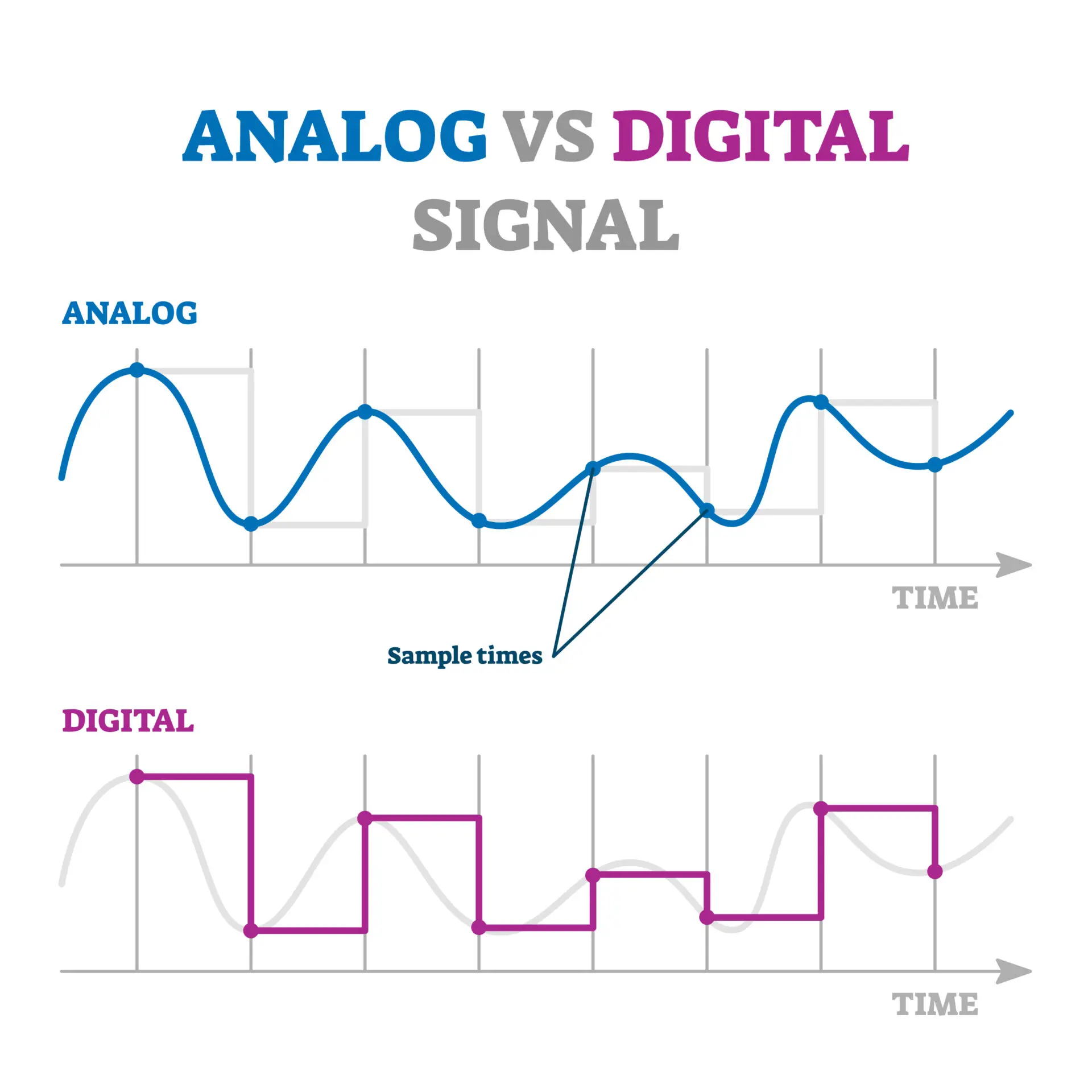


Basics

* Sound: Vibrations that travel through the air or another medium and can be heard when they reach a person's or animal's ear.
* Audio Signal: A representation of sound, typically as a waveform (amplitude vs. time).
* Frequency: The number of vibrations per second, measured in Hertz (Hz).
* Amplitude: The height of the waveform, representing the loudness of the sound.
* Sampling: Converting a continuous audio signal into a discrete signal by measuring the amplitude at regular intervals.
* Quantization: Mapping the amplitude of each sample to a fixed set of values (bit depth).
* Fourier Transform Converts a time-domain signal into its frequency components.
* Mathematics: The Fourier Transform of a signal x(t) is X(f)=∞∫−∞​x(t)e^−j2πft dt.
* Spectogram is a visual representation of the spectrum of frequencies in a signal as it varies with time.Shows how the energy of different frequencies evolves over time.
* Mel Scale:
* Perceptual Scale: Reflects how humans perceive pitch.
* Conversion: m=2595log10​(1+700f​) where m is the Mel frequency and f is the regular frequency in Hz.
* Filter Banks:
* Purpose: Apply a series of band-pass filters to the signal to emphasize certain frequency ranges.
* Mel Filter Bank: A set of filters spaced according to the Mel scale, typically implemented as triangular filters.
* Cepstrum is the result of taking the inverse Fourier transform of the logarithm of the spectrum of a signal. Separates the source and the filter characteristics in speech signals.

MFCC Calculation:

* Step-by-Step Process:
  1. Pre-emphasis: Apply a high-pass filter to emphasize higher frequencies.
  2. Framing: Divide the signal into overlapping frames.
  3. Windowing: Apply a window function to each frame.
  4. Fast Fourier Transform (FFT): Compute the spectrum of each frame.
  5. Mel Filter Bank: Apply the Mel filter bank to the spectrum.
  6. Logarithm: Take the logarithm of the filter bank energies.
  7. Discrete Cosine Transform (DCT): Convert the log energies to the cepstral domain.
  8. Liftering (optional): Apply a sinusoidal lifter to smooth the coefficients.

*1*.***Conversion(a to d).***

To be processed by the machine the audio signal is converted from analog to digital format with a sampling frequency of 8khz or 16khz (Sampling at 16 kHz captures frequencies up to 8 kHz, providing better audio quality. It is often used in applications requiring better voice clarity, such as high-quality VoIP and some speech recognition systems.The range 4khz mostly covers the majority of the human voices fundamental frequencies and some harmonics also but the higher range is kep with reapect to Nyquist Theorem)

*2.****Preemphasis***

* Higher Frequencies tend to have lower energy than lower frequencies, especially in voiced segements like vowels .So preemphasis increases the magnitued of the higher frequencies making them more promiment in the feature extraction process.
* By amplifying higher frequencies, preemphasis helps in better detection of phonetic elements, improving the overall performance of speech recognition models.
* Preemphasis is done by the first order high-pass filter as given below(first order means that the filter just has only one step to decide which frequencies to let through and frequencies to be blocked)

*3.****Windowing.***

* The audio is broken down into many segments and a mathematcal function is applied on them called a window function and this process is known as Windowing.This function shapes the segment to reduce the impact of sudden changes at its boundaries.
* When the audio is chooped into segments it can create discontinuties, due to these discontinutiesit can lead to unwanted artifcats in the frequency analysis such as noise and supurion frequecies , windowing reduces these artifacts .
* The audio signals are segemented into overlapping segments which are 25 ms long ans 10 ms apart.
* This function shapes the segment to reduce the impact of sudden changes at its boundaries, and from each segment 39 features are extracted.
* Commong windowing functions are Hamming and Hanning.

4.***DFT(discreate fourier transform).***

DFT simplifies the analysis by breaking down the signal into its constituent frequencies, making it easier to identify specific tones, harmonics, and patterns within the audio.

It transforms each segment of the audio signal into the frequency domain, where further feature extraction techniques can be applied to capture relevant characteristics.

The output of DFT, often represented as a spectrogram, provides a visual representation of how signal energy is distributed across different frequencies over time.

*5****.Mel-Filter Bank***

* Mel-Scale is a scale that approximates human auditory systems response to frequencies.It is based on emprical data showing how humans perceive pitch intervals.
* We use the Mel scale to map frequencies in Hertz (Hz) to Mel units, which correspond more closely to how humans perceive pitch differences. The formula for converting Hz to Mel is:
* Mel(f)=2595⋅log10​(1+f/700)

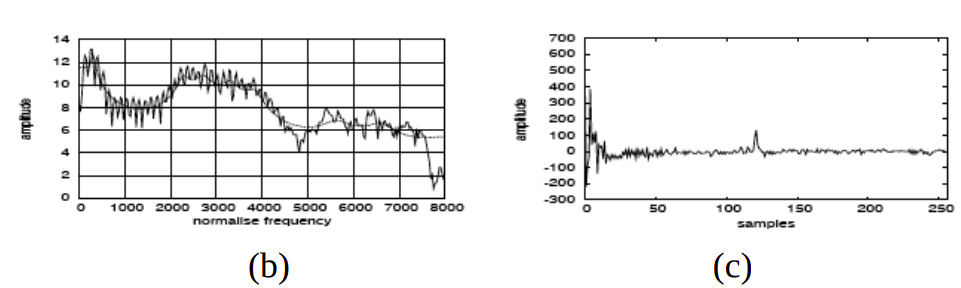
here:

* f is the frequency in Hertz (Hz) and Mel(f) is the frequency in Mel units.

***6.Applying Log***

* Human hearing is less sensitive to change in audio signal at higher energies than compared to lower energies.
* Log function also has a similar property, at a low value of input x gradient of log function will be higher but at high value of input gradient value is less, so we apply log to the output of the mel-filter to mimci the human hearing system.

***7.IDFT***

* Smallest frequency that is produced is called the fundamental frequency and all the remaining frequencies are multiples of the fundamental frequency
* The below figure shows the signal sample before and after the idft operation.
* The peak frequency at the rightmost in figure(c) is the fundamental frequency and it will provide information about the pitch and frequencies at the rightmost will provide information about the phones(phoned-distinct speech or sound that distinguishes one word from another). We will discard the fundamental frequency as it is not providing any information about phones.

13 Original MFCCs:

* Capture static spectral information of the audio signal.
* Represent the characteristics of the sound segment in terms of its frequency components.

26 Derivative Features (Delta and Delta-Delta MFCCs):

* Capture dynamic changes and temporal patterns in the audio signal.
* Provide information about how the spectral features are evolving over time, which is crucial for speech recognition tasks.
* Help in differentiating between phonemes and words by capturing transitions and nuances in speech.

6th step in depth

Logarithm Operation:

* + After obtaining the Mel-filterbank energies (which represent how much energy is in each frequency band), we take the logarithm (commonly base 10 or natural logarithm) of these energies.
  + Purpose: This logarithm operation compresses the range of energy values, emphasizing differences at lower energy levels more than at higher ones, similar to how our ears perceive sound.

1. Discrete Cosine Transform (DCT):
   * Once we have the log-filterbank energies, we apply the Discrete Cosine Transform (DCT).
   * The DCT converts these log-filterbank energies into a set of coefficients known as Mel-Frequency Cepstral Coefficients (MFCCs).
   * Purpose: The MFCCs represent the spectral characteristics of the audio signal in a compact form, capturing features like pitch, timbre, and spectral envelope.

### Simplified Steps:

* Logarithm: Compresses the energy values to better match human auditory perception.
* DCT: Transforms the log-energy values into MFCCs, which are used as features in speech recognition and other audio processing tasks.

These steps are crucial in transforming raw audio signals into a form that is more suitable for machine processing, effectively capturing the essential aspects of speech that are important for tasks like automatic speech recognition.The DCT(applined after log) in MFCC computation can be viewed as a process of combining information from segmented log Mel-filterbank energies into a more concise and informative representation—Mel-Frequency Cepstral Coefficients (MFCCs). This transformation not only condenses the spectral information but also facilitates effective modeling of speech characteristics for various audio processing applications.

IDFT (Inverse Discrete Fourier Transform): In the context of MFCC computation, IDFT is not typically applied after the logarithm operation. The confusion might arise from the terminology used.

Cepstrum: The term "cepstrum" refers to the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of a signal. This concept is related to MFCCs in that MFCCs involve taking the DCT of the logarithm of filterbank energies, which is analogous to a cepstrum-like operation but is not the same as applying IDFT directly to the log-spectrum.

IDFT is a fundamental tool in signal processing for transforming signals from frequency to time domain.