

## Computational Models of Music Similarity

**Elias Pampalk**

National Institute for Advanced Industrial Science and Technology (AIST)

### **Abstract**

The perceived similarity of two pieces of music is multi-dimensional, subjective, and context-dependent. This talk focuses on simplified computational models of similarity based on audio signal analysis. Such models can be used to help users discover, organize, and enjoy the contents of large music collections.

The topics of this talk include an introduction to the topic, a review of related work, a review of current state-of-the-art technologies, a discussion of evaluation procedures, a demonstration of applications (including playlist generation and the organization of music collections), and finally a discussion of limitations, opportunities, and future directions.

2005/10/27, Osaka, SIGMUS



## Outline

2

### **1. Introduction**

- Context
- Definition of similarity
  - Playlist generation demonstration
- Alternative approaches
- Related research, history

### 2. Techniques

### 3. Evaluation

### 4. Application (MusicRainbow)

# Context

3

- **Abundance of (Digital) Music**

- new commercial music released every week
- back-catalogues
- creative commons (garage bands etc.)
- library music, ...

- **Technological Possibilities**

- **storage** → practically unlimited size of music collections
- **bandwidth** → music can be accessed via Internet, mobile phones, ...
- **portable music players** etc. → music is always present
- **CPU** → complex computations are feasible
- **algorithms** (many years of related research, e.g. MFCCs) → ...

→ **GOAL:**

use existing and develop new technologies to  
**make music more accessible**  
for active exploration as well as passive consumption

## Perception of Music Similarity

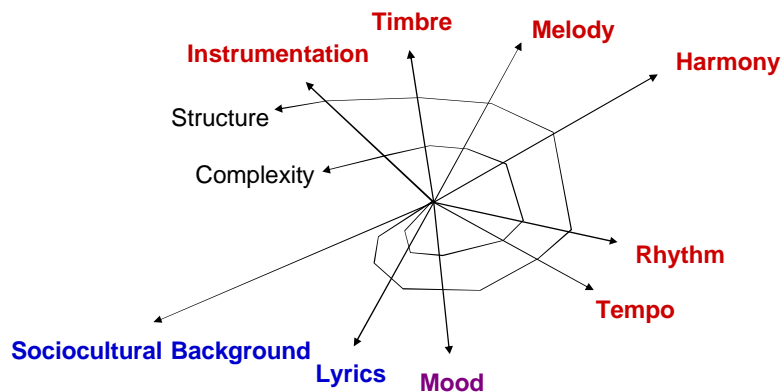
4

1. subjective

2. context-dependant

3. multi-dimensional

↙ E.g.:



## Music Similarity: Definition

5

### Songs A and song B are similar if ...

- **Playlist generation:**  
... **users** think A and B fit into the same playlist.
- **Recommendation:**  
... **users** who like A also like B.
- **Organization:**  
... **users** would expect to find A in the same category as B.

→ **User centered view**

**Problem:** difficult to evaluate

## Music Similarity: Definition

6

Example: **playlist generation**

### Specific Scenario

- Music: private collection (< 20,000 songs)
- Hardware: e.g. mobile audio player
- User: minimal interaction ("lazy")

### Basic Idea

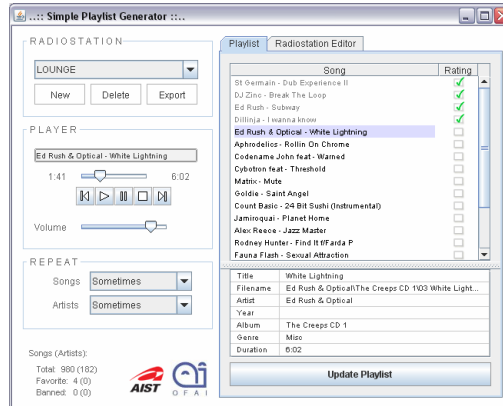
use audio-based similarity and user feedback to create playlist

(Demonstration uses state of the art similarity measure.)

## Music Similarity: Definition

7

### Demonstration: "Simple Playlist Generator"



[Pampalk & Gasser, ISMIR 2006]

## Alternatives to Audio-based Music Similarity

8

- Specific case of **playlist generation**:  
(personalized internet radio)
  - Experts (e.g. <http://pandora.com>)  
BUT: expensive! (human: 20-30 minutes per song)
  - Communities (e.g. <http://last.fm>)  
BUT: many problems with collaborative approaches

**Ideal Solution:**  
Combination with audio-based approaches

## Advantages of Audio-based Similarity

9

### - Fast & Cheap

On this laptop (Centrino 2GHz):

- < 2 seconds to analyze one song

- ~ 0.1 milliseconds to compare two songs

→ can be applied to huge music collections

### - Objective & consistent

## Audio-based Similarity: Related Fields

10

### **Audio** (signal processing)

Self-similarity, segmentation, summarization,  
extracting semantic descriptors (rhythm, harmony, melody, ...),  
genre classification, ...

### **Web** (collaborative filtering, web-crawling, ...)

Artist similarity, lyrics similarity, describing music with words, ...

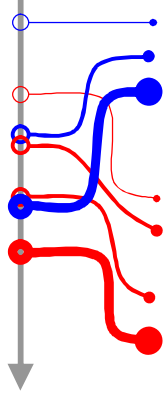
### **Symbolic** (MIDI etc.)

Melodic similarity, genre classification, ...

## Audio-based Similarity: Brief History

11

### Genre classification

- 
- 1996: audio classification (Wold et al.)
  - 2001: music classification (Tzanetakis & Cook)
  - 2004: first genre classification contest (ISMIR)

### Music similarity

- 1999: retrieval (Foote)
- 2001: organization (Frühwirth; Pampalk)  
playlist generation (Logan & Salomon)
- 2004: **“glass ceiling”** (Aucouturier & Pachet)
- 2006: first music similarity contest (MIREX)

- Young research field
- BUT: no major quality improvements since 2004!

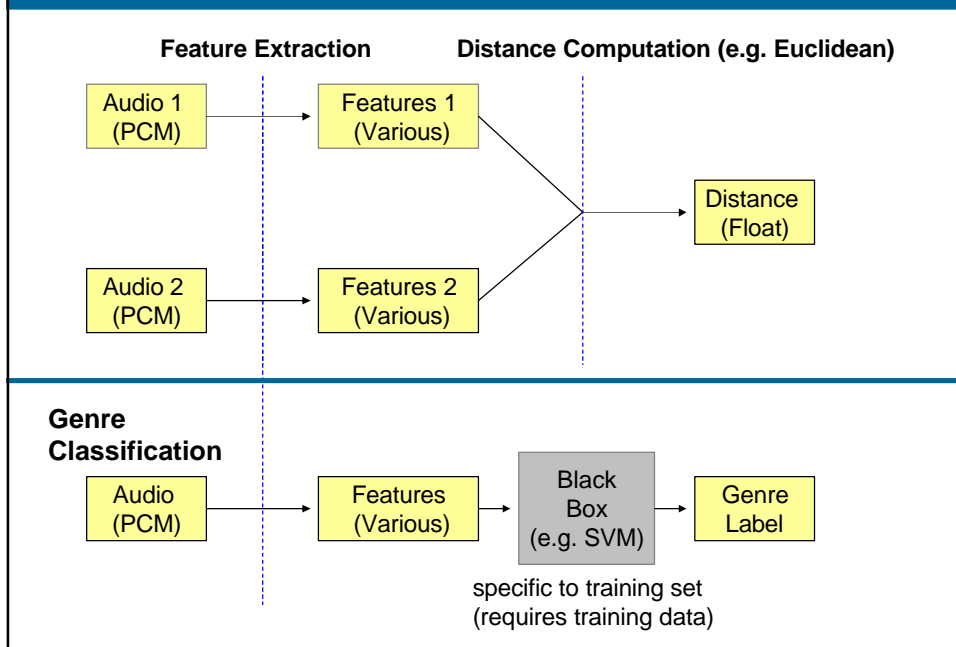
## Outline

12

1. Introduction
2. **Techniques**
  - Basics
  - Zero Crossing Rate (ZCR) walkthrough
  - Spectral similarity
  - Fluctuation patterns
  - Combination of different similarity measures
3. Evaluation
4. Application

## Music Similarity: Schema

13



## Audio Features: Type and Scope

14

### Type

- single numerical value (e.g. ZCR)
- vector (e.g. MFCCs)
- matrix or n-dimensional histograms (e.g. fluctuation patterns)
- multivariate probability distribution (e.g. spectral similarity)
- anything else (e.g. sequence of chords)

### Scope

- frame (e.g. 20ms, usually: 10ms-100ms)
- segment (e.g. note, bar, phrase, chorus...)
- song
- set of songs (e.g. album, artist, collection...)

## Distance Computation

15

Features: numerical, vector, matrix

→ **Euclidean, cosine, Minkowski,...**

Features: probability distributions

→ **Earth Mover's distance, Monte Carlo sampling, Kullback Leibler divergence, ...**

**Alternatives (e.g.):**

- use genre classification results to compute similarity
- use any form of combination

## Audio Features in this Talk

16

- **Zero Crossing Rate (ZCR)**

- simple walkthrough
- illustrates problem of generalization

- **Timbre related**

- introduction to MFCCs
- spectral similarity

State of the Art

- **Rhythm related**

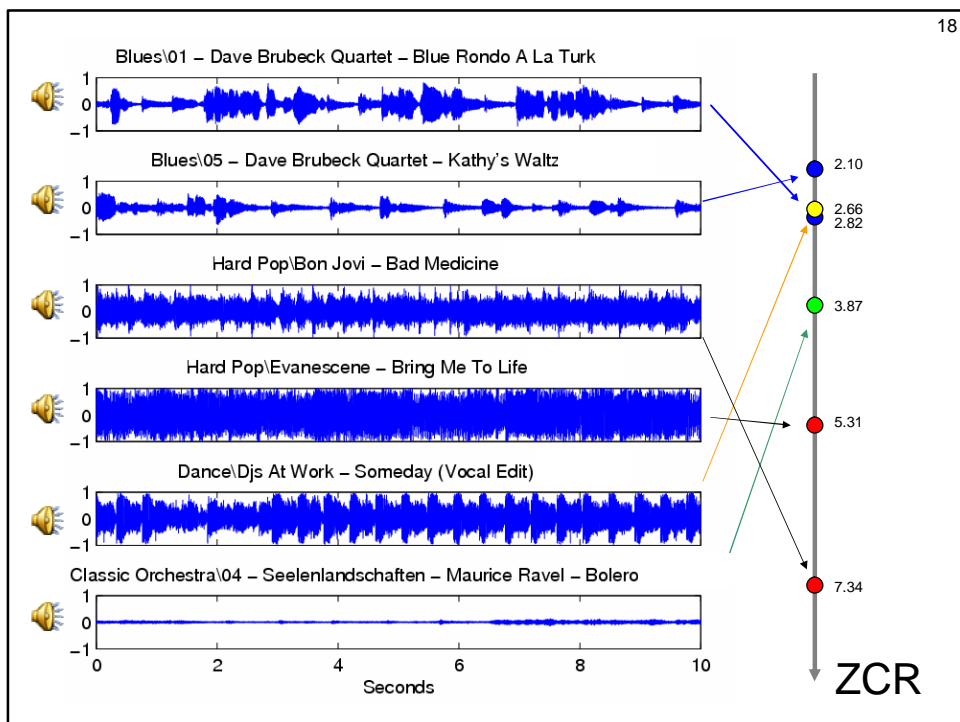
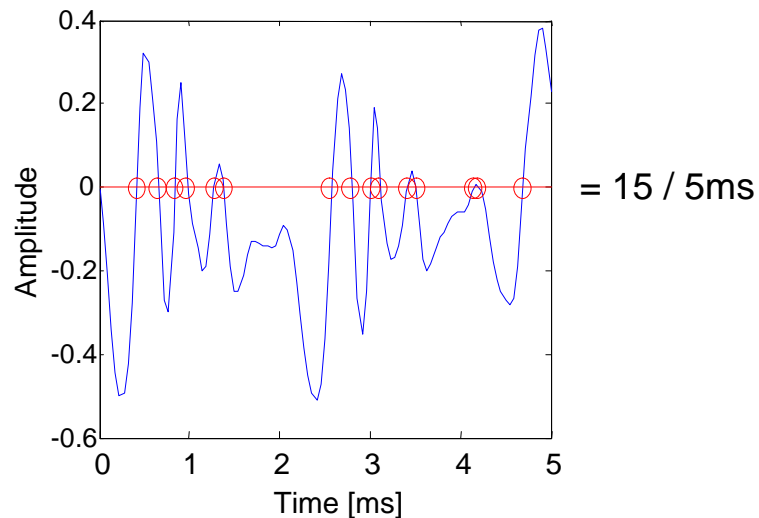
- fluctuation patterns



# Audio-based Music Similarity: Walkthrough

17

Zero Crossing Rate (ZCR) = 3/ms



18

Similarity = **Feature Extraction** + Distance Computation

Typical schema in feature extraction research (**generalization problem**)

1. find feature that works good on current set of music (e.g. 4 pieces)
2. later on, find out that there are other pieces where feature fails  
(→ go back to step 1)

ZCR (and many other low-level audio statistics, incl. e.g. RMS)

- + simple
- + can create interesting results sometimes
- only weakly connected (if at all) to human perception of audio
- generally **musically not really meaningful** (noise/pitch?)

→ meaningful descriptors **require higher level analysis.**

one typical intermediate representation is the spectrogram ...  
(time domain → frequency domain)

## Spectral Similarity (Timbre Related)

Spectrum



### References:

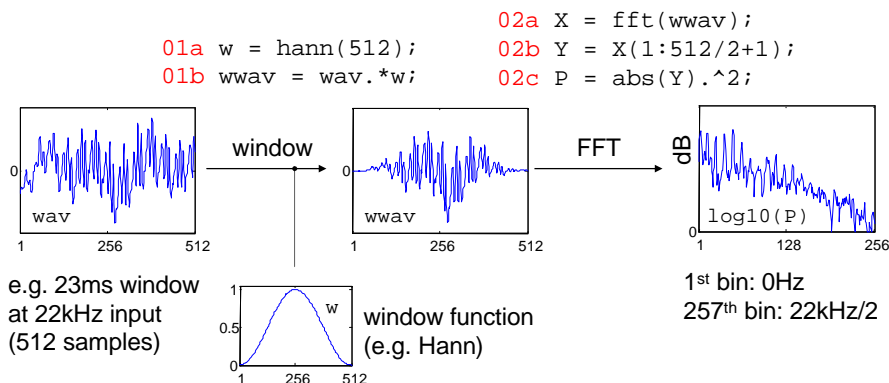
- Logan & Salomon, ICME 2001 (+ Patent)
- Aucouturier & Pachet, ISMIR 2002
- Mandel & Ellis, ISMIR 2005

## Mel Frequency Cepstrum Coefficients (MFCCs)<sup>21</sup>

MFCCs are one of the most common representations used for Spectra in MIR

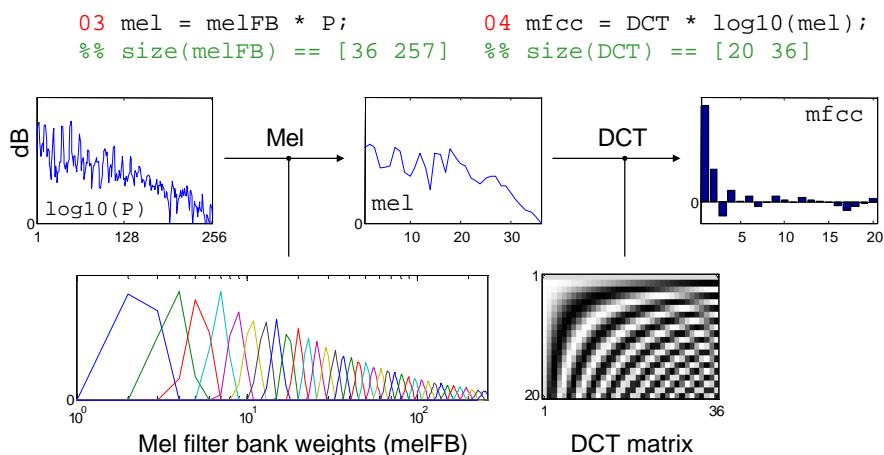
Given audio signal (e.g. 23 milliseconds, 22kHz mono)

1. apply window function
2. compute power spectrum (with FFT)



## Mel Frequency Cepstrum Coefficients (MFCCs)<sup>22</sup>

3. apply Mel filter bank
4. apply Discrete Cosine Transform (DCT) → MFCCs



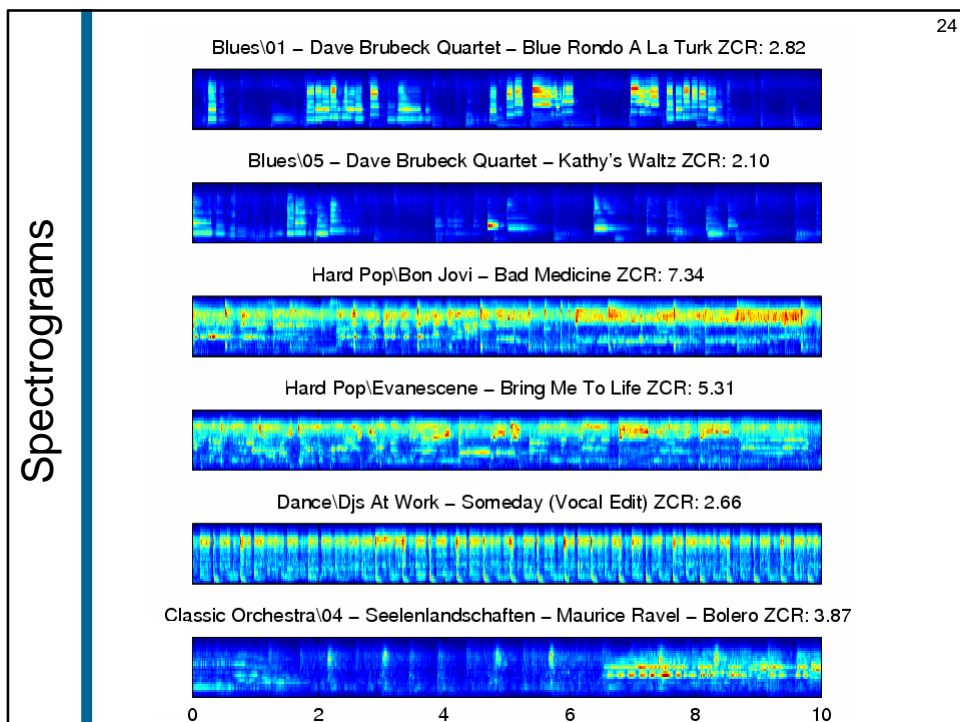
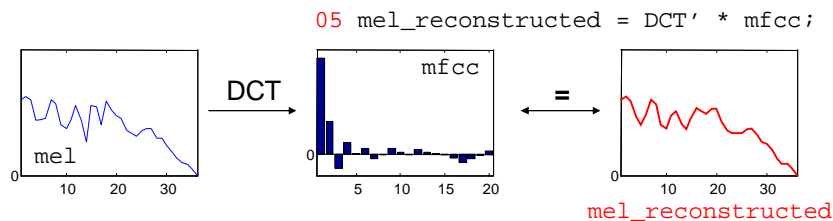
## Mel Frequency Cepstrum Coefficients (MFCCs)<sup>23</sup>

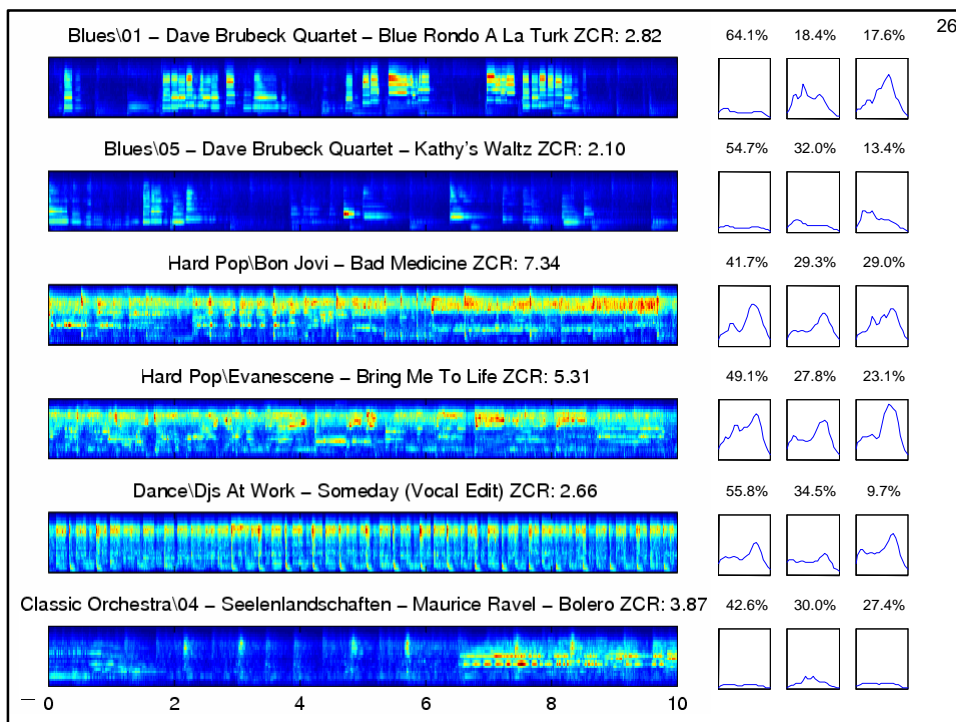
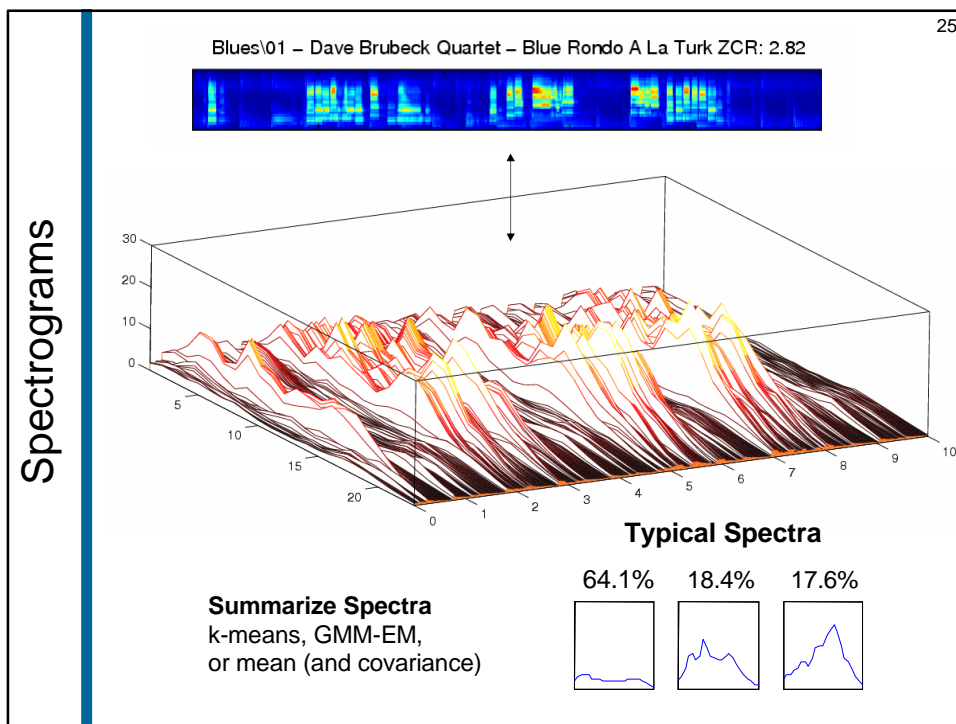
### Advantages

- simple and fast (compared to other auditory models)
- well tested, many implementations available (speech processing)
- compressed representation, yet easy to handle  
(e.g. Euclidean distance can be used on MFCCs)

### Important characteristics

- non-linear loudness (usually dB)
- non-linear filter bank (Mel scale)
- spectral smoothing (DCT; depends on number of coefficients used)  
simple approximation of psychoacoustic spectral masking effects





## Computing Distances between Typical Spectra <sup>27</sup>

### 1. Earth Mover's Distance + Kullback Leibler Divergence

(k-means clustering, diagonal covariance)

Logan & Salomon, ICME'01

### 2. Monte Carlo sampling

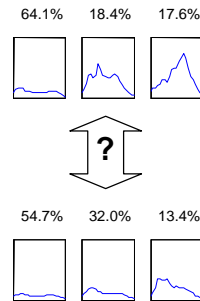
(GMM-EM, diagonal covariance)

Aucouturier & Pachet, ISMIR'02

### 3. Kullback Leibler Divergence

(mean, full covariance)

Mandel & Ellis, ISMIR'05

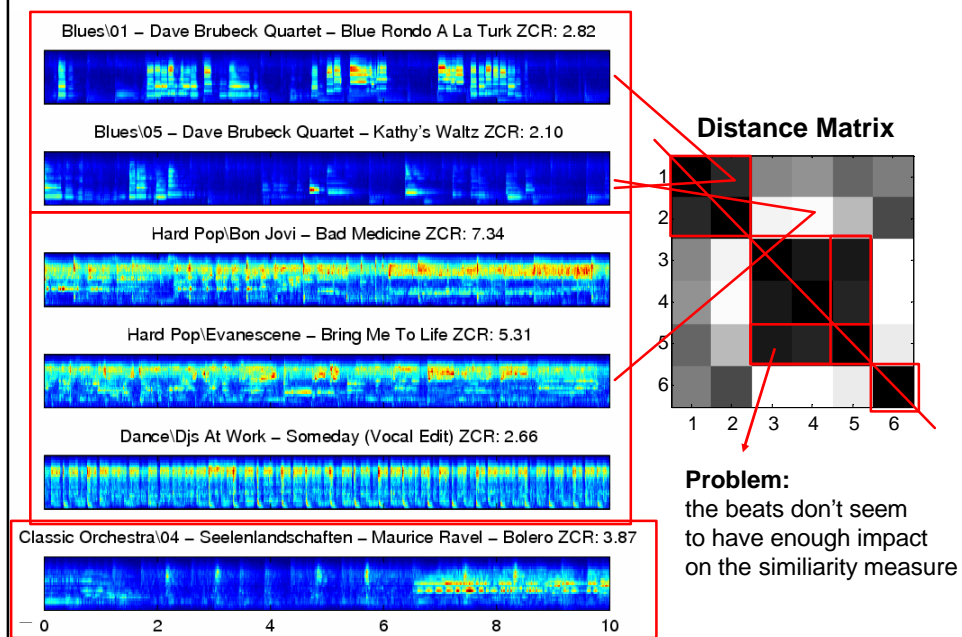


### Recommended article

Aucouturier & Pachet: "Improving timbre similarity: How high is the sky?"

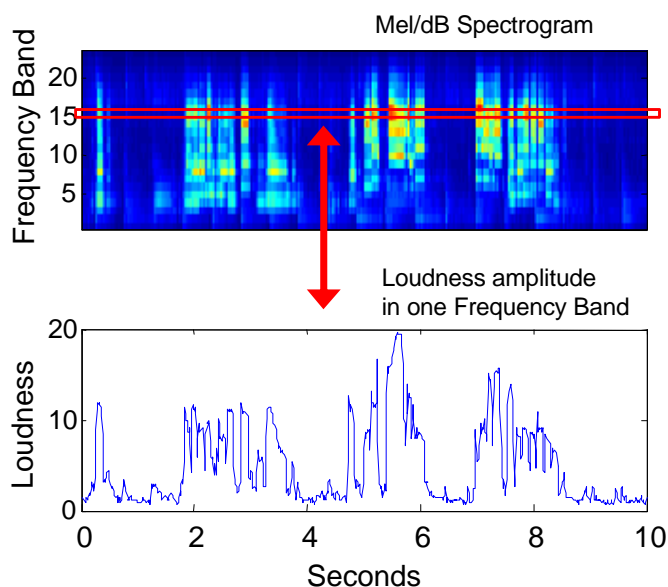
*Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

## Spectral Similarity, Distance Matrix <sup>28</sup>



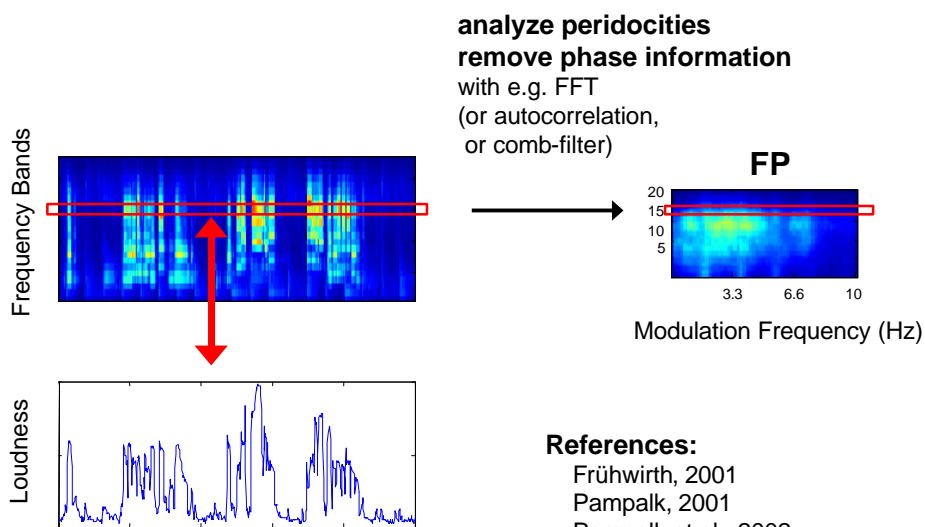
## Fluctuation Patterns (Rhythm Related)

29

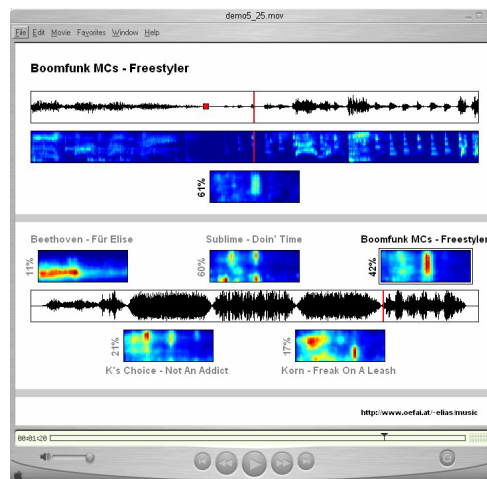


## Fluctuation Patterns (Rhythm Related)

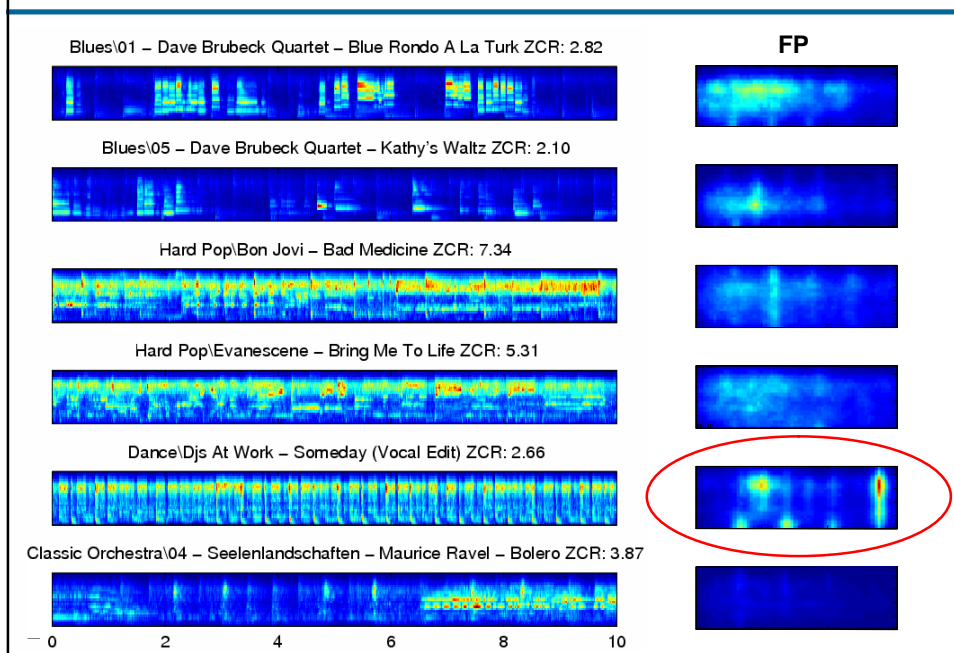
30



## Fluctuation Patterns: Demonstration



## Fluctuation Patterns (Rhythm Related)

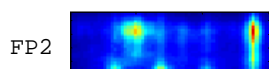
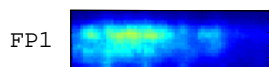




## Fluctuation Patterns (Rhythm Related)

33

### Distance computation



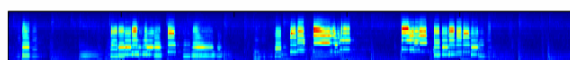
#### Euclidean distance (L2 norm)

```
d = sqrt(sum((FP1(:)-FP2(:)).^2));
%% e.g. size(FP1)      == [24 60]
%%      size(FP1(:)) == [1440 1]
```

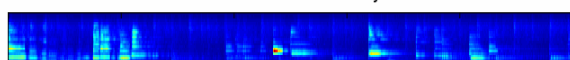
## Fluctuation Patterns (Rhythm Related)

34

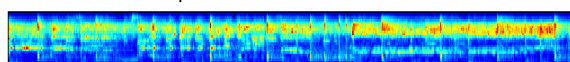
Blues\01 – Dave Brubeck Quartet – Blue Rondo A La Turk ZCR: 2.82



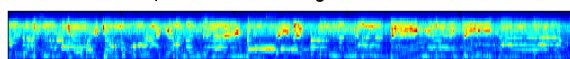
Blues\05 – Dave Brubeck Quartet – Kathy's Waltz ZCR: 2.10



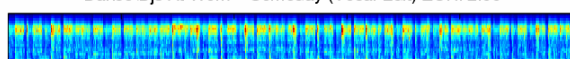
Hard Pop\Bon Jovi – Bad Medicine ZCR: 7.34



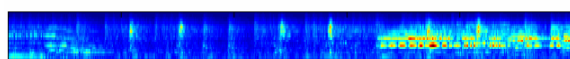
Hard Pop\Evanescence – Bring Me To Life ZCR: 5.31



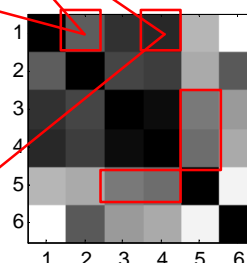
Dance\DJs At Work – Someday (Vocal Edit) ZCR: 2.66



Classic Orchestra\04 – Seelenlandschaften – Maurice Ravel – Bolero ZCR: 3.87

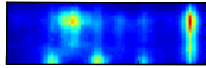


— 0 2 4 6 8 10



→ combine with spectral similarity

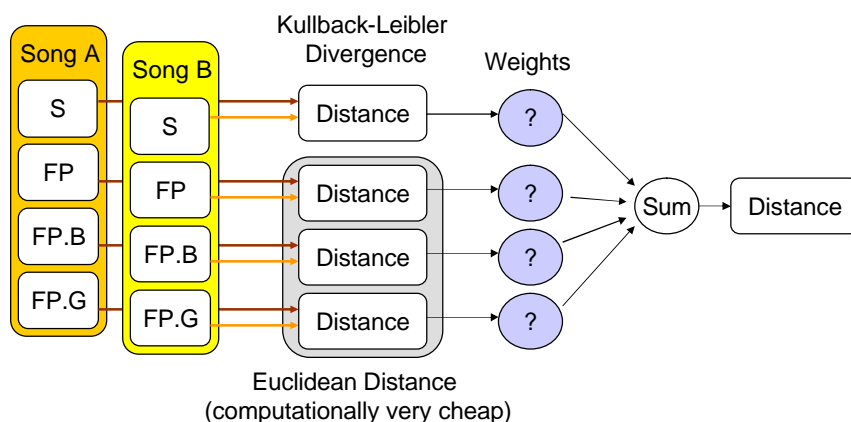
# Features Extracted from FPs



- **FP.B:** Modulations in bass frequency bands (e.g. <200Hz)
- **FP.G:** “Center of Gravity” on the horizontal axis  
(related to perceived tempo)
- Max, mean, variance, ...

[Pampalk 2001; Pampalk et al. 2005; Lidy & Rauber 2005; Pampalk 2006]

## Linearly Combined Distances



## Outline

37

1. Introduction
2. Techniques
- 3. Evaluation (and Optimization)**
  - Different types of evaluations
  - Genre-based evaluation
  - Listening tests, MIREX'06
4. Application

## 4 Basic Evaluation Types

38

- Evaluation within context of application
  - only way to find out about “acceptance”
  - very specific (results cannot be generalized to other applications)
  - **very difficult** to evaluate a large number of similarity measures
- Listening test: full similarity matrix
  - seems **infeasible** for larger numbers of songs
  - once similarity matrix is defined: fast & cheap evaluation and measuring perceptual significance of differences
- Listening test: based on rankings by algorithms
  - **allows measuring perceptual significance of differences**
  - difficult to evaluate a large number of similarity measures
- Genre-based
  - **fast & cheap**
  - can be used to evaluate very large parameter spaces
  - **DANGER:** very easy to do overfitting &  
not so easy to measure performance correctly

- **Assumption:** similar pieces belong to the same genre.

**Seems to hold in general!**

[Pampalk 2006; Novello et al. 2006; MIREX 2006]

- Basic Procedure (e.g.):
  1. Given a query song:
  2. Count number of pieces from the same genre within top N results

Typical genres used include

rock, classic, jazz, blues, rap, pop, electronic, heavy metal, ...

### + Advantages

genre labels easy to collect, cheap, fast

→ possible to evaluate large parameter spaces!

→ should always be the first sanity check of a similarity measure (before using listening tests!)

**if done correctly, good approximation of results from listening test!** [Pampalk 2006; MIREX 2006]

### - Problems

- danger of overfitting!!

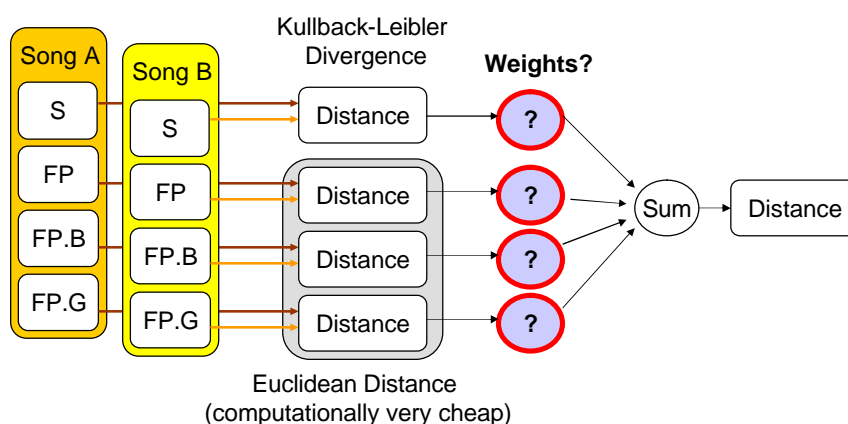
- genre taxonomies are inconsistent,

- similarity is not measured directly, ...  
(assumption does not always hold)

- Artist filter:  
**test set and training set must not contain pieces from the same artist.**  
otherwise “artist identification” performance is measured (focus on singers voice etc.). In addition: production effects (record studio etc.) might have unwanted effects on the evaluation.
- Different music collections (3 or more):  
**from different sources.** Performance of similarity measure can change a lot depending on the collection used. at least 2 collections should be used for development, and at least 1 for final conclusions (to test generalization).

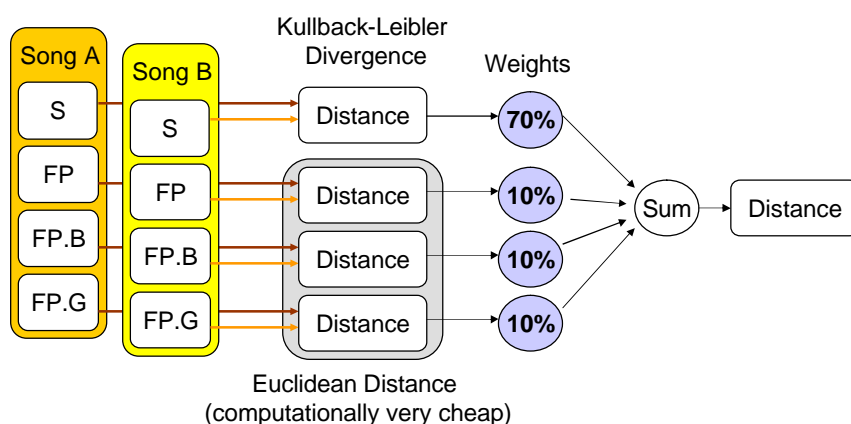
[Pampalk et al. 2005; Pampalk 2006]

## Linearly Combined Distances



Rank	Noisiness	Perc.	FP	FP Gravity	FP Bass	FP DLF	Spec. Sim.	DB-MS		DB-L		Score
								G1	G30S	G1	G30S	
1			<b>10</b>	<b>10</b>	<b>10</b>		<b>70</b>	67.4	67.4	32.4	35.2	<b>6.14</b>
2		10	10	10	10		60	67.1	66.4	33.0	34.6	5.83
3		10		10	10		70	66.8	66.4	31.8	34.7	5.46
4			10	10			80	67.4	65.7	32.1	34.4	5.44
5			10	10	20		60	66.1	66.9	31.5	34.9	5.42
6		10	20		10		60	65.7	66.4	32.6	34.5	5.36
7	10		10	10	10		60	63.9	66.1	<b>33.6</b>	<b>35.6</b>	5.35
8		10		10			80	66.8	66.1	31.8	34.1	5.26
9	10		20	10		10	50	64.9	66.1	32.7	35.1	5.25
10			10	10	10	10	60	67.2	66.8	30.9	33.9	5.25
11				10	10		80	<b>68.2</b>	66.7	31.0	32.9	5.25
<b>25</b>	10		20	10			60	64.1	65.2	32.7	<b>35.6</b>	<b>4.92</b>
515				30		20	50	66.0	68.4	26.5	29.5	3.15
2666							100	62.8	62.4	27.6	25.0	0.00

## Linearly Combined Distances (G1C)



**State-of-the art:** highest score at MIREX'06 audio-based similarity evaluation

## Listening Tests

45

allows measuring the perceptual significance of differences

- Select query song
- Ask algorithms to retrieve most similar songs
- Ask human listeners to rate similarity of these given the query

### **Assumption:**

Different people rate similarity of songs consistently.

### **Seems to hold in general!**

[Logan & Salomon 2001; Pampalk 2006;  
Novello et al. 2006; MIREX 2006]

- What scale should be used to rate similarity?
- What about the context of the question?
- Which songs should be selected? (Stimuli)

## Listening Test: G1 vs. G1C

46

- **100** queries
- **2** algorithms (G1, G1C)
- for each query each algorithm retrieves the most similar song from the music collection (using artist filter)
- given **3** songs (query Q, A, B) listeners are asked to rate the similarity of Q-A, and Q-B on a scale from 1 to 9.  
(1 = terrible, 9 = perfect)
- **3** listeners per song pair  
(to measure consistency)

[Pampalk 2006]

47

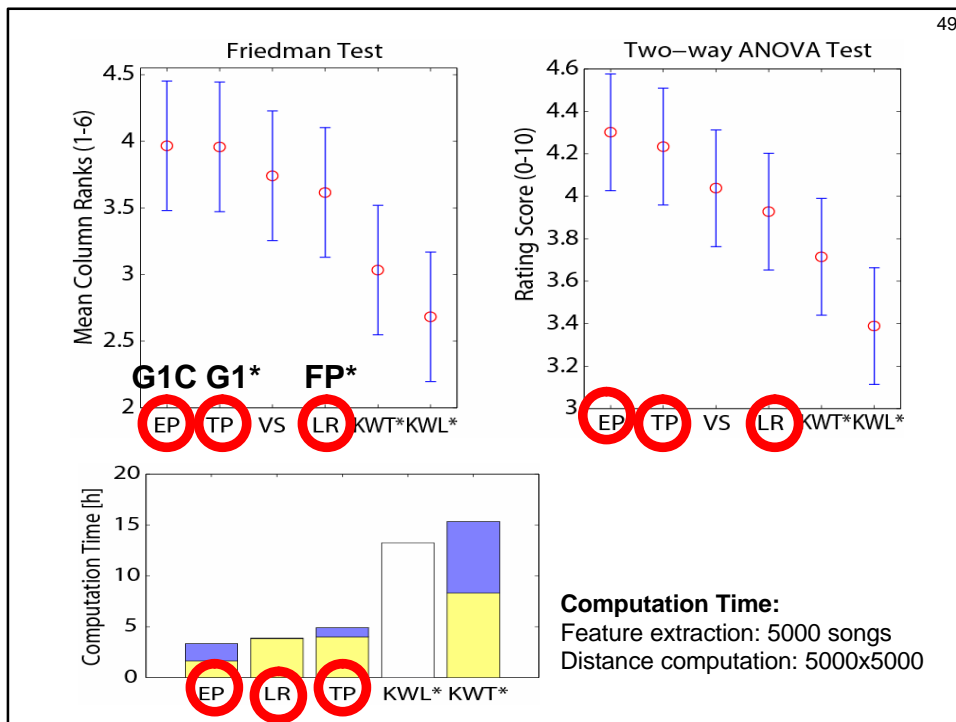
Rank	Noisiness	Perc.	FP	FP Gravity	FP Bass	P DLF	pec. Sim.	DB-MS		DB-L		Score
1	<b>G1C</b>	10	10	10	10			<b>G1C average rating: 6.37</b>				<b>6.14</b>
2		10	10	10	10		60	67.1	66.4	33.0	34.6	5.83
3		10		10	10		70	66.8	66.4	31.3	34.7	5.46
4			10	10			80	67.4	65.7	32.1	34.4	5.44
5										31.5	34.9	5.42
6										32.6	34.5	5.36
7	10									33.6	<u>35.6</u>	5.35
8										31.8	34.1	5.26
9	10									32.7	35.1	5.25
10										30.9	33.9	5.25
11				10	10		80	<u>68.2</u>	66.7	31.0	32.9	5.25
25	10		20	10			60	64.1	65.2	32.7	<u>35.6</u>	4.92
515				30		20						3.15
2666	<b>G1</b>							<b>G1 average rating: 5.73</b>				<b>0.00</b>

**Listening test result:**  
On a scale from 1 to 9 the difference is only about 0.6!

48

Listening Test: MIREX'06	
<ul style="list-style-type: none"> <li>• <b>60</b> queries</li> <li>• <b>6</b> algorithms (4 different research groups)</li> <li>• for each query, each algorithm retrieved the <b>5</b> most similar songs (using artist filter)</li> <li>• given <b>31</b> songs (query + 6 x 5 candidates) listeners are asked to rate the similarity of each query/candidate pair on a scale from 0 to 10. (0 = terrible, 10 = perfect)</li> <li>• <b>3</b> listeners per query/candidate pair</li> </ul>	



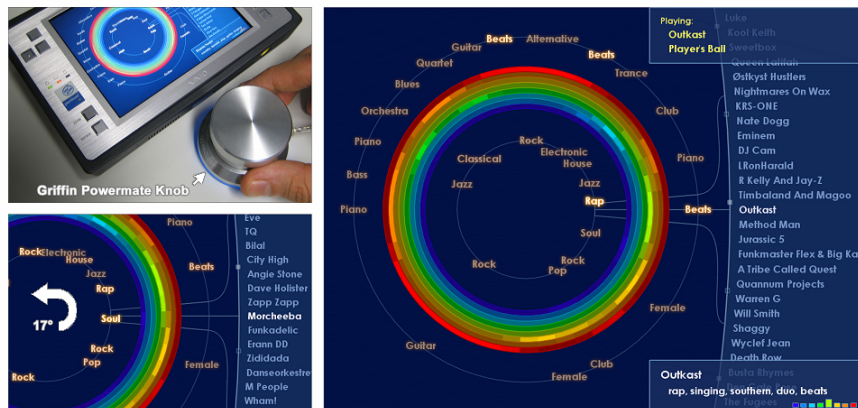


50

## Outline

1. Introduction
  - Playlist generation
2. Techniques
3. Evaluation
- 4. Application**
  - MusicRainbow

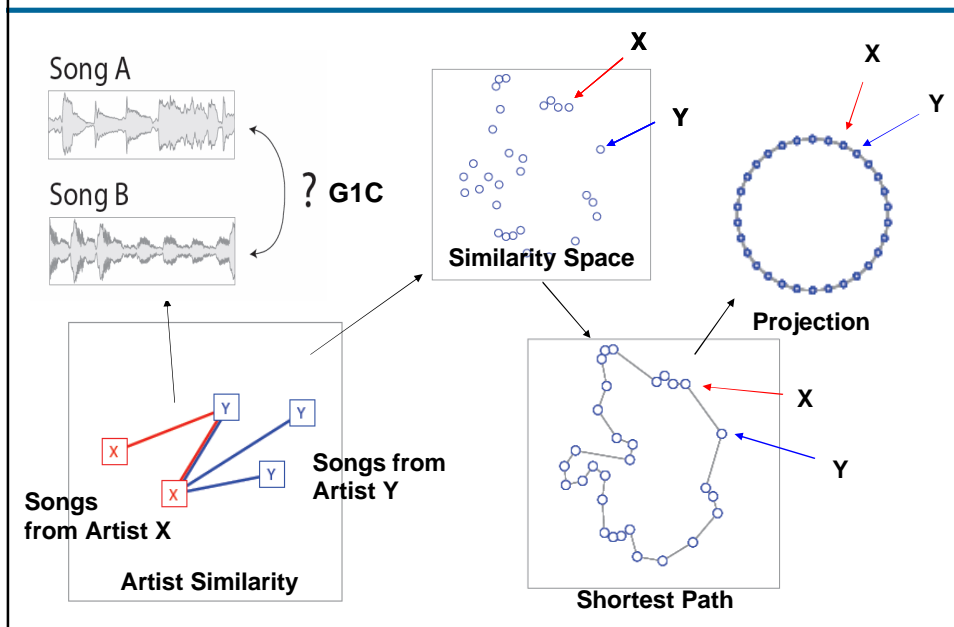
# MusicRainbow



Use audio-based similarity measure to compute artist similarity.

[Pampalk & Goto, ISMIR 2006]

## Artist Similarity and Organization



## Conclusions

### Current Situation:

- Low-level features are not enough
- Slow progress in the last years  
“glass ceiling” since 2004  
however, computational complexity has been  
reduced by several magnitudes (factor 1000 faster!)
- Many unexplored questions ...  
[Novello et al., ISMIR 2006]

## Similarity: Future Directions

- Improve linear combination model
- Use higher level semantic descriptors  
Rhythm, harmony, ...
- Context-dependant similarity  
Different parameters for different types of music and different users
- Combine audio-based similarity with other sources (e.g. collaborative filtering)  
e.g. [Yoshii et al., ISMIR 2006]
- Explore applications which can deal with erroneous similarity measures (e.g. playlist generation)

## References: Starting Points

---

- ISMIR Proceedings
- MIREX 2006 webpages
- J.-J. Aucouturier: "Ten Experiments on the Modelling of Polyphonic Timbre", PhD Thesis, 2006
- E. Pampalk: "Computational Models of Music Similarity and their Application in Music Information Retrieval", PhD Thesis, 2006