

AI-based Audio Analysis of Music and Soundscapes

Audio Processing

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

jakob.abesser@idmt.fraunhofer.de

Seminar Structure

- Sounds – Generation & Recording & Processing
- Time-Frequency Representations
 - Short-time Fourier Transform (STFT)
 - Mel-Spectrogram
 - Constant-Q Transform
- Audio Features
 - Low-level Features
 - Mid-level Features
 - Mel-Cepstral Coefficients (MFCCs)
 - Chroma Features

Audio Processing

Sound

- Sound [[Merriam-Webster](#)]
 - "Sensation perceived by the sense of hearing"
 - "A particular auditory impression"
 - "Mechanical radiant energy that is transmitted by longitudinal pressure waves in a material medium (such as air) and is the objective cause of hearing"

Audio Processing

Sound

- Mechanical vibration with contact to air
- Rapid modulation of airflow

Audio Processing

Sound Waves

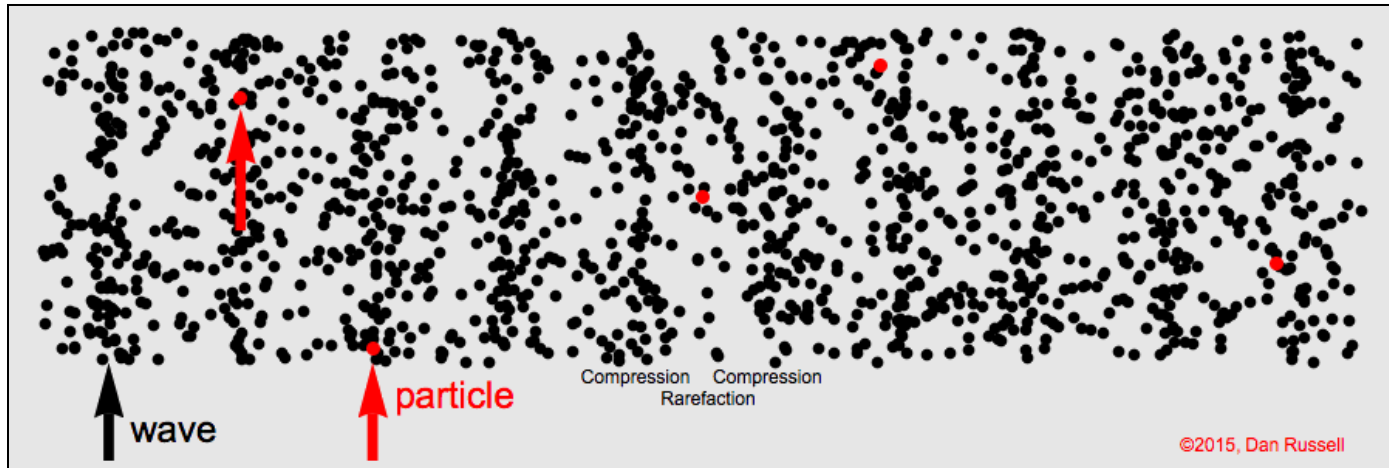


Fig. 1

Audio Processing

Sound Waves

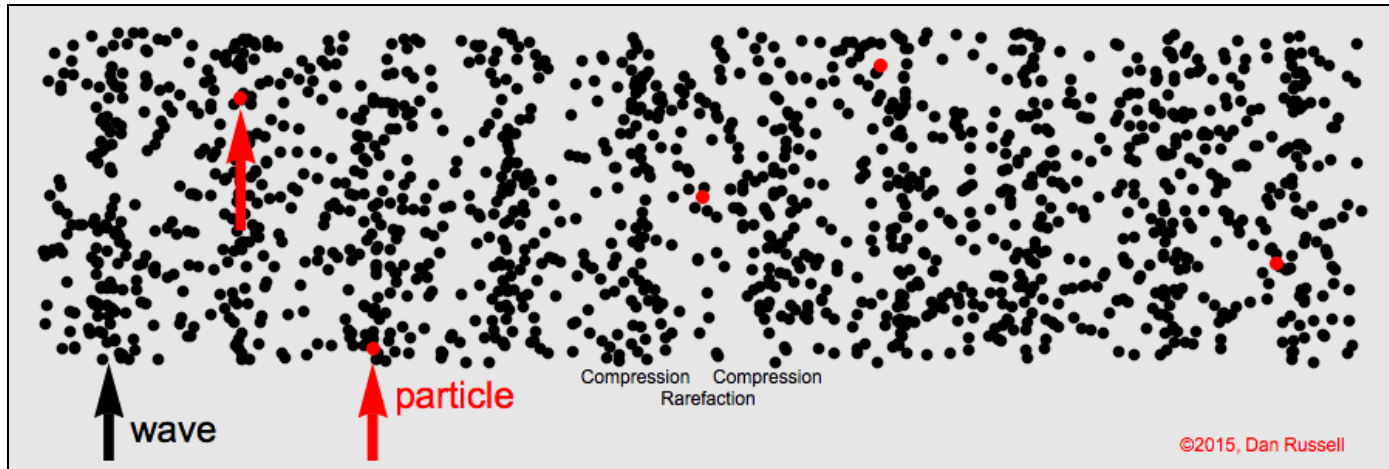


Fig. 1

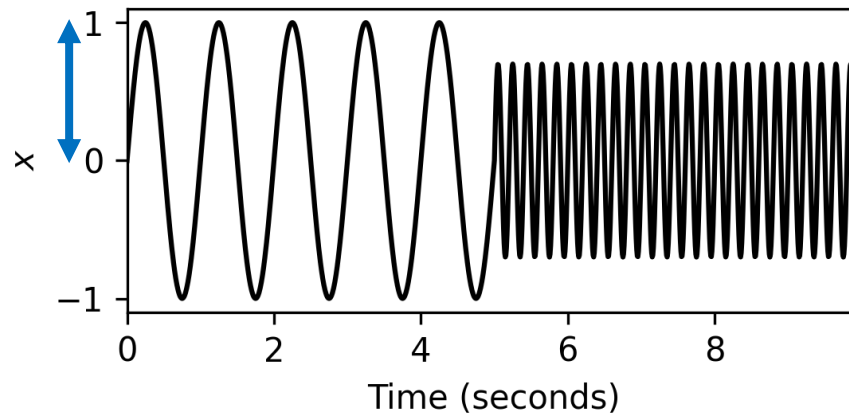
- Oscillation of air pressure
- Propagate through medium (air)
- Converted into physical motion by ear / microphone

Audio Processing

Waveform

- Waveform $x(t)$

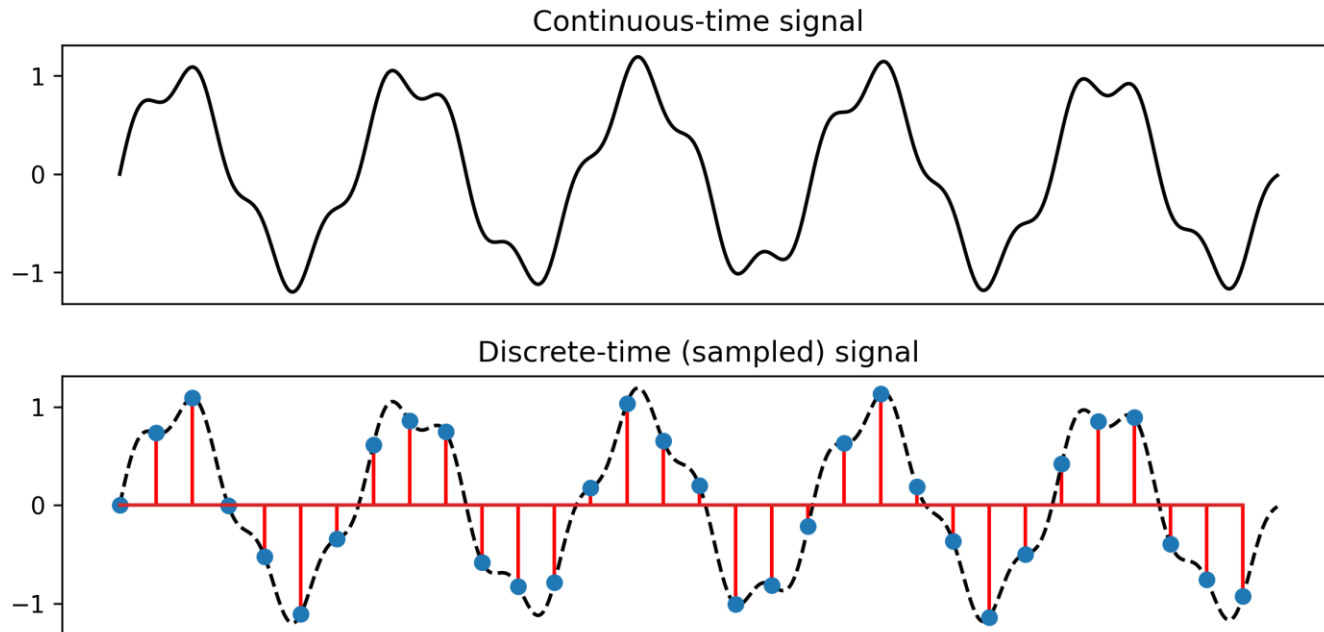
- Amplitude (vertical displacement) of pressure vs. time



Audio Processing

Sampling

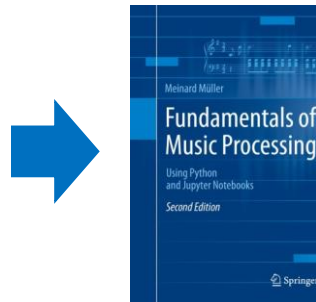
- Converting (continuous-time) analog signals into (discrete-time) digital signal
- Sampling frequency f_s : Number of samples per seconds [Hz]



Audio Processing

Sampling

- (Nyquist-Shannon) sampling theorem: $f_{\max} < f_s/2$
 - Transformation between continuous-time and discrete-time signals
 - Signal must be band-limited
 - If sampling rate is too slow, aliasing occurs (higher frequencies can not be reconstructed properly)



FMP Notebooks

Audio Processing

Sound Level

- Sound level [dB]

- $L_{\text{dB}} = 20 \log_{10} x_{\text{RMS}}$

- Root mean square $x_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_i x_i^2}$

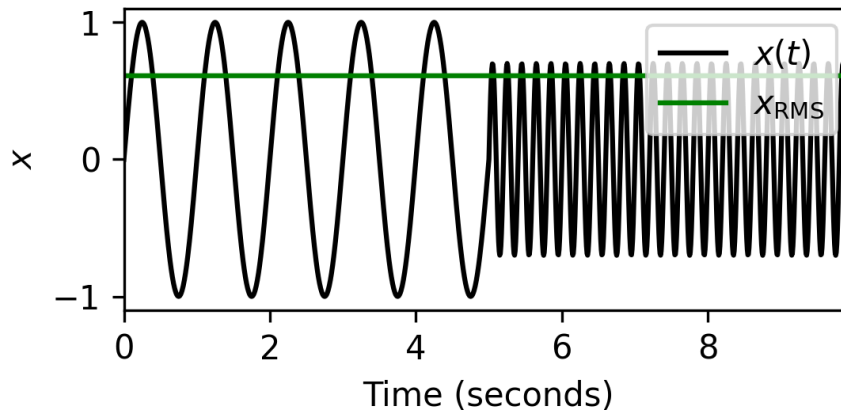
Audio Processing

Sound Level

- Sound level [dB]

- $L_{\text{dB}} = 20 \log_{10} x_{\text{RMS}}$

- Root mean square $x_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_i x_i^2}$



Audio Processing

Sound Intensity

- Dynamics
 - Volume of a sound
- Sound power
 - Energy per time emitted by sound source (in all directions)

Audio Processing

Sound Intensity

- Dynamics
 - Volume of a sound
- Sound power
 - Energy per time emitted by sound source (in all directions)
- Sound intensity
 - Sound power per unit area
 - Minimum perceivable sound intensity = threshold of hearing
 - $I_{\text{TOH}} = 10^{-12} \text{W/m}^2$
 - Intensity is computed using reference (TOH)
 - $I[\text{dB}] = 10 \cdot \log_{10} \left(\frac{I}{I_{\text{TOH}}} \right)$

Audio Processing

Sound Intensity

■ Examples

Source	Intensity	Intensity level	× TOH
Threshold of hearing (TOH)	10^{-12}	0 dB	1
Whisper	10^{-10}	20 dB	10^2
Pianissimo	10^{-8}	40 dB	10^4
Normal conversation	10^{-6}	60 dB	10^6
Fortissimo	10^{-2}	100 dB	10^{10}
Threshold of pain	10	130 dB	10^{13}
Jet take-off	10^2	140 dB	10^{14}
Instant perforation of eardrum	10^4	160 dB	10^{16}

Table 1.1 from [Müller, FMP, Springer 2015]

Fig. 1.1

Audio Processing

Loudness

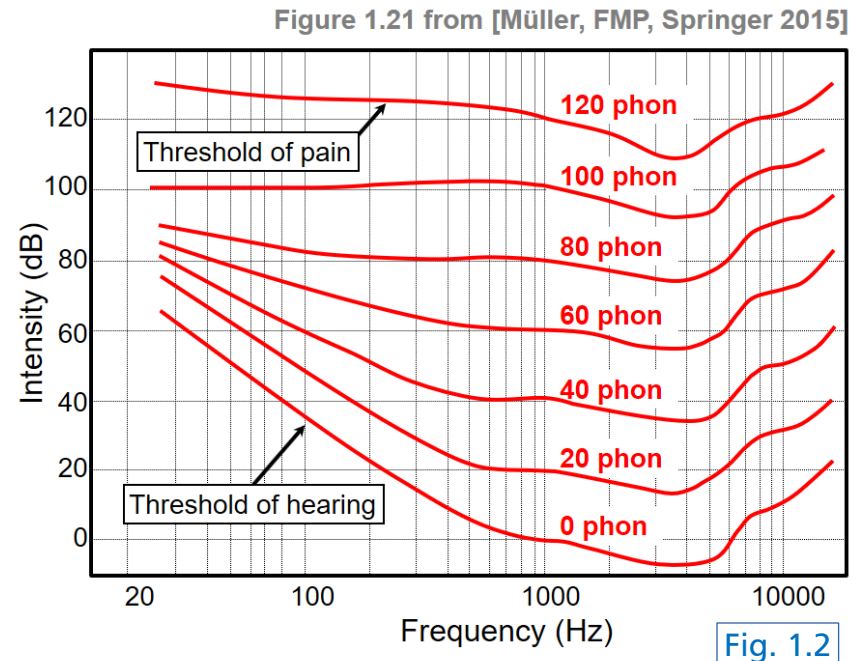
- Perceptual property (sort sounds from quiet to loud)
- Correlates with sound intensity
- Subjective, further depends on sound duration & frequency

Audio Processing

Loudness

- Perceptual property (sort sounds from quiet to loud)
- Correlates with sound intensity
- Subjective, further depends on sound duration & frequency

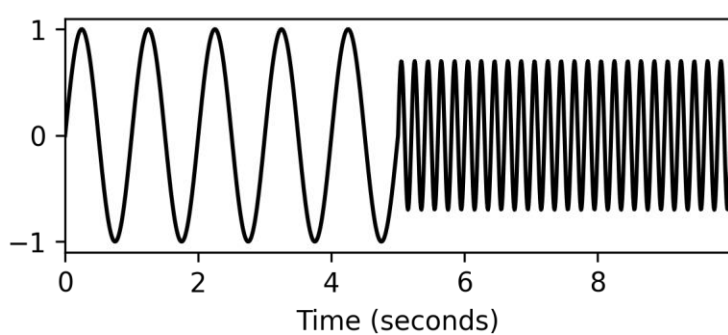
- Equal loudness curves
 - Perceived loudness [phon]



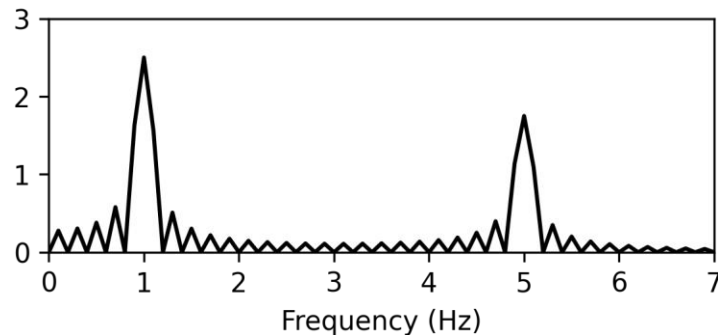
Audio Processing

Time-Frequency Decomposition

- Fourier Transform
 - Decompose signal into sum of sinusoids
 - Amplitude, frequency, phase



Waveform

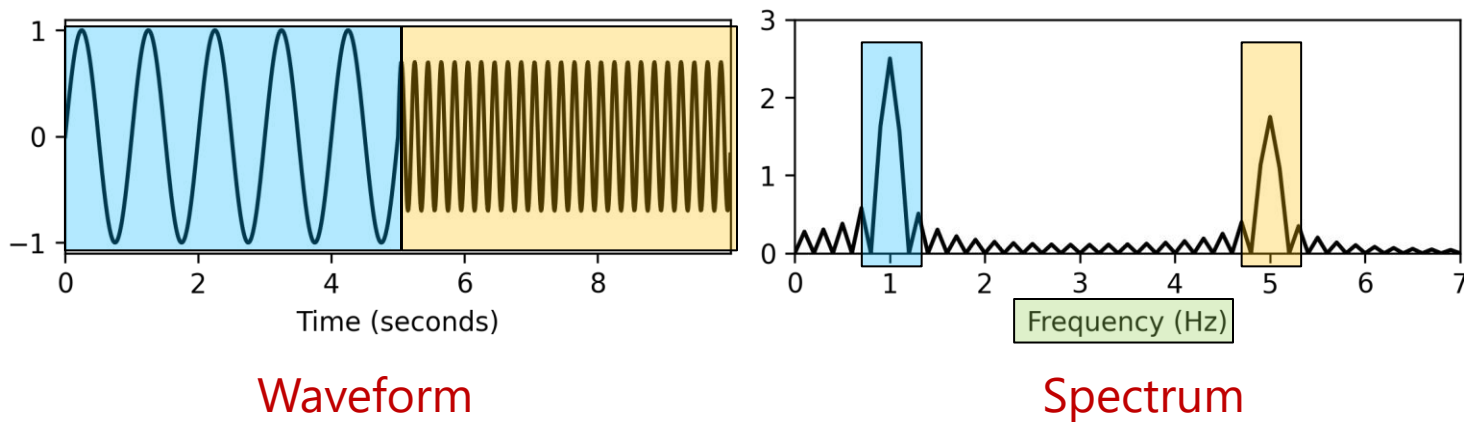


Spectrum

Audio Processing

Time-Frequency Decomposition

- Fourier Transform
 - Decompose signal into sum of sinusoids
 - Amplitude, frequency, phase



Audio Processing

Time-Frequency Decomposition

- Short-Time Fourier Transform (STFT)
 - Moving analysis window

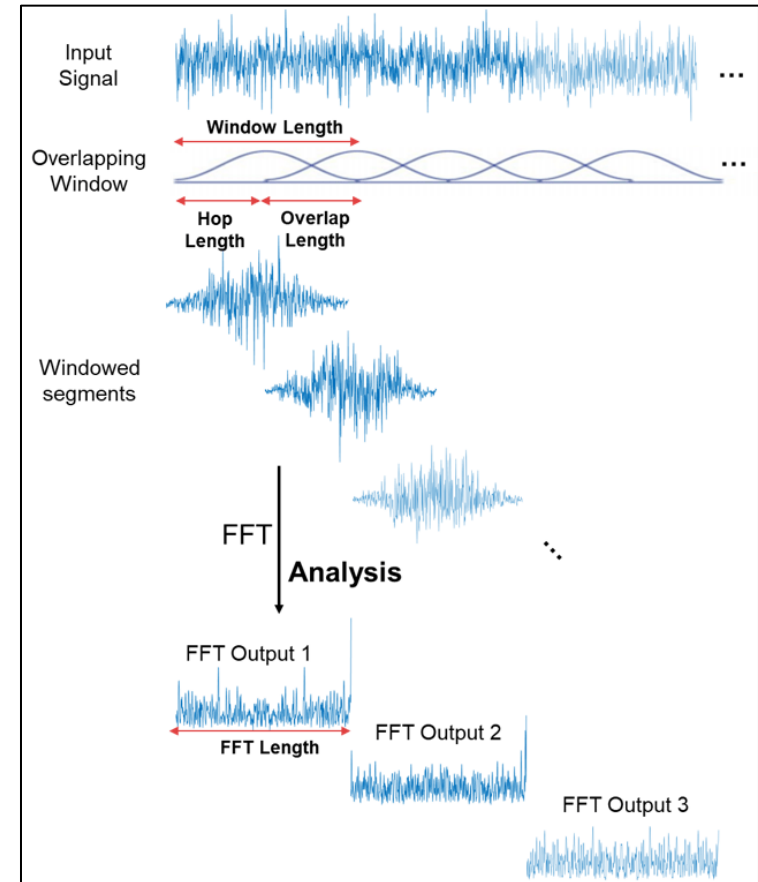
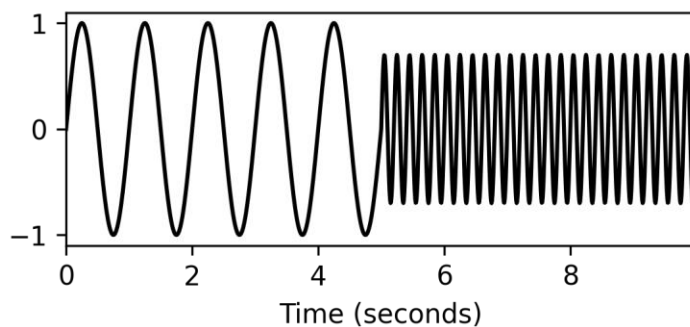


Fig. 2

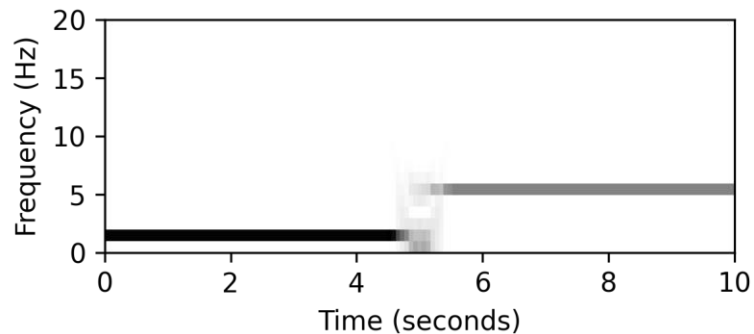
Audio Processing

Time-Frequency Decomposition

- Short-Time Fourier Transform (STFT)
 - Moving analysis window
 - Time-frequency energy distribution in audio signal



Waveform

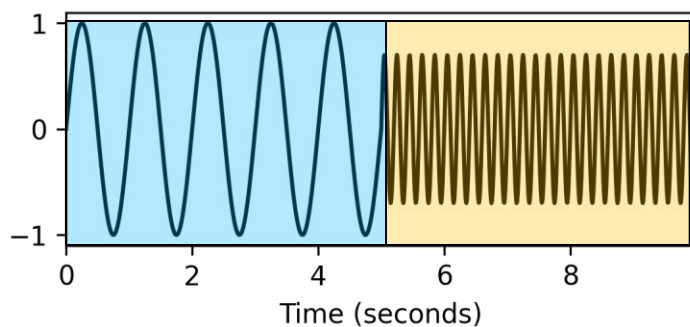


Spectrogram

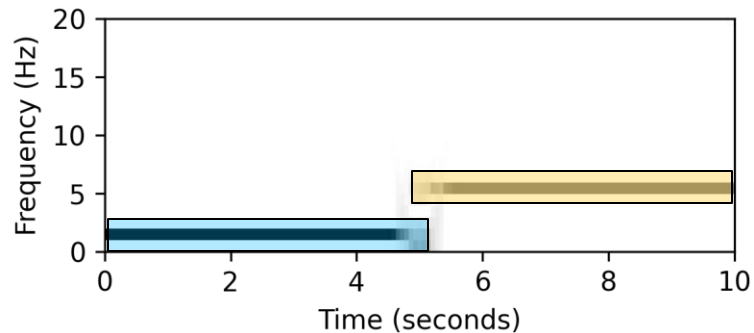
Audio Processing

Time-Frequency Decomposition

- Short-Time Fourier Transform (STFT)
 - Moving analysis window
 - Time-frequency energy distribution in audio signal



Waveform



Spectrogram

Audio Processing

Programming Session #1



Fig. 2.1

Audio Processing

Mel Frequency Scale

- Logarithmic frequency mapping (human pitch perception)

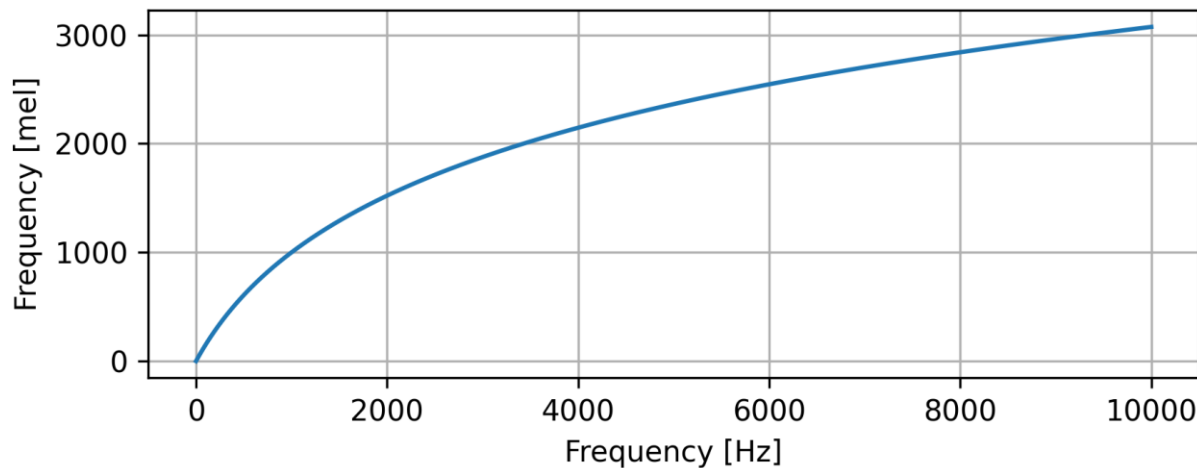
- $f[\text{mel}] = 2595 \cdot \log_{10} \left(1 + \frac{f[\text{Hz}]}{700} \right)$

Audio Processing

Mel Frequency Scale

- Logarithmic frequency mapping (human pitch perception)

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



Audio Processing

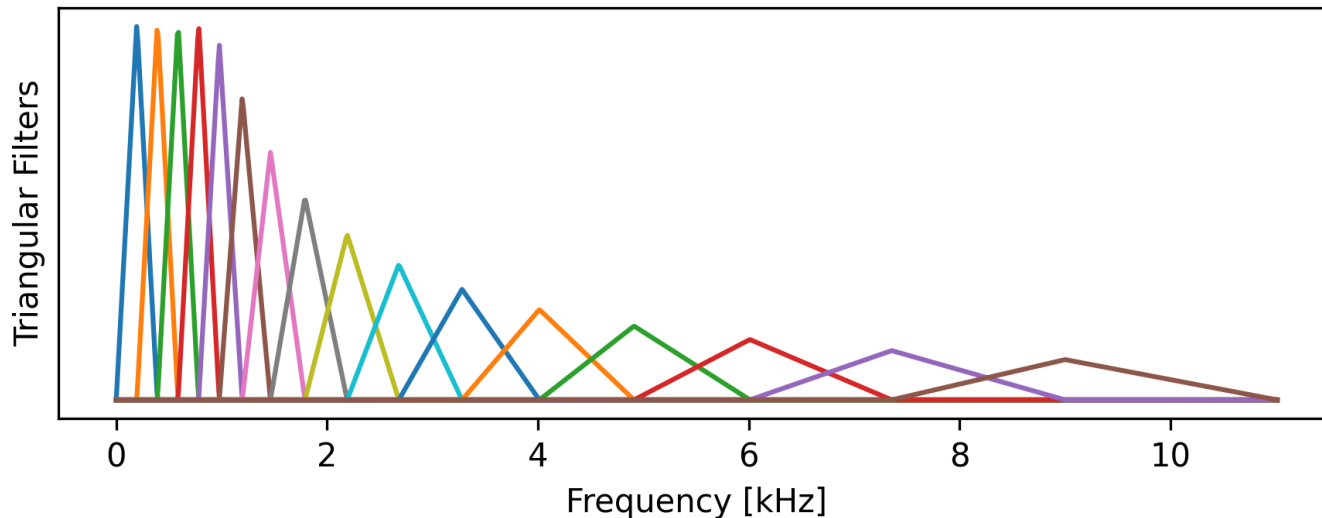
Mel Spectrogram

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

Audio Processing

Mel Spectrogram

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



Audio Processing

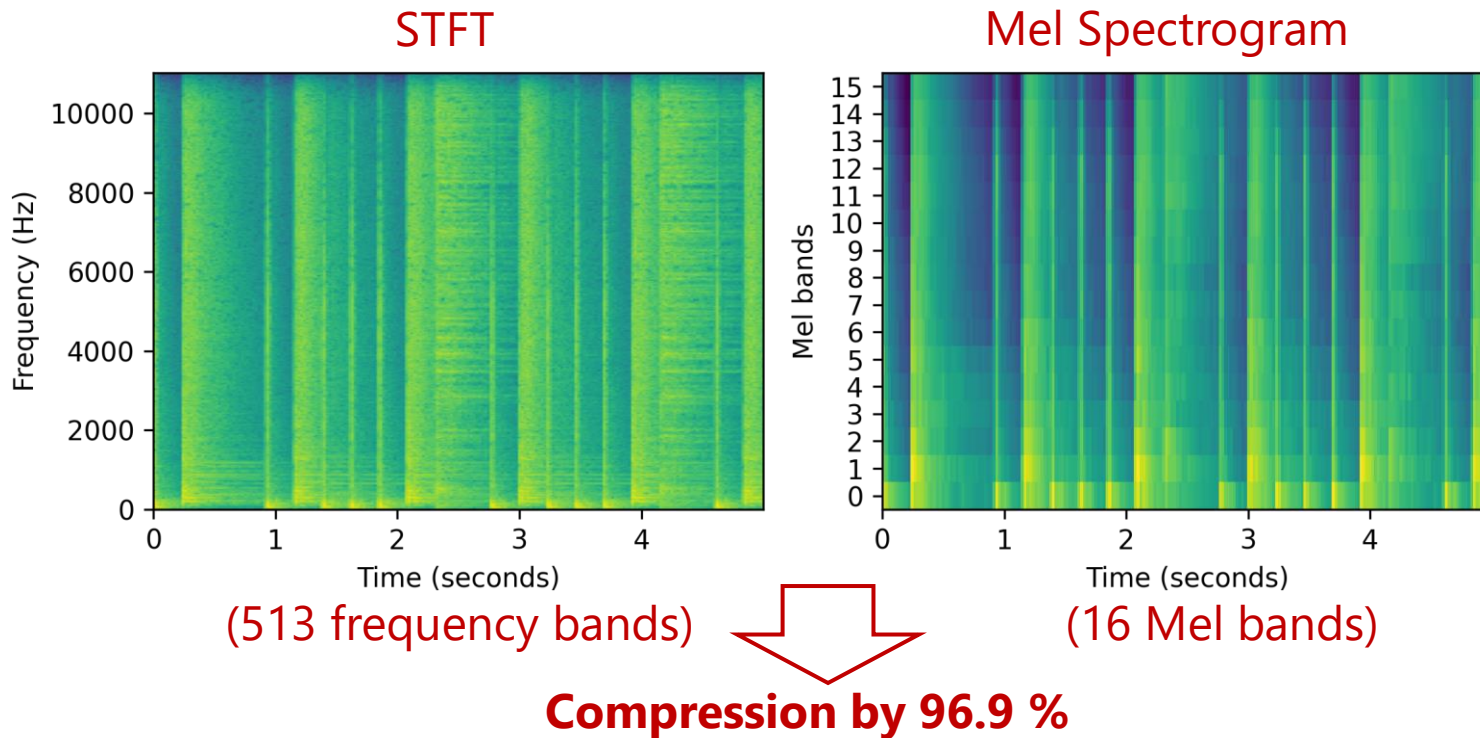
Mel Spectrogram

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

Audio Processing

Mel Spectrogram

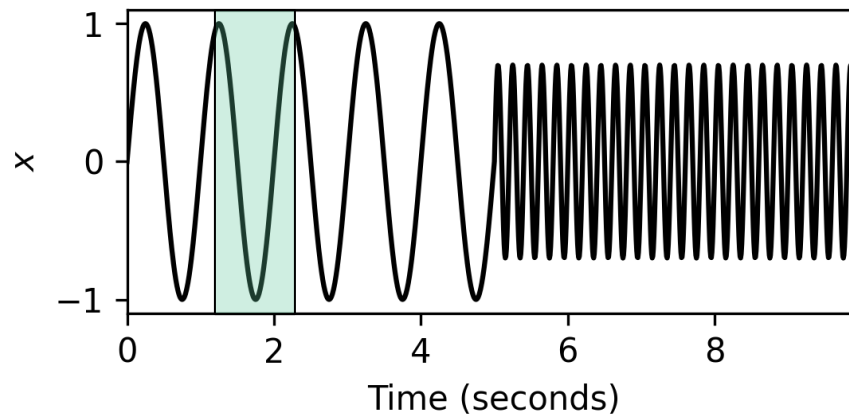
- More efficient representation (fewer frequency bands)
- Still captures perceptually relevant information



Audio Processing

Periodic Signals

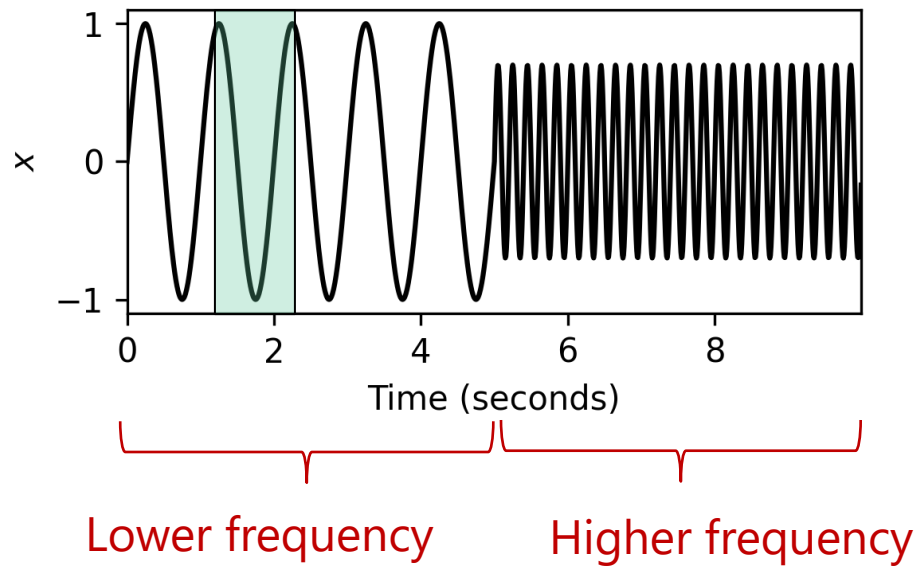
- Period T [s] – Duration of an elementary waveform



Audio Processing

Periodic Signals

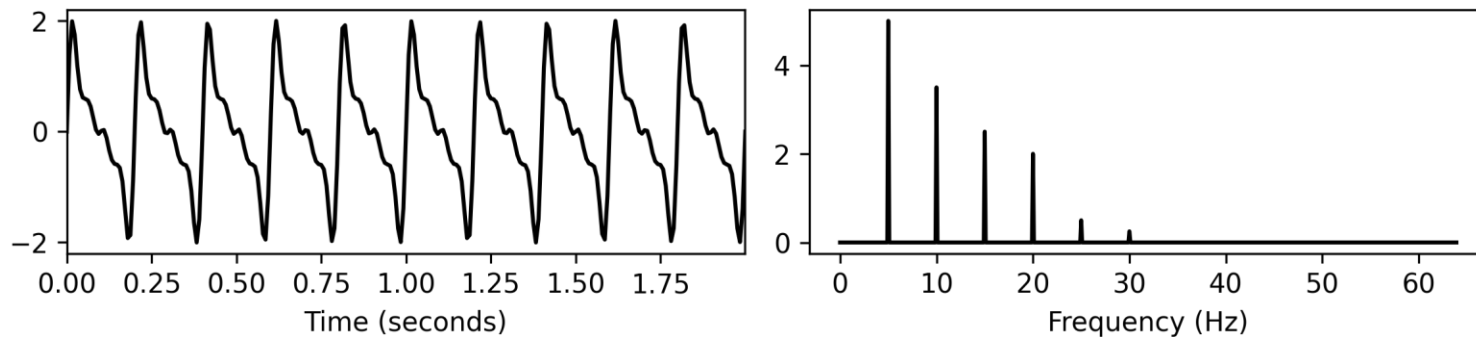
- Period T [s] – Duration of an elementary waveform
- Frequency f – Inverse of period ($f = \frac{1}{T}$ [Hz])



Audio Processing

Periodic Signals

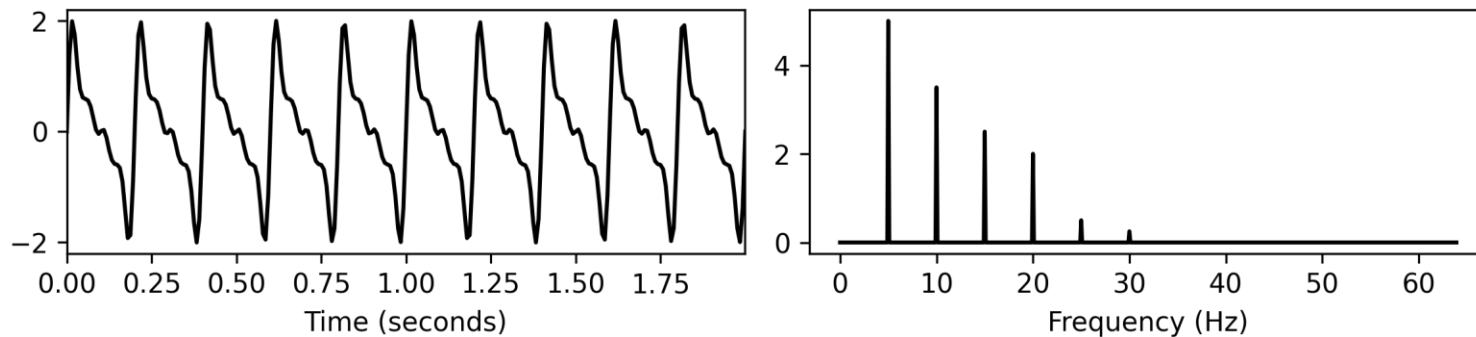
- Periodic signals:
 - Sum of pure tones (partials)
 - Fundamental frequency f_0



Audio Processing

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 Processing

Periodic Signals

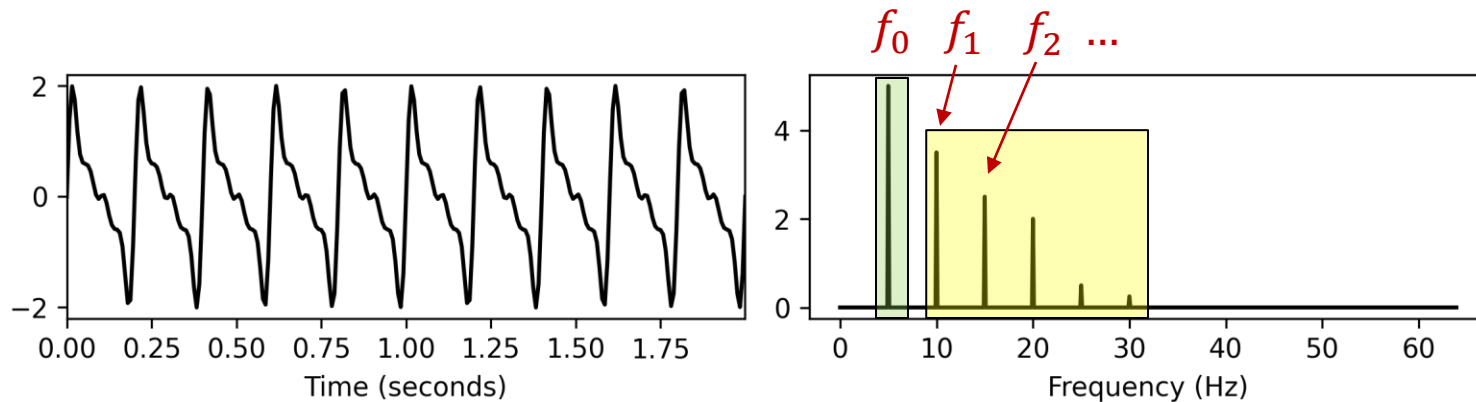
- Periodic signals:

- 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 Processing

Pitch

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

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

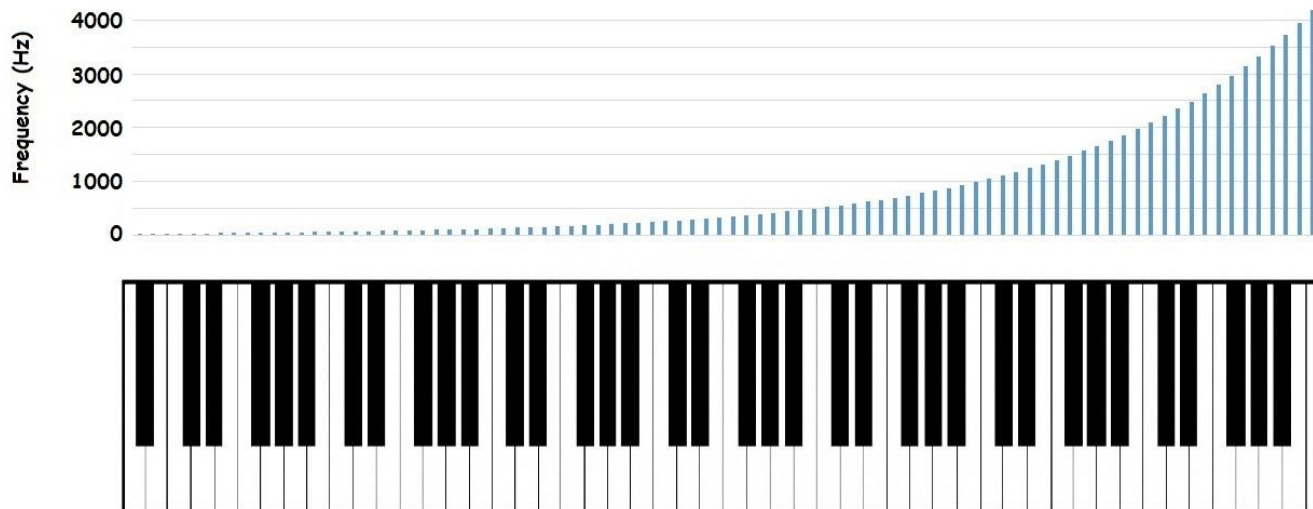


Fig. 2.5

Audio Processing

Pitch Distance (Intervals)

- Depend on ratio between pitch frequencies
- Examples

Note	Pitch p	Frequency f
A3	57	220 Hz
A4	69	440 Hz
A5	81	880 Hz

Octave intervals

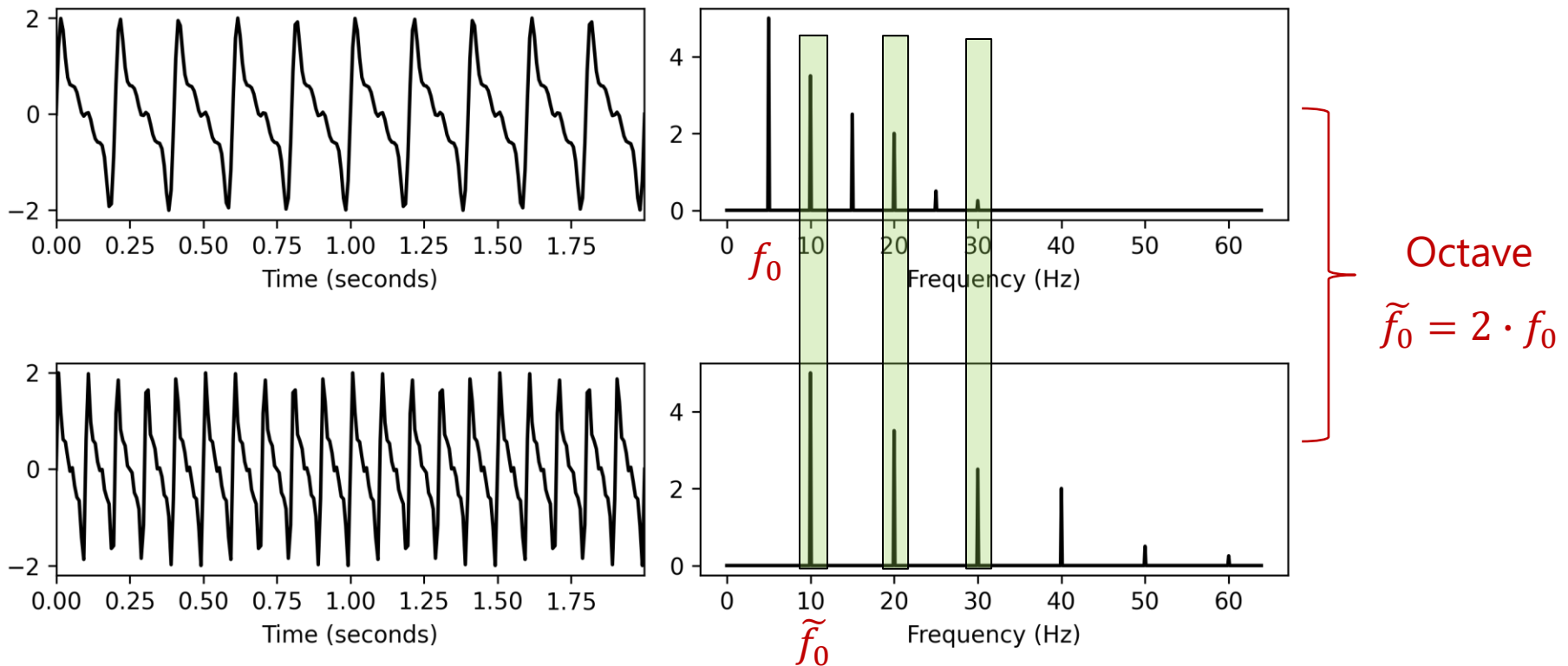
$f(A4) = 2 \cdot f(A3)$

$f(A5) = 2 \cdot f(A4)$

Audio Processing

Pitch Distance (Intervals)

- Note: Consonant intervals share partial frequencies

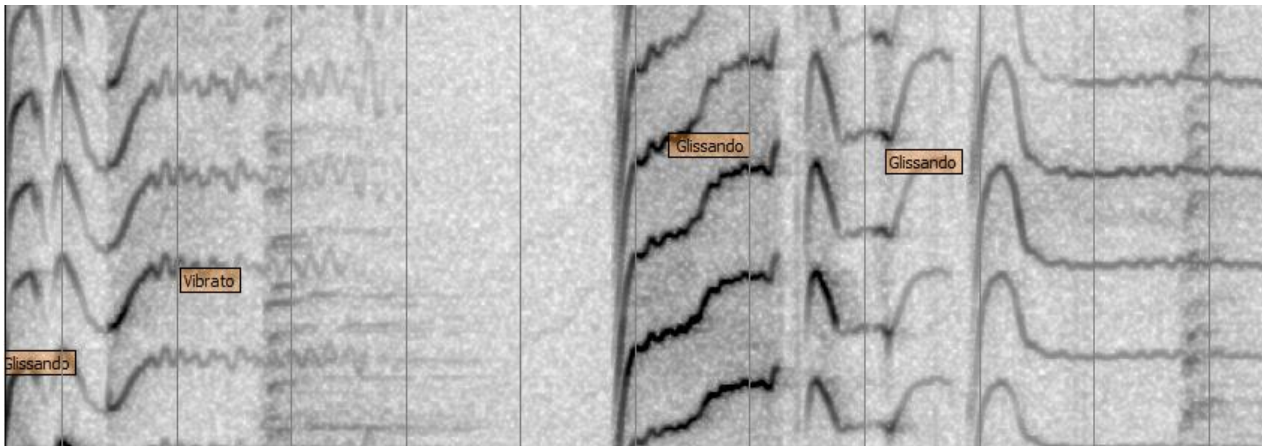


Audio Processing

Frequency Modulation

■ Techniques

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



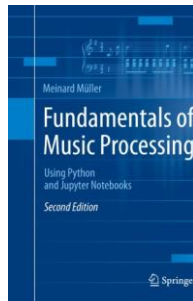
Spectrogram example (frequency x time)

Fig. 2.6

Audio Processing

Frequency Modulation

- Example: Opera singing
 - Estimation & sonification of fundamental frequency contours



FMP Notebooks

Audio Processing

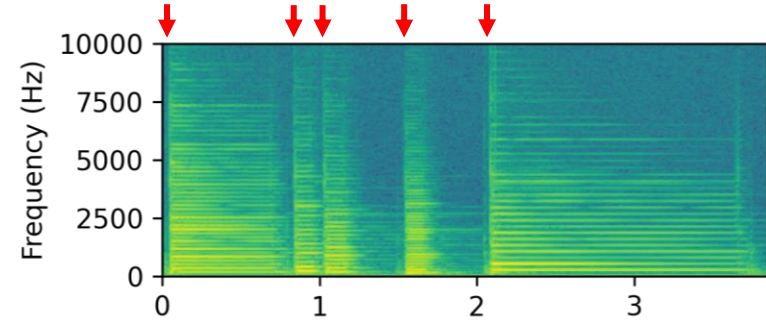
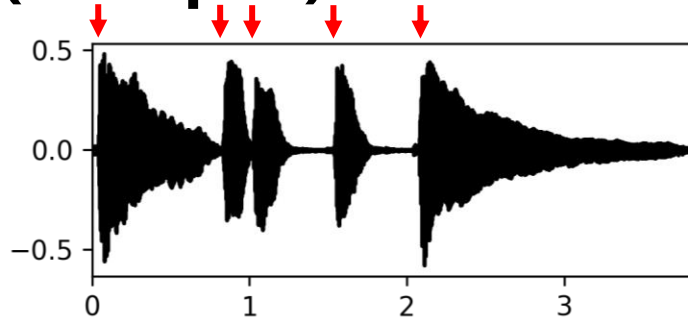
Transients

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

Audio Processing

Transients (Examples)

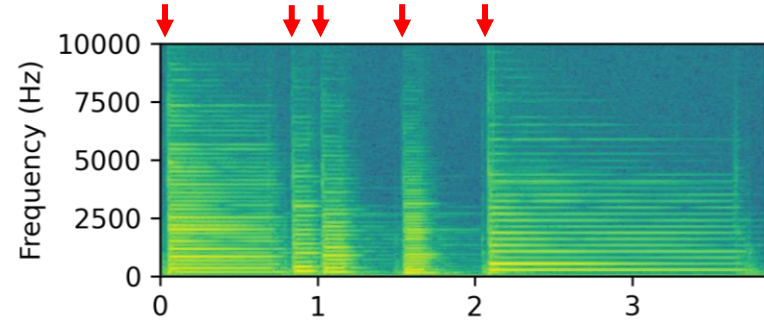
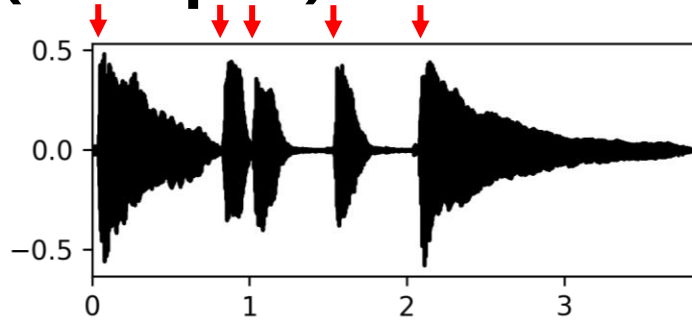
- String instruments
- 🔊 [Audio 1](#)



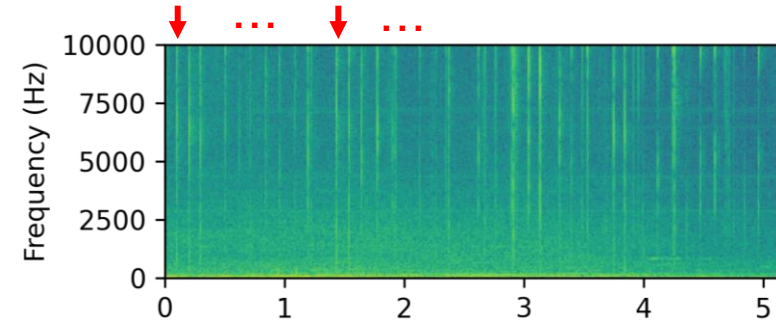
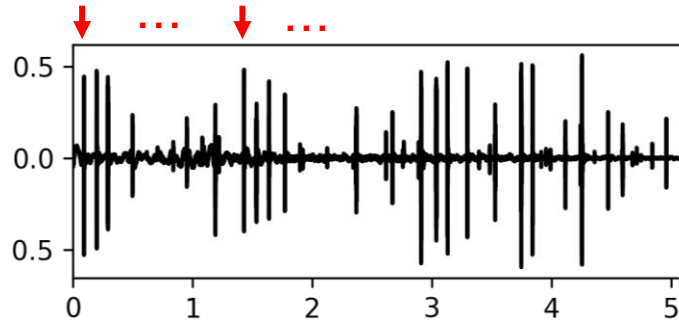
Audio Processing

Transients (Examples)

- String instruments



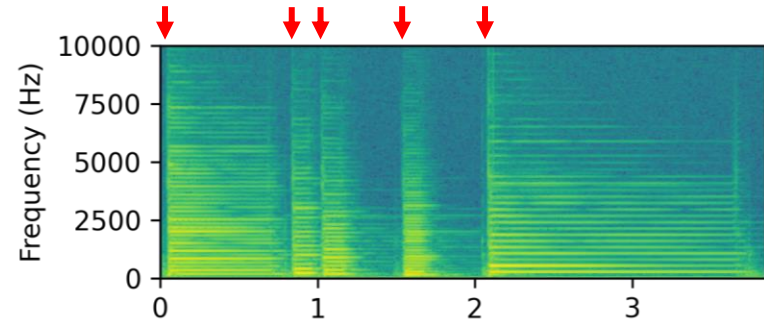
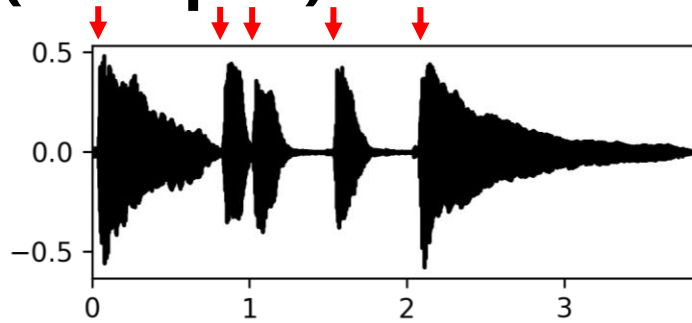
- Bat vocalizations



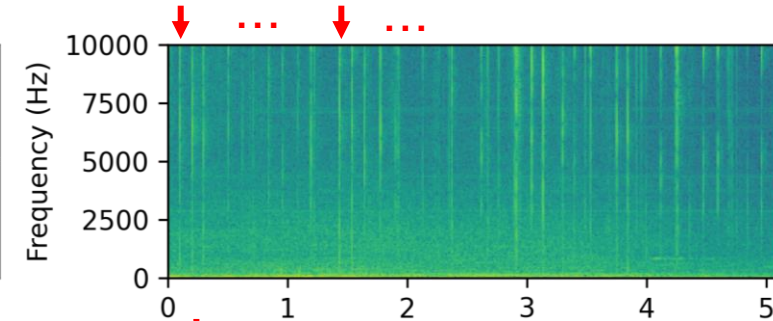
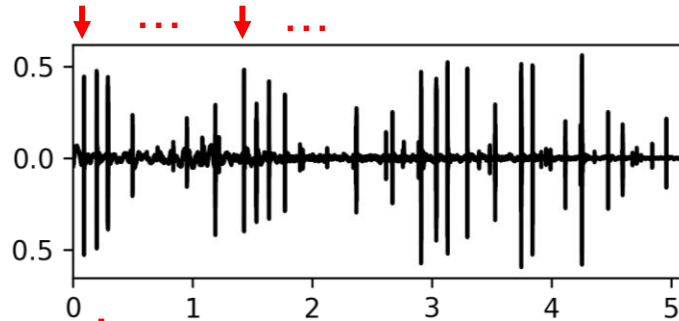
Audio Processing

Transients (Examples)

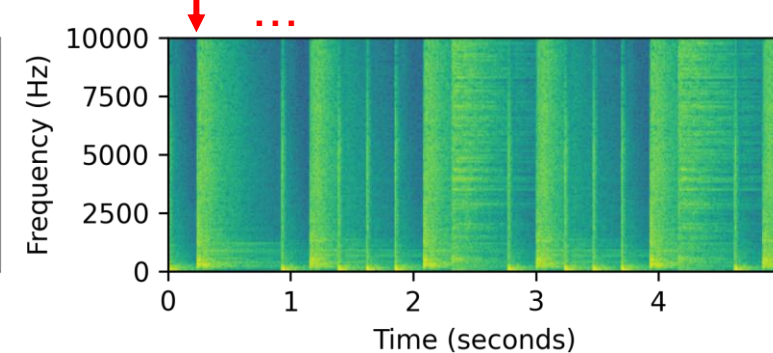
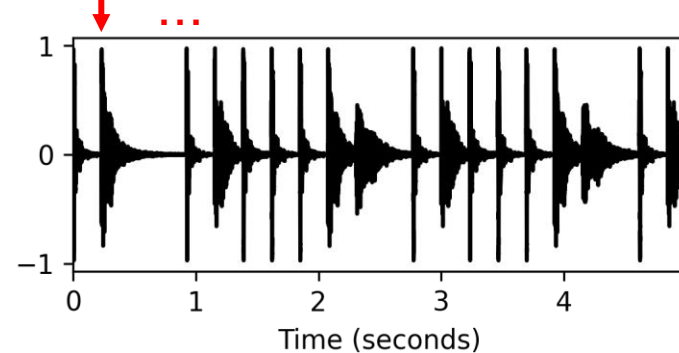
- String instruments



- Bat vocalizations



- Drum instruments




Audio Processing

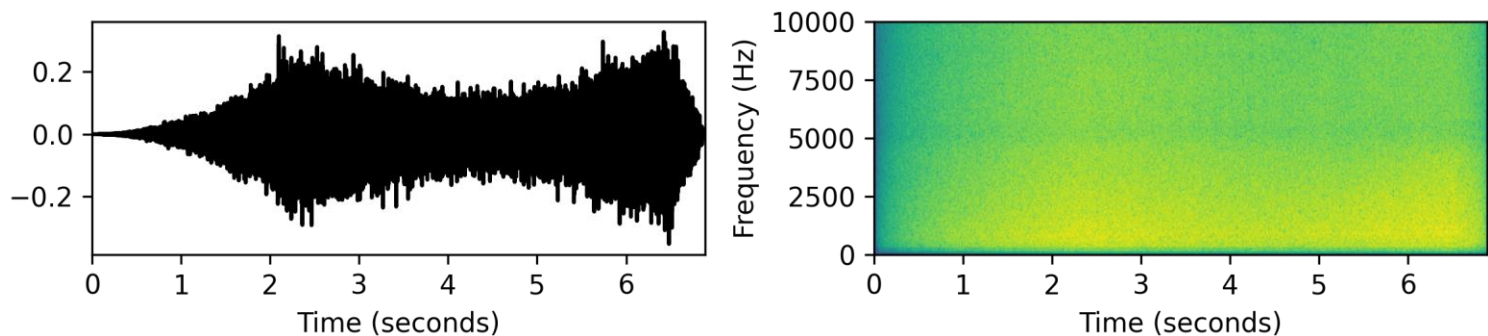
Noise

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

Audio Processing

Noise

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



Audio Processing

Temporal Envelope

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

Audio Processing

Temporal Envelope

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

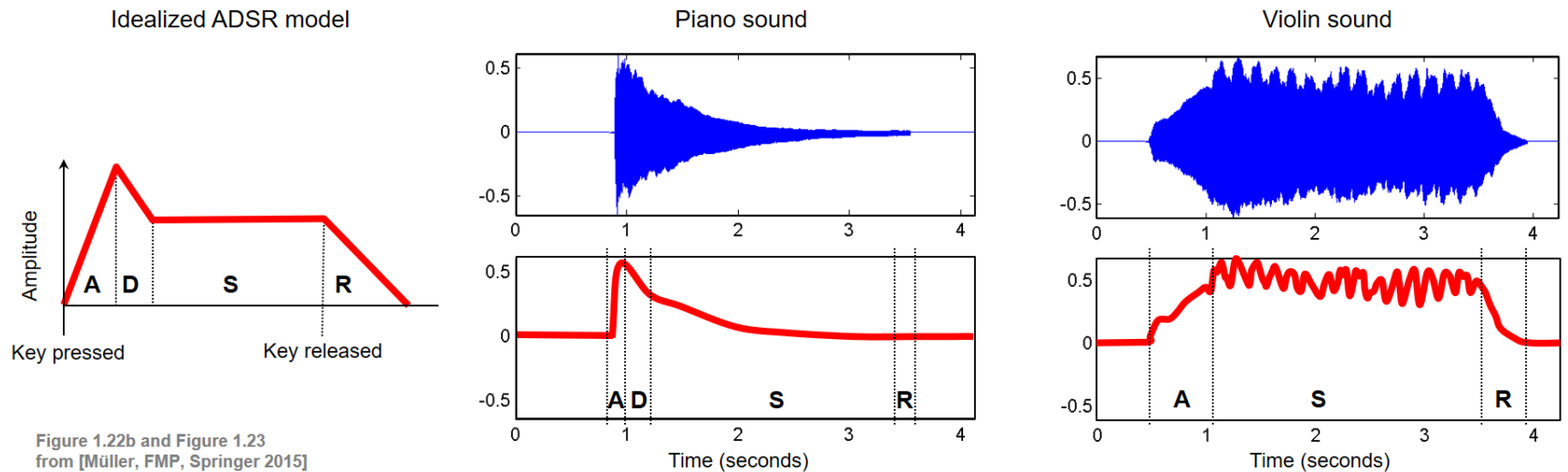
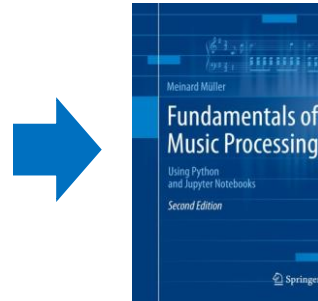


Fig. 2.7

Audio Processing

Temporal Envelope

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



FMP Notebooks

Fig. 2.7

Audio Processing

Timbre

- Perceptual attribute (complements pitch, loudness, duration)
- Difference between musical tones of same pitch & loudness

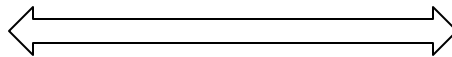
Audio Processing

Timbre

- Perceptual attribute (complements pitch, loudness, duration)
- Difference between musical tones of same pitch & loudness
- Timbre research

**(Subjective)
perceptual attributes**

Correlations?



(Objective) sound characteristics

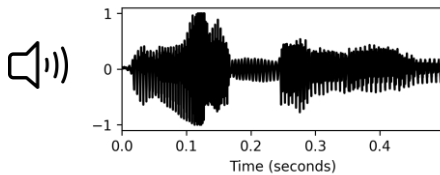
- Temporal / spectral envelope
- Tonal / noise-like components
- Partial (frequency) energies
- ...

Audio Processing

Mel-Frequency Cepstral Coefficients (MFCC)

- Compact representation of spectral envelope

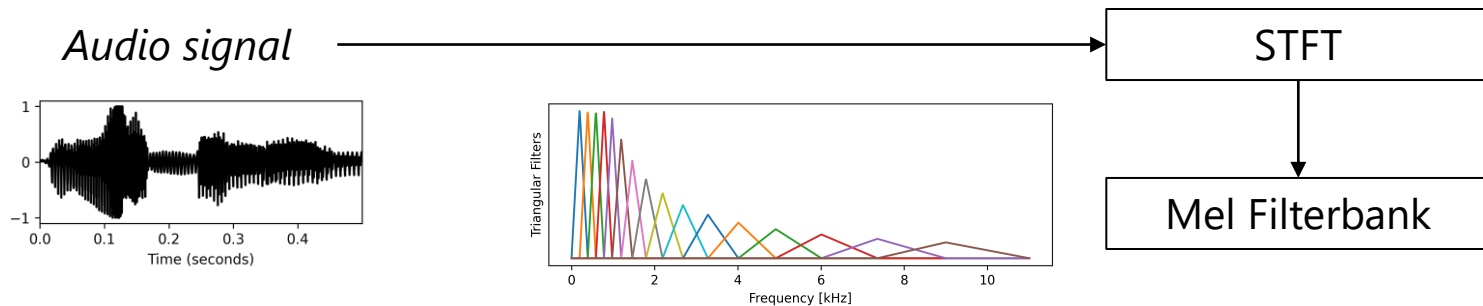
Audio signal



Audio Processing

Mel-Frequency Cepstral Coefficients (MFCC)

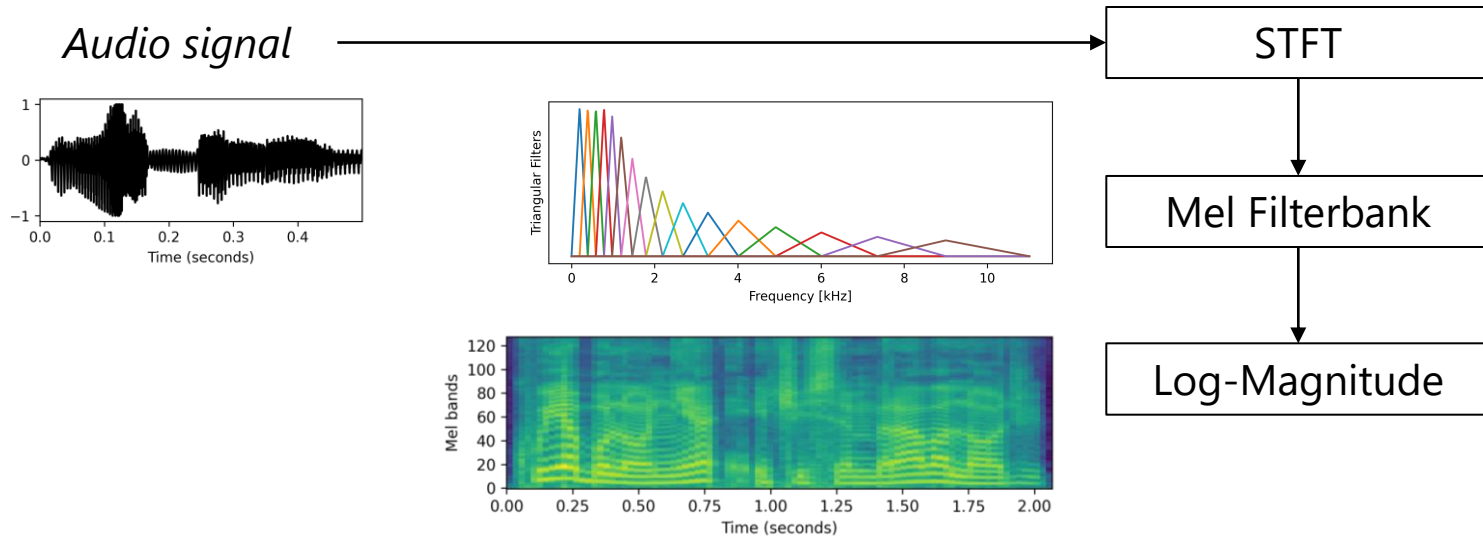
- Compact representation of spectral envelope



Audio Processing

Mel-Frequency Cepstral Coefficients (MFCC)

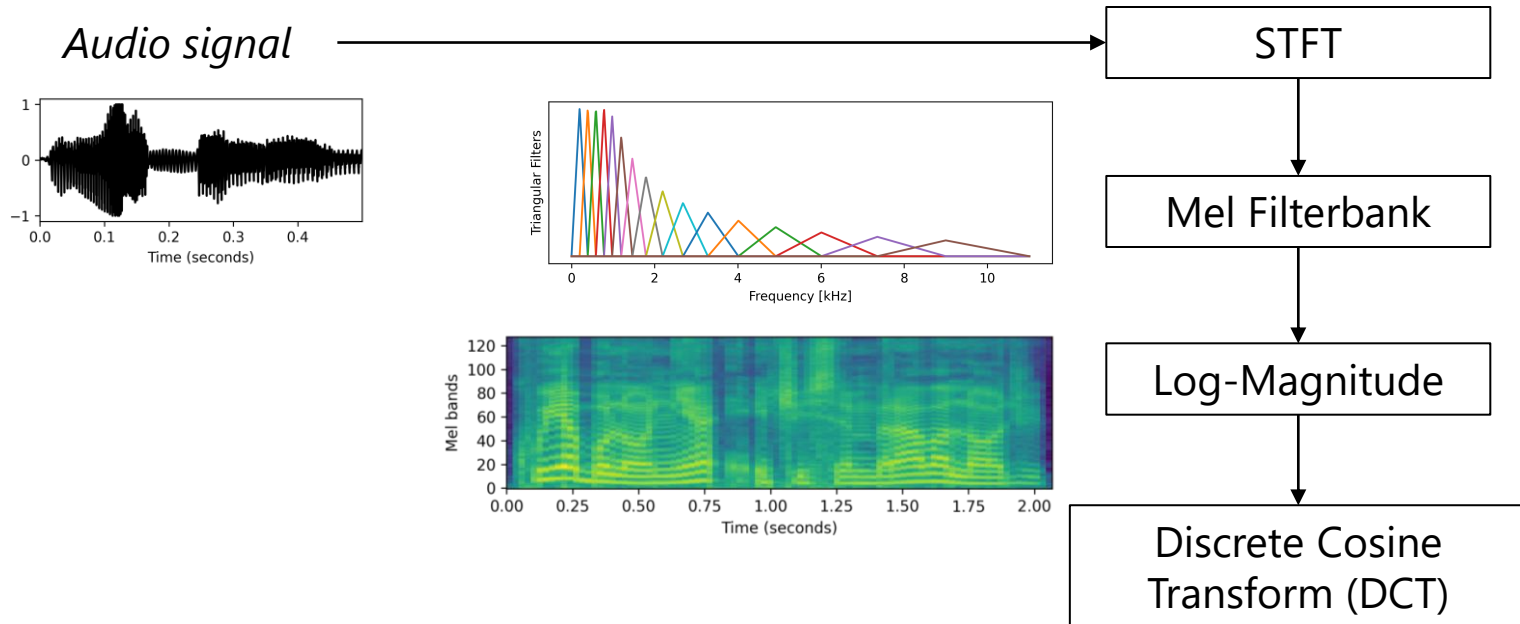
- Compact representation of spectral envelope



Audio Processing

Mel-Frequency Cepstral Coefficients (MFCC)

- Compact representation of spectral envelope

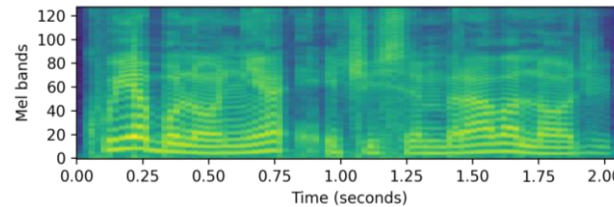
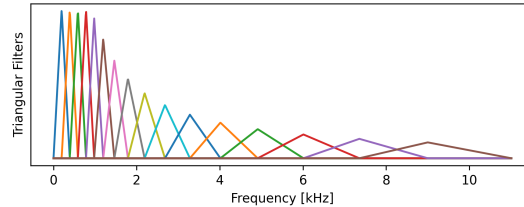
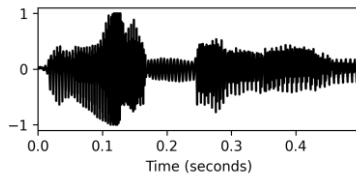


Audio Processing

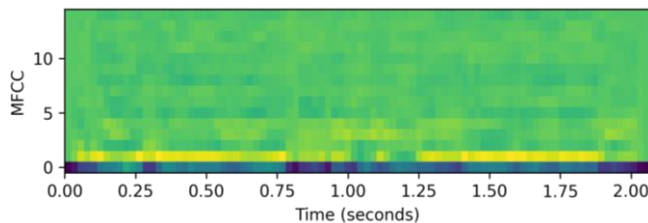
Mel-Frequency Cepstral Coefficients (MFCC)

- Compact representation of spectral envelope

Audio signal



MFCC



STFT

Mel Filterbank

Log-Magnitude

Discrete Cosine Transform (DCT)

Audio Processing

Constant-Q Transform

- STFT (linearly-spaced frequencies)
- CQT (logarithmically-spaced, closer to human auditory perception)

Audio Processing

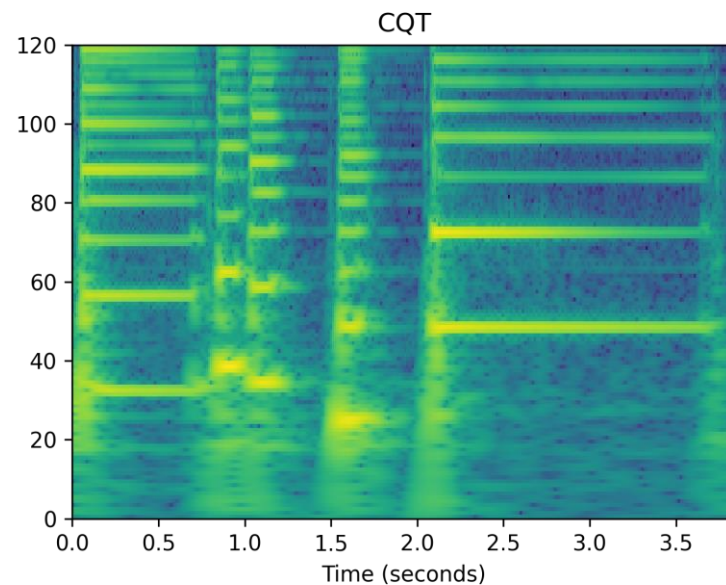
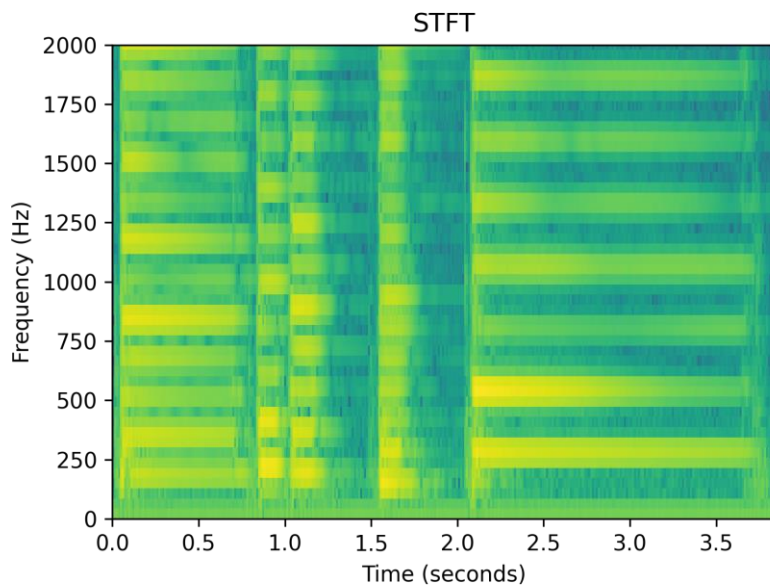
Constant-Q Transform

- 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 Processing

Constant-Q Transform

- 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

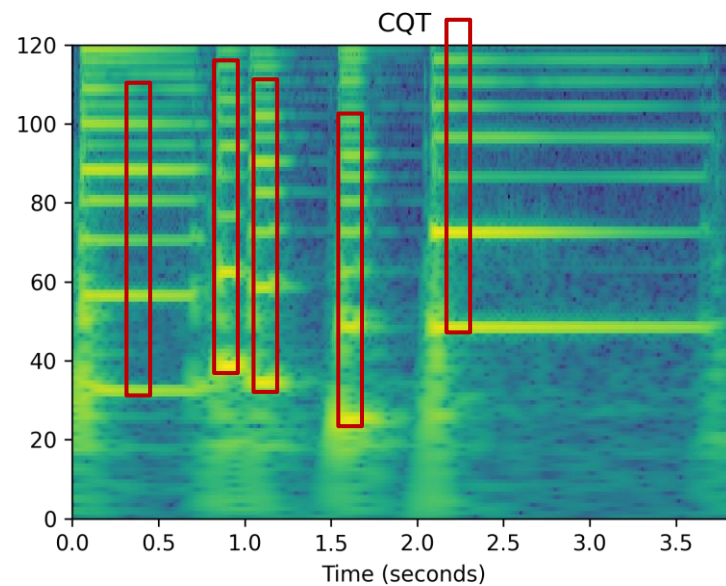
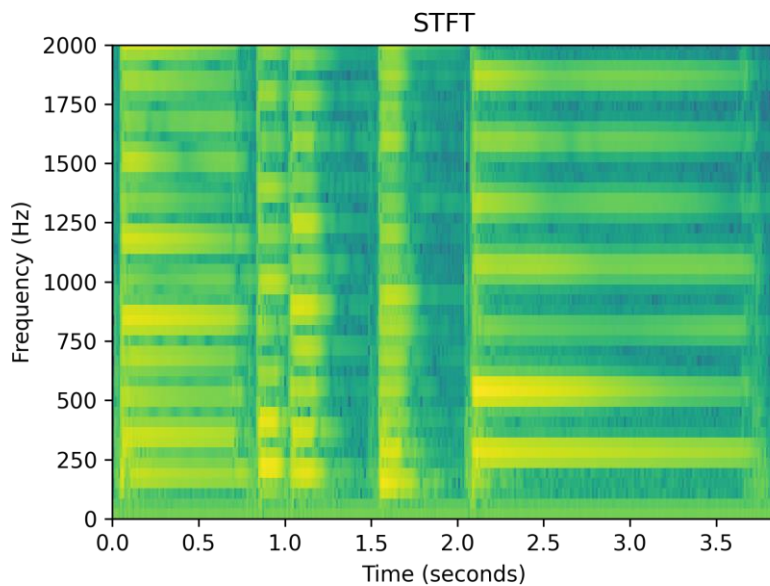


Audio 1

Audio Processing

Constant-Q Transform

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



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]

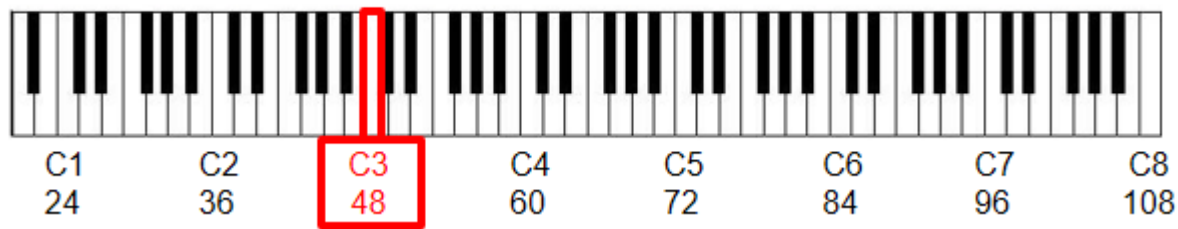


Fig. 2.8

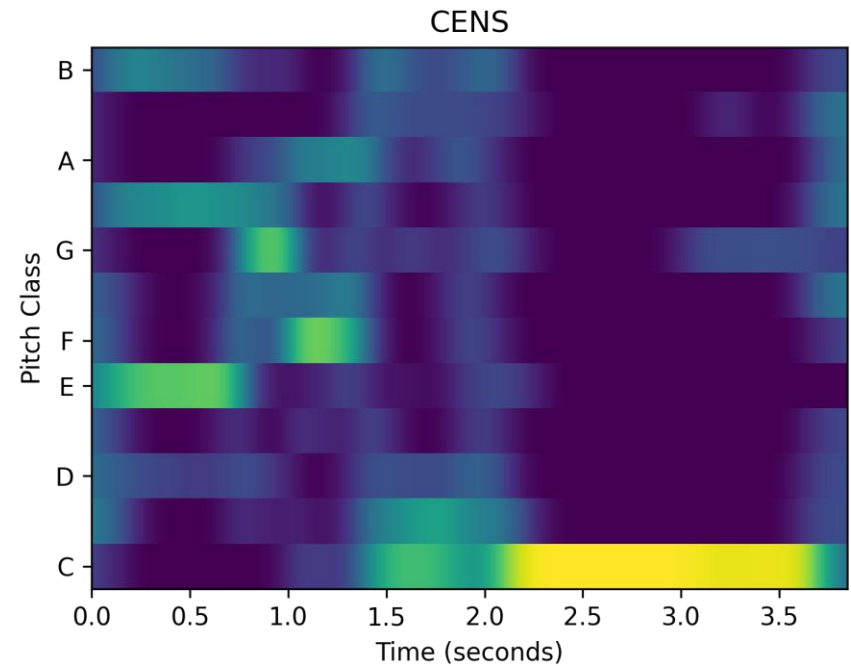
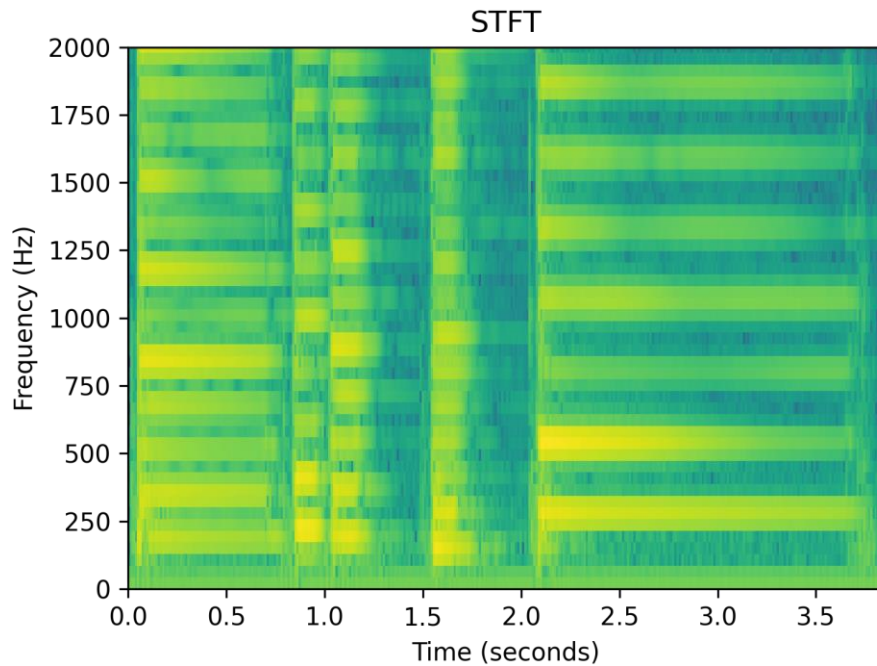
Audio Processing

Chroma Features

■ Example



[Audio 1](#)



Audio Processing

Programming Session #2

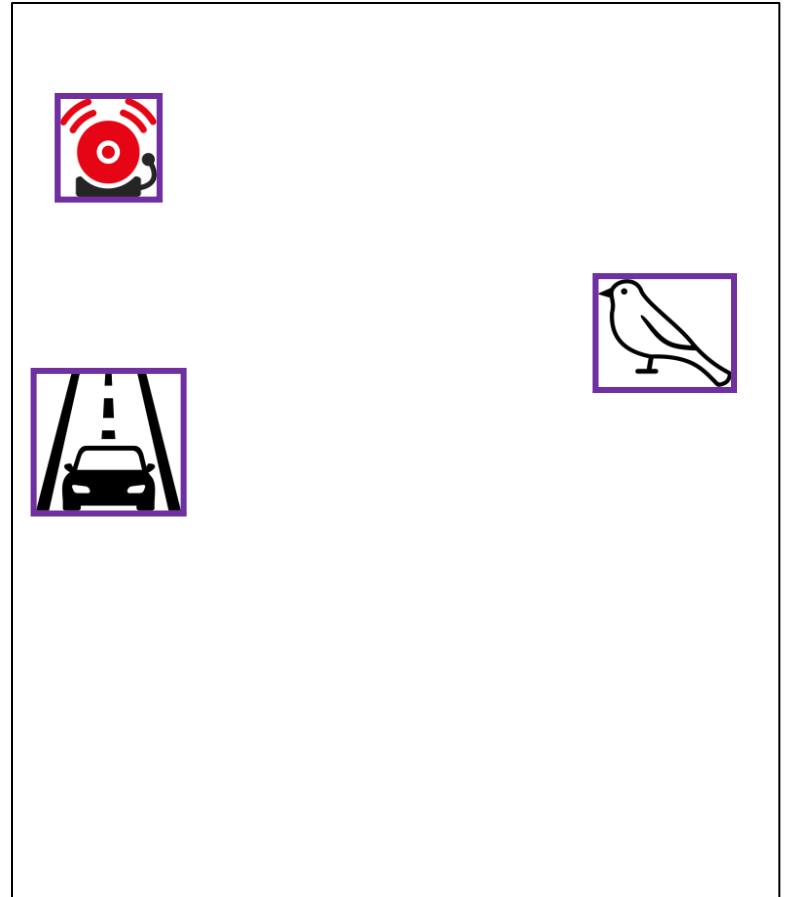


Fig. 2.1

Audio Processing

Auditory Scene

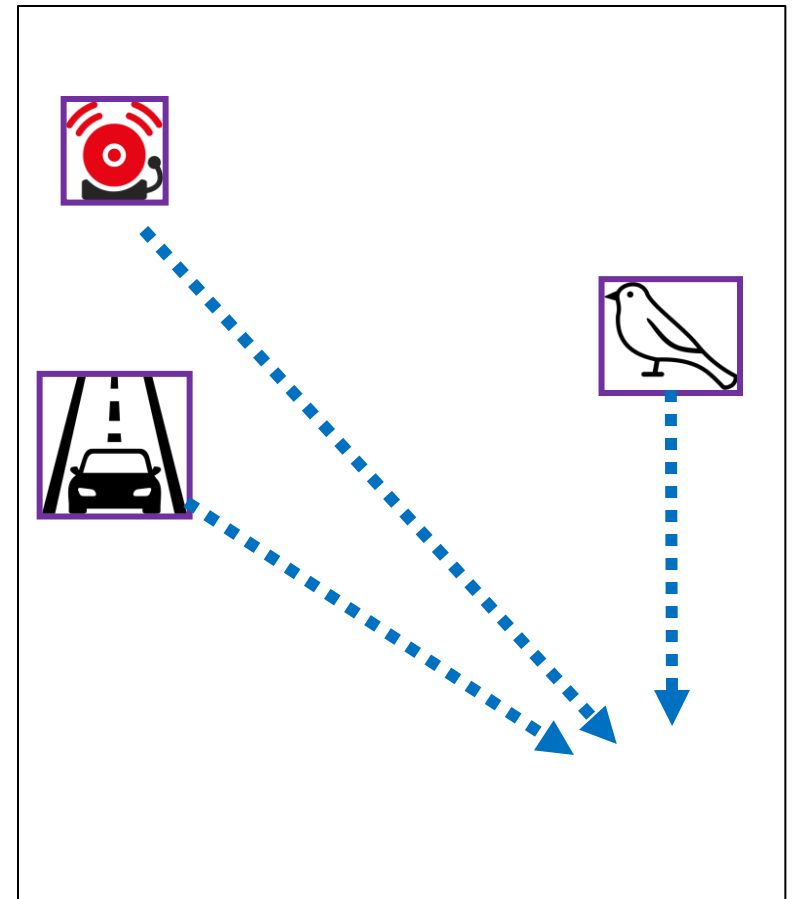
- Distributed sound producing events
(sound sources)



Audio Processing

Auditory Scene

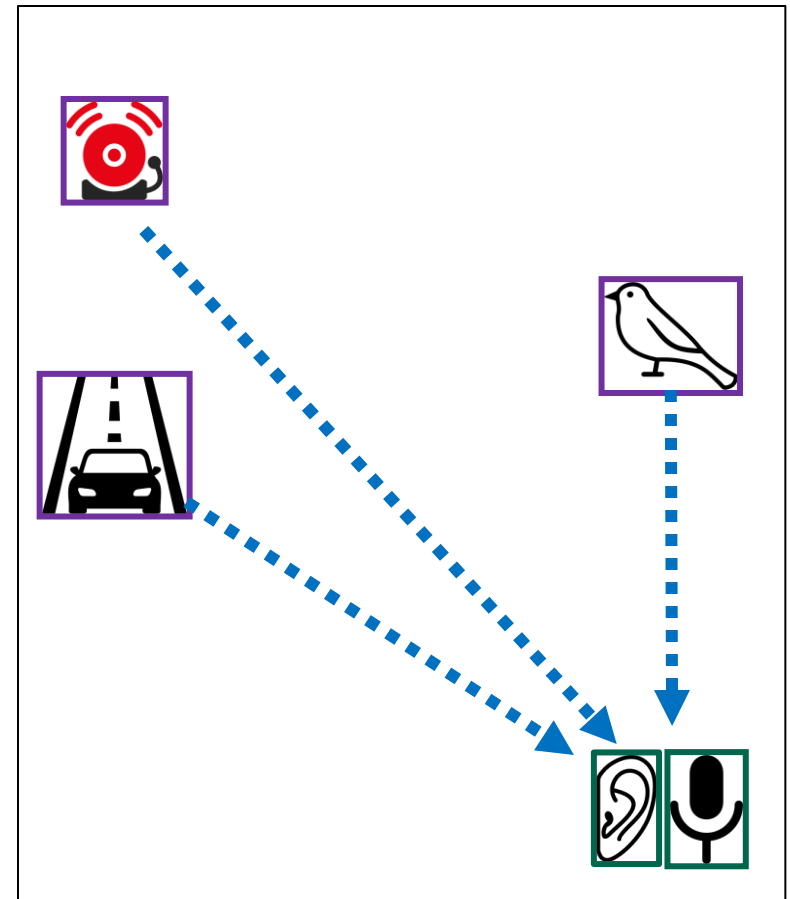
- Distributed sound producing events (sound sources)
- Sound propagation through space



Audio Processing

Auditory Scene

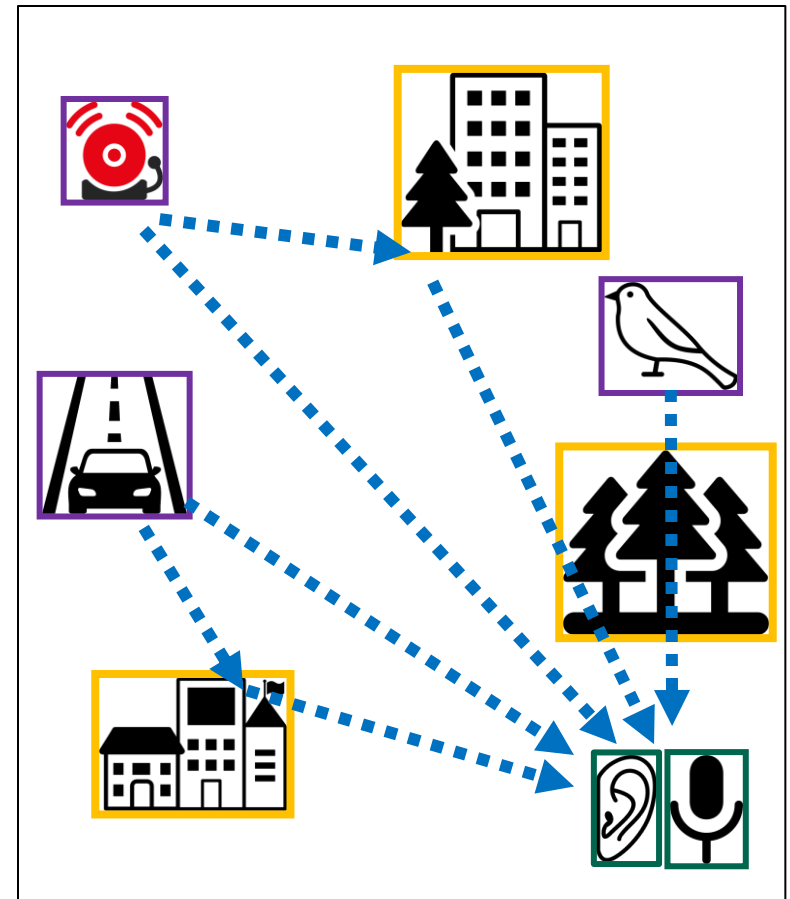
- Distributed sound producing events (sound sources)
- Sound propagation through space
- Perceived (ear) or recorded (microphone)



Audio Processing

Auditory Scene

- Distributed sound producing events (sound sources)
- Sound propagation through space
- Perceived (ear) or recorded (microphone)
- Room acoustics
 - Reflection (surfaces)
 - Diffraction (objects)



Audio Processing

Auditory Scene Analysis

- Acoustic signals are processed by peripheral auditory system (pinna, eardrum, ossicles, cochlea, auditory nerve)
- Auditory scene analysis
 - Segregation and grouping of spectro-temporal regularities into auditory objects (percept)
 - Identify corresponding sound events
 - Can be simple and complex patterns / sequences



Images

Fig. 1: <https://www.acs.psu.edu/drussell/Demos/waves-intro/Lwave-Red-2.gif>

Fig. 1.1: TODO FMP

Fig. 2: https://www.mathworks.com/help/dsp/ref/stft_output.png

Fig. 2.1: https://upload.wikimedia.org/wikipedia/commons/thumb/3/38/Jupyter_logo.svg/1200px-Jupyter_logo.svg.png

Fig. 2.5: https://pressbooks.pub/app/uploads/sites/140/2022/07/Piano_to_F.jpg

Fig. 2.6: https://www.hfm-weimar.de/popvoices/media/_glossar/BH8.png

Fig. 2.7: TODO FMP

Fig. 2.8: TODO FMP (https://www.audiolabs-erlangen.de/resources/MIR/FMP/data/C3/FMP_C3_F03a.png)

Fig. 3: <https://i.makeagif.com/media/9-11-2015/6HmpFN.gif>

Fig. 4: <https://prezigram-assets.prezicdn.net/e53764d415cd58a530e5f66144779100cc9bdc843686bbb9ea5f5c273ef1d1784bcc63c4f75847717119716f5b62701de16797ec8a51ca9a9247981613460ebc>

References

[Merriam-Webster] <https://www.merriam-webster.com/dictionary/sound>

Audio

[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)