Machine Listening for Music and Sound Analysis

Lecture 1 – Audio Representations

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://machinelistening.github.io

Fraunhofer IDMT

Learning Objectives

- Sound categories
- Music representations
- Audio representations
- Audio signal decomposition
- Audio features

Sound Categories
Environmental Sounds

- Sound sources
  - Nature, climate, humans, machines
- Sound characteristics
  - Structured or unstructured, stationary or non-stationary, repetitive or without any predictable nature
- Sound duration
  - From very short (gun shot, door knock, shouts) to very long and almost stationary (running machines , wind, rain)

https://freesound.org/people/{InspectorJ/sounds/416529, prometheus888/sounds/458461, MrAuralization/sounds/317361}
https://ccsearch-dev.creativecommons.org/photos/{39451123-ee45-4ec3-ad8d-b42d856bca06, c69d3b07-76bd-43e2-a44e-8742edc8447a, ab3062ab-fe0f-420d-b93d-7451db166b4e}

Fraunhofer IDMT

Sound Categories
Music signals

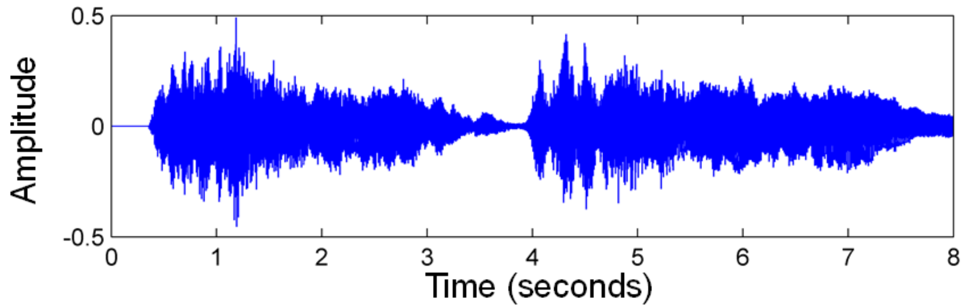
- Sound sources
  - Music instruments
    - Sound production mechanisms (brass, wind, string, percussive)
  - Singing Voice
- Sound characteristics
  - Mostly well structured along
    - Frequency (pitch, overtone relationships, harmony)
    - Time (onset, rhythm, structure)



https://freesound.org/people/{whatsanickname4u/sounds/127337, jcveliz/sounds/92002, klankbeeld/sounds/192691}
https://ccsearch-dev.creativecommons.org/photos/{a27a7541-45f5-4176-91a4-e2cb70eea266, photos/79d466c1-cfa6-417e-9832-34438678bf5d, 269394a4-5803-47fd-abaa-57ef92735e24}

Music Representations
Recording & Notation

- Music recording (waveform)



- Music notation (score)

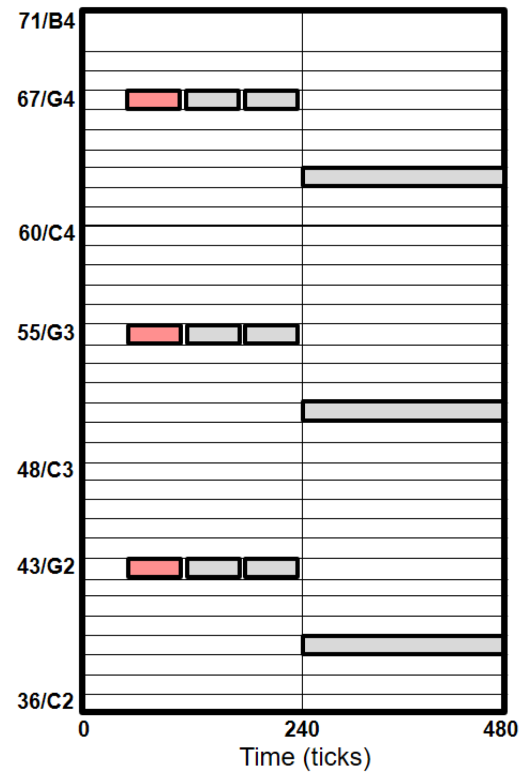
Figure 1.1 from [Müller, FMP, Springer 2015]

Music Representations
MIDI

- Sequence of note events (MIDI)

Figure 1.13 from [Müller, FMP, Springer 2015]

Time (Ticks)	Message	Channel	Note Number	Velocity
60	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	63	100
0	NOTE ON	2	51	100
0	NOTE ON	2	39	100
240	NOTE OFF	1	63	0
0	NOTE OFF	2	51	0
0	NOTE OFF	2	39	0



Music Representations
MusicXML

- Sequence of note events (MusicXML)

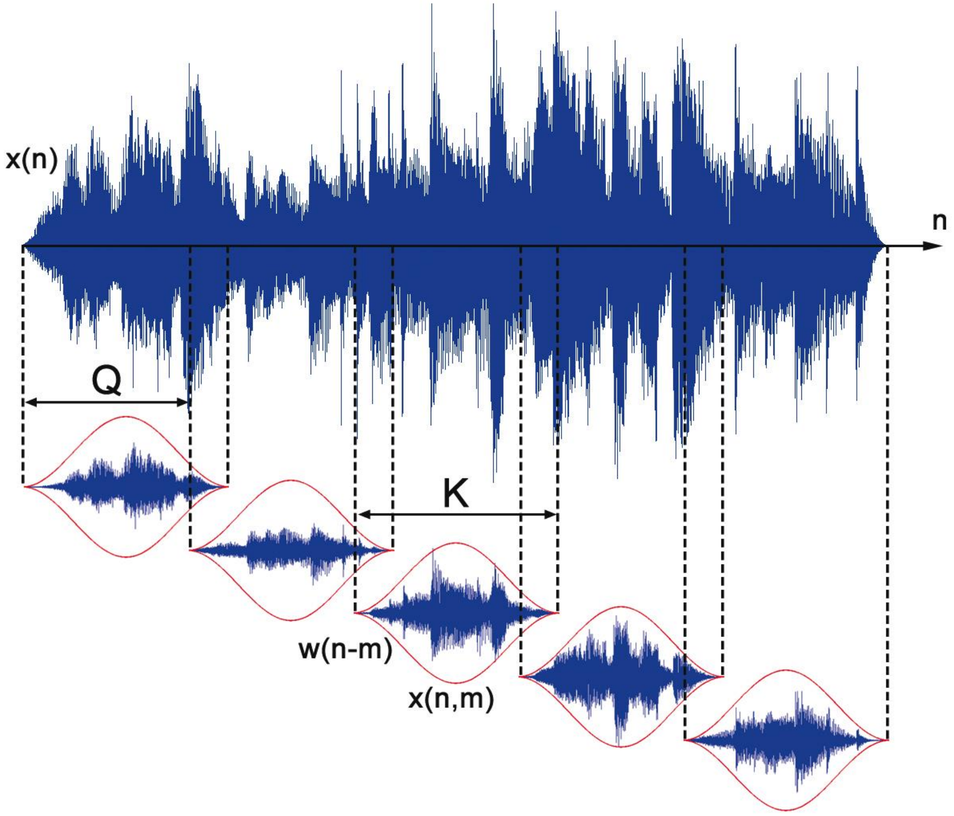
```
<note>
  <pitch>
    <step>E</step>
    <alter>-1</alter>
    <octave>4</octave>
  </pitch>
  <duration>2</duration>
  <type>half</type>
</note>
```

Meinard Müller, Fundamentals of Music Processing, Figure 1.15, Springer 2015

Audio Representations

- Short-term (frame-by-frame) analysis of audio signals

x(n)	digital audio signal
w(n)	windowing function
Q	hopsize
K	windowsize / blocksize



Fraunhofer IDMT

Audio Representations
Short-term Fourier Transform (STFT)

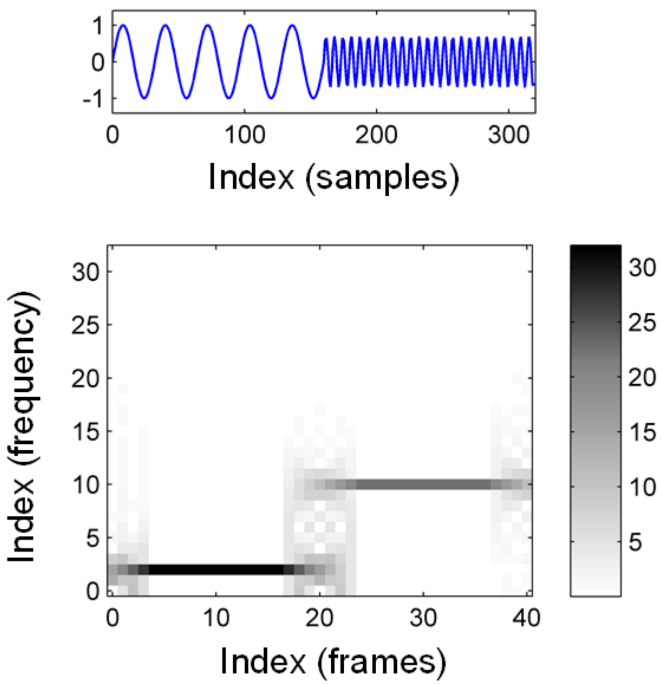
- Short-term Fourier Transform (STFT)
  - Discrete Fourier Transform (DFT)

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, k = 0, 1, \ldots, N-1$$

  - Instead of full signal, short (overlapping) windowed segments are used
  - Linearly-spaced frequency axis & fixed resolution
  - Trade-off between
    - Frequency resolution (separate close frequency components)
    - Time resolution

Audio Representations
Short-term Fourier Transform (STFT)

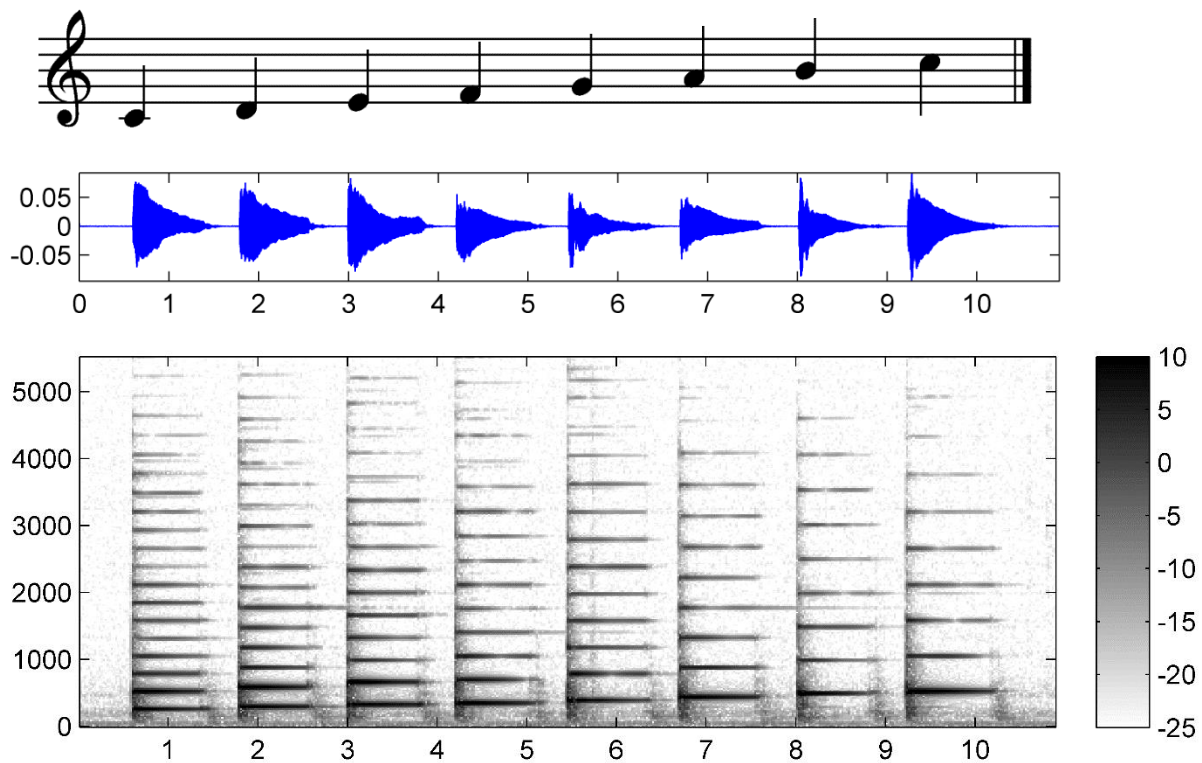
- Example: Sinosoid signal, two frequencies



Meinard Müller, Fundamentals of Music Processing, Figure 2.9, Springer 2015

Audio Representations
Short-term Fourier Transform (STFT)

- Example: C major scale, fundamental frequency & overtones



Audio Representations
Constant-Q Transform (CQT)

- Bank of filters with geometrically spaced center frequencies

$$f_k = f_0 \cdot 2^{k/b}$$

k - Filter index

b - Number of filters per octave

- Filter bandwidth (for adjacent filters)

$$\Delta_k = f_{k+1} - f_k = f_k \left(2^{\frac{1}{b}} - 1\right)$$

  - Increasing time resolution towards higher frequencies
  - Resembles human auditory perception

Audio Representations
Constant-Q Transform (CQT)

- Constant frequency-to-resolution ratio

$$Q = \frac{f_k}{\Delta_k} = \frac{1}{2^{\frac{1}{b}} - 1}$$

- Correspondence to musical note frequencies

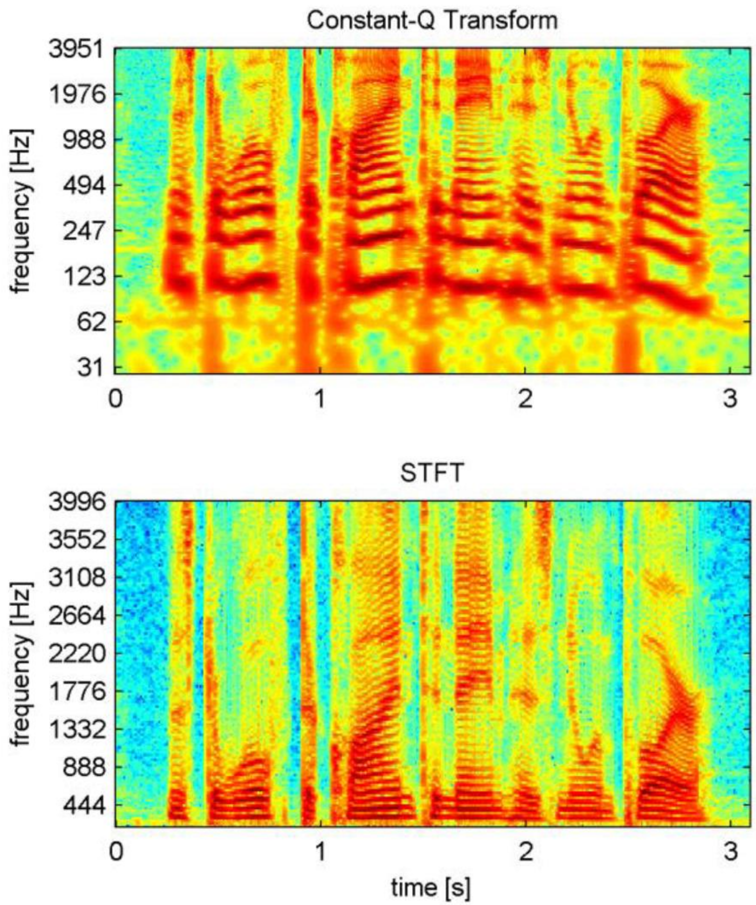
$$f_m[\mathrm{Hz}] = 440 \cdot 2^{\frac{m-69}{12}}$$

m – MIDI pitch

A4 (440 Hz) – reference pitch

Audio Representations
Constant-Q Transform (CQT)

- Example signal (CQT vs. STFT)



Audio Representations
Mel Spectrogram

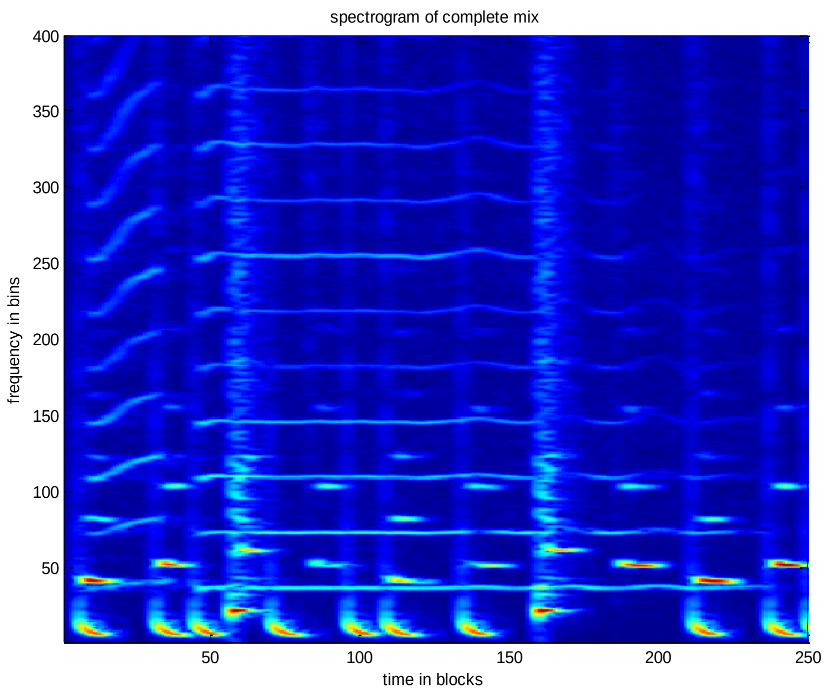
- Mel frequency scale (Stevens et al., 1937)

$$f[\mathrm{Mel}] = 2595 \cdot \log_{10}(1 + \frac{f[\mathrm{Hz}]}{700})$$

  - Describes perceived pitch of sinosoidal frequencies
- Mel spectrogram
  - Time-frequency representation sampled around
    - Equally spaced times
    - Frequency points along the mel-scale

Audio Signal Decomposition
Music mixtures

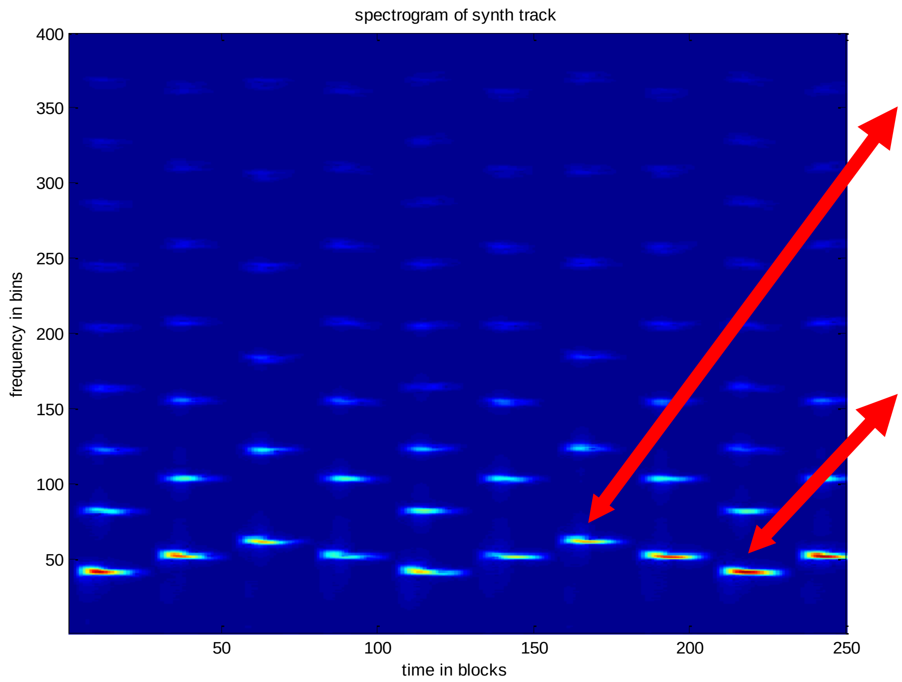
- Instrument mixture (STFT magnitude spectrogram)
    - Bass + melody (saxophone) + drums



Audio Signal Decomposition

Music mixtures

- Bass
  - Harmonic structure, stable tones

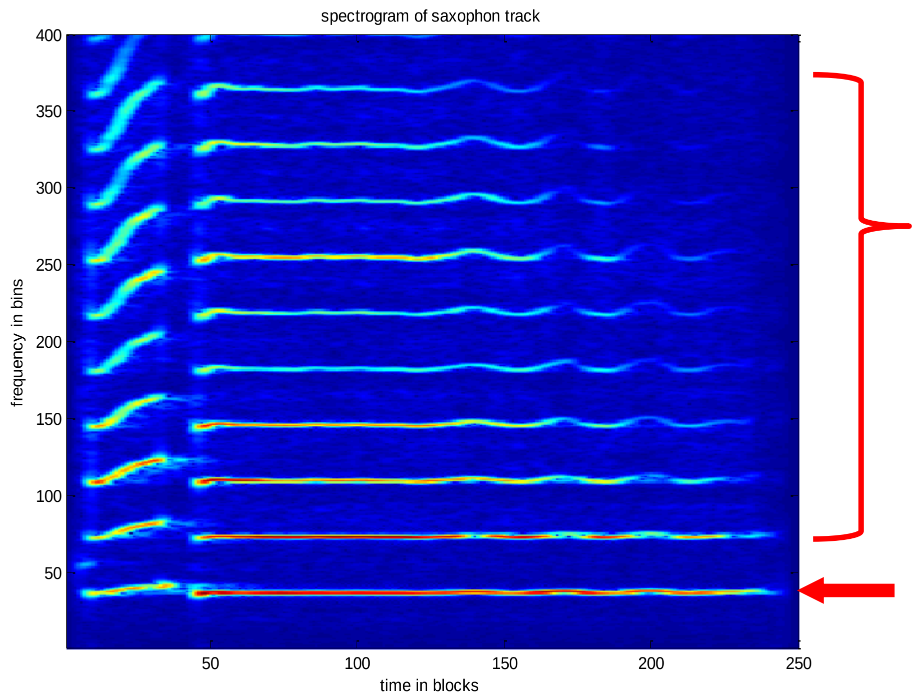


Higher pitch - spectrum stretched

Lower pitch - spectrum compressed

Audio Signal Decomposition
Music mixtures

- Melody (saxophone)
  - Harmonic components (melody)

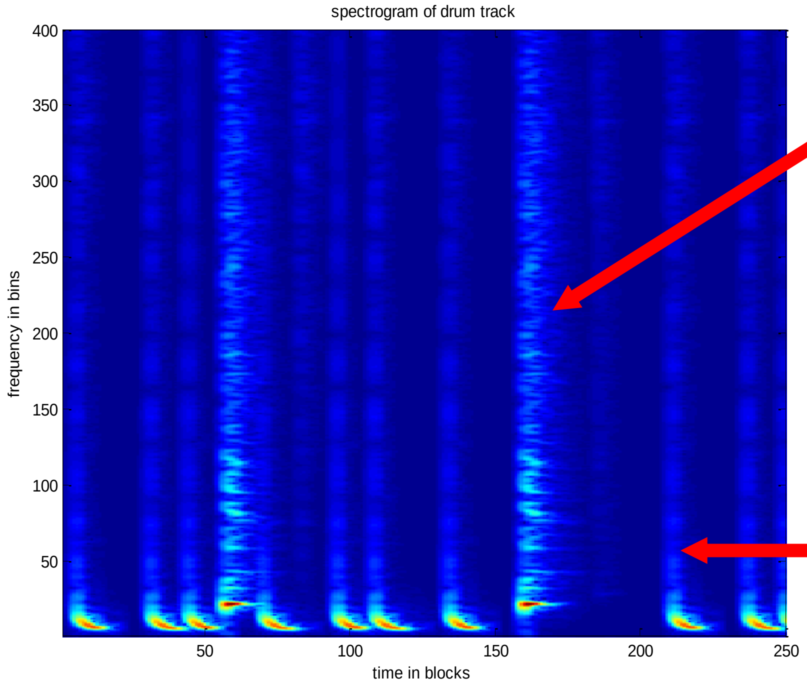


Harmonic overtones ≈ timbre

Fundamental frequency f_0 ≈ pitch

Audio Signal Decomposition
Music mixtures

- Drums
  - Percussive components (noise-like, inharmonic spectra)

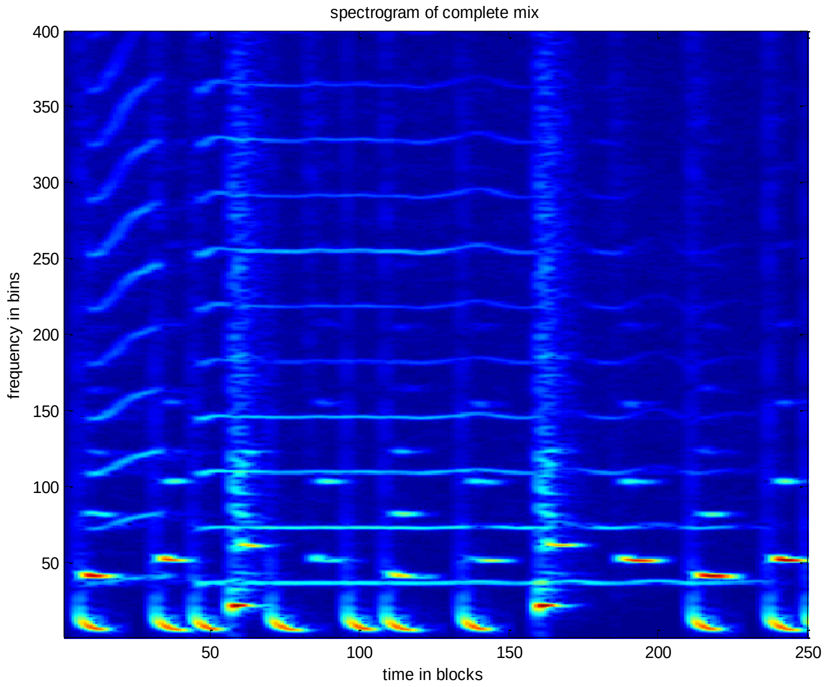


Snare drum (low frequencies)

Bass drum (low frequencies)

Audio Signal Decomposition
Music mixtures

- Instrument mixture (magnitude STFT)
    - All components add up to the mix signal



Audio Features
Motivation

- Compact representation of audio signal for machine learning applications
- Capture different properties at different semantic levels
    - Timbre – perceived sound, instrumentation
    - Rhythm – tempo, meter
    - Melody/Tonality – pitches, harmonies
    - Structure - repetitions

Audio Features
Categorization

	Timbre	Rhythm	Tonality
Low-level (Q~10 ms)	- Zero Crossing Rate (ZCR) - Linear Predictive Coding (LPC) - Spectral centroid / flatness		
Mid-level (Q ~ 2.5s)	- Mel-frequency Cepstral Coefficients (MFCC) - Octave-based Spectral Contrast (OSC) - Loudness	- Tempogram - Log-lag Autocorrelation (ACF)	- Chromagram - Enhanced Pitch Class Profiles (EPCP)
High-level	- Instrumentation	- Tempo - Time signature - Rhythm patterns	- Key - Scales - Chords

Fraunhofer IDMT

Audio Features
Timbre

- Timbre
  - Timbre distinguishes musical sounds that have the same pitch (fundamental frequency) and loudness
  - Affected by different acoustic phenomena such as
    - Spectral structure / envelope of overtones
    - Noise-like components
    - Formants (speech)
    - Inharmonicity (inharmonic relationship between overtones)
    - Variations over time: frequency (vibrato) or loudness (tremolo)

https://www.audiolabs-erlangen.de/resources/MIR/FMP/C1/C1S3_Timbre.html

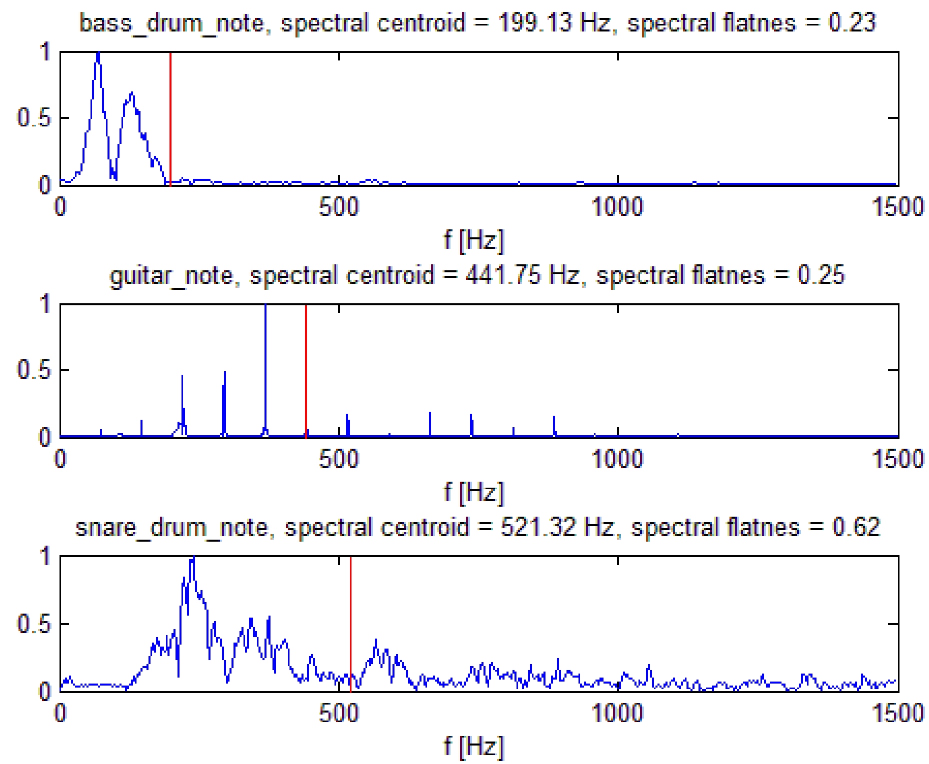
Audio Features
Timbre

- Timbre
    - When looking at musical instruments, we need to consider
        - Instrument construction
        - Sound production principles
            - Membranophones, chordophones, aerophones, electrophones
        - Human performance
            - Playing techniques, expressivity, dynamics, style

- How do design features to quantify these acoustic phenomena?

Audio Features
Low-level Audio Features

- Spectral Centroid (SC):
  - Center of mass in the magnitude spectrogram
  - Low-pitched vs. high-pitched sounds
- Spectral Flatness Measure (SFM)
  - Measure of flatness
  - Harmonic sounds (sparse energy distribution) vs. percussive sounds (wideband energy distribution)



Audio Features

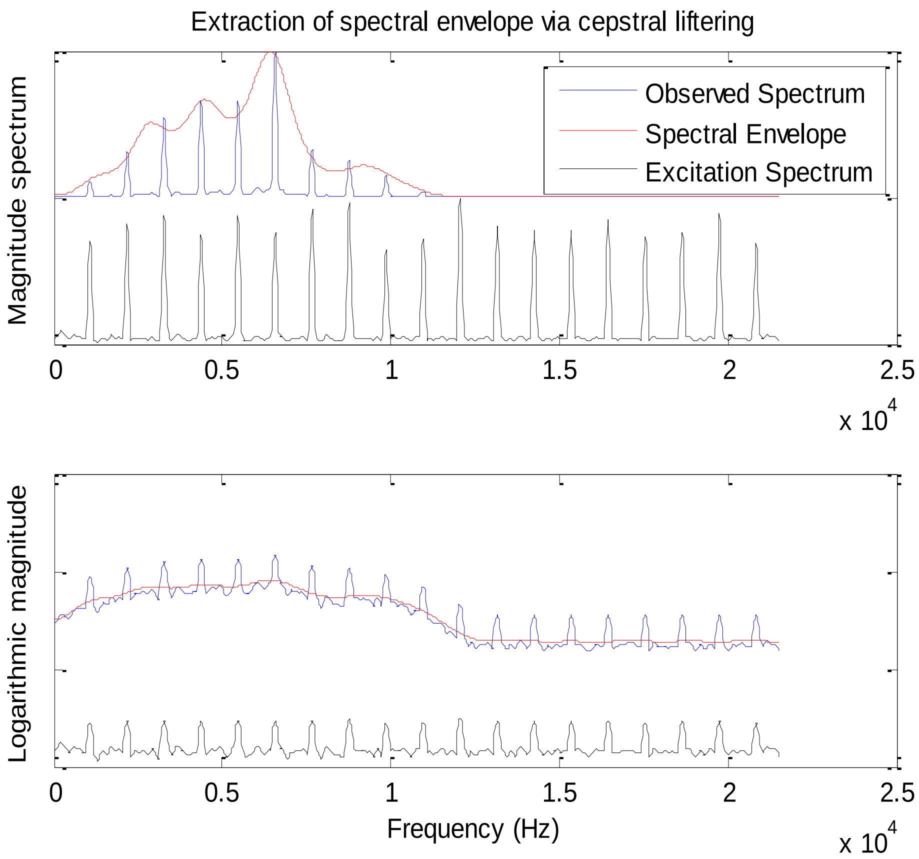
Timbre Mid-level Audio Features: MFCC

- Convolutive **excitation * filter** model
  - Excitation: vibration of vocal folds
  - Filter: resonance of the vocal tract

- FFT magnitude spectrum
  - Multiplicative **excitation · filter** model
- Logarithm of magnitude spectrum
  - Additive **excitation + filter** model
- Separation into
  - Smooth spectral envelope
  - Fine-structured excitation spectrum via „liftering"→ commonly done via Discrete Cosine Transform (and inverse)

Audio Features
Timbre Mid-level Audio Features: MFCC

- Example

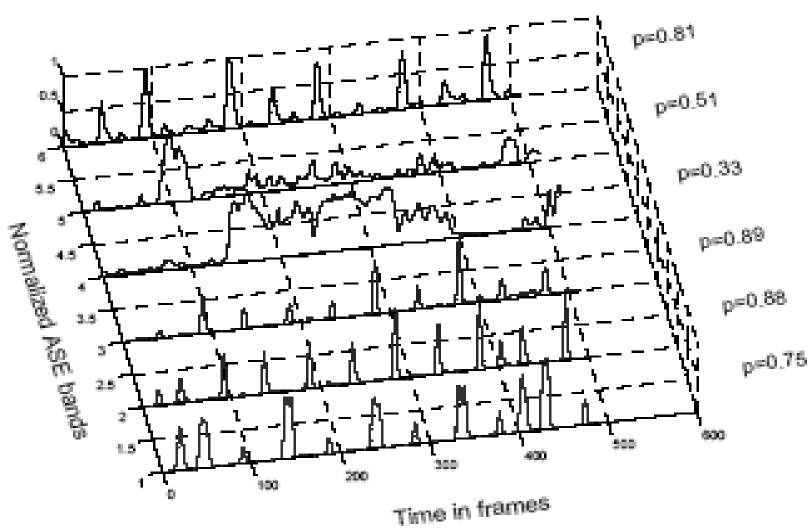


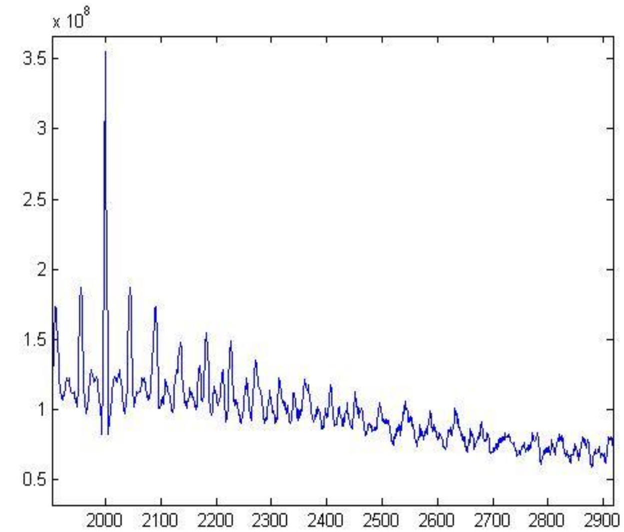
Fraunhofer IDMT

Audio Features

Rhythmic Mid-level Audio Features

- Rhythmic properties important for audio classification
- Audio Spectral Energy (ASE)
    - Weighted sum of energy slope in single bands
    - Find Periodicities via auto-correlation Function (ACF) on resulting detection function

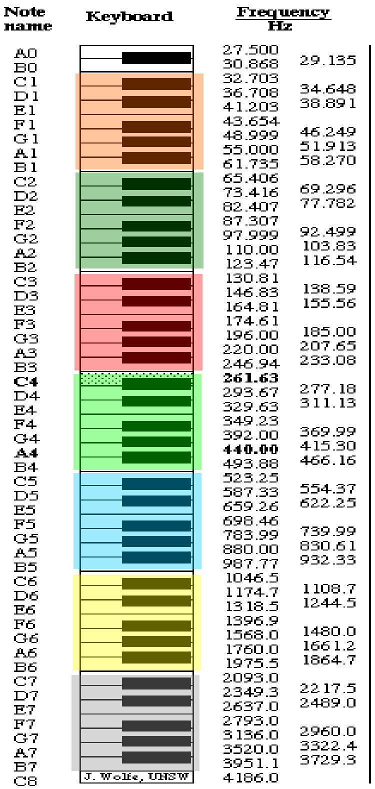


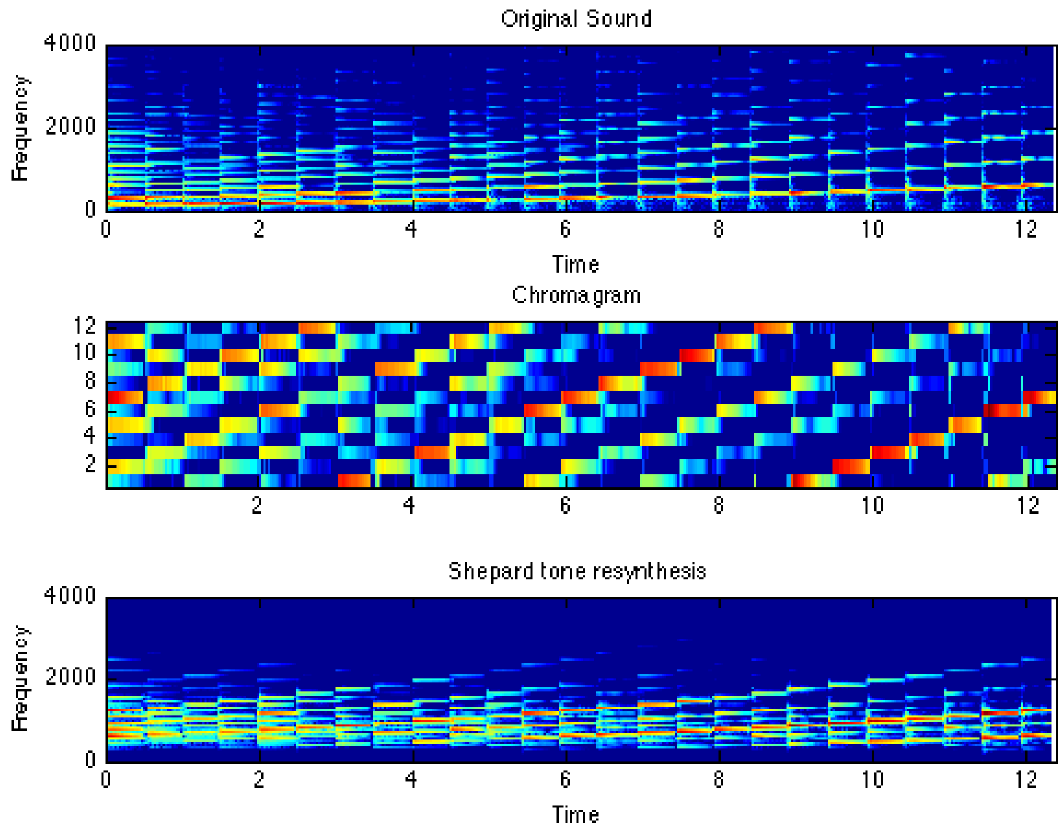


Audio Features

Tonal Mid-level Audio Features

Chromagram





Fraunhofer IDMT

Summary

- Sound categories
- Music representations
- Audio representations
- Audio signal decomposition
- Audio features

Thank you!

- Any questions?

Dr.-Ing. Jakob Abeßer
Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://machinelistening.github.io

Fraunhofer IDMT