CHAPTER 2

Features Extraction

What is Feature Extraction in audio signals ?

Definition:

Feature extraction is the process of transforming raw audio signals into a set of informative and compact representations (features) that can be used as input to a machine learning model.

Purpose:

It involves selecting and computing features that capture the essential characteristics of the sound or speech signal, reducing the complexity of the data while preserving important information.

Common Features

Spectrograms:

Visual representations of the spectrum of frequencies in a signal as it varies over time.

Mel-Frequency Cepstral Coefficients (MFCCs):

Features that represent the short-term power spectrum of a sound, often used in speech and audio processing.

Chroma Features:

Capture the 12 different pitch classes and are useful in music analysis.

Zero-Crossing Rate:

The rate at which the signal changes sign, indicating the noisiness or percussiveness of a signal.

Pitch and Formants:

Fundamental frequency and resonant frequencies in speech analysis.

Why is Feature Extraction Necessary? (i)

Dimensionality Reduction:

Raw audio signals are high-dimensional and can be noisy. Feature extraction reduces the dimensionality, making it computationally feasible to process the data.

Improving Model Performance:

By extracting relevant features, models can learn more efficiently and accurately, focusing on the aspects of the signal that are most important for the task.

Noise Reduction:

It helps to filter out irrelevant or redundant information, enhancing the signal-to-noise ratio, which is crucial for tasks like speech recognition.

Why is Feature Extraction Necessary? (ii)

Capturing Important Characteristics:

Different features capture different properties of the sound, such as frequency content, temporal structure, and timbre, which are important for distinguishing between different types of sounds or speech patterns.

Facilitating Learning:

Extracted features provide a more structured and meaningful representation of the data, which can help deep learning models converge faster and achieve better generalization.

Domain-Specific Insights:

In speech processing, certain features like MFCCs are known to be effective in capturing the properties of human speech, aiding in tasks like speaker recognition and emotion detection.

LIBROSA Library

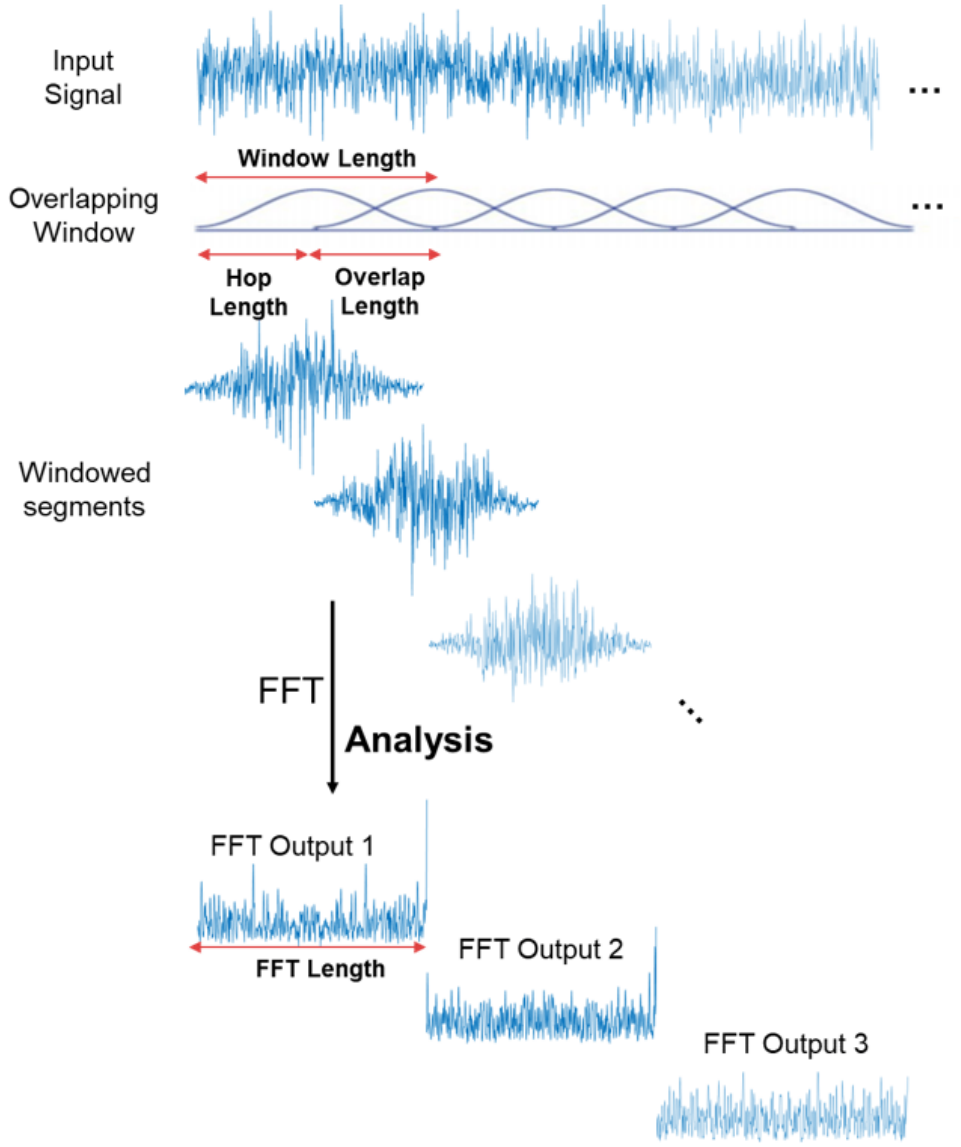
LIBROSA Library

- It is a library with audio functions and tools
- It is available for many languages including Python
- Among it, we have spectrogram, MEL & MFCC calculation functions
- It also has functions for plotting spectrograms

- Example of use:
- https://librosa.org/doc/main/auto_examples/plot_display.html

Spectrogram

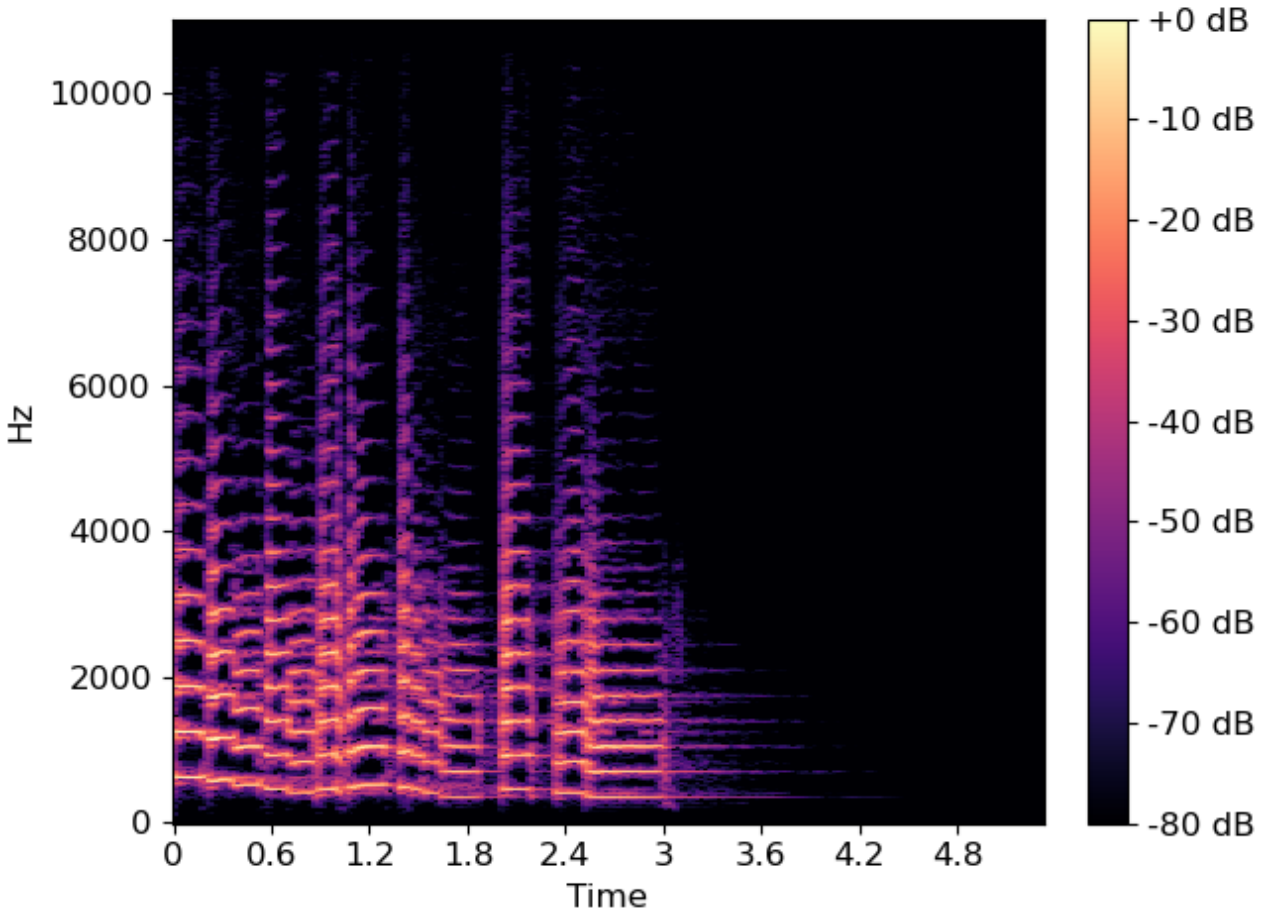
- The audio signal is divided into overlapping segments called "windows".
- Each segment is transformed from the time domain to the frequency domain using the Fourier Transform.
- This process results in a series of frequency spectra.
- These spectra are then arranged sequentially to form a time-frequency representation.
- The amplitude of each frequency component is typically displayed as a color intensity or grayscale value.



Spectrogram

- A spectrogram provides a comprehensive view of how the frequency content of a signal changes over time.
- This allows for detailed analysis of how different frequencies evolve, which is crucial for understanding complex audio signals like speech and music.
- It helps in identifying transient events (e.g., sudden noises) and patterns (e.g., musical notes, speech phonemes) that occur at specific times.
- It separates the time-domain and frequency-domain aspects of the signal, providing a clearer understanding of how different frequencies contribute to the overall signal over time.

Spectrogram



Problem with RAW spectrogram in AI

- Traditional spectrograms use a linear frequency scale, which can be too detailed in the high-frequency range where human hearing is less sensitive.
- This results in an uneven representation of frequencies, with too much detail in frequencies that are less perceptually relevant.
- Raw spectrograms can include noise and less relevant frequency information, which can negatively impact the performance of machine learning models.
- A Raw spectrogram includes almost the same samples that a PCM signal, so the number of input parameters to a NN is not reduced.
- We are looking for a Dimensionality Reduction as an input of the NN.

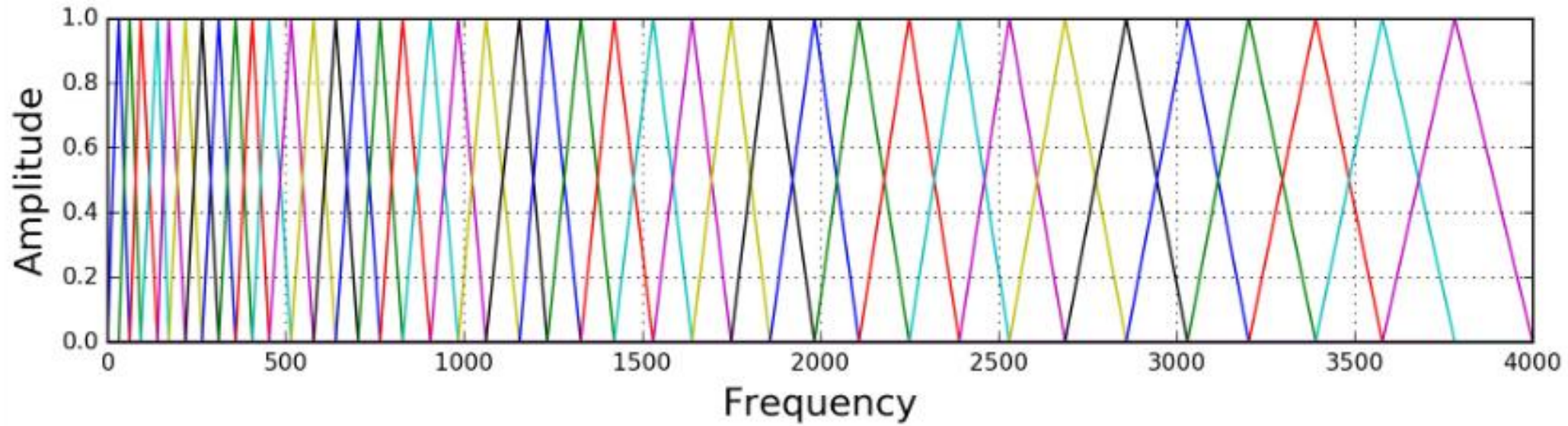
MEL Scale

The Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies. We can convert between Hertz (f) and Mel (m) using the following equations:

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

$$f = 700(10^{m/2595} - 1)$$

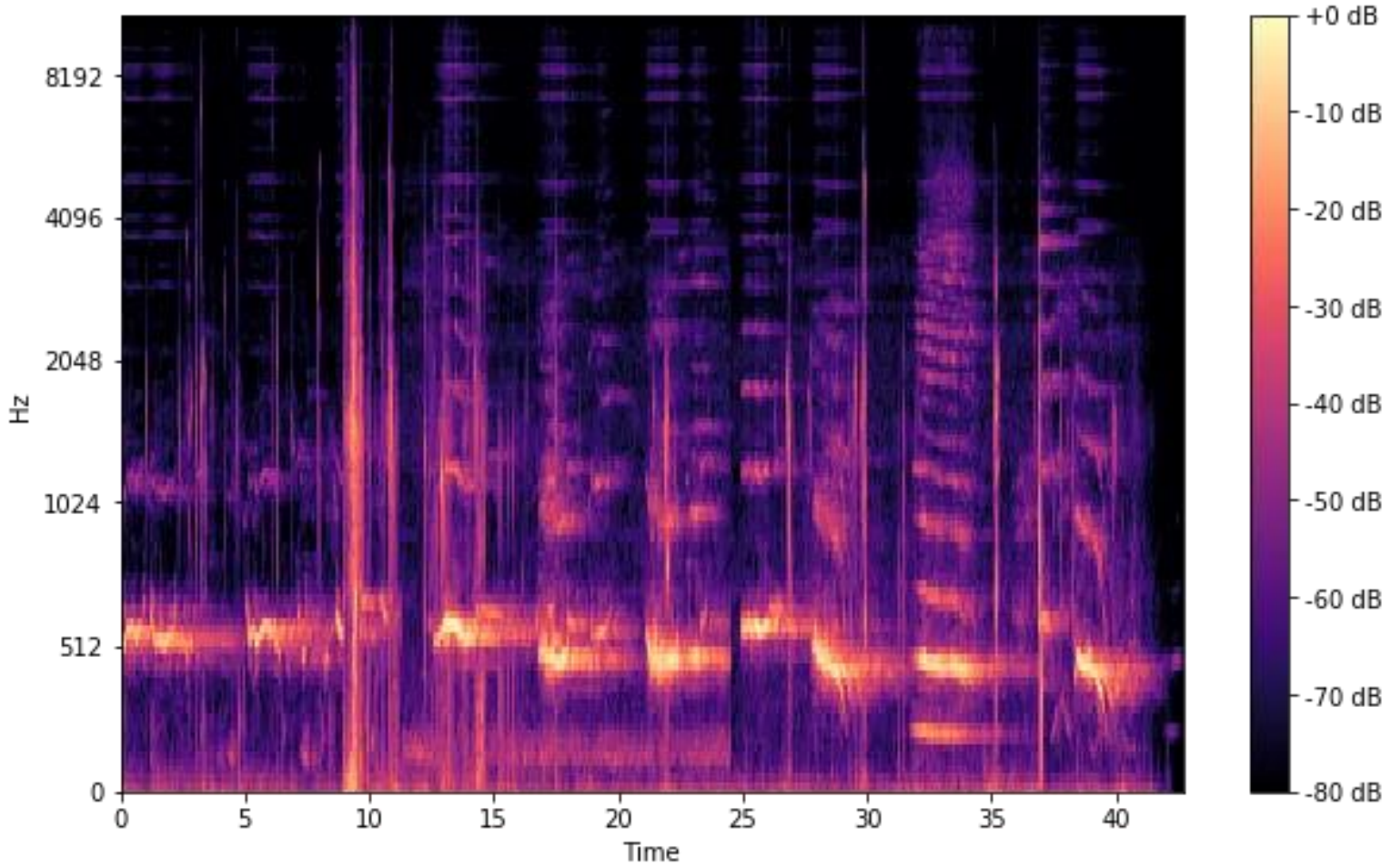
Each filter in the filter bank is triangular having a response of 1 at the center frequency and decrease linearly towards 0 till it reaches the center frequencies of the two adjacent filters where the response is 0, as shown in this figure:



MEL Scale advantages

- Mel-spectrograms help to focus on features that are important for perception and recognition tasks.
- By using a perceptual scale, Mel-spectrograms provide features that are more robust and informative for tasks like speech recognition and music classification.
- Models trained on raw spectrograms may struggle with generalization and are not consistently relevant across different audio conditions or environments.
- Mel-spectrograms can improve generalization by providing a representation that aligns more closely with human auditory processing. This can lead to better performance and robustness in machine learning models.
- In summary
  - Align more closely with human auditory perception.
  - Reduce dimensionality and computational complexity.
  - Capture perceptually relevant features while discarding less useful high-frequency details.

MEL Spectrogram



MEL Spectrogram in Librosa

- Documentation
- https://librosa.org/doc/main/generated/librosa.filters.mel.html#librosa.filters.mel

- How to call the function
- `mel_spec = librosa.feature.melspectrogram(y=audio, sr=sr, n_mels=NMELS)`

MEL Cepstrum (MFCC)

- MEL Cepstrum Parameters (MFCCs) are an alternative to MEL bands
- They provide a compact representation of the audio signal's power spectrum by applying a series of transformations.
- These coefficients can be thought of as capturing the "envelope" of the log MEL spectra.
- They represent the main components of the spectral content in a lower-dimensional space (Dim. Reduction)
- Computation:
  - The logarithm of the MEL parameters is taken to compress the dynamic range and emphasize variations.
  - The Discrete Cosine Transform (DCT) is applied to the log Mel spectra to obtain the MFCCs.
  - Normally, only the 10 to 15 first MFCC parameters are keep as an input for NN (HF are discarded)
- The DCT decorrelates the features and compresses the information into a set of coefficients.
- MFCCs are generally more robust to variations in recording conditions and background noise compared to MEL.
- Very use full in speech recognition, but also used for general audio purposes

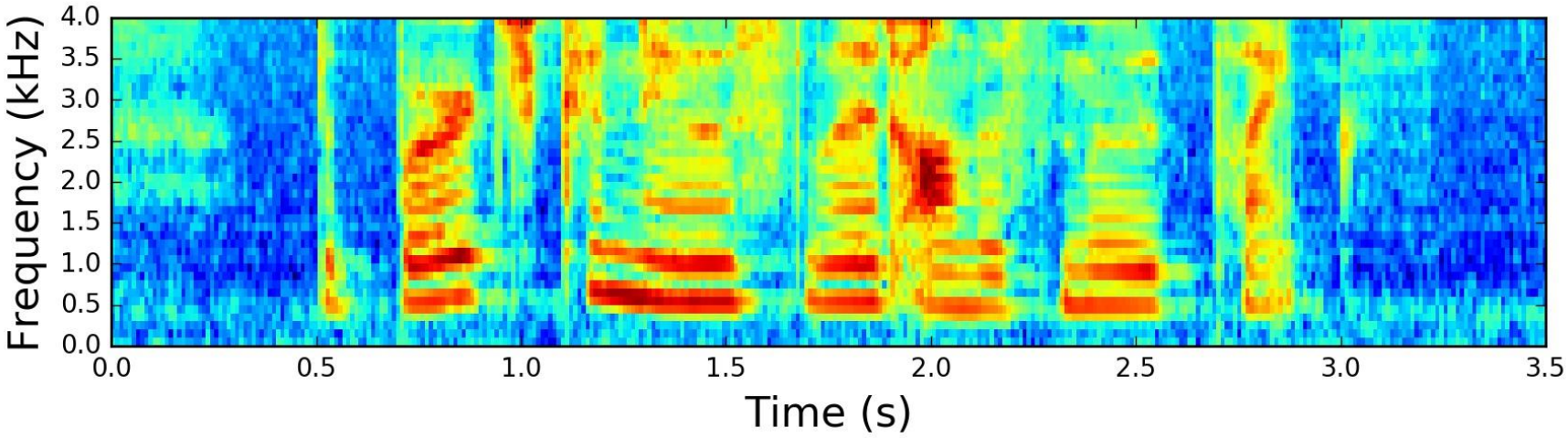
MEL Cepstrum (MFCC)

- Tutorial
  - https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

- How to call the function
- `mfcc = librosa.feature.mfcc(y=signal, sr=sample_rate, n_mfcc=num_mfcc, n_fft=n_fft)`

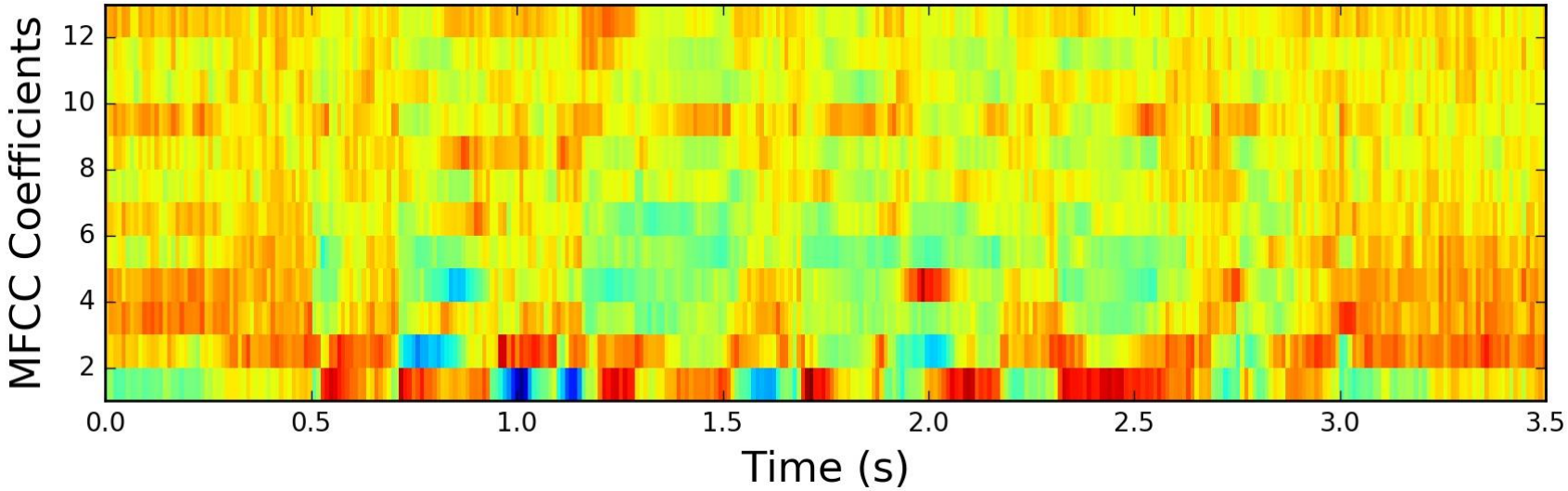
Espectrograma MFCC

MEL



MFCC

First 13 param.



Other features

https://www.kaggle.com/code/andradaolteanu/work-w-audio-data-visualise-classify-recommend

- Zero Crossing Rate
  - the rate at which the signal changes from positive to negative or back.
- Harmonics and Percusive decomposition
  - Harmonics are characteristics that represents the sound color
  - Percusive components represents the sound rhythm and emotion
- Tempo BMP (beats per minute)
  - Dynamic programming beat tracker.
- Spectral Centroid
  - indicates where the ”centre of mass” of the spectrum of a sound is located and is calculated as the weighted mean of the frequencies present in the sound.
- Spectral Contrast
  - measures the difference in amplitude between peaks and valleys in the audio spectrum
- Spectral Rolloff
  - is a measure of the shape of the signal spectrum. It represents the frequency below which a specified percentage of the total spectral energy, e.g. 85%, lies
- Chroma Frequencies
  - Interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

Applications of feature extraction in Audio

1. **Speech Recognition:** Extracting MFCCs and other features for automatic speech recognition systems.
2. **Music Analysis:** Analyzing music files to extract features for genre classification, music recommendation, and tempo estimation.
3. **Sound Event Detection:** Detecting and classifying sound events in environmental audio, such as identifying bird species in bird songs.
4. **Emotion Recognition:** Analyzing speech features for emotion detection and sentiment analysis.
5. **Instrument Recognition:** Identifying musical instruments from audio recordings.
6. **Voice Biometrics:** Extracting voice features for speaker identification and verification.
7. **Anomaly Detection:** Detecting unusual or abnormal sounds in audio streams, such as in industrial settings.
8. **Audio Synthesis:** Using extracted features to synthesize new audio content, like generating music or voice.