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)

AUD-1

Fig. 1

Fig. 2

Fig. 3

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)

AUD-2

Fig. 4

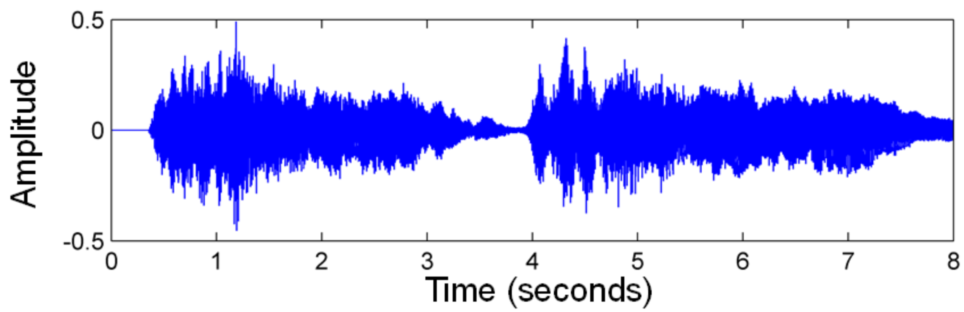
Fig. 5

Fig. 6

Fraunhofer IDMT

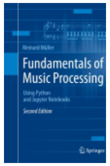
Music Representations
Recording & Notation

- Music recording (waveform)



- Music notation (score)

Fig. 7



FMP Notebooks

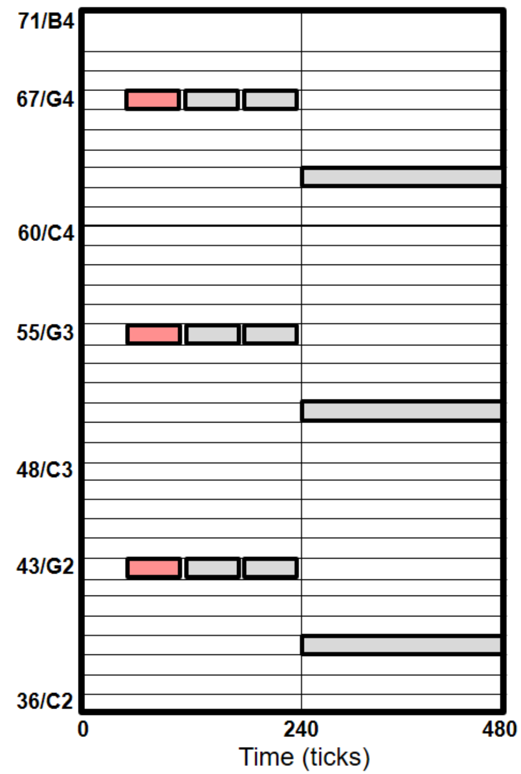
Fraunhofer IDMT

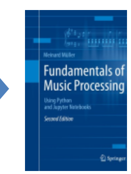
Music Representations

MIDI

- Sequence of note events (MIDI)

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





FMP Notebooks

Fig. 8

Fraunhofer IDMT

Music Representations
MusicXML

- Textual description of note events (MusicXML)

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

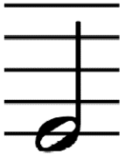


Fig. 9

Audio Representations
Short-term Fourier Transform (STFT)

- Discrete Short-term Fourier Transform (STFT)

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)e^{-2\pi ikn/N}$$

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

Audio Representations
Short-term Fourier Transform (STFT)

- Example: Sinosoid signal, two frequencies

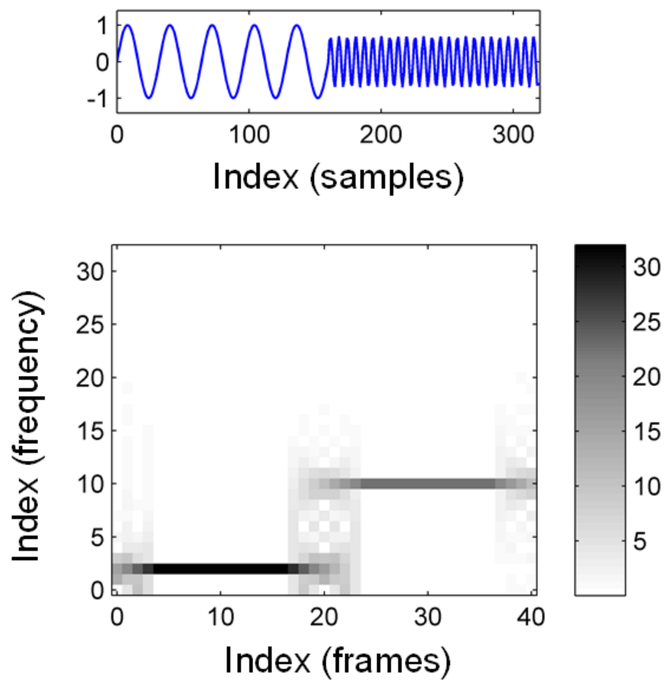
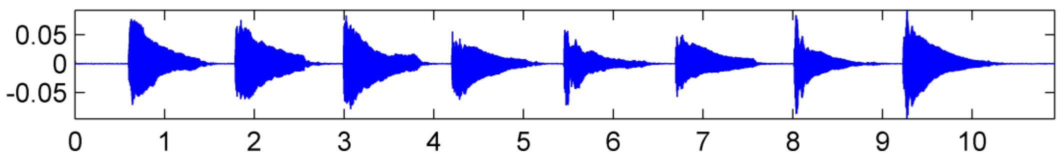
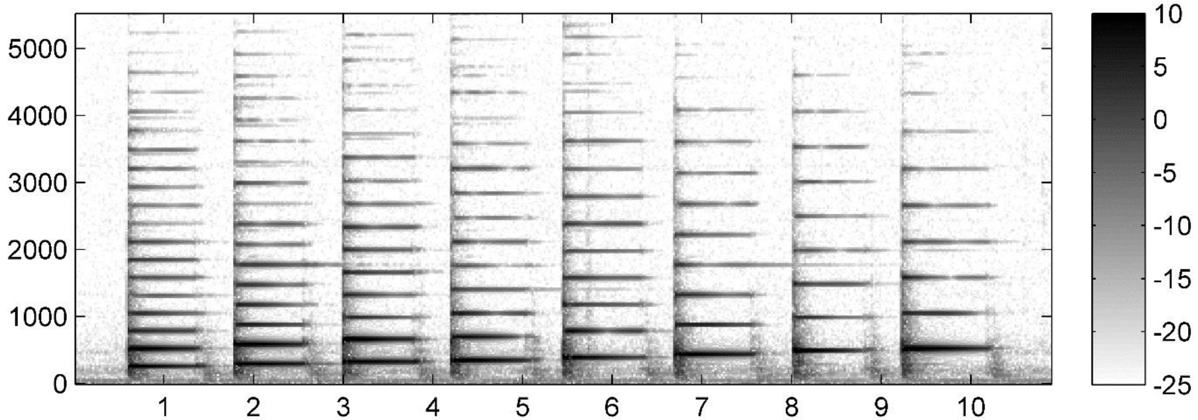


Fig. 10

Audio Representations
Short-term Fourier Transform (STFT)

- Example: C major scale, fundamental frequencies & overtones





FMP Notebooks

Fig. 11

Fraunhofer IDMT

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)

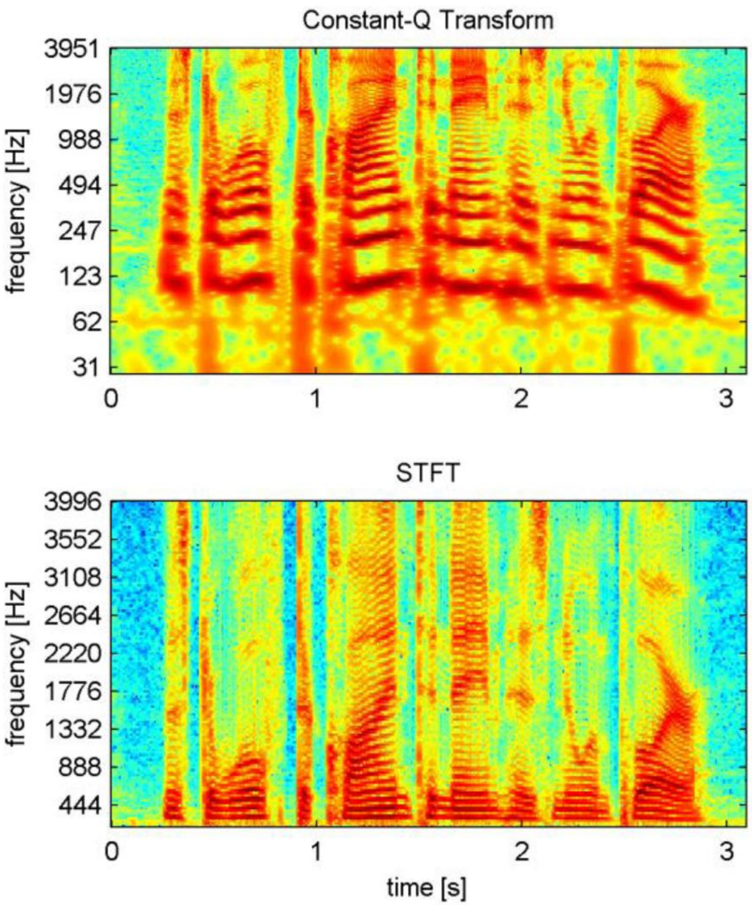


Fig. 12

Audio Representations
Mel Spectrogram

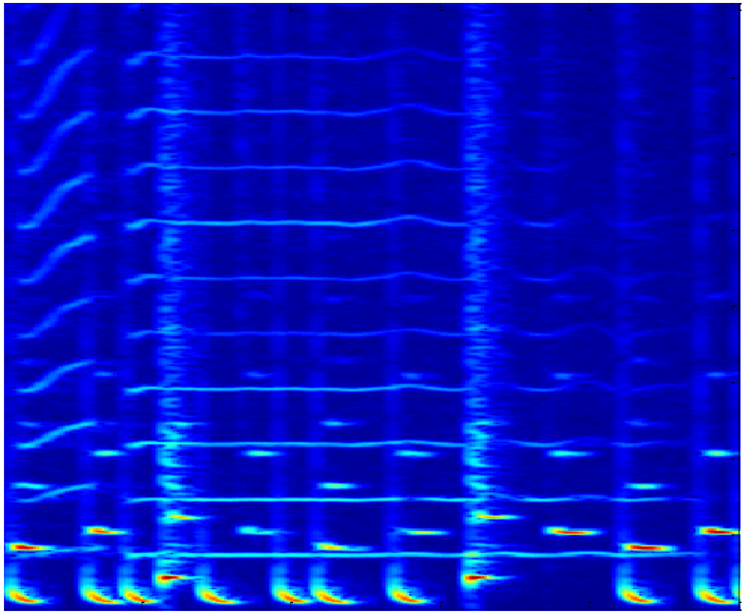
- Mel frequency scale (Stevens et al., 1937)

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

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

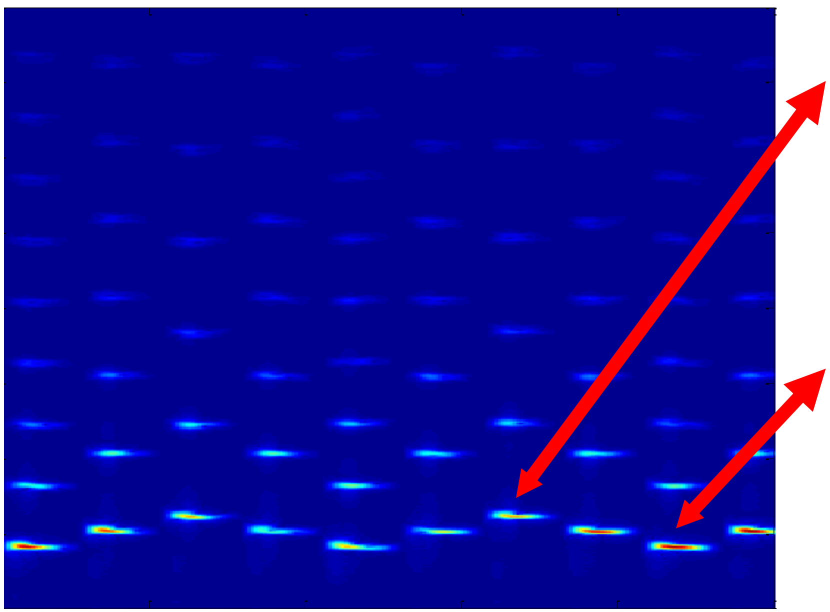
Audio Signal Decomposition
Music mixtures

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



Audio Signal Decomposition
Music mixtures

- Bass
  - Harmonic structure, stable tones

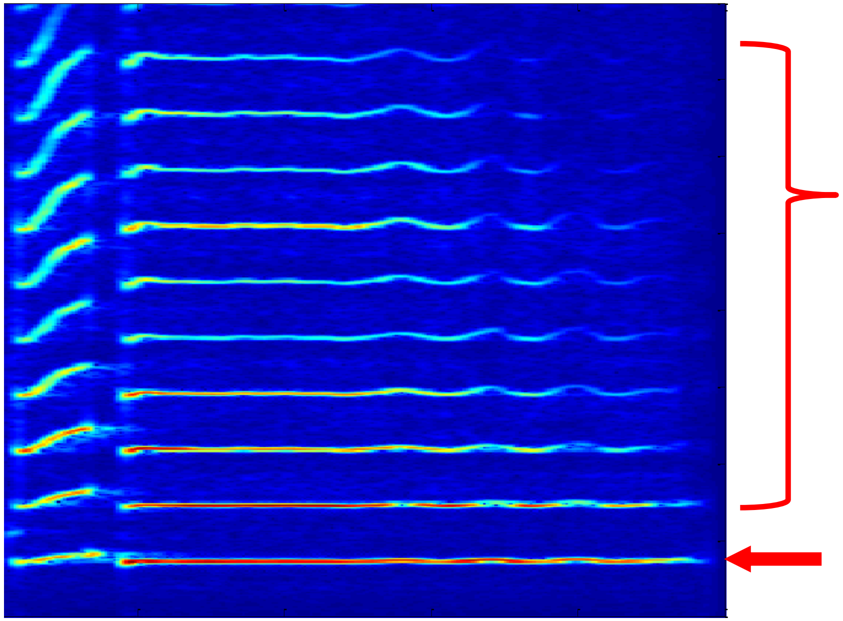


Higher pitch -
spectrum stretched

Lower pitch -
spectrum compressed

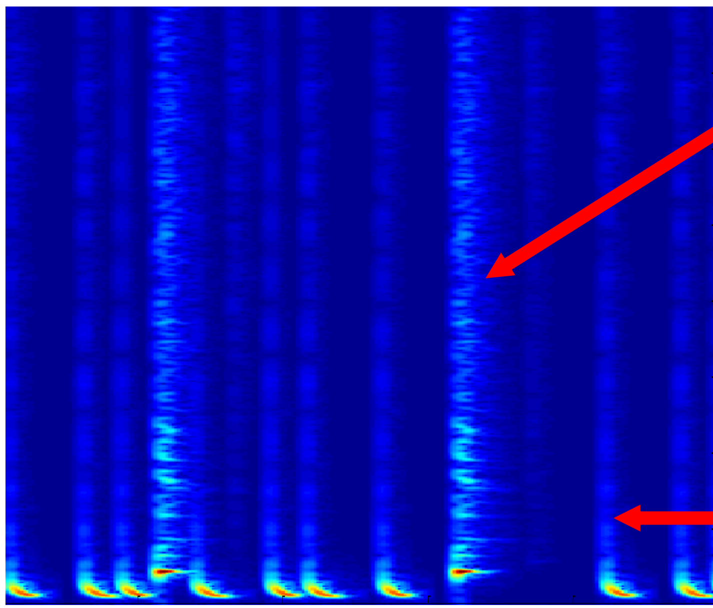
Audio Signal Decomposition
Music mixtures

- Melody (saxophone)
  - Harmonic components (melody)



Audio Signal Decomposition
Music mixtures

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

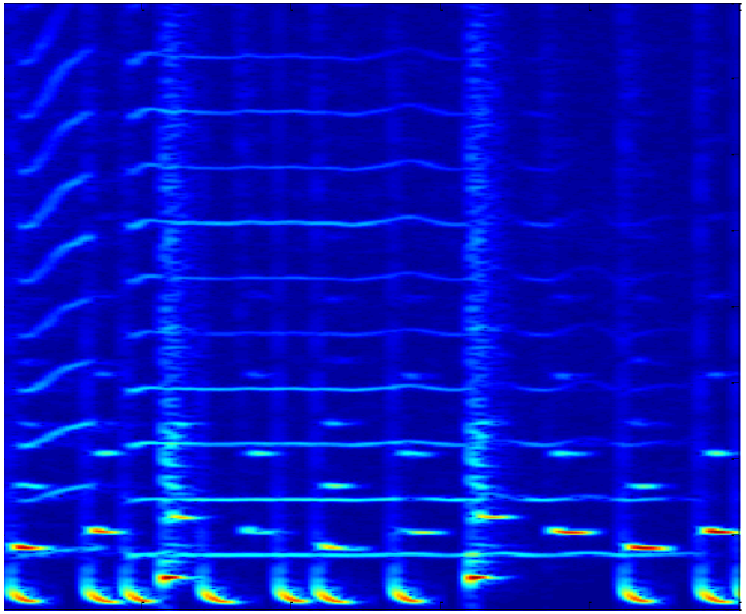


Snare drum (low frequencies)

Bass drum (low frequencies)

Audio Signal Decomposition
Music mixtures

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



Audio Features
Motivation

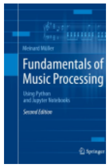
- Compact representation of audio signal for machine learning applications
- Capture different properties at different semantic levels
    - Timbre – perceived sound, instrumentation
    - Rhythm – tempo, meter
    - Melody/Tonality – pitches, harmonies
    - Structure - repetitions

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

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)



FMP Notebooks

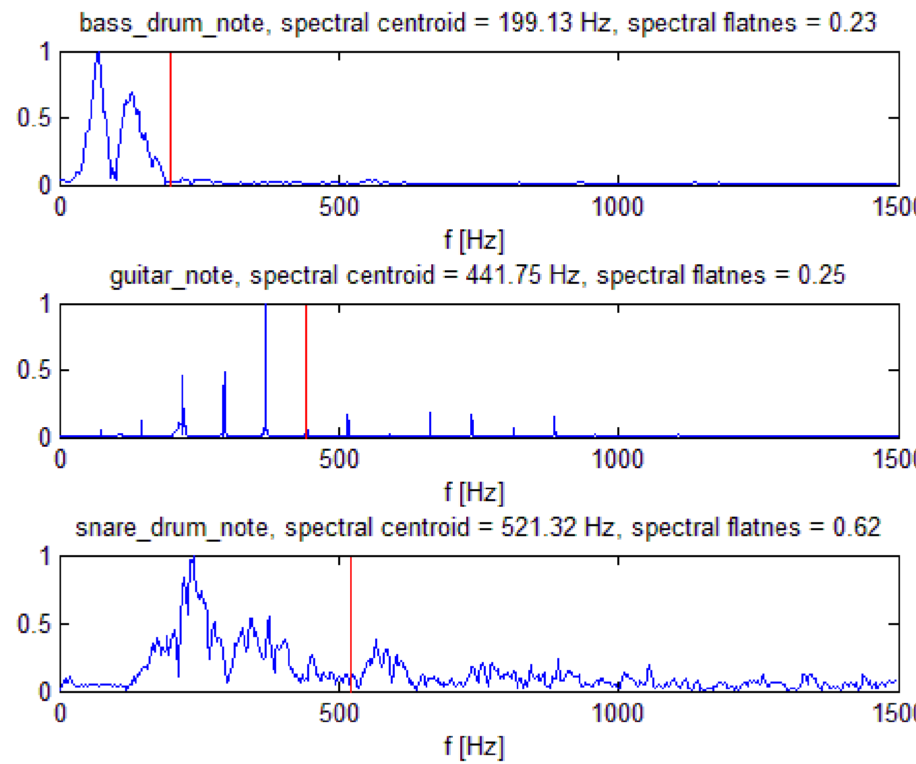
Audio Features
Timbre

- 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

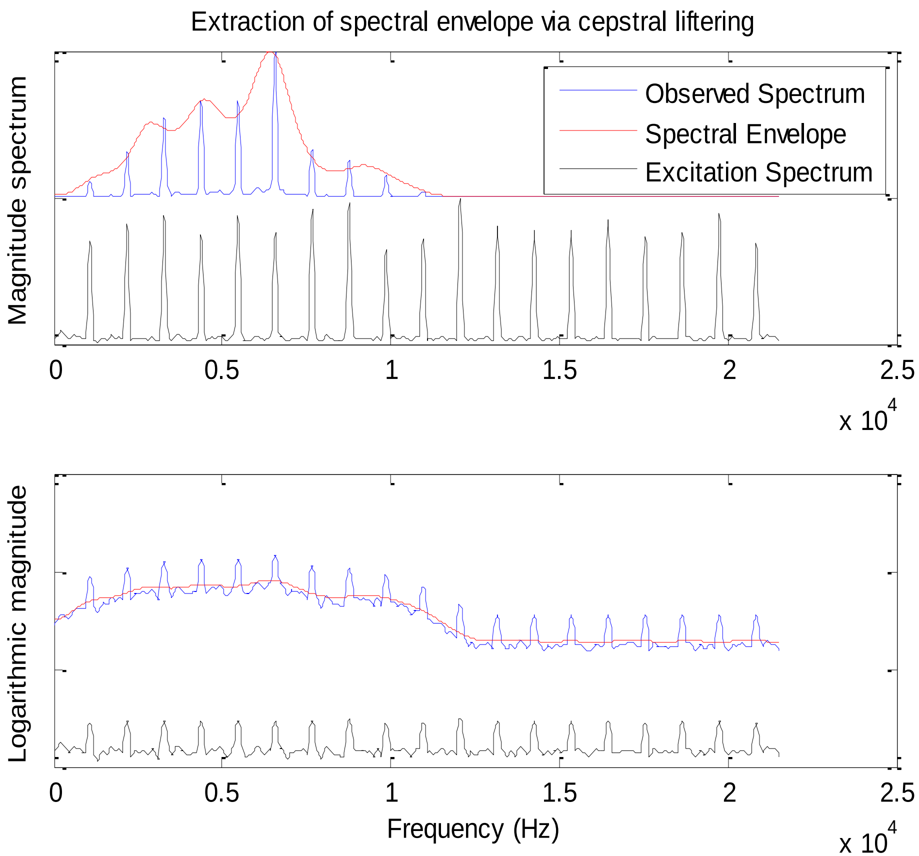
Timbre Mid-level Audio Features: MFCC

- 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)

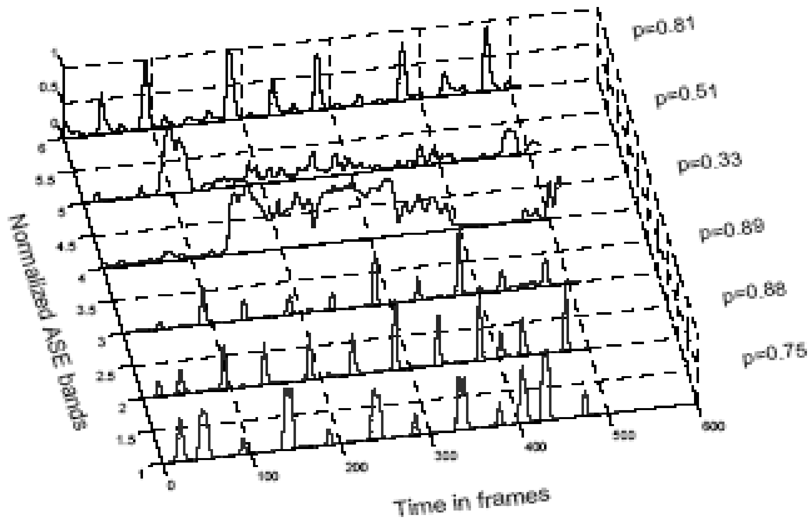
Audio Features
Timbre Mid-level Audio Features: MFCC

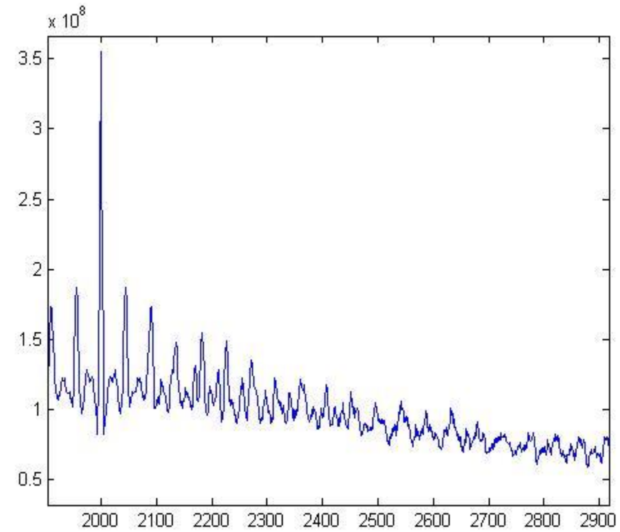
- Example



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

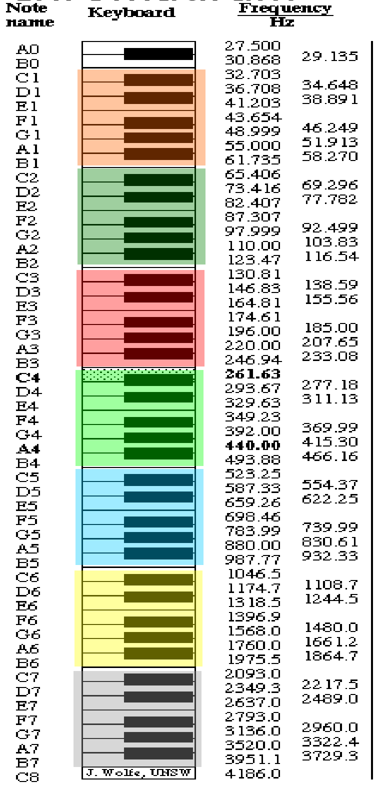


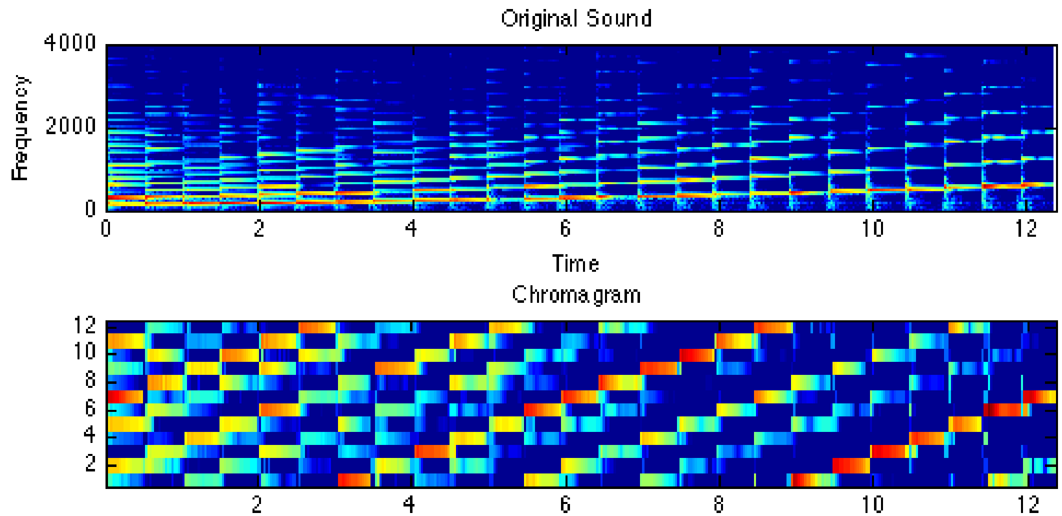


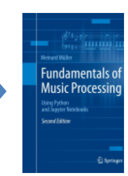
Audio Features

Tonal Mid-level Audio Features

Chromagram







FMP Notebooks

Fraunhofer IDMT

Summary

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

References

Müller, M. (2021). *Fundamentals of Music Processing - Using Python and Jupyter Notebooks* (2nd ed.). Springer.

Shi, Z., Lin, H., Liu, L., Liu, R., & Han, J. (2019). Is CQT More Suitable for Monaural Speech Separation than STFT? An Empirical Study. *ArXiv Preprint ArXiv:1902.00631*.

Images

Fig. 1: https://ccsearch-dev.creativecommons.org/photos/39451123-ee45-4ec3-ad8d-b42d856bca06

Fig. 2: https://ccsearch-dev.creativecommons.org/photos/c69d3b07-76bd-43e2-a44e-8742edc8447a

Fig. 3: https://ccsearch-dev.creativecommons.org/photos/ab3062ab-fe0f-420d-b93d-7451db166b4e

Fig. 4: https://ccsearch-dev.creativecommons.org/photos/a27a7541-45f5-4176-91a4-e2cb70eea266

Fig. 5: https://ccsearch-dev.creativecommons.org/photos/79d466c1-cfa6-417e-9832-34438678bf5d

Fig. 6: https://ccsearch-dev.creativecommons.org/photos/269394a4-5803-47fd-abaa-57ef92735e24

Fig. 7: [Müller, 2021], p. 2, Fig. 1.1

Fig. 8: [Müller, 2021], p. 14, Fig. 1.13

Fig. 9: [Müller, 2021], p. 17, Fig. 1.15

Fig. 10: [Müller, 2021], p. 56, Fig. 2.9

Fig. 11: [Müller, 2021], p. 57, Fig. 2.10

Fig. 12: [Shi, 2019], p. 3, Fig. 2

Sounds

AUD-1: Medley: https://freesound.org/people/InspectorJ/sounds/416529, https://freesound.org/people/prometheus888/sounds/458461, https://freesound.org/people/MrAuralization/sounds/317361

AUD-2: Medley: https://freesound.org/people/whatsanickname4u/sounds/127337, https://freesound.org/people/jcveliz/sounds/92002, https://freesound.org/people/klankbeeld/sounds/192691

Thank you!

- Any questions?

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://machinelistening.github.io