Part II

MUSIC CONTENT ANALYSIS AND SIMILARITY

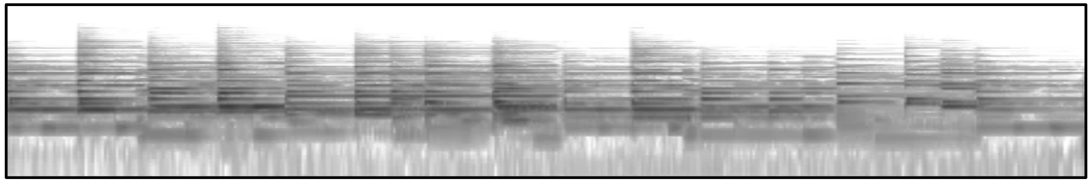
Department of
Computational
Perception

Categorization of Content-Based Features

Domain:

- **Time domain**
  consider signal in time/amplitude representation ("waveform")

- **Frequency domain**
  consider signal in frequency/magnitude representation



Transformation from time to frequency domain using, e.g., Fast Fourier Transform (FFT)

Department of Computational Perception

Categorization of Content-Based Features

Temporal scope:

- **Instantaneous**
  feature is valid for a “point in time” (NB: time resolution of ear is several msec!)
- **Segment**
  feature is valid for a segment, e.g., phrase, chorus (on a high level), or a chunk of *n* consecutive seconds in the audio signal
- **Global**
  feature is valid for whole audio excerpt or piece of music

Categorization of Content-Based Features

Level of abstraction:

- **Low-level**
  properties of audio signal (e.g., energy, zero-crossing-rate)
- **Mid-level**
  aggregation of low-level descriptors,
  applies psycho-acoustic models (cf. MFCC, FP);
  typically the level used when estimating similarity
- **High-level**
  musically meaningful to listener, e.g., melody, themes, motifs;
  "semantic" categories, e.g., genre, time period, mood, …
  (cf. semantic tags learned from audio features)

Department of
Computational
Perception

How to Describe Audio Content?

Possible idea: get features that describe music the way humans do and compute similar songs based on this information

Unfortunately we are are not able to extract most of these features reliably (or at all…)

- even “simple” human concepts are difficult to model (**“semantic gap”**)
- even tempo estimation is very hard…
- NB: a human annotation approach is done in the Music Genome Project (cf. Pandora’s automatic radio station service)

Furthermore some of these features are quite subjective (e.g., mood)

Need to find computable descriptors that capture these dimensions somehow (…and work acceptably)

Descriptors of Content

Acoustic property to describe: ♪

- **Loudness:** perceived strength of sound; *e.g., energy*
- **Pitch:** frequency, psychoacoustic ordering of tones (on scale; from low to high); *e.g., chroma-features*
- **Timbre:** "tone color", what distinguishes two sounds with same pitch and loudness; *e.g., MFCCs*
- **Chords and harmony:** simultaneous pitches
- **Rhythm:** pattern in time; *e.g., FPs*
- **Melody:** sequence of tones; combination of pitch and rhythm

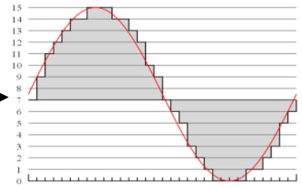
cf. (Casey et al.; 2008)

Department of Computational Perception

Scheme of Content-Based Feature Extraction

analog signal

sampling
quantization

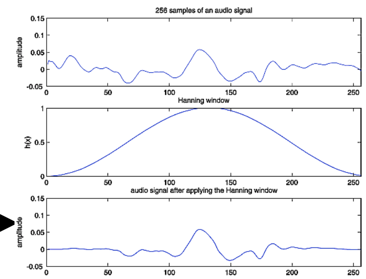


Pulse Code Modulation (PCM)

framing

frame 1: e.g. sample 1...256
frame 2: e.g. sample 129...384
frame 3: e.g. sample 257...512
...
frame n

windowing



FFT

time domain feature calculation

frequency domain feature calculation

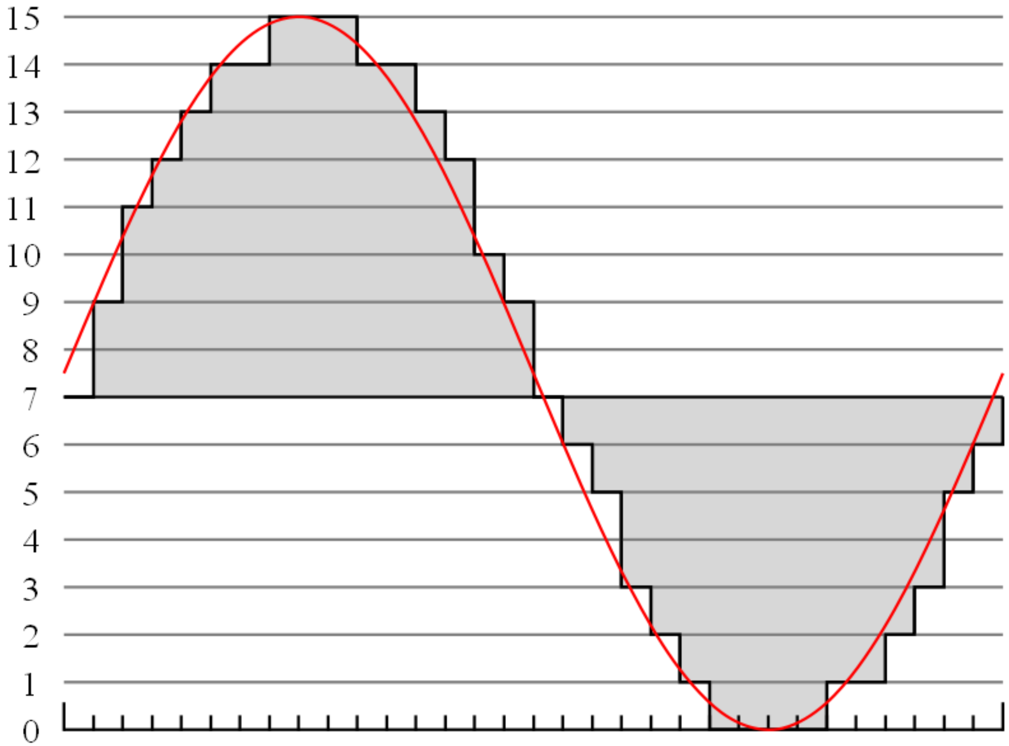
aggregation, model building (mean, median, sum, GMM, HMM)

feature value, vector, or matrix

♪

Department of Computational Perception

Analog-Digital-Conversion (ADC)

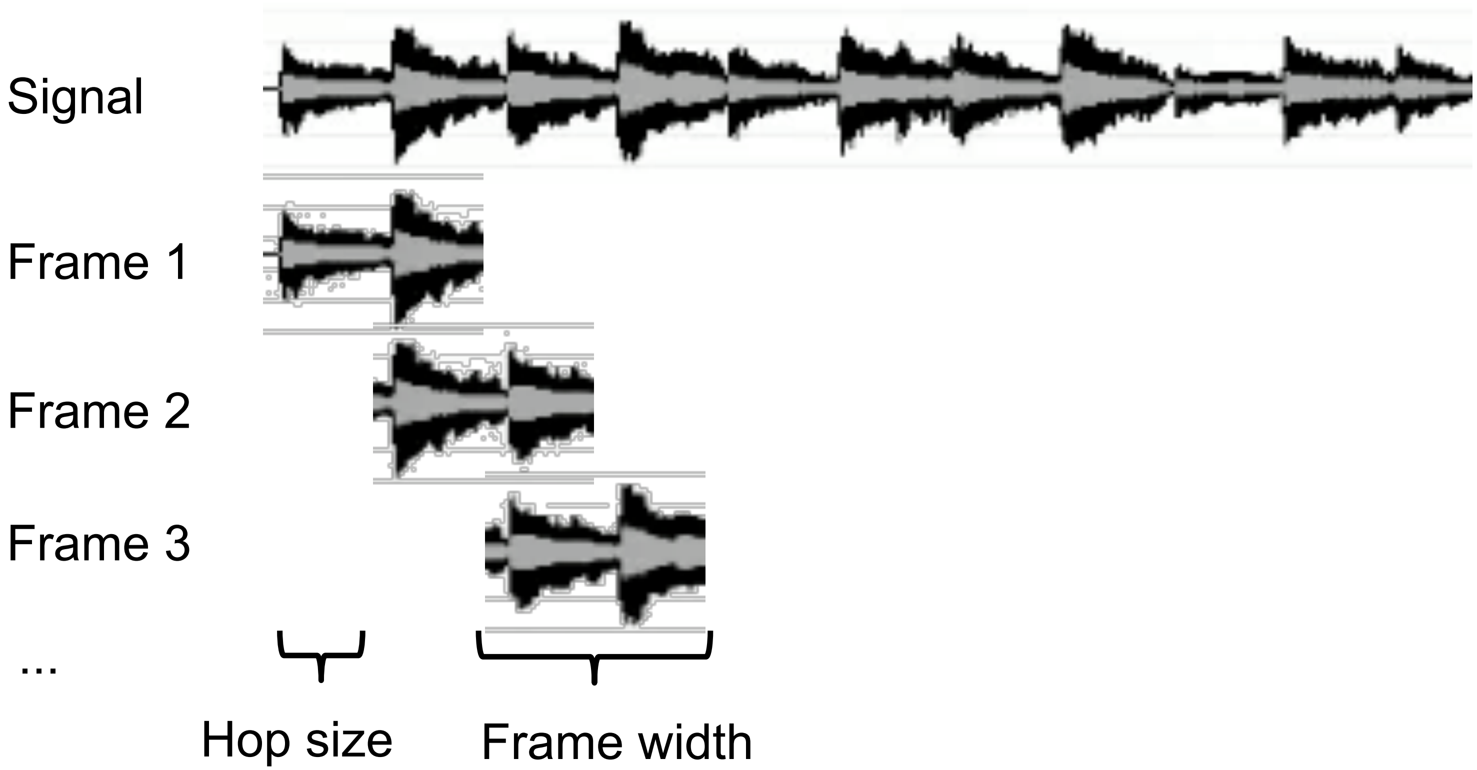


PCM: analog signal is sampled at equidistant intervals and quantized in order to store it in digital form (here with 4 bits)

Problems that may occur in ADC:

- **Quantization error**: difference between the actual analog value and quantized digital value
- *Solution: finer resolution (use more bits for encoding), common choice in music encoding: 16 bits per channel*
- Due to **Nyquist–Shannon Sampling Theorem**, frequencies above ½ of sampling frequency (Nyquist frequency) are discarded or heavily distorted
- *Solution: choose a sampling frequency that is high enough (e.g. 44,100 Hz for Audio CDs)*

Framing



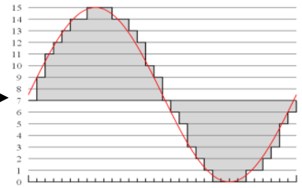
In short-time signal processing, pieces of music are cut into segments of fixed length, called frames, which are processed one at a time; typically, a frame comprises 256 - 4096 samples.

Department of Computational Perception

Scheme of Content-Based Feature Extraction

analog signal

sampling
quantization

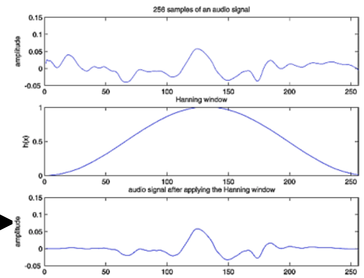


Pulse Code Modulation (PCM)

framing

frame 1: e.g. sample 1...256
frame 2: e.g. sample 129...384
frame 3: e.g. sample 257...512
...
frame n

windowing



FFT

time domain feature calculation

frequency domain feature calculation

aggregation, model building (mean, median, sum, GMM, HMM)

feature value, vector, or matrix

♪

Department of Computational Perception

Low-Level Feature: Zero Crossing Rate

Scope: time domain

s(k)...amplitude of k^{th} sample in time domain
K...frame size (number of samples in each frame)

Calculation:

$$ZCR_t = \frac{1}{2} \cdot \sum_{k=t \cdot K}^{(t+1)\cdot K - 1} \left| \operatorname{sgn}(s(k)) - \operatorname{sgn}(s(k+1)) \right|$$

Description:

number of times the amplitude value changes its sign within frame *t*

Remarks:

commonly used as part of a low-level descriptor set

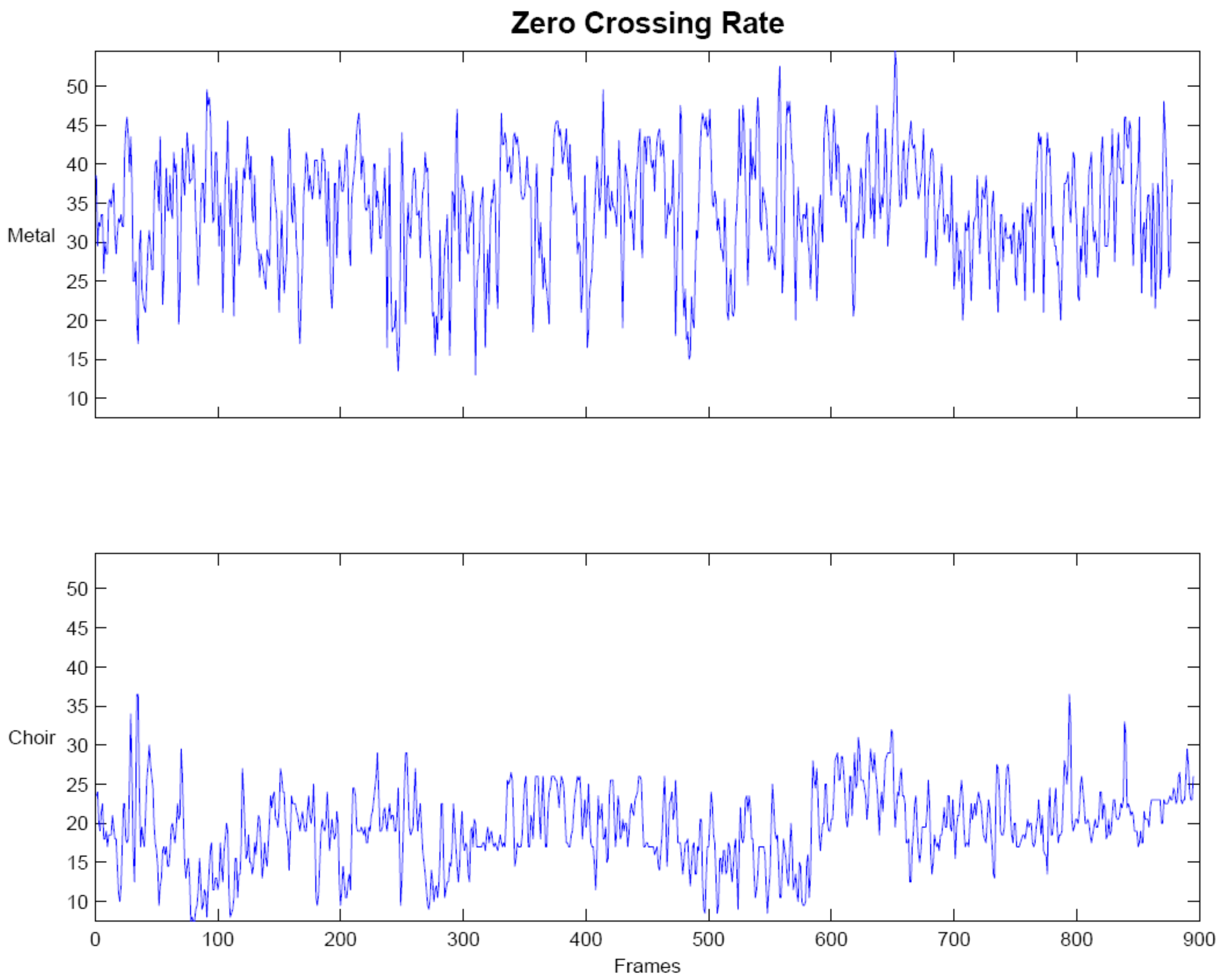
\+ might be used as an indicator of pitch

\+ sometimes stated to be an approximate measure of the signal's noisiness

– in general, low discriminative power

Department of Computational Perception

Zero Crossing Rate: Illustration



Low-Level Feature: Amplitude Envelope

Scope: time domain

s(k)...amplitude of k^{th} sample in time domain
K...frame size (number of samples in each frame)

Calculation:

$$AE_t = \max_{k=t\cdot K}^{(t+1)\cdot K-1} |s(k)|$$

Description:
maximum amplitude value within frame *t*

Remarks:

similar to RMS energy (see next), but less stable

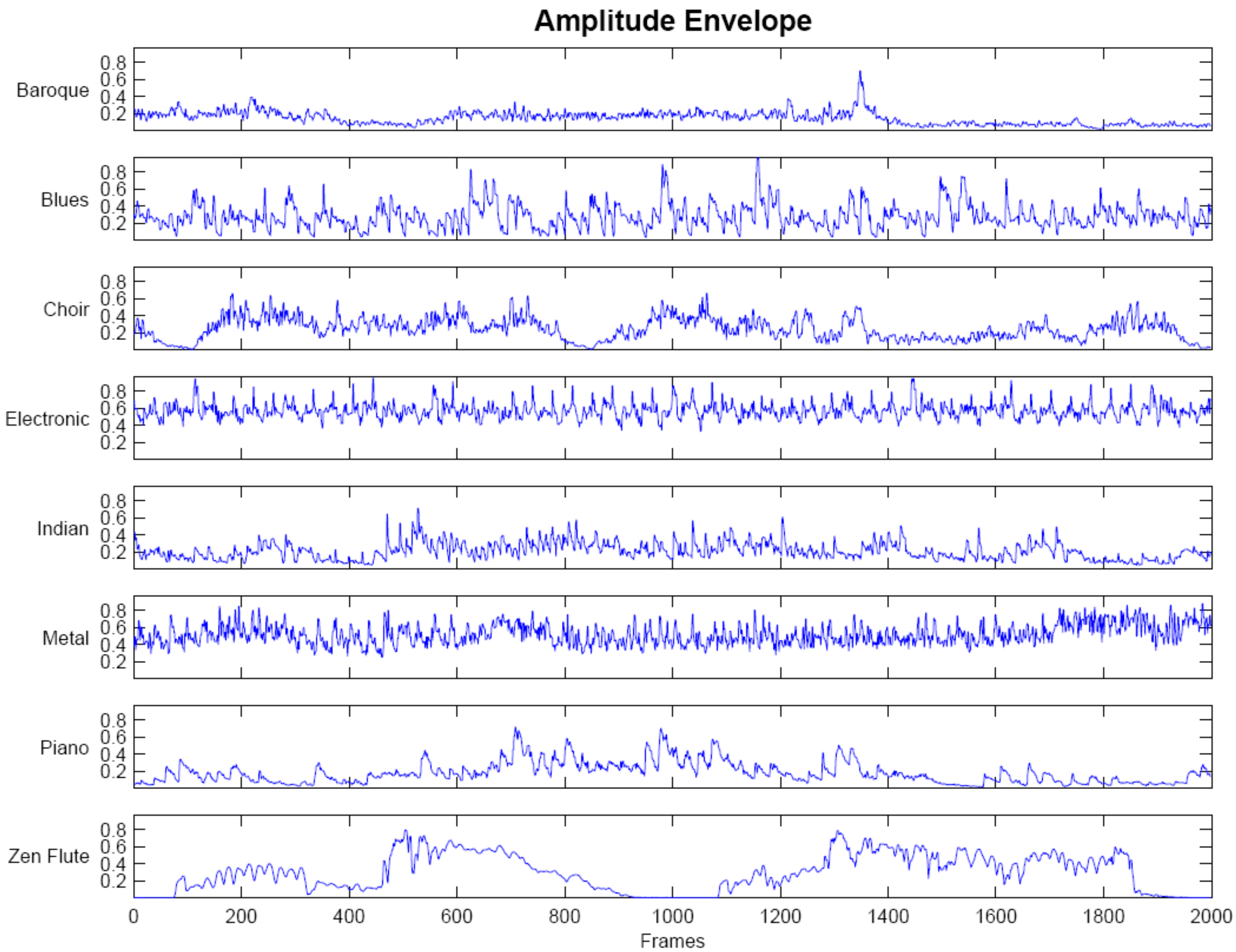
+ important for beat-related feature calculation, e.g. for beat detection

– discriminative power not clear

– sensitive to amplitude outliers

Department of
Computational
Perception

Amplitude Envelope: Illustration



artment of
putational
eption

Low-Level Feature: RMS Energy

Root-Mean-Square Energy (aka RMS power, RMS level, RMS amplitude)

Scope: time domain

Calculation:

$$RMS_t = \sqrt{\frac{1}{K} \cdot \sum_{k=t\cdot K}^{(t+1)\cdot K-1} s(k)^2}$$

s(k)...amplitude of k^{th} sample in time domain
K...frame size (number of samples in each frame)

Remarks:

+ beat-related feature, can be used for beat detection

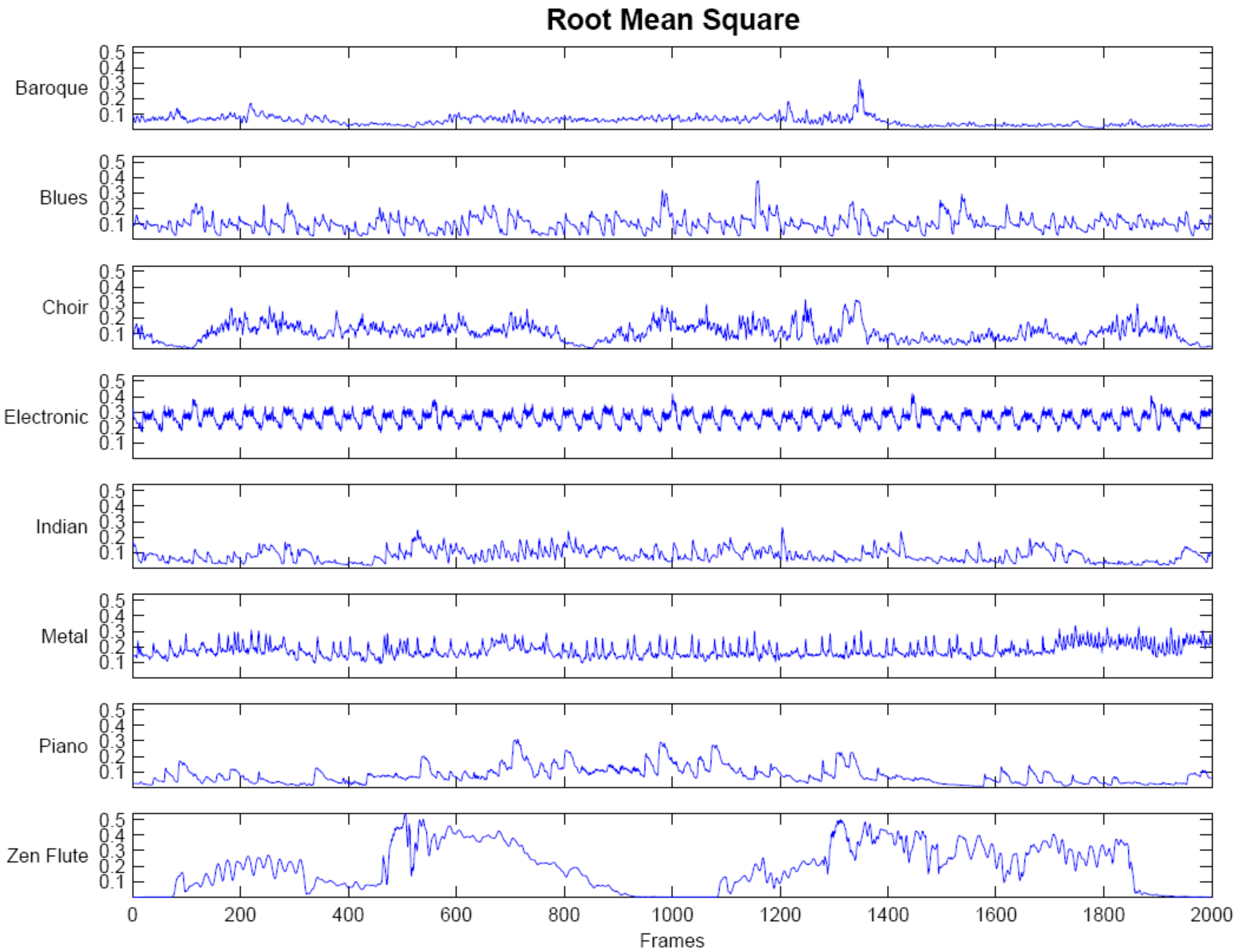
+ related to perceived intensity

+ good loudness estimation

– discriminative power not clear

Department of Computational Perception

RMS Energy: Illustration

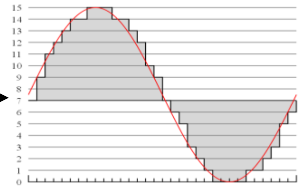


artment of
putational
eption

Scheme of Content-Based Feature Extraction

analog signal

sampling
quantization

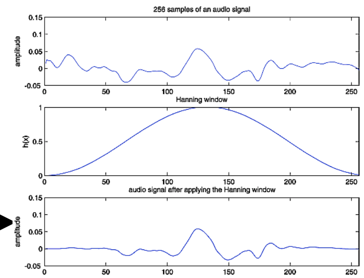


Pulse Code Modulation (PCM)

framing

frame 1: e.g. sample 1...256
frame 2: e.g. sample 129...384
frame 3: e.g. sample 257...512
...
frame n

windowing



FFT

time domain feature calculation

frequency domain feature calculation

aggregation, model building (mean, median, sum, GMM, HMM)

feature value, vector, or matrix

Department of Computational Perception

Fourier Transform

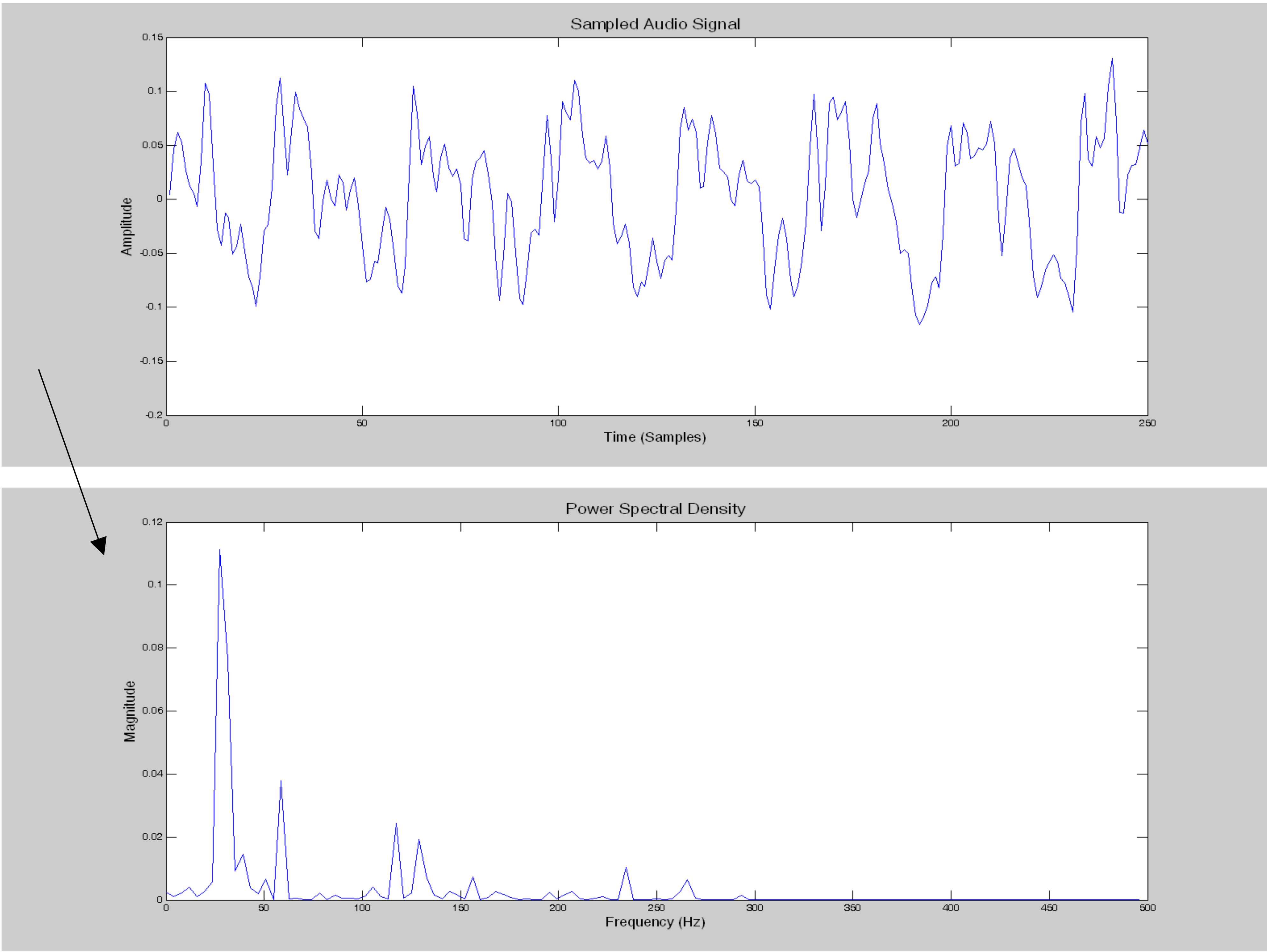
Transformation of the signal
from **time domain** (time vs. amplitude)
to **frequency domain** (frequency vs. magnitude)

- Theorem: any continuous periodic function with a period of 2π can be represented as the sum of sine and/or cosine waves (of different frequencies)
- Implication: any audio signal can be decomposed into an infinite number of overlapping waves when periodic

Jean Baptiste Joseph Fourier

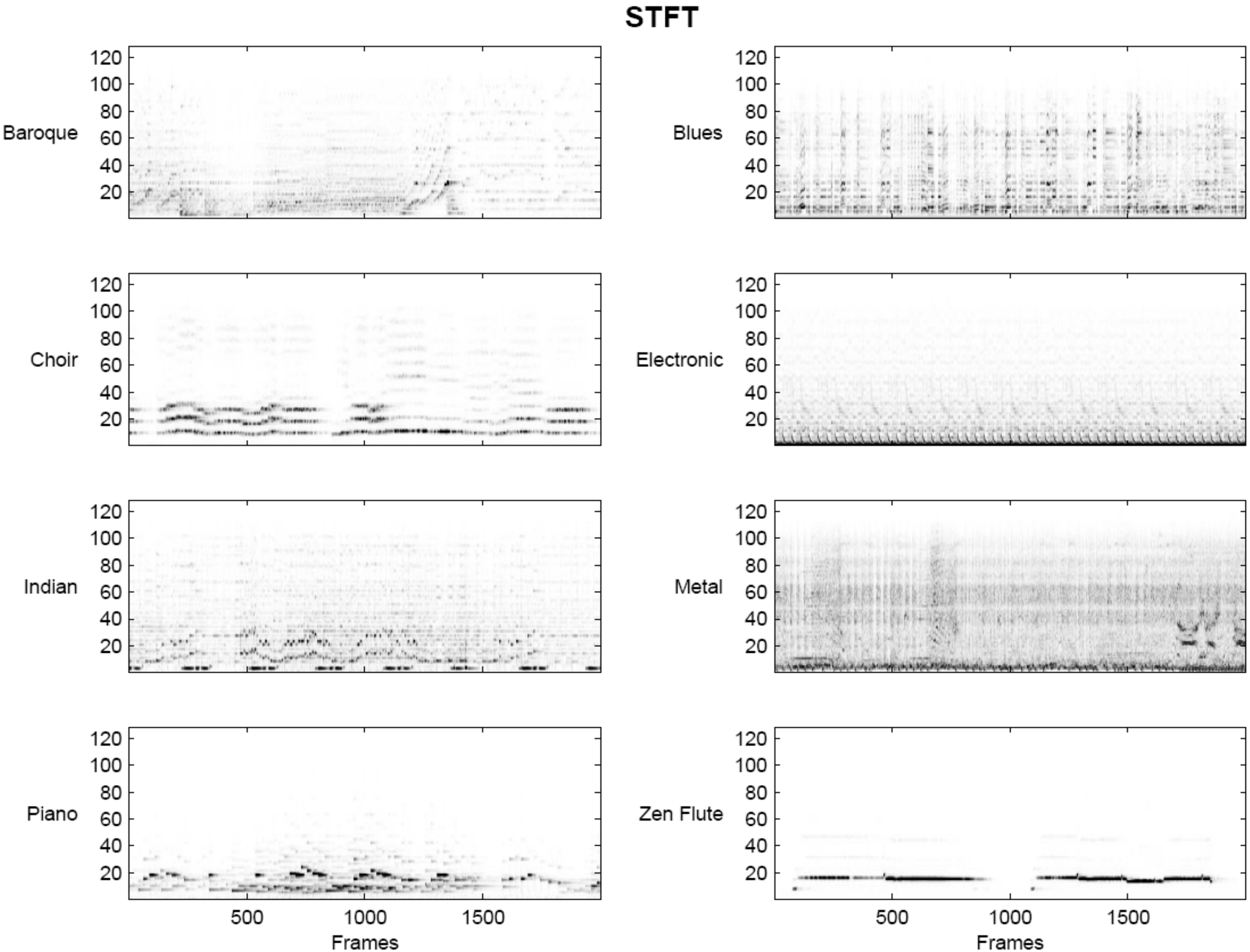
- Periodicity is achieved by multiplying the PCM magnitude values of each frame with a suited function, e.g., a Hanning window (**windowing**)
- In our case: **Discrete Fourier Transform (DFT)**
- In practice efficiently calculated via **Fast Fourier Transform (FFT)** (Cooley, Tukey; 1965)

Department of Computational Perception



Sampled Audio Signal
Amplitude
Time (Samples)
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0
50
100
150
200
250
Power Spectral Density
Magnitude
Frequency (Hz)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200
250
300
350
400
450
500

Representation as STFT



nt of
onal
n

Low-Level Feature: Spectral Centroid

Scope: frequency domain

Calculation:

$$C_t = \frac{\sum_{n=1}^{N} M_t(n) \cdot n}{\sum_{n=1}^{N} M_t(n)}$$

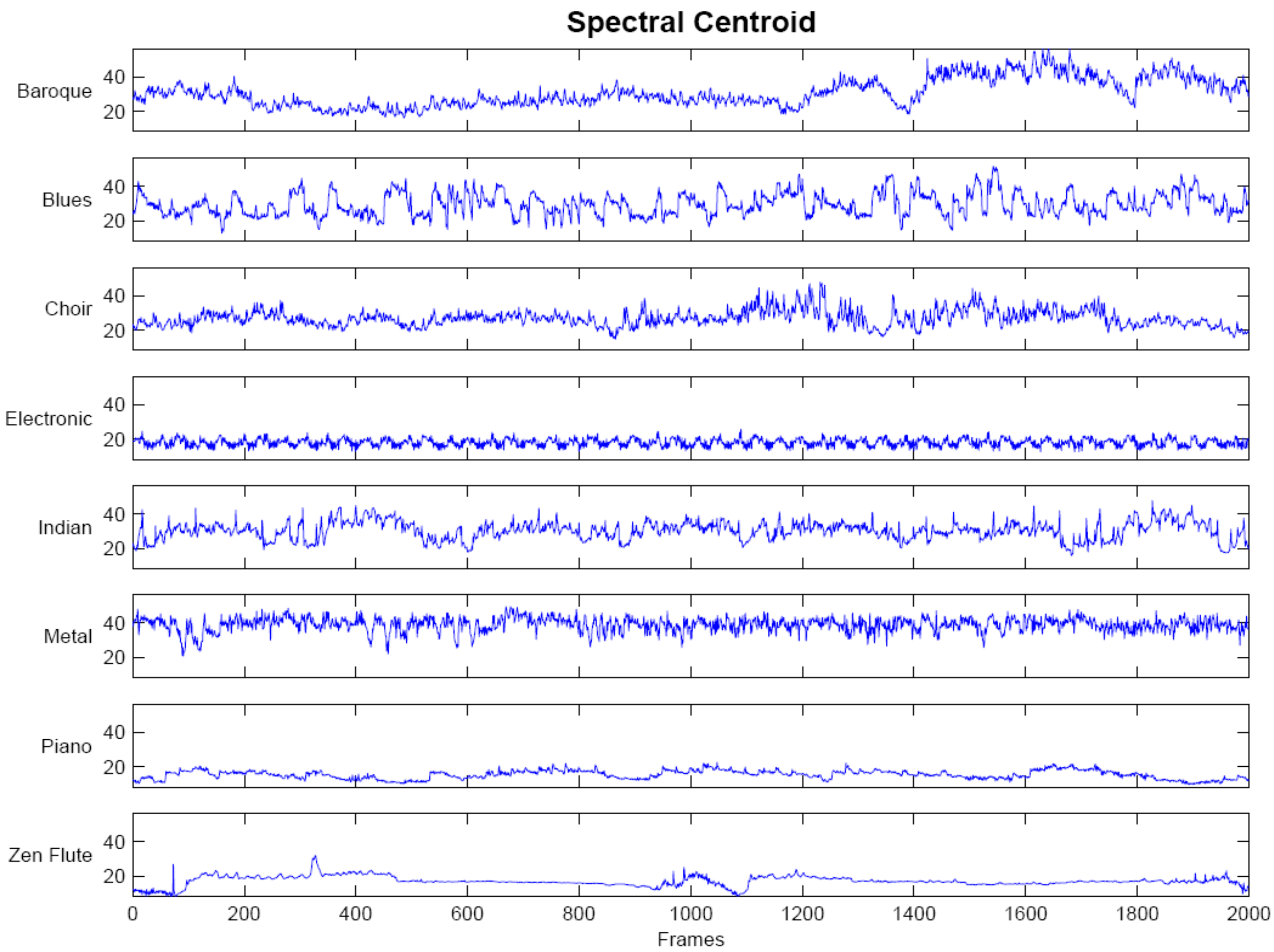
$M_t(n)$...magnitude in frequency domain at frame t and frequency bin n
N...number of highest frequency band

Description: center of gravity of the magnitude spectrum of the DFT, i.e. the frequency (band) region where most of the energy is concentrated

Remarks:
used as measure of sound sharpness (strength of high frequency energy)
– sensitive to low pass filtering (downsampling) as the high frequency bands are given more weight
– sensitive to white noise (for the same reason)

Department of Computational Perception

Spectral Centroid: Illustration



tment of
utational
ption

Low-Level Feature: Bandwidth

Scope: frequency domain

Calculation:

$$BW_t^2 = \frac{\sum_{n=1}^{N} (n - C_t)^2 \cdot M_t(n)}{\sum_{n=1}^{N} M_t(n)}$$

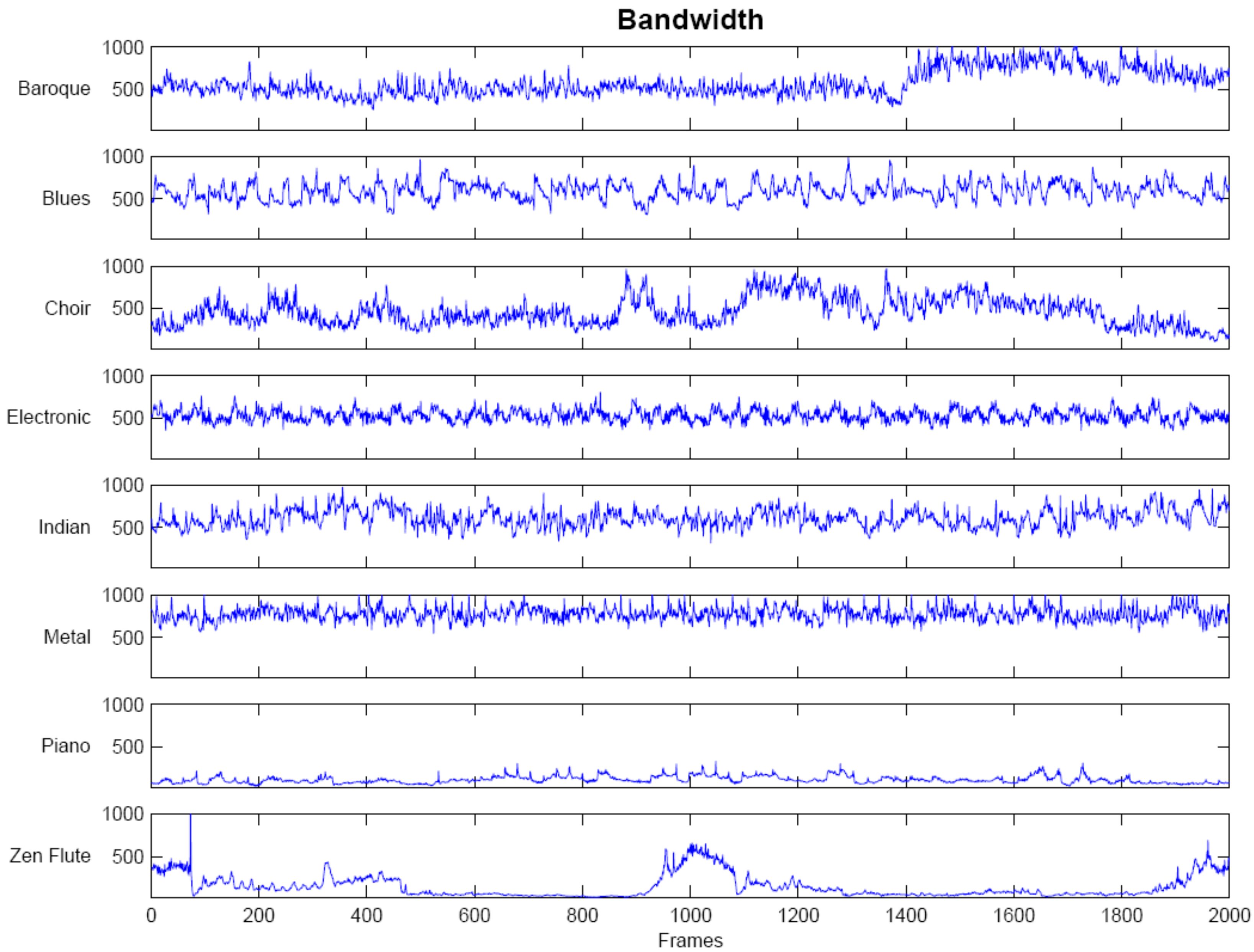
$M_t(n)$...magnitude in frequency domain at frame *t* and frequency bin *n*
N...number of highest frequency band
C_t...Spectral Centroid

Description: describes the spectral range of the interesting parts of the signal

Remarks:
+ average bandwidth of a piece of music may serve as indicator of aggressiveness
– no information about perceived rhythmic structure
– not suited to distinguish different parts of a piece of music
(cf. vocal part in metal piece not visible)

Department of Computational Perception

Bandwidth: Illustration



artment of
putational
eption

Low-Level Feature: Spectral Flux

(aka Delta Spectrum Magnitude)

Scope: frequency domain

Calculation:

$$F_t = \sum_{n=1}^{N} \left(N_t(n) - N_{t-1}(n)\right)^2$$

N_t...frame-by-frame normalized frequency distribution in frame *t*
N...number of highest frequency band

Description:
measures the rate of local spectral change, big spectral change from frame *t-1* to *t* → high F_t value

Remarks:

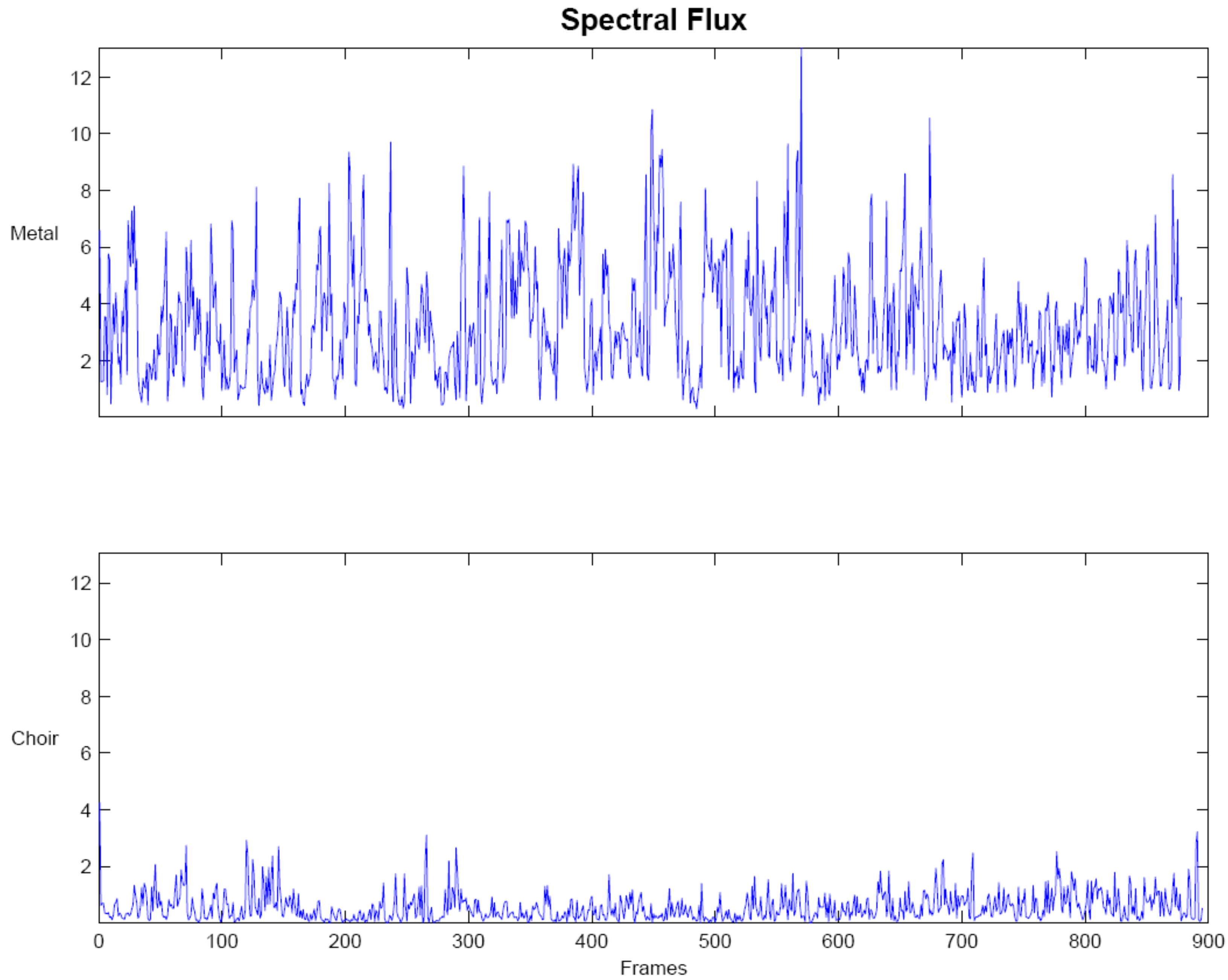
• commonly used as part of a low-level descriptor set

\+ may be used to distinguish between aggressive and calm music

\+ may serve as speech detector

Department of Computational Perception

Spectral Flux: Illustration



Advanced Music Content Analysis

Markus Schedl
Peter Knees

{markus.schedl, peter.knees}@jku.at

Department of Computational Perception
Johannes Kepler University (JKU)
Linz, Austria

Outline

Mid-level feature extraction and similarity calculation

- **Pitch Class Profiles**: related to Western music tone scale, melodic retrieval
- **MFCCs**: related to timbral properties
- **Block-Level Features**
  - Fluctuation Patterns: related to rhythmic/periodic properties
  - Correlation Patterns: temporal relation of frequencies
  - Spectral Contrast Patterns: related to "tone-ness"

Throughout: Examples and Applications

Department of Computational Perception

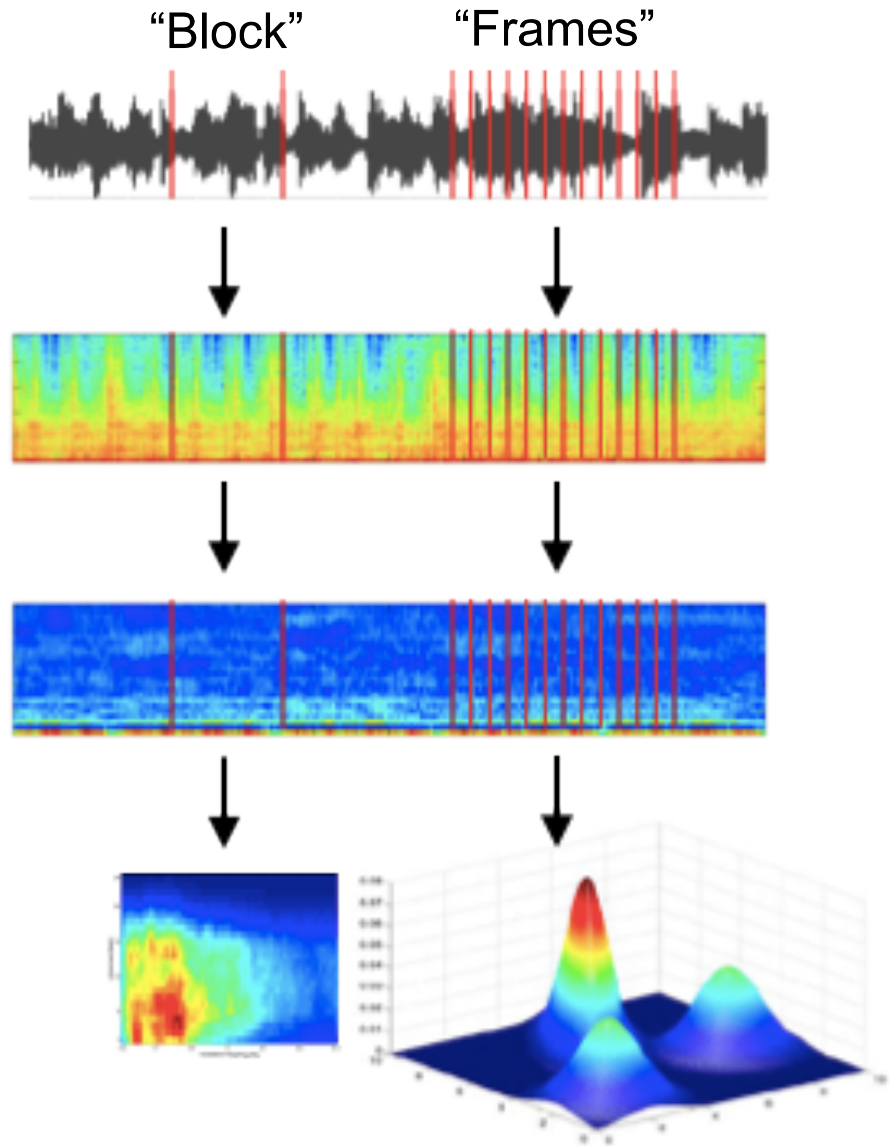
Mid-level Feature Processing Overview

Convert signal to *frequency domain*, e.g., using an FFT

(Psycho)acoustic transformation (Mel-scale, Bark-scale, Cent-scale, ...): mimics human listening process (not linear, but logarithmic!), removes aspects not perceived by humans, emphasizes low frequencies

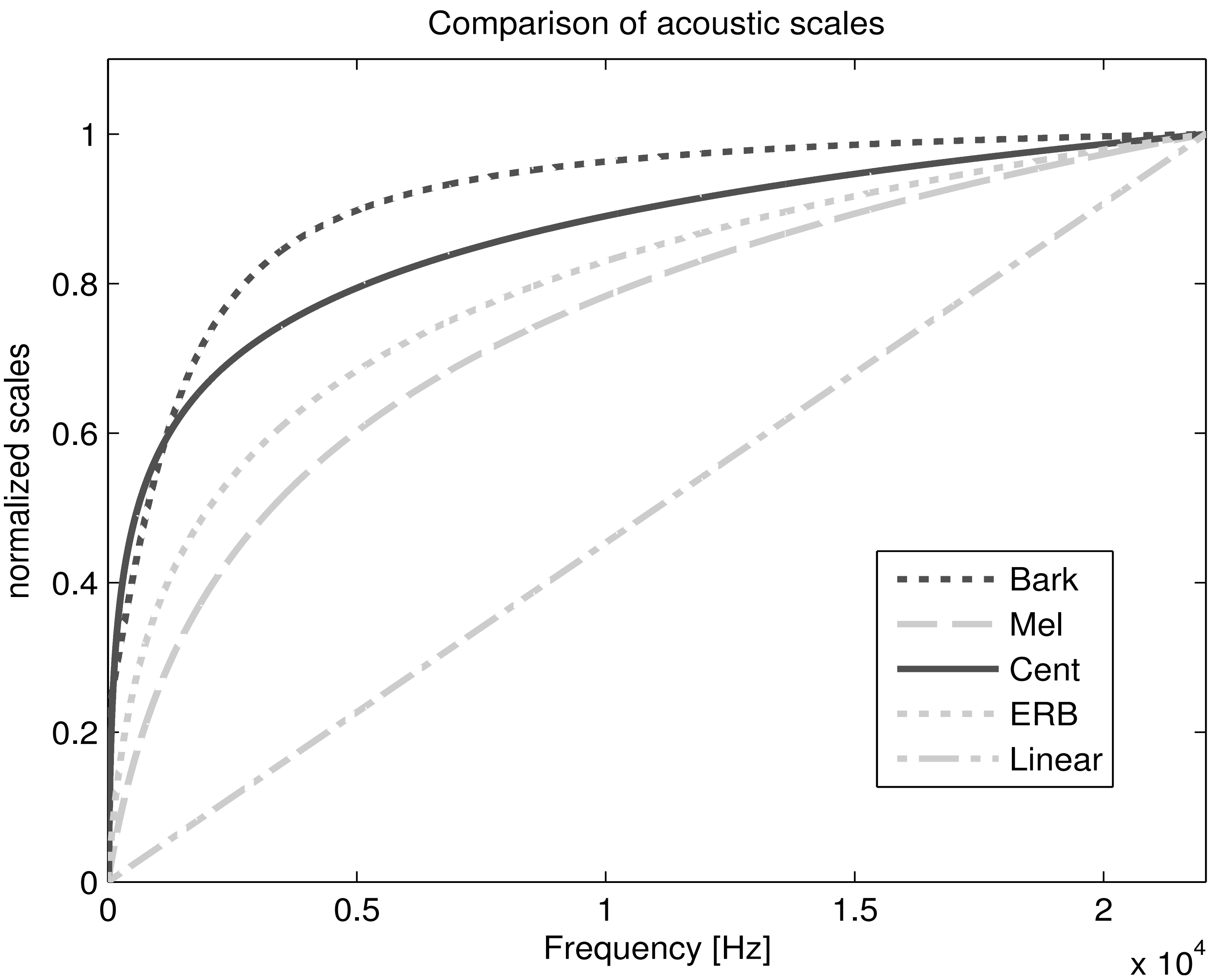
Extract features

- *Block-level*
  (large time windows, e.g., 6 sec)
- *Frame-level*
  (short time windows, e.g., 25 ms)
  needs feature distribution model



Department of
Computational
Perception

Acoustic Scales



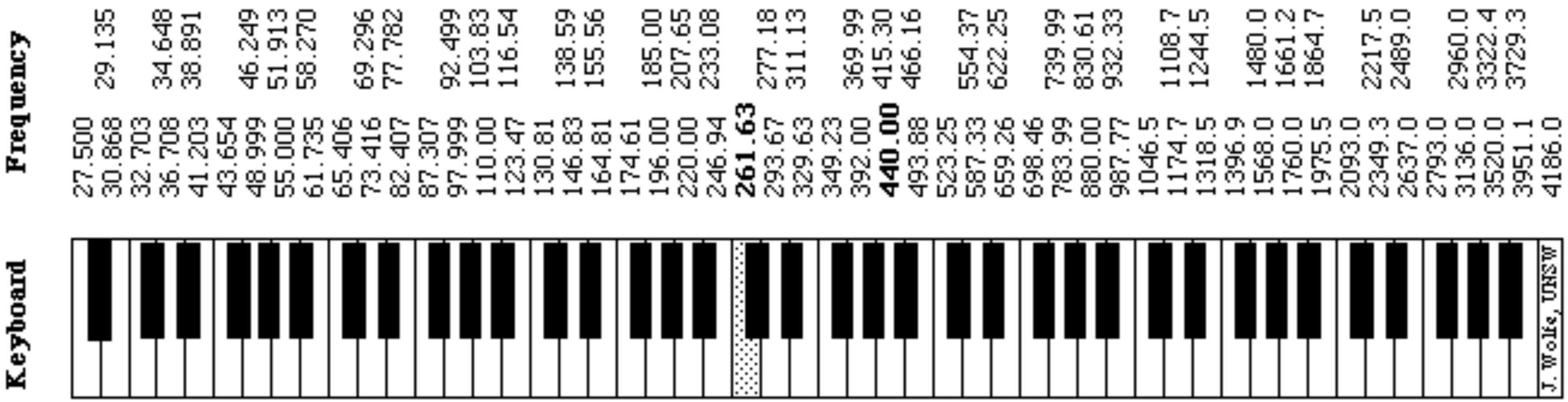
Department of
Computational
Perception

Pitch Class Profiles

(Fujishima; 1999)

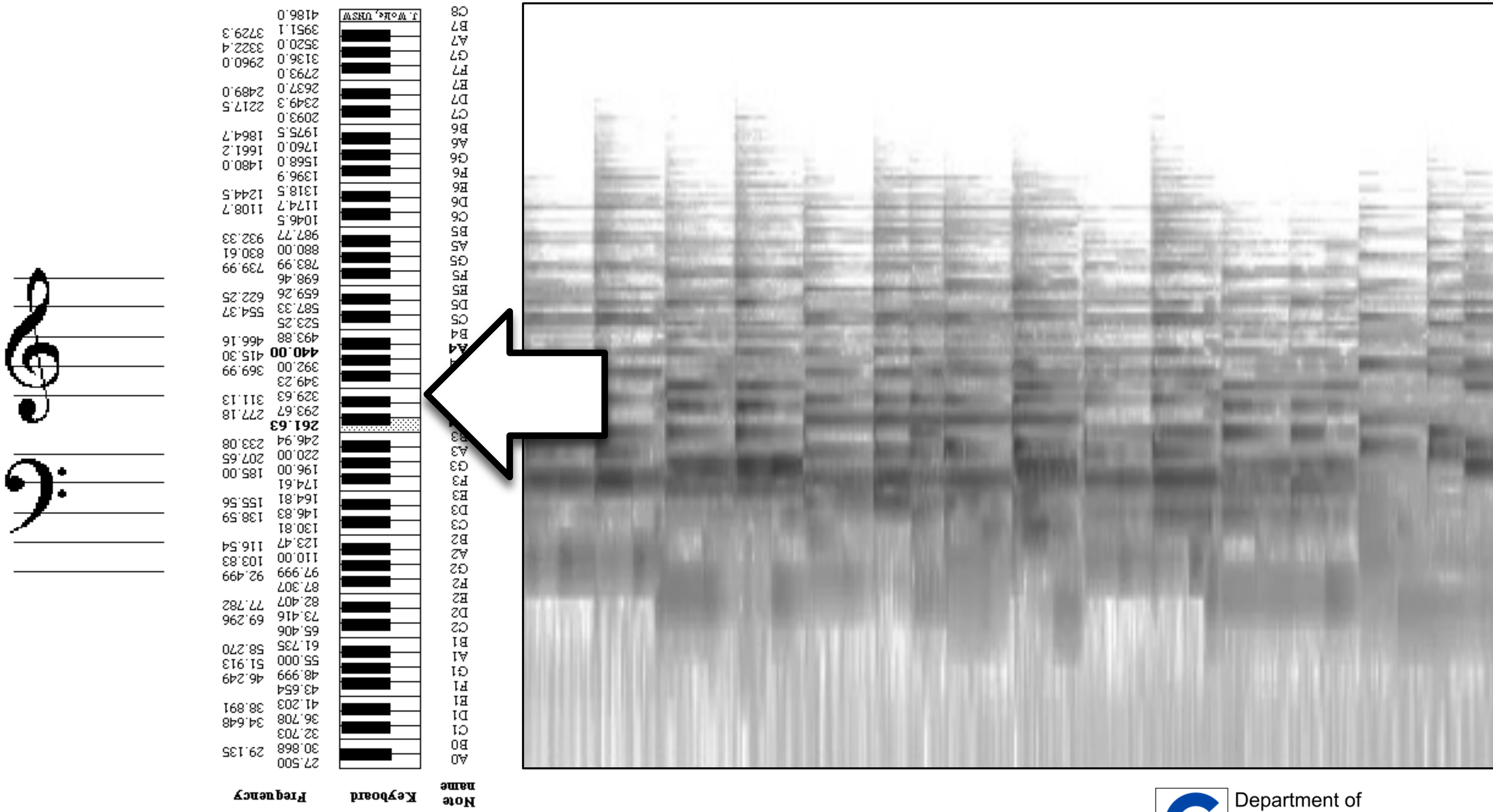
(aka *chroma vectors*)

- Transforming the frequency activations into well known musical system/representation/notation
- Mapping to the equal-tempered scale (each semitone equal to one twelfth of an octave)
- For each frame, get intensity of each of the 12 semitone (pitch) classes



Department of Computational Perception

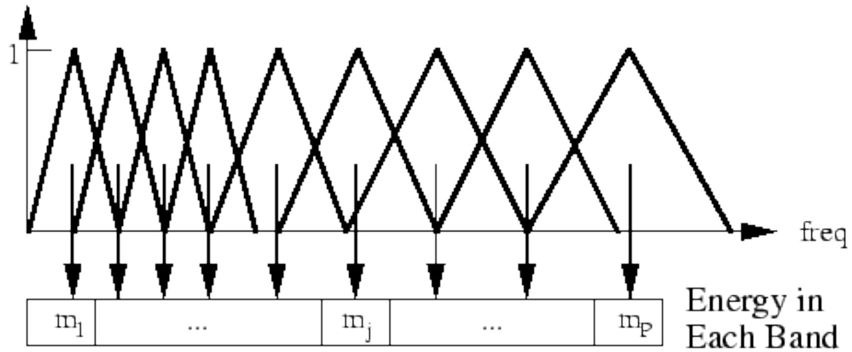
Mapping Frequencies to Semitones

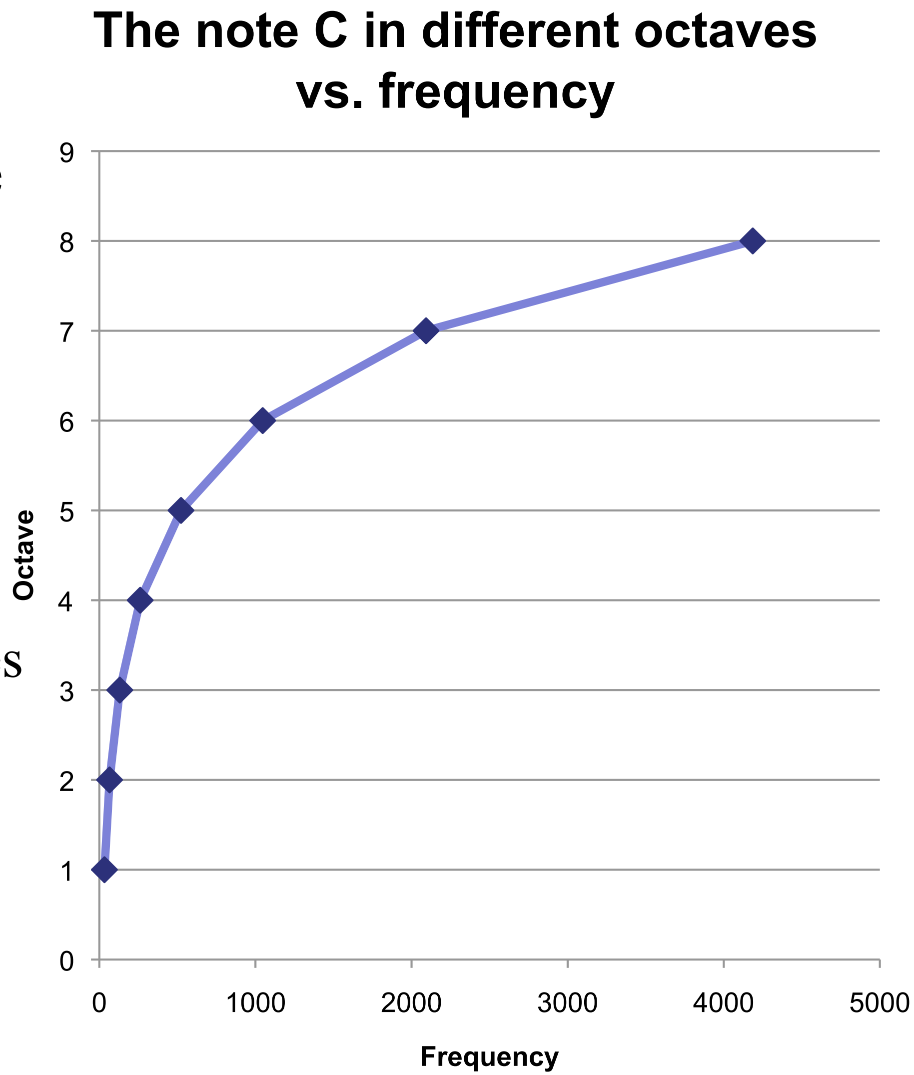


Department of
Computational
Perception

Semitone Scale

- Map data to semitone scale to represent (western) music
- Frequency doubles for each octave
  - e.g. pitch of A3 is 220 Hz, compared to 440 Hz of A4
- Mapping, e.g., using filter bank with triangular filters
  - centered on pitches
  - width given by neighboring pitches
  - normalized by area under filter

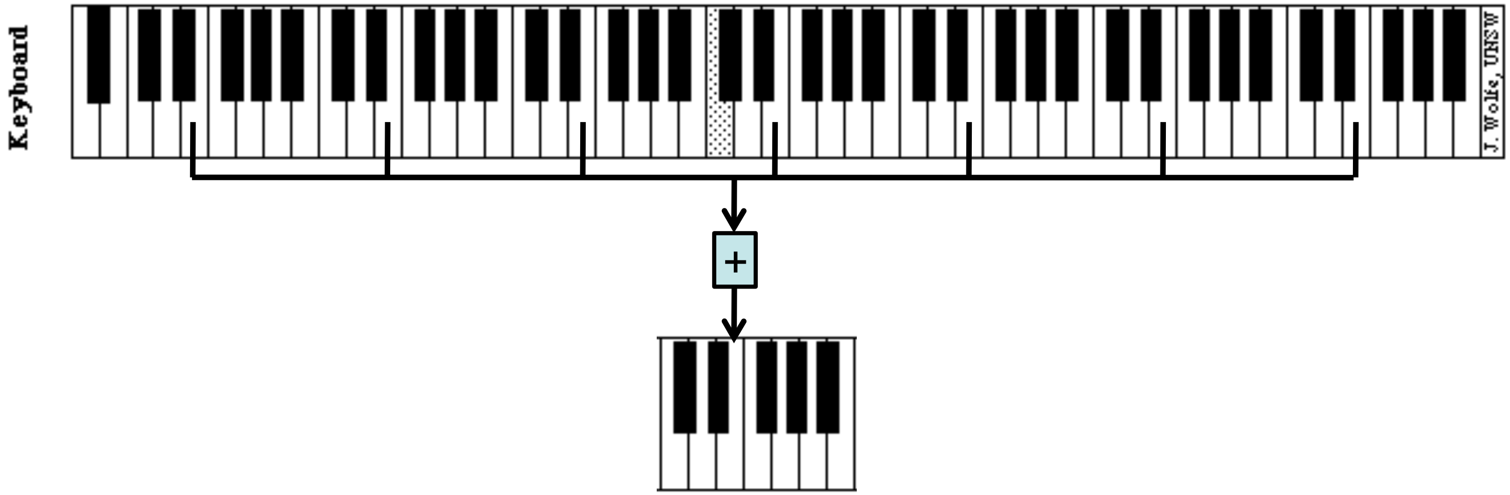




Department of Computational Perception

Pitch Class Features

Sum up activations that belong to the **same class of pitch** (e.g., all A, all C, all F#)



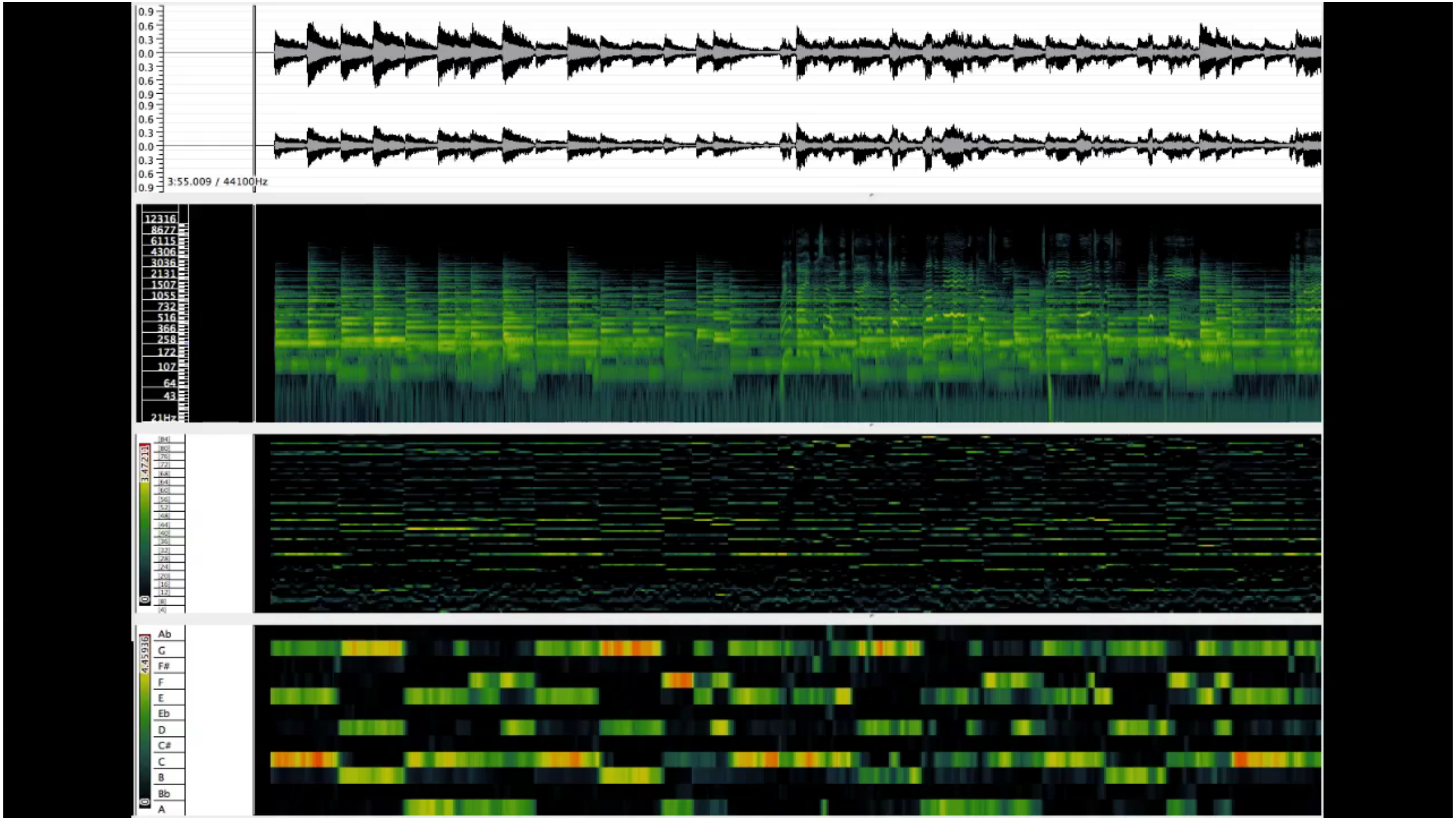
Results in a 12-dimensional feature vector for each frame

PCP feature vectors describe tonality

- Robust to noise (including percussive sounds)
- Independent of timbre (~ played instruments)
- Independent of loudness

Department of Computational Perception

Pitch Class Profiles in Action

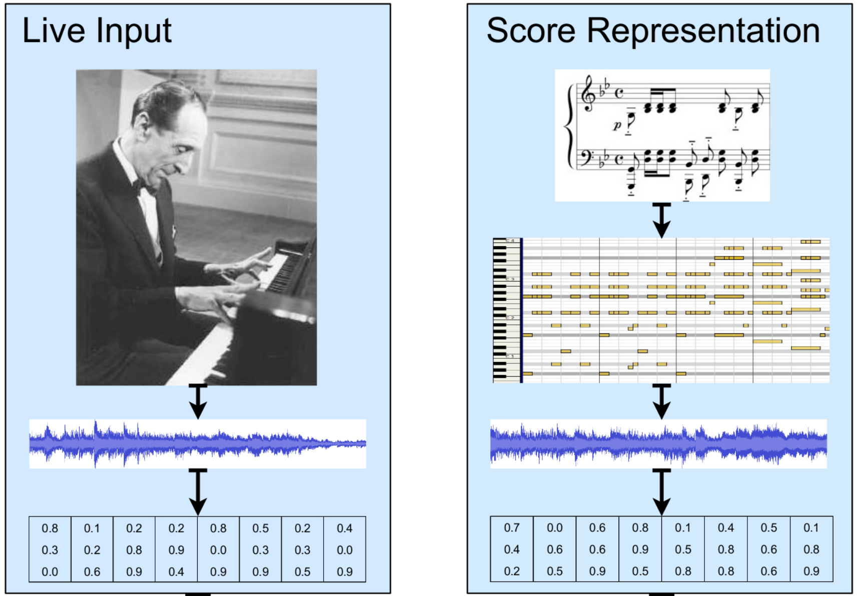


Sonic Visualizer by QMUL, C4DM; http://www.sonicvisualiser.org

Department of Computational Perception

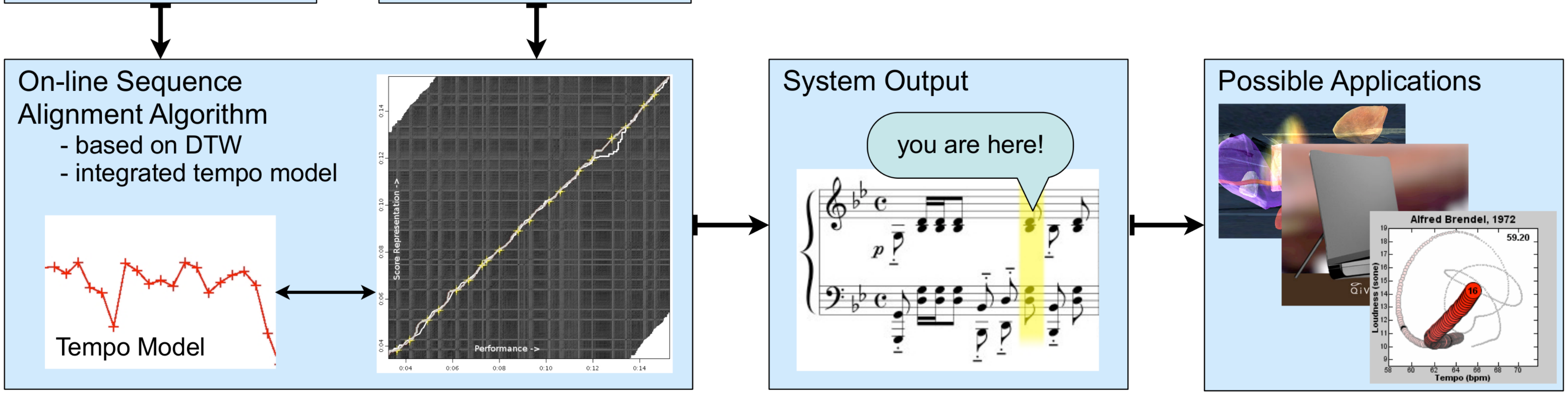
Real-Time Score Following

(Arzt, Widmer; 2010)



Tracks the position of a piano player in the score while playing

- Uses a combination of spectral flux and PCPs as features
- Dynamic Time Warping (DTW) to match recorded live performance with dead-pan synthesized version



Department of Computational Perception

(Arzt, Widmer; 2010)

Application: Automatic Page Turner

Department of
Computational
Perception