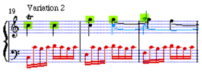


Tutorial

A Music-oriented Approach to Music Signal Processing

Meinard Müller

Saarland University and MPI Informatik
meinard@mpi-inf.mpg.de

Anssi Klapuri

Queen Mary University of London
anssi.klapuri@elec.qmul.ac.uk

Overview

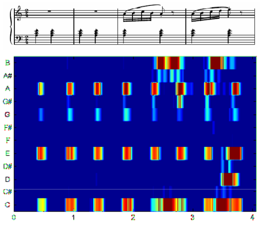
Part I: Pitch and Harmony

Part II: Tempo and Beat

Coffee Break

Part III: Timbre

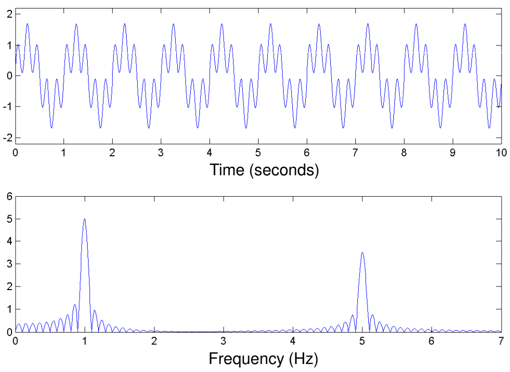
Part IV: Melody



Fourier Transform

f

$|\hat{f}|$



Fourier Transform

Signal $\quad f : \mathbb{R} \to \mathbb{R}$

Fourier representation $\quad f(t) = \int_{\omega \in \mathbb{R}} c_\omega e^{2\pi i \omega t} d\omega \ , \ c_\omega = \hat{f}(\omega)$

Fourier transform $\quad \hat{f}(\omega) = \int_{t \in \mathbb{R}} f(t) e^{-2\pi i \omega t} dt$

Fourier Transform

Signal $\quad f : \mathbb{R} \to \mathbb{R}$

Fourier representation $\quad f(t) = \int_{\omega \in \mathbb{R}} c_\omega e^{2\pi i \omega t} d\omega \ , \ c_\omega = \hat{f}(\omega)$

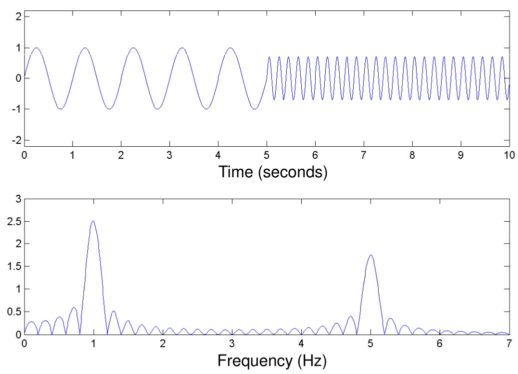
Fourier transform $\quad \hat{f}(\omega) = \int_{t \in \mathbb{R}} f(t) e^{-2\pi i \omega t} dt$

- Tells which frequencies occur, but does not tell when the frequencies occur
- Frequency information is averaged over the entire time interval
- Time information is hidden in the phase

Fourier Transform

f

$|\hat{f}|$



Short Time Fourier Transform

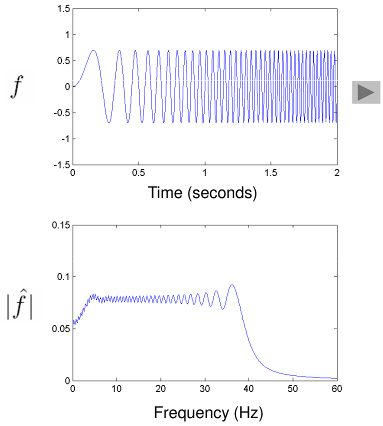
Idea (Dennis Gabor, 1946):

- Consider only a small section of the signal for the spectral analysis

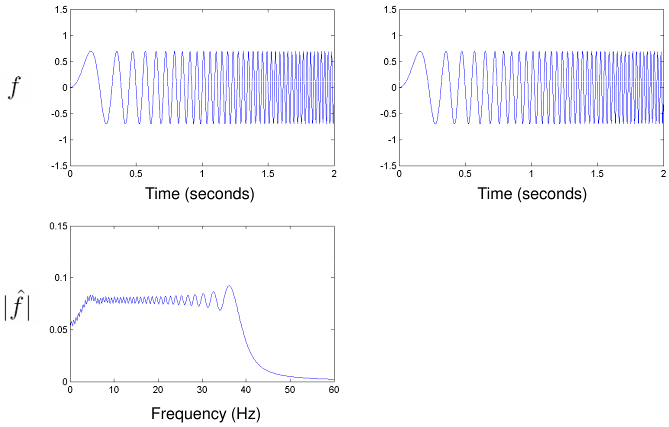
  → recovery of time information

- Short Time Fourier Transform (STFT)
- Section is determined by pointwise multiplication of the signal with a localizing window function

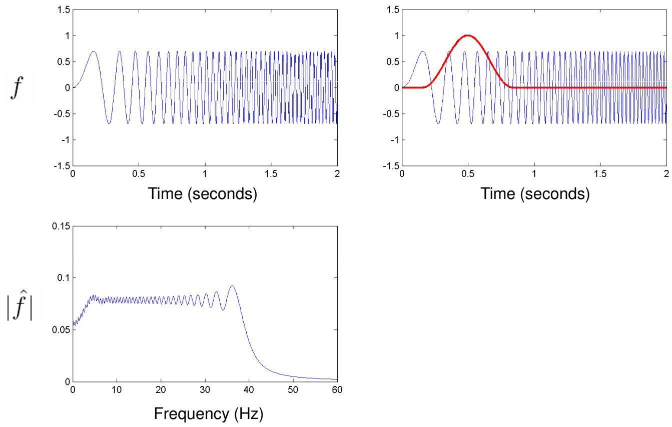
Short Time Fourier Transform



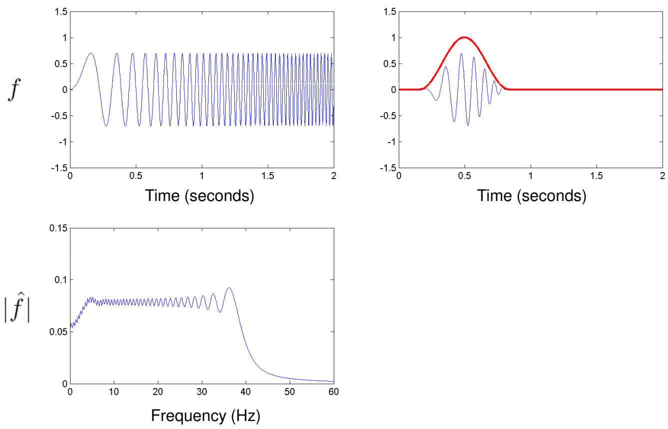
Short Time Fourier Transform



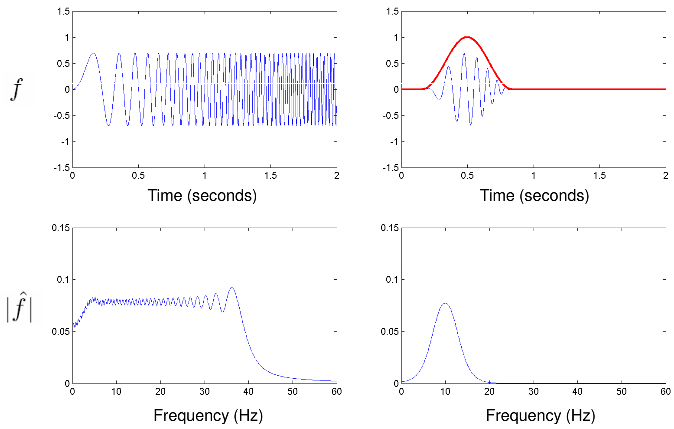
Short Time Fourier Transform



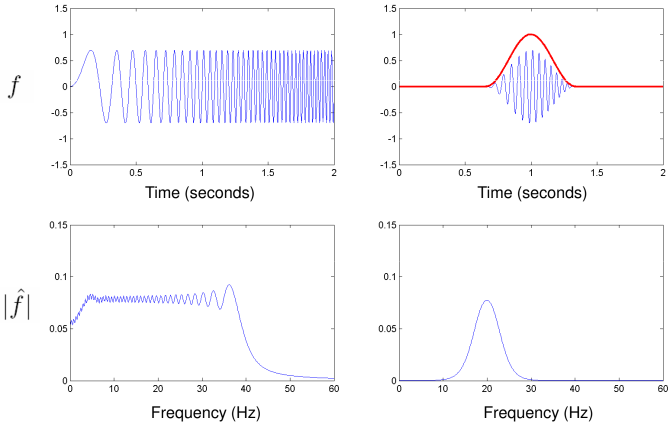
Short Time Fourier Transform



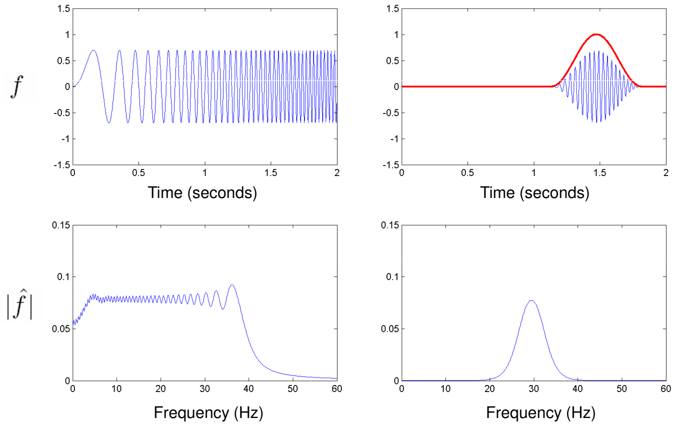
Short Time Fourier Transform



Short Time Fourier Transform



Short Time Fourier Transform



Short Time Fourier Transform

Definition

- Signal $f : \mathbb{R} \to \mathbb{R}$
- Window function $g : \mathbb{R} \to \mathbb{R}$ $(g \in L^2(\mathbb{R})\ , \|g\| = 1\ )$
- STFT $\tilde{f}(\omega, t) := \int_{\mathbb{R}} f(u)\bar{g}(u-t)e^{-2\pi i \omega u} du = \langle f | g_{\omega,t} \rangle$

  with $g_{\omega,t}(u) := e^{2\pi i \omega u} g(u-t), \quad u \in \mathbb{R}$

Short Time Fourier Transform

Intuition:

- $g_{\omega,t}$ is ``sound event´´ of frequency ω , which oscillates within the translated window $u \to g(u-t)$

Short Time Fourier Transform

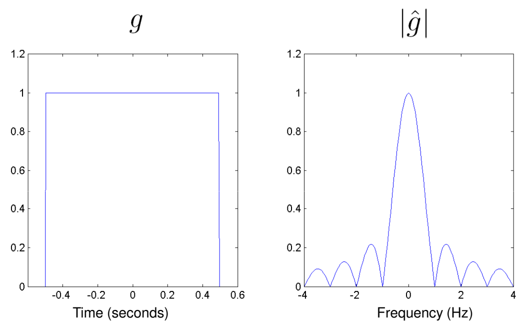
Intuition:

- $g_{\omega,t}$ is ``sound event´´ of frequency ω , which oscillates within the translated window $u \to g(u-t)$

- Inner product $\langle f | g_{\omega,t} \rangle$ measures the correlation between the sound event $g_{\omega,t}$ and the signal f

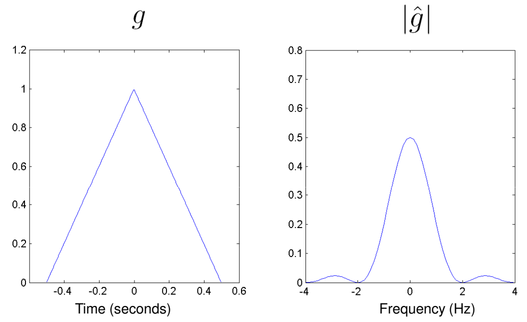
Window Function

Box window



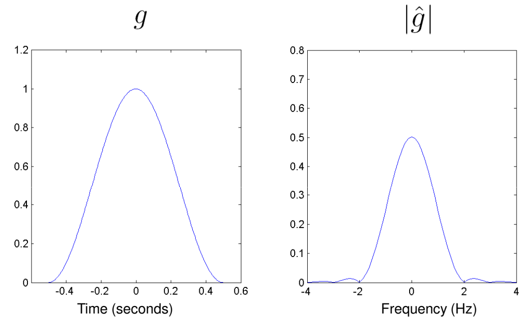
Window Function

Triangle window

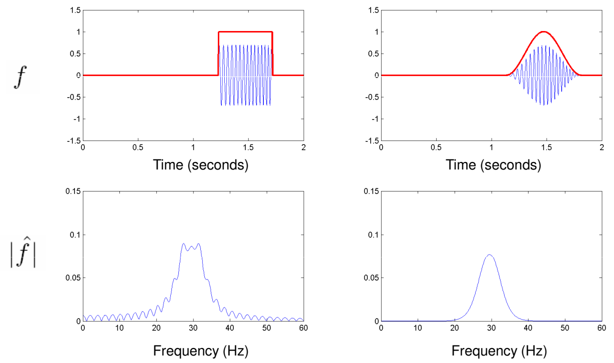


Window Function

Hann window

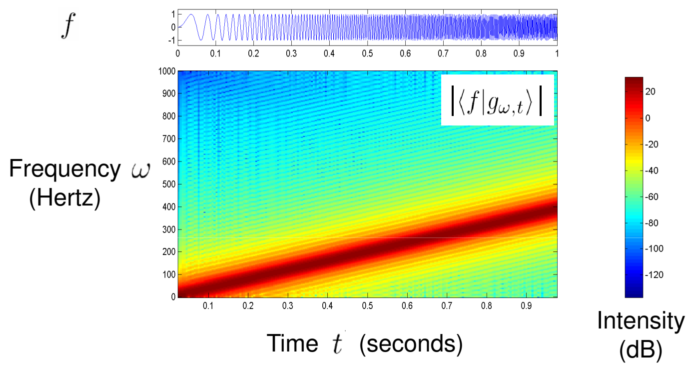


Window Function

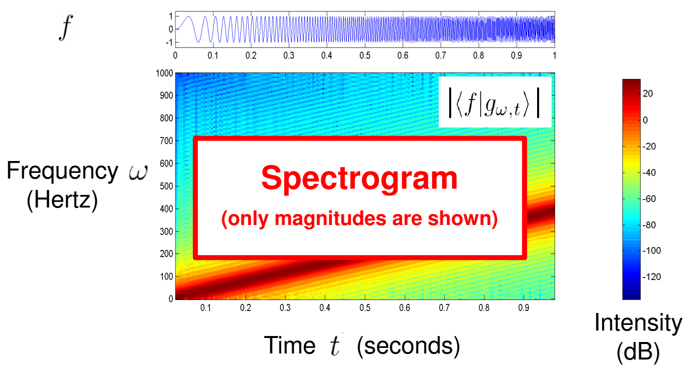


Trade off between smoothing and „ringing“

Time-Frequency Representation

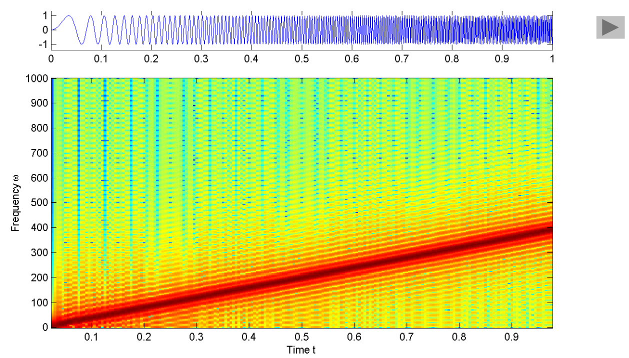


Time-Frequency Representation



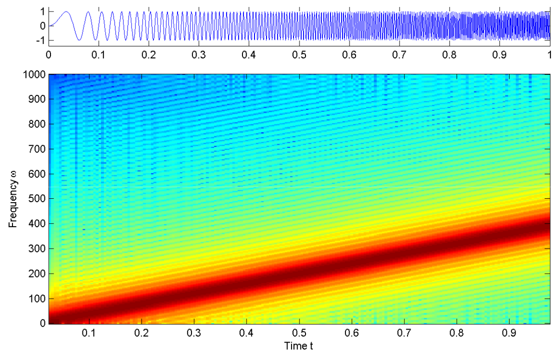
Time-Frequency Representation

Chirp signal and STFT with box window of length 0.05



Time-Frequency Representation

Chirp signal and STFT with Hann window of length 0.05



Time-Frequency Localization

- Size of window constitutes a trade-off between time resolution and frequency resolution:

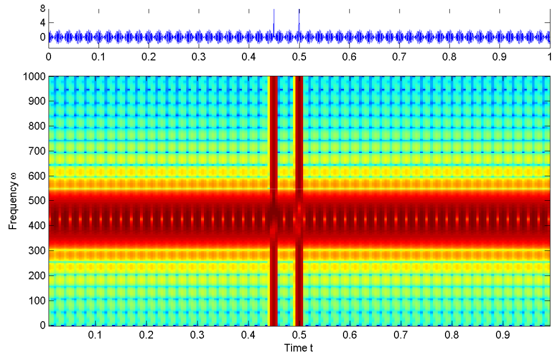
  Large window : poor time resolution
  good frequency resolution

  Small window : good time resolution
  poor frequency resolution

- Heisenberg Uncertainty Principle: there is no window function that localizes in time and frequency with arbitrary precision.

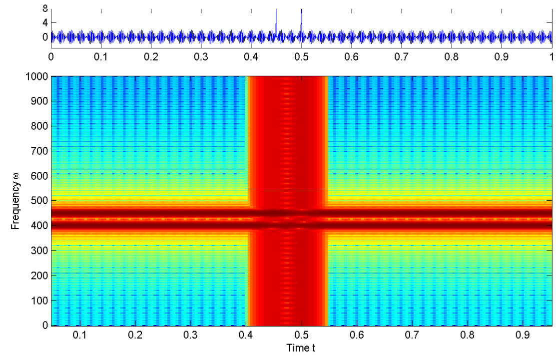
Short Time Fourier Transform

Signal and STFT with Hann window of length 0.02



Short Time Fourier Transform

Signal and STFT with Hann window of length 0.1



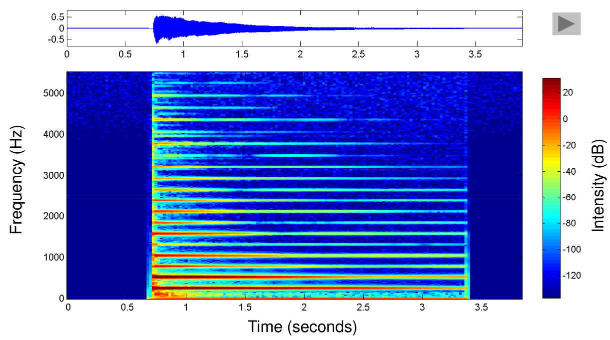
Short Time Fourier Transform

Note: Playing a single note on an instrument may result in a complex superposition of different frequencies.

Pitch and frequency are two different concepts!

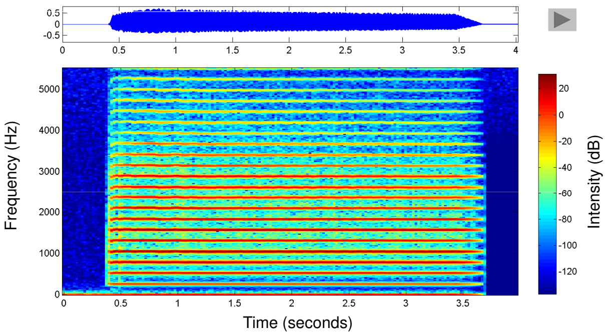
Short Time Fourier Transform

Example: Piano



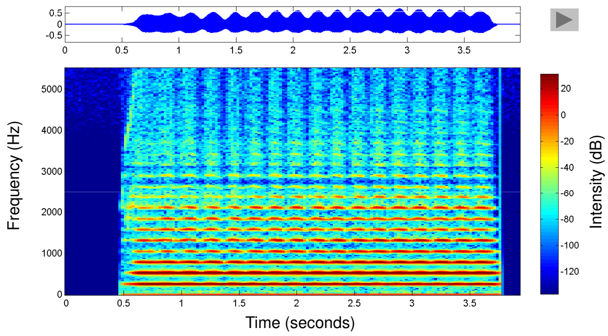
Short Time Fourier Transform

Example: Trumpet



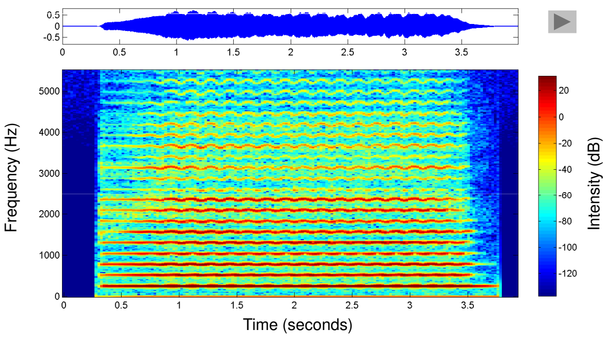
Short Time Fourier Transform

Example: Flute

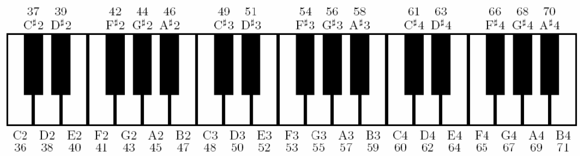


Short Time Fourier Transform

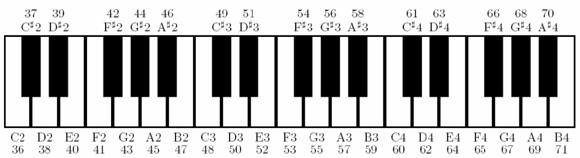
Example: Violine



Pitch Features



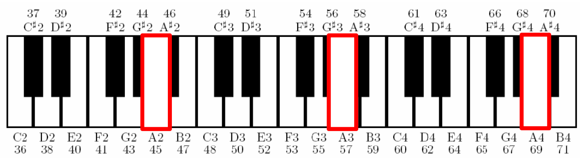
Pitch Features



Model assumption: Equal-tempered scale

- MIDI pitches: $p \in [1:128]$
- Piano notes: $p = 21\ \text{(A0)}\ \ \text{to}\ \ p = 108\ \text{(C8)}$
- Concert pitch: $p = 69\ \text{(A4)} = 440\ \text{Hz}$
- Center frequency: $f_{\mathrm{MIDI}}(p) = 2^{\frac{p-69}{12}} \cdot 440$ Hz

Pitch Features



A2 110 Hz

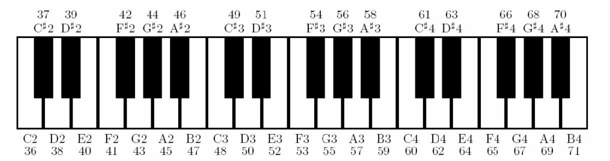
A3 220 Hz

A4 440 Hz

Logarithmic frequency distribution

Octave: doubling of frequency

Pitch Features

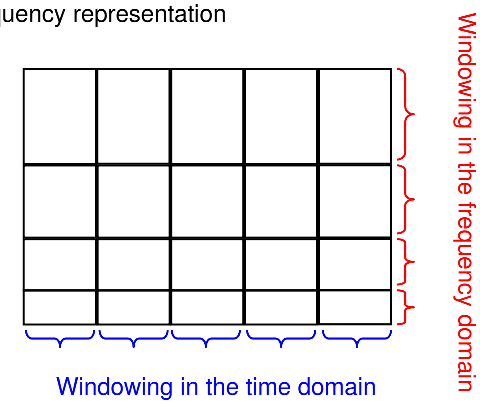


Idea: Binning of Fourier coefficients

Divide up the fequency axis into logarithmically spaced „pitch regions" and combine spectral coefficients of each region to form a single pitch coefficient.

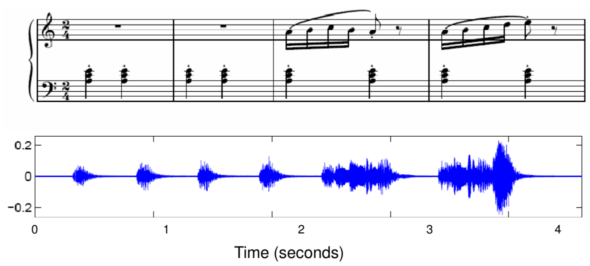
Pitch Features

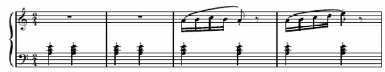
Time-frequency representation

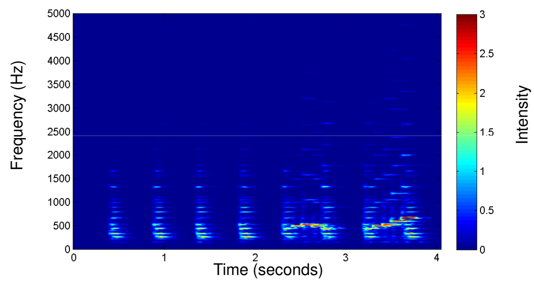


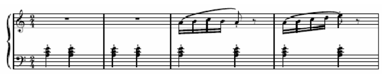
Pitch Features

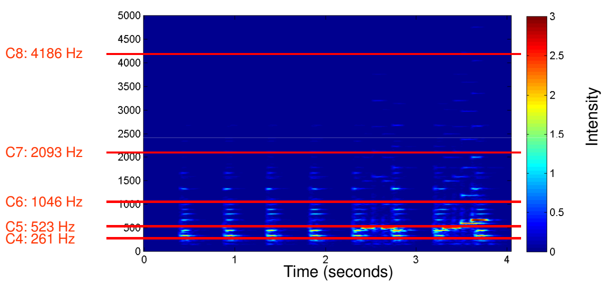
Example: Friedrich Burgmüller, Op. 100, No. 2

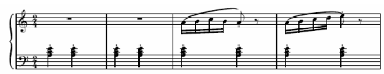


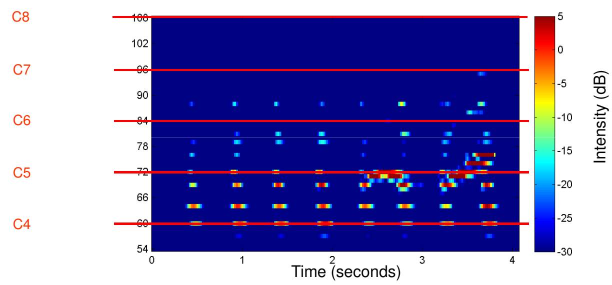
Pitch Features

Spectrogram

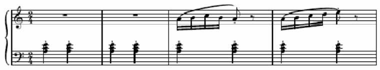


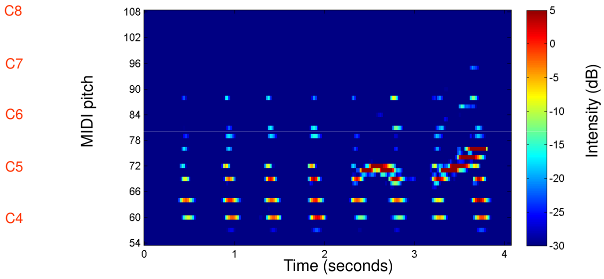
Pitch Features

Spectrogram



Pitch Features

Pitch representation



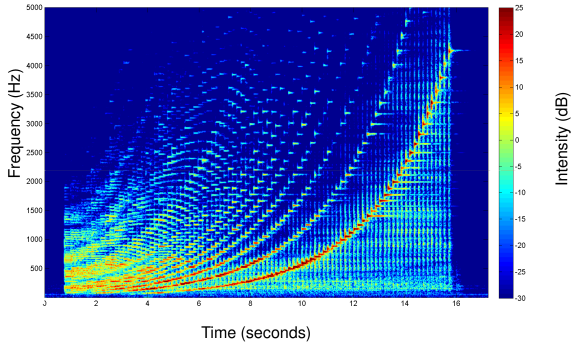
Pitch Features

Pitch representation



Pitch Features

Example: Chromatic Scale

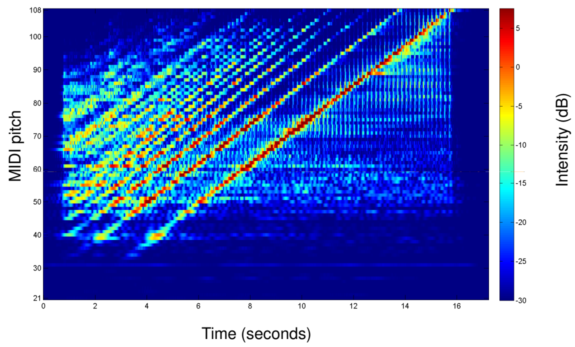
Spectrogram



Pitch Features

Example: Chromatic Scale

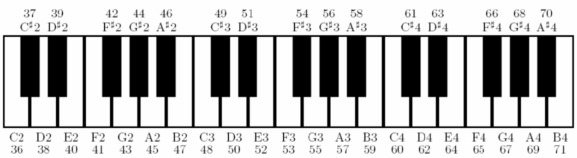
Pitch representation



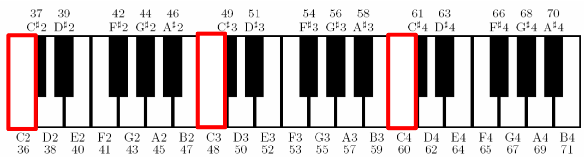
Chroma Features

- Pitches are perceived as related (harmonically similar) if they differ by an octave
- Idea: through away information which is difficult to estimate and not so important for harmonic analysis
- Separation of pitch into two components: tone height (octave number) and chroma
- Chroma: 12 traditional pitch classes of the equal-tempered scale. For example:
  Chroma C $\widehat{=} \{\ldots , \mathrm{C0} , \mathrm{C1} , \mathrm{C2} , \mathrm{C3} , \ldots\}$
- Computation: pitch features → chroma features
  Add up all pitches belonging to the same class
- Result: 12-dimensional chroma vector

Chroma Features



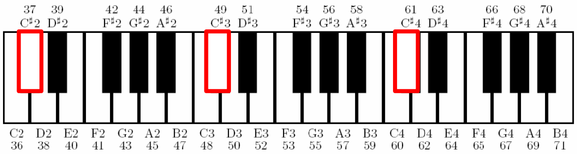
Chroma Features



C2 C3 C4

Chroma C

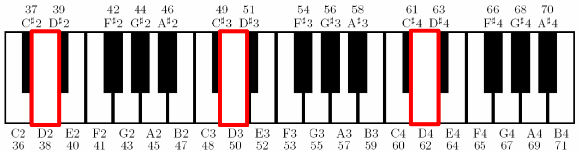
Chroma Features



C#2 C#3 C#4

Chroma C#

Chroma Features

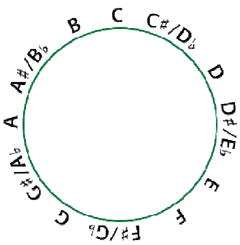


D2 D3 D4

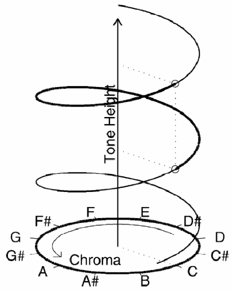
Chroma D

Chroma Features

Chromatic circle



Shepard's helix of pitch perception



http://en.wikipedia.org/wiki/Pitch_class_space

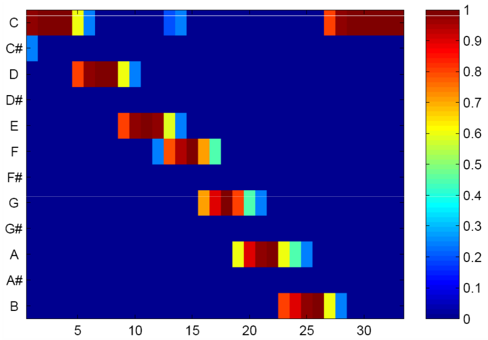
[Gómez, PhD 2006][Bartsch/Wakefield, IEEE-TMM 2005]

Chroma Features

- Sequence of chroma vectors correlates to the harmonic progression
- Normalization $v \rightarrow \frac{v}{\|v\|}$ makes features invariant to changes in dynamics
- Further quantization and smoothing
- Taking logarithm before adding up pitch coefficients accounts for logarithmic sensation of intensity

Chroma Features

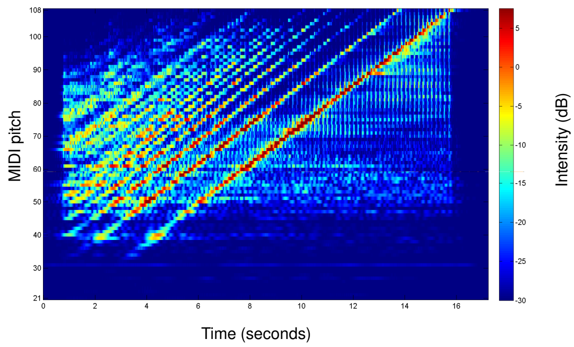
Example: C-Major Scale



Chroma Features

Example: Chromatic Scale

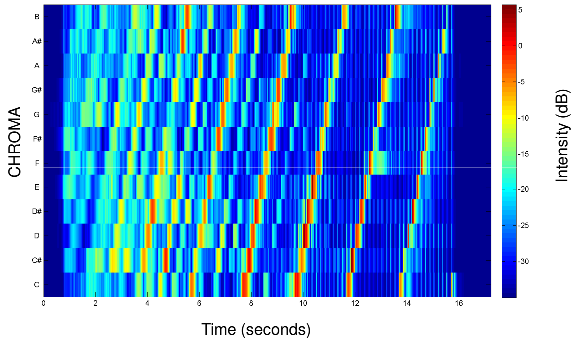
Pitch representation



Chroma Features

Example: Chromatic Scale

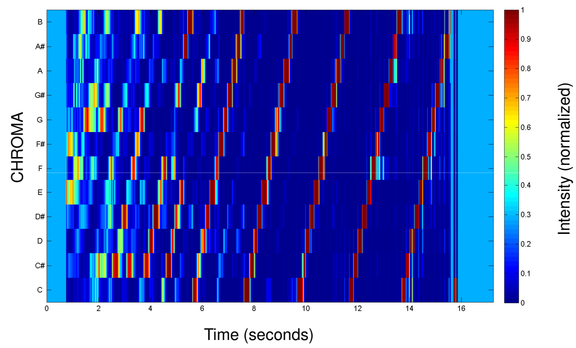
Chroma representation



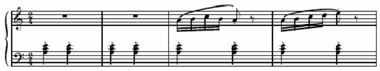
Chroma Features

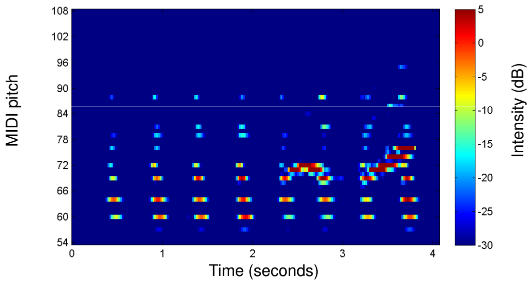
Example: Chromatic Scale

Chroma representation (normalized)

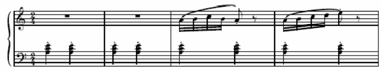


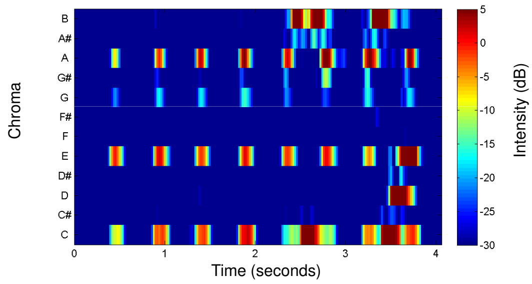
Chroma Features

Example: Friedrich Burgmüller, Op. 100, No. 2

Pitch representation

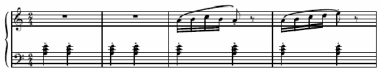


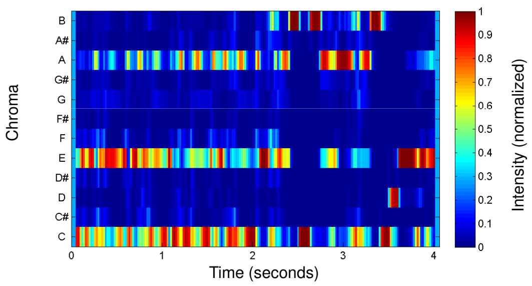
Chroma Features

Example: Friedrich Burgmüller, Op. 100, No. 2

Chroma representation



Chroma Features

Example: Friedrich Burgmüller, Op. 100, No. 2

Chroma representation (normalized)



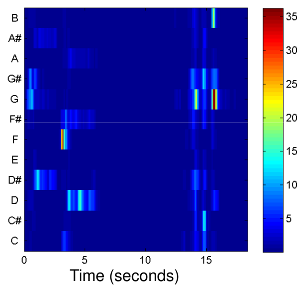
Chroma Features

Example: Beethoven's Fifth

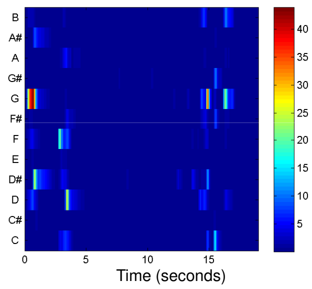
Chroma representation

Feature resolution: 10 Hz

Karajan



Scherbakov



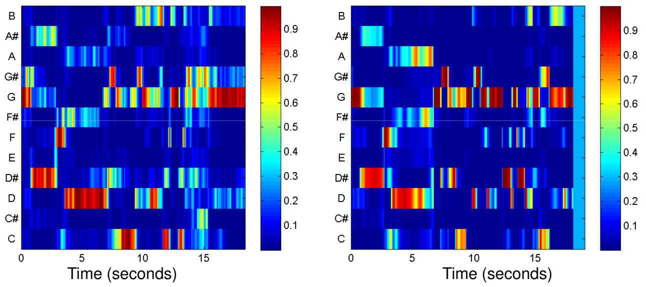
Chroma Features

Example: Beethoven‘s Fifth

Chroma representation (normalized)

Feature resolution: 10 Hz

Karajan Scherbakov



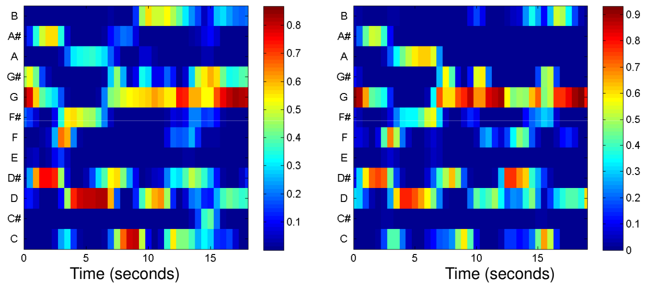
Chroma Features

Example: Beethoven‘s Fifth

Chroma representation (normalized)

Feature resolution: 2 Hz

Karajan Scherbakov



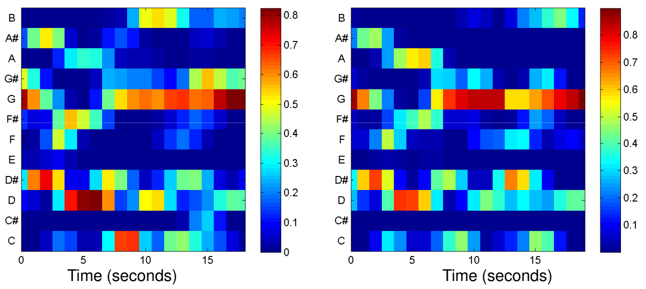
Chroma Features

Example: Beethoven‘s Fifth

Chroma representation (normalized)

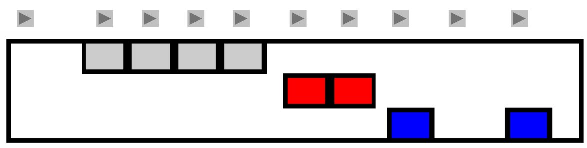
Feature resolution: 1 Hz

Karajan Scherbakov



Chroma Features

Example: Zager & Evans “In The Year 2525”

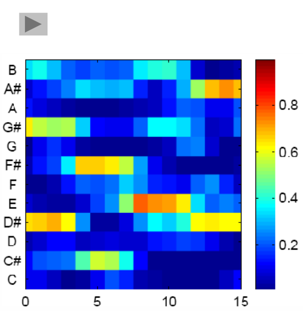


How to deal with transpositions?

[Goto, ICASSP 2003]

Chroma Features

Example: Zager & Evans “In The Year 2525”

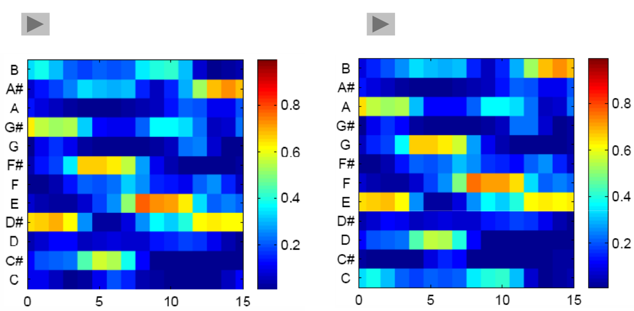


Original: $(v^1, \ldots, v^N)$

[Goto, ICASSP 2003]

Chroma Features

Example: Zager & Evans “In The Year 2525”



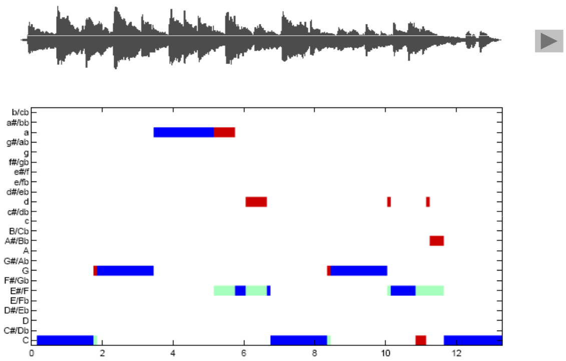
Original: $(v^1, \ldots, v^N)$ Shifted: $(\sigma(v^1), \ldots, \sigma(v^N))$

[Goto, ICASSP 2003]

Application: Chord Recognition

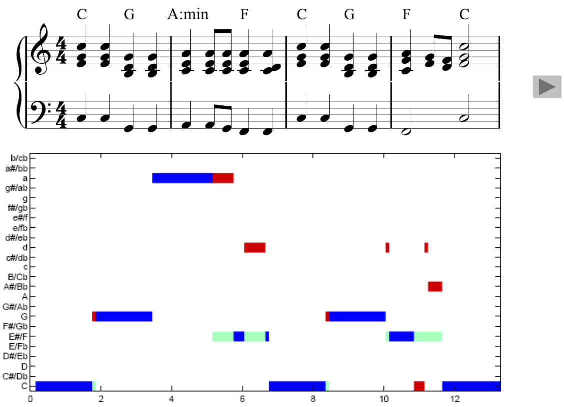
[Ueada et al., ICASSP 2010][Sheh/Ellis, ISMIR 2003]

Application: Chord Recognition



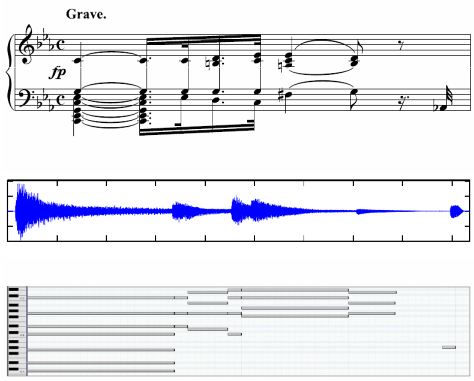
[Ueada et al., ICASSP 2010][Sheh/Ellis, ISMIR 2003]

Application: Chord Recognition



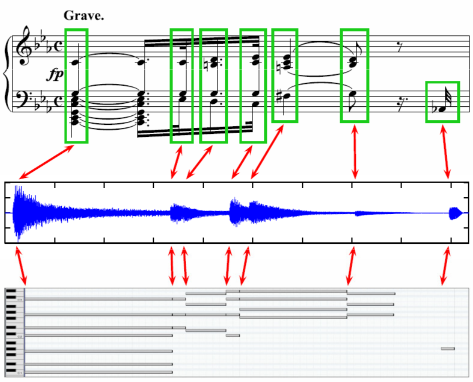
[Ueada et al., ICASSP 2010][Sheh/Ellis, ISMIR 2003]

Application: Music Synchronization



[Müller, Springer-Monograph 2007]

Application: Music Synchronization



[Müller, Springer-Monograph 2007]

Application: Music Synchronization

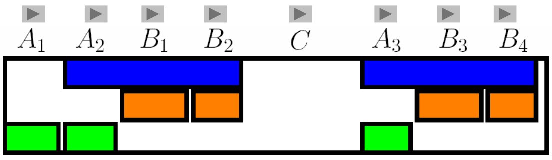
System: Interpretation Switcher (Beethoven-Haus)

Application: Audio Structure Analysis

Given: CD recording

Goal: Automatic extraction of the repetitive structure
(or of the musical form)

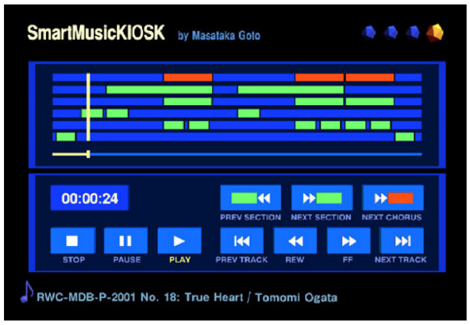
Example: Brahms Hungarian Dance No. 5 (Ormandy)



[Dannenberg/Hu, ISMIR 2002]

Application: Audio Structure Analysis

System: SmartMusicKiosk



[Goto, ICASSP 2003]

Application: Cover Song Identification

Goal: Given a music recording of a song or piece of music, find all corresponding music recordings within a huge collection that can be regarded as a kind of version, interpretation, or cover song.

- Live versions
- Versions adapted to particular country/region/language
- Contemporary versions of an old song
- Radically different interpretations of a musical piece

Instance of document-based retrieval!

[Ellis/Poliner, ICASSP 2007] [Serrà et al., IEEE-TASLP 2009]

Application: Cover Song Identification

Query: Bob Dylan – Knockin' on Heaven's Door

Retrieval result:

Rank	Recording	Score
1.	Guns and Roses: Knockin' On Heaven's Door	94.2
2.	Avril Lavigne: Knockin' On Heaven's Door	86.6
3.	Wyclef Jean: Knockin' On Heaven's Door	83.8
4.	Bob Dylan: Not For You	65.4
5.	Guns and Roses: Patience	61.8
6.	Bob Dylan: Like A Rolling Stone	57.2
7.-14.	…	

[Ellis/Poliner, ICASSP 2007] [Serrà et al., IEEE-TASLP 2009]

Application: Audio Matching

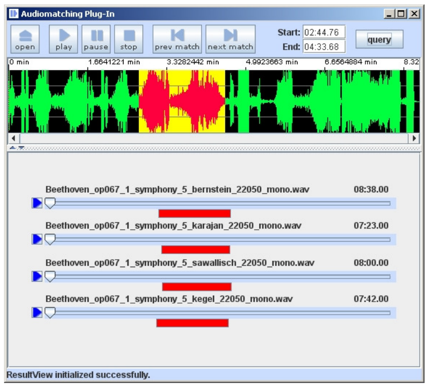
Given: Large music database containing several
- recordings of the same piece of music
- interpretations by various musicians
- arrangements in different instrumentations

Goal: Given a short query audio clip, identify all corresponding audio clips of similar musical content
- irrespective of the specific interpretation and instrumentation
- automatically and efficiently

Query-by-Example paradigm

[Müller et al., ISMIR 2005]

Application: Audio Matching



[Kurth/Müller, IEEE-TASLP 2008]

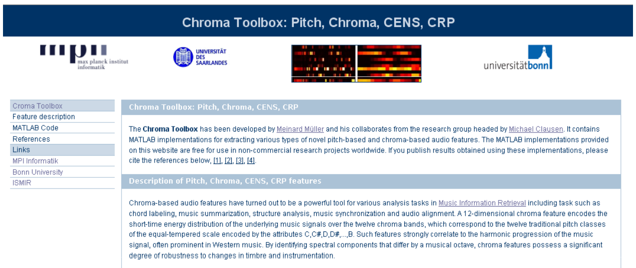
Application: Audio Matching

[Damm et al., ICMI 2008]

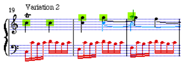
Conclusions (Chroma Features)

- Chroma features capture harmonic information
- High robustness to changes in timbre and instrumentation
- Many chroma variants with different properties
- Various implementations publically available

Chroma Toolbox



- Freely available Matlab toolbox
- Feature types: Pitch, Chroma, CENS, CRP
- http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/



Tutorial

A Music-oriented Approach to Music Signal Processing

Meinard Müller

Saarland University and MPI Informatik
meinard@mpi-inf.mpg.de

Anssi Klapuri

Queen Mary University of London
anssi.klapuri@elec.qmul.ac.uk

Overview

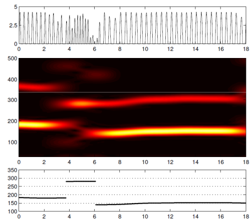
Part I: Pitch and Harmony

Part II: Tempo and Beat

Coffee Break

Part III: Timbre

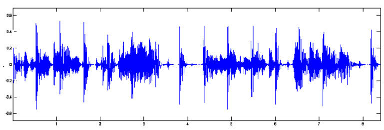
Part IV: Melody



Introduction (Tempo and Beat)

Basic Task: Given a recording of a piece of music, determine the periodic sequence of beat positions.

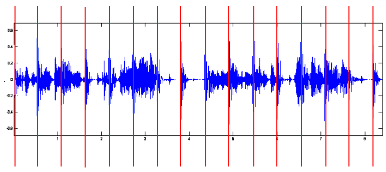
Tapping the foot when listening to music.



Introduction (Tempo and Beat)

Basic Task: Given a recording of a piece of music, determine the periodic sequence of beat positions.

Tapping the foot when listening to music.



Introduction (Tempo and Beat)

Example 1: Queen – Another One Bites The Dust

Pulse level: Quarter note

Tempo: 110 BPM

Introduction (Tempo and Beat)

Example 1: Queen – Another One Bites The Dust

Pulse level: Eighth note

Tempo: 220 BPM

Introduction (Tempo and Beat)

Example 2: Chopin – Mazurka Op. 68-3

Pulse level: Quarter note

Tempo: ???

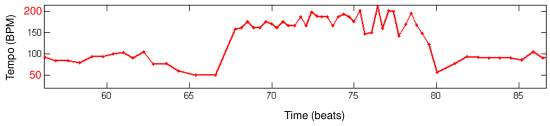
Introduction (Tempo and Beat)

Example 2: Chopin – Mazurka Op. 68-3

Pulse level: Quarter note

Tempo: 50-200 BPM

Tempo curve



Introduction (Tempo and Beat)

Example 2: Borodin – String Quartet No. 2

Pulse level: Quarter note

Tempo: 120-140 BPM (roughly)

Introduction (Tempo and Beat)

Tasks

- Onset detection
- Beat tracking
- Tempo estimation

Introduction (Tempo and Beat)

Tasks

- Onset detection
- Beat tracking
- Tempo estimation

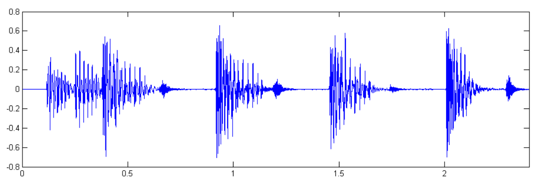
Challenges

- Non-percussive music
- Soft note onsets
- Time-varying tempo

Introduction (Tempo and Beat)

Tasks

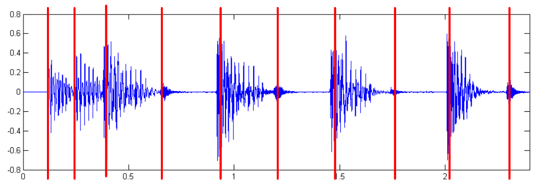
- Onset detection
- Beat tracking
- Tempo estimation



Introduction (Tempo and Beat)

Tasks

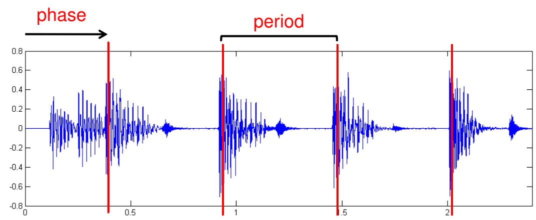
- Onset detection
- Beat tracking
- Tempo estimation



Introduction (Tempo and Beat)

Tasks

- Onset detection
- Beat tracking
- Tempo estimation



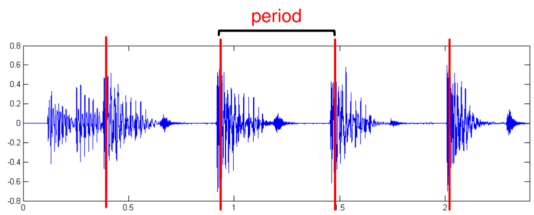
Introduction (Tempo and Beat)

Tasks

- Onset detection
- Beat tracking
- Tempo estimation

Tempo := 60 / period

Beats per minute (BPM)



Overview (Tempo and Beat)

Tasks

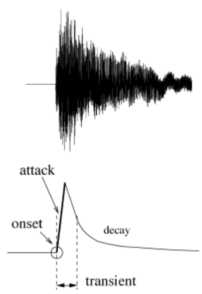
- Onset detection
- Beat tracking
- Tempo estimation

Onset Detection

- Finding perceptually relevant impulses in a music signal
- Onset is the time position where a note is played
- Onset typically goes along with a change of the signal's properties:
  - energy or loudness
  - pitch or harmony
  - timbre

Onset Detection

- Finding perceptually relevant impulses in a music signal
- Onset is the time position where a note is played
- Onset typically goes along with a change of the signal's properties:
  - energy or loudness
  - pitch or harmony
  - timbre

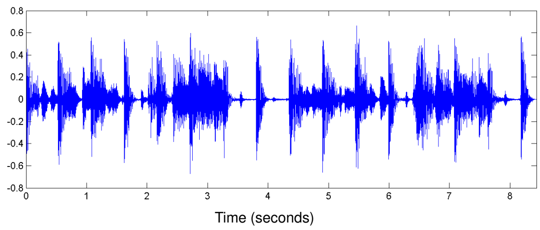


[Bello et al., IEEE-TASLP 2005]

Onset Detection (Energy-Based)

Steps

Waveform

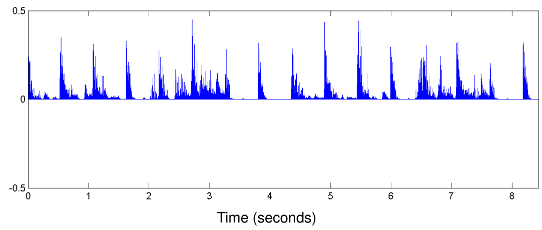


Onset Detection (Energy-Based)

Steps

1. Amplitude squaring

Squared waveform

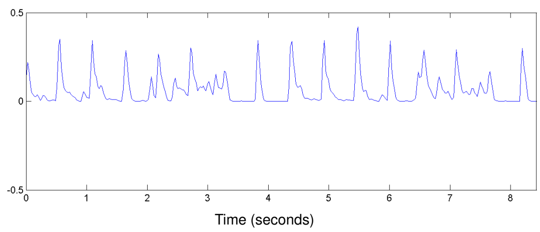


Onset Detection (Energy-Based)

Steps

1. Amplitude squaring
2. Windowing

Energy envelope



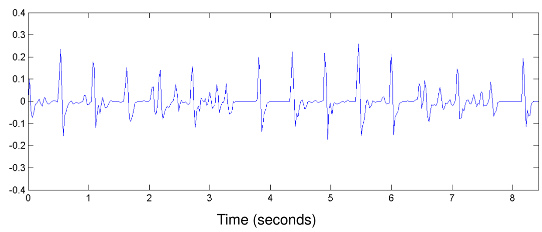
Onset Detection (Energy-Based)

Steps

1. Amplitude squaring
2. Windowing
3. Differentiation

Capturing energy changes

Differentiated energy envelope



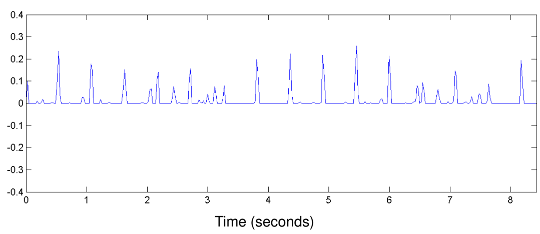
Onset Detection (Energy-Based)

Steps

1. Amplitude squaring
2. Windowing
3. Differentiation
4. Half wave rectification

Only energy increases are relevant for note onsets

Novelty curve

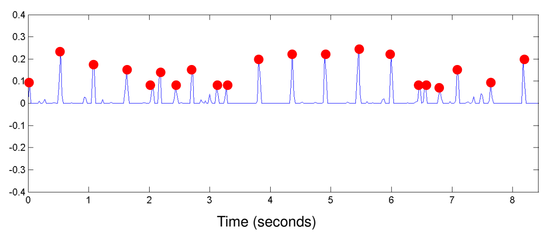


Onset Detection (Energy-Based)

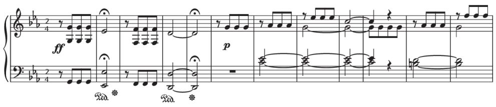
Steps

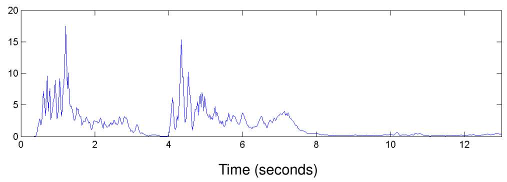
1. Amplitude squaring
2. Windowing
3. Differentiation
4. Half wave rectification
5. Peak picking

Peak positions indicate note onset positions

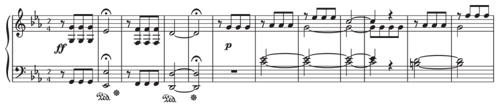


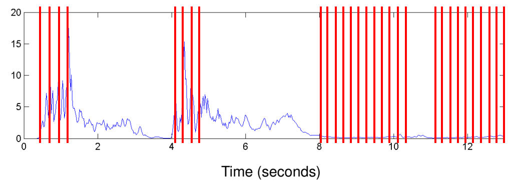
Onset Detection (Energy-Based)





Onset Detection (Energy-Based)



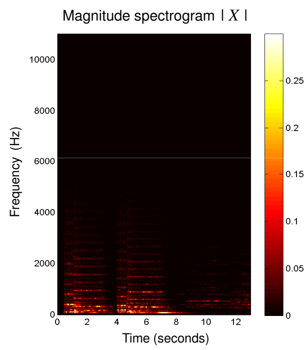


Onset Detection (Energy-Based)

- Energy curves often only work for percussive music
- Many instruments such as strings have weak note onsets
- No energy increase may be observable in complex sound mixtures
- More refined methods needed that capture
  - changes of spectral content
  - changes of pitch
  - changes of harmony

[Bello et al., IEEE-TASLP 2005]

Onset Detection (Spectral-Based)

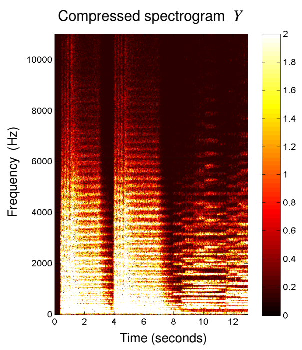


Steps:

1. Spectrogram

- *Aspects concerning pitch, harmony, or timbre are captured by spectrogram*
- *Allows for detecting local energy changes in certain frequency ranges*

Onset Detection (Spectral-Based)



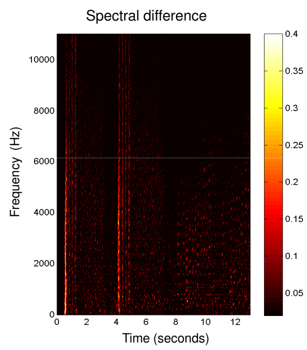
Steps:

1. Spectrogram
2. Logarithmic compression

$$Y = \log(1 + C \cdot | X |)$$

- *Accounts for the logarithmic sensation of sound intensity*
- *Dynamic range compression*
- *Enhancement of low-intensity values*
- *Enhancement of high-frequency spectrum*

Onset Detection (Spectral-Based)

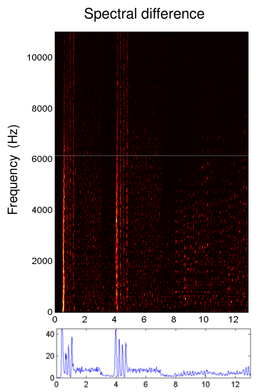


Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation

- *First-order temporal difference*
- *Captures changes of the spectral content*
- *Only positive intensity changes considered*

Onset Detection (Spectral-Based)



Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation
4. Accumulation

- *Frame-wise accumulation of all positive intensity changes*
- *Encodes changes of the spectral content*

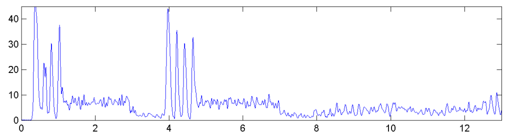
Novelty curve

Onset Detection (Spectral-Based)

Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation
4. Accumulation

Novelty curve



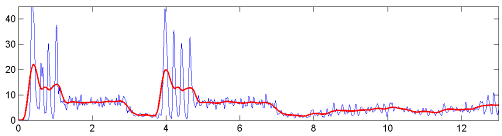
Onset Detection (Spectral-Based)

Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation
4. Accumulation
5. Normalization

Novelty curve

Substraction of local average

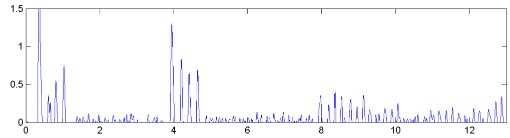


Onset Detection (Spectral-Based)

Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation
4. Accumulation
5. Normalization

Normalized novelty curve

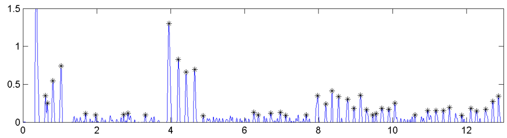


Onset Detection (Spectral-Based)

Steps:

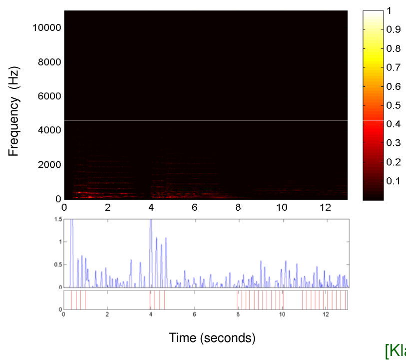
1. Spectrogram
2. Logarithmic compression
3. Differentiation
4. Accumulation
5. Normalization
6. Peak picking

Normalized novelty curve



Onset Detection (Spectral-Based)

Logarithmic compression is essential



| X |

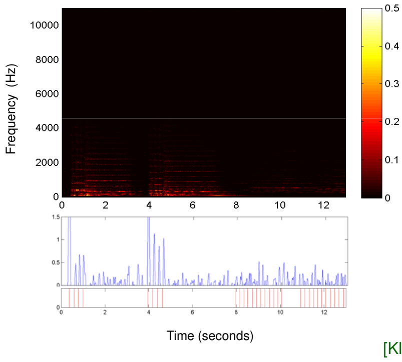
Novelty curve

Groundtruth onsets

[Klapuri et al., IEEE-TASLP 2006]

Onset Detection (Spectral-Based)

Logarithmic compression is essential



$$Y = \log(1 + C \cdot |X|)$$

C = 1

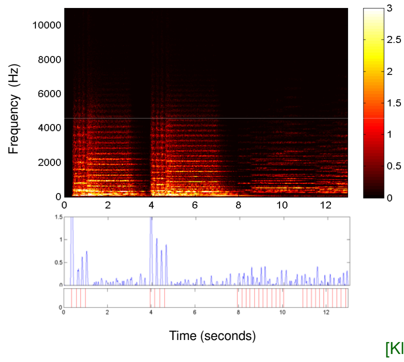
Novelty curve

Groundtruth onsets

[Klapuri et al., IEEE-TASLP 2006]

Onset Detection (Spectral-Based)

Logarithmic compression is essential



$$Y = \log(1 + C \cdot |X|)$$

C = 10

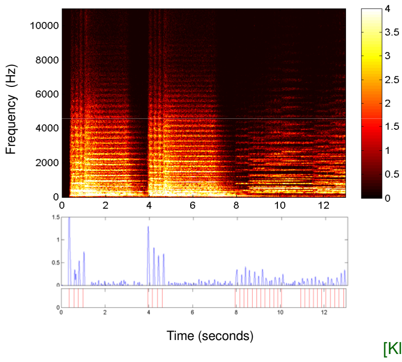
Novelty curve

Groundtruth onsets

[Klapuri et al., IEEE-TASLP 2006]

Onset Detection (Spectral-Based)

Logarithmic compression is essential



$$Y = \log(1 + C \cdot |X|)$$

C = 1000

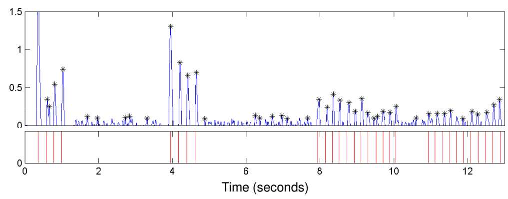
Novelty curve

Groundtruth onsets

[Klapuri et al., IEEE-TASLP 2006]

Onset Detection

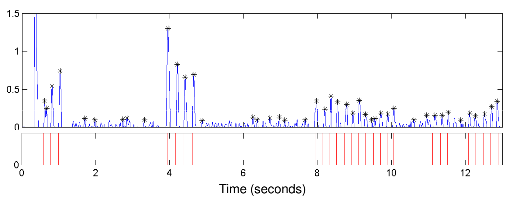
Peak picking



- Peaks of the novelty curve indicate note onset candidates

Onset Detection

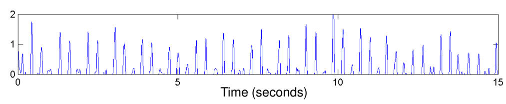
Peak picking



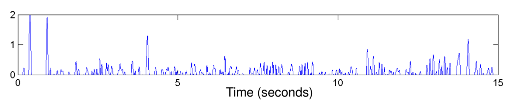
- Peaks of the novelty curve indicate note onset candidates
- In general many spurious peaks
- Usage of local thresholding techniques
- Peak-picking very fragile step in particular for soft onsets

Onset Detection

Shostakovich – 2nd Waltz



Borodin – String Quartet No. 2



Overview (Tempo and Beat)

Tasks

- Onset detection
- Beat tracking
- Tempo estimation

Beat and Tempo

What is a beat?

- Steady pulse that drives music forward and provides the temporal framework of a piece of music
- Sequence of perceived pulses that are equally spaced in time
- The pulse a human taps along when listening to the music

[Parncutt 1994]
[Sethares 2007]
[Large/Palmer 2002]
[Lerdahl/ Jackendoff 1983]
[Fitch/ Rosenfeld 2007]

The term tempo then refers to the speed of the pulse.

Beat and Tempo

Strategy

- Analyze the novelty curve with respect to reoccurring or quasi-periodic patterns
- Avoid the explicit determination of note onsets (no peak picking)

Beat and Tempo

Strategy

- Analyze the novelty curve with respect to reoccurring or quasi-periodic patterns
- Avoid the explicit determination of note onsets (no peak picking)

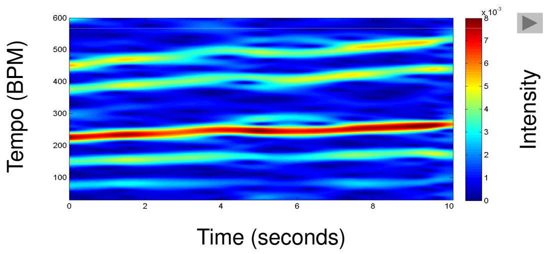
Methods

- Comb-filter methods
- Autocorrelation
- Fourier transfrom

[Scheirer, JASA 1998]
[Ellis, JNMR 2007]
[Davies/Plumbley, IEEE-TASLP 2007]
[Peeters, JASP 2007]
[Grosche/Müller, ISMIR 2009]

Tempogram

Definition: A tempogram is a time-tempo representation that encodes the local tempo of a music signal over time.



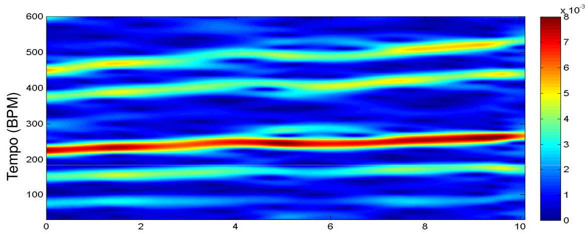
Tempogram (Fourier)

Definition: A tempogram is a time-tempo represenation that encodes the local tempo of a music signal over time.

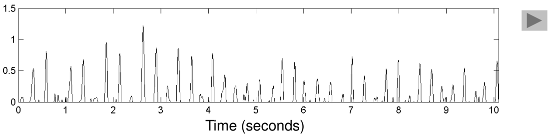
Fourier-based method

- Compute a spectrogram (STFT) of the novelty curve
- Convert frequency axis (given in Hertz) into tempo axis (given in BPM)
- Magnitude spectrogram indicates local tempo

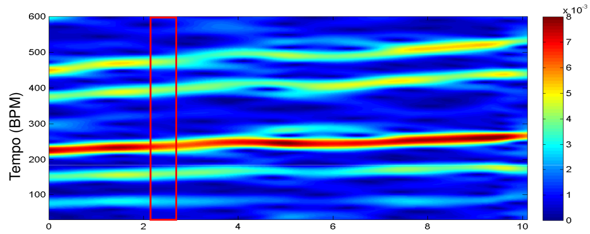
Tempogram (Fourier)



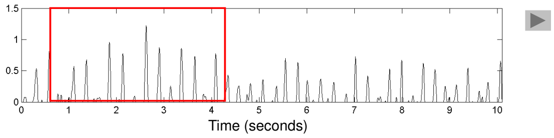
Novelty curve



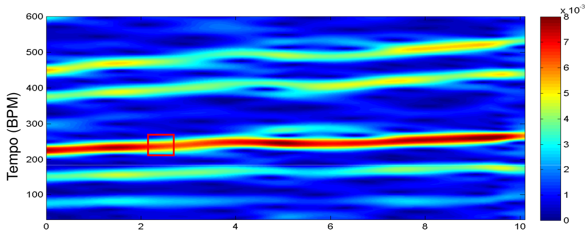
Tempogram (Fourier)



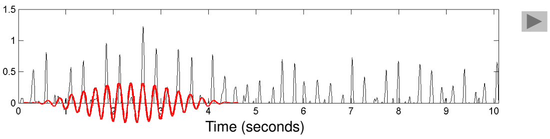
Novelty curve (local section)



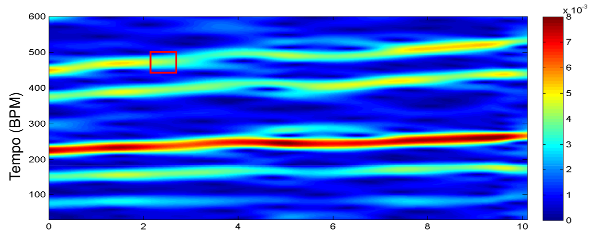
Tempogram (Fourier)



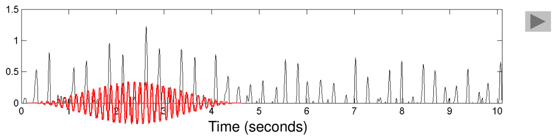
Windowed sinusoidal



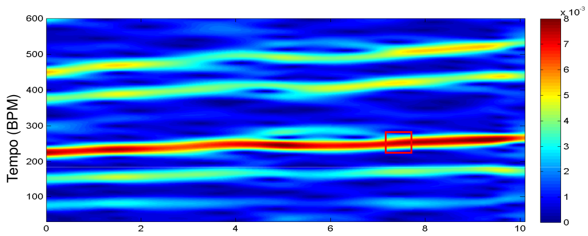
Tempogram (Fourier)



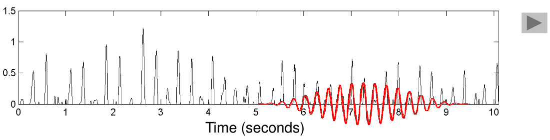
Windowed sinusoidal



Tempogram (Fourier)



Windowed sinusoidal



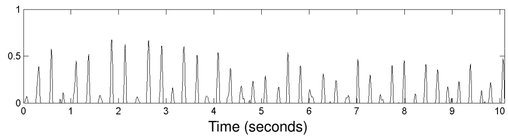
Tempogram (Autocorrelation)

Definition: A tempogram is a time-tempo represenation that encodes the local tempo of a music signal over time.

Autocorrelation-based method

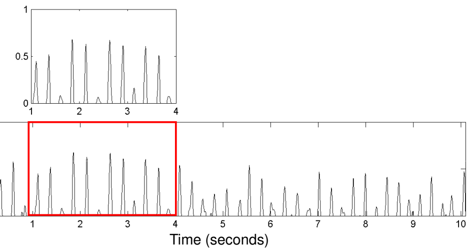
- Compare novelty curve with time-lagged local sections of itself
- Convert lag-axis (given in seconds) into tempo axis (given in BPM)
- Autocorrelogram indicates local tempo

Tempogram (Autocorrelation)



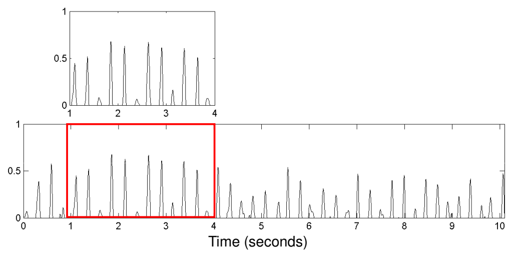
Novelty curve

Tempogram (Autocorrelation)



Novelty curve (local section)

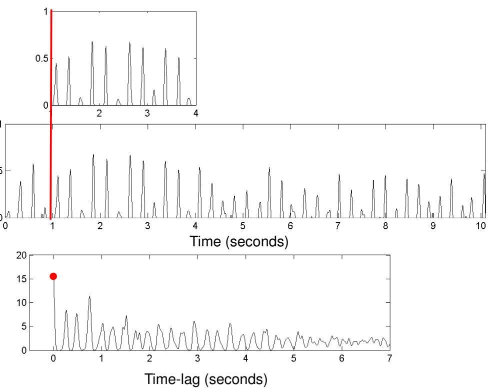
Tempogram (Autocorrelation)



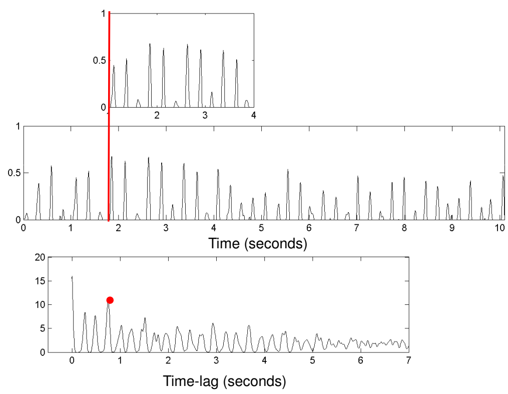
Novelty curve (local section)

Compare novelty curve with time-lagged local sections

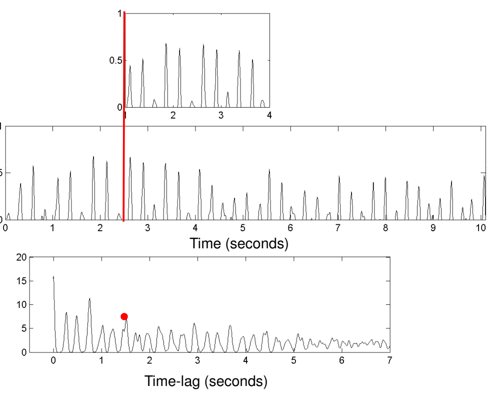
Tempogram (Autocorrelation)



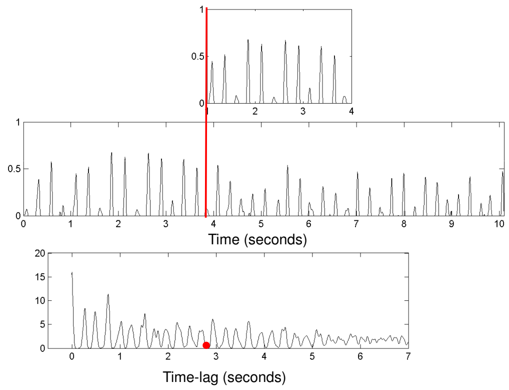
Tempogram (Autocorrelation)



Tempogram (Autocorrelation)



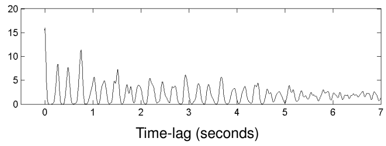
Tempogram (Autocorrelation)



Tempogram (Autocorrelation)

- Time lag of high value indicates high correlation
- Autocorrelation reveals periodic self-similarities
- Maximum for a lag of zero (no shift)

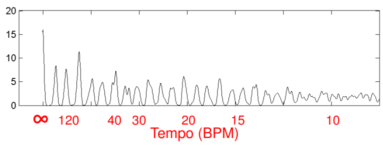
Windowed autocorrelation



Tempogram (Autocorrelation)

- Time lag of high value indicates high correlation
- Autocorrelation reveals periodic self-similarities
- Maximum for a lag of zero (no shift)
- Convert time-lag axis (seconds) into tempo axis (BPM)

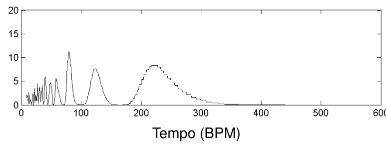
Windowed autocorrelation



Tempogram (Autocorrelation)

- Time lag of high value indicates high correlation
- Autocorrelation reveals periodic self-similarities
- Maximum for a lag of zero (no shift)
- Convert time-lag axis (seconds) into tempo axis (BPM)
- Convert into linear tempo axis

Windowed autocorrelation



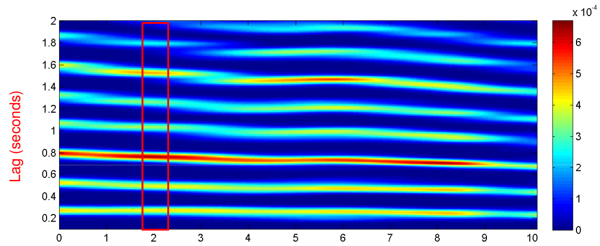
Tempogram (Autocorrelation)

- Time lag of high value indicates high correlation
- Autocorrelation reveals periodic self-similarities
- Maximum for a lag of zero (no shift)
- Convert time-lag axis (seconds) into tempo axis (BPM)
- Convert into linear tempo axis

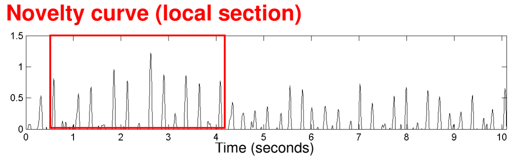
Do this for a sliding window.

Every window defines a local section for which a windowed autocorrelation is computed.

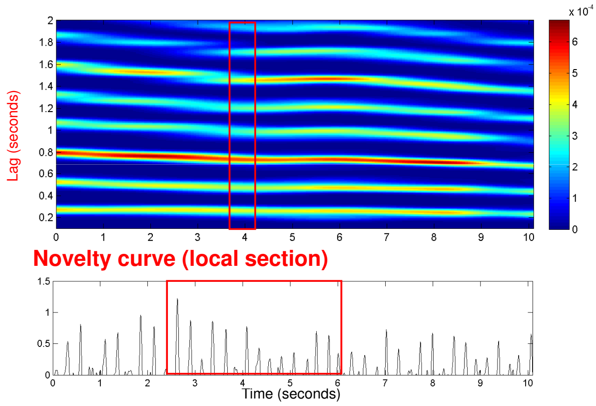
Tempogram (Autocorrelation)



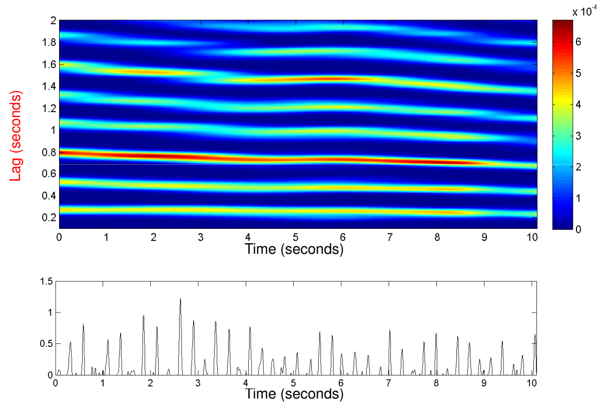
Novelty curve (local section)



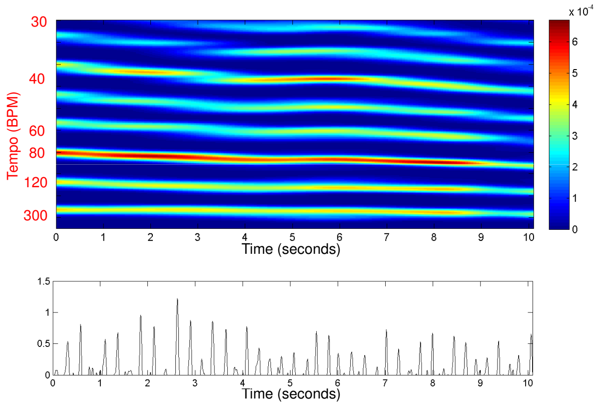
Tempogram (Autocorrelation)



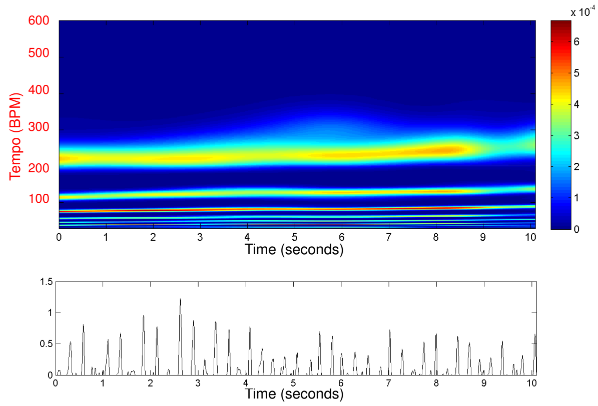
Tempogram (Autocorrelation)



Tempogram (Autocorrelation)



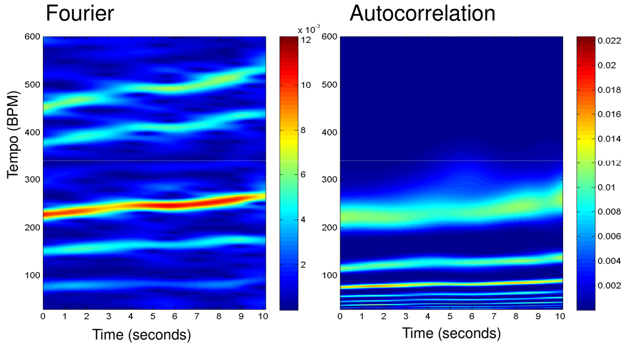
Tempogram (Autocorrelation)



Tempogram

Fourier

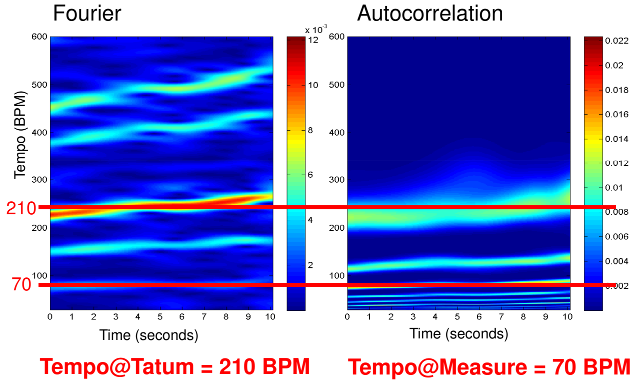
Autocorrelation



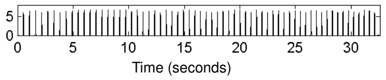
Tempogram

Fourier

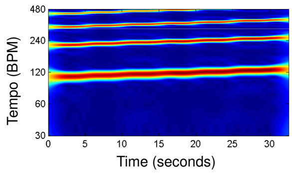
Autocorrelation



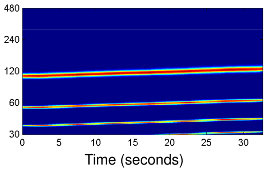
Tempogram



Fourier



Autocorrelation



Emphasis of tempo harmonics (integer multiples)

Emphasis of tempo subharmonics (integer fractions)

[Peeters, JASP 2007][Grosche et al., ICASSP 2010]

Tempogram (Summary)

Fourier	Autocorrelation
Novelty curve is compared with sinusoidal kernels each representing a specific tempo	Novelty curve is compared with time-lagged local (windowed) sections of itself
Convert frequency (Hertz) into tempo (BPM)	Convert time-lag (seconds) into tempo (BPM)
Reveals novelty periodicities	Reveals novelty self-similarities
Emphasizes harmonics	Emphasizes subharmonics
Suitable to analyze tempo on tatum and tactus level	Suitable to analyze tempo on tatum and measure level

Overview (Tempo and Beat)

Tasks

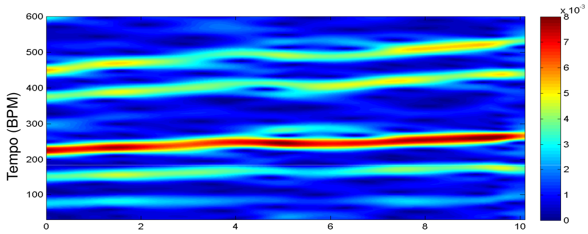
- Onset detection
- Beat tracking
- Tempo estimation

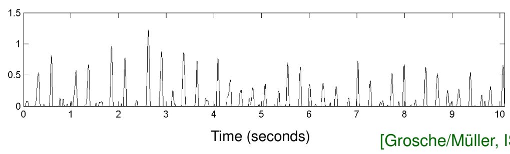
Beat Tracking

- Given the tempo, find the best sequence of beats
- Complex Fourier tempogram contains magnitude and phase information
- The magnitude encodes how well the novelty curve resonates with a sinusoidal kernel of a specific tempo
- The phase optimally aligns the sinusoidal kernel with the peaks of the novelty curve

[Peeters, JASP 2007]

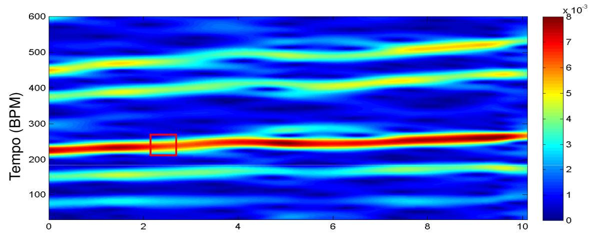
Beat Tracking



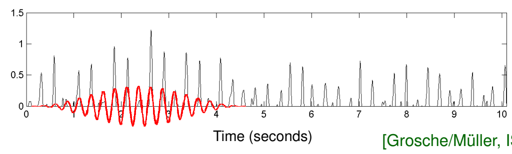


[Grosche/Müller, ISMIR 2009]

Beat Tracking

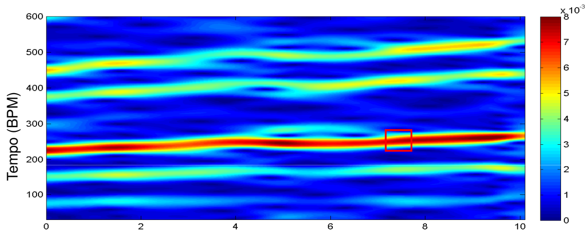


Optimal sinusoidal kernel (obtained from complex Fourier tempogram)

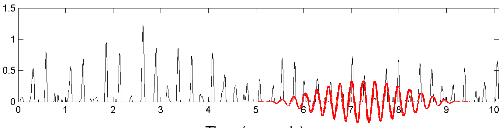


[Grosche/Müller, ISMIR 2009]

Beat Tracking



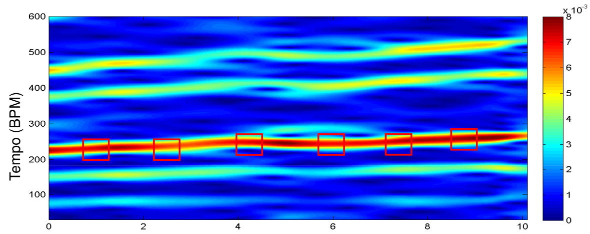
Optimal sinusoidal kernel (obtained from complex Fourier tempogram)



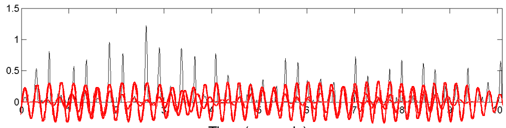
Time (seconds)

[Grosche/Müller, ISMIR 2009]

Beat Tracking



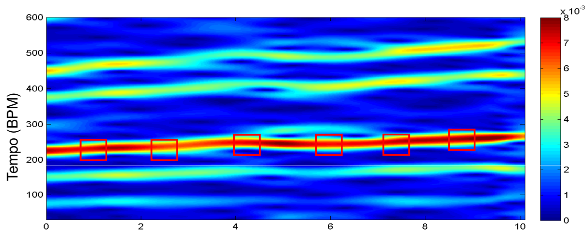
Accumulation of kernels



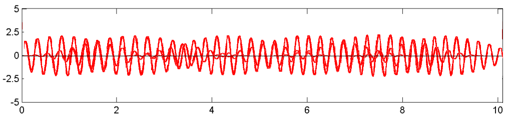
Time (seconds)

[Grosche/Müller, ISMIR 2009]

Beat Tracking



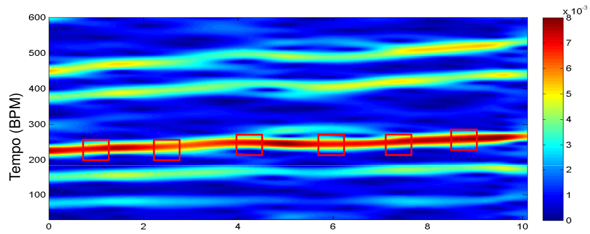
Accumulation of kernels



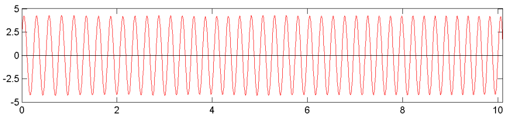
Time (seconds)

[Grosche/Müller, ISMIR 2009]

Beat Tracking



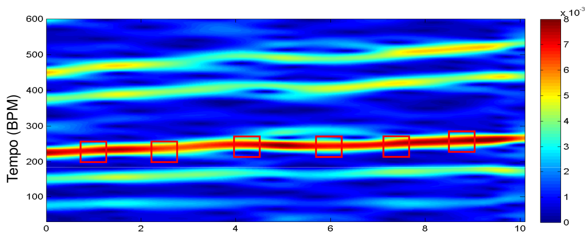
Accumulation of kernels



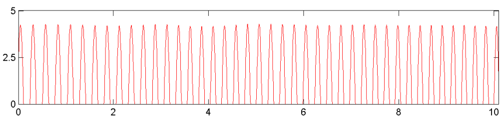
Time (seconds)

[Grosche/Müller, ISMIR 2009]

Beat Tracking



Halfwave rectification

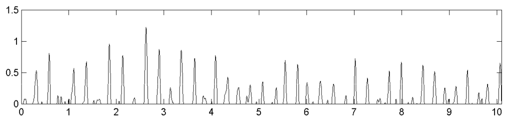


Time (seconds)

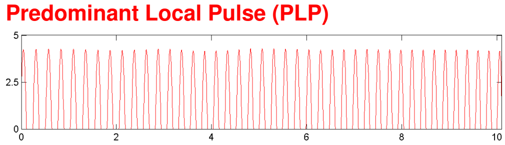
[Grosche/Müller, ISMIR 2009]

Beat Tracking

Novelty Curve



Predominant Local Pulse (PLP)



Time (seconds)

[Grosche/Müller, ISMIR 2009]

Beat Tracking

Novelty Curve

- Indicates note onset candidates
- Extraction errors in particular for soft onsets
- Simple peak-picking problematic

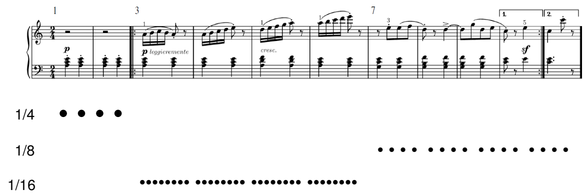
Predominant Local Pulse (PLP)

- Periodicity enhancement of novelty curve
- Accumulation introduces error robustness
- Locality of kernels handles tempo variations

[Grosche/Müller, ISMIR 2009]

Pulse Levels

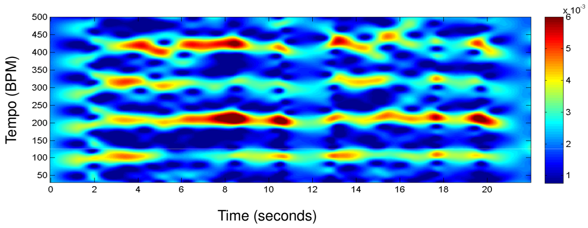
Piano Etude Op. 100 No. 2 by Burgmüller



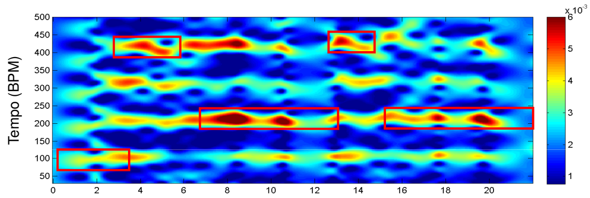
What is the pulse level: Measure – Tactus – Tatum?

[Klapuri et al., IEEE-TASLP 2006]

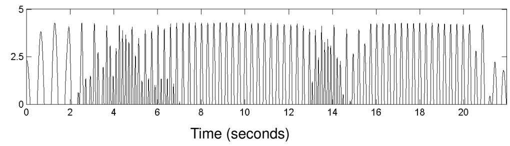
Pulse Levels



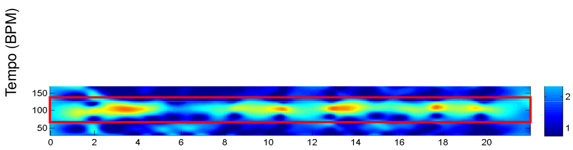
Pulse Levels



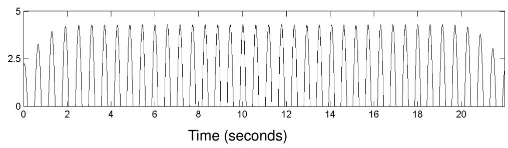
Switching of predominant pulse level



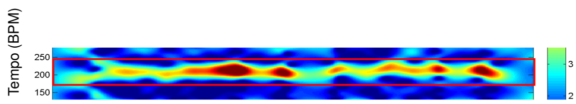
Pulse Levels



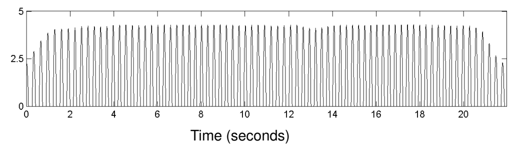
1/4 note pulse level



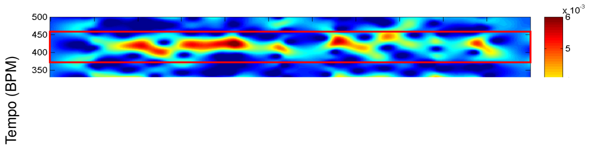
Pulse Levels



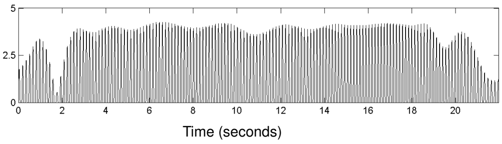
1/8 note pulse level



Pulse Levels

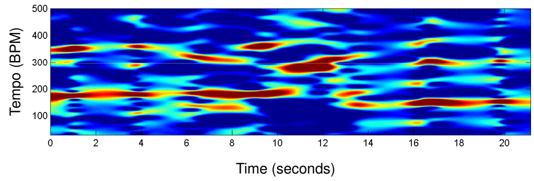


1/16 note pulse level



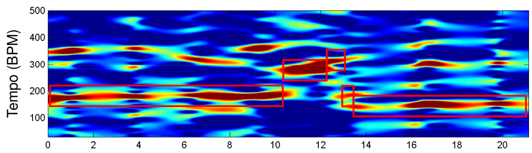
Evaluation

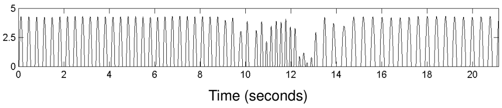
Brahms Hungarian Dance No. 5



Evaluation

Brahms Hungarian Dance No. 5

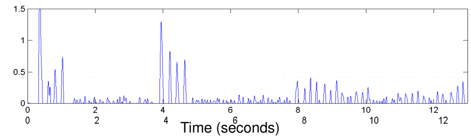




Evaluation

Beethoven Symphony No. 5

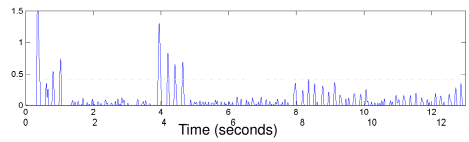
Novelty Curve



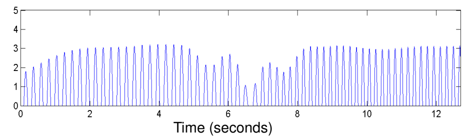
Evaluation

Beethoven Symphony No. 5

Novelty Curve

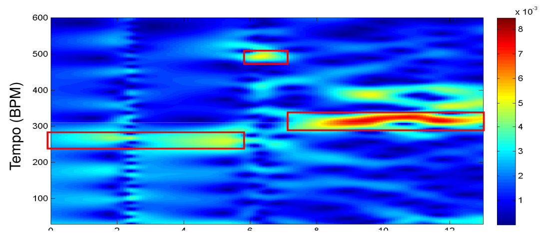


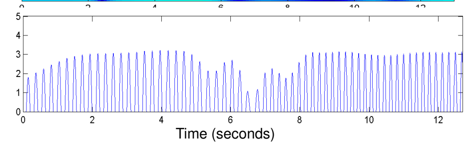
Predominant Local Pulse (PLP)



Evaluation

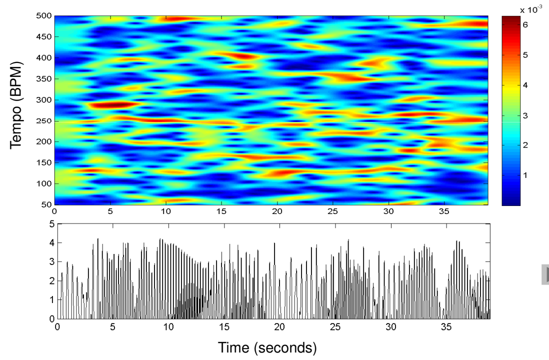
Beethoven Symphony No. 5





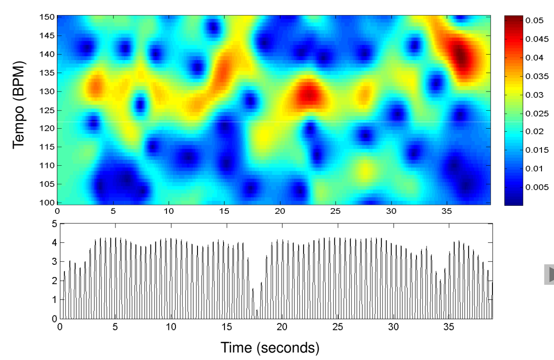
Evaluation

Borodin – String Quartet No. 2



Borodin – String Quartet No. 2

Borodin – String Quartet No. 2



Conclusions (PLP)

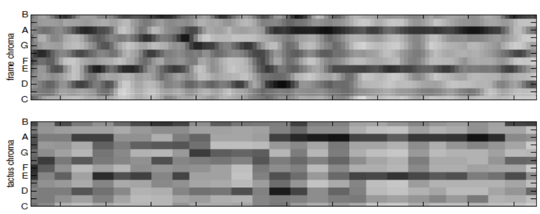
- Predominant local pulse (PLP)
- Reveals pulse rate (tempo) and pulse positions
- Periodicity enhancement of novelty curves
- Suitable for non-percussive music with tempo variations
- Combination with autocorrelation methods
- Tempo-based audio segmentation

[Peeters, JASP 2007]
[Jensen, JASP 2007]
[Müller/Grosche, ICASSP 2010]
[Paulus/Klapuri, IEEE-TASLP 2009]

Applications (Beat and Tempo)

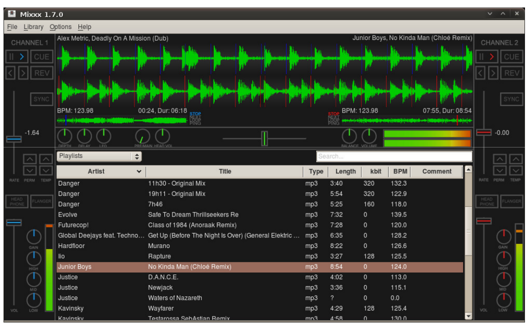
- Feature design
  (usage of beat-synchronous windows of adaptive size)
- Digital DJ / audio editing
  (mixing and blending of audio material)
- Music classification
- Music recommendation
- Performance analysis
  (extraction of tempo curves)

Application: Beat-Synchronous Features



[Bello/Pickens, ISMIR 2005]

Application: Audio Editing (Digital DJ)



http://www.mixxx.org/

Application: Beat-Synchronous Light Effects

Tutorial

A Music-oriented Approach to Music Signal Processing

Meinard Müller
Saarland University and MPI Informatik
meinard@mpi-inf.mpg.de

Anssi Klapuri
Queen Mary University of London
anssi.klapuri@elec.qmul.ac.uk

Overview

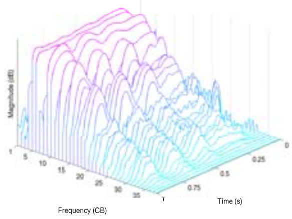
Part I: Pitch and Harmony

Part II: Tempo and Beat

Coffee Break

Part III: Timbre

Part IV: Melody



Timbre

- Characterizes the identity of a sound source
- Perceptual attribute of sounds, separate from pitch, loudness and duration
- Examples of sounds with the same pitch and root-mean-square (RMS) levels, but different timbre:
- In MIR, the term is usually stretched to refer to the instrumentation aspects of a polyphonic signal
- Recent MIR PhD theses addressing timbre: [Kitahara-PhD-07], [Eronen-PhD-10]
- Focus here: what is unique for music compared to speech

Acoustic features underlying timbre

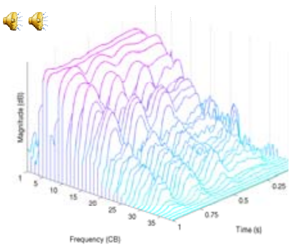
- Timbre is an inherently multidimensional concept
- Several underlying acoustic features of both spectral and temporal types
- Schouten's [1968] list of the five attributes of timbre:
  1. its character ranging from "tonal" to "noiselike"
  2. spectral envelope
  3. time envelope in terms of rise, duration, and decay
  4. fluctuations of spectral envelope and pitch
  5. onset differing notably from the steady state

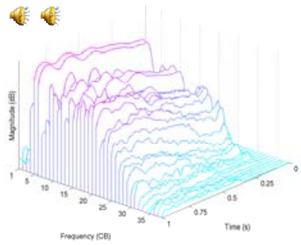
Acoustic features underlying timbre

- Usually when signal processing people (like me) talk about timbre, they think about the spectral envelope
- Stems from speech recognition
- Limited view, but good as a first approximation

Time-varying spectral envelope

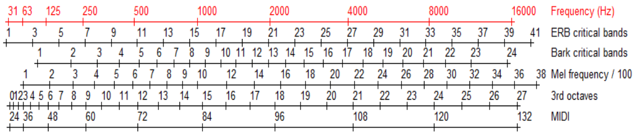
- As a first approximation, let us assume "timbre ≈ levels at critical bands as a function of time"
- Flute (left) and violin (right) spectra





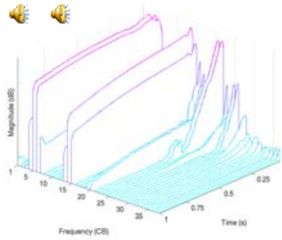
Critical-band scales

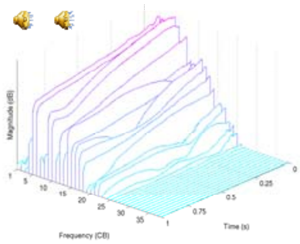
- Critical-band scales describe the frequency resolution of the auditory system
- On the previous slide, ERB scale was used
  $f_{\mathrm{ERB}} = 21.4\times\log10(0.00437\, f_{\mathrm{Hz}} + 1)$
- Mel-scale is often used too
  $f_{\mathrm{Mel}} = 2595\times\log10(0.00143\, f_{\mathrm{Hz}} + 1)$
- Bark scale is very similar, see comparison below



Time-varying spectral envelope

- More examples: vibraphone (left) and piano (right)
- On Schouten's list, this representation covers 2 (spectral envelope), 3 (time envelope), part of 4 (fluctuations) and much of 5 (onset vs. steady)





Variation from "tonal" to "noiselike"

- Time-varying spectral envelope is the main determinant of timbre, but it is not all
- In music, there are other important factors too

- Consider the variation from "tonal" to "noiselike"
- In the following examples, the proportion of tonal vs. noisy spectral components is varied, keeping the time-varying spectral envelope unchanged
  - Flute   Singing

Variation from "tonal" to "noiselike"

- The above suggests that we should break a music signal into its tonal and noisy components and then attach "proportion of tonal vs. noisy" descriptor to each critical band (in addition to its level)
- Useful tools for doing this
  - Sinusoids + noise model [Serra-1997]
  - Harmonic and percussive separation [Ono-2008]

Timbre features beyond time-varying spectral envelope

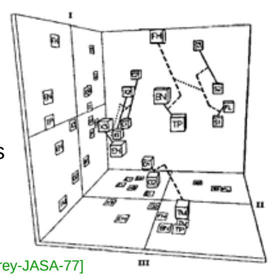
- In examples below, the time-varying spectral envelope of one sound ("mould") is imposed on another sound ("material"), without changing the spectral fine structure or phases of the latter sound
- Does the identity of the source change?

Material	Mould sound			
	trumpet	clarinet	piano	flute
trumpet				
clarinet				
piano				
flute				

- Conclusion: spectral fine structure and phases affect timbre too

Main acoustic factors of timbre

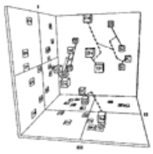
- The above timbre representations are not very compact
- What are the main acoustic factors of timbre differences?
- Multidimensional scaling (MDS) experiments address this question:
  1. Let subjects rate the dissimilarity of timbre pairs
  2. Squeeze the data into a low-dimensional space, trying to preserve distances
  3. Find acoustic correlates to the dimensions of this perceptual space



[Grey-JASA-77]

Main acoustic factors of timbre

- Note that MDS is based on distances only, not on absolute positions (→ rotational ambiguity etc.)
- Main acoustic factors of timbre found in MDS experiments [Grey-77, Krumhansl-89, McAdams-95, Caclin-07]
  - Spectral centroid (center of gravity): $\Sigma\, kX(k) \,/\, \Sigma\, X(k)$
  - Log attack time ($\log(t_{max} - t_{thresh})$)
  - Spectral irregularity (≈ amplitude difference of neighbouring harmonic partials)
  - Spectral flux (irregularity over time)

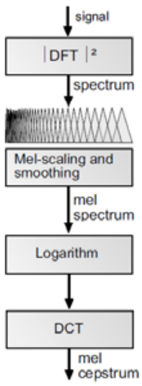


Acoustic feature extraction for timbre

- Let us move on from timbre perception to the practical extraction of acoustic features from audio for timbre description
- Emphasis here is on musical and perceptual relevance of the features

Mel-frequency cepstral coefficients (MFCC)

- MFCCs describe the spectral envelope and are the most widely used feature for recognizing speech or instruments
- Calculation
  1. Compute a power spectrogram
  2. Warp to Mel-frequency scale
  3. Log of the powers at Mel bands→dB
  4. Discrete cosine transform→decorrelate
- Toolboxes: see e.g. [LabRosa code page]



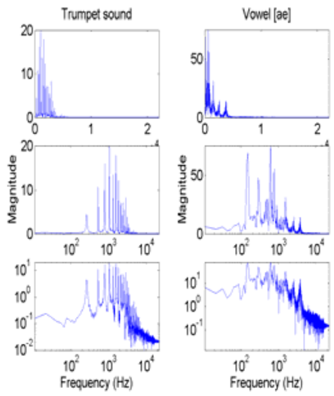
Mel-frequency cepstral coefficients (MFCC)

- Reasons why MFCCs are popular:
  - Straightforward to calculate
  - Mel-frequency scale
  - Log of magnitudes

    } *Large (small) numerical change ↔ large (small) perceptual change*
  - Discrete cosine transform } *Decorrelation, energy compaction*
- The amount of MFCC coefficients included controls the frequency resolution of the modelled spectral envelope

MFCC: Motivation for frequency and magnitude warping

- Linear scale
  - usually hard to "see" anything
- Log-frequency
  - each octave is approximately equally important perceptually
- Log-magnitude
  - perceived change from 50 to 60dB about the same as from 60 to 70dB

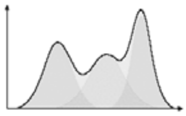


Other acoustic features

- A lot of different features have been used for instrument classification
- See [Peeters-TechReport-2004] for a comprehensive list
- However, many features are redundant with MFCCs and do not make a substantial improvement in instrument classification, for example
- When using several features, it is important to decorrelate them and reduce the dimensionality by principal component analysis (PCA) or linear discriminant analysis (LDA) or independent component analysis (ICA) [Matlab, Duda-Hart-book-2001]

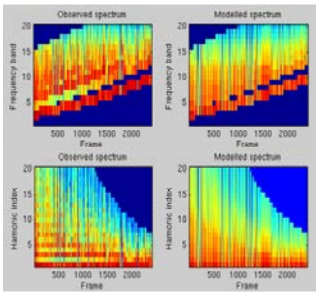
Timbre model for a sound source

- Note that acoustic features typically describe the properties of a short segment of one sound
- A timbre model should represent all sounds emitted by the modeled sound source (instrument)
- Typical approach
  - extract acoustic features from several example sounds
  - use a statistical model to represent the distribution of the features for a given source



Is time-frequency plane the right place for timbre modeling?

- Left: observed; Right: modeled with MFCCs
  - upper: as a function of frequency
  - lower: –"– of harm. index

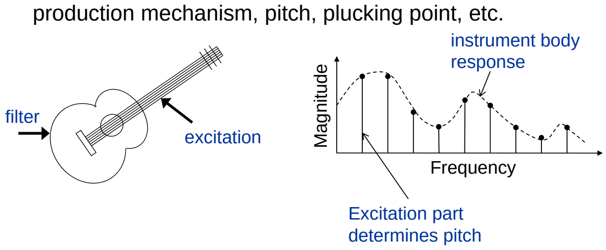


→ MFCCs do not capture the properties that vary as a function of harmonic index

→ Need to represent spectrum both as a function of frequency and as a function of harmonic index

Structured timbre models: Excitation-filter

- Excitation represents a vibrating object such as a guitar string and filter refers to the resonance structure of the rest of the instrument which colors the produced sound
- Excitation contains information about the sound production mechanism, pitch, plucking point, etc.



Excitation-filter signal model

- The magnitude spectrum $|S_t(f)|$ is modeled as
  $|S(f)| \approx \gamma X(h) B(f_h)$
  where $f_h \approx hF$ is the frequency of h-th overtone
  - γ represents the overall gain
  - $X(h)$ represents harmonic amplitudes at excitation
  - $B(f_h)$ represents the frequency response of the body
- Consider $|S_t(f_h)|$ on a decibel scale:
  $|S_{\mathrm{dB}}(f)| \approx \gamma_{\mathrm{dB}} + X_{\mathrm{dB}}(h) + B_{\mathrm{dB}}(f_h)$
  where
  $|S_{\mathrm{dB}}(f)| = 10\log_{10}(|S(f)|^2)$
  → Logarithm renders the model linear

Task

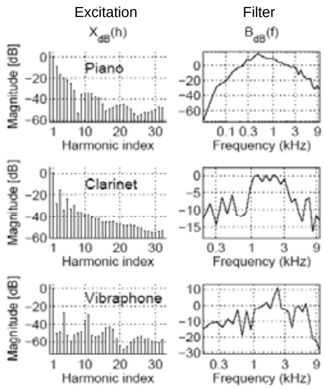
- Learn such $X(h)$ and $B(f_h)$ that all sounds emitted by the instrument can be approximated with the model
- Harmonic levels $X_{dB}(h)$ and body response $B_{dB}(f)$ are further represented with a linear model so that the number of free parameters can be controlled:

$$X_{dB}(h) = \sum_{i=1}^{C_x} \xi_i x_i(h) \qquad B_{dB}(f) = \sum_{j=1}^{C_b} \beta_j b_j(f)$$

→ Parameters to be estimated are the coefficients ξ_i and β_j

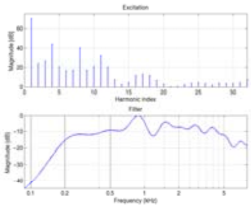
- Note: the vector of ≈ 30 numbers $[\xi_i, \beta_j]$ represents all sounds of the instrument (even without further statistical modeling) → compact model

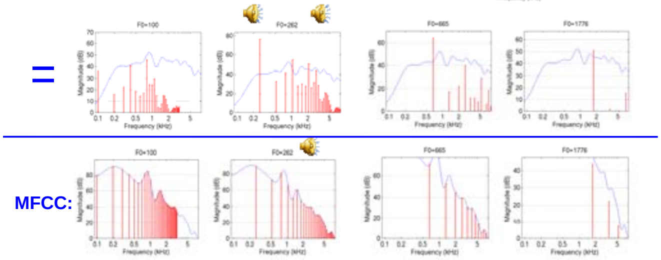
Models learned for piano, clarinet, and vibraphone



Marimba

- Example spectra (red) obtained by varying the pitch



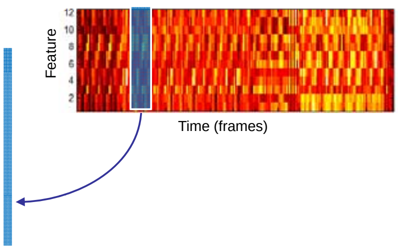


Temporal evolution

- So far we have discussed only models for the short-time spectrum within individual frames
- From previous examples (static spectral), it is clear that temporal evolution is very important
- Auditory system is quickly "exhausted" when listening to static spectra

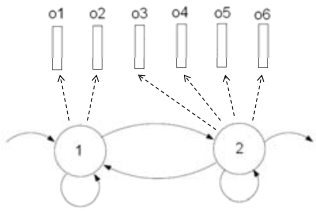
Conventional ways of representing temporal evolution

- Calculate time differential of features and append them to the feature vector (e.g. MFCC and ∆MFCC)
- Stack feature vectors from M successive frames into a single long vector, "audio shingle"



Conventional solutions: HMM

- Hidden Markov models (HMM)
  - state-conditional observation densities describe the observations generated by each state
  - transition probabilities control switching between the states
- HMM takes into account temporal structure while allowing duration variation

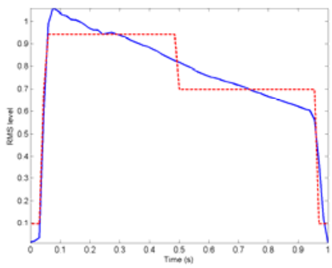


Conventional state models (HMM)

- Many musical sounds are poorly modeled using a convenional state model, where time-varying spectra are modeled by switching between states
- Figure: piano energy envelope modeled with three states

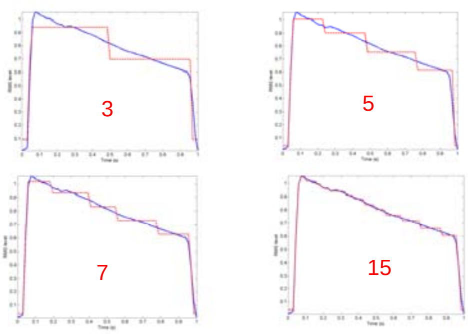
original

model



Conventional state models (HMM)

- Adding more states helps, but is inefficient

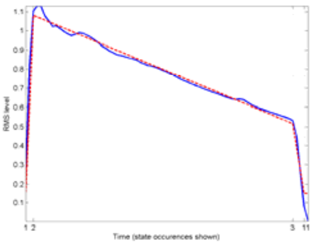


Interpolating state model

- Contrary to the above models, most musical sounds can be represented efficiently by interpolating between suitably selected spectra
  - Several examples of this in sound synthesis
- Figure: piano energy envelope modeled with three states
- Occurrence times of the three states are indicated on the x-axis

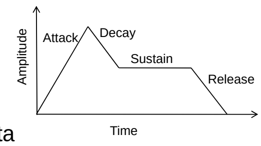
original

model



Interpolating state model

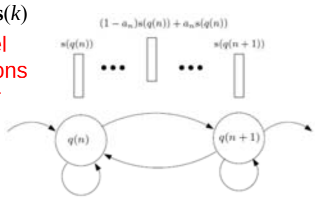
- The interpolating state model can be seen as a generalization of the attack-decay-sustain-release (ADSR) paradigm



- Generalizations
  - Multi-dimensional data
  - Turning points and levels are automatically estimated
  - Not specific to audio data (generalization of VQ)

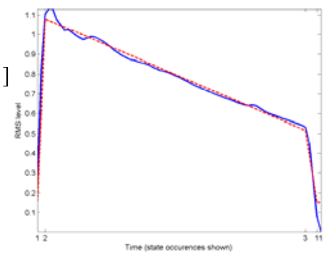
Interpolating state model

- Input data to be modelled are sequence of feature vectors $\mathbf{x}(i)$ extracted in frames $i=0,1,\ldots,T\text{-}1$
- Idea: find a small number of "state vectors" (anchor points in the feature space) so that the data can be approximated by interpolating between these
- There are $K \ll T$ states and each has its characteristic state vector $\mathbf{s}(k)$
- Figure: Output of the model is generated at the transitions between states, as a linear interpolation of the state vectors at the two ends



Interpolating state model

- During the transition, the model moves with a constant speed towards the next state
- The occurrence times of the states in their pure form are called nodes
- Nodes $n = 0,\ldots,N-1$ are characterized by a time stamp $t(n) \in [0, T-1]$ and state number that occurred $q(n) \in [0, K-1]$
- Figure: 3 states, 5 nodes

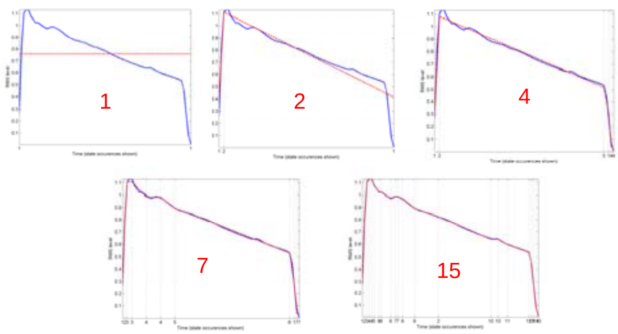


Parameter estimation

- Parameters of the model can be estimated in $T \log T$ time, where T is the length of the feature sequence [Klapuri-TASLP-2010]

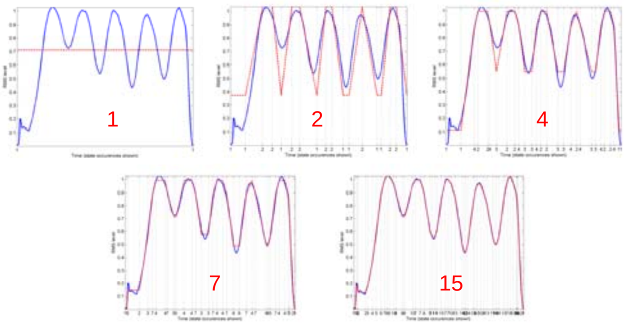
Interpolating state model

- Piano: varying the number of states



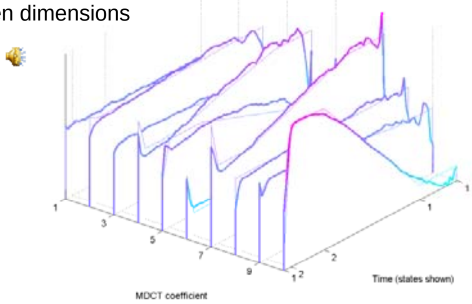
Interpolating state model

- Flute: varying the number of states



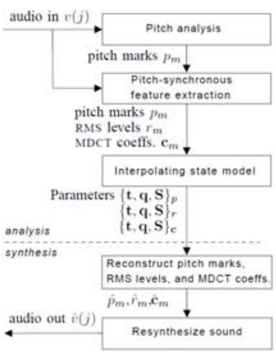
Multidimensional data

- Modeling piano MDCT coefficients (with 2 states)
- Note that state occurrence times are common to all ten dimensions



Audio coding with interpolating state model

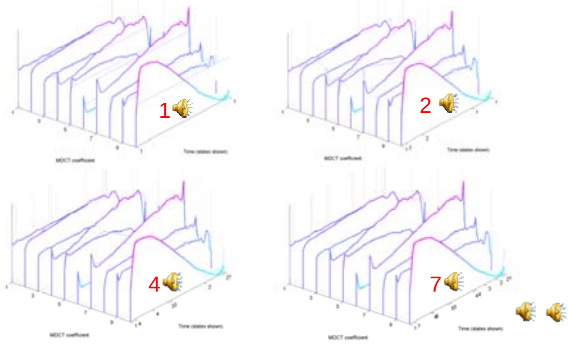
- Pitch-synchronous waveform modeling: assumes only one sound is playing at a time (monophonic)
- Pitch (period length), energy, and waveshape are each encoded separately



Audio coding

original

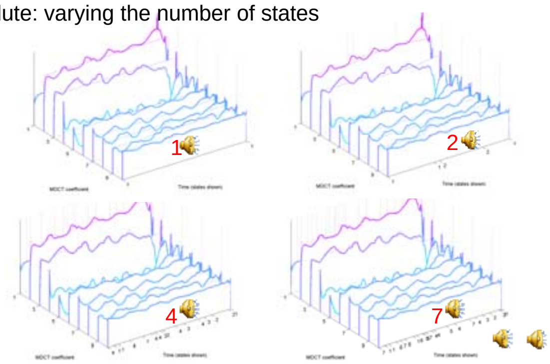
- Piano: varying the number of states



Audio coding

original

- Flute: varying the number of states

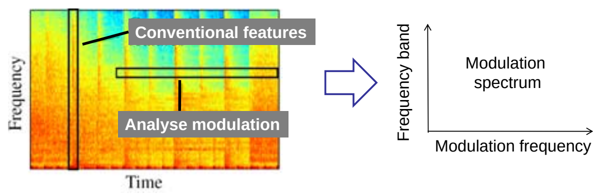


Interpolating state model: Summary

- Interpolation might be a good idea in music
- About 3dB better SNR than using vector quantization with the same model order
- The method has not been used for audio classification so far
- The model is completely deterministic, therefore further statistical modeling of the parameter distributions is required

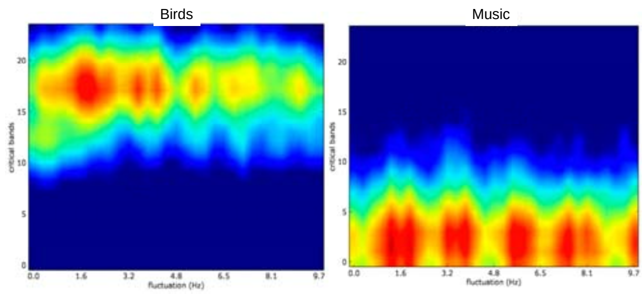
Modulation spectrum: *Texture* of music

- Modulation spectrum is a promising way of modeling the texture of complex music pieces, and complex timbres, such as animal sounds
- A.k.a. fluctuation patters [Pampalk-MSc, Dixon-03]
- Shift-invariance



Modulation spectrum

- Video examples here are courtesy of Thomas Grill [grrrr.org]



Applications of timbre analysis and modeling

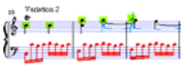
- Instrument recognition
- Sound source separation and streaming
- Sound synthesis and composition
- Analysis of instrument acoustics

Remaining challenges

- Polyphonic instrument recognition
  - would have implications on robust speech recognition and sound separation
  - see [Kitahara-06, Essid-06, Burred-09, Heittola-09]
- Polyphonic recognition and sound separation are closely related problems
  - solve one and you have solved the other
  - recognition allows generating a spectro-temporal mask

Conclusions

- Basics of timbre modeling stem from hearing and are therefore common to speech and music: critical-band scales and log-magnitude scale
- Musical instruments comprise several sound production mechanisms. Excitation-filter model is needed to capture aspects of excitation well.
- Musical sounds are generally more slowly-varying than speech, therefore interpolating models are well-suited in music
- Modulation spectra have attractive properties for modeling the texture of music

Tutorial

A Music-oriented Approach to Music Signal Processing

Meinard Müller
Saarland University and MPI Informatik
meinard@mpi-inf.mpg.de

Anssi Klapuri
Queen Mary University of London
anssi.klapuri@elec.qmul.ac.uk

Overview

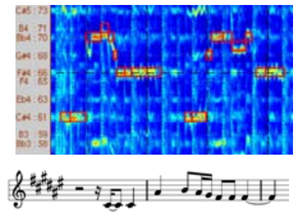
Part I: Pitch and Harmony

Part II: Tempo and Beat

Coffee Break

Part III: Timbre

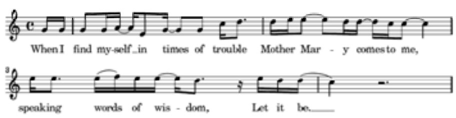
Part IV: Melody

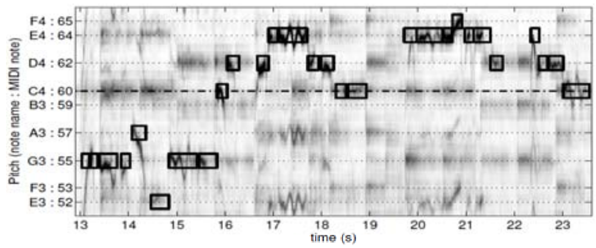


Melody

- Oxford English Dictionary: "A series of single notes arranged in musically expressive succession"
- Usually performed by a lead singer or by a solo instrument
- The part of music that listeners tend to remember and are innately able to reproduce by humming
- Recent MIR PhD theses addressing melody and vocals extraction: [Paiva-06, Ryynanen-08, Fujihara-10]

Example: "Let It Be"





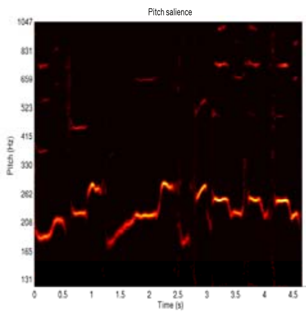
[Ryynänen-PhD-08]

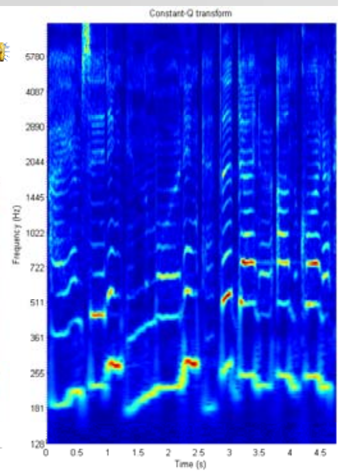
Difference between audio and written music

- Note how far the sung melody is from the idealized written music
- Vibrato, glissando, ambiguities (see e.g. E4 note "Let" at 22.4 s)
- This is not because the singing is below ideal, but because written notation is so limited
- Deriving discrete notation from a singing performance requires heavy use of musical context

"Tom's Diner" by Suzanne Vega

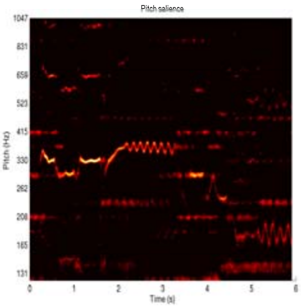
- About as "right-angled" a performance as it gets

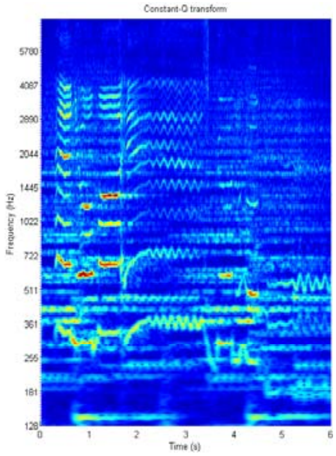




“Summertime“ by Ella Fitzgerald

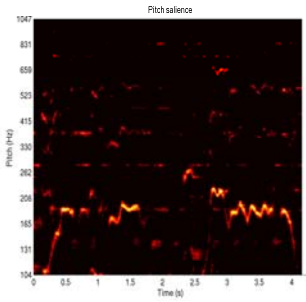
- Vibrato and glissandi (2 s)
- Formant structure

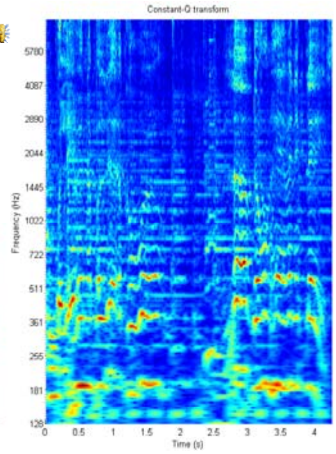




“Political World“ by Bob Dylan

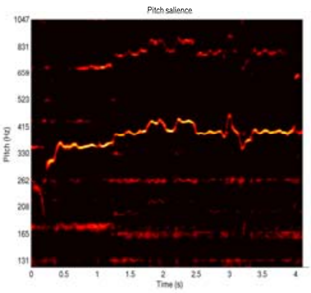
- Short low-pitched notes at the beginning

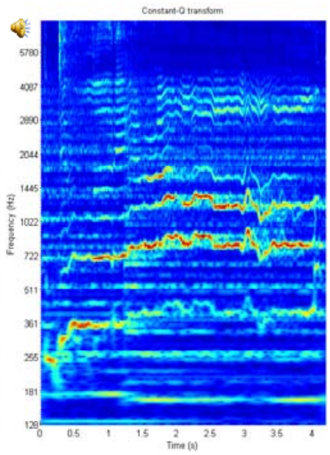




“Nothing Compares 2 U“ by Sinéad O‘Connor

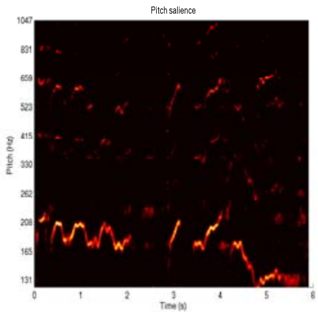
- Trick at 3s, falsetto at end
- Formant around 3kHz

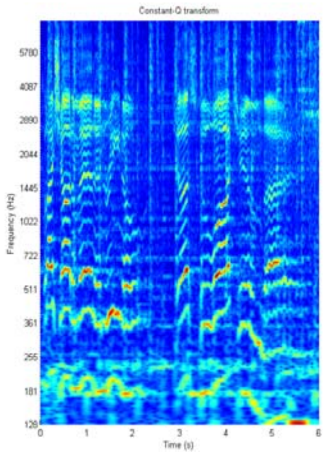




“Folsom Prison Blues“ by Johnny Cash

- Glissandi near the end





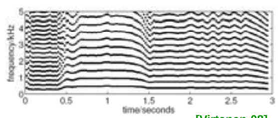
Acoustic characteristics of melodies

- Pitch range typically 100Hz–1kHz (Ab2–C6)
- Relatively prominent (loud) compared to other instruments
- Vocal timbre: varying but identifiable
- Usually panned at the center of the stereo field
- Vibrato and pitch glides make the vocals stand out from among the accompaniment

- All these can be utilized in melody/vocals extraction

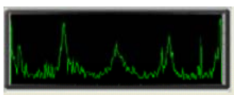
Approaches to vocals extraction

1. *Track the pitch* of melody (and select corresponding spectral components) [Goto-04, Paiva-05, Fujihara-07, [Mesaros-07, Li&Wang-06, Virtanen-08]



[Virtanen-08]

2. Train two *timbre models*, one for vocals and one for the accompaniment, and use these to pull out the vocal components [Ozerov-05, Durrieu-10]
3. Use *stereo information* to pick a source at certain angle of arrival [Barry-2004]

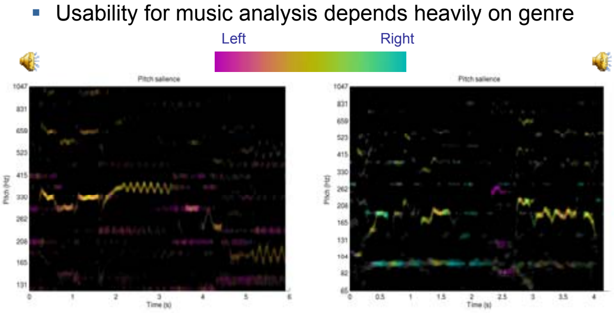


[www.audioresearchgroup.com]

4. Data-driven [Poliner-06]

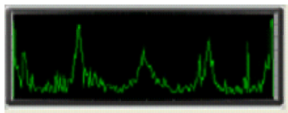
Stereo information

- *Stereo* info can be used to pick a source at a certain angle
- Spatial information is important for human scene analysis
- Usability for music analysis depends heavily on genre



Stereo information

- For an example method, see [Barry-2004]
- Select spectrogram components based on their interaural intensity difference (amplitude difference in the left- and right-channel spectrogram)



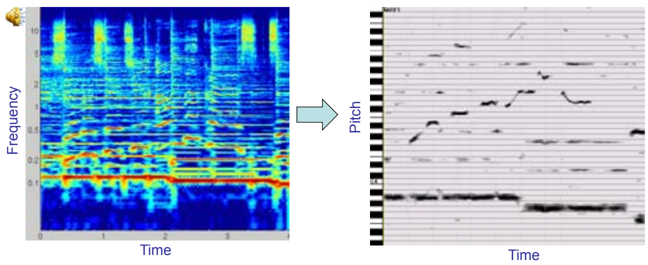
www.audioresearchgroup.com → Demos

Timbre models

- Consider, for example [Durrieu-TASLP-10]:
  - Input power spectrogram is modeled as the sum of the leading voice and the accompaniment
  - source-filter model for vocals, implemented in the statistical framework of mixture models
  - model for the accompaniment derived from non-negative matrix factorization
  - Pitch obtained as a side-information
  - Results highly ranked at MIREX'09 (#2 and #3)
- Melody transcribers of Dressler [Dressler-MIREX-09] and Goto [PreFEst-SC-04] utilize timbre too

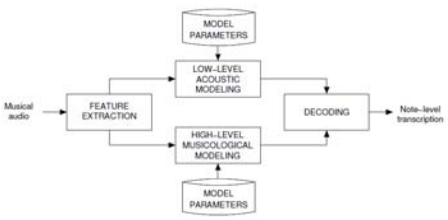
Pitch information

- Pitch content is central for a melody
- Can extract using a multipitch estimator, or by performing mapping from time-frequency to time-pitch [Klapuri-ISMIR-09]



Acoustic modeling

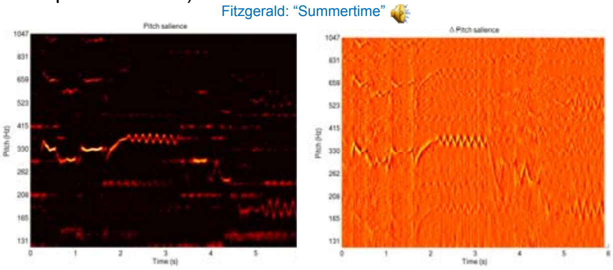
- For acoustic and musicological modeling of melodies, consider as an example the method [Ryynänen-CMJ-08]
- Focus on pitch information: no timbre or stereo features included in the feature vector



Time differential of pitch salience

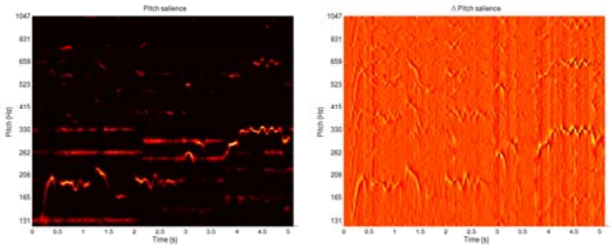
- Take advantage of the fact that vocal pitch is highly time-varying → vocals stand out in ΔSalience
- Stable-pitched instruments filter out (except at the point of onset)

Fitzgerald: "Summertime"



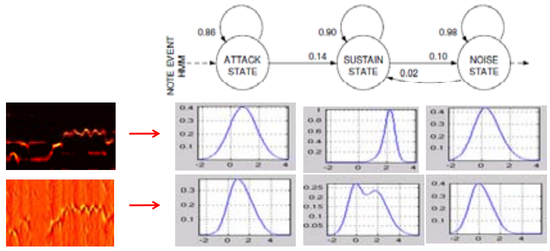
Time differential of pitch salience

The Beatles: "Let It Be"



Hidden Markov model for acoustic features

1. Extract frame-wise acoustic features:
   pitch salience, Δsalience, onset accent (not shown)
2. Use training data (RWC Pop with time-aligned audio and MIDI) to learn HMM parameters for note events

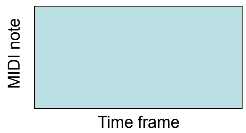


Side-note about HMMs

- Cambridge hidden Markov model toolbox (HTK)
  - excellent toolbox for training and using HMMs
  - well-documented, convenient to use, and supports cutting-edge stuff
  - (althought was not used for the described system)

Acoustic model for melody versus background models

- Separate models trained for
  - melody notes
  - bass notes
  - other instruments' notes
  - silence/noise
- In the time-pitch plane, each MIDI note in each frame must be classified into one of the above categories



Musicological modeling

- Musical context and assumptions about "typical" melodies can be used to resolve otherwise ambiguous situations

Utilizing musical context

- Guess the next note

No context → have to rely on the (often ambiguous) observation only

$P(n_t \mid \mathbf{o})$

Key (scale) information helps to resolve pitch inaccuracies (C vs C#)

$P(n_t \mid \mathbf{o}, k)$

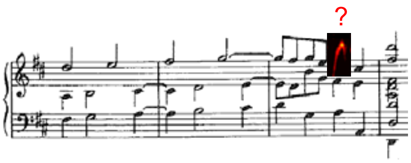
Preceding note helps to remove octave errors and spurious short detections (melodic continuity)

$P(n_t \mid \mathbf{o}, n_{t-1}, k)$

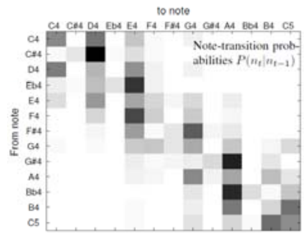
Several preceding notes implicitly encode some of the chord context

$P(n_t \mid \mathbf{o}, n_{t-1}, n_{t-2}, k)$

Utilizing musical context

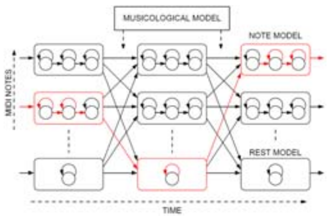
- In principle, the larger the context the better, but in practice, large models are hard to train and use (decode)

Musicological modeling: N-gram models

- N-gram models the probability of the note pitch based on N−1 previous note pitches: $P(n_t \mid n_{1:t-1}) \approx P(n_t \mid n_{n-N+1:t-1})$
- Figure: Key-dependent note bigram probabilities for C major / A minor key pair $P(n_t \mid n_{t-1}, k)$
- Bigram probabilities estimated as
  $P(n_t \mid n_{t-1})$
  $= P(n_t, n_{t-1}) / P(n_{t-1})$
  $\approx Cnt(n_t, n_{t-1}) / Cnt(n_{t-1})$
- Smoothing is needed to avoid zero probabilities (e.g. Witten-Bell)

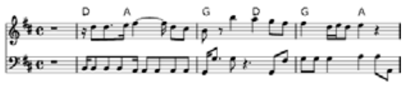
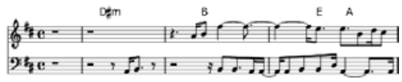
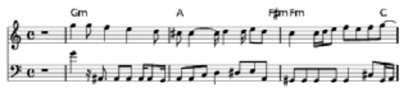


Combining acoustic and musicological models

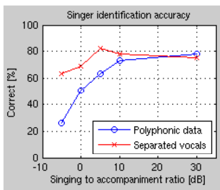
- Hierarchical HMM is an option widely used in speech rec.
- Musicological model operates at a higher level, assigning probabilities for transitions between note events
- Task: find the most probable path given observed data and the model parameters → Viterbi algorithm



Transcription examples

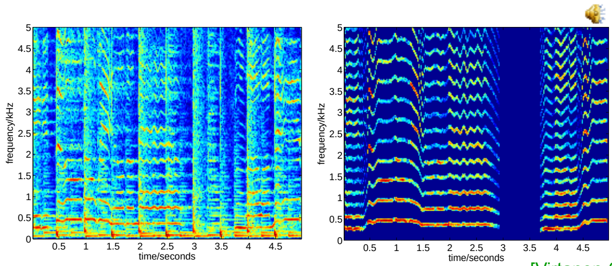
- RWC pop 70

- RWC pop 38

- RWC pop 12

Vocals separation

- Vocals carry a lot of meaning besides the pitch contour
  - lyrics
  - identity of the singer
  - vocal timbre characteristics
  - musical and emotional expression
- Analysis becomes easier if vocals can be separated from the rest
- Figure: singer identification in polyphonic music with/without vocals separation [Mesaros-2007]



Vocals separation based on melody pitch

- Binary masking: estimate pitch and then predict time-frequency points where vocals are present



[Virtanen-08]

Overlap in time-frequency

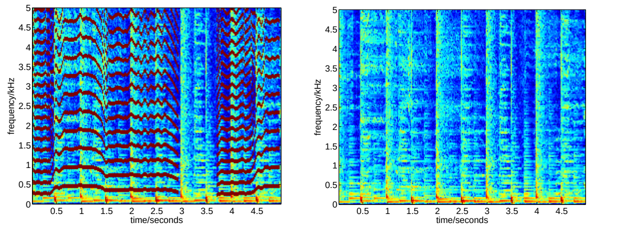
- The above methods assign all the energy at harmonic frequencies to vocals
- When sounds overlap in time and frequency, separation quality degrades
  - Consonant musical intervals cause partials of different instruments to overlap
  - Wideband percussive sources

Estimation and removal of accompaniment

- Learn model for the accompaniment using the non-vocal regions and a binary-weighted NMF [Virtanen-08]

→ Predict & subtract the accompaniment from vocal regions

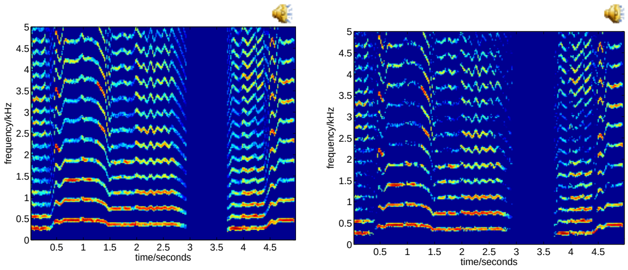
- Some similarity with the approach of [Durrieu-TASLP-10]



[Virtanen-08]

Effect of removing the accompaniment

- Left: vocals obtained using binary masking only
- Right: vocals after subtracting the accompaniment

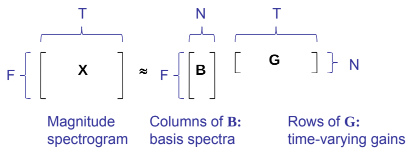


[Virtanen-08]

Using non-negative matrix factorization as a background model

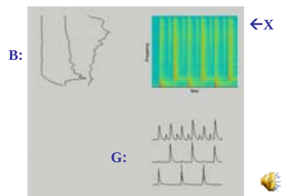
- Signal model

$$\mathbf{X} \approx \mathbf{BG}$$



→ NMF represents matrix as a product of two lower-rank matrices

- Figure: NMF for drum track spectrogram



How many NMF components are needed to represent the accompaniment?

- In these examples, magnitude spectrograms of music are factorized with NMF and original phases are used for resynthesis

Number of components in factorization

	orig	1	2	4	8	16	32
Drums [Weckl]							
Classical [Vivaldi]							
Rock [Santana]							
Rock [U2]							
Bass [Laboriel]							

Applications of melody and vocals extraction

- Karaoke
- Music-oriented games
- Replace vocals on an existing recording with user input
- Alignment of textual lyrics with audio
- Singer identification
- Query by humming

Conclusions

- Melody and lead vocals are a central part of many music types
- Vocal melodies have acoustic and musical characteristics that can be modeled meaningfully
- Utilization of musical context improves the robustness of analysis considerably
- Vocals separation can be done to a reasonable degree, and by using various different approaches