Music Information Retrieval:
Part 2
Feature Extraction, Evaluation, Applications

http://www.ifs.tuwien.ac.at/mir

Alexander Schindler
Research Assistant
Institute of Software Technology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~schindler

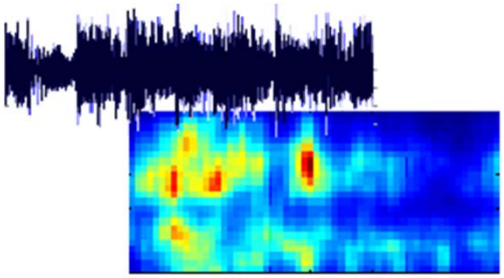
Outline

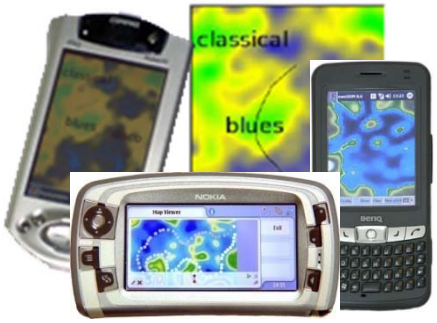
- 1. Recap Music IR Part 1
- 2. Feature Extraction Algorithms
- 3. Feature Extraction Tools
- 4. Benchmarking in Music IR
- 5. MIR Research at IFS

1.
Recap Part 1

Recap Part 1

- Sound as acoustic wave
- Representation of sound in digital formats
- Other representation of music
- **What is Music IR?**
- Introduction of Audio Features
- Music Clustering: Music Maps
- Combining Audio + Lyrics
- Audio Segmentation
- Chord Detection
- Source Separation





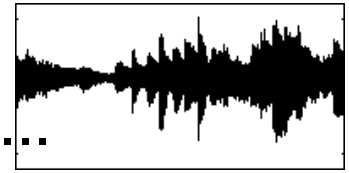
What is Music?

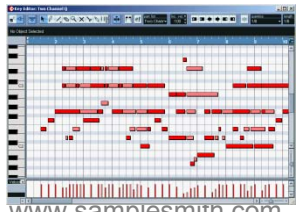
- Music

  Audio: wav, au, mp3, ...

  Symbolic: MIDI, mod, ...

  Scores: Scan, MusicXML





www.samplesmith.com

www.westminster.gov.uk

- Text
  - Song lyrics
  - Artist Biographies
  - Websites: Fanpages, Blogs, Album Reviews, Genre descriptions
- Community data
  - Market basket
  - Tags
  - Social Networks
    - Spotify
    - Last.fm
- Video/Images
  - Album covers
  - Music videos

What is Music IR?

- Searching for Music
  - Searching for music on the Web
  - Query by Humming
  - Similarity Retrieval
  - Identity detection (fingerprinting)

- Extraction of information from music

  → plenty of other tasks!

Music IR – Tasks

- Genre classification
- Mood classification
- Music Recommendation
- Artist identification
- Artist similarity
- Cover song detection
- Rhythm and beat detection
- Score following
- Chord detection
- Organization of music
- Audio Fingerprinting
- Audio segmentation
- Instrument detection
- Automatic source separation
- Onset detection
- Optical music recognition
- Melody transcription …..

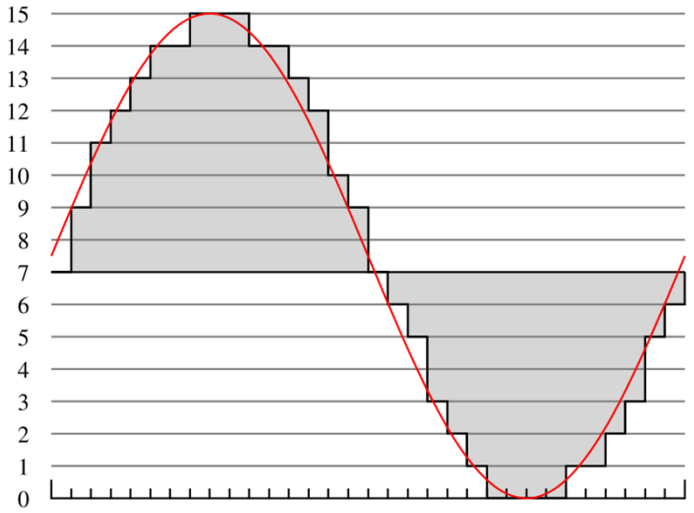
2.
Feature Extraction from Music

Feature Extraction from Music

- Purpose?
  - Information Retrieval! (see tasks before)
- Text
  - bag of words
  - n-grams
  - Phrases
  - POS
  - ...
- Music: ??
- Challenge: too much audio data

Too much Audio Data

- Digital Audio
  - Sampling Rate: 44,100 Hz
  - 16-bit resolution for each channel
    - 2 channels for stereo
  - 88,200 Integers per second



Excercise: Find Documents Containing the Word „Music“

Document 1:

*“Most of these issues stem from the commercial interest in **music** by record labels, and therefore imposed rigid copyright issues, that prevent researchers from sharing their **music** collections with others. Subsequently, only a limited number of data sets has risen to a pseudo benchmark level, i.e. where most of the researchers in the field have access to the same collection.“*

Document 2:

*“The Echonest Analyzer [5] is a **music** audio analysis tool available as a free Web service accessible over the Echonest API and as a commercially distributed standalone command line tool. The Analyzer implements an onset detector which is used for segmentation.”*

Document 3:

*“The Million Song Dataset (MSD), a collection of one million **music** pieces, enables a new era of research of **Music** Information Retrieval methods for large-scale applications. It comes as a collection of meta-data such as the song names, artists and albums, together with a set of features extracted with the The Echo Nest services, such as loudness, tempo, and MFCC-like features.”*

Excercise: Find Songs with Strings

Song 1:

83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98

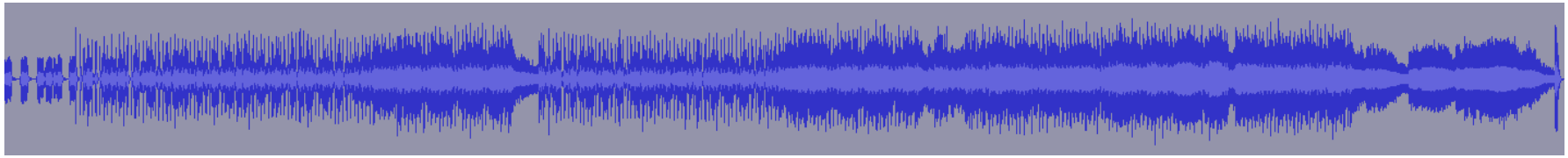
Song 2:

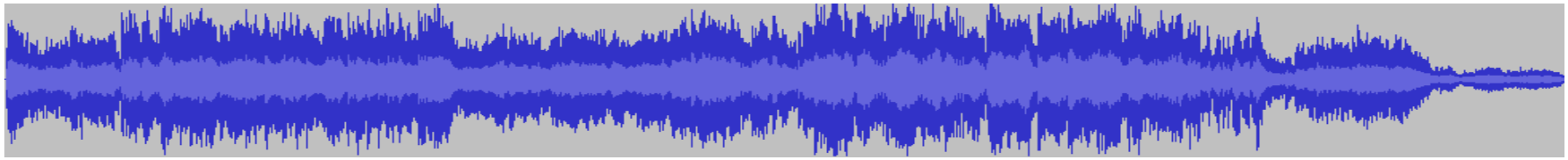
55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7

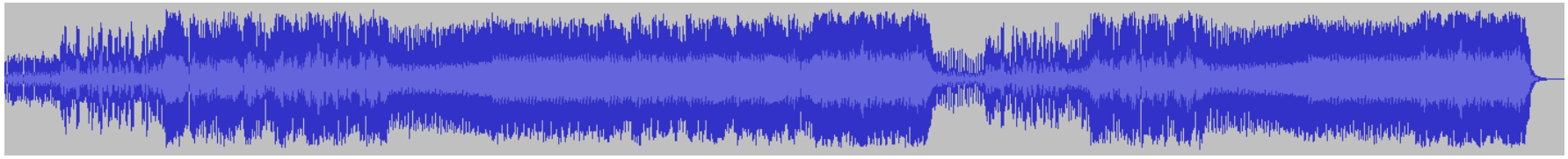
Song 3:

66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2

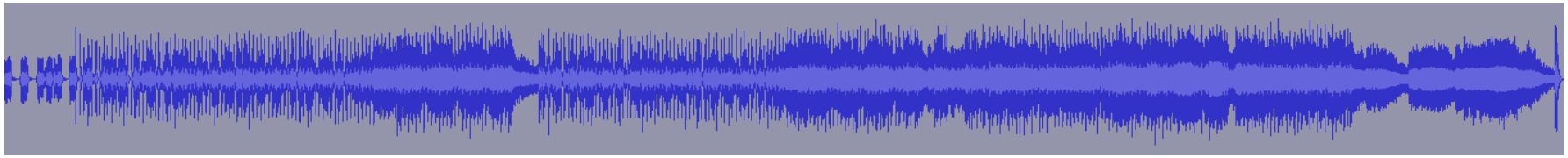
Excercise: Same Genre?



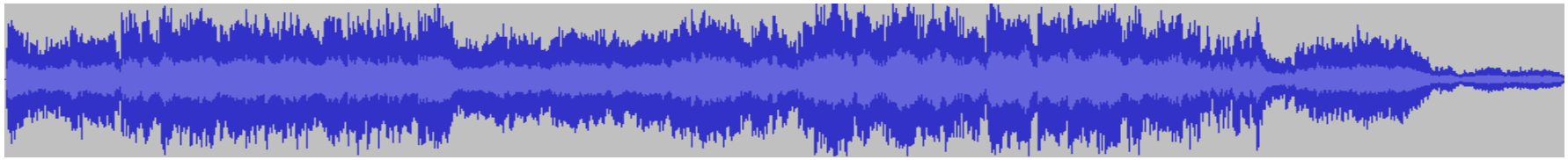




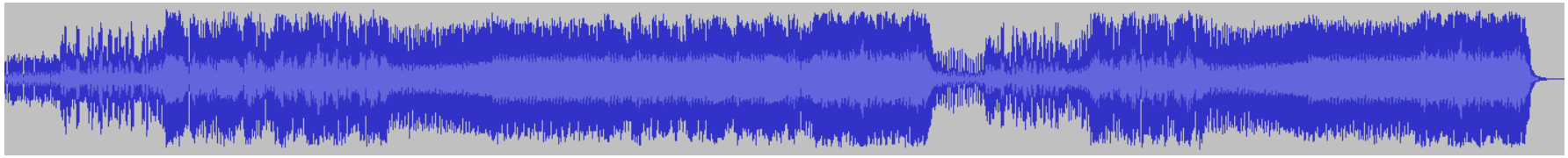
Excercise: Identify Songs



AC-DC – Highway to Hell



John Williams – Star Wars Main Theme

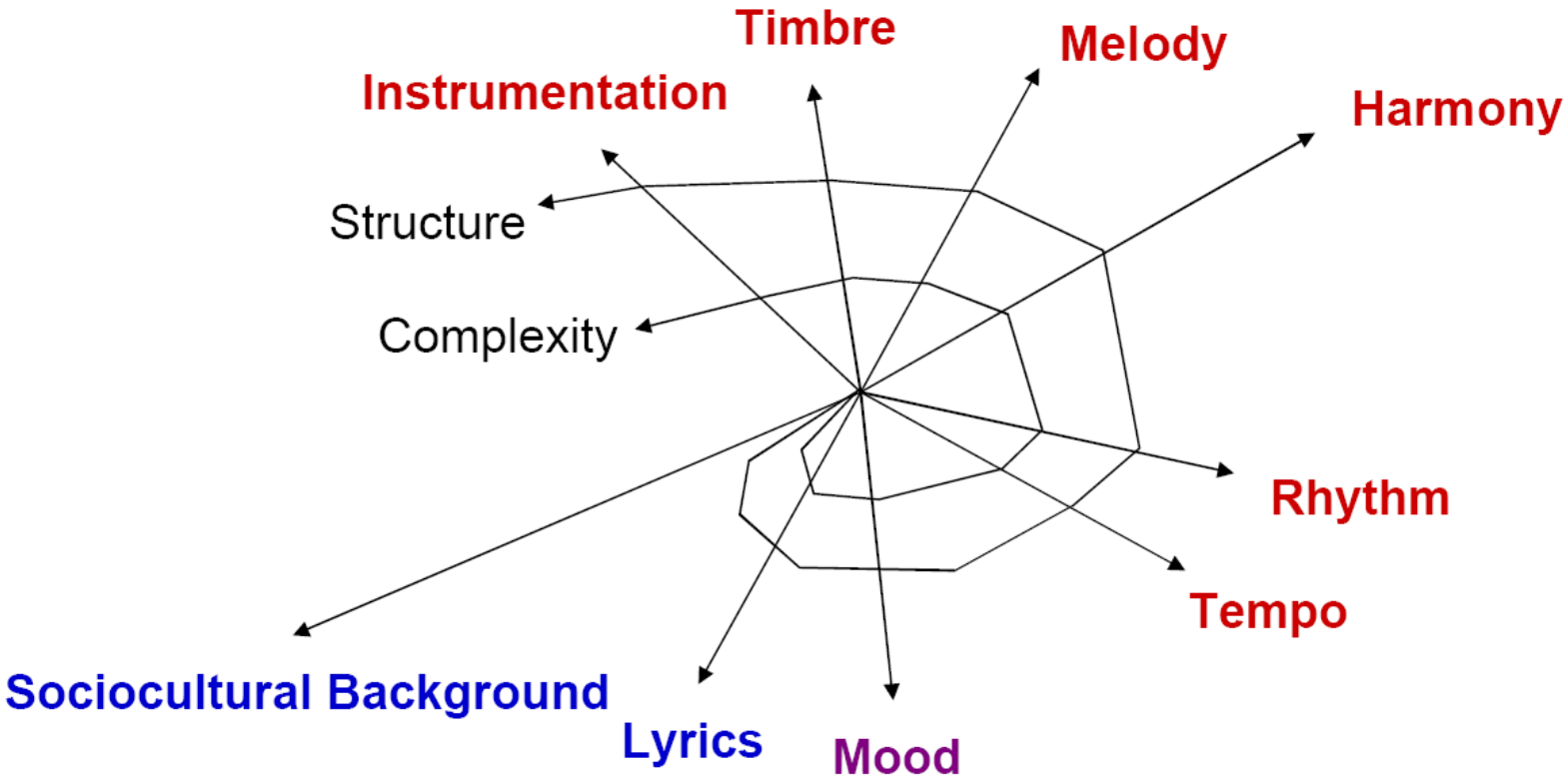


Rihanna feat. Calvin Harris – We Found Love

Audio Feature Extraction

- Reduce audio data by extracting information about:
  - Pitch
  - Timbre
  - Rhythm
  - etc.
- → extract „audio descriptors“

Dimensions of Music

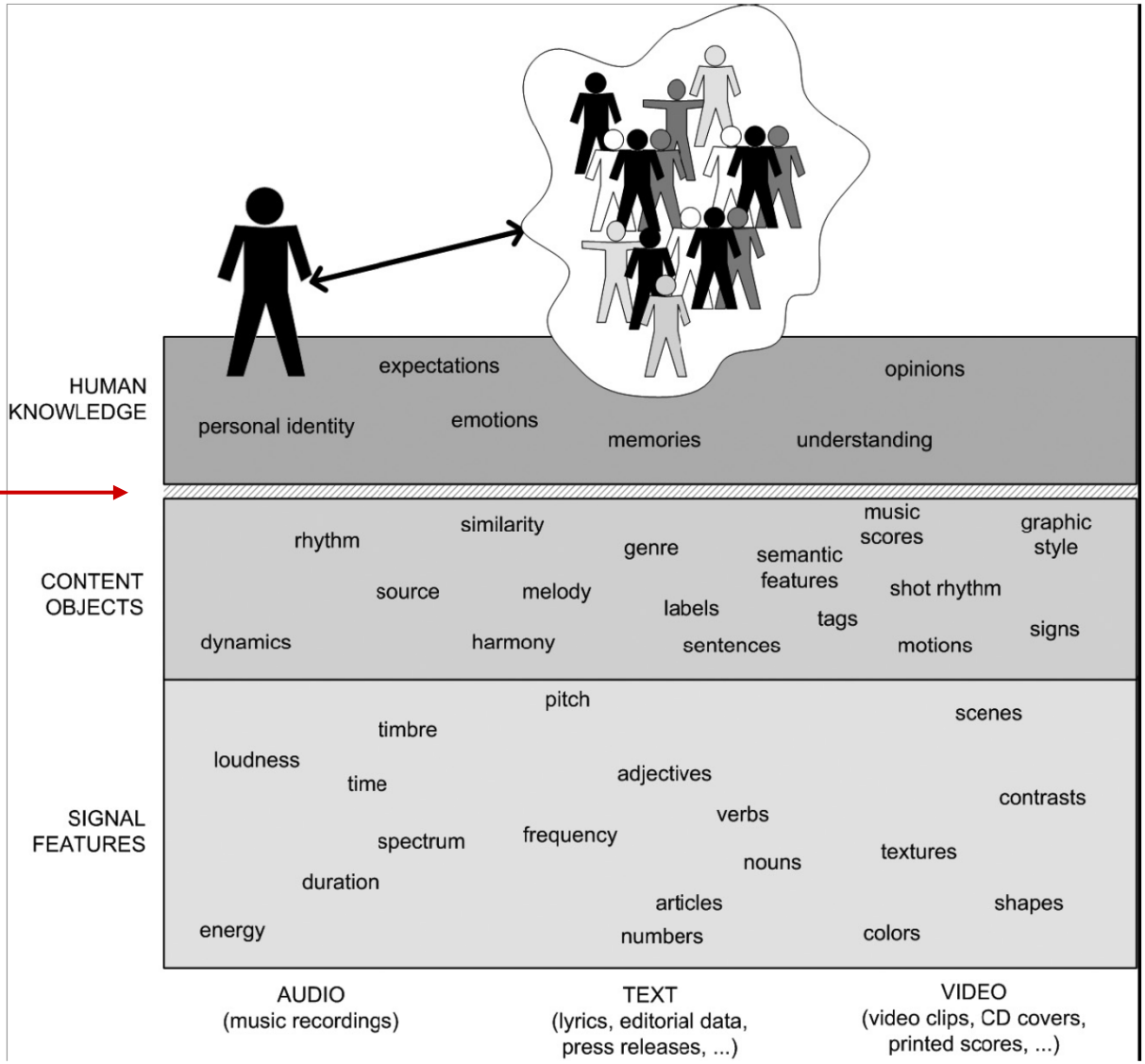


Matching users and items

users

semantic gap

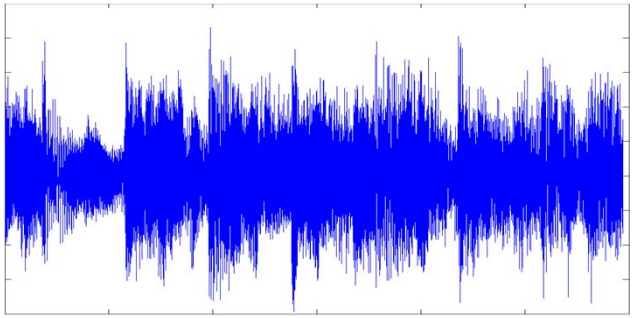
items

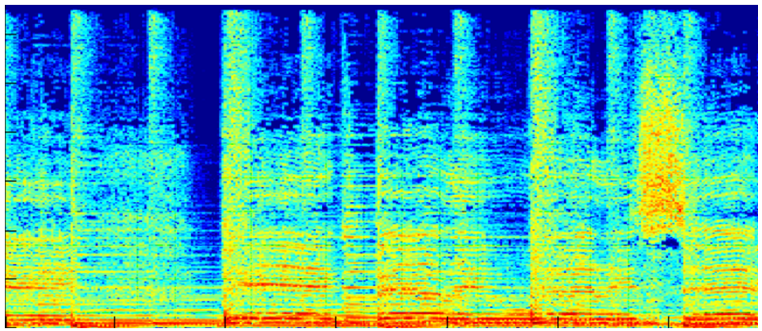


Audio Features

- „Low-level“ features (Zero Crossings, RMS Energy, Spectral Centroid, Rolloff, Flux, ...)
- MPEG-7 descriptors (temporal, spectral, timbral, ...)
- Mel-frequency Cepstral Coefficients (MFCCs)
- MARSYAS Features (Spectral, MPEG-compression-based, Wavelet-based, Beat and Pitch Histograms)
- Rhythm Patterns, Statistical Spectrum Descriptors, Rhythm Histograms

Signal Processing





Time Domain
(„Wave Form“)

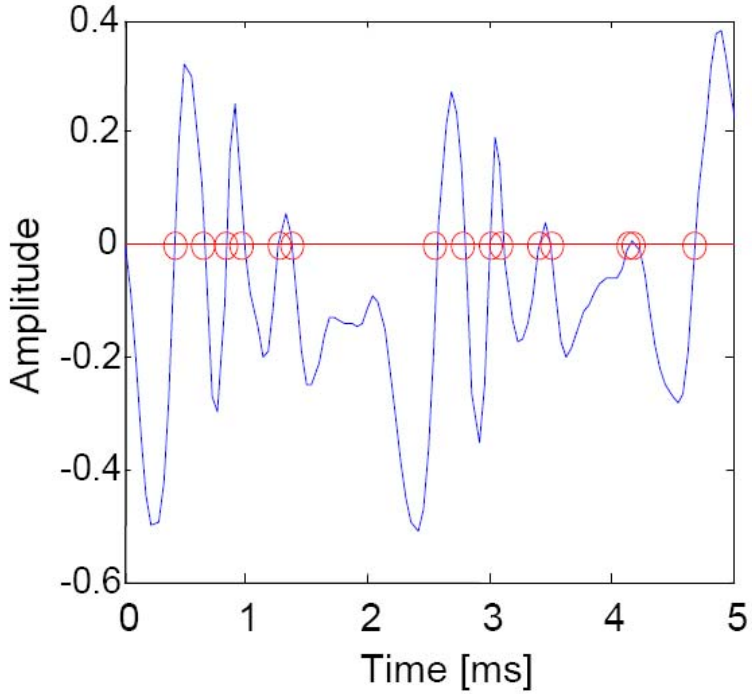
Frequency Domain
(„Spectrum“)

Time-Frequency Transformation

Fourier Transform (FFT)
Discrete Cosine Transform (DCT)
Wavelet Transform

Time-Domain Features

Zero Crossing Rate (ZCR) = 3/ms



= 15 / 5ms

measures noisiness

"crossing zero" is defined as:

(x[n-1] < 0 and x[n] > 0)

or (x[n-1] > 0 and x[n] < 0)

or (x[n-1] ≠ 0 and x[n] = 0).

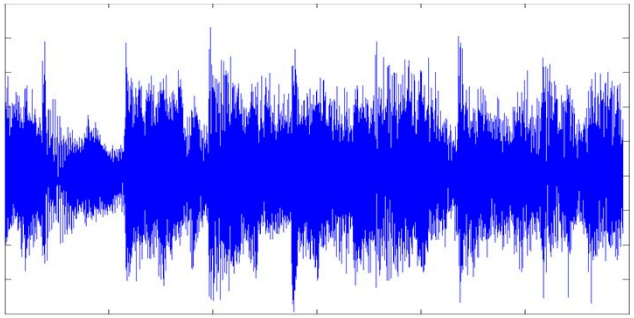
Time-Domain Features

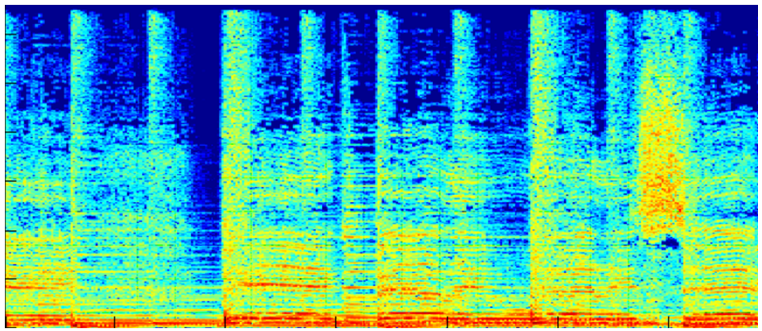
Root Mean Square (RMS)

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

indication of loudness

Signal Processing





Time Domain
(„Wave Form“)

Frequency Domain
(„Spectrum“)

Time-Frequency Transformation

Fourier Transform (FFT)
Discrete Cosine Transform (DCT)
Wavelet Transform

Fast Fourier Transform (FFT)

- Fourier Transform:

$$f_j = \sum_{k=0}^{2n-1} x_k \, e^{-\frac{2\pi i}{2n} jk} \qquad j = 0, \ldots, 2n-1.$$

- Fast Fourier Transform (FFT):
  - efficient algorithm to compute the discrete Fourier transform (DFT)
  - divide and conquer algorithm
  - O(N log N) instead of $O(N^2)$
  - N must be a power of 2

Spectra

- Magnitude Spectrum vs. Power Spectrum
  - the power spectrum is the magnitude spectrum **squared** (calculated by, for each bin, by summing the square of the imaginary output of the FFT with the square of the real value)
  - magnitude spectrum and power spectrum rarely used directly as features (too much raw information)
  - many spectral features are derived from either the power spectrum or the magnitude spectrum

Spectral Features

- Spectral Centroid
  - center of gravity (balancing point of the spectrum)
  - gives an indication of how “dark” or “bright” a sound is

$$SC = \frac{\sum_{n=1}^{N} P_t[n] * n}{\sum_{n=1}^{N} P_t[n]}$$

$P_t[n]$ … n^{th} frequency bin of **power spectrum** (with N bins)
t … timeframe

Cory McKay, Automatic Music Classification with jMIR (PhD thesis, 2010)

Spectral Features

- Spectral Rolloff
  - the frequency below which some fraction, k (typically 0.85, 0.9 or 0.95 percentile), of the cumulative spectral power resides
  - measure of the skewness of the spectral shape
  - indication of how much energy is in the lower frequencies

$$\sum_{n=1}^{SR_t} P_t[n] = k \sum_{n=1}^{N} P_t[n]$$

$P_t[n]$ … n^{th} frequency bin of **power spectrum** (with N bins)
t … timeframe

Cory McKay, Automatic Music Classification with jMIR (PhD thesis, 2010)

Spectral Features

- Spectral Flux
  - squared differences in frequency distribution of two successive time frames
  - measures the rate of local change in the spectrum

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

computed from the **normalized magnitude spectrum** $N_t[n]$

Spectral Features

- Spectral Variability:
  - standard deviation of the bin values of the magnitude spectrum
  - provides an indication of how flat the spectrum is and if some frequency regions are much more prominent than others
- Strongest Partial:
  - center frequency of the bin of the magnitude or power spectrum with the greatest strength
  - can provide a primitive form of pitch tracking
- and others …

MPEG7 Features

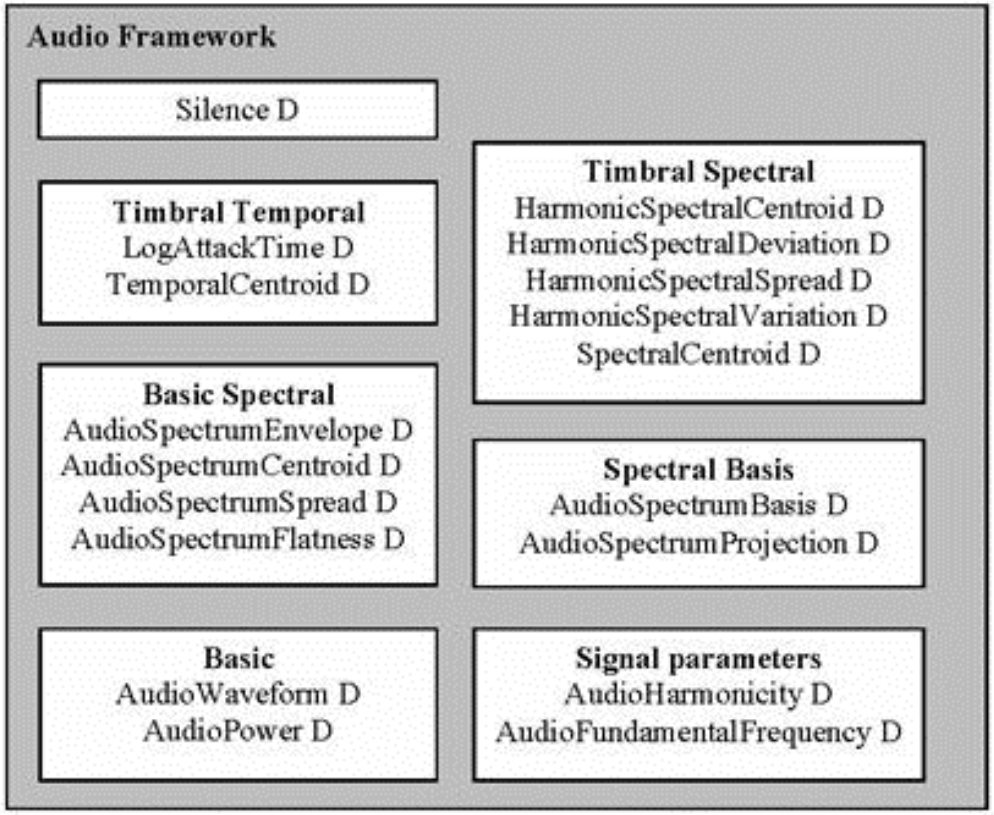
- **"Multimedia Content Description Interface“**
- ISO/IEC standard by MPEG (Moving Picture Experts Group)
- Providing meta-data for multimedia
- MPEG-1, -2, -4: make content available
- MPEG-7: makes content accessible, retrievable, filterable, manageable (via device / computer).

- Details:
  ISO/IEC JTC1/SC29/WG11N6828; editor:José M. Martínez
  Palma de Mallorca, Oct. 2004, MPEG-7 Overview (version 10)
  http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

MPEG7 Features

- Low-level descriptors
  - spectral, parametric, and temporal features of a signal
- High-level description tools:
  specific to a set of applications
  - general sound recognition and indexing
  - instrumental timbre
  - spoken content
  - audio signature description scheme
  - melodic description tools to facilitate query-by-humming

MPEG7 Features



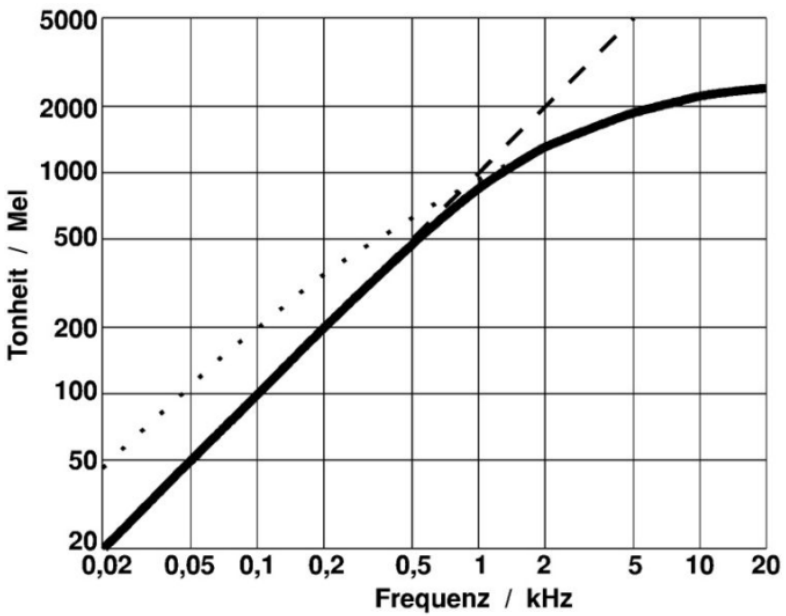
MFCC Features

Mel-Frequency Cepstral Coefficients (MFCC)

- used previously in speech recognition
- model human auditory response (Mel scale)
- „cepstrum“ (s-p-e-c reversed): result of taking the Fourier transform (FFT) of the decibel spectrum as if it were a signal
- show rate of change in the different spectrum bands
- good timbre feature

MFCC: Mel Scale

- perceptually motivated
- Mel comes from the word *melody* to indicate that the scale is based on pitch comparisons
- reference point:
  1000 Hz tone, 40 dB above listener's threshold = 1000 Mels

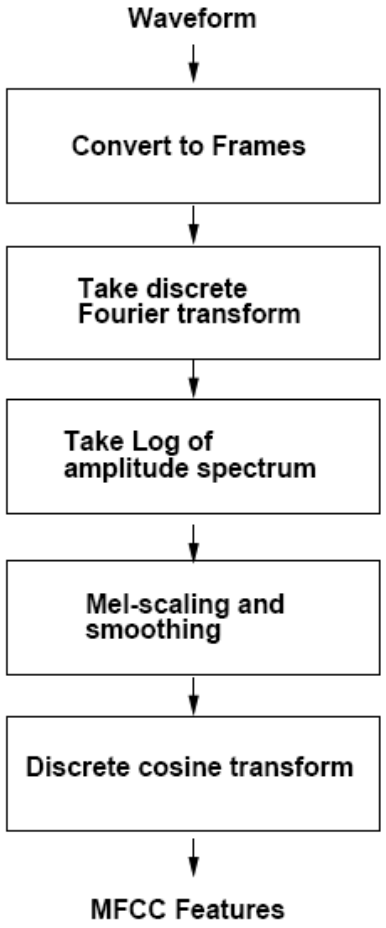


Mel Scale:
< 500 Hz linear
> 500 Hz non-linear

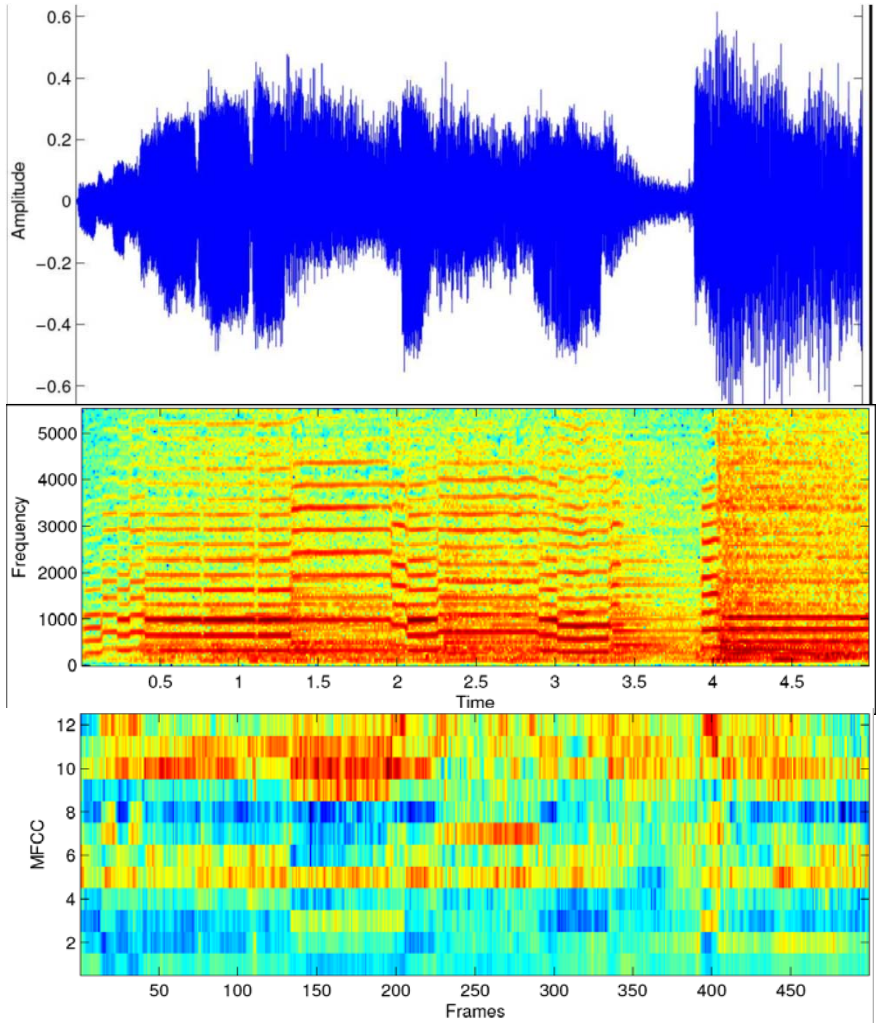
http://en.wikipedia.org/wiki/Mel_scale

MFCC Features

- MFCC feature calculation:
  - Fourier transform of signal window
  - Mapping of frequency bins to Mel scale (using triangular overlapping windows)
  - log of powers at each of the Mel frequencies
  - discrete cosine transform of the list of Mel log powers („as if it were a signal“)
  - MFCCs are the amplitudes of the resulting spectrum



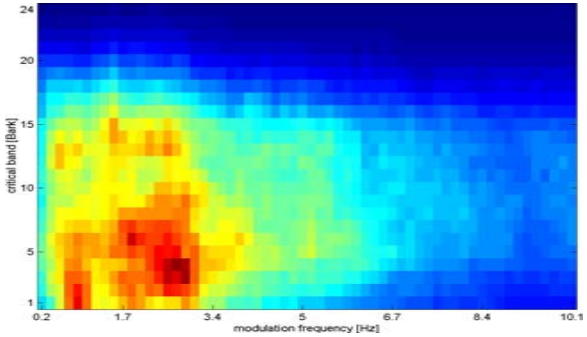
MFCC Features



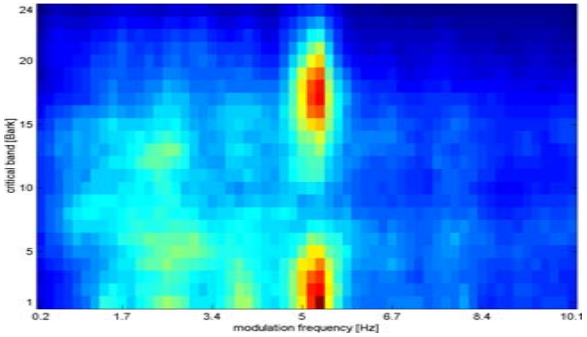
Rhythm Pattern (RP)

- fluctuations on critical frequency bands
  (a.k.a. Fluctuation Pattern)

- covers rhythm in the broad sense

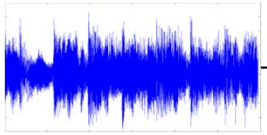


Classical



Rock

Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

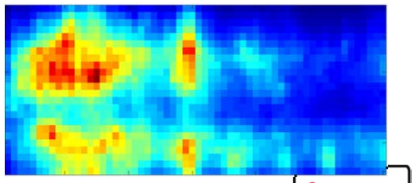
Modulation Amplitudes

Fluctuation Strength

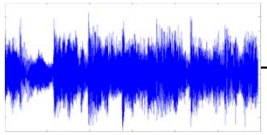
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Rhythm Pattern (RP)



Segmentation (6 s)

- pre-prosessing step

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

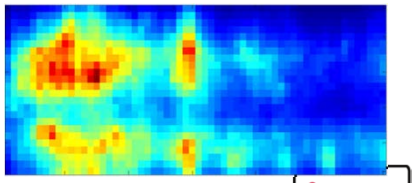
Modulation Amplitudes

Fluctuation Strength

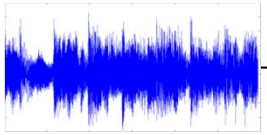
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

- Fourier Transform / Spectral Analysis

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

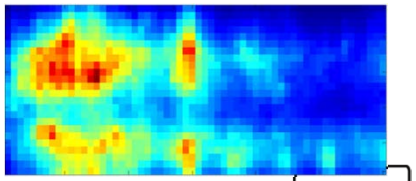
Modulation Amplitudes

Fluctuation Strength

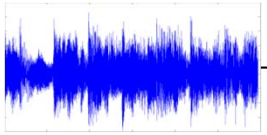
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

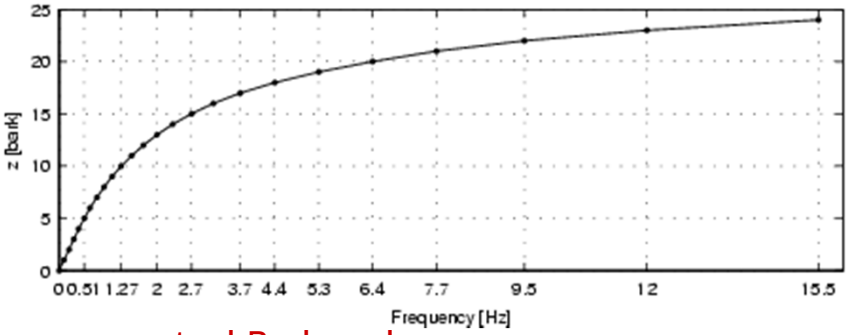
Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone



perceptual Bark scale

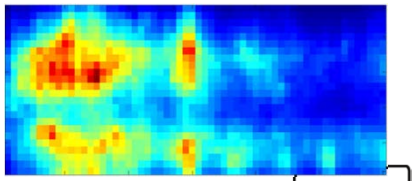
Modulation Amplitudes

Fluctuation Strength

Filtering/Smoothing

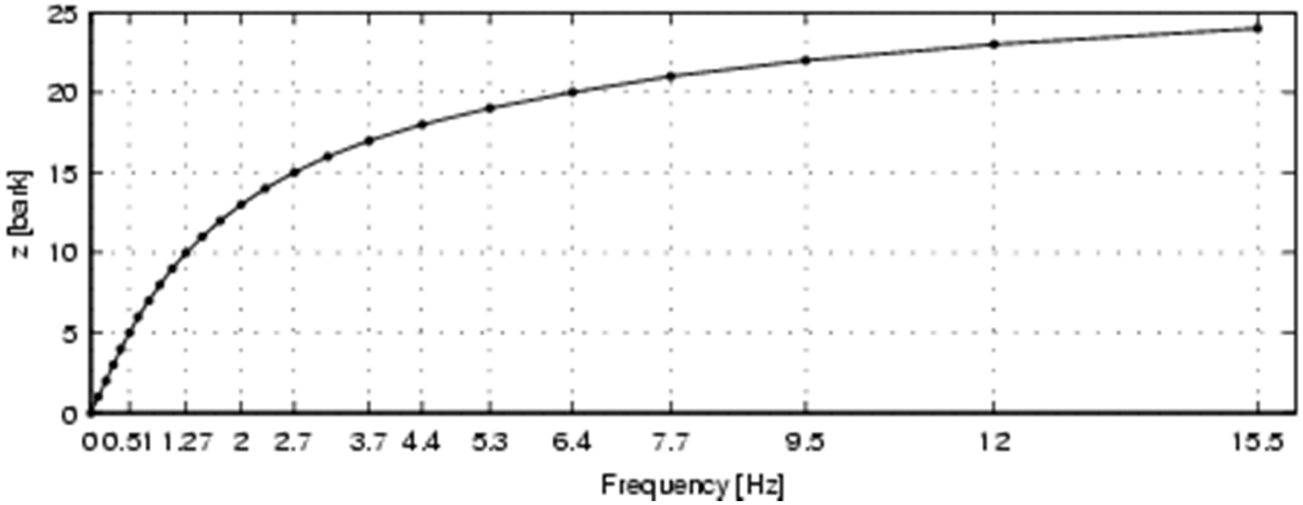
Rhythm Pattern features

(1440 dimensions)



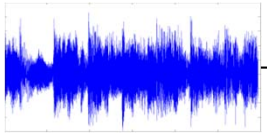
Bark Scale

- psychoacoustical scale (related to Mel scale)
- 24 „critical bands“ of hearing (non-linear)
- proposed by Eberhard Zwicker in 1961



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

Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

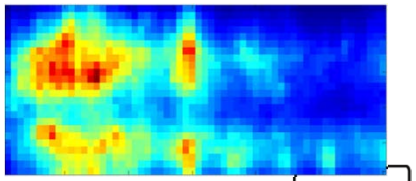
Modulation Amplitudes

Fluctuation Strength

Filtering/Smoothing

Rhythm Pattern features

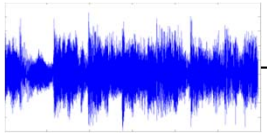
(1440 dimensions)



Spectral Masking

- Occlusion of a quiet sound by a louder sound when both sounds are present simultaneously and have similar frequencies
  - Simultaneous masking: two sounds active simultaneously
  - Post-masking: a sound closely following it (100-200 ms)
  - Pre-masking: a sound preceding it (usually neglected, only measured during about 20ms)
- Spreading function defining the influence of the *j*-th critical band on the *i*-th

Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

$$L_{\mathrm{dB}} = 10 \log_{10} \left(\frac{P_1}{P_0} \right)$$

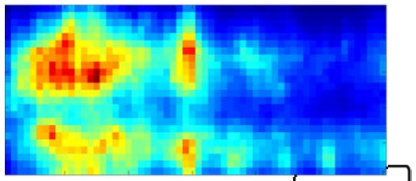
Modulation Amplitudes

Fluctuation Strength

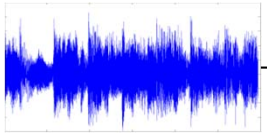
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

equal loudness curves - Phon

Sone

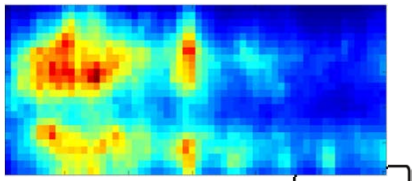
Modulation Amplitudes

Fluctuation Strength

Filtering/Smoothing

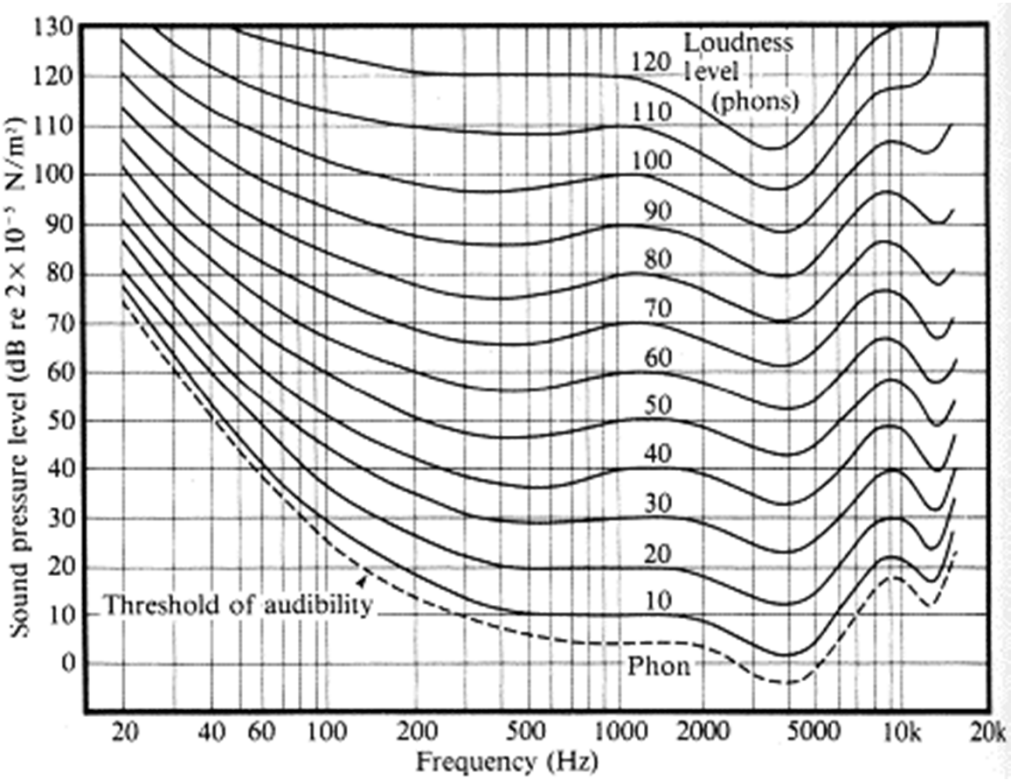
Rhythm Pattern features

(1440 dimensions)



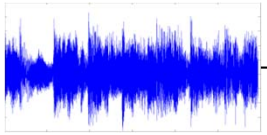
Equal loudness curves (Phon)

- Relationship between sound pressure level in decibel and hearing sensation is not linear
- Perceived loudness depends on frequency of the tone
- equal loudness contours for 3, 20, 40, 60, 80, 100 phon



on-line test: http://www.phys.unsw.edu.au/jw/hearing.html

Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

specific loudness sensation -

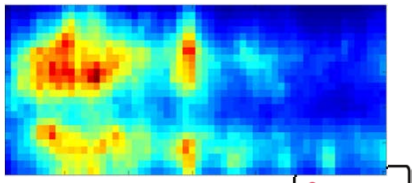
Modulation Amplitudes

Fluctuation Strength

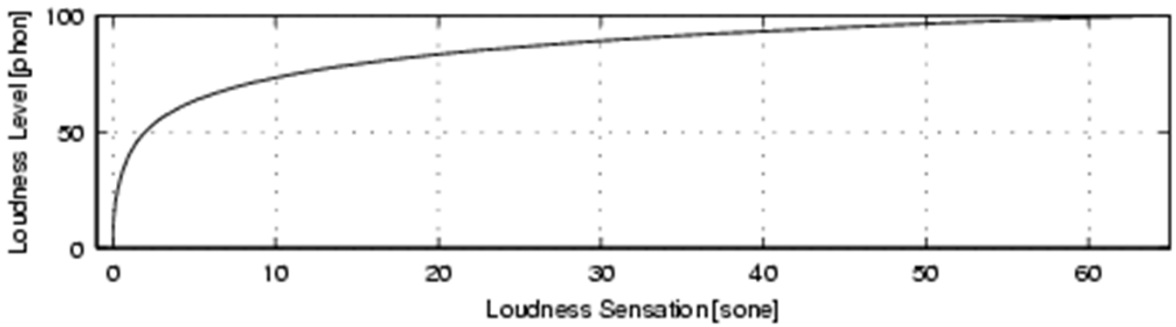
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Sone Transformation



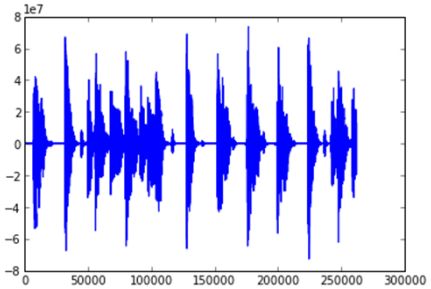
Sone	1	2	4	8	16	32	64
Phon	40	50	60	70	80	90	100

- Perceived loudness measured in Phon does not increase linearly
- Transformation into Sone
- Up to 40 phon slow increase in perceived loudness, then drastic increase
- Higher sensibility for certain loudness differences

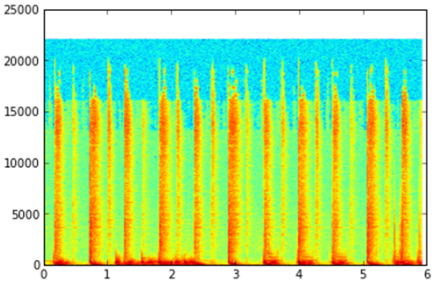
Rhythm Pattern (RP): 2 examples

Queen – Another One Bites The Dust (first 6 seconds)

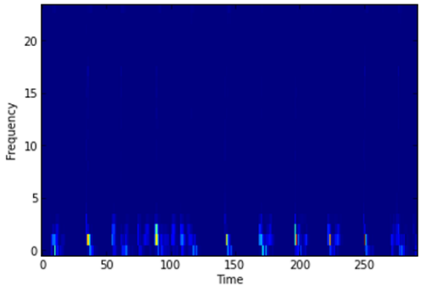
PCM Audio Signal



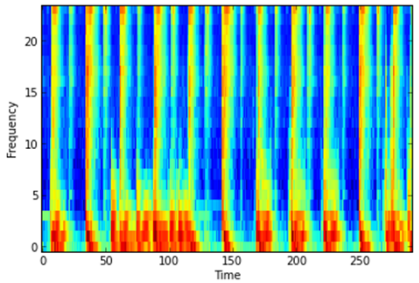
Power Spectrum



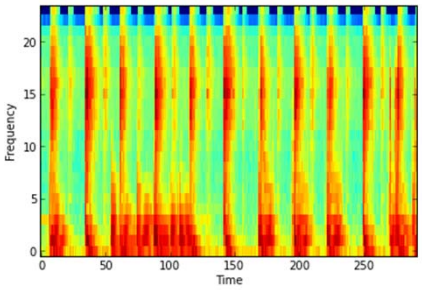
Bark Scale



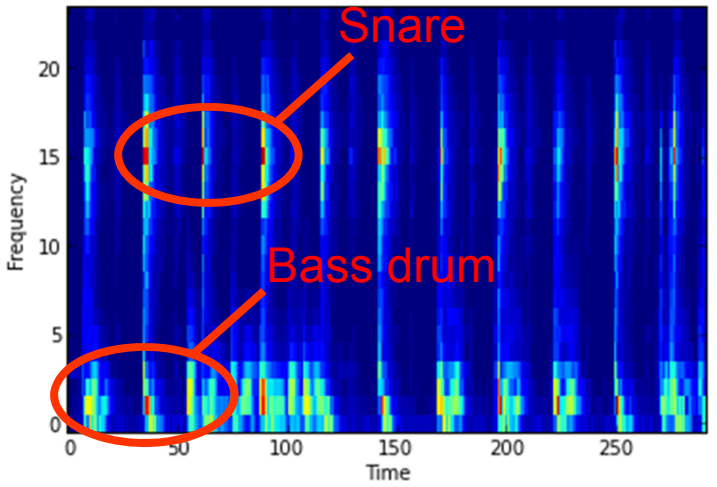
Decibel



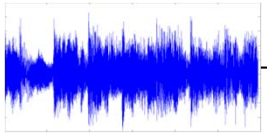
Phon



Sone



Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

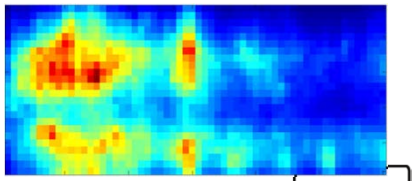
Modulation Amplitudes - 2nd FFT

Fluctuation Strength

Filtering/Smoothing

Rhythm Pattern features

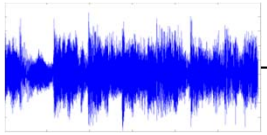
(1440 dimensions)



Amplitude modulation

- Loudness of a critical band usually rises and falls several times
- Periodical pattern, a.k.a. rhythm
- another Fourier Transform retrieves magnitude of modulation for various repetition rates (modulation frequencies) (from 0 to 43Hz) (a.k.a. „cepstrum“)
- A modulation frequency of 43Hz corresponds to almost 2600bpm → cut-off at 10 Hz (600 bpm)
- 60 bins per frequency band

Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

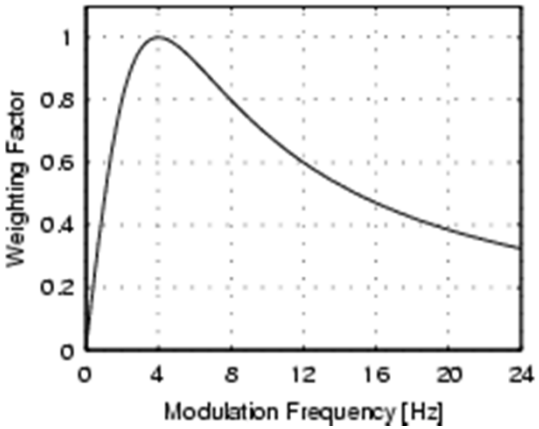
Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone



weighting according to human perception

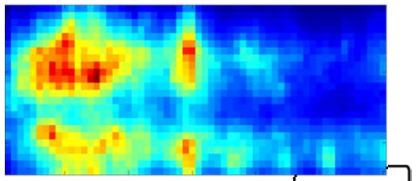
Modulation Amplitudes

Fluctuation Strength

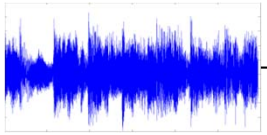
Filtering/Smoothing

Rhythm Pattern features

(1440 dimensions)



Rhythm Pattern (RP)



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

Spectral Masking

Decibel

Phon

Sone

Modulation Amplitudes

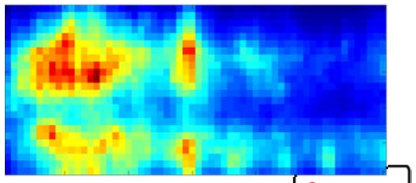
Fluctuation Strength

Filtering/Smoothing

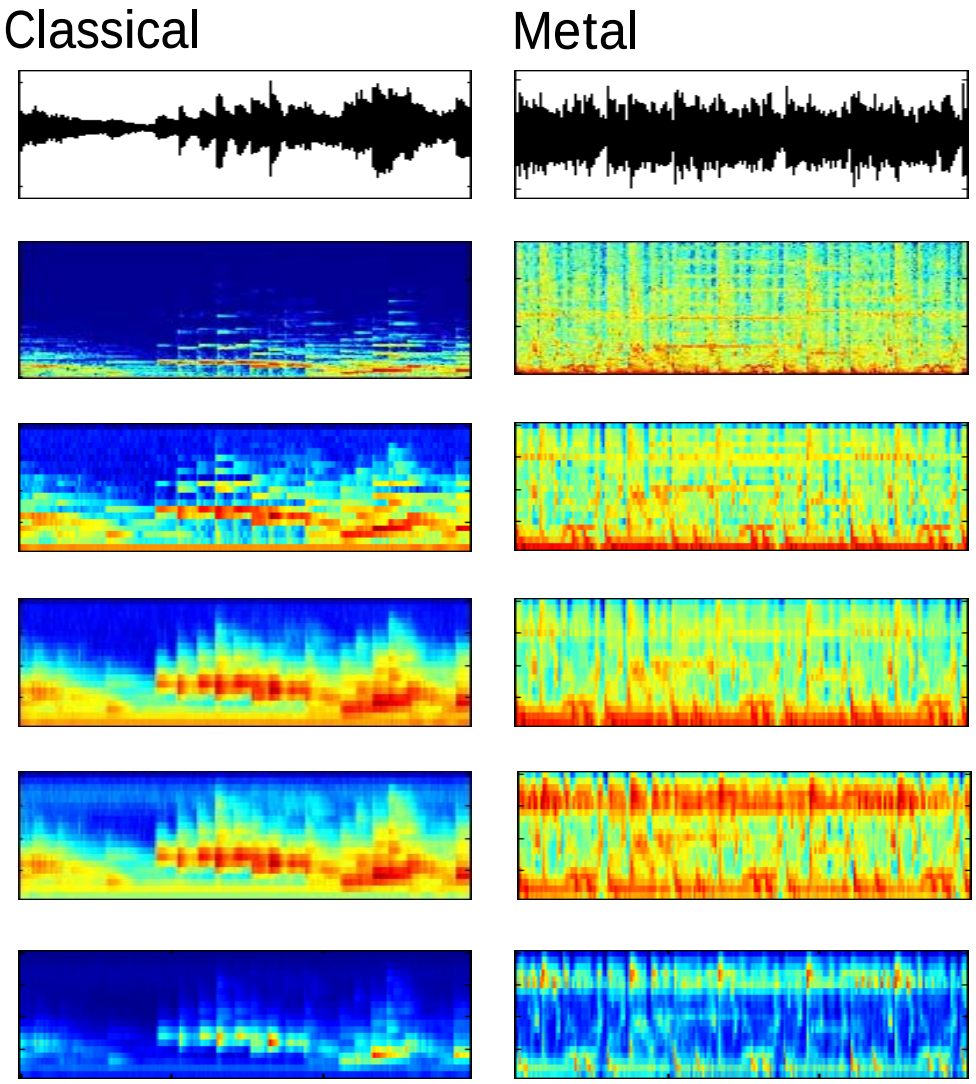
Rhythm Pattern features

(1440 dimensions)

- Gradient filter to emphasize distinctive beats
- Gaussian smoothing to blur slightly



Rhythm Pattern (RP): 2 examples



PCM Audio Signal

Power Spectrum

Frequency Bands

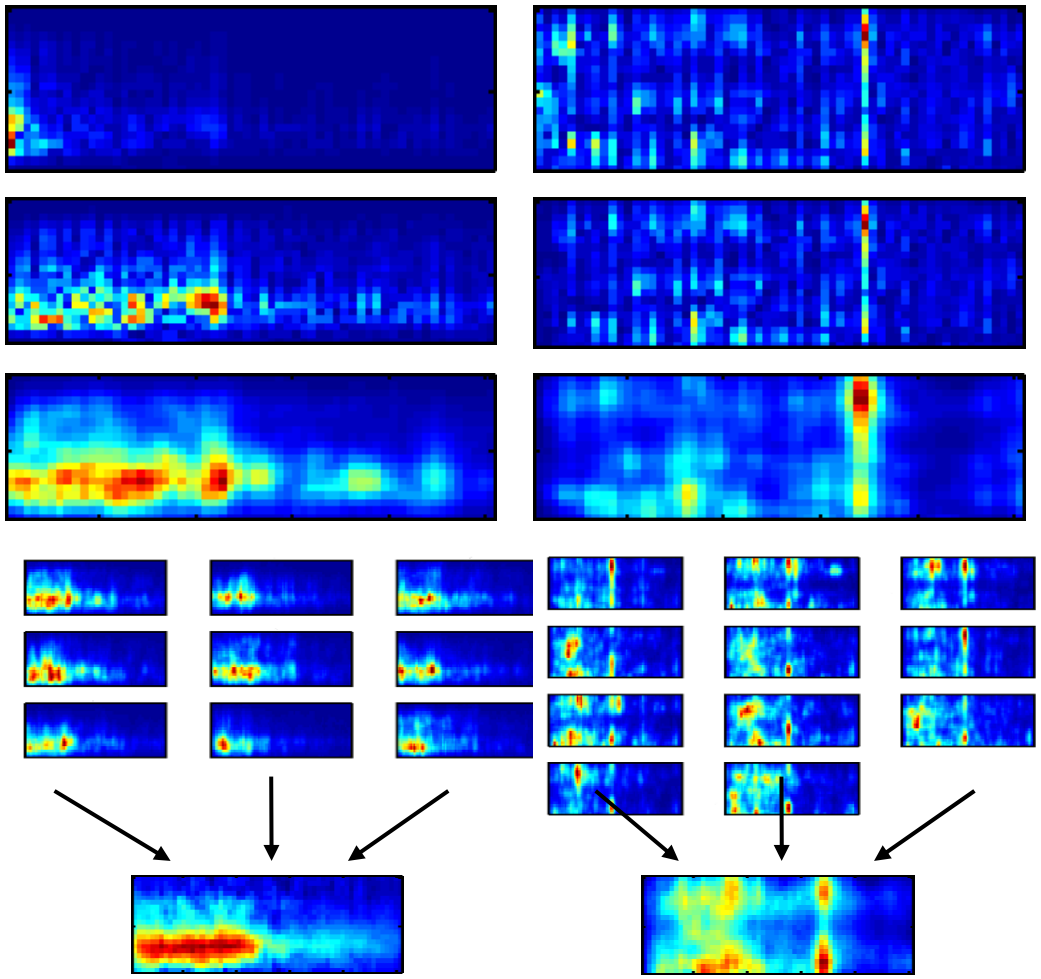
Masking Effects

Phon

Sone

Rhythm Pattern (RP): 2 examples

Classical Metal



modulation amplitude spectrum (“cepstrum”)

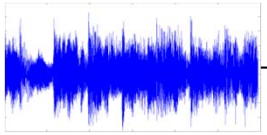
Fluctuation Strength

Filter (Gradient, Gauss)

Median

24*60=
1.440-dim feature vec.

Importance of Psycho-Acoustics



Segmentation (6 s)

Spectrogram (FFT)

Bark scale (24 bands)

problematic → Spectral Masking

crucial → Decibel

crucial → Phon

crucial → Sone

Modulation Amplitudes

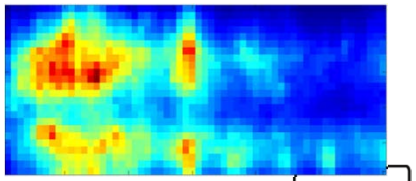
Fluctuation Strength

Filtering/Smoothing

dependent on music data

Rhythm Pattern features

(1440 dimensions)



Lidy and Rauber, ISMIR 2005

Rhythm Pattern
Feature Re-Synthesis

- What do the features really capture?
- It is not
  - Rhythm
  - Pitch/melody
  - Energy
- It is all of the above to some degree:
  complex rhythmic/fluctuation patterns
- Re-synthesis


original re-synthesised

hiphop

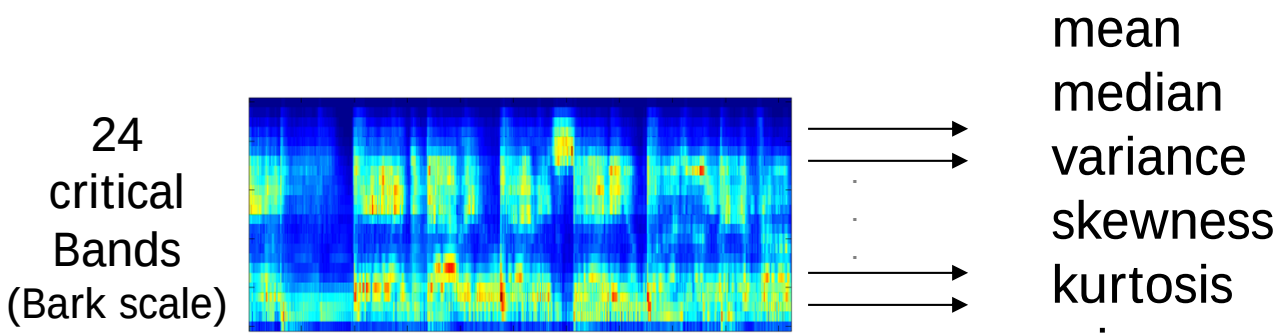

original re-synthesised

reggae

SSD Features

Statistical Spectrum Descriptor (SSD):

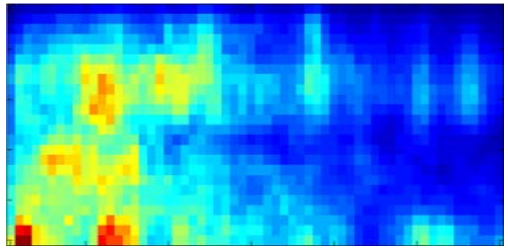
- description of each of the 24 critical bands of the Sonogram by 7 statistical measures
- 168 feature attributes (24x7)



Bark-scale Sonogram (after Sone Step of RP)

Rhythm Histogram (RH)

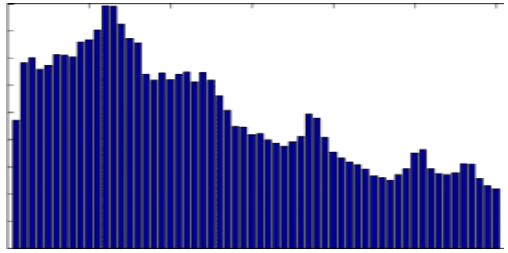
critical bands



modulation frequency

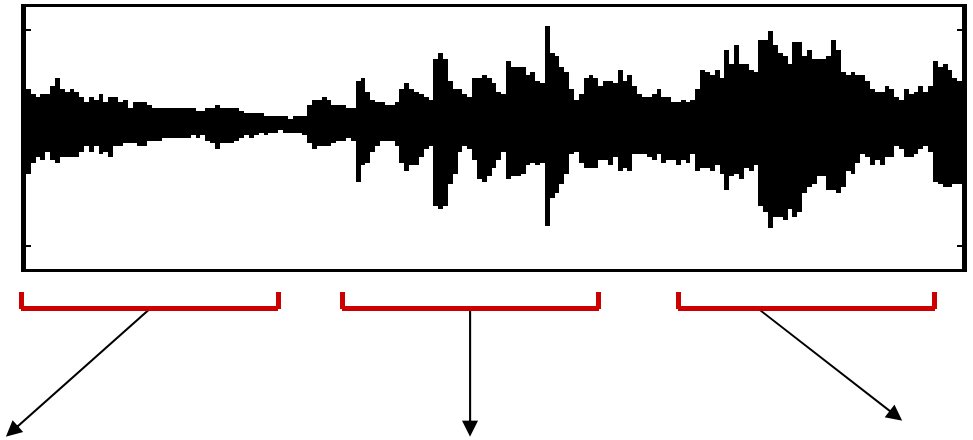
Modulation frequency „cepstrum“
(after 2nd FFT of RP)

aggregation
(before weighting and blurring)



- histogram of modulation magnitude per modulation frequency
- 60 bins -> 60 feature attributes

Time Aspect in Features



- features usually extracted from segments in time
- finally summarized by aggregation (mean, median)
- temporal aspects and information about variations lost
- why not measure these variations?

Temporal Features

- TSSD Features:
  - Extraction of multiple SSD features over time (various 6 seconds segments)
  - Statistical measures* of **changes** of SSD feature attributes (for each of the 168) **over time**
  - → 1176 attributes in final feature set
- TRH Features:
  - analogously for RH features
  - → 420 attributes in final feature set

*
mean
median
variance
skewness
kurtosis
min
max

Comparison of Feature Sets

	3 benchmark music collections		
	GTZAN	ISMIRrhythm	ISMIRgenre
RP (1440 dim)	64.4	**82.8**	75.0
SSD (168 dim)	**72.7**	54.7	**78.5**
RH (60 dim)	44.1	79.9	63.2
RP+SSD	72.3	83.5	**80.3**
RP+RH	64.2	83.7	75.5
SSD+RH	**74.9**	82.7	79.6
RP+SSD+RH	72.4	**84.2**	80.0
ensemble class.	**77.5**	**89.1**	**84.0**

Classification Accuracy in %

Feature Set Applications

- Retrieval of Music by Sound Similarity
- Music Recommendation
- Semantic Content Description
- Automatic Genre Classification
- etc. (more later)

Retrieval by Similarity

Example 1: classical song
Query Song: classical_2-fruhlingsnacht.mp3

Top 5 and **#10** similar songs according to different feature sets:

Features:	**RH**	**SSD**	**RP**	**MFCC**
Rank:				
1.	classic	classic	classic	classic
2.	classic	classic	classic	classic
3.	classic	classic	classic	classic
4.	world	classic	classic	classic
5.	classic	classic	classic	classic
10.	classic	classic	classic	classic

Retrieval by Similarity

Example 2: rock/pop song
Query Song: rock_pop_1-nocturne.mp3

Top 5 and **#10** similar songs according to different feature sets:

Features:	**RH**	**SSD**	**RP**	**MFCC**
Rank:				
1.	rock_pop	rock_pop	rock_pop	rock_pop
2.	world	rock_pop	rock_pop	electronic
3.	electronic	world	world	electronic
4.	jazz_blues	electronic	jazz_blues	electronic
5.	rock_pop	electronic	metal_punk	electronic
10.	rock_pop	rock_pop	metal_punk	electronic

Retrieval by Similarity

Example 3: electronic music
Query Song: electronic_10-walking_safely.mp3

Top 5 and **#10** similar songs according to different feature sets:

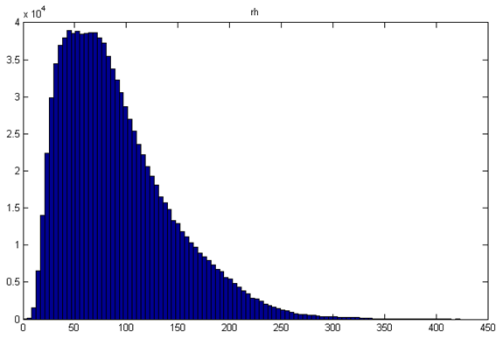
Features:	**RH**	**SSD**	**RP**	**MFCC**
Rank:				
1.	rock_pop	electronic	rock_pop	rock_pop
2.	rock_pop	jazz_blues	rock_pop	world
3.	world	rock_pop	world	electronic
4.	electronic	rock_pop	rock_pop	rock_pop
5.	world	rock_pop	world	metal_punk
10.	electronic	rock_pop	metal_punk	rock_pop

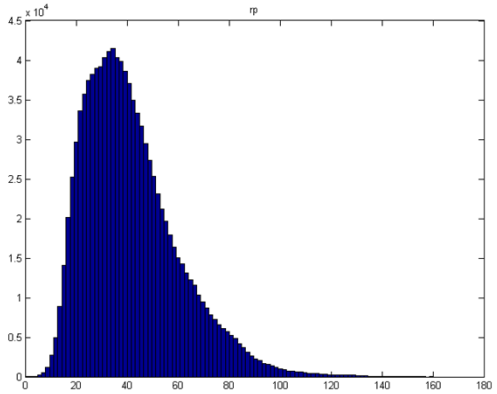
Similarity Measurement

- different distribution of distances

and different distance value ranges for each feature set

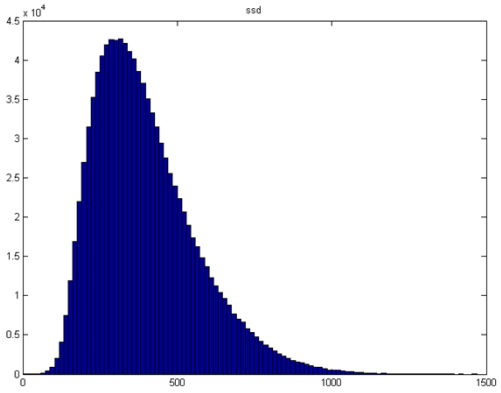
RH

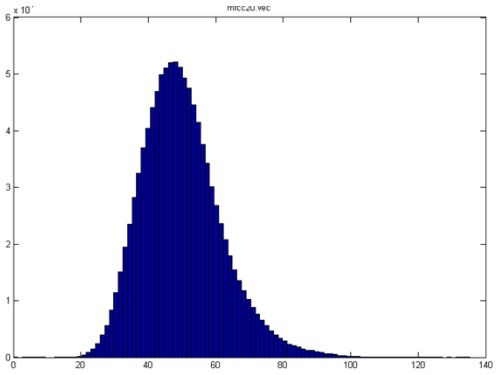




RP

SSD





MFCC

Mid and High-Level Features

- Onset detection
- Beat histograms, Pitch histograms
- Beat detection, BPM recognition
- Detection of pitch, harmony
- Structure detection (chorus, verse, …)
- Key detection
- Chord recognition
- Instrument recognition
- …

Boundary between low/mid/high blurry!

towards semantic concepts

→ creation of larger systems from individual features

Symbolic Features

- Features extracted from MIDI or MusicXML files
- based on note and pitch statistics, variations etc.
- advantages compared to audio:
  - no audio signal analysis necessary
  - note frequencies automatically contained in file
  - no source separation necessary, etc.
- disadvantages compared to audio:
  - actual sound is missing!
    (difficult for detecting instruments, timbre, …)

Symbolic Features (examples)

- Pitch: occurrence rates of diff. notes, pitch classes, ranges, variety
- Rhythm: time intervals, attacks, duration of notes, meters and rhythmic patterns
- Melody: melodic intervals, variation, melodic contours, phrases
- Chords: types of chords, vertical intervals, harmonic movement
- Instrumentation: types of instruments, importance, pitched vs. non-pitched, …
- Texture: # + rel. importance of independent voices, polyphonic, homophonic
- Dynamics: loudness of notes, variations in dynamics

Audio Features - Summary

- numerous features can be calculated from audio
- presented a selection of low/mid-level features
- many further features exist or can be calculated
- features capture different characteristics of sound
- have different dimensionality
- perform differently on different tasks
- and on different audio collections
- are joined for larger music information retrieval systems
- usually with the aim to detect higher semantics:
  chords, key, instruments, genre, mood …

Audio Features - Problematics

- undesired side-effects in features
  - volume (loudness) dependence
  - noise (clicks) can have an impact (also pitch/tempo changes)
  - „production effect“ (artefacts from (CD-)mastering „visible“ in features)
- different feature sets – different distance values and measures
- how to combine feature sets?
  - simply concatenating feature vectors? – weighting necessary
- proper normalization needed (but which?)
  - attribute normalization - zero mean/unit length normalization
  - (vector normalization) …

Audio Features - Problematics

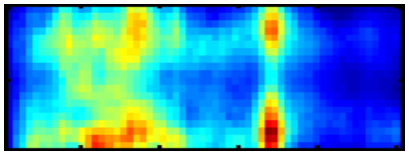
- testing on real-world datasets
  - copyright issues also for researchers (transfer of datasets…)
  - benchmark datasets often too small
  - too few genres (or „wrong“ genre assignments)
  - real-world feedback for similarities needed (manpower!)
  - artist/album effect („production effect“):
    train and test dataset should not contain tracks from same album or artist, because it might be „artificially easier“ for algorithm

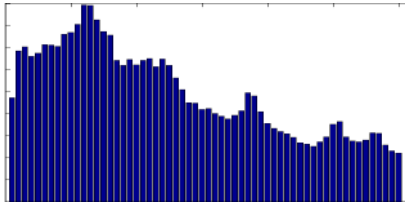
Questions?

3.
Audio Feature Extraction Tools

Rhythm Pattern Extractor

- Extractor for
  - Rhythm Patterns
  - Statistical Spectrum Descriptors
  - Rhythm Histograms
  - TSSD, TRH



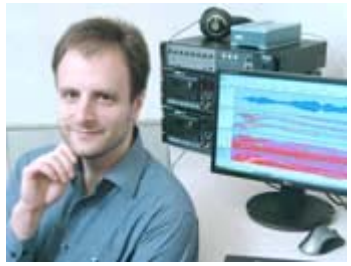


- by Pampalk, Lidy, Rauber et al., TU Wien
- available in Matlab and Java

  http://www.ifs.tuwien.ac.at/mir/downloads.html

MARSYAS

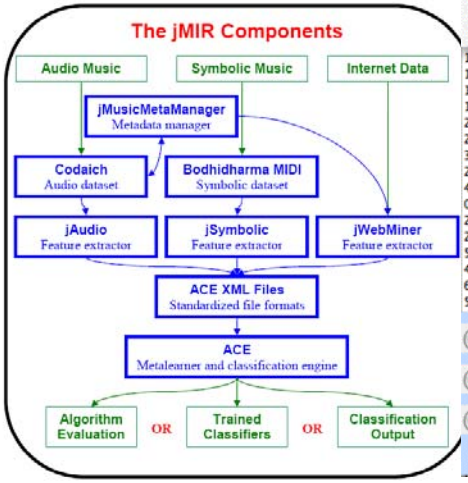
- Music Analysis, Retrieval and Synthesis for Audio Signals
- implements a range of functions and feature extractors:
  - Zero Crossings, Spectral Centroid, Rolloff, Flux, ...
  - MPEG-compression-based, Wavelet-based, Beat and Pitch Histograms
- by George Tzanetakis (Univ. of Victoria, Canada), now Open Source (C++)

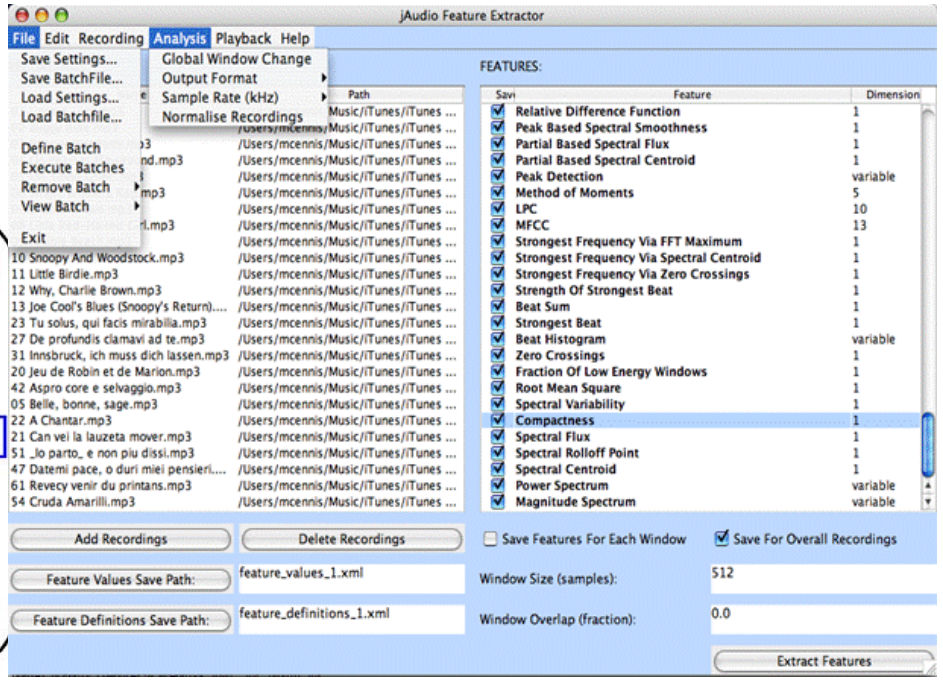


- http://marsyas.sness.net/
  http://sourceforge.net/projects/marsyas

jMIR (jAudio, jSymbolic)

- open-source feature extraction + classification in Java
- initiated by Cory McKay, McGill University
- 28 audio features
- 160 symbolic features

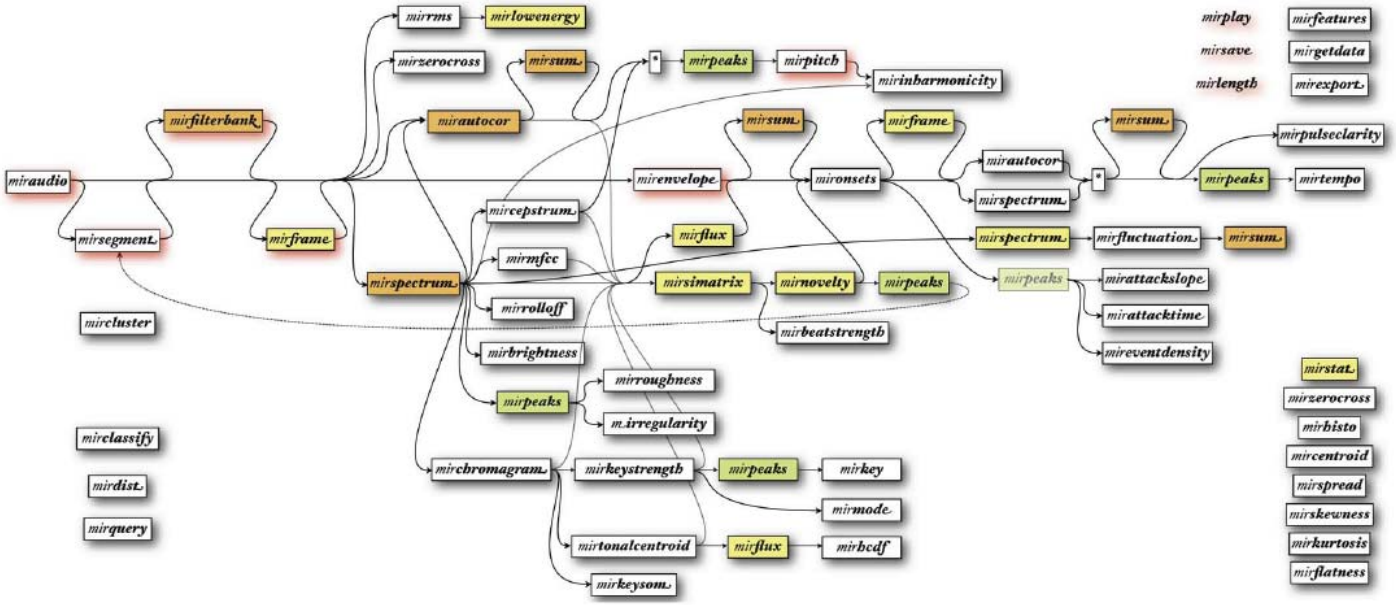




http://jmir.sourceforge.net/

mirtoolbox

- large number of state-of-the-art audio features
- conveniently usable as Matlab functions
- well-suited for experiments



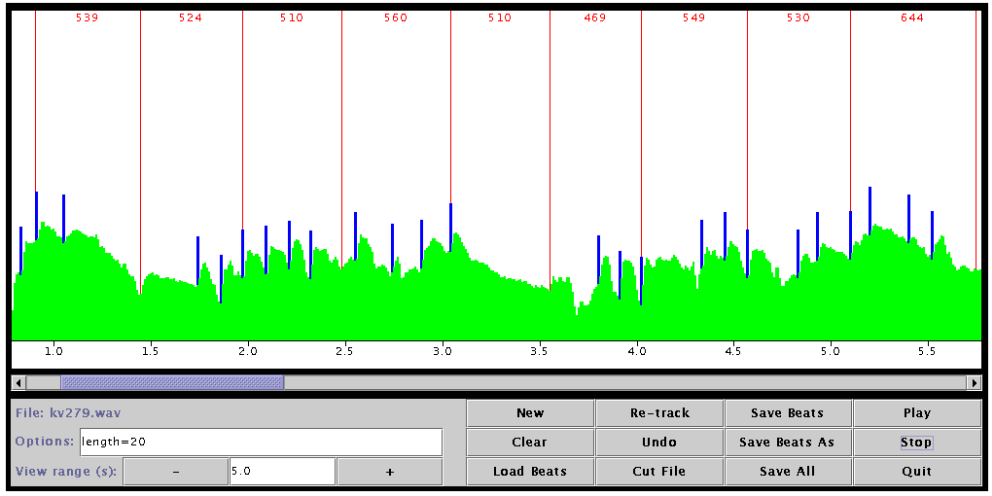
Synthetic overview of the features available in MIRtoolbox 1.2

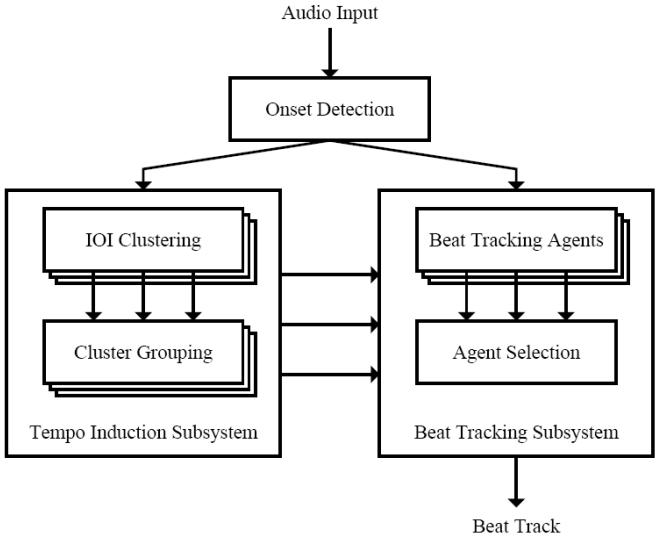
http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

by Olivier Lartillot et al., University of Jyväskylä, Finland ....

BeatRoot

- (Interactive) Beat Tracking and Visualisation (Java)
- winner of MIREX 2006 Audio Beat Tracking Competition
- by Simon Dixon, Queen Mary University London





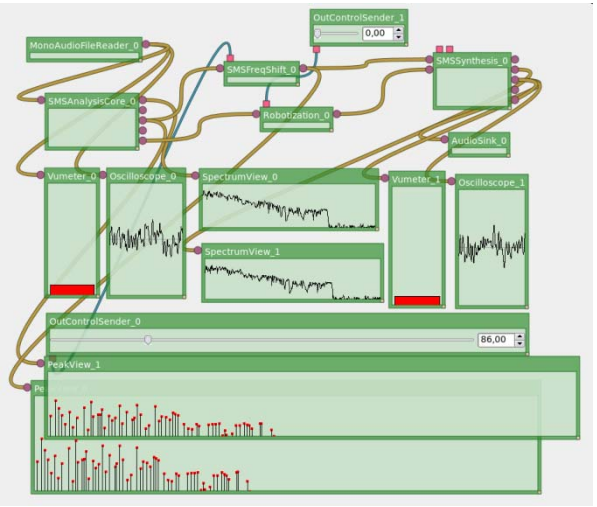
http://www.elec.qmul.ac.uk/people/simond/beatroot/index.html

Echonest

- Online Web API for audio feature extraction + meta-data (commercial, but limited free usage)
- delivers beats, key, tempo, segments etc. from audio
- delivers news, blogs, reviews etc.

CLAM

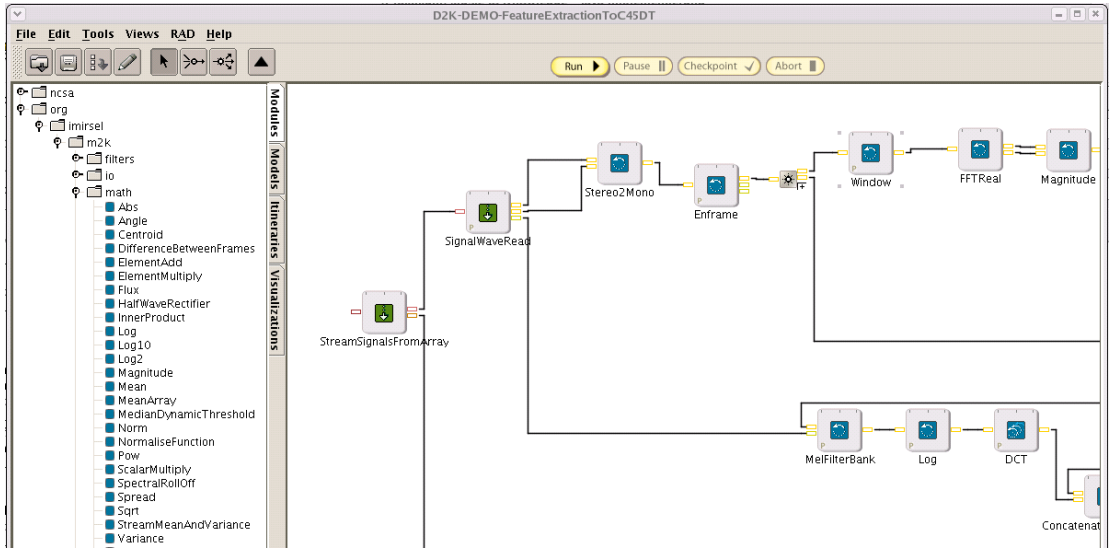
- software framework for research and application development in the Audio and Music domain
- by Music Technology Group (MTG), UPF Barcelona
- CLAM stands for C++ Library for Audio and Music
- in Catalan it also means something like "a continuous sound produced by a large number of people as to show approval or disapproval of a given event"



http://clam-project.org/

M2K

- Music to Knowledge (Java)
- Flow-Chart-based connection of algorithms



- Stephen Downie et al., University of Illinois

http://www.music-ir.org/evaluation/m2k/

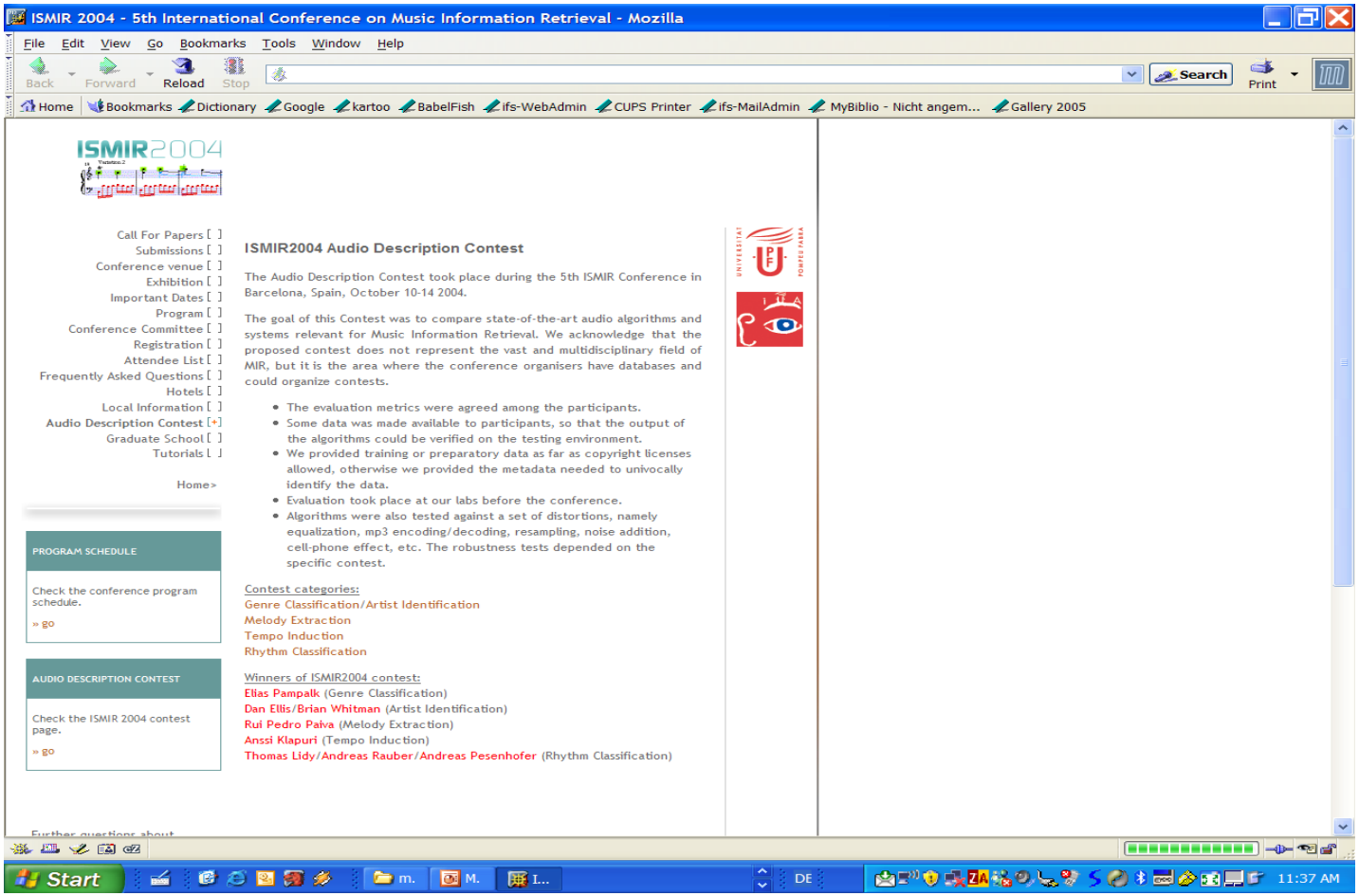
Questions?

4.
Benchmarking in Music IR

Evaluation in Music IR: History

- Discussion started at ISMIR 2001
  - evaluation frameworks
  - standardized test collections
  - tasks and evaluation metrics
- IMIRSEL project started 2002:
  (International Music Information Retrieval Systems Evaluation Laboratory), Univ. of Illinois, Stephen Downie
- First Audio Description contest at ISMIR 2004
- MIREX (Music Information Retrieval Exchange) started in 2005
- Annual, in connection with ISMIR conferences
- Evaluating many approaches of the MIR domain

2004 Audio Description Contest



http://ismir2004.ismir.net/ISMIR_Contest.html

2004 Audio Description Contest

- First attempt towards comparative benchmarking of MIR algorithms
- Five different tasks
  - Genre Classification
  - Artist Identification
  - Melody Extraction
  - Tempo induction
  - Rhythm Classification
- Some training/test data made available to participants
- Automatic evaluation
- Test for robustness of algorithms

MIREX

- Music Information Retrieval Evaluation eXchange
- annually before the ISMIR conference (deadline July/August)
- conducted by the IMIRSEL team, Univ. of Illinois (Stephen Downie)
- everyone can participate!
- democratic process:
  - suggestion and discussion of tasks via a mailing list months before
  - discussion of evaluation strategies on MIREX Wiki

http://www.music-ir.org/mirex/wiki/MIREX_HOME

MIREX 2012 Tasks

- Audio Classification
  - Audio US Pop Genre Classification
  - Audio Latin Genre Classification
  - Audio Music Mood Classification
  - Audio Classical Composer Identification
- Audio Cover Song Identification
- Audio Tag Classification
- Audio Music Similarity and Retrieval
- Symbolic Melodic Similarity
- Audio Onset Detection
- Audio Key Detection
- Real-time Audio to Score Alignment
- Query by Singing/Humming
- Audio Melody Extraction
- Multiple Fundamental Frequency Estimation & Tracking
- Audio Chord Estimation
- Query by Tapping
- Audio Beat Tracking
- Structural Segmentation
- Audio Tempo Estimation

MIREX Classification Tasks

- Audio Genre Classification
- Audio Artist Identification
- Audio Music Mood Classification
- Audio Classical Composer Identification

- genre classification typically used to evaluate performance of feature sets
- measured by Accuracy of recognition in %

Classification Tasks

- for stable results, cross-validation is used:
  - full data set is split into n folds
  - in n iteratons, n-1 parts of the data set are used for training the algorithm (learning), the remaining part is used for testing
  - final result is average performance of n folds
- in publications, usually 10 folds are used,
  in MIREX typically 3
- significance tests performed
  (or standard deviation of folds given)

Audio Music Similarity and Retrieval

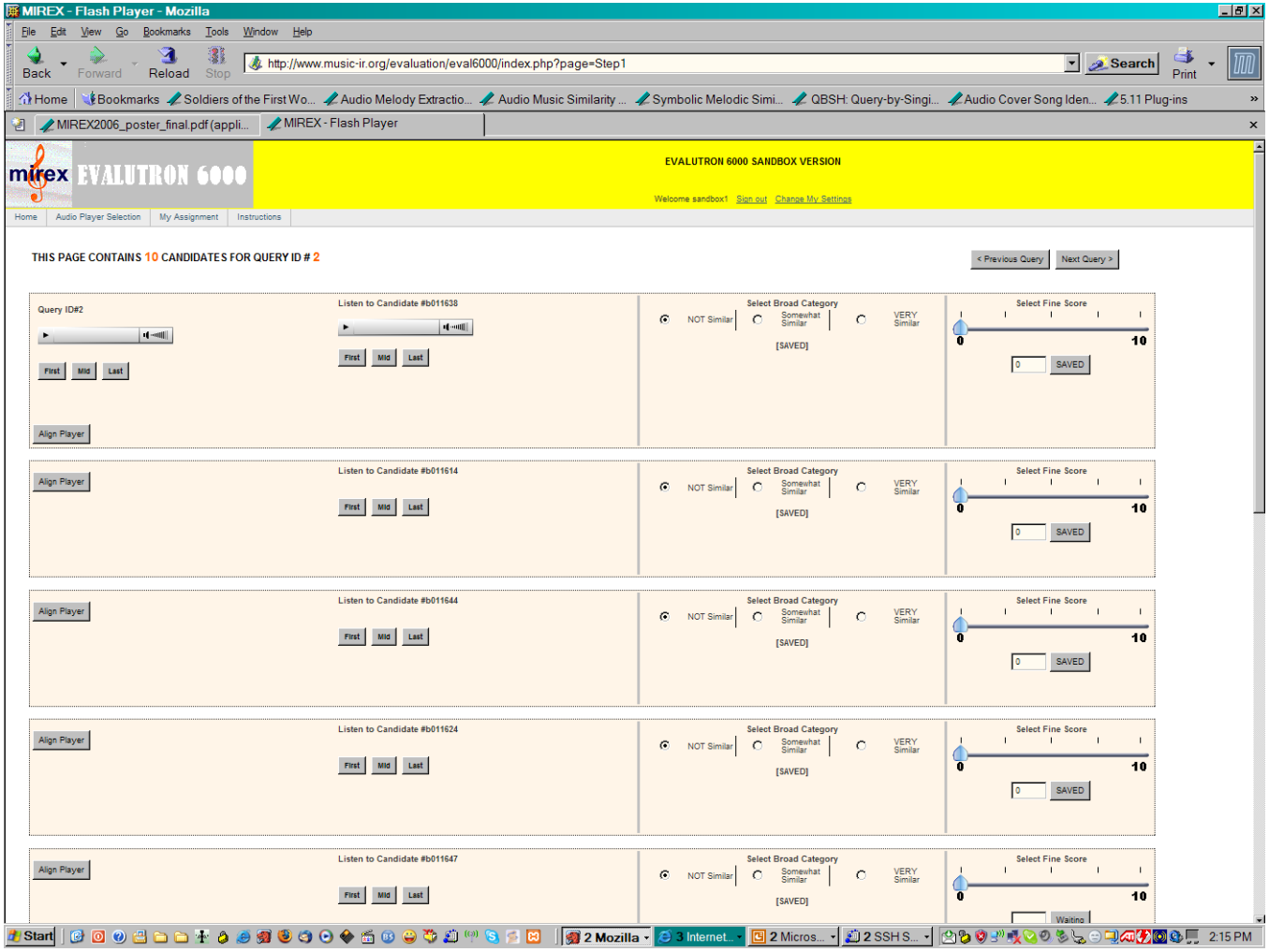
Audio Music Similarity and Retrieval Task

- Similarity retrieval rather than classification
- evaluated by human judgements:
  human listening tests (first @ MIREX 2006)
- Evalutron 6000:
  http://www.music-ir.org/evaluation/eval6000
- Test of statistical significance: Friedman test

Audio Music Similarity and Retrieval

- large scale music similarity evaluation
- 5000 music files, 9 genres
- Task:
  - apply feature extraction for audio similarity
  - compute distance matrix between all 5000 songs
- Evaluation
  - human listening tests on similarity
  - objective statistics based on meta-data

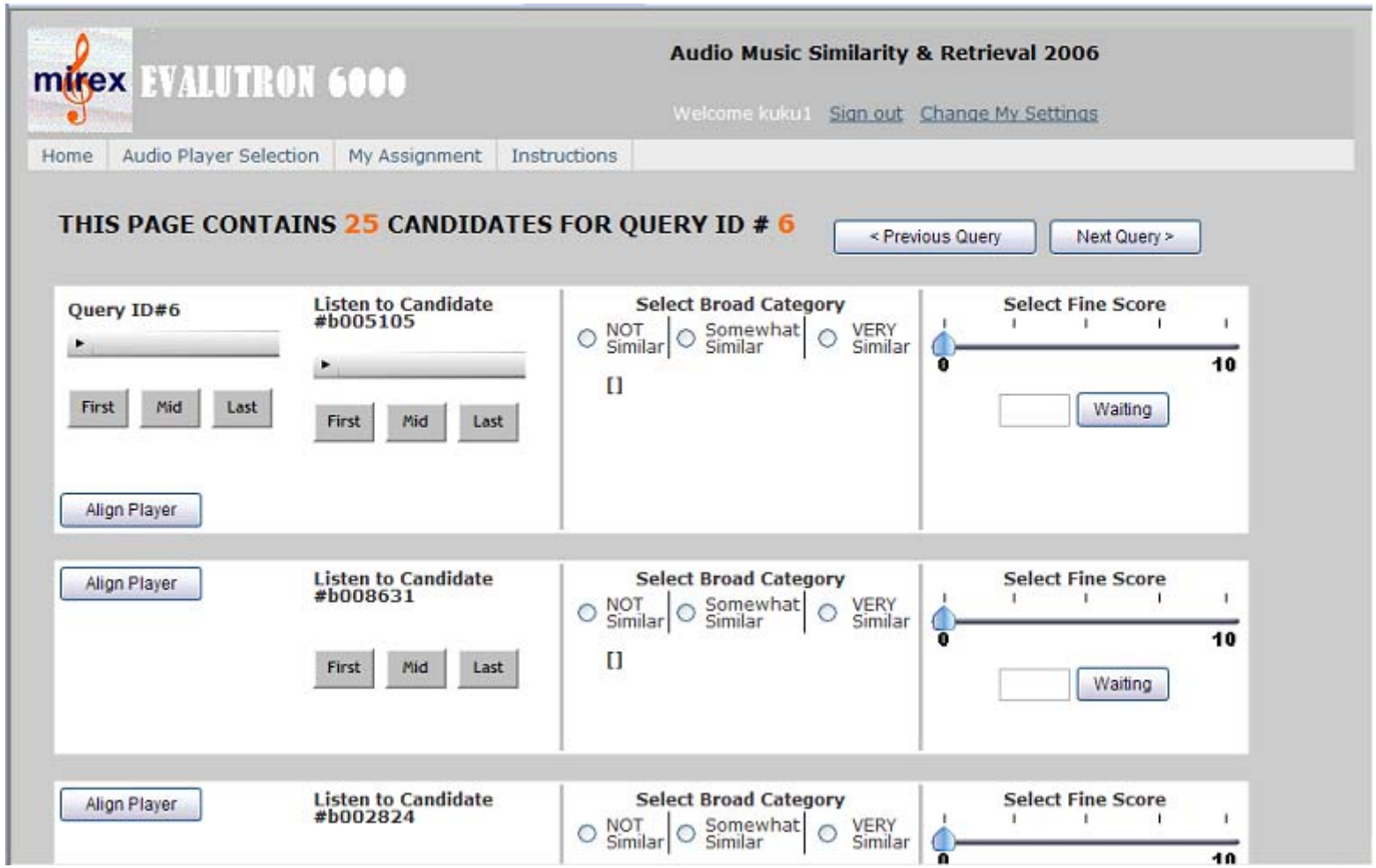
MIREX Evalutron 6000



MIREX Human Evaluation

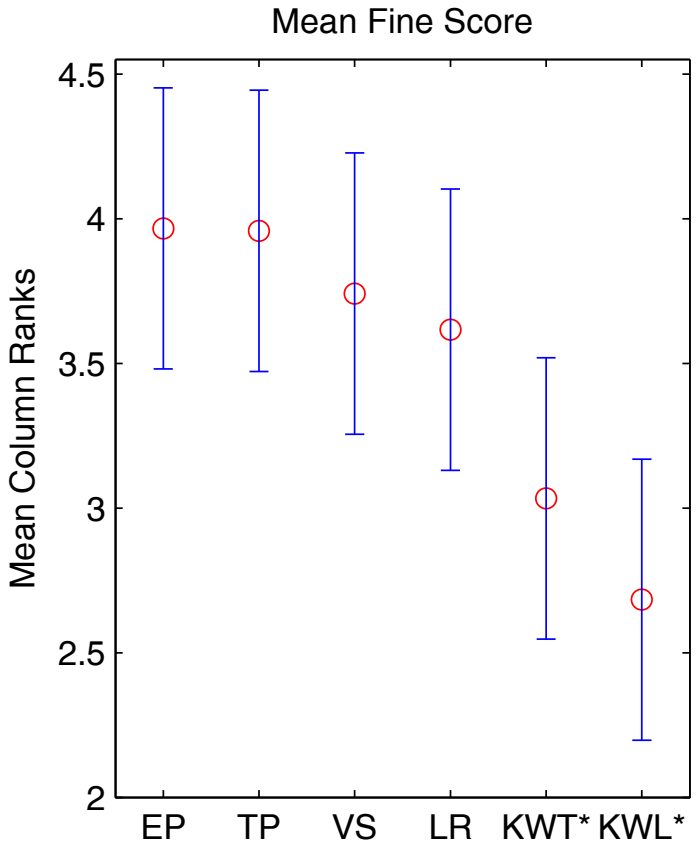
- 60 randomly selected queries
- ~ 20 human evaluators
- 7-8 ranked lists per evaluator
- 3 evaluations per ranked list
- 2 evaluation scales:
  - broad scale: very/somewhat/not similar
  - fine scale: between 0 and 10 (10 = best)

MIREX Evalutron 6000



MIREX2006:
Human Evaluation - Results

- 6 participating approaches
- Friedman test on fine scale
- no significant differences between first 5 algorithms
- LR = Lidy & Rauber

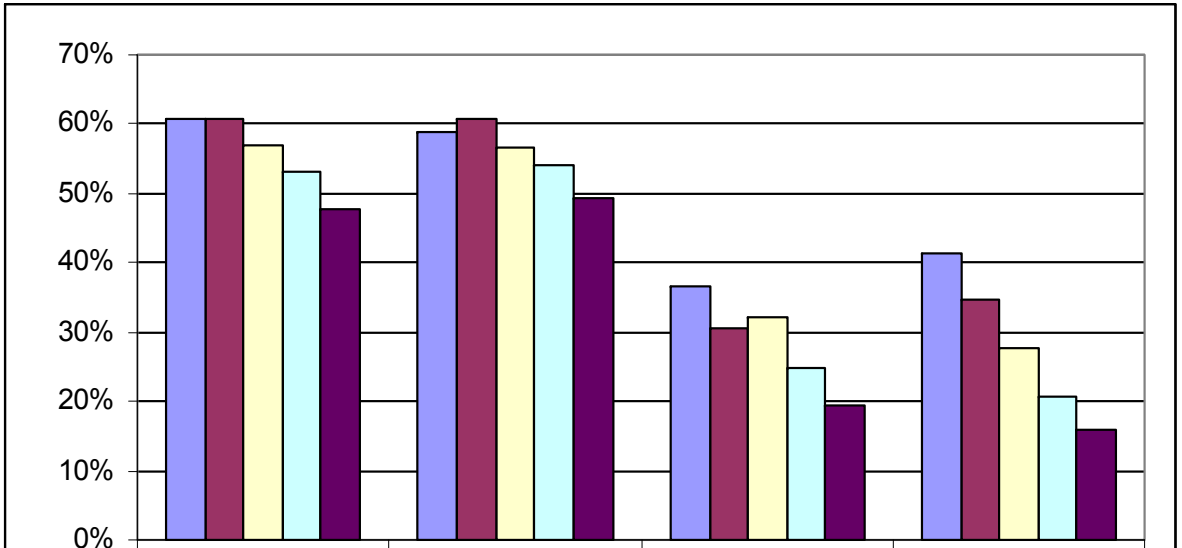


MIREX 2006:
MusicSim: Metadata Statistics

- Retrieval of the top 5, 10, 20 & 50 most similar to each file in the database
- Evaluation of the average % match of same
  - Genre
  - Genre after filtering out the query artist
  - Artist
  - Album title

MIREX2006:
MusicSim: Metadata Statistics - Results

- Results on the top 20 most similar



	genre	filtered genre	album	artist
TP	60,8%	58,9%	36,6%	41,3%
EP	60,6%	60,7%	30,5%	34,7%
LR	57,0%	56,7%	32,2%	27,7%
KWT	53,2%	54,1%	24,7%	20,7%
KWL	47,8%	49,2%	19,4%	15,9%

Audio Cover Song Identification

- 30 cover songs of a variety of genres
- 11 versions each (i.e. 330 audio files)
- embedded in 5000 song collection
- used a reduced data set of 1000 songs
- Task:
  - 30 cover song queries
  - return the 10 correct cover songs

MIREX 2006:
Audio Cover Song Identification

8 participants:

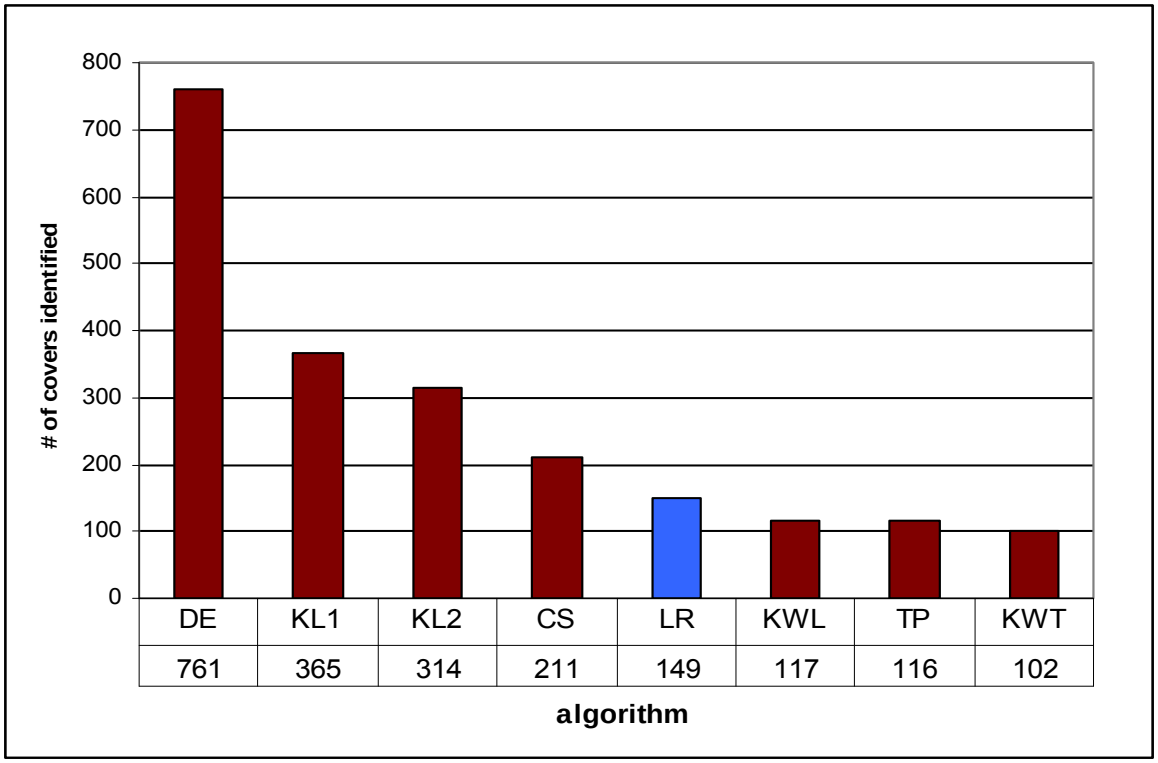
- 4 cover song detection algorithms
- 4 music similarity algorithms

Evaluation:

- Total number of covers identified
- Mean number of covers identified
- Mean of maxima (average of best-case perform.)
- Mean reciprocal rank (MRR) of first correctly identified cover

MIREX2006:
Audio Cover Song Identification - Results

- Number of identified covers:



Friedman Test on MRR:

DE is significantly better than others

no significant difference between remaining algorithms

Challenges for Music Retrieval Benchmarking

- Data - and access to it
  - real-world data set (- but how to get & use legally?)
  - sufficient (?) size
  - sufficient quality
- Metadata
  - high-quality labels (production-style)
  - ground truth annotation (can be very very time-consuming!!)
- Evaluation
  - automatic vs. human evaluation
  - which are the proper evaluation measures?
    how to perform tests properly?

MIREX Summary

- growing number of tasks
- growing number of people interested
- growing size of data sets
- data set issues remain (copyright, distribution, insufficient)

- MIREX is open to everyone interested in evaluation
- democratic process via mailing list and Wiki

Million Song Dataset Challenge

- Large Scale Evaluation Campaign for Music Recommendation Systems
- Based on the Million Song Dataset (MSD)
- Goal: Predict which songs a user will listen to
- Joint effort between
  - Computer Audition Lab at UC San Diego
  - LabROSA at Columbia University.
- Hosted on Kaggle.com
  - http://www.kaggle.com/c/msdchallenge

MSD Challenge Data

- Million Song Dataset
  - Metadata of 1M Songs
  - 30 Seconds Audio Snippets
- Full Listening History of 1M Users
  - Anonymized
- Half Listening History of 110.000 Users
  - 100.000 Training Set
  - 10.000 Validation Set

MSD Challenge Task

- Predict the second half of the 110.000 Users
- Any type of algorithm can be used
  - collaborative filtering
  - content-based methods
  - web crawling
- Create a list of Songs for each user
- Submit as Textfile
  - Results within Minutes
  - Multiple Submissions possible

Questions?

5. Current Research at IFS

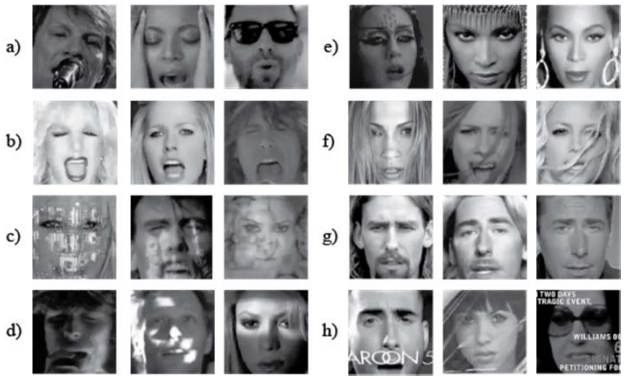
Music Video Information Retrieval

- Exploit the visual domain of videos
  - Image processing
  - Video retrieval
- Combine multiple modalities
  - Audio
  - Video
  - Lyrics
  - Social data

Music Video Information Retrieval

- Artist Identification
  - Identify the performing artist of a track
  - Problems
    - Features do not extract artist characteristics
    - Music style changes

- Possible Solutions in MVIR
  - Face recognition



Music Video Information Retrieval

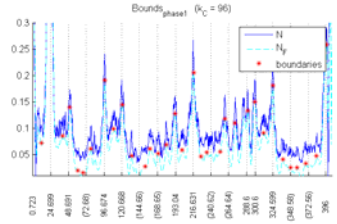
- Genre Classification
  - Color
  - Cut frequency
  - Objects
- General Insights
  - Correlations between
    - Sound and video progression
    - Sound and color
  - Director Effect
  - Music Video Similarity

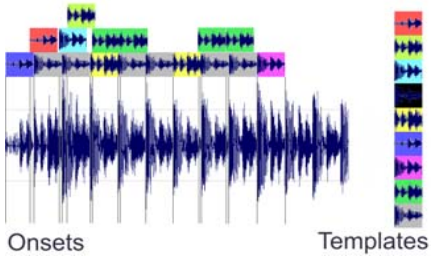
Benchmarking

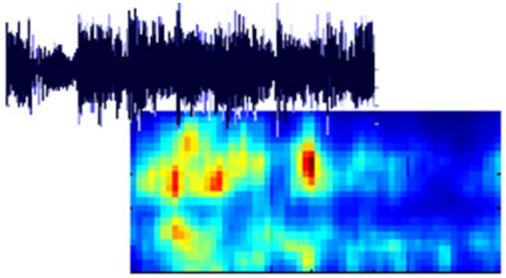
- Million Song Dataset
  - Downloaded all samples
  - Extraction of additional features
  - Generation of Ground Truth data
  - Large scale benchmarking experiments
- Further Datasets
  - Free Music Archive

Summary

- numerous Music IR algorithms exist already
- numerous commercial applications based on Music IR already exist as well
- but still there is a large number of open issues
- benchmarking and evaluation is important, but also faces challenges
- it‘s a very interesting domain!
  … still a lot of research to do! get involved!



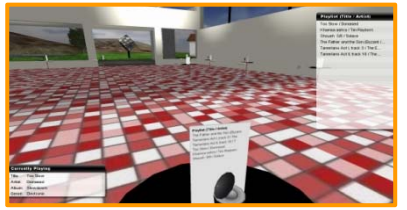


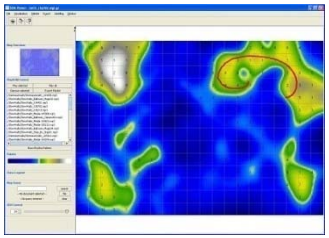


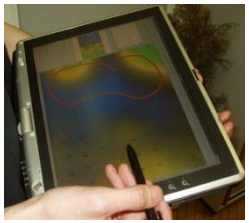
Thank You !

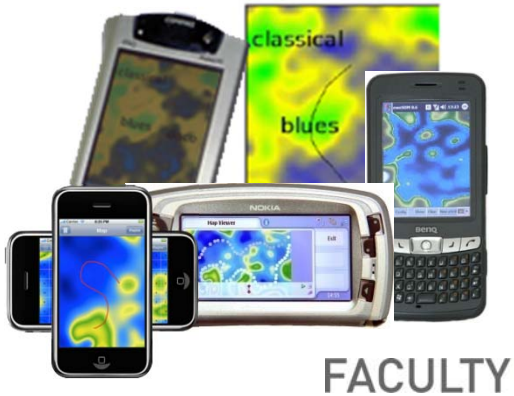
Alexander Schindler - schindler@ifs.tuwien.ac.at

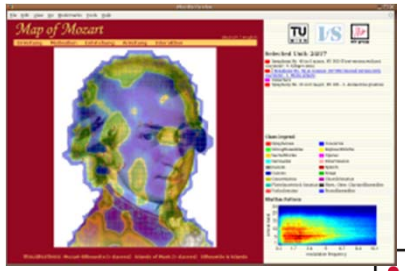
http://www.ifs.tuwien.ac.at/mir











FACULTY OF !NFORMATICS

ifs