
Machine Listening for Music and Sound Analysis

Lecture 1 – Audio Representations

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

<https://www.machinelisting.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

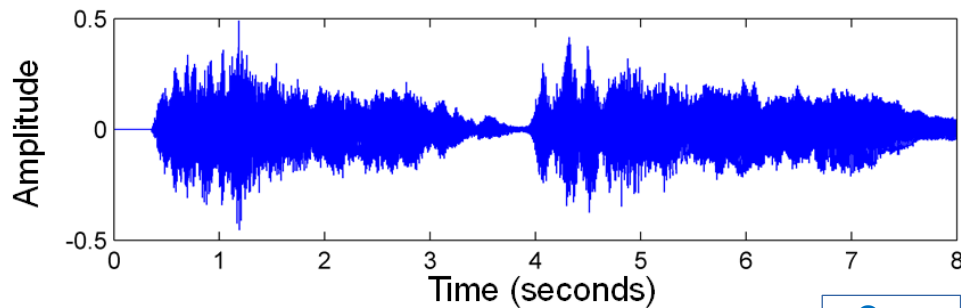


Fig. 6

Music Representations

Recording & Notation

■ Music recording (waveform)

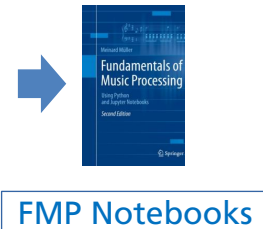


Own

■ Music notation (score)



Fig. 7



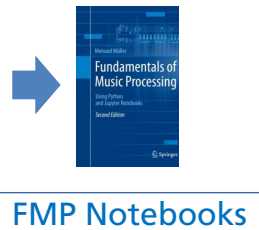
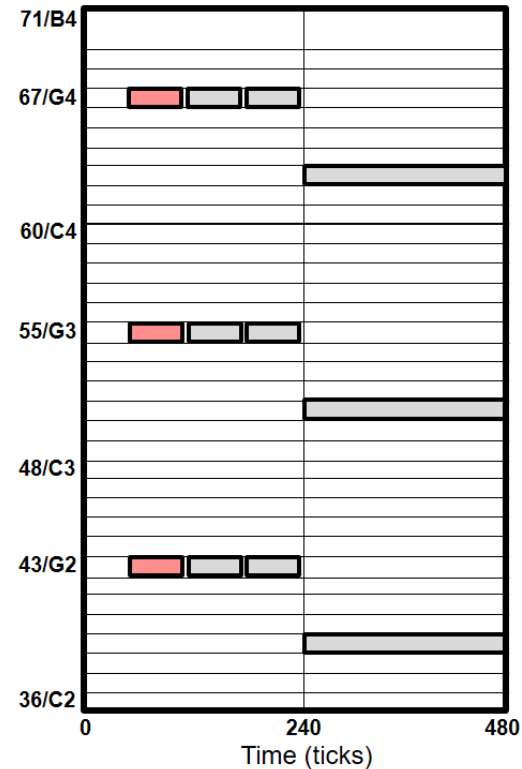
Music Representations

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)
- Windowed analysis of audio signals

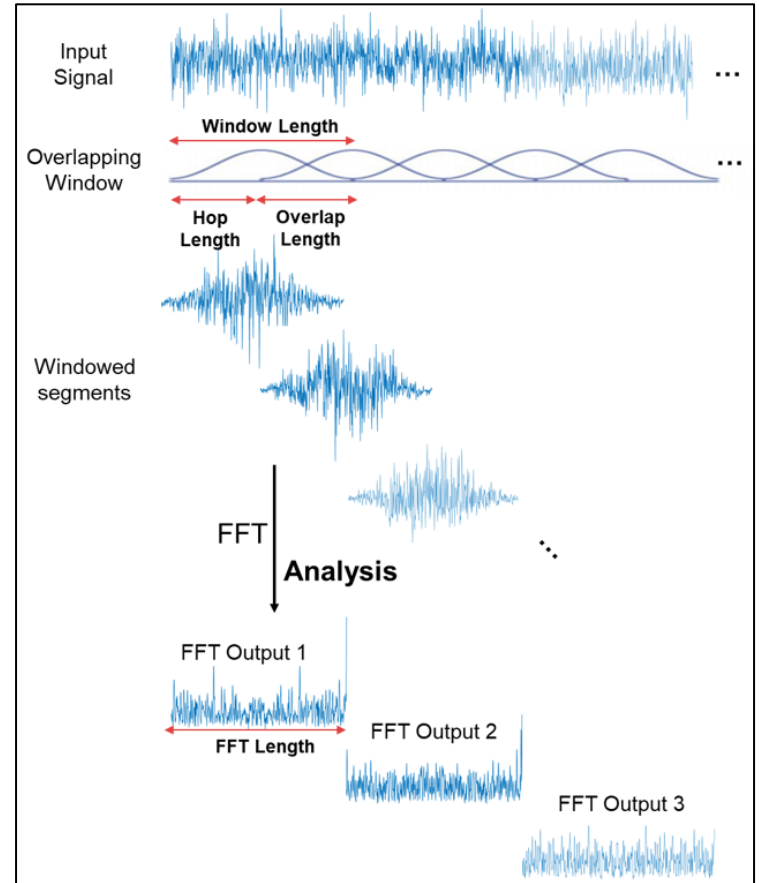


Fig. 9.5

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

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
- Linearly-spaced frequency axis
- Trade-off between
 - Frequency resolution
 - Time resolution

Audio Representations

Short-term Fourier Transform (STFT)

- Example: Sinusoid signal, two frequencies

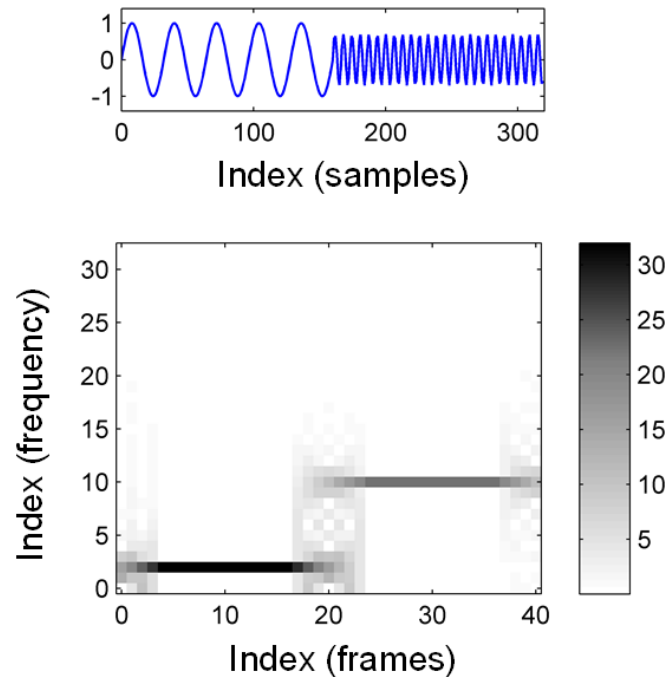


Fig. 10

Audio Representations

Short-term Fourier Transform (STFT)

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

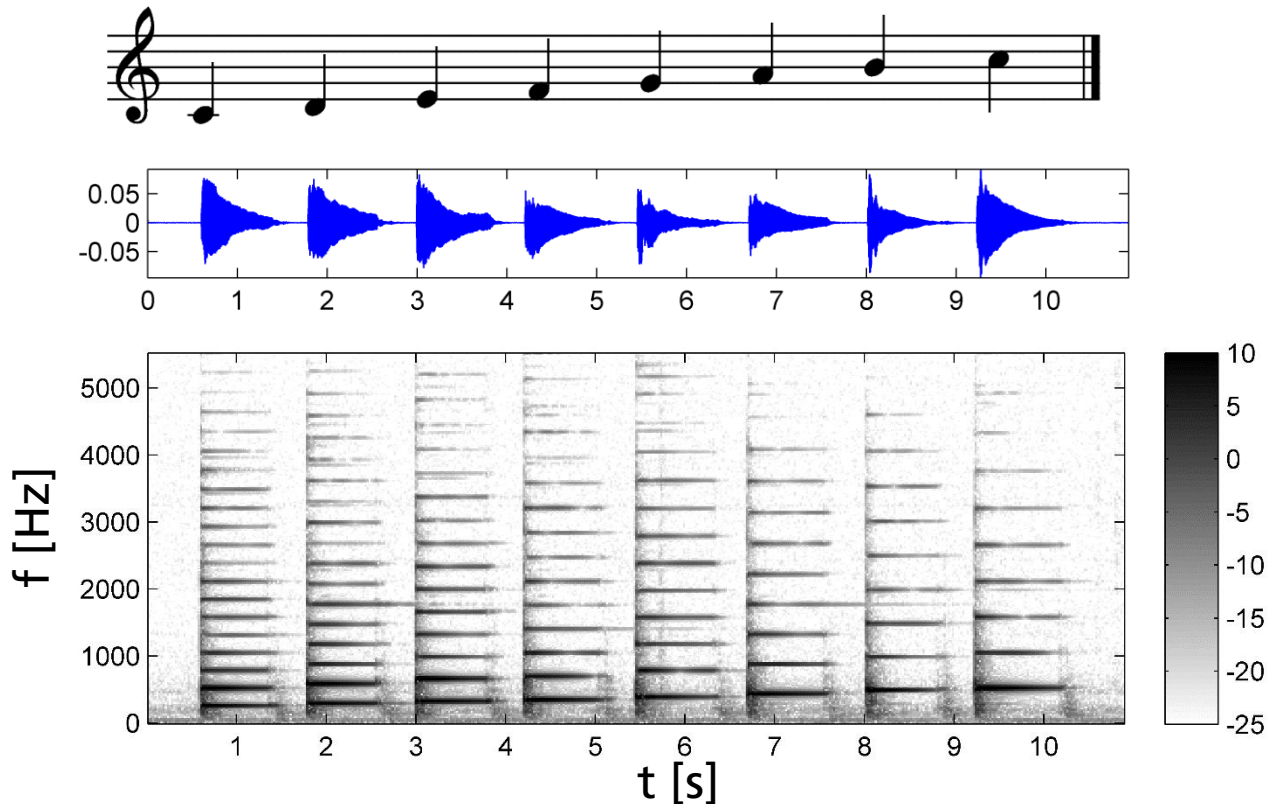


Fig. 11

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

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

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

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

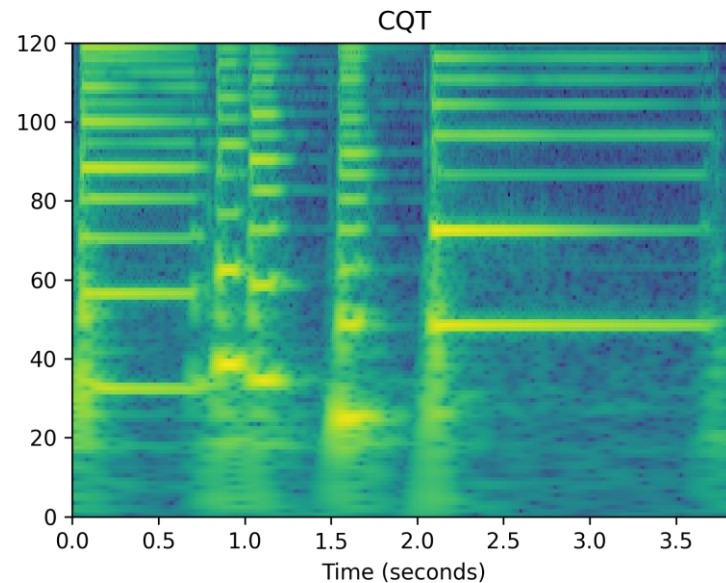
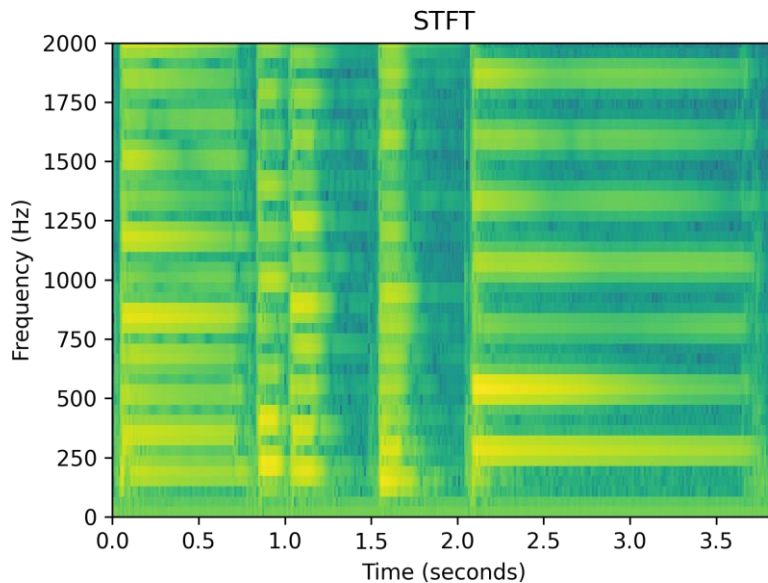
Constant-Q Transform (CQT)

- STFT (linearly-spaced frequencies)
- CQT (logarithmically-spaced, closer to human auditory perception)
 - Variable number of frequency bins per octave
 - Increasing time resolution towards higher frequencies

Audio Representations

Constant-Q Transform (CQT)

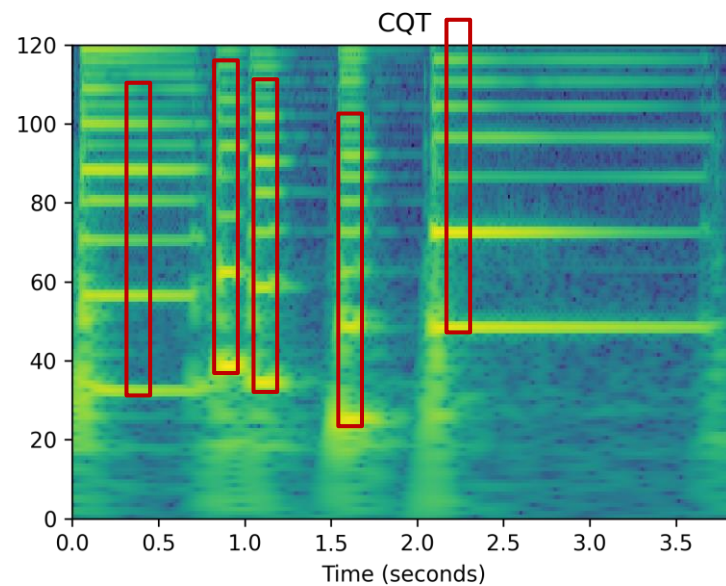
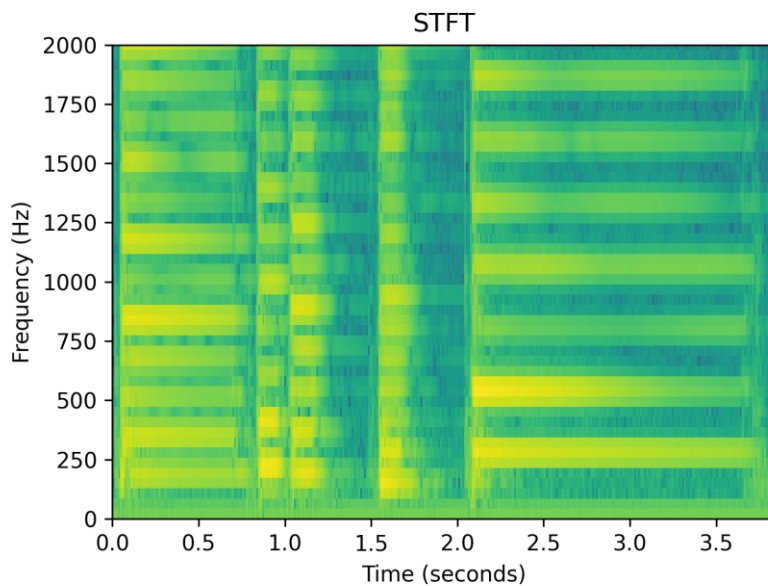
- STFT (linearly-spaced frequencies)
- CQT (logarithmically-spaced, closer to human auditory perception)
 - Variable number of frequency bins per octave
 - Increasing time resolution towards higher frequencies



Audio Representations

Constant-Q Transform (CQT)

- Suitable for music transcription
- Partial have a constant frequency pattern
 - Vertically shifted
 - Pitch-independent



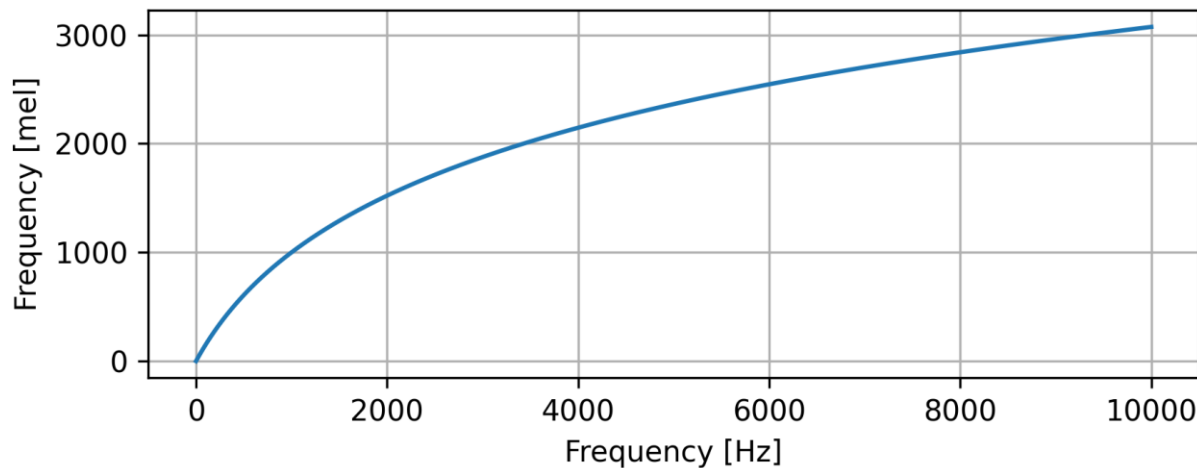
Audio 1

Audio Representations

Mel Spectrogram

- Logarithmic frequency mapping (human pitch perception)

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



Audio Representations

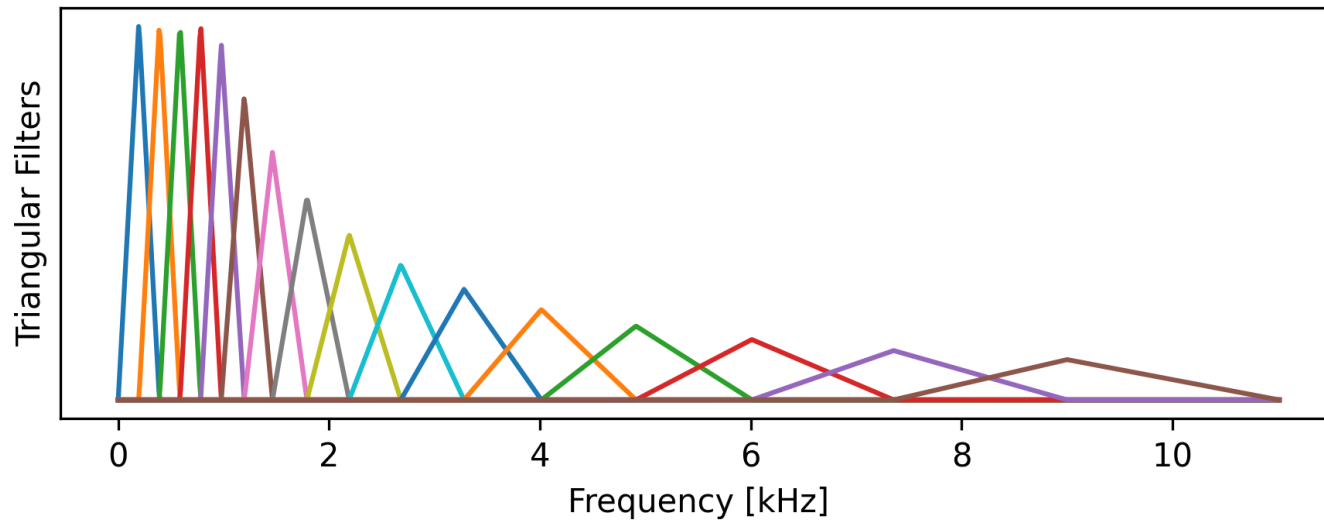
Mel Spectrogram

- Mapping from STFT magnitude spectrogram to Mel spectrogram
 - Triangular filterbank + Matrix multiplication

Audio Representations

Mel Spectrogram

- Mapping from STFT magnitude spectrogram to Mel spectrogram
 - Triangular filterbank + Matrix multiplication
- Example: 16 mel bands, $f_s = 22.05$ kHz



Audio Representations

Mel Spectrogram

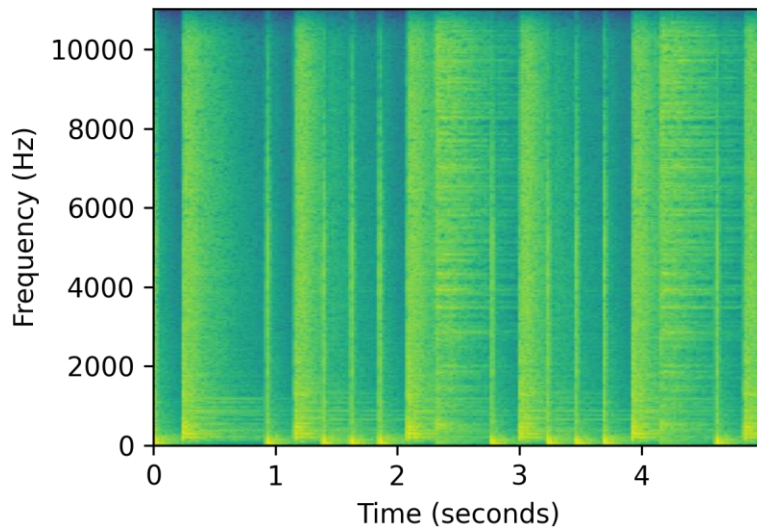
- More efficient representation (fewer frequency bands)
- Still captures perceptually relevant information

Audio Representations

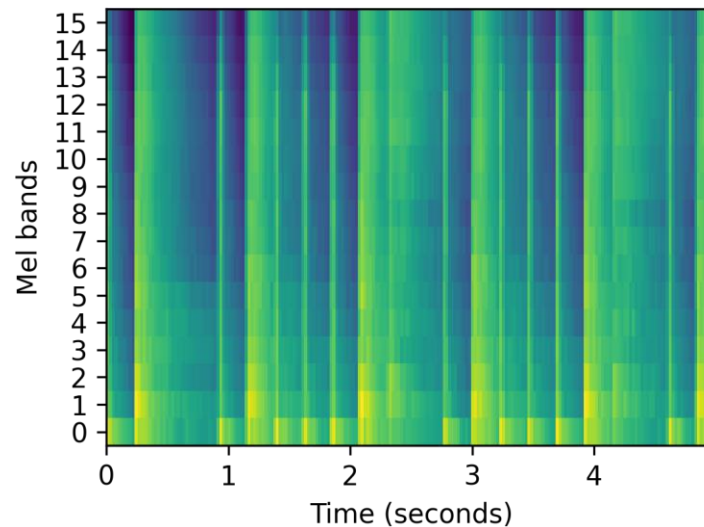
Mel Spectrogram

- More efficient representation (fewer frequency bands)
- Still captures perceptually relevant information

STFT



Mel Spectrogram



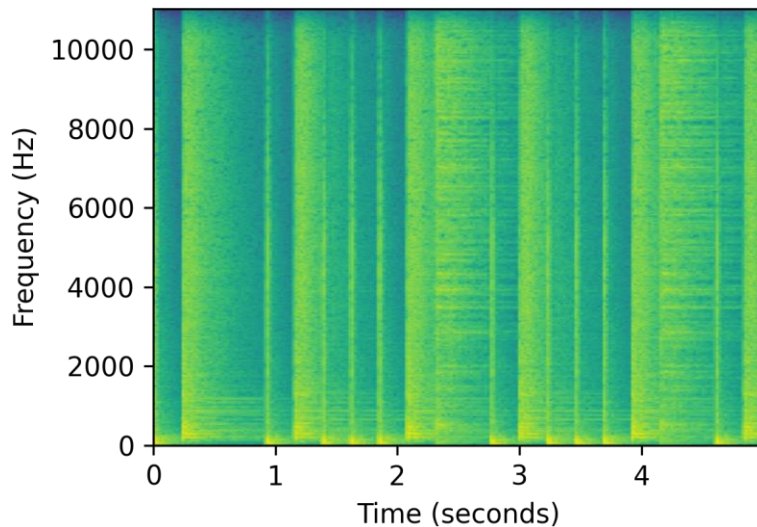
Audio 3

Audio Representations

Mel Spectrogram

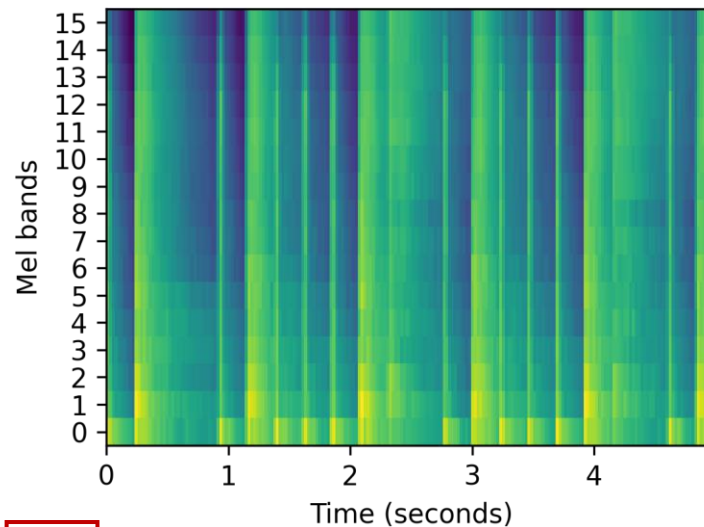
- More efficient representation (fewer frequency bands)
- Still captures perceptually relevant information

STFT



(513 frequency bands)

Mel Spectrogram



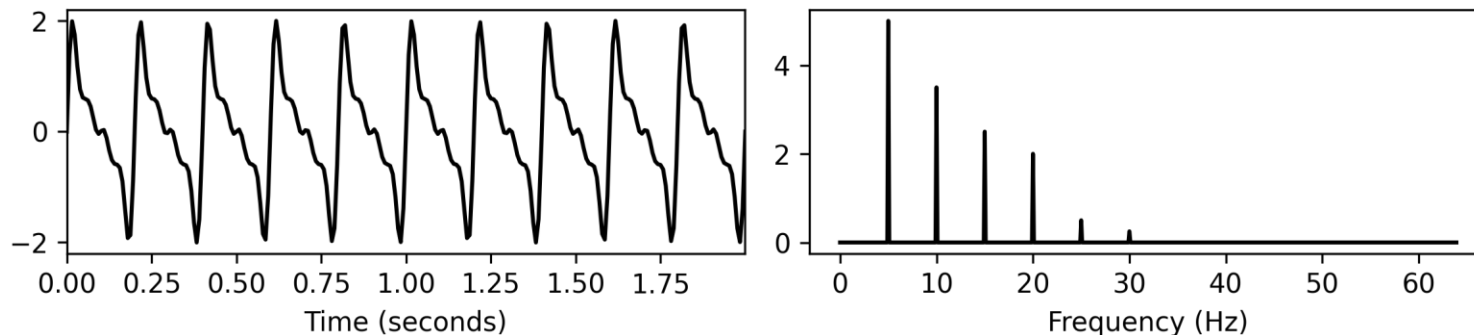
(16 mel bands)

Compression by 96.9 %

Audio Signal Decomposition

Periodic Signals

- Periodic signals:
 - Sum of pure tones (partials)
 - Fundamental frequency f_0
 - Harmonics f_k (approx. integer multiples of f_0):
 - $f_k \approx (k + 1) \cdot f_0$



Audio Signal Decomposition

Periodic Signals

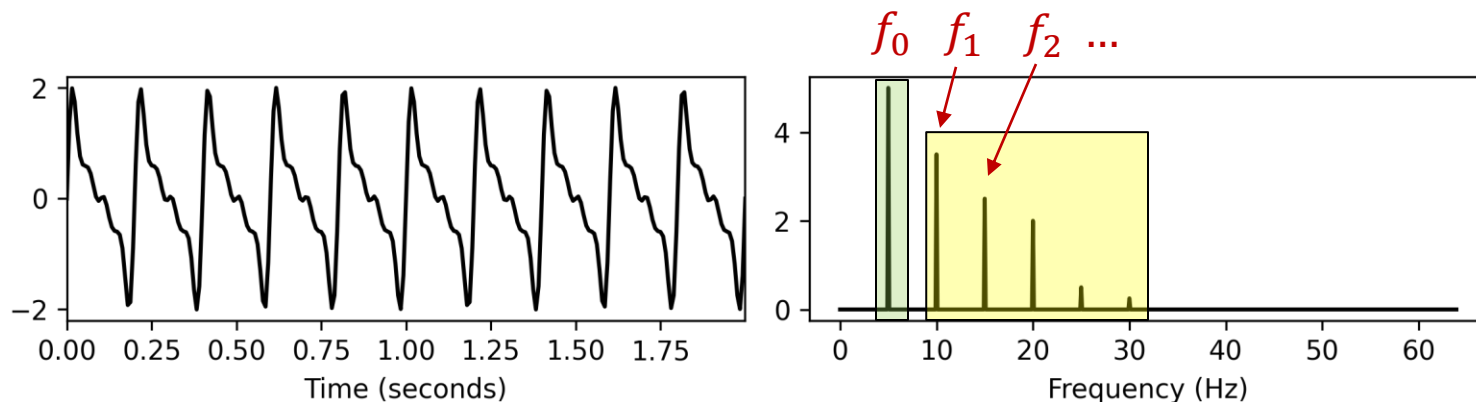
- Periodic signals:

- Sum of pure tones (partials)

- Fundamental frequency f_0

- Harmonics f_k (approx. integer multiples of f_0):

- $f_k \approx (k + 1) \cdot f_0$



Audio Signal Decomposition

Pitch

- Perceptual property (sort sounds from low to high pitch)
- Closely related to frequency

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

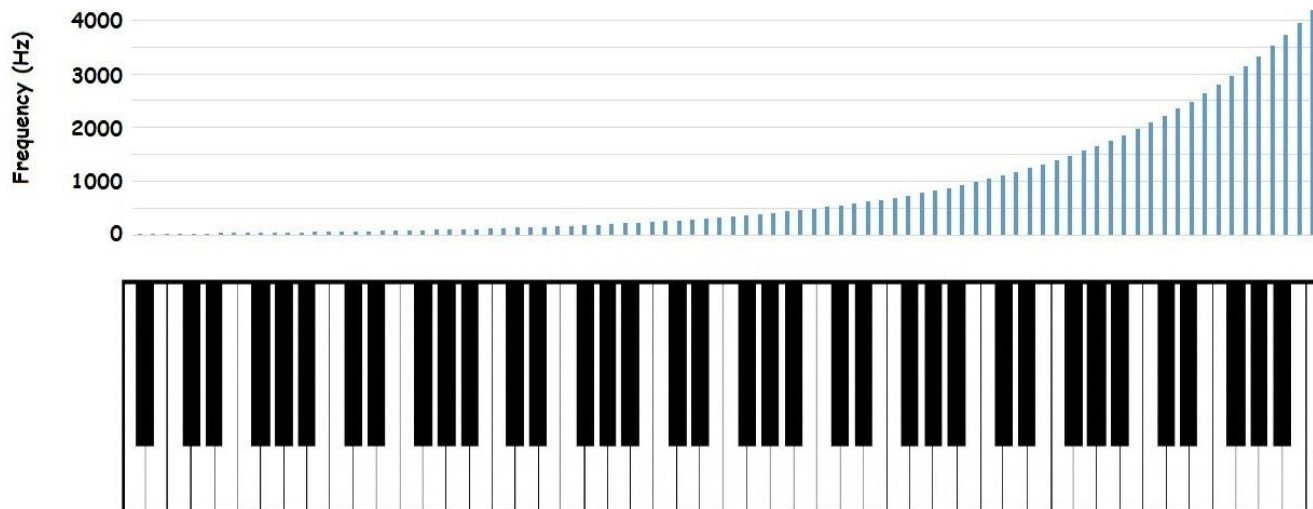


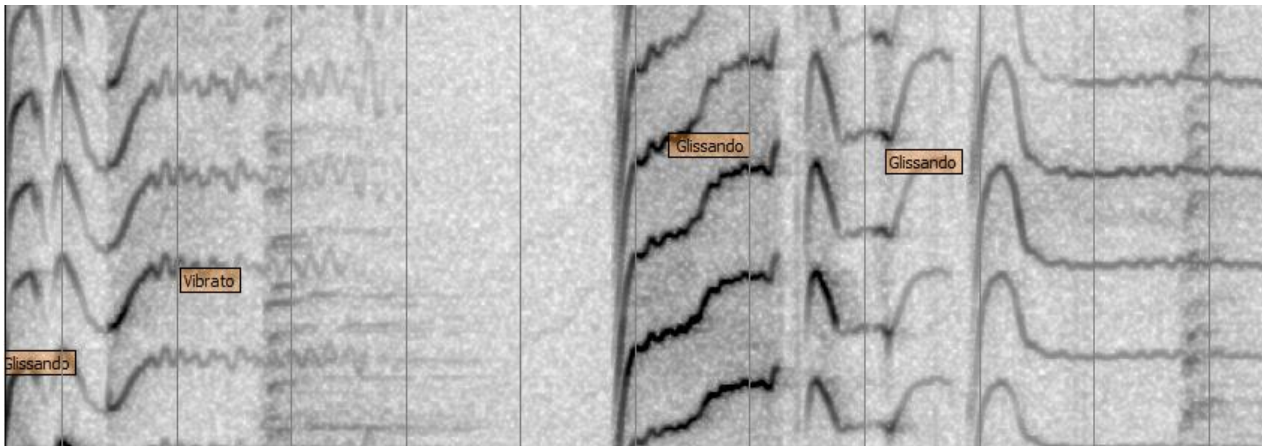
Fig. 2.5

Audio Signal Decomposition

Frequency Modulation

■ Techniques

- Glissando – continuous transition between note pitches
- Vibrato – periodic frequency modulation



Spectrogram example (frequency x time)

Fig. 2.6

Audio Signal Decomposition

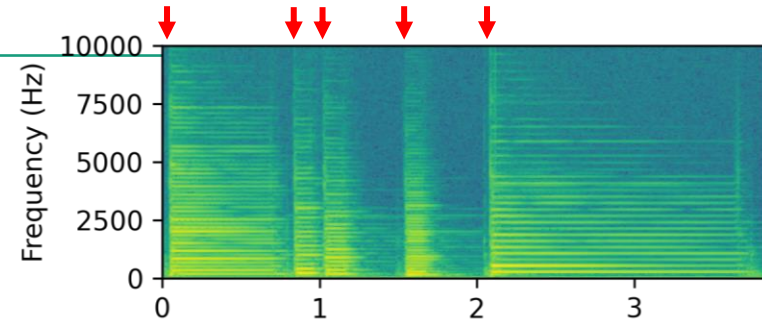
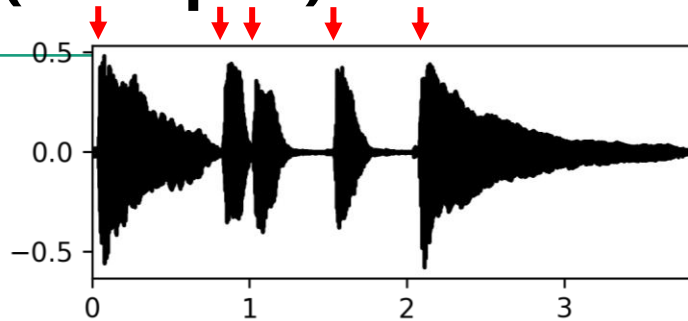
Transients

- Sound characteristics
 - High amplitude
 - Short duration
 - Wide-band signal
 - Energy distributed over large frequency range (not just a few frequencies)

Audio Signal Decomposition

Transients (Examples)

■ String instruments
🔊 [Audio 1](#)



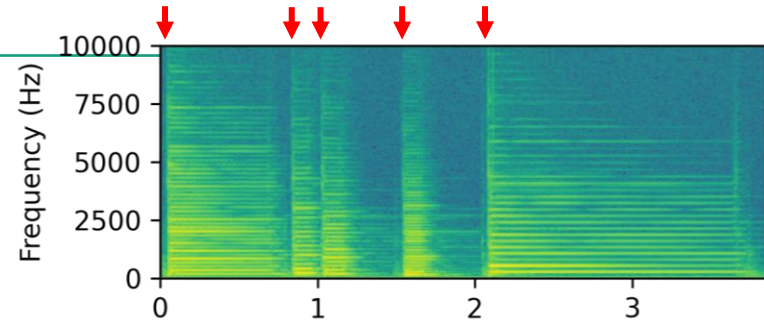
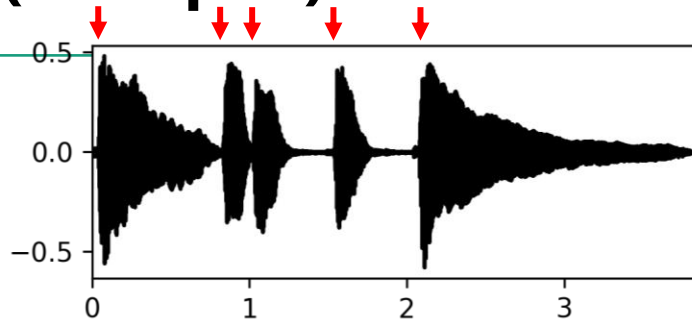
Audio Signal Decomposition

Transients (Examples)

- String instruments



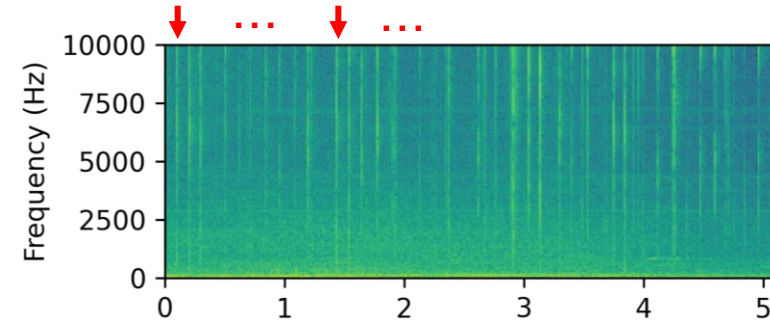
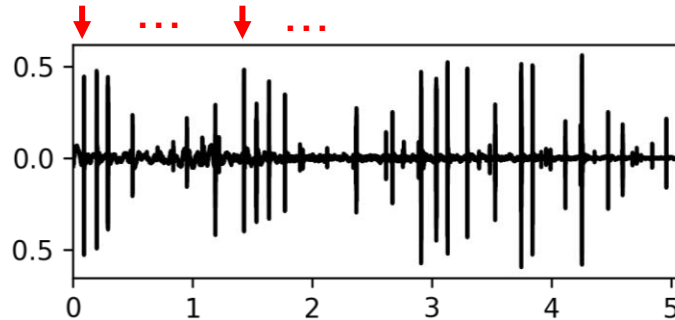
Audio 1



- Bat vocalizations



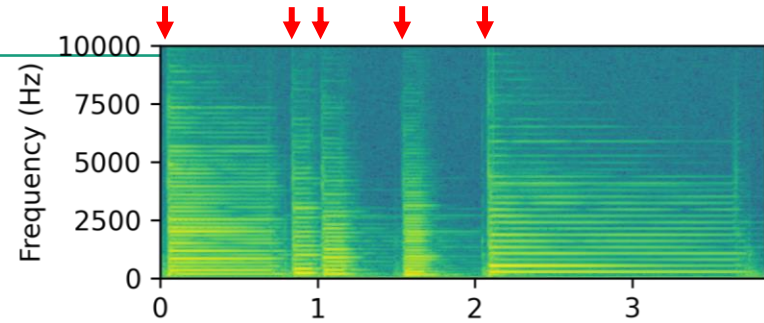
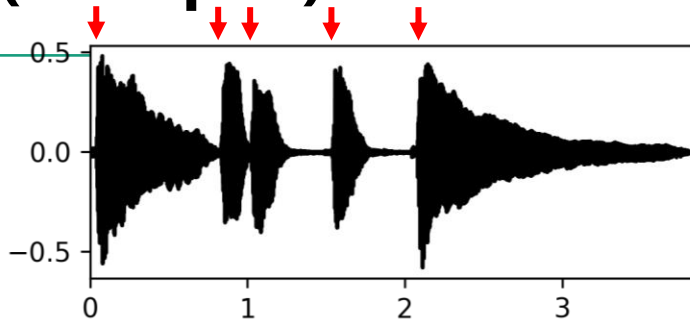
Audio 2



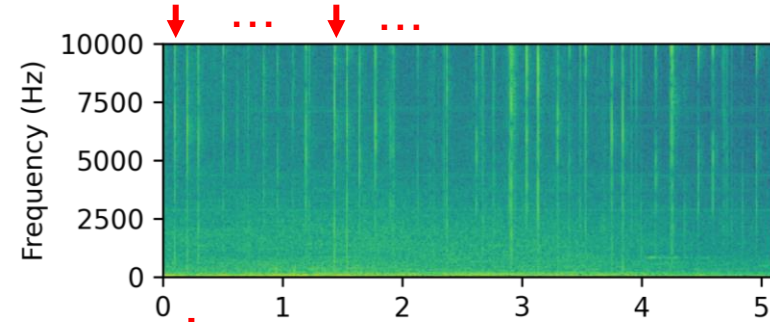
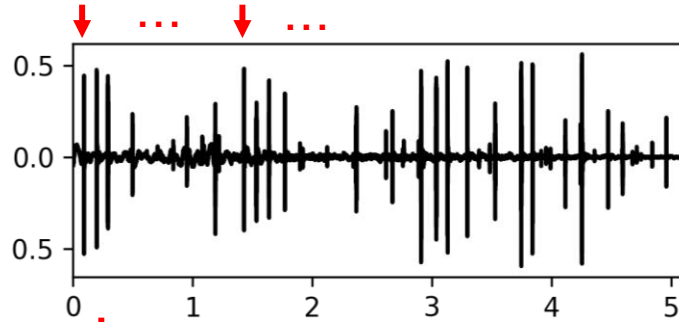
Audio Signal Decomposition

Transients (Examples)

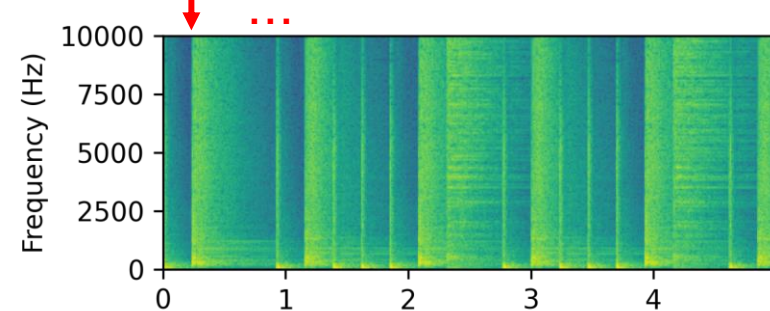
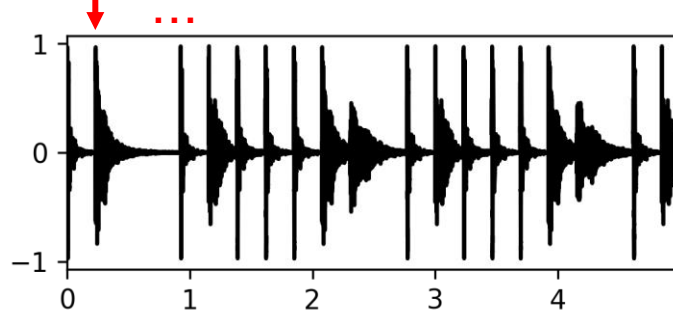
- String instruments



- Bat vocalizations



- Drum instruments



Time (seconds)

Time (seconds)


Audio Signal Decomposition

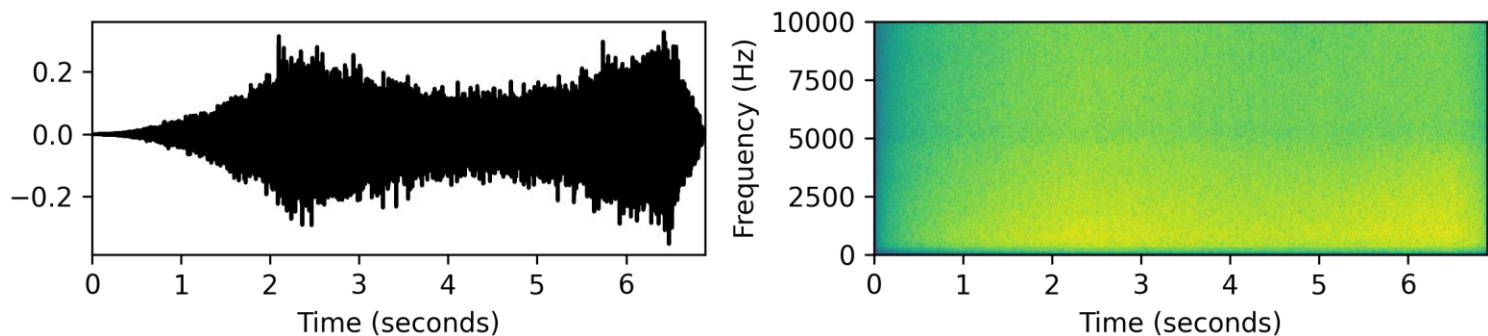
Noise

- Sound characteristics
 - Non-periodic, texture-like
 - Random fluctuations of air pressure

Audio Signal Decomposition

Noise

- Sound characteristics
 - Non-periodic, texture-like
 - Random fluctuations of air pressure
- Examples
 - Consonants (speech)
 - Wind (random aerodynamic turbulences)
 - Waves (ocean)  [Audio 4](#)



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

Timbre

- Timbre

- Timbre distinguishes musical sounds that have the same pitch (fundamental frequency) and loudness

Audio Features

Timbre

■ Timbre

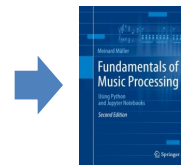
- Timbre distinguishes musical sounds that have the same pitch (fundamental frequency) and loudness
- Affected by different acoustic phenomena such as
 - Spectral structure / envelope of overtones
 - Noise-like components

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

Audio Features

Timbre

- Timbre
 - When looking at musical instruments, we need to consider
 - Instrument's construction
 - Sound production principles
 - Membranophones, chordophones, aerophones, electrophones

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

Audio Features

Temporal Envelope

- Smooth curve outlining the signal extreme points
- ADSR envelope model (also used for audio synthesis)
 - Attack, Decay, Sustain, Release

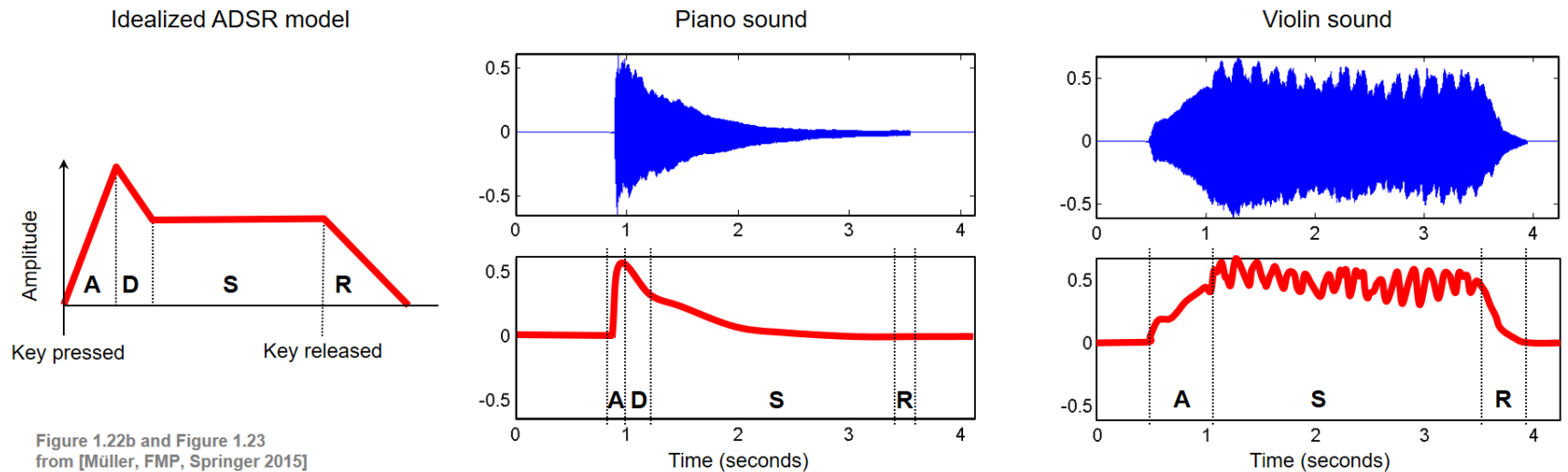
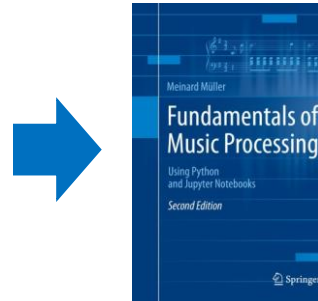


Fig. 2.7

Audio Features

Temporal Envelope

- Tremolo
 - Periodic amplitude modulation
 - Often coincides with frequency modulation (vibrato)
 - Examples: instrument sounds



FMP Notebooks

Fig. 2.7

Audio Features

Categorization

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

Audio Features

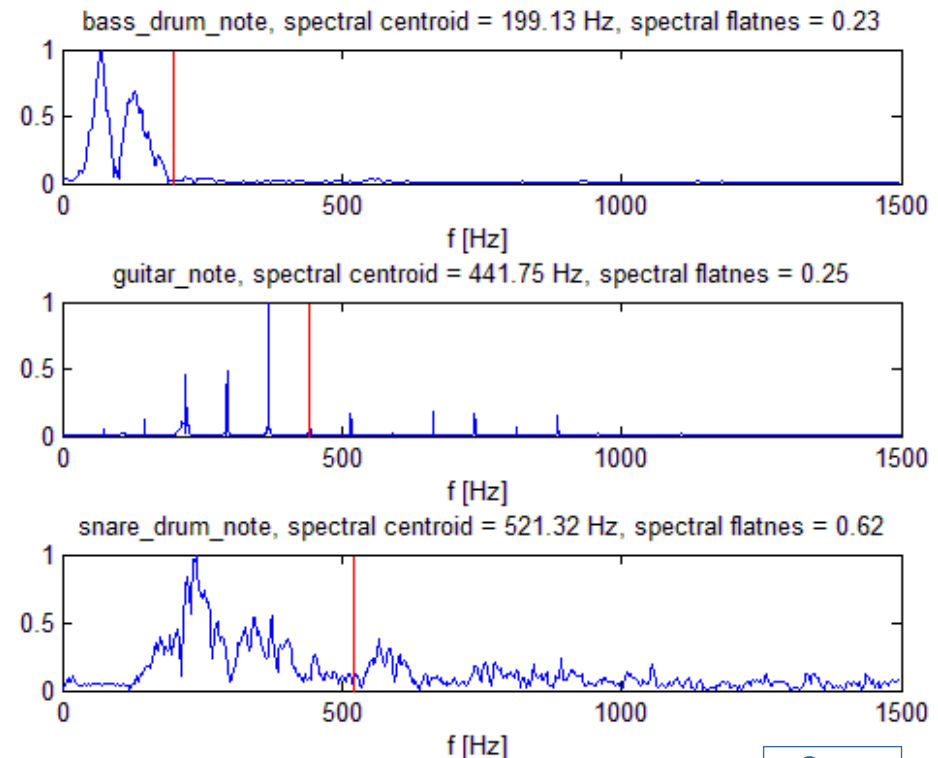
Categorization

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

Audio Features

Timbre Low-level Audio Features

- Spectral Centroid (SC):
 - Center of mass in the magnitude spectrogram
 - Low-pitched vs. high-pitched sounds



Own

Audio Features

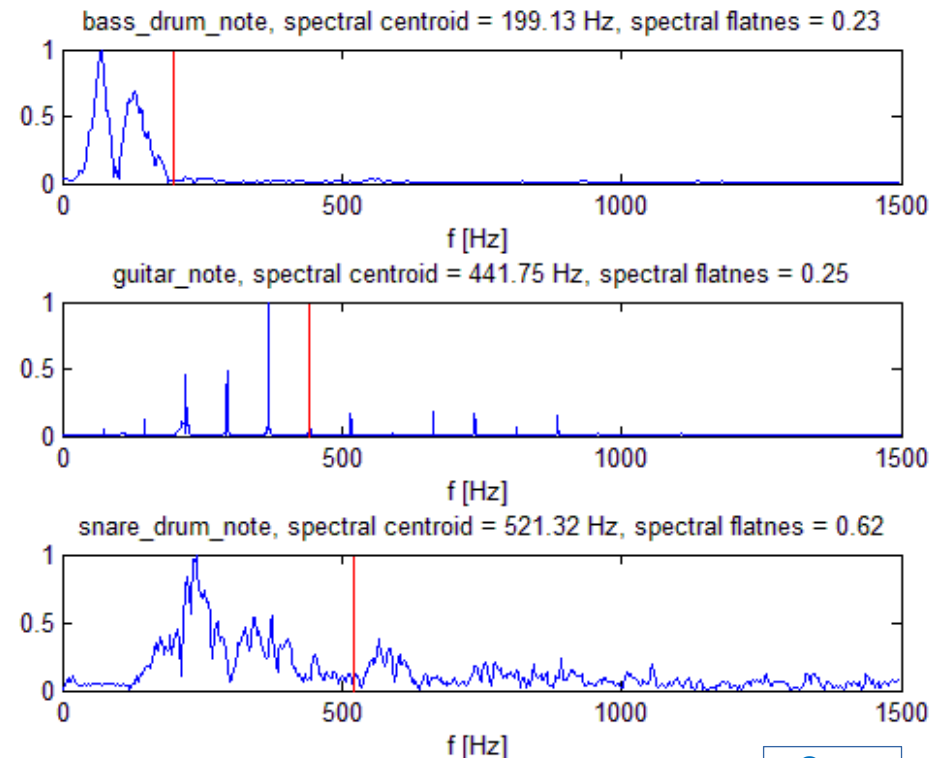
Timbre Low-level Audio Features

■ 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

Mel-Frequency Cepstral Coefficients (MFCC)

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

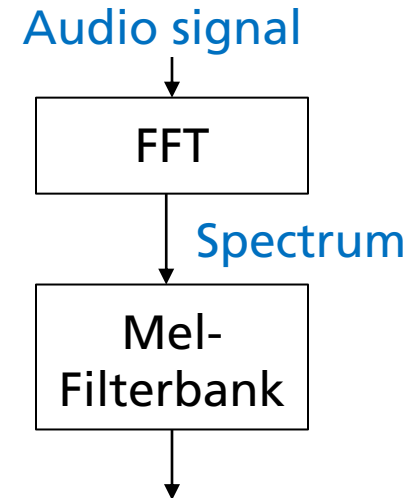
Audio signal

Own

Audio Features

Mel-Frequency Cepstral Coefficients (MFCC)

- Convolutional **excitation** * **filter** model
 - Excitation: vibration of vocal folds
 - Filter: resonance of the vocal tract
- FFT magnitude spectrum
 - Multiplicative **excitation** · **filter** model

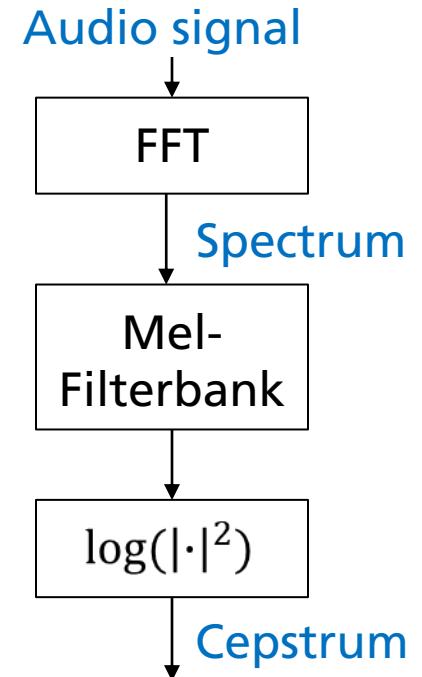


Own

Audio Features

Mel-Frequency Cepstral Coefficients (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

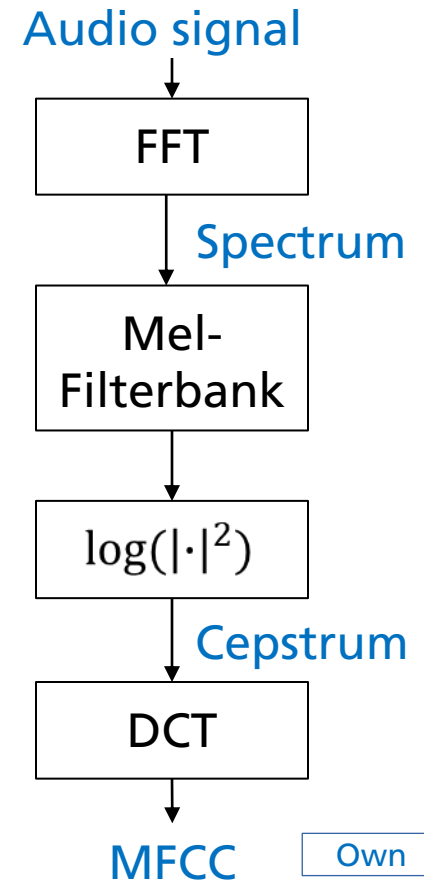


Own

Audio Features

Mel-Frequency Cepstral Coefficients (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

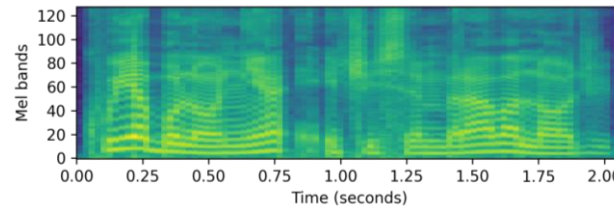
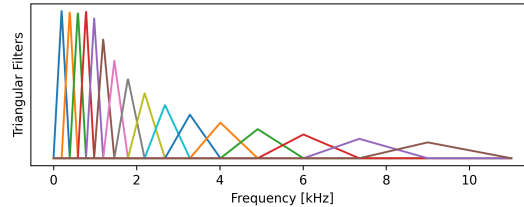
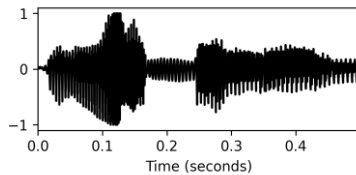


Audio Features

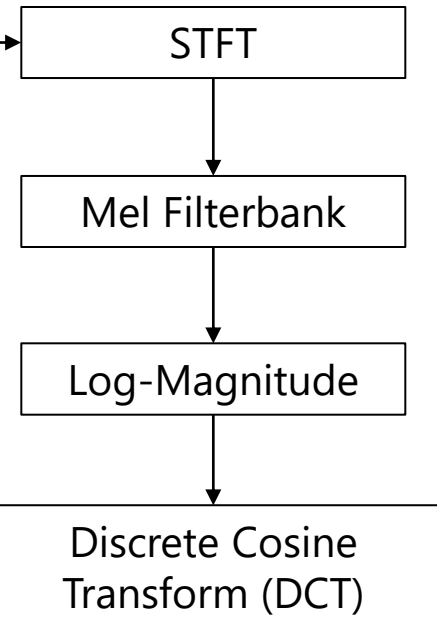
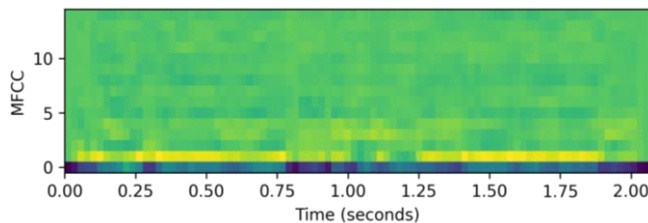
Mel-Frequency Cepstral Coefficients (MFCC)

- Compact representation of spectral envelope

Audio signal



MFCC



Audio Processing

Chroma Features

- Human pitch perception is periodic
- 2 pitches one octave apart are perceived as similar

Audio Processing

Chroma Features

- Human pitch perception is periodic
- 2 pitches one octave apart are perceived as similar
- Pitch = chroma + tone height
 - Chroma: C, C#, D, D#, ..., B (12)
 - Tone height: Octave number

Figure 3.3a from [Müller, FMP, Springer 2015]

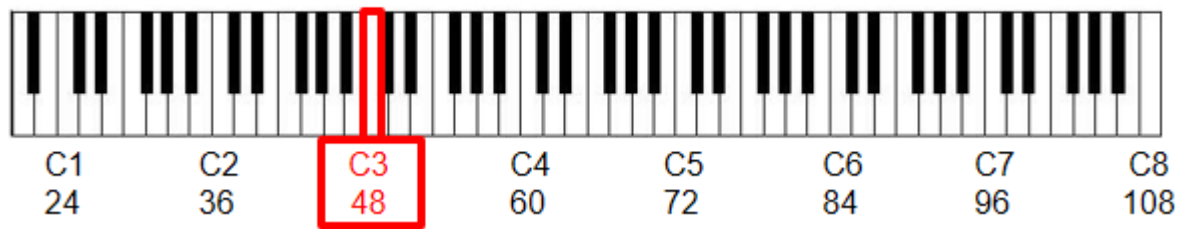


Fig. 2.8

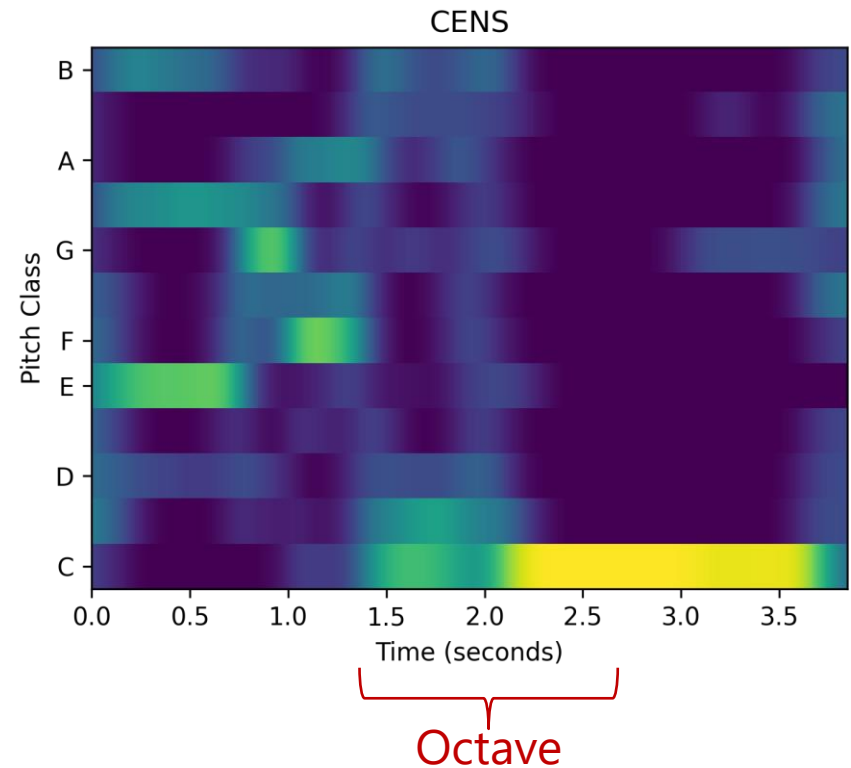
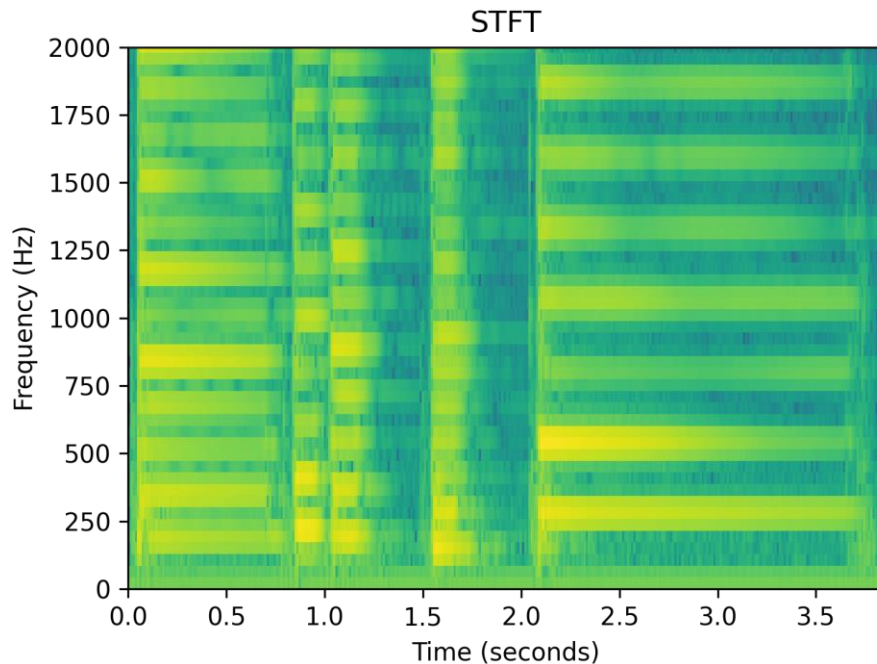
Audio Processing

Chroma Features

■ Example



Audio 1



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. 2.8: [Müller, 2015]: Fundamentals of Music Processing (FMP), Springer, 2015, Fig. 3.3a

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. 9.5: https://www.mathworks.com/help/dsp/ref/stft_output.png

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

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

Fig. 13: <https://newt.phys.unsw.edu.au/jw/graphics/notes.GIF>

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>

[Audio 1] <https://freesound.org/people/xserra/sounds/196765/>

[Audio 2] <https://freesound.org/people/IliasFlou/sounds/498058/> (~0:00 – 0:05)

[Audio 3] <https://freesound.org/people/danlucaz/sounds/517860/> (~0:00 – 0:05)

[Audio 4] <https://freesound.org/people/IENBA/sounds/489398/> (~0:00 – 0:07)

Thank you!

■ Any questions?

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

<https://www.machinelistinging.de>