

GCT634/AI613: Musical Applications of Machine Learning

# Music Classification: Overview & Audio Features



**Juhan Nam**

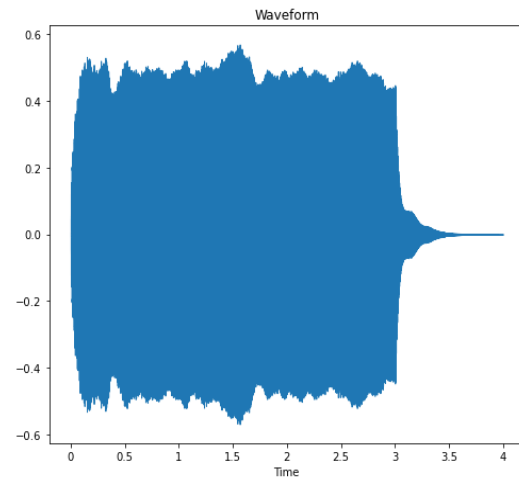
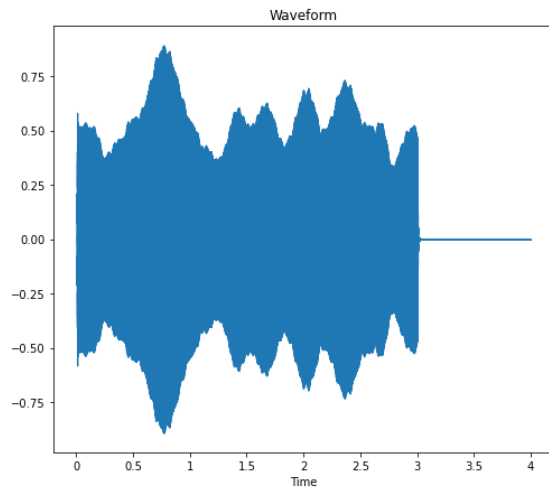
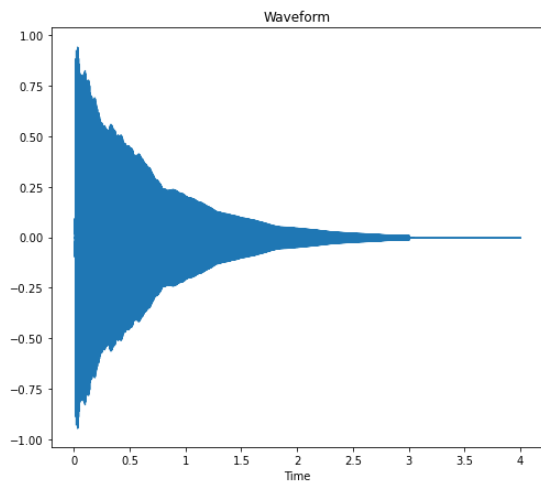
# Musical Instrument Classification

- Classify the musical instrument samples into one of the categories
  - 1) organ, 2) guitar, 3) brass



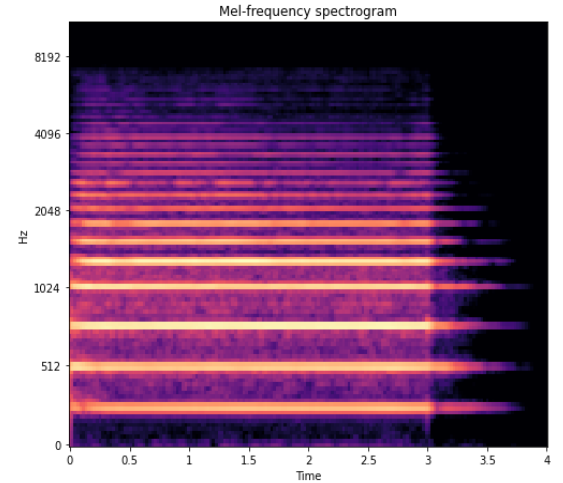
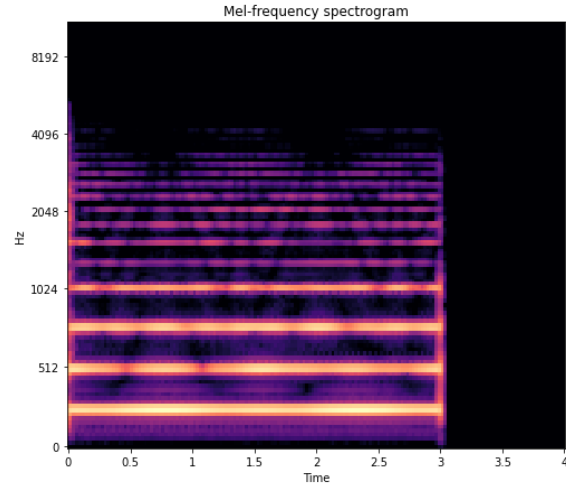
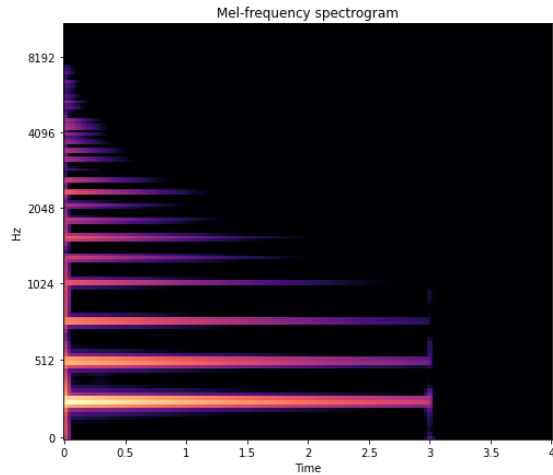
# Musical Instrument Classification

- Classify the musical instrument samples into one of the categories
  - 1) organ, 2) guitar, 3) brass



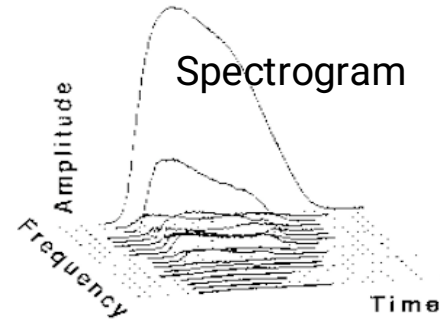
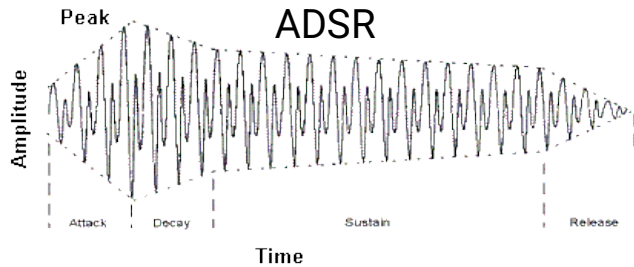
# Musical Instrument Classification

- Classify the musical instrument samples into one of the categories
  - 1) organ, 2) guitar, 3) brass



# Audio Features that Determine “Timbre”

- Time envelope: ADSR
- Spectral envelope and its temporal change
- The ratio between tonal and noise-like character
- Changes of fundamental frequency: expressions such as vibrato
- Inharmonicity: the amount of non-integer partials (e.g., bell, marimba)



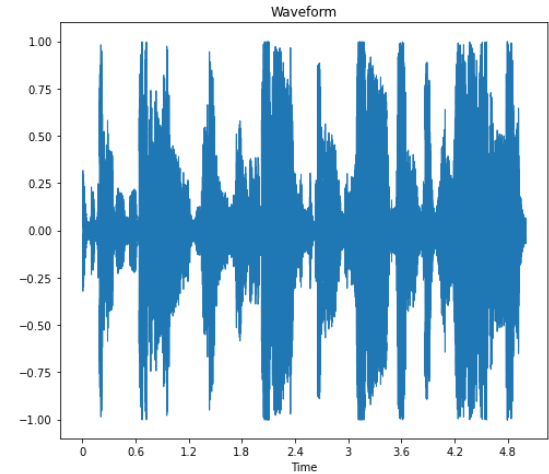
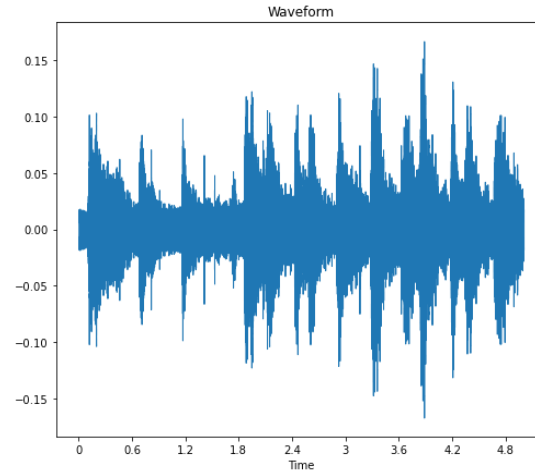
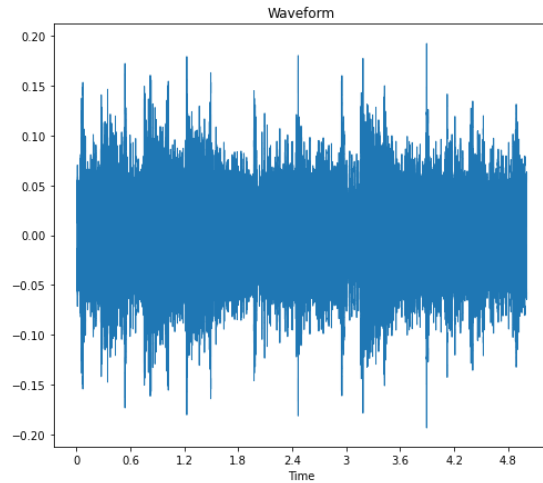
# Musical Genre Classification

- Classify the musical instrument samples into one of the categories
  - 1) rock, 2) jazz, and 3) hiphop



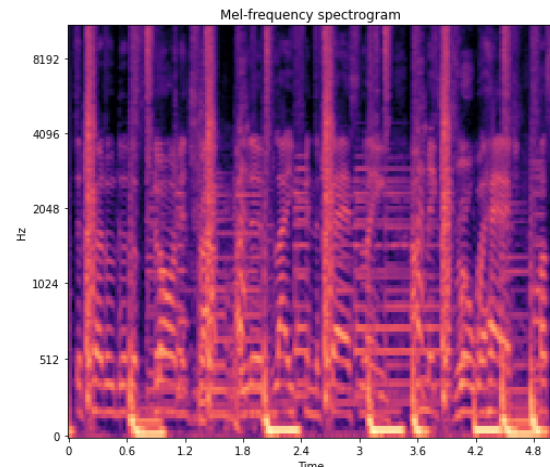
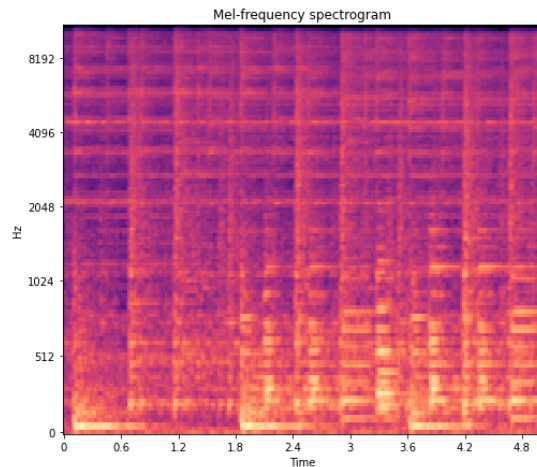
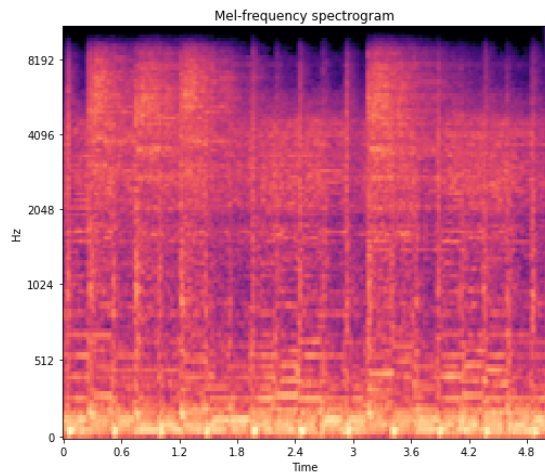
# Musical Genre Classification

- Classify the musical instrument samples into one of the categories
  - 1) rock, 2) jazz, and 3) hiphop



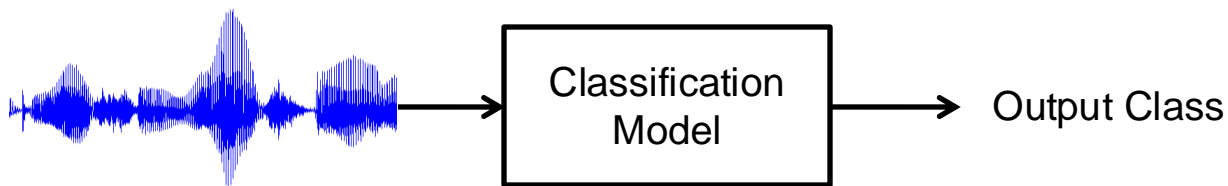
# Musical Genre Classification

- Classify the musical instrument samples into one of the categories
  - 1) rock, 2) jazz, and 3) hiphop



# Music Classification

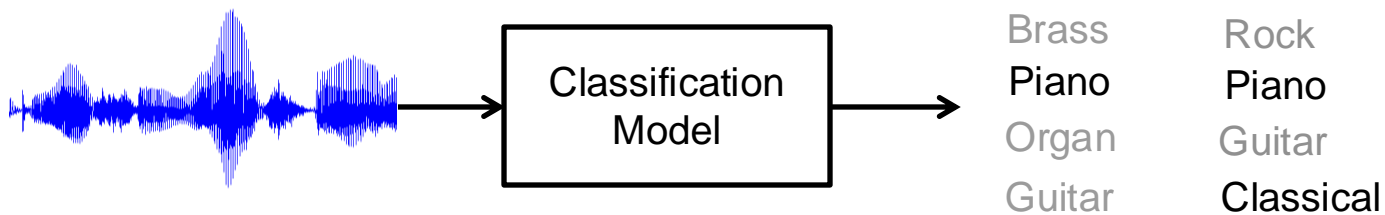
- A family of tasks that classify music data into a set of classes



- Input data
  - **Audio: waveform**
  - More generally, MIDI or other symbolic music data can be included but we will not handle them here
- Output class
  - **Instrument, genre, mood, or other musical attributes**
  - More generally, pitch, chord, beat (on/off) can be included but such sequential output will be handled in automatic music transcription.

# Music Classification

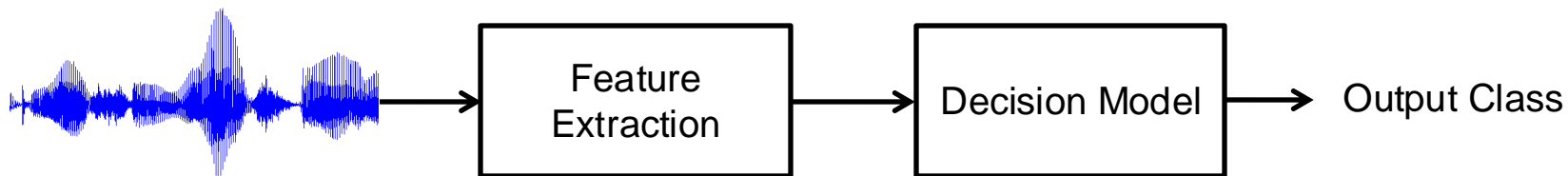
- They can be defined as a **single-label** or **multi-label** classification task



- Tasks in Music Information Retrieval (MIR) research
  - Genre classification
  - Mood classification
  - Instrument classification
  - Music tagging: all types of music attributes (genre, mood, instrument, ...)

# Music Classification Model

- General pipeline



## Acoustic Level

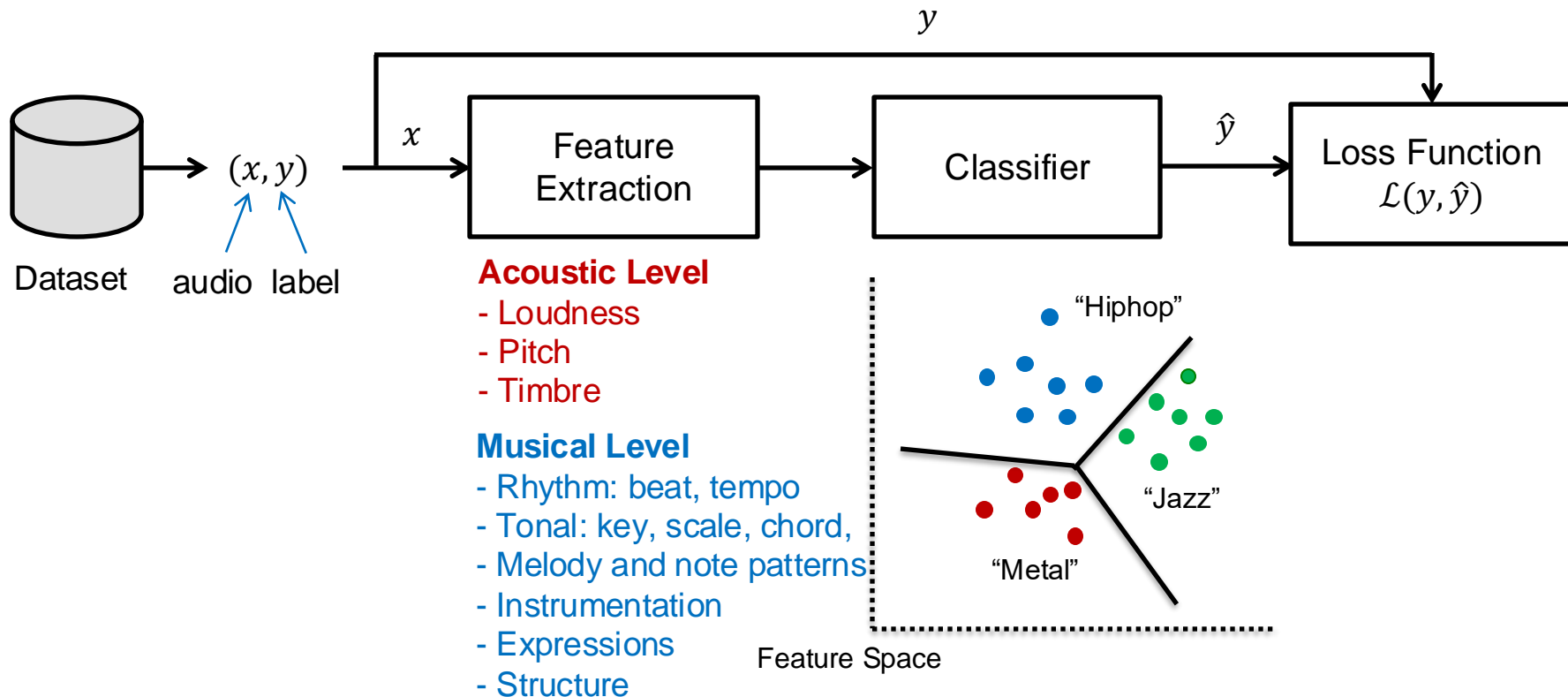
- Loudness
- Pitch
- Timbre

## Musical Level

- Rhythm: beat, tempo
- Tonal: key, scale, chord,
- Melody and note patterns
- Instrumentation
- Expressions
- Structure

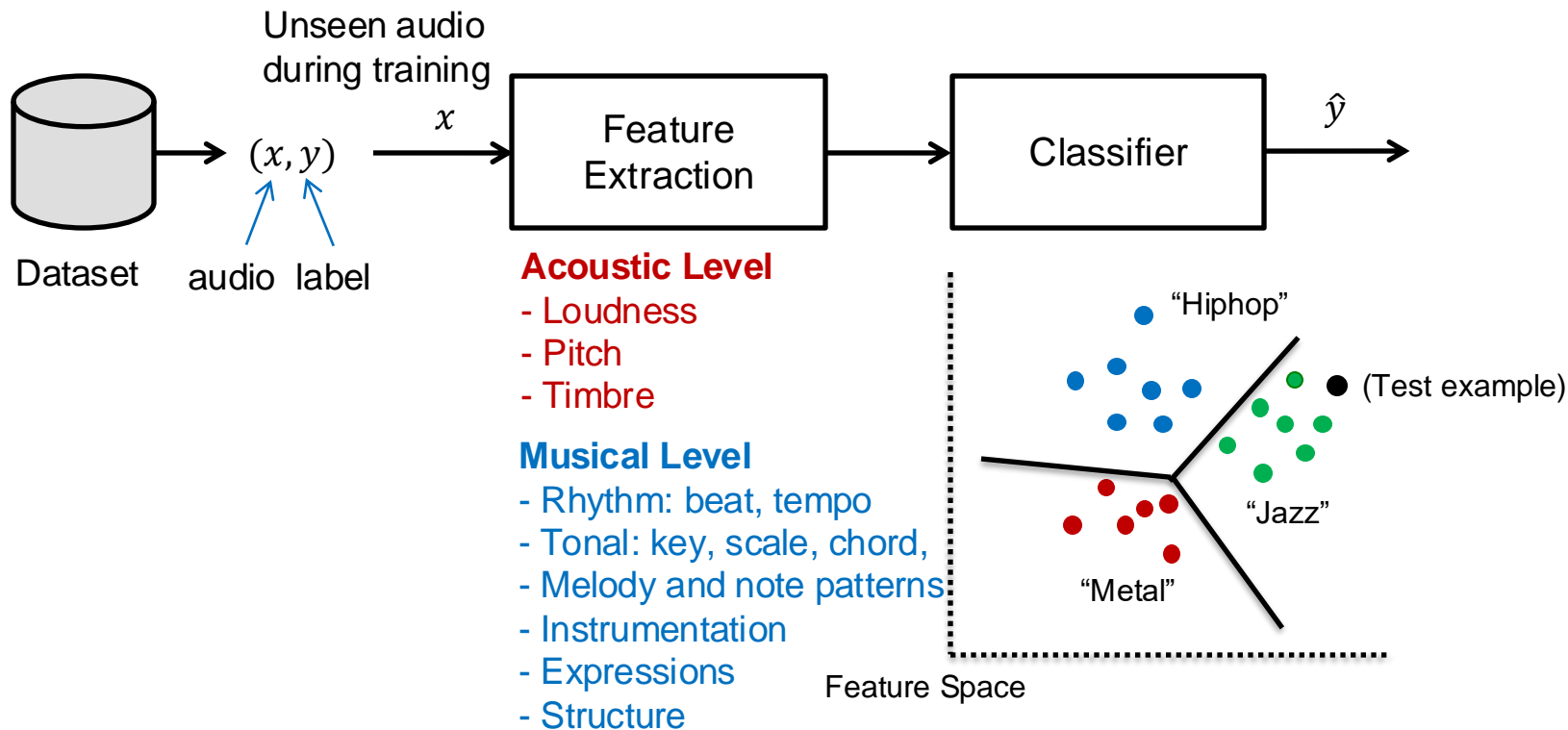
# Machine Learning for Music Classification

- Training phase



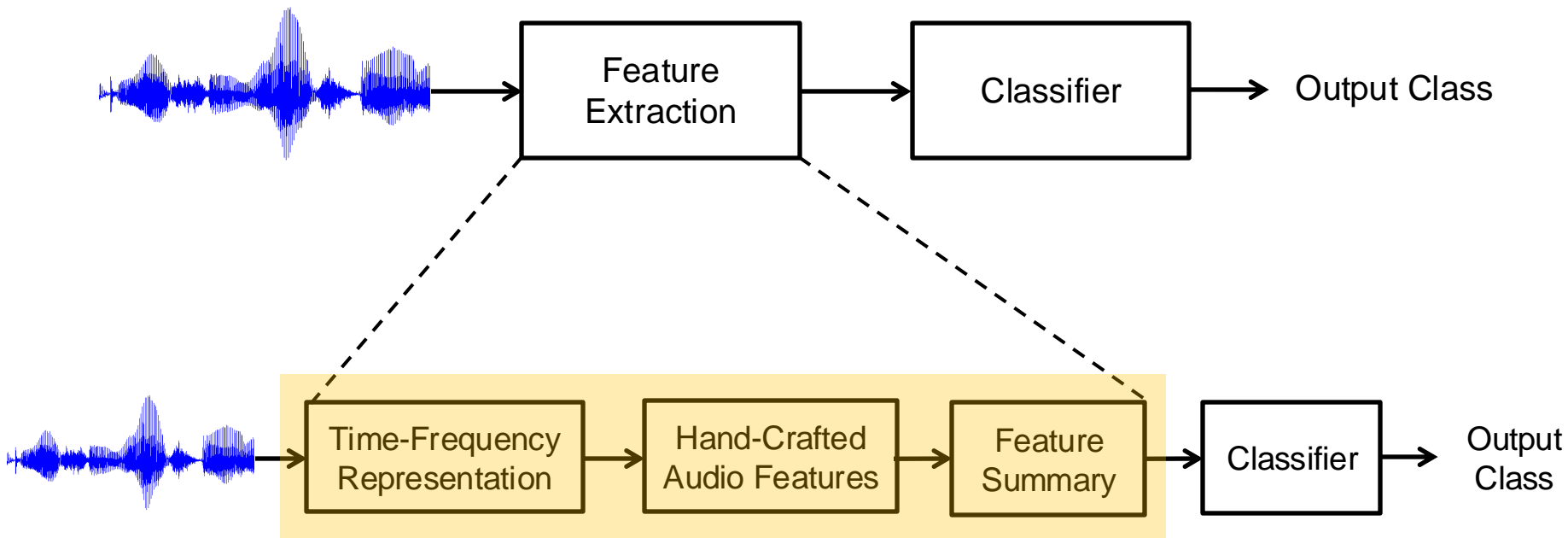
# Machine Learning for Music Classification

- Test phase



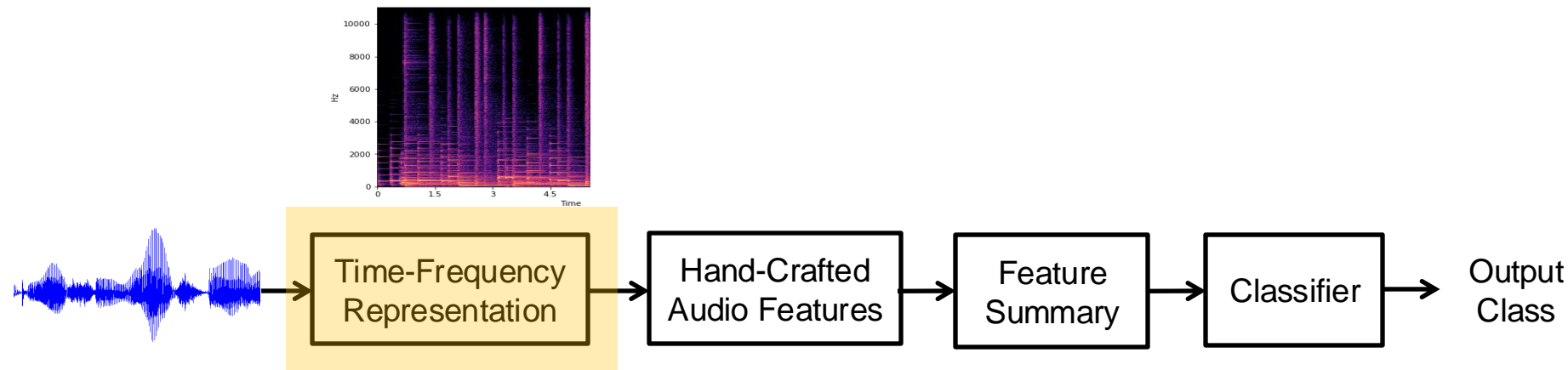
# Traditional Machine Learning Pipeline

- Traditional music classification models



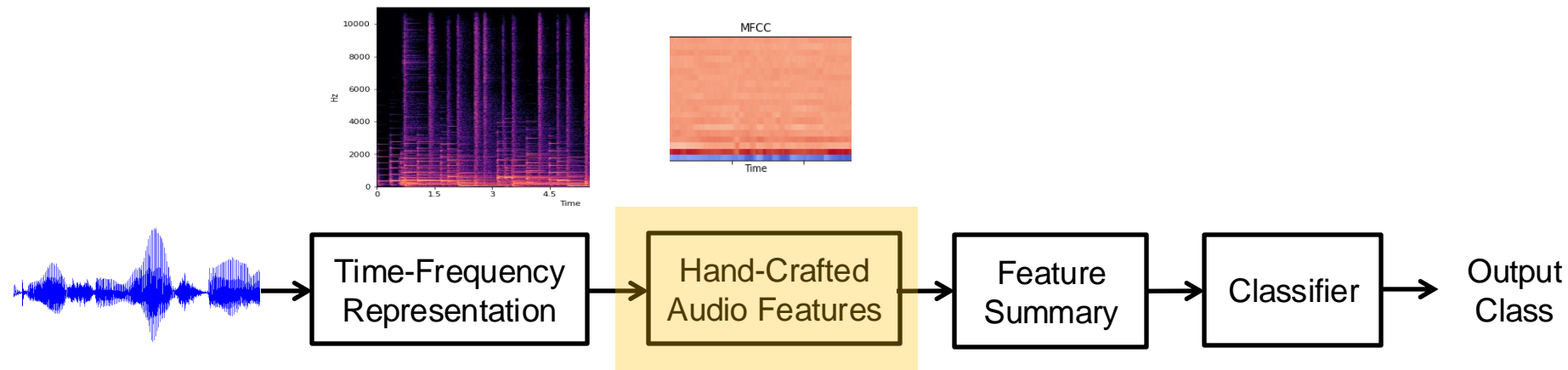
# Time-Frequency Representation

- Transform the waveforms into time-frequency representations
  - “2D image” that shows harmonic and percussive patterns of sounds well
  - Example: spectrogram, mel-spectrogram, constant-Q transform



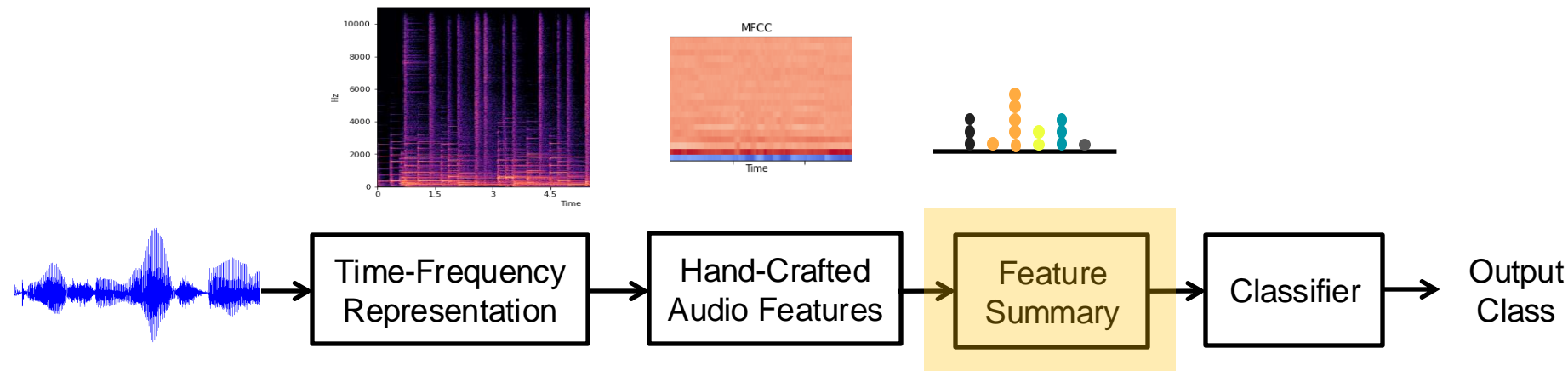
# Hand-Crafted Audio Features

- Extract certain characteristics of audio in a low-dimensional form
  - Mostly frame-level vectors
  - Engineered based on domain knowledge
  - Examples: Mel-Frequency Cepstral Coefficient (MFCC), Chroma, spectral statistics



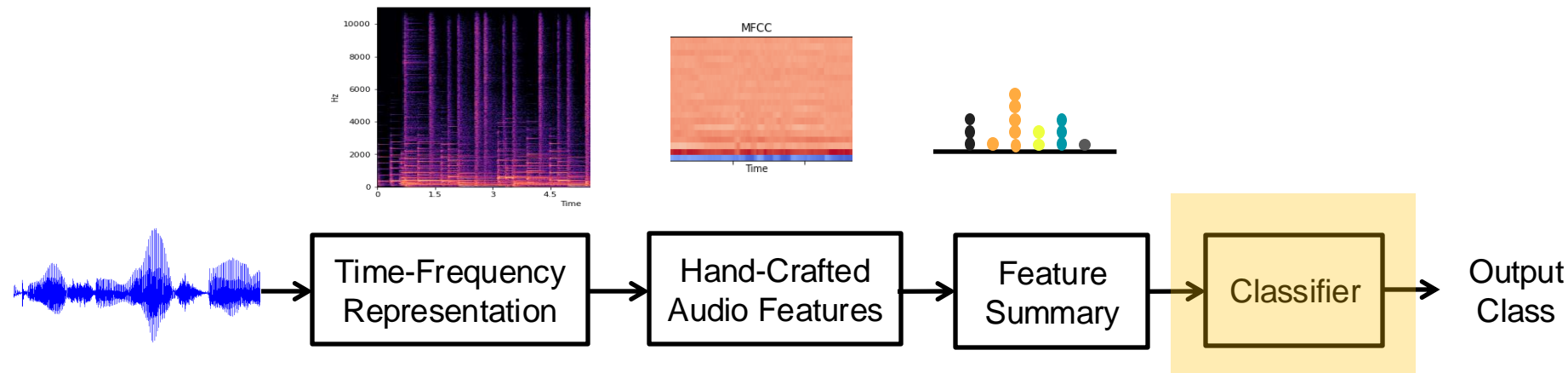
# Feature Summary

- Concatenate all audio features and summarize them over time
  - Statistical summary: average, max
  - Unsupervised learning algorithms such as K-means or PCA are often used for vector quantization (a-bag-of-words approach) and dimensionality reduction



# Classifiers

- Conduct supervised learning with the feature vectors and labels
  - Off-the-shelf classifiers are usually used
  - Examples: K-NN, Logistic regression, support vector machine, multi-layer-perceptron



# Traditional Machine Learning

- Advantages
  - A small dataset is fine
  - The classifiers are fast to train
  - The hand-engineered features are interpretable
- Disadvantages
  - Requires domain knowledge
  - The feature design is an art
  - The two-stage approach is sub-optimal
- Good as a baseline method

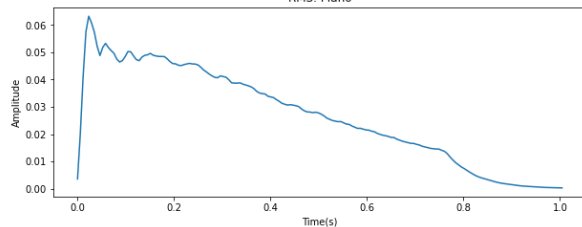
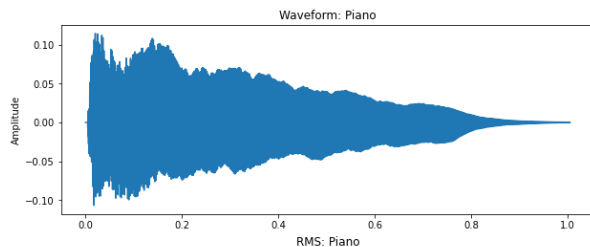
# Hand-Crafted Audio Features

- RMS: loudness
- Spectral statistics: timbre
- MFCC: timbre
- Chroma: tonality (pitch, key, chord)

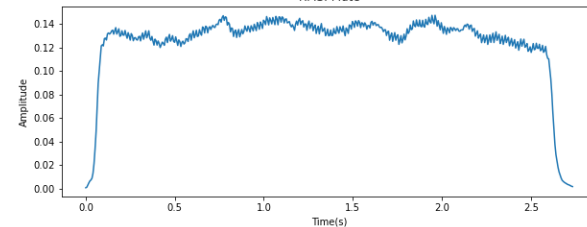
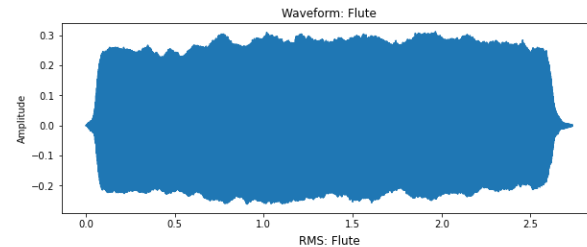
# Root-Mean-Square (RMS)

- Computes the amplitude envelope of waveforms

- Frame-by-frame computation: 
$$\text{RMS}(l) = \sqrt{\frac{1}{N} \sum_{m=0}^{N-1} (x(m - l \cdot R)w[n])^2}$$
  - Can be computed from STFT: 
$$\text{RMS}(l) = \sqrt{\frac{1}{N} \sum_{k=0}^{N-1} |X(k, l)|^2}$$
- $w[n]$ : window  
 $N$ : window size  
 $R$ : hop size



Piano (single note)



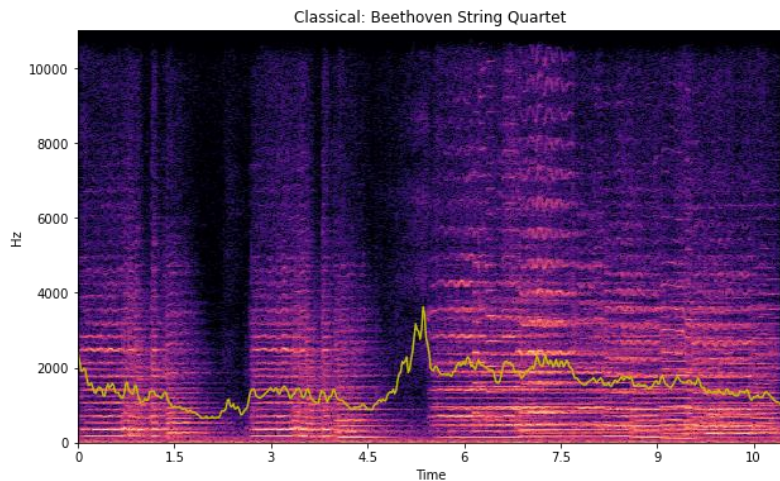
Flute (single note)

# Spectral Statistics: Centroid

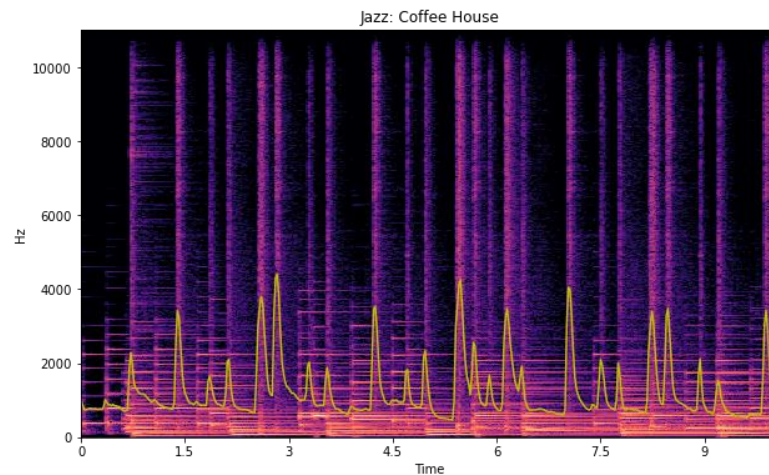
- “Center of mass” of the spectrum
  - Associated with the brightness of sounds

$$SC(l) = \frac{\sum_{k=0}^{N-1} f_k \cdot |X(k, l)|}{\sum_{k=0}^{N-1} |X(k, l)|}$$

$$f_k = \frac{k}{N} f_s$$



Classical Music



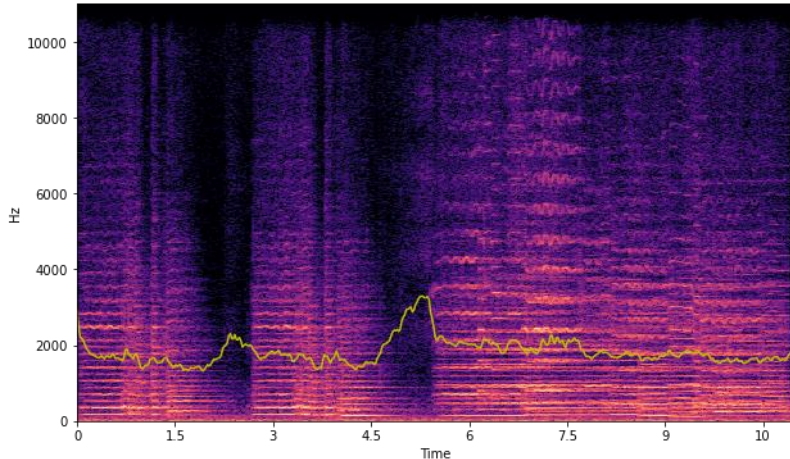
Jazz Music

# Spectral Bandwidth

- A measure of the bandwidth of the spectrum

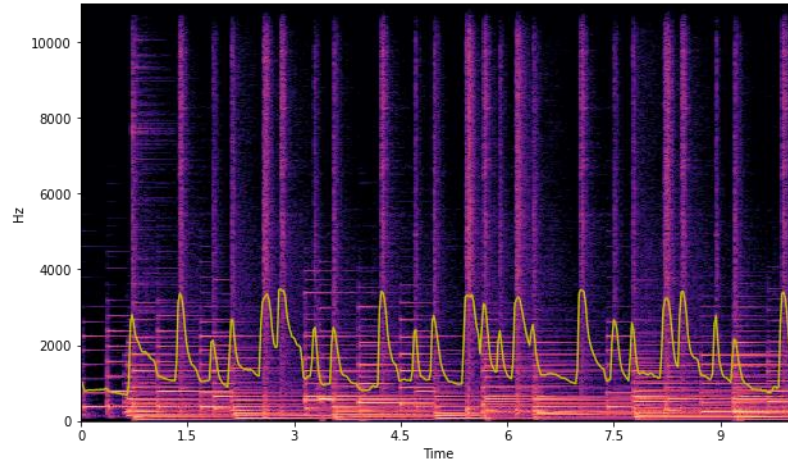
$$SB(l) = \frac{\sum_{k=0}^{N-1} (f_k - SC(l))^2 \cdot |X(k, l)|}{\sum_{k=0}^{N-1} |X(k, l)|}$$

Classical: Beethoven String Quartet



Classical Music

Jazz: Coffee House



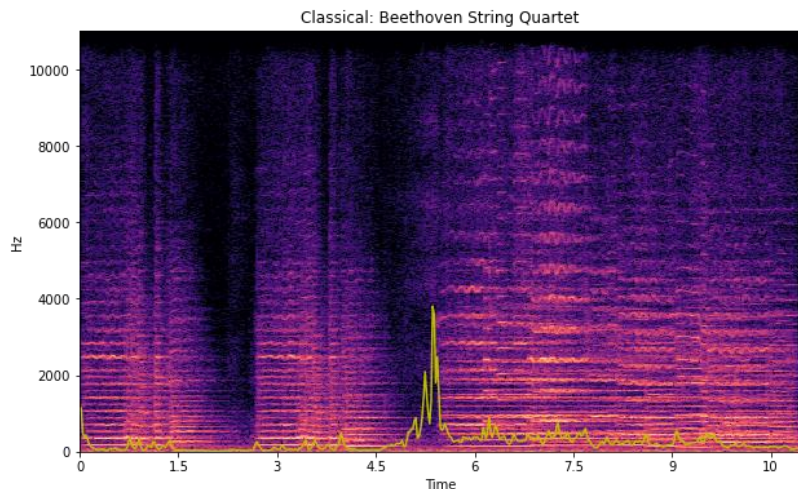
Jazz Music

# Spectral Flatness

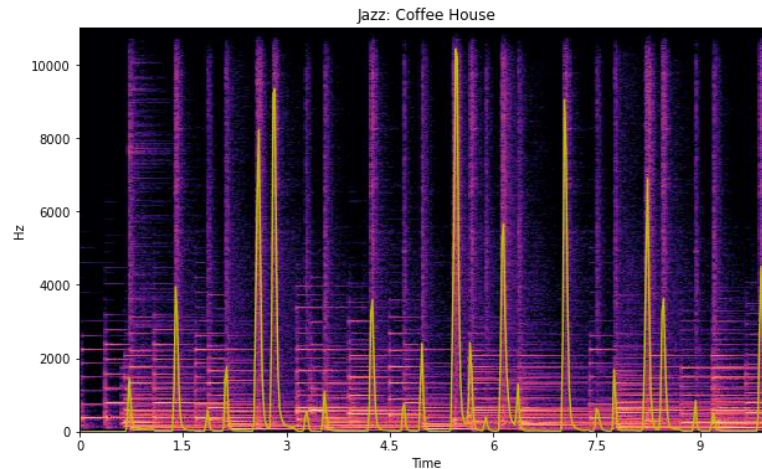
- A measure of the noisiness (or tonality) of the spectrum
  - The ratio between the geometric and arithmetic means

$$SF(l) = \frac{\sqrt[N]{\prod_{k=0}^{N-1} |X(k, l)|}}{\frac{1}{N} \sum_{k=0}^{N-1} |X(k, l)|}$$

Close to 1 for white noise  
(maximum flatness)



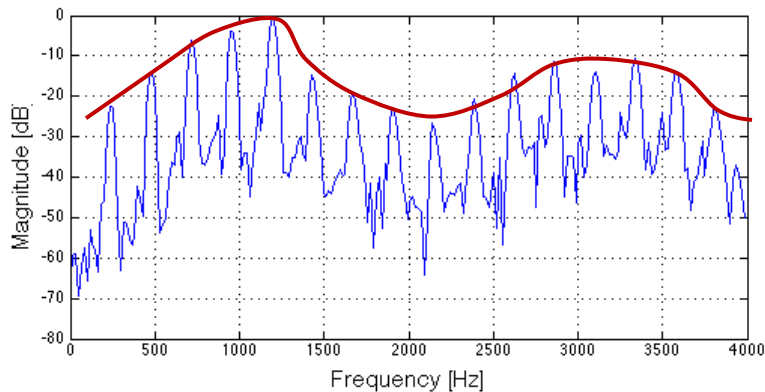
Classical Music



Jazz Music

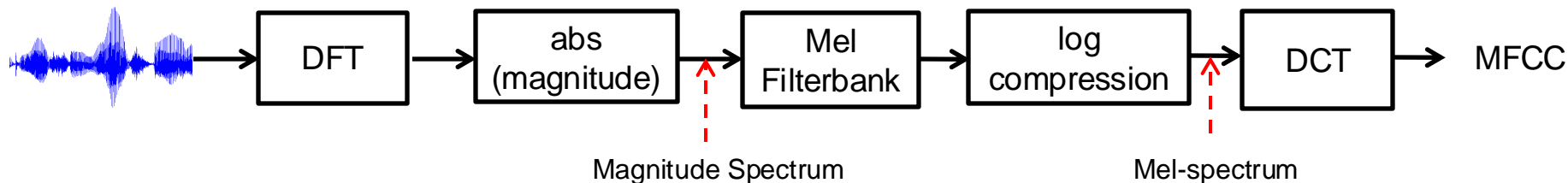
# Mel-Frequency Cepstral Coefficient (MFCC)

- Most popularly used audio feature to extract “timbre”
  - Pitch-invariant: extract the **spectrum envelop** from an audio frame
  - Standard audio feature in the legacy speech/speaker recognition systems: captures phonetic information and speaker information
  - Low-dimensional: typically 13 to 20-dim vectors

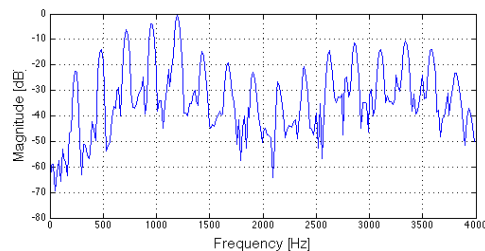


# Mel-Frequency Cepstral Coefficient (MFCC)

- Computation steps
  - Mel-spectrum: reduce the dimensionality (40 - 128 bins)
  - Log compression
  - Discrete Cosine Transform (DCT):
    - A small set of cosine kernels with low frequencies to reduce the dimensionality again (typically, 13 – 20 bins)
    - Captures a slowly varying trend of mel-spectrum over frequency which corresponds to the spectrum envelope

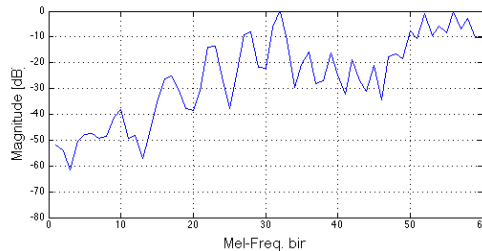


# Mel-Frequency Cepstral Coefficient (MFCC)



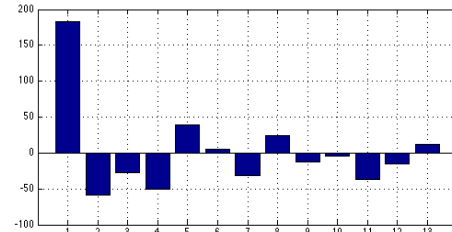
Magnitude spectrum  
(512 bins)

Mel  
filterbank

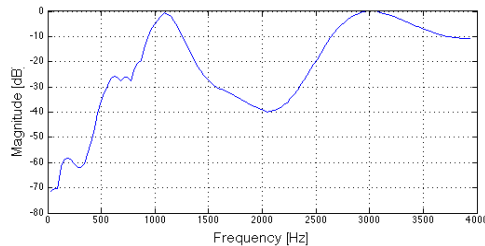


Frequency spectrum  
(mel-scaled, 60 bins)

DCT

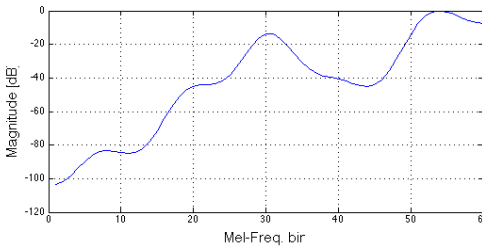


MFCC  
(13 dim)



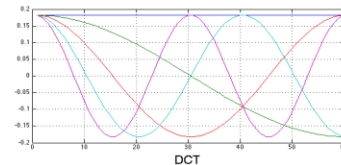
Reconstructed  
Magnitude spectrum

Inverse mel  
filterbank

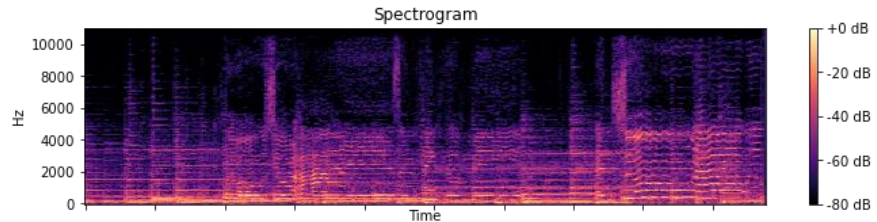


Reconstructed  
Mel spectrum

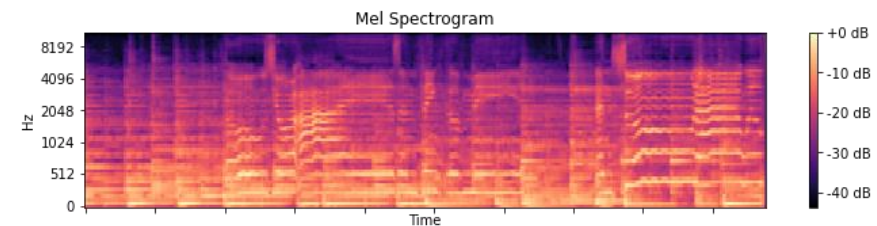
Inverse  
DCT



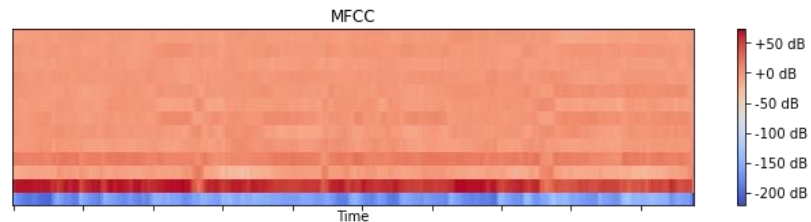
# Mel-Frequency Cepstral Coefficient (MFCC)



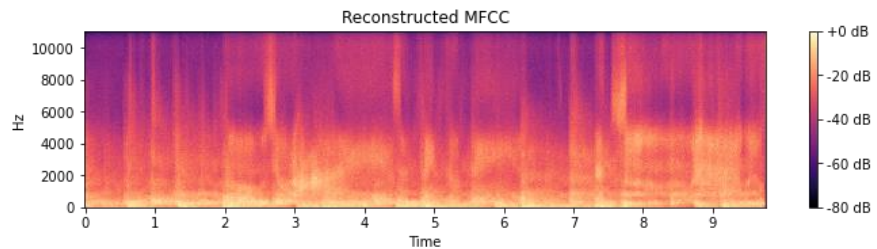
Spectrogram



Mel-frequency  
Spectrogram



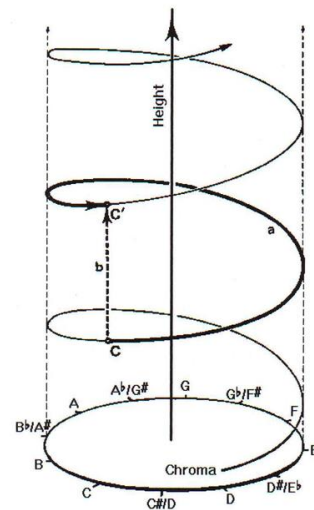
MFCC



Reconstructed  
Spectrogram  
from MFCC

# Chroma

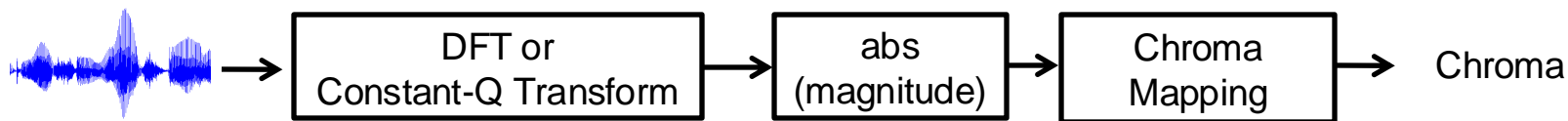
- Musical notes are denoted with a pitch class and an octave number
  - Pitch class: C, C#, D, D#, E, F, F#, G, G#, A, A#, B
  - Octave number: 0, 1, 2, 3, 4, 5, ...
  - Example: C4 (middle C), E3, G5
- The octave difference is the most consonant pitch interval
  - Therefore, they belong to the same pitch class
- This can be represented with “pitch helix”
  - Chroma: inherent circularity of pitch organization
  - Height: naturally increase and have one octave above for one rotation



Pitch Helix and Chroma  
(Shepard, 2001)

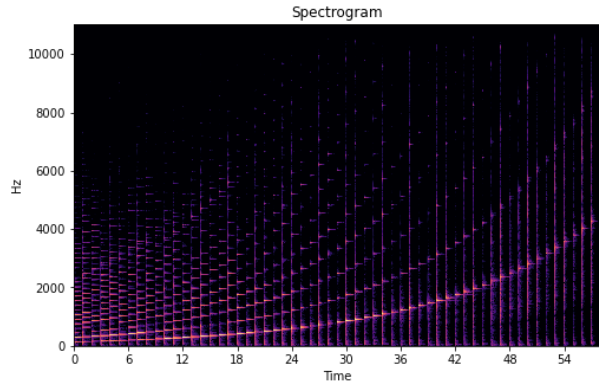
# Chroma

- Compute the energy distribution of an audio frame on 12 pitch classes
  - Convert the frequency to a musical note ( $= 12 \log_2 \left( \frac{f}{440} \right) + 69$ ) and take the pitch class from the musical note (e.g.  $69 \rightarrow A4 \rightarrow A$ )
  - Extract tonal characteristics, ideally removing timbre information
  - Useful in music synchronization, chord recognition, music structure analysis, music genre classification
- Computation Steps
  - Projecting the DFT or Constant-Q transform onto 12 pitch classes

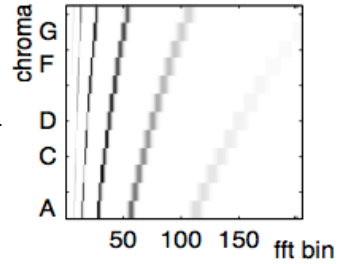


# Chroma

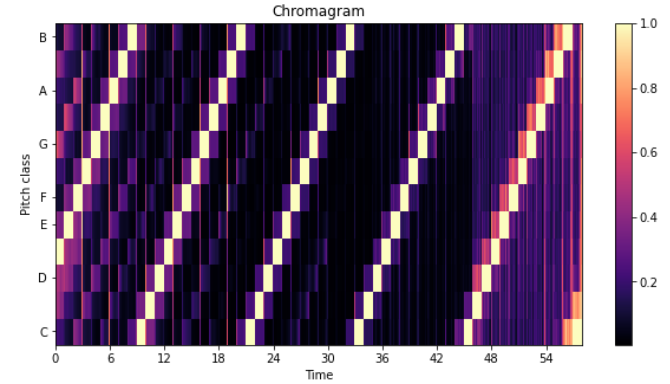
Spectrogram



Chroma mapping



Chroma



(Reconstructed Chroma: Shepard tone)

