

---

# Machine Listening for Music and Sound Analysis

## Lecture 1 – Audio Representations

---

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

<https://www.machinelistinging.de>

---

# Learning Objectives

---

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

# Sound Categories

## Environmental Sounds

- Sound sources
  - Animals, 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

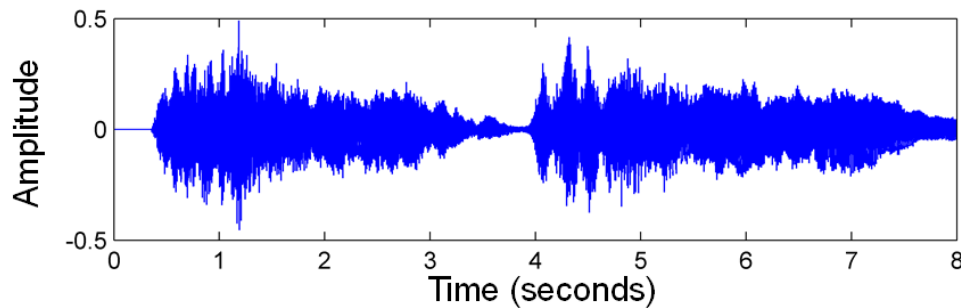


Fig. 6

# Music Representations

## Recording & Notation

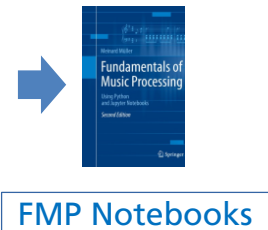
### ■ Music recording (waveform)



### ■ Music notation (score)



Fig. 7



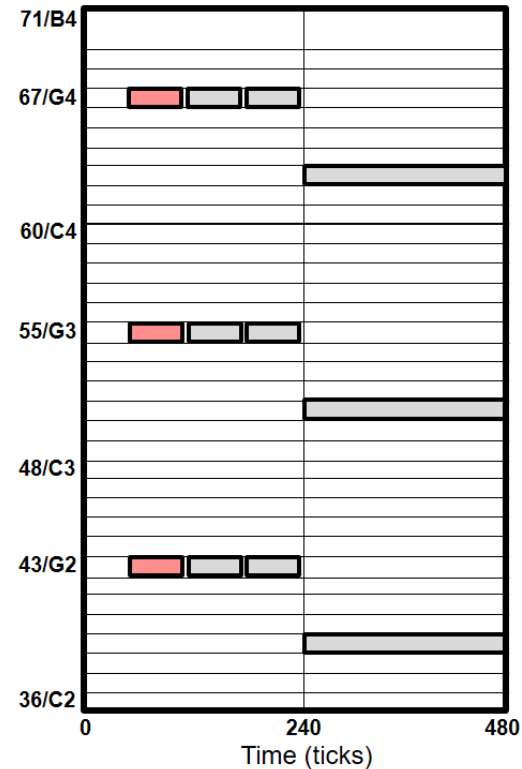
# Music Representations

## 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

# 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)

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

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



# Audio Representations

## Short-term Fourier Transform (STFT)

- Example: Sinusoid signal, two frequencies

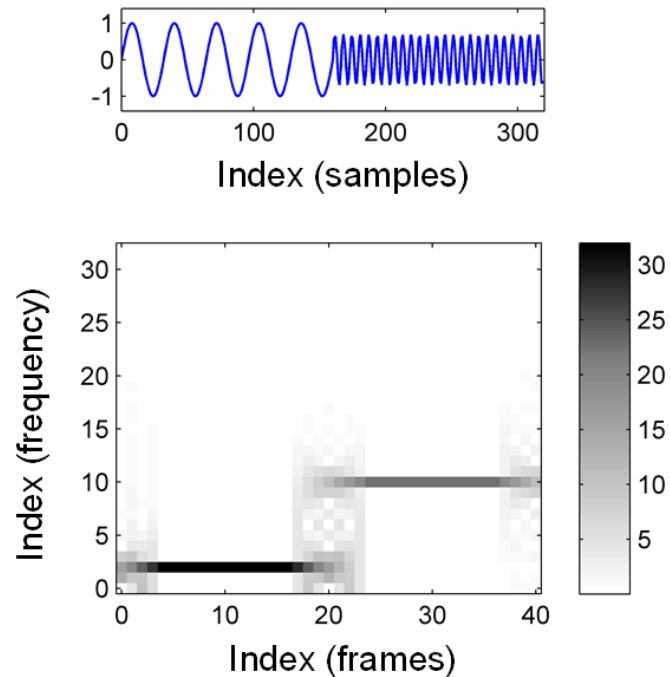
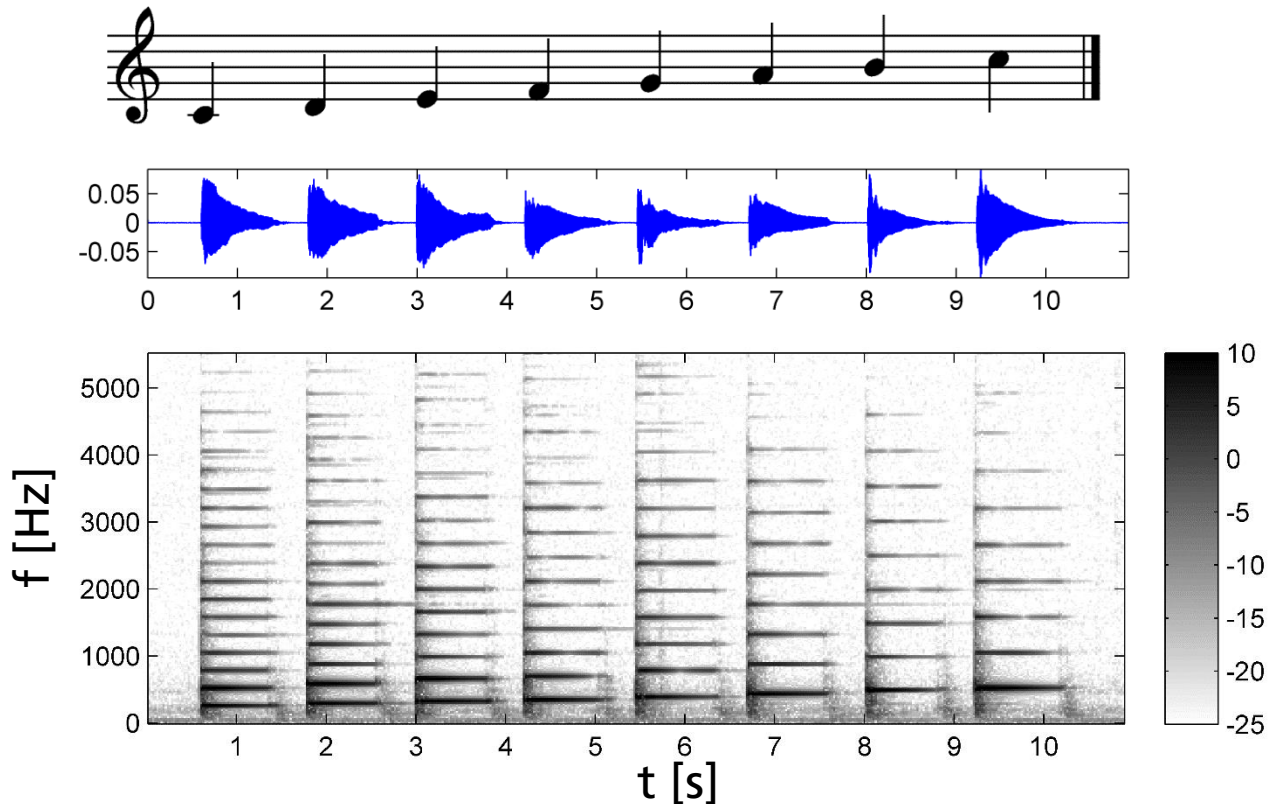


Fig. 10

# Audio Representations

## Short-term Fourier Transform (STFT)

- Example: C major scale, fundamental frequencies ( $f_0$ ) & overtones



FMP Notebooks

Fig. 11

---

# 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

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

$m$ : MIDI pitch

A4 (440 Hz): reference pitch

# Audio Representations

## Constant-Q Transform (CQT)

- Example signal (speech)
- CQT (top)
- STFT (bottom)

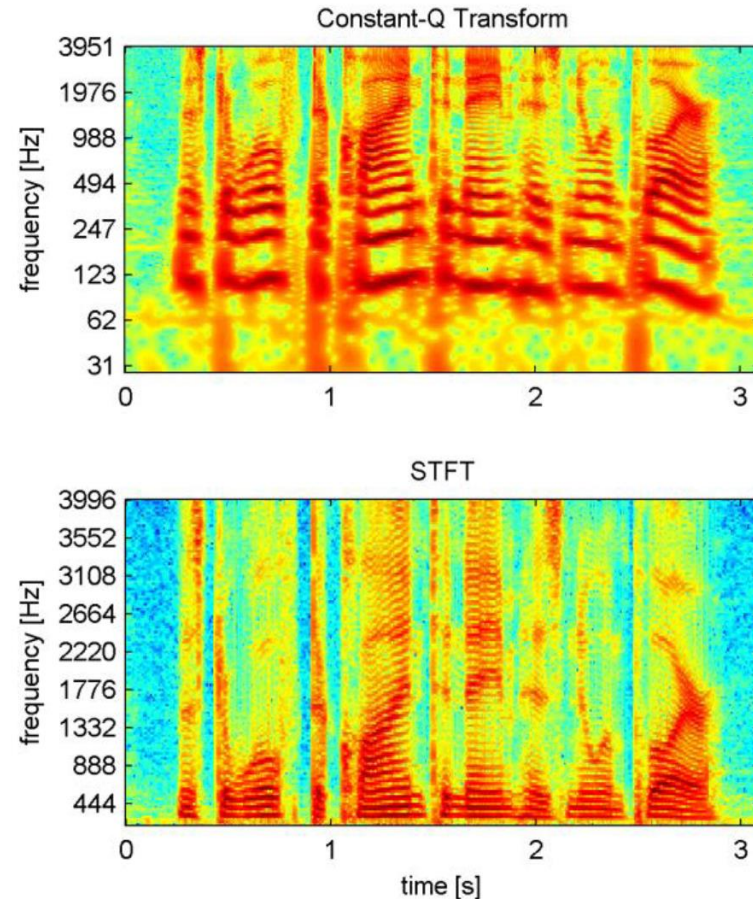


Fig. 12

---

# Audio Representations

## Mel Spectrogram

---

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

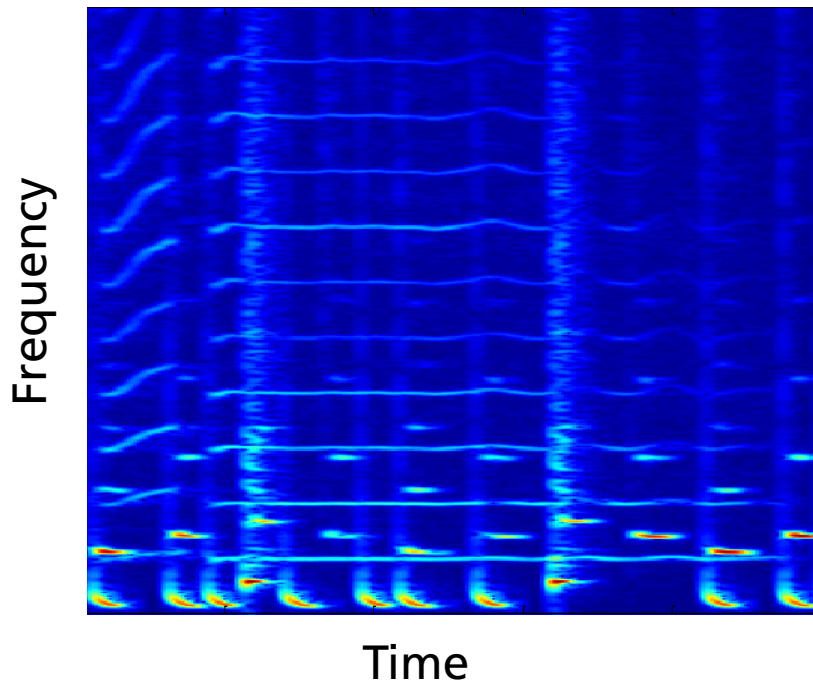
$$f[\text{Mel}] = 2595 \cdot \log_{10}\left(1 + \frac{f[\text{Hz}]}{700}\right)$$

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

# Audio Signal Decomposition

## Music mixtures

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

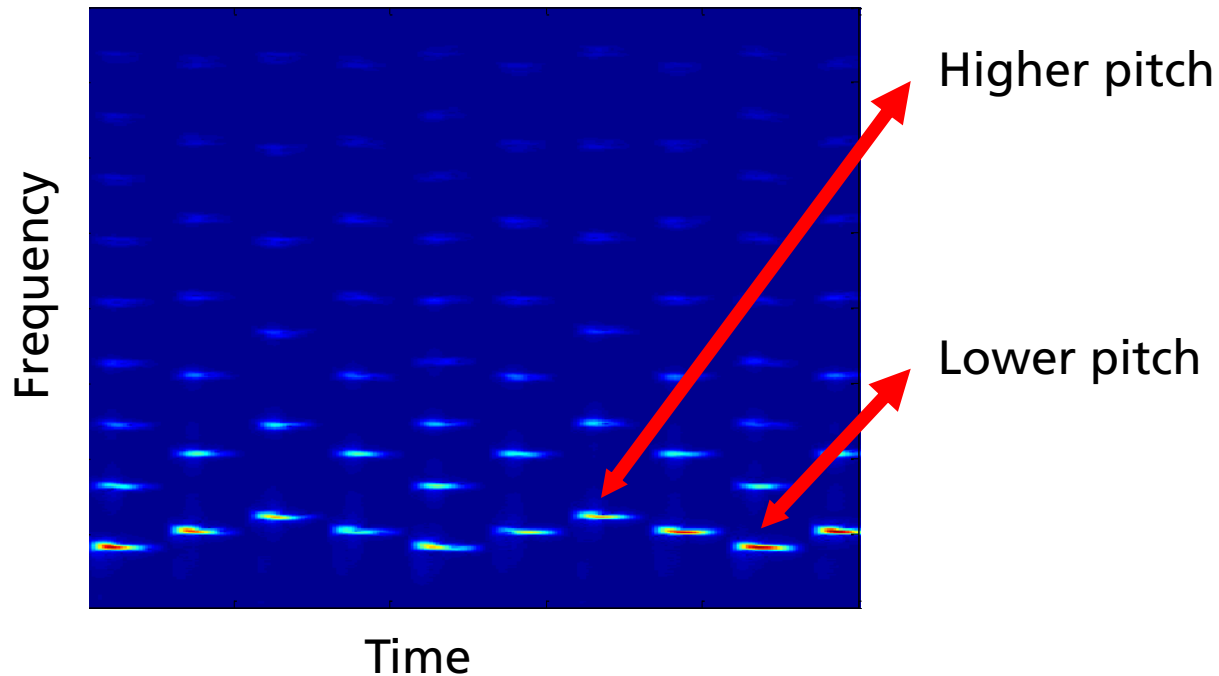


# Audio Signal Decomposition

## Music mixtures

■ Bass

■ Harmonic structure, stable tones

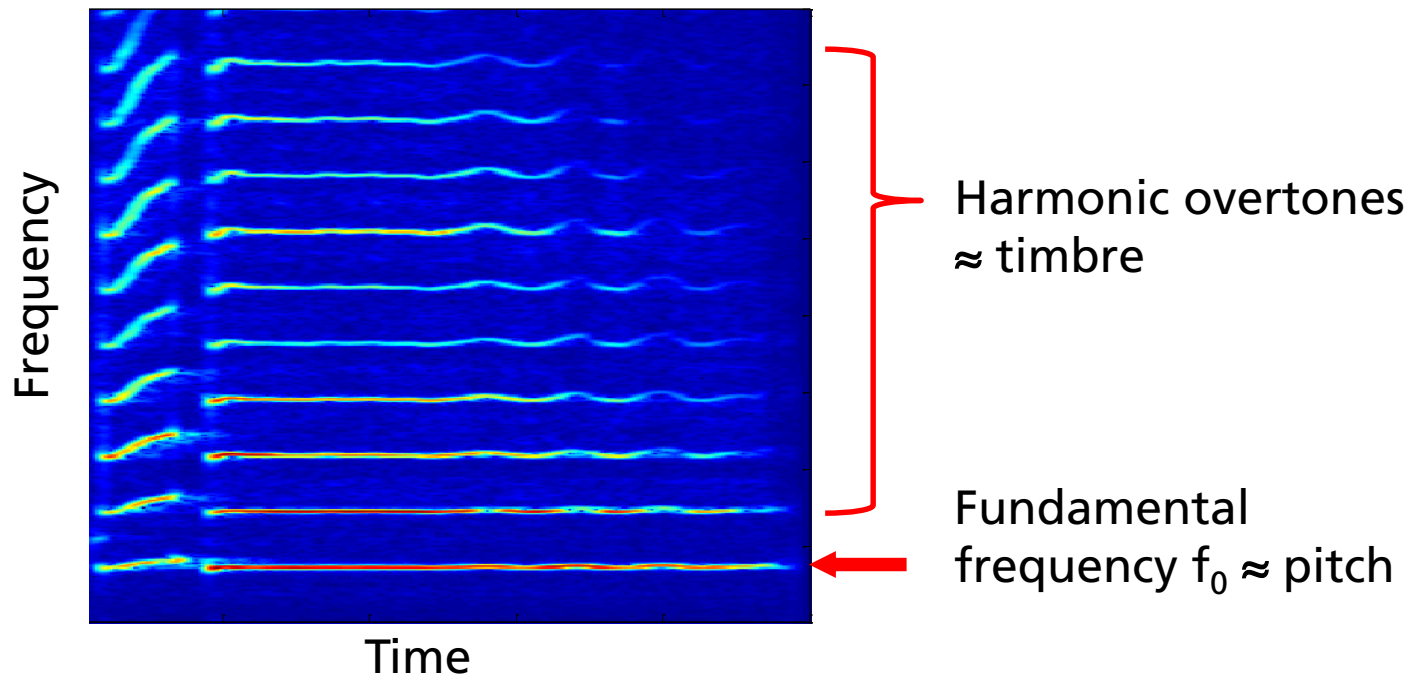




# Audio Signal Decomposition

## Music mixtures

- Melody (saxophone)
  - Harmonic components (melody)

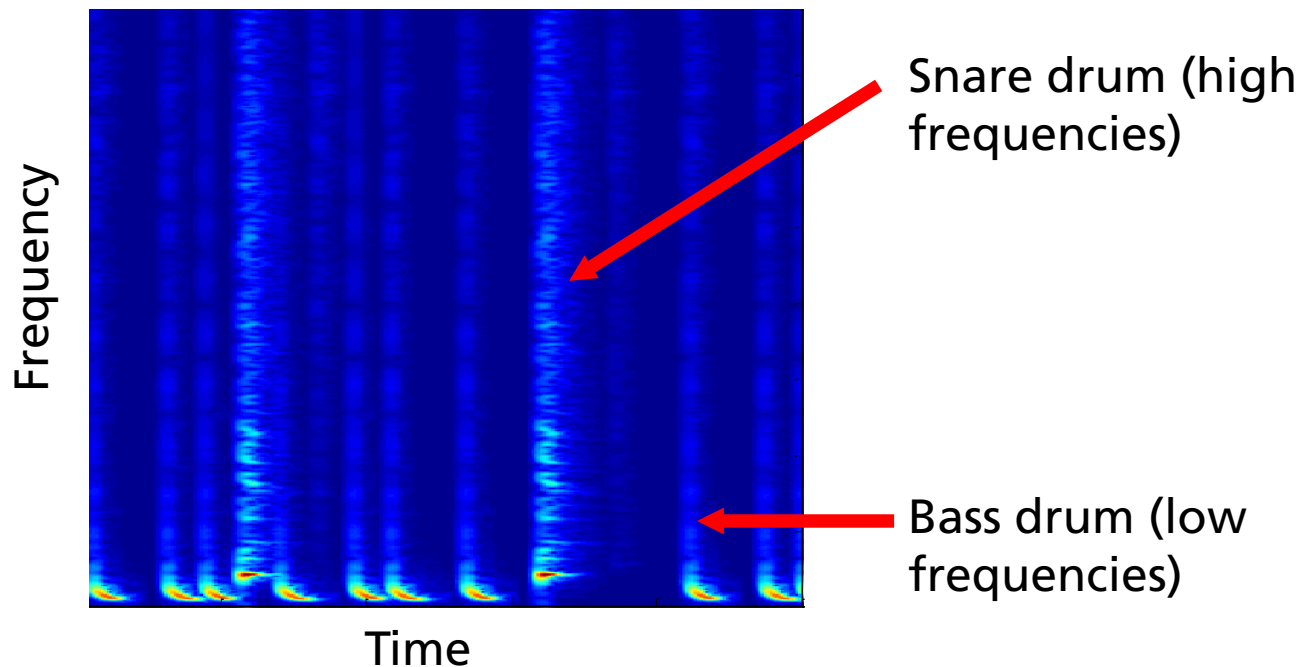


# Audio Signal Decomposition

## Music mixtures

- Drums

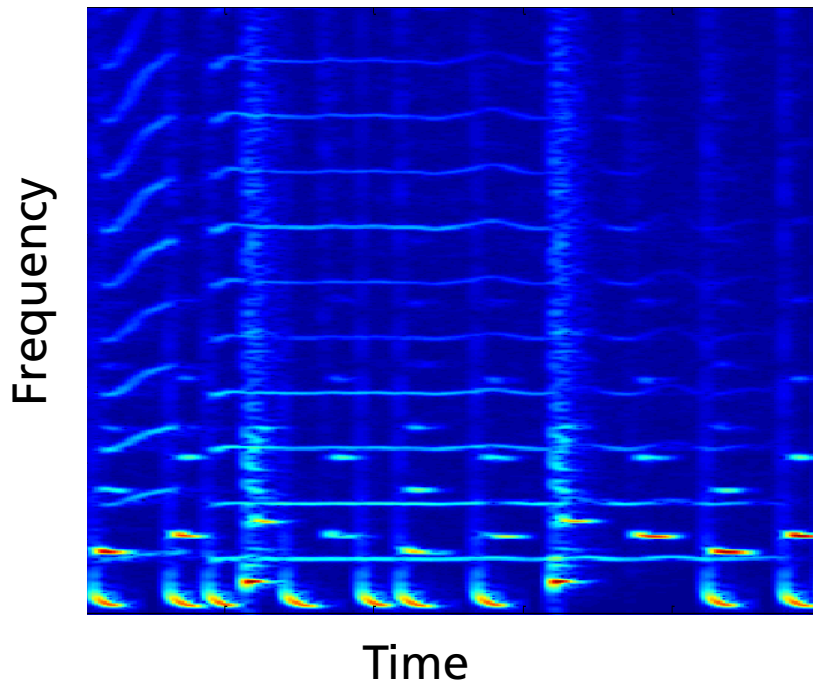
- Percussive components (noise-like, inharmonic spectra)



# Audio Signal Decomposition

## Music mixtures

- Instrument mixture (magnitude STFT)
  - All components add up



---

# 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, novelty, homogeneous segments

# Audio Features

## Categorization

	Timbre	Rhythm	Tonality
Low-Level (Q~10 ms)	<ul style="list-style-type: none"><li>- Zero Crossing Rate (ZCR)</li><li>- Linear Predictive Coding (LPC)</li><li>- Spectral Centroid / Spectral Flatness</li></ul>		
Mid-Level (Q ~ 2.5s)	<ul style="list-style-type: none"><li>- Mel-Frequency Cepstral Coefficients (MFCC)</li><li>- Octave-Based Spectral Contrast (OSC)</li><li>- Loudness</li></ul>	<ul style="list-style-type: none"><li>- Tempogram</li><li>- Log-Lag Autocorrelation (ACF)</li></ul>	<ul style="list-style-type: none"><li>- Chromagram</li><li>- Enhanced Pitch Class Profiles (EPCP)</li></ul>
High-Level	<ul style="list-style-type: none"><li>- Instrumentation</li></ul>	<ul style="list-style-type: none"><li>- Tempo</li><li>- Time Signature</li><li>- Rhythm Patterns</li></ul>	<ul style="list-style-type: none"><li>- Key</li><li>- Scales</li><li>- Chords</li></ul>

---

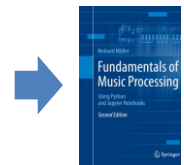
# Audio Features

## 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 (non-integer relationship between partials)
  - Variations over time: frequency (vibrato) or loudness (tremolo)



FMP Notebooks

---

# Audio Features

## Timbre

---

- Timbre
  - When looking at musical instruments, we need to consider
    - Instrument's construction
    - Sound production principles
      - Membranophones, chordophones, aerophones, electrophones
    - Human performance
      - Playing techniques, expressivity, dynamics, style
- How to design features to quantify these acoustic phenomena?

# Audio Features

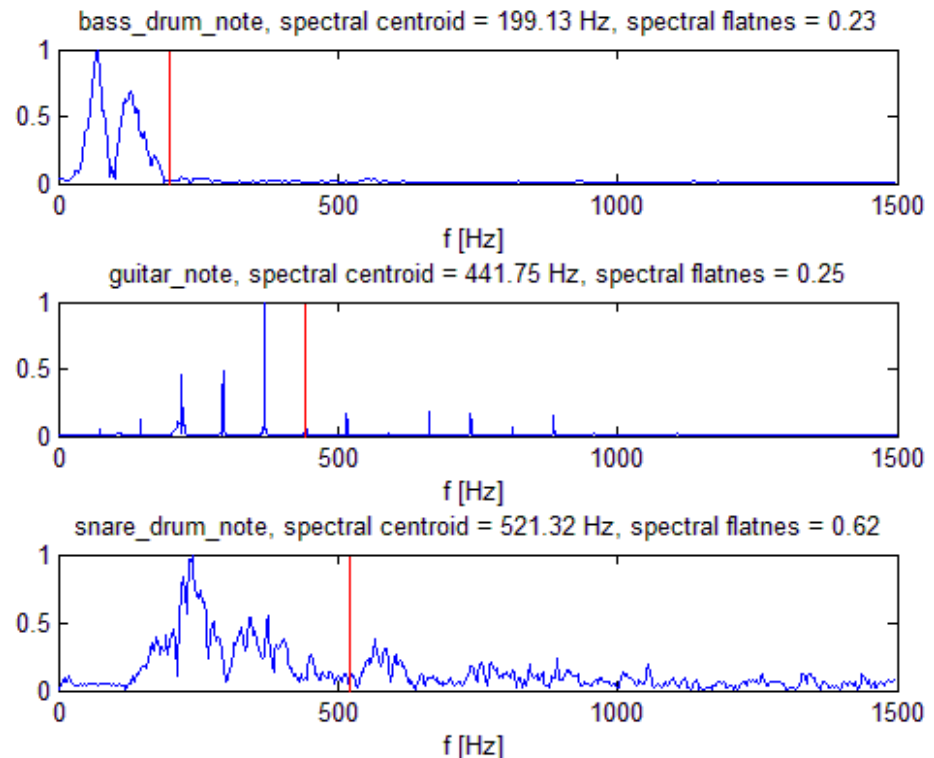
## Timbre Low-level Audio Features

### ■ Spectral Centroid (SC):

- Center of mass in the magnitude spectrogram
- Low-pitched vs. high-pitched sounds

### ■ Spectral Flatness Measure (SFM)

- Harmonic sounds (sparse energy distribution)
- Percussive sounds (wideband energy distribution)

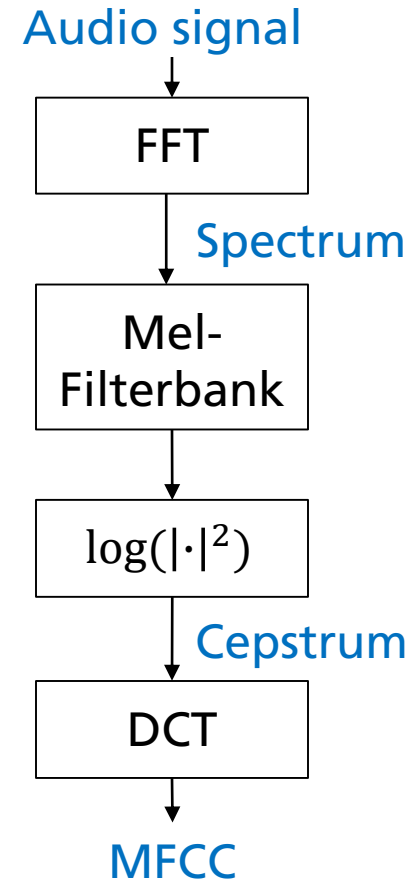




# Audio Features

## Timbre Mid-level Audio Features: MFCC

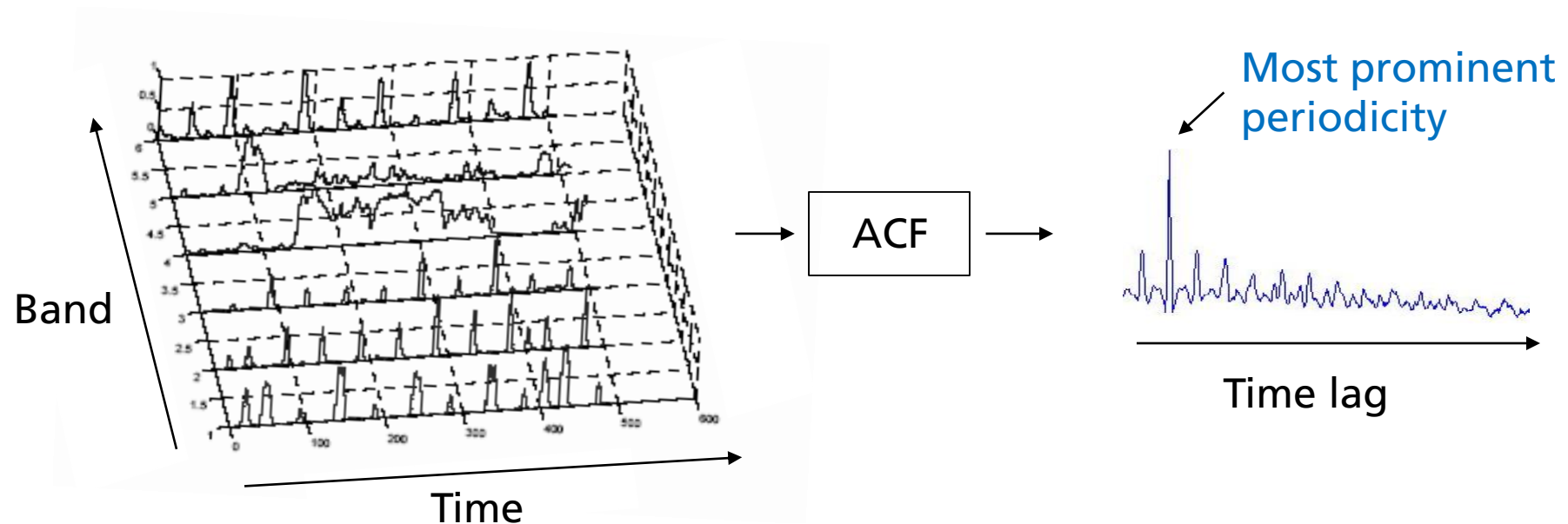
- Convolutional **excitation** \* **filter** model
  - Excitation: vibration of vocal folds
  - Filter: resonance of the vocal tract
- FFT magnitude spectrum
  - Multiplicative **excitation** · **filter** model
- Logarithm of magnitude spectrum
  - Additive **excitation** + **filter** model
- Discrete Cosine Transform (DCT)
  - First coefficients allow for a compact description of the spectral envelope shape



# Audio Features

## Rhythmic Mid-level Audio Features

- Rhythmic properties are important for audio classification
- Audio Spectral Energy (ASE)
  - Analyze energy slopes in different frequency bands
  - Find periodicities via auto-correlation function (ACF)



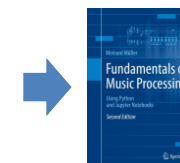
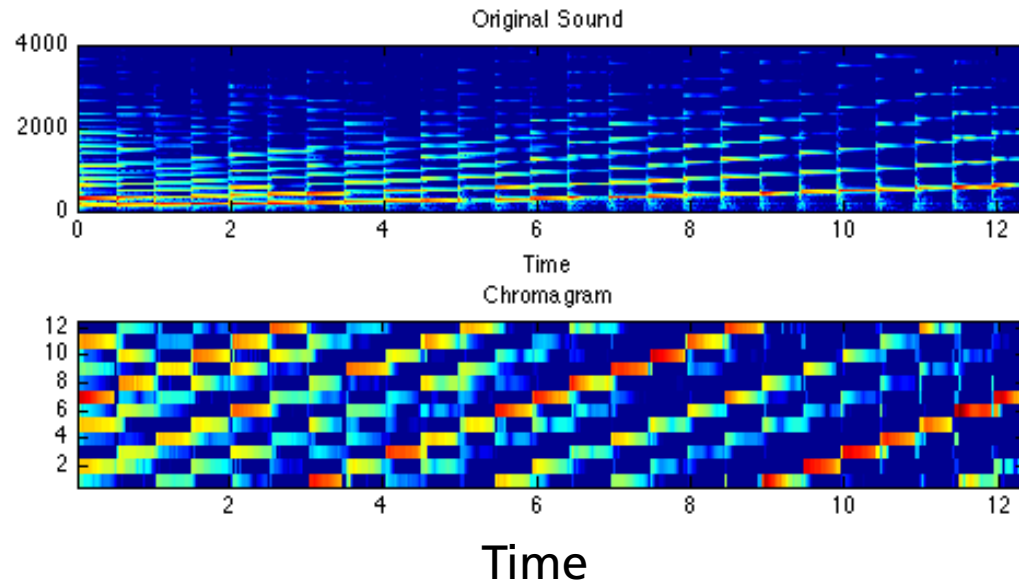
# Audio Features

## Tonal Mid-level Audio Features: Chromagram

Note name	Keyboard	Frequency Hz
A0		27.500
B0		30.868
C1		32.703
D1		36.708
E1		41.203
F1		43.654
G1		48.999
A1		55.000
B1		61.735
C2		65.406
D2		73.416
E2		82.407
F2		87.307
G2		97.999
A2		110.00
B2		123.47
C3		130.81
D3		146.83
E3		164.81
F3		174.61
G3		196.00
A3		220.00
B3		246.94
C4		<b>261.63</b>
D4		293.67
E4		329.63
F4		349.23
G4		392.00
A4		<b>440.00</b>
B4		493.88
C5		523.25
D5		587.33
E5		659.26
F5		698.46
G5		783.99
A5		880.00
B5		987.77
C6		1046.5
D6		1174.7
E6		1318.5
F6		1396.9
G6		1568.0
A6		1760.0
B6		1975.5
C7		2093.0
D7		2349.3
E7		2637.0
F7		2793.0
G7		3136.0
A7		3520.0
B7		3951.1
C8		4186.0

J. Wolfe, UNSW

Frequency  
Pitch Class



FMP Notebooks

---

# 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://www.machinelistinging.de>