



Tutorial

Cross-Modal Music Retrieval and Applications

Part I: Classical Approaches

Meinard Müller

International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

Andreas Arzt, Stefan Balke

Johannes Kepler University
andreas.arzt@jku.at, stefan.balke@jku.at



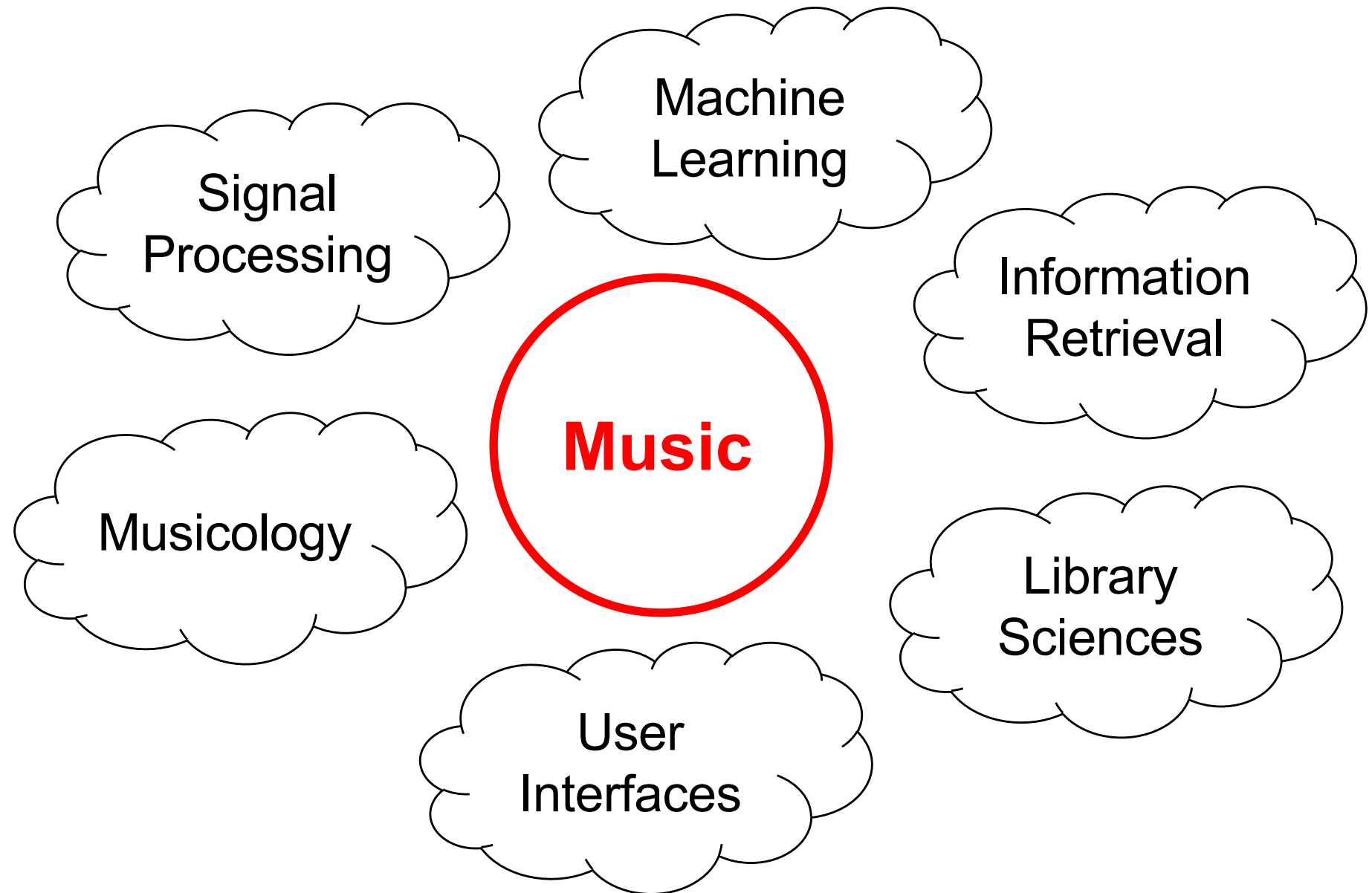
Overview (Part I)

- Music representations
- Retrieval scenarios (modality , specificity, granularity)
- Music synchronization (chroma features, dynamic time warping)
- Audio matching (subsequence DTW)
- Cover song retrieval
- Shingle-based retrieval (embedding techniques, PCA, deep learning)
- Cross-modal retrieval (challenges, enhanced representations)
- ...

Music



Music Information Retrieval (MIR)

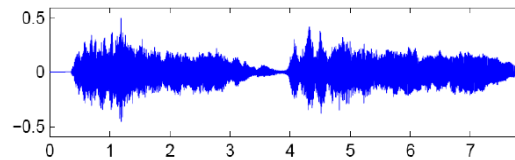


Music Information Retrieval (MIR)

Sheet Music (Image)



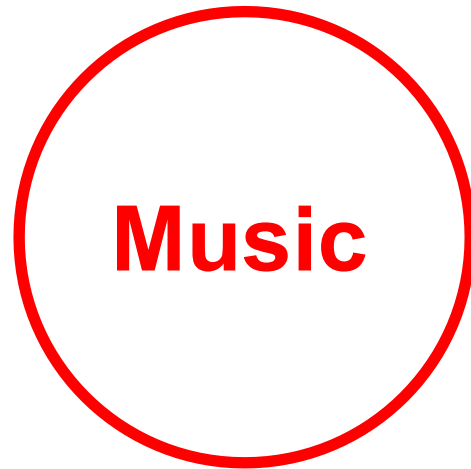
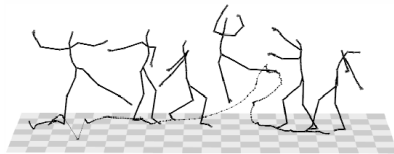
CD / MP3 (Audio)



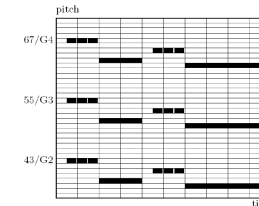
MusicXML (Text)

```
<note>  
  <pitch>  
    <step>E</step>  
    <alter>-1</alter>  
    <octave>4</octave>  
  </pitch>  
  <duration>2</duration>  
  <type>half</type>  
</note>
```

Dance / Motion (Mocap)



MIDI



Singing / Voice (Audio)



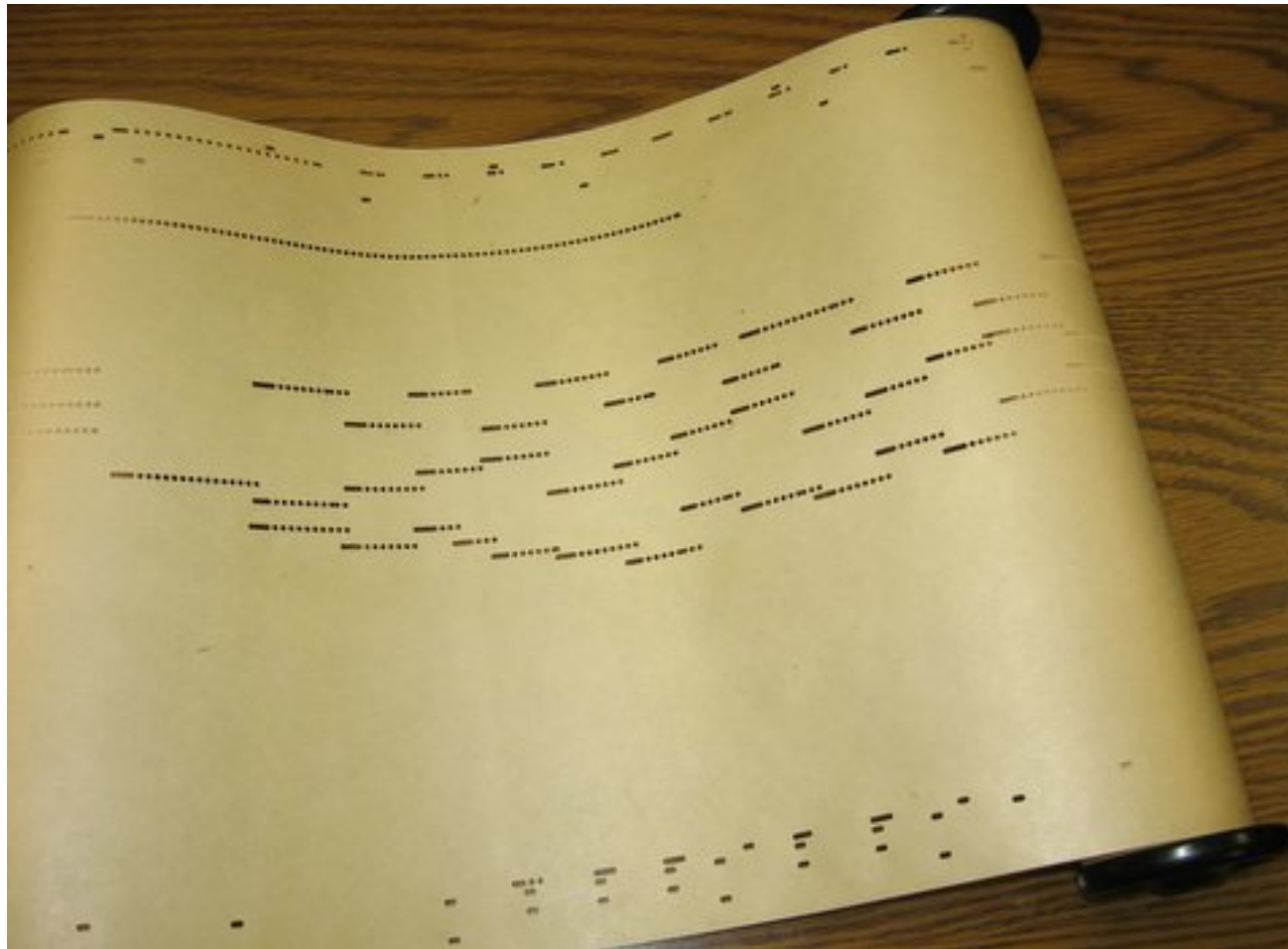
Music Film (Video)



Music Literature (Text)



Piano Roll Representation



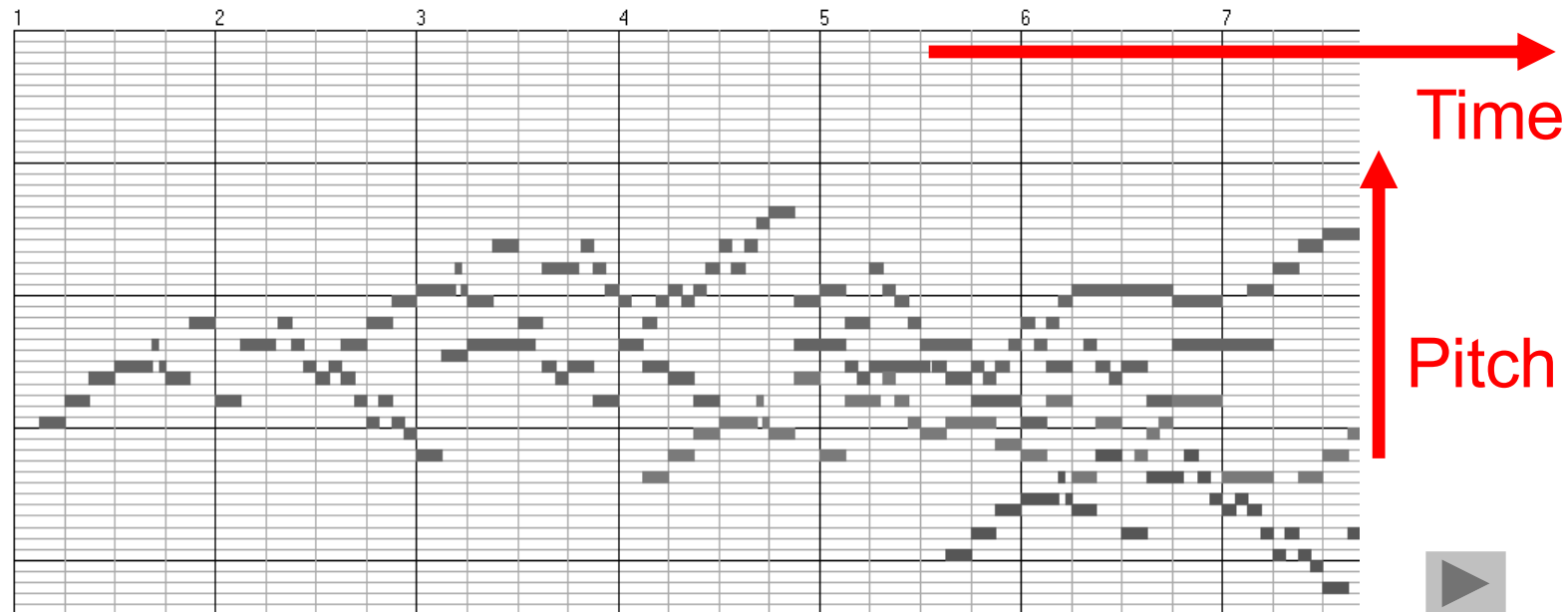
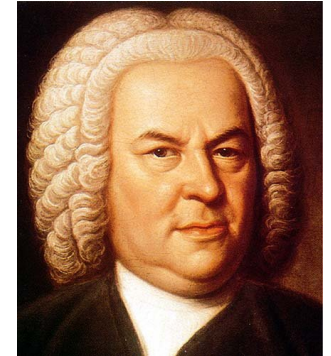
Player Piano (1900)



Piano Roll Representation (MIDI)

J.S. Bach, C-Major Fuge

(Well Tempered Piano, BWV 846)

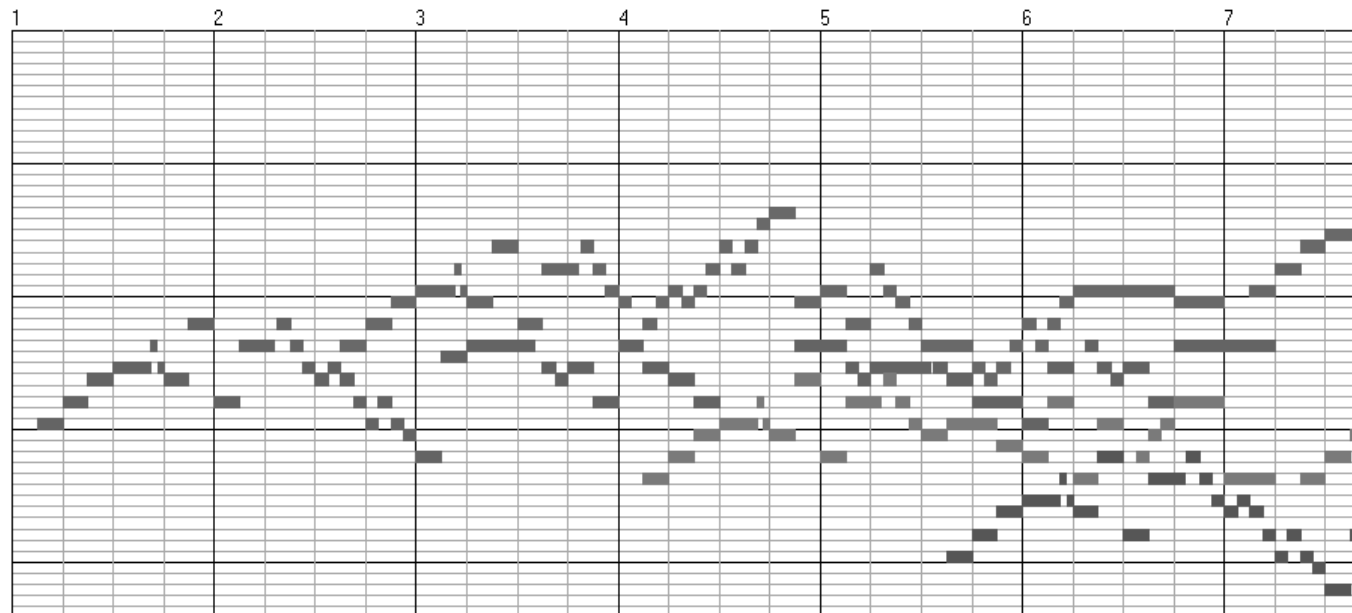
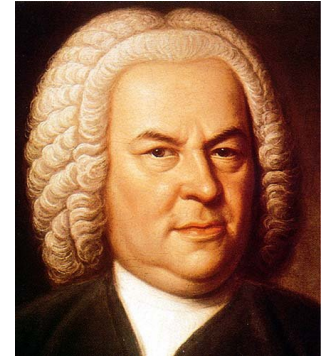


Piano Roll Representation (MIDI)

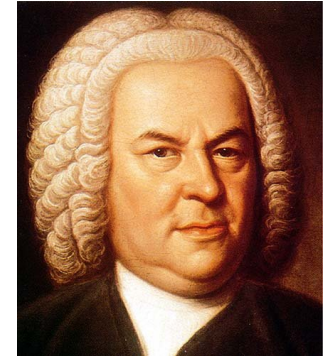
Query:



Goal: Find all occurrences of the query



Piano Roll Representation (MIDI)

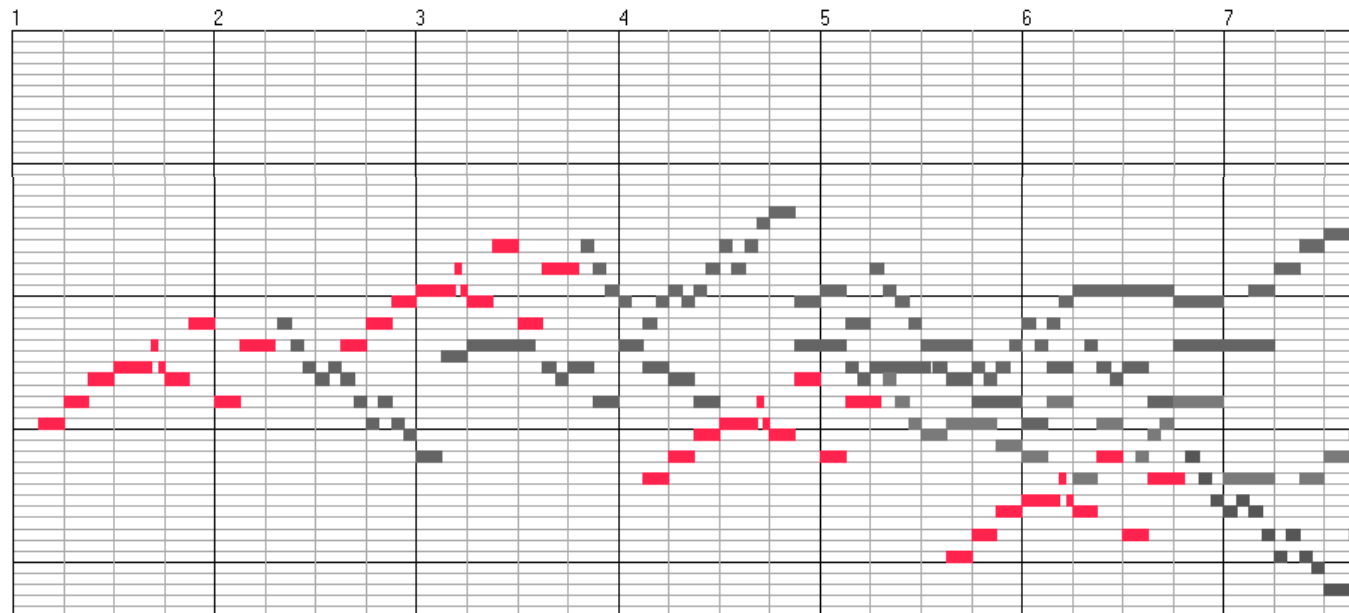


Query:



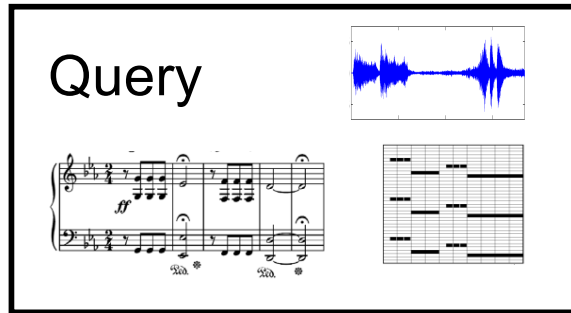
Goal: Find all occurrences of the query

Matches:

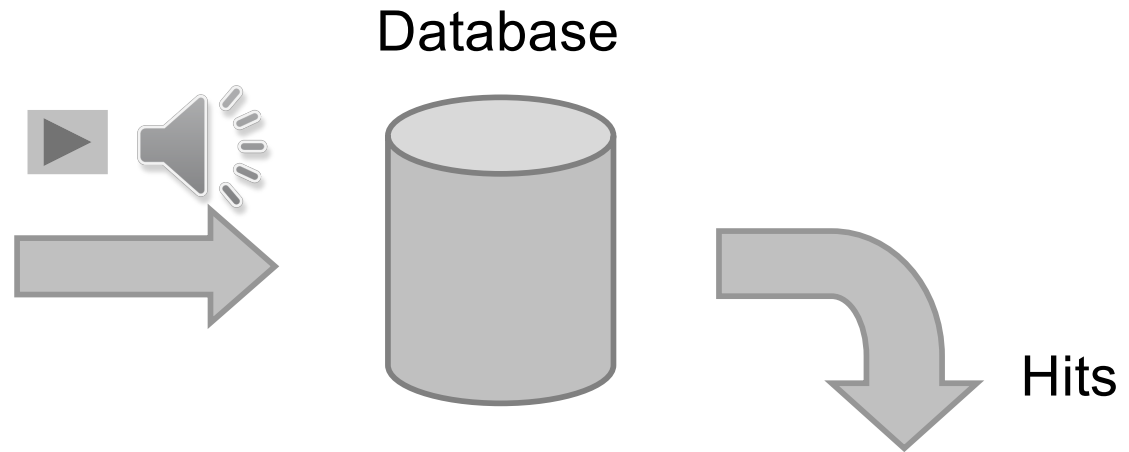


Music Retrieval

Query



The query box contains three visual representations of music: a musical score with treble and bass clefs, a blue waveform representing the audio signal, and a piano roll showing pitch and duration over time.



Retrieval tasks:

Audio identification

Audio matching

Version identification

Category-based music retrieval

Bernstein (1962)
Beethoven, Symphony No. 5

Beethoven, Symphony No. 5:

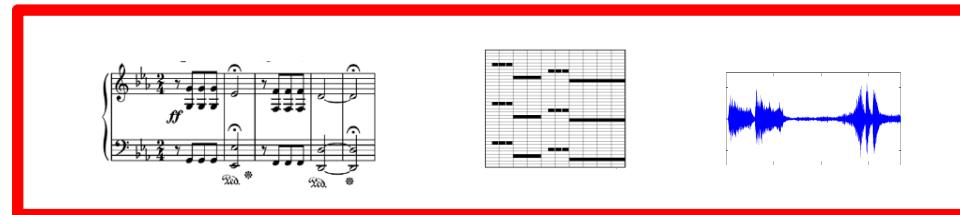
- Bernstein (1962)
- Karajan (1982)
- Gould (1992)

- Beethoven, Symphony No. 9
- Beethoven, Symphony No. 3
- Haydn Symphony No. 94



Music Retrieval

Modalities



Retrieval tasks:

Audio identification

Audio matching

Version identification

Category-based music retrieval

Specificity

High specificity



Low specificity

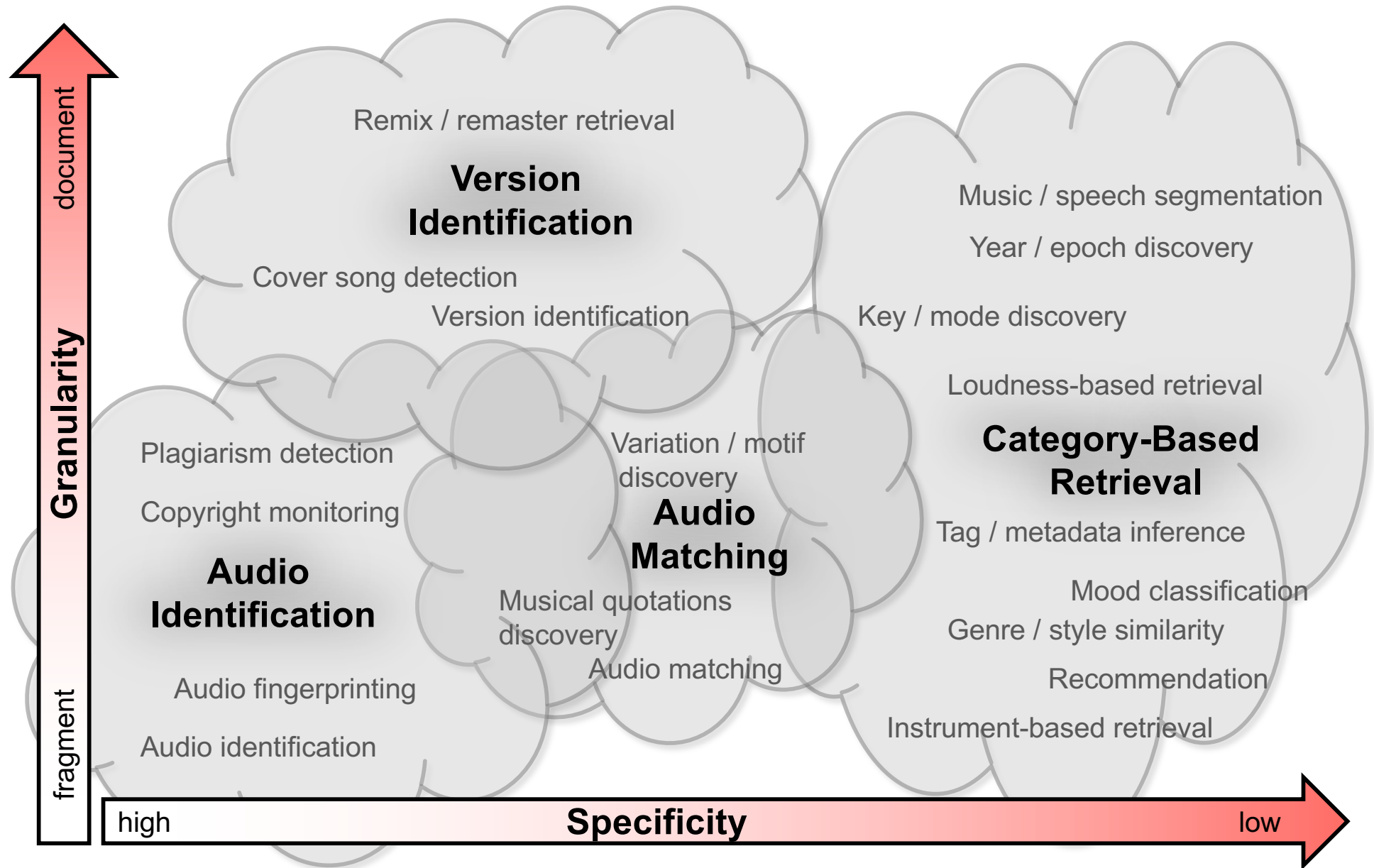
Granularity

Fragment-based retrieval

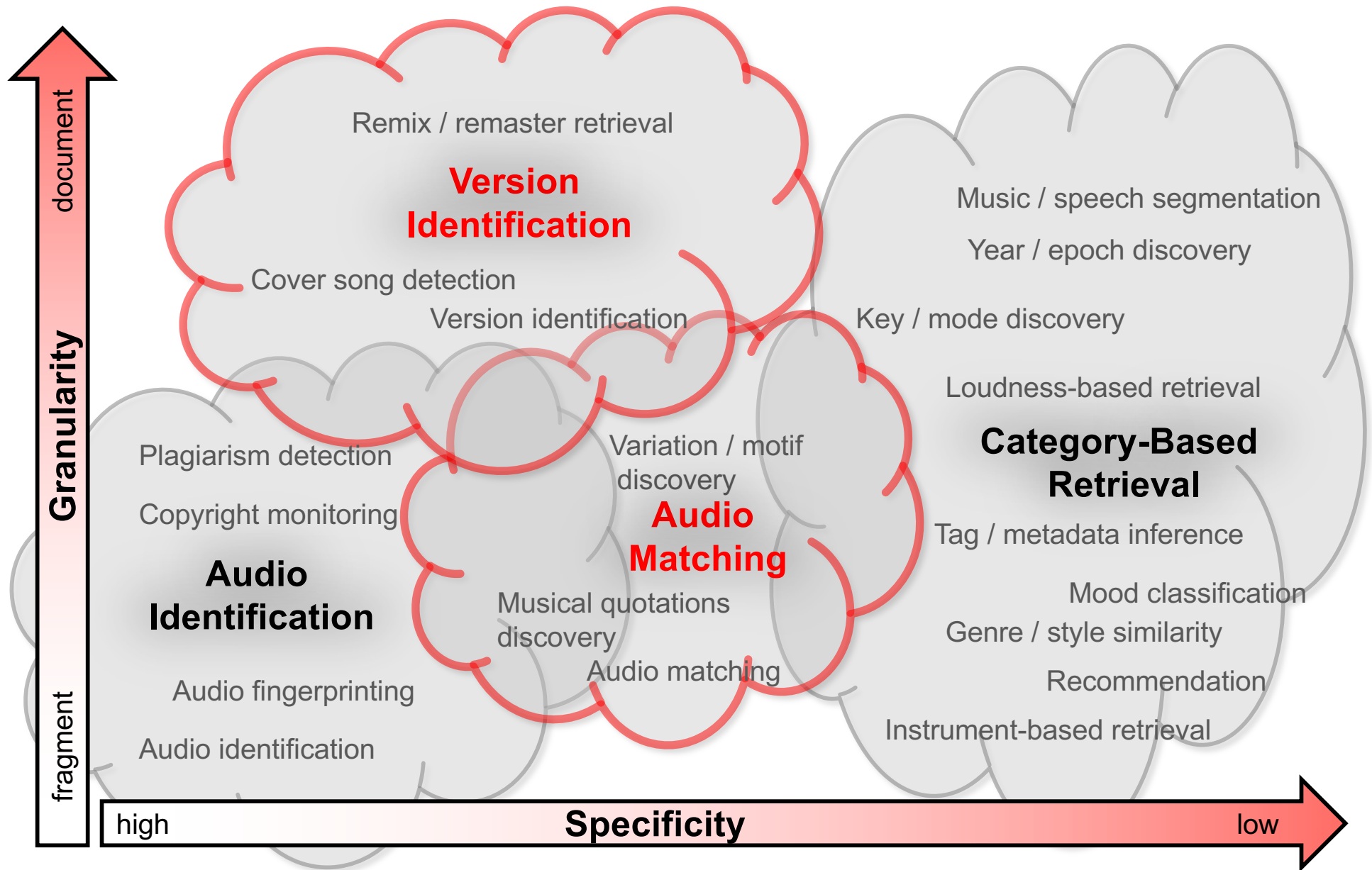


Document-based retrieval

Music Retrieval



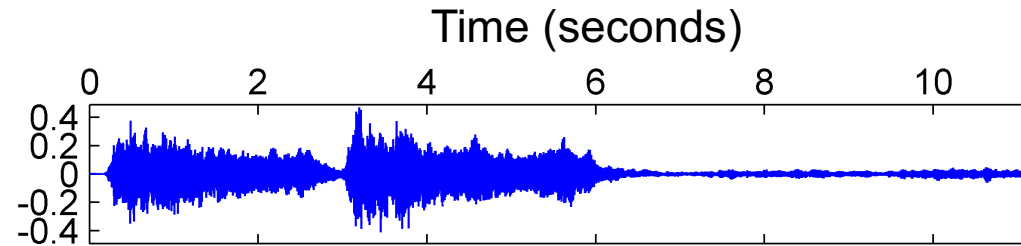
Music Retrieval



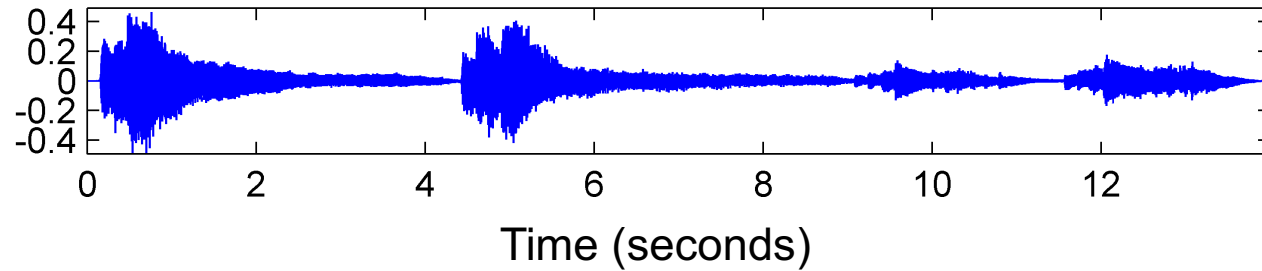
Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan



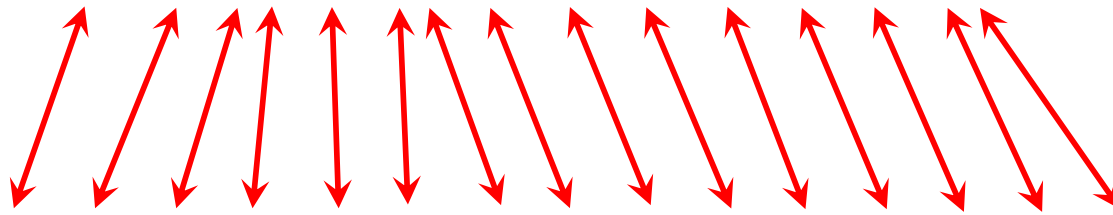
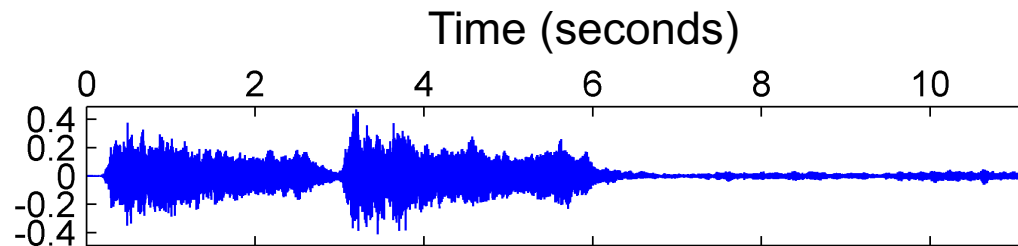
Gould



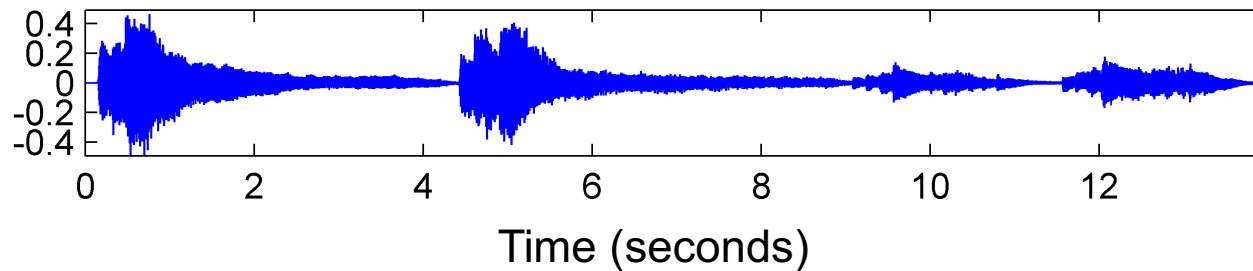
Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan 



Gould 



Music Synchronization: Audio-Audio

Task

Given: Two different audio recordings (two versions) of the same underlying piece of music.

Goal: Find for each position in one audio recording the **musically** corresponding position in the other audio recording.

Music Synchronization: Audio-Audio

Two main steps:

1.) Feature extraction

- Robust to variations (e.g., instrumentation, timbre, dynamics)
- Discriminative (e.g., capturing harmonic, melodic, tonal aspects)

➡ **Chroma features**

2.) Temporal alignment

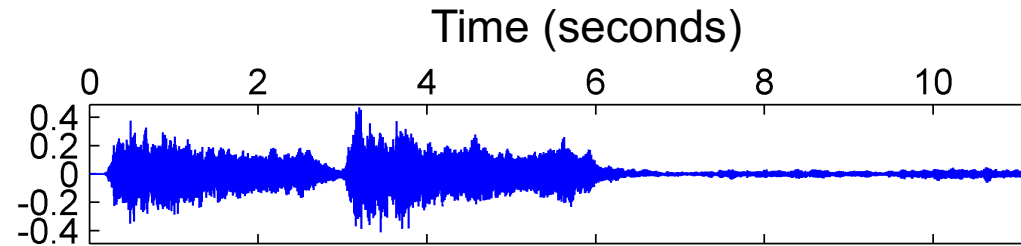
- Capturing local and global tempo variations
- Trade-off: Robustness vs. accuracy
- Efficiency

➡ **Dynamic time warping (DTW)**

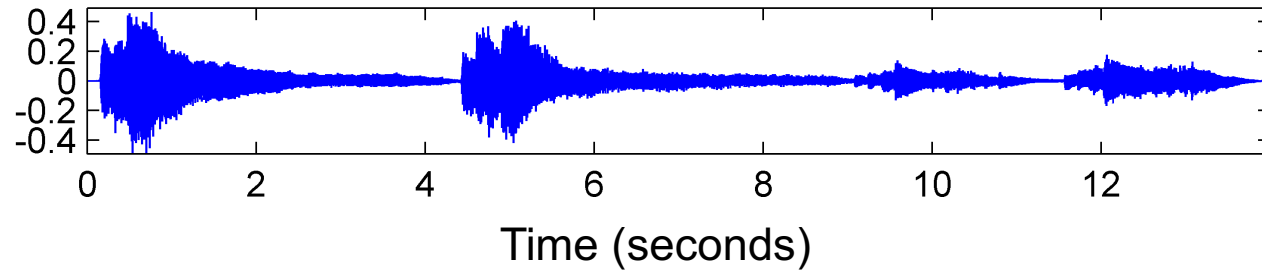
Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan



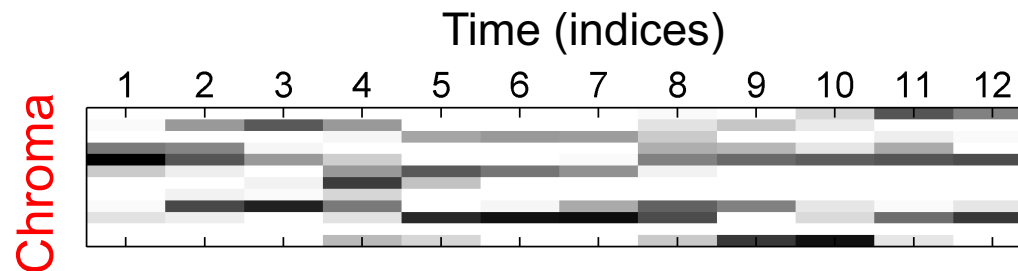
Gould



Music Synchronization: Audio-Audio

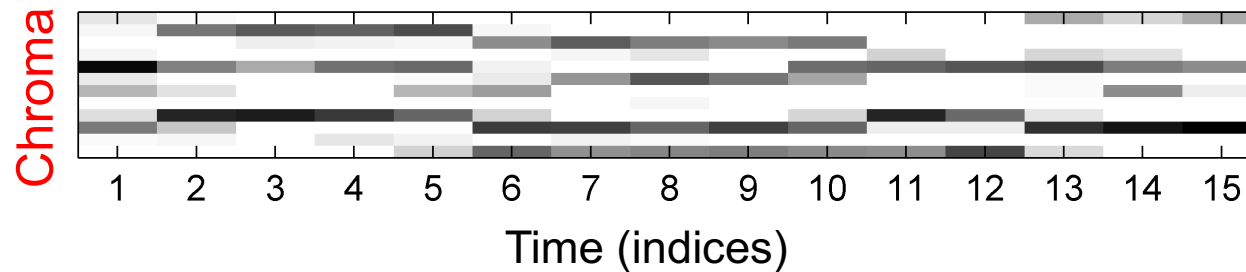
Beethoven's Fifth

Karajan



Time–chroma representations

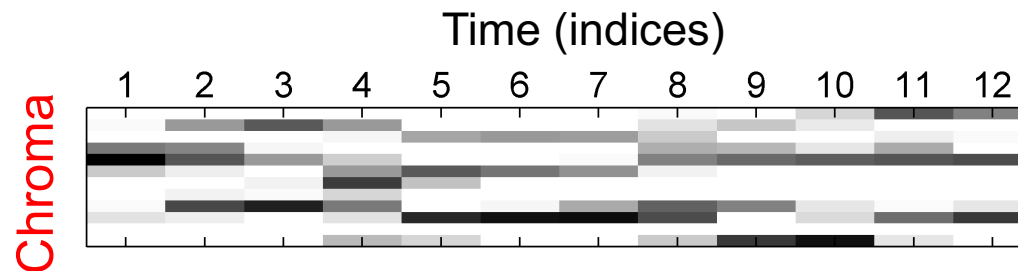
Gould



Music Synchronization: Audio-Audio

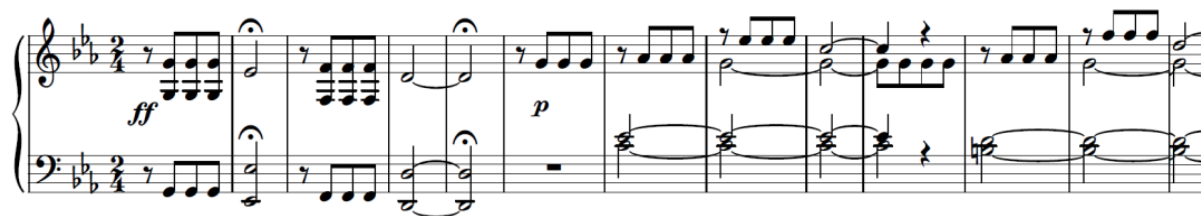
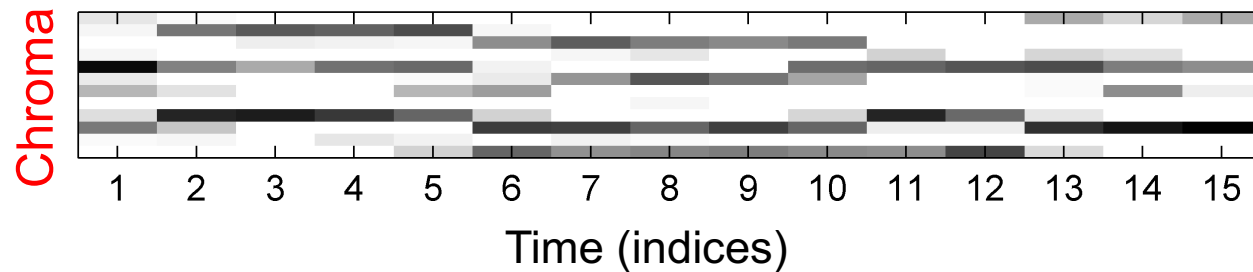
Beethoven's Fifth

Karajan



Time–chroma representations

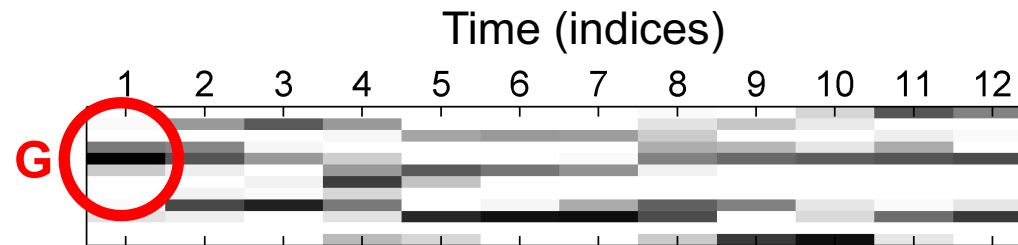
Gould



Music Synchronization: Audio-Audio

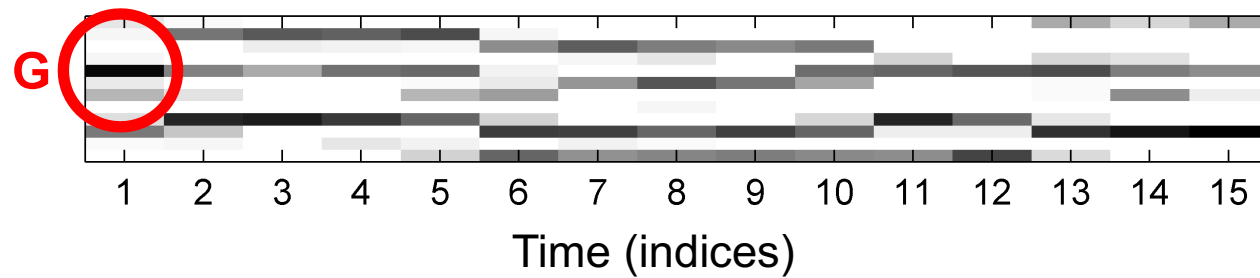
Beethoven's Fifth

Karajan



Time–chroma representations

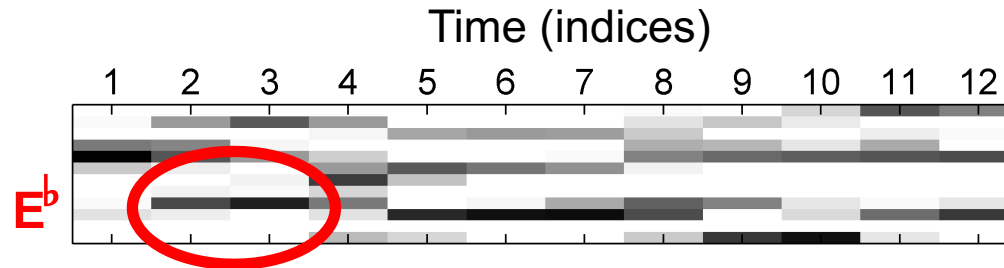
Gould



Music Synchronization: Audio-Audio

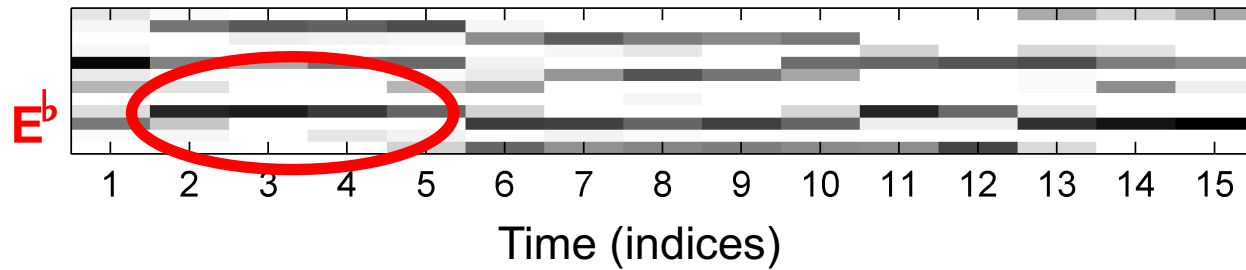
Beethoven's Fifth

Karajan

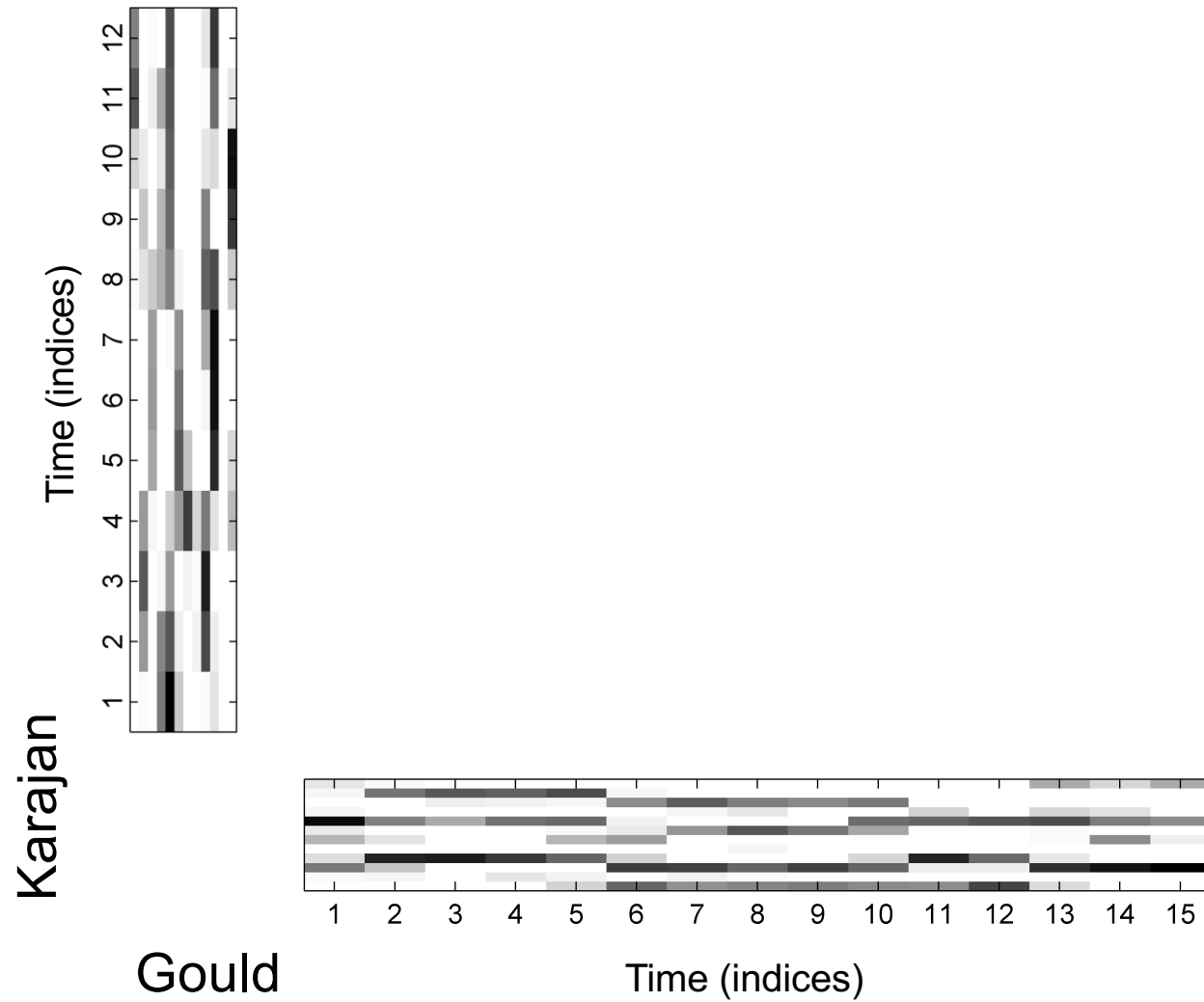


Time–chroma representations

Gould

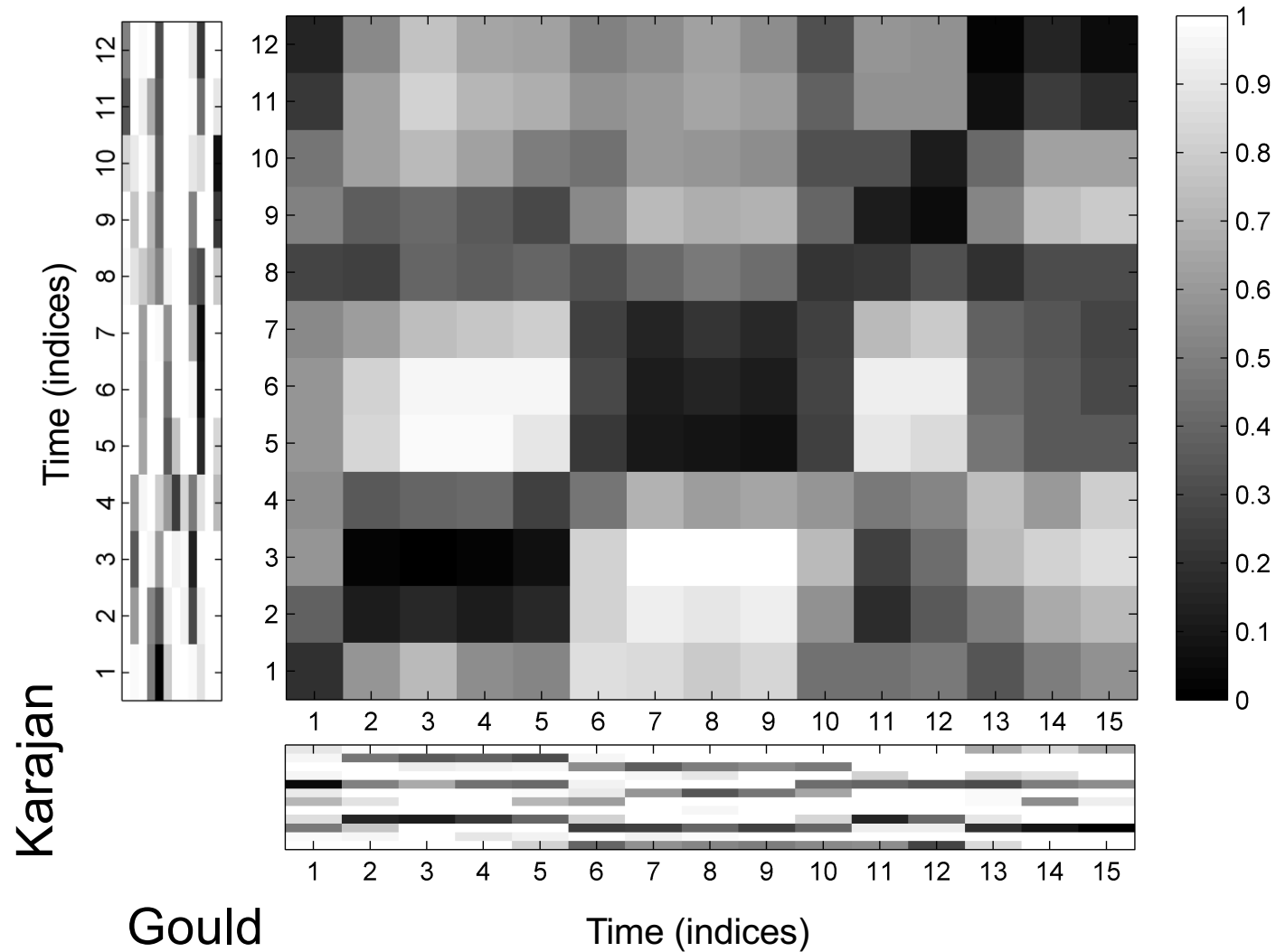


Music Synchronization: Audio-Audio



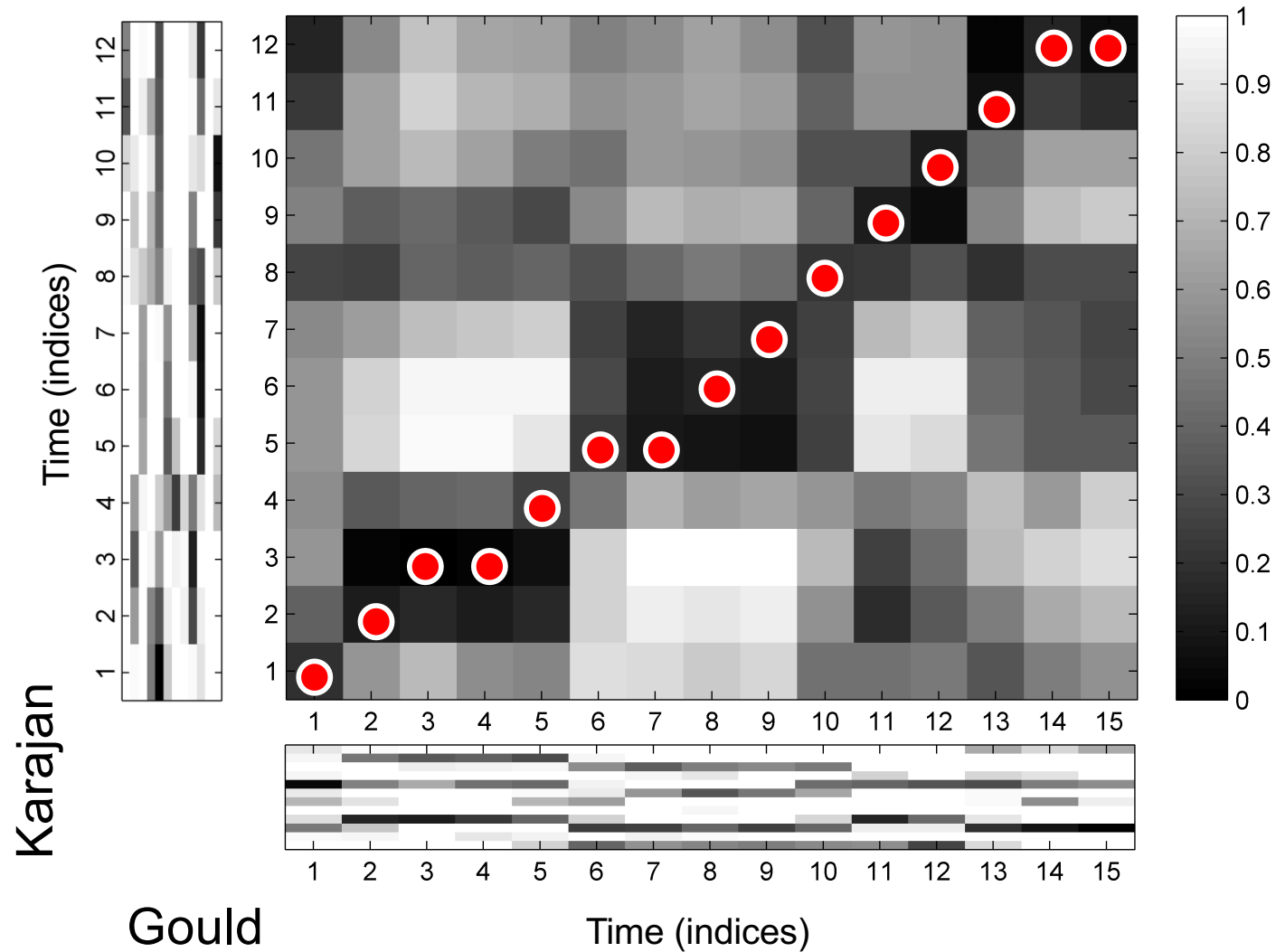
Music Synchronization: Audio-Audio

Cost matrix



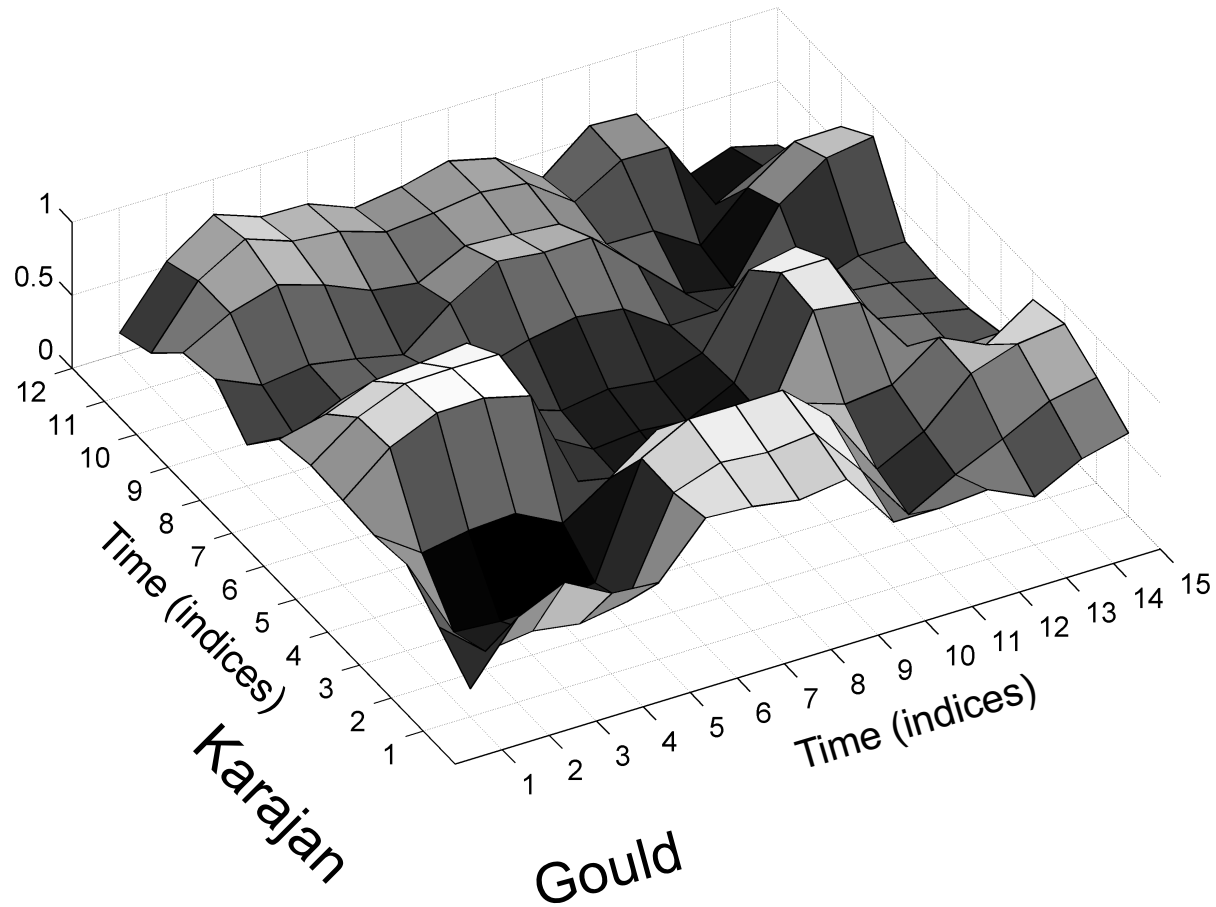
Music Synchronization: Audio-Audio

Cost-minimizing warping path



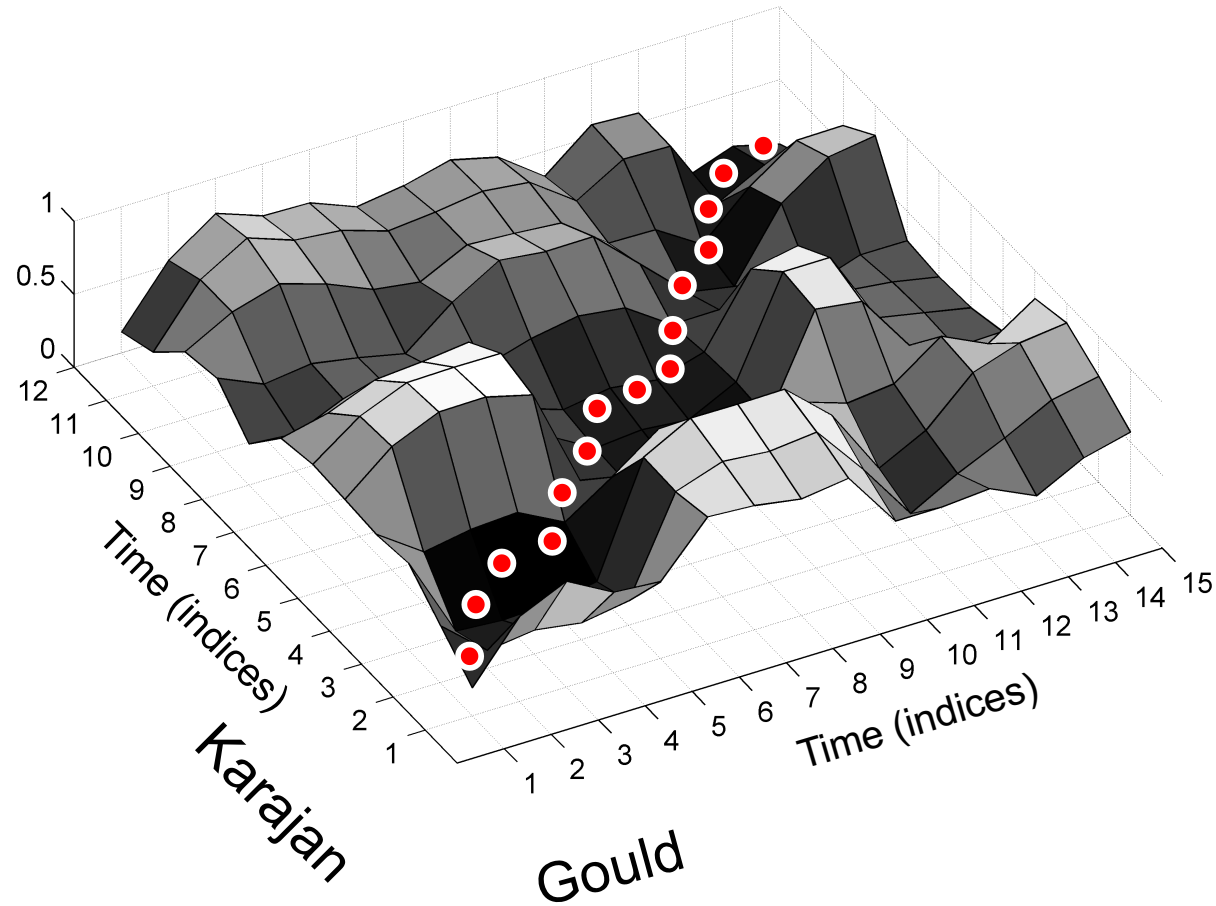
Music Synchronization: Audio-Audio

Cost matrix



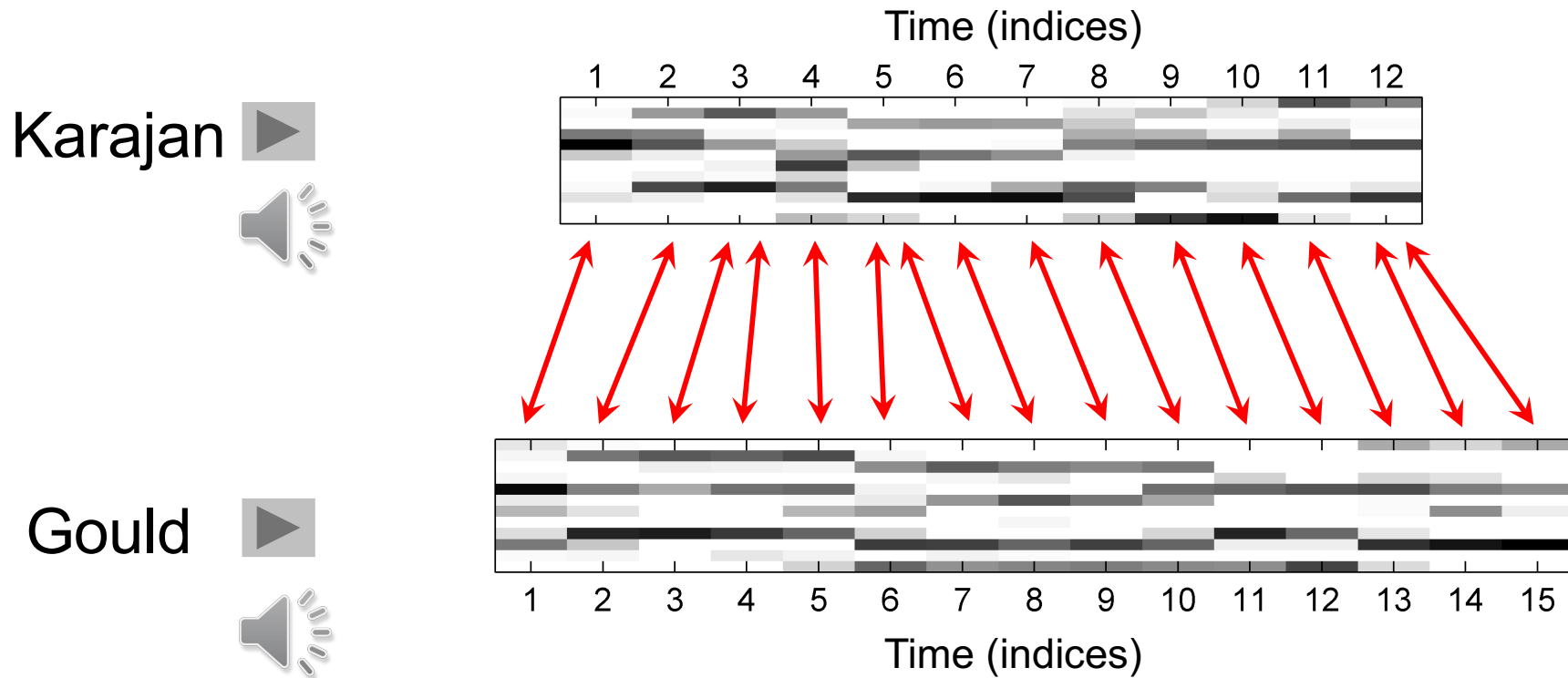
Music Synchronization: Audio-Audio

Cost-minimizing warping path



Music Synchronization: Audio-Audio

Optimal alignment (cost-minimizing warping path)



Application: Interpretation Switcher

The screenshot shows a software window titled "Interpretation Switcher" for the piece "Beethoven, Op067-1_Symphony5". The interface features four horizontal progress bars, each representing a different interpretation: "midi", "Bernstein", "Sawallisch", and "Scherbakov". Each bar is divided into three segments: blue, red, and green. A playhead icon is positioned at the start of each bar, and a timestamp of "00:00.00" is displayed at the end of each bar. On the right side, a list of checkboxes allows users to select or deselect interpretations: "midi", "Bernstein", "Sawallisch", and "Scherbakov", all of which are currently checked. A "Deselect all" button is located below this list. The bottom of the window contains a control panel with the following elements: a radio button for "Absolute" (checked) and "Relative Reference"; a play button icon; a square stop button icon; a "Movement selection" button with an upward-pointing arrow icon; a checkbox for "Interval Repeat" (unchecked); and an "Info" button with a question mark icon. A separate play button icon is also visible on the right side of the slide.

Music Synchronization: Image-Audio

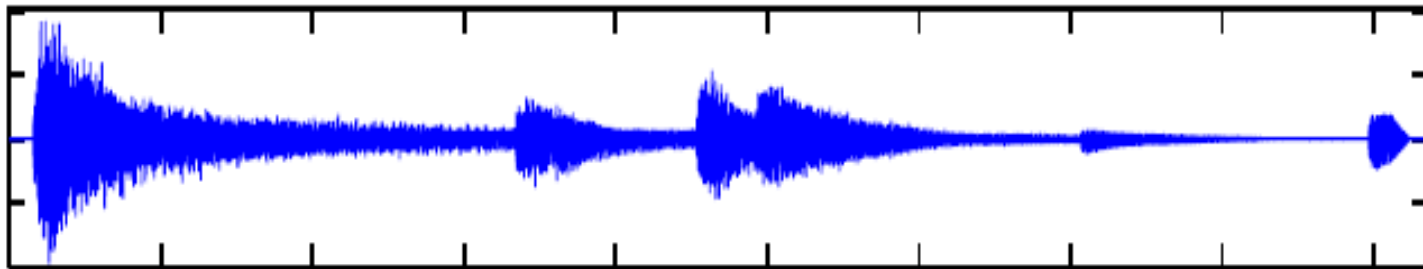
Image

Grave.

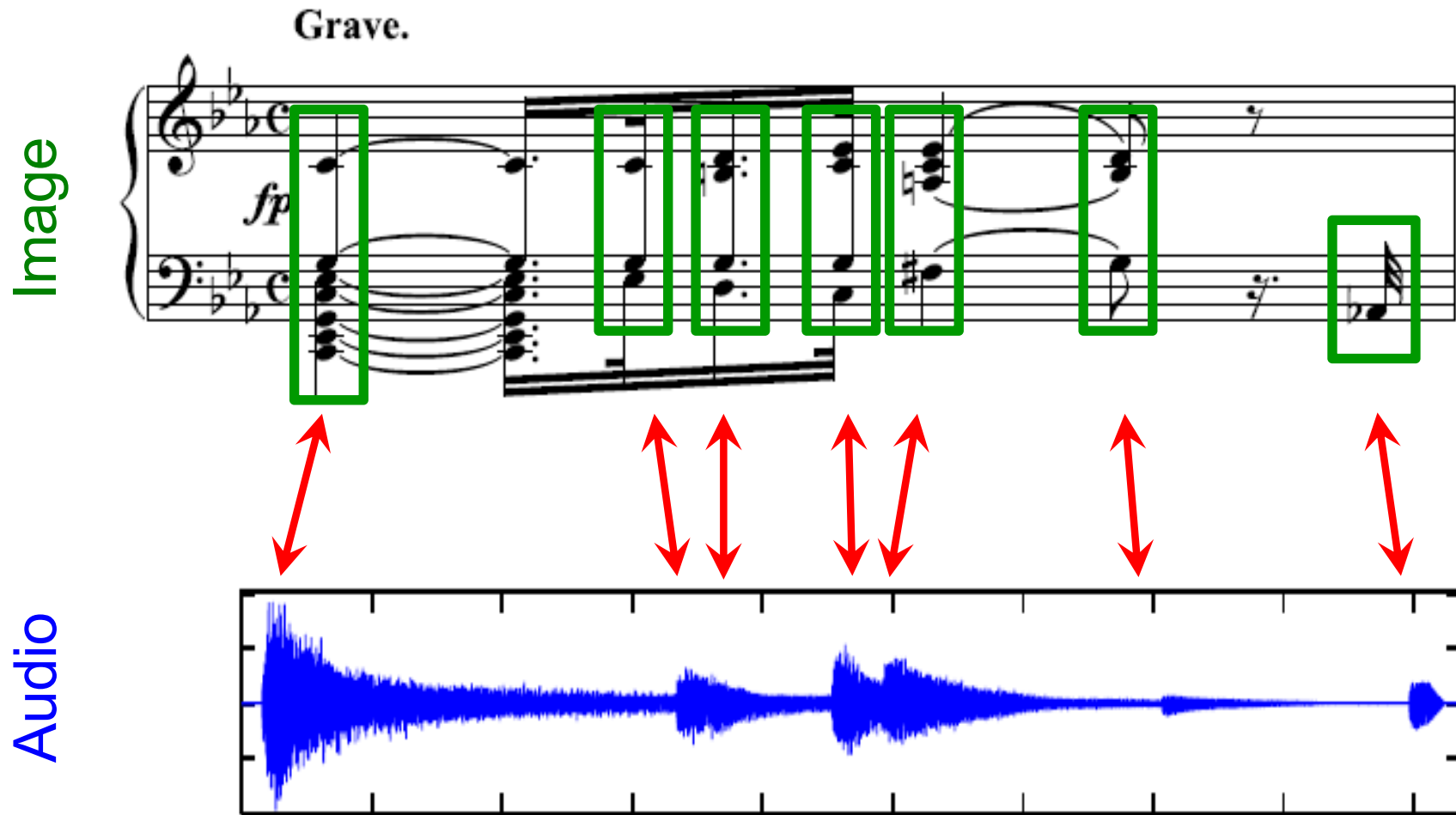


The image shows a musical score for piano, marked "Grave." and "fp". The score is written in a grand staff with a treble and bass clef. The time signature is common time (C). The music features a slow, somber melody with a prominent bass line. The first measure is marked "fp" (fortissimo piano). The score consists of several measures, including a long note in the treble clef and a complex bass line with many notes.

Audio



Music Synchronization: Image-Audio

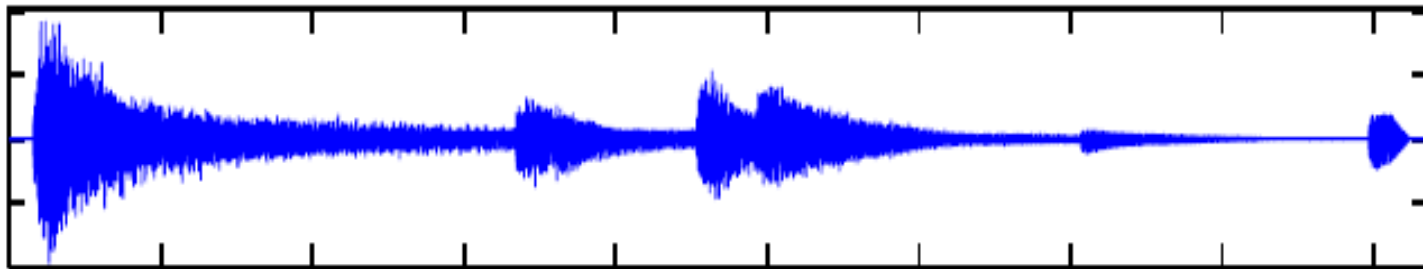


How to make the data comparable?

Image



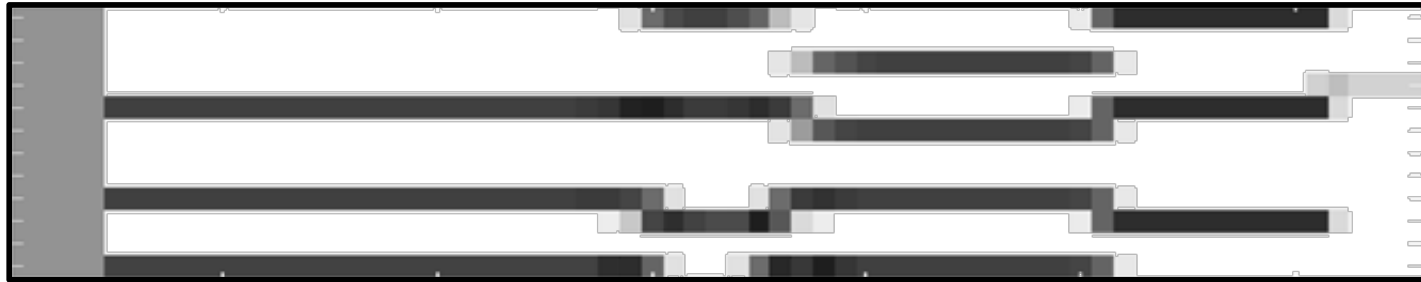
Audio



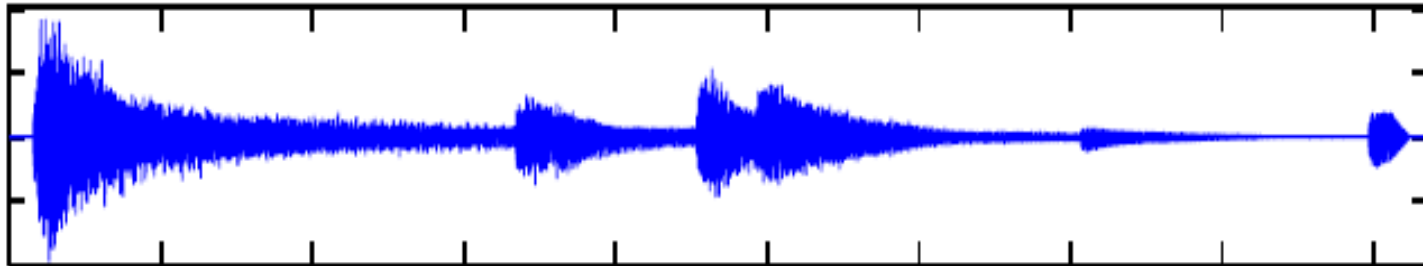
How to make the data comparable?

Image Processing: Optical Music Recognition

Image



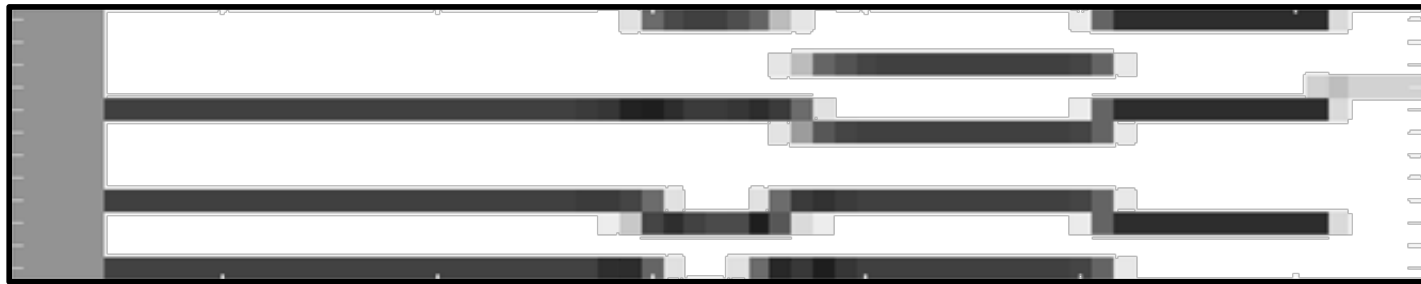
Audio



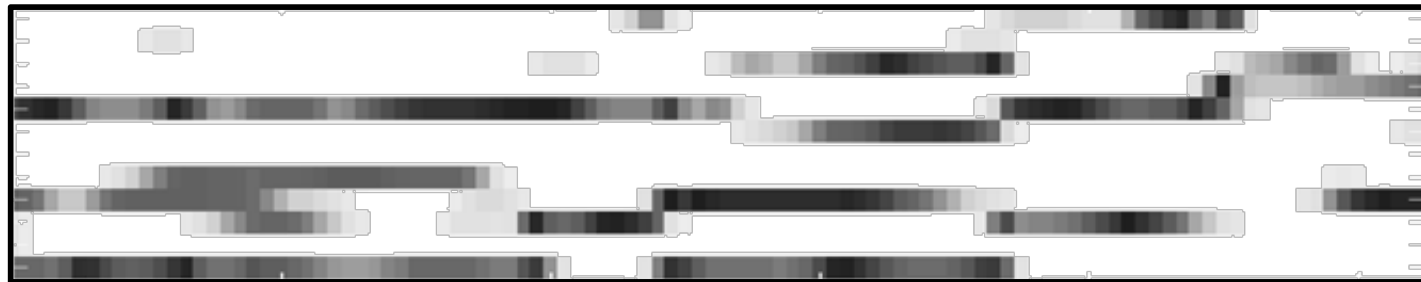
How to make the data comparable?

Image Processing: Optical Music Recognition

Image



Audio



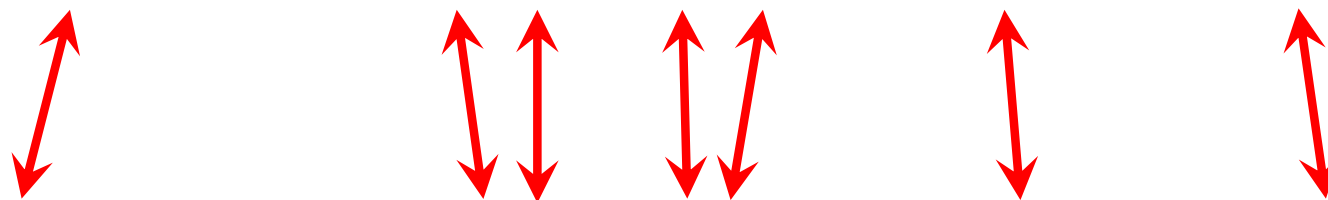
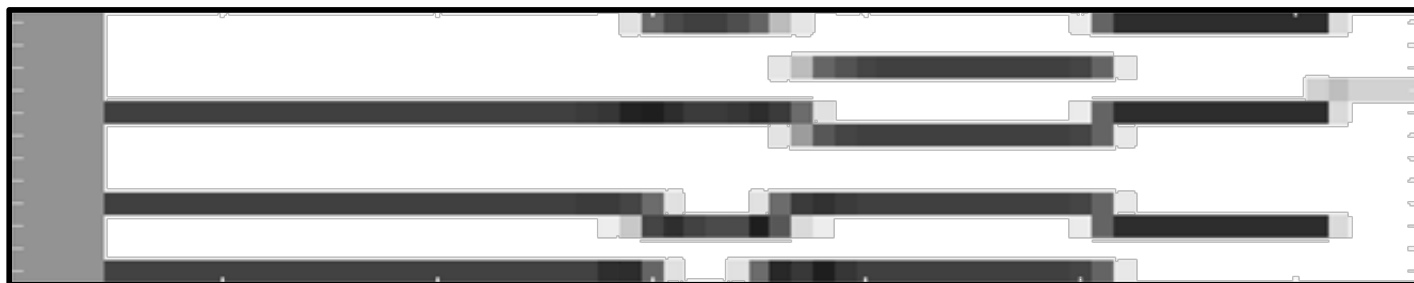
Audio Processing: Fourier Analysis



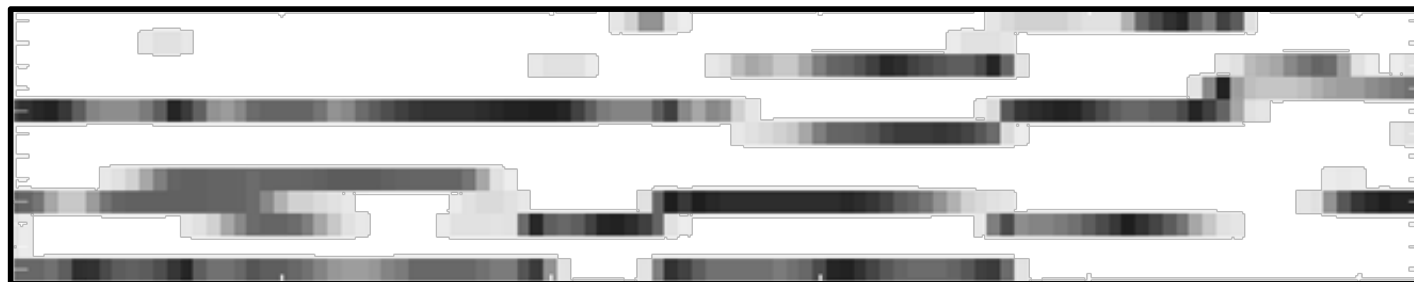
How to make the data comparable?

Image Processing: Optical Music Recognition

Image



Audio



Audio Processing: Fourier Analysis



Application: Score Viewer

The screenshot displays two windows from a music application. The top window, titled "ScoreViewer", shows a musical score for "Beethoven - Klaviersonaten Band 1 - Henle". The score is for "Sonata no.8 in C minor, op.13 'Pathétique' / Rondo (Allegro)". The score is displayed in a split view, with the left page showing the beginning of the piece and the right page showing the Rondo section. The score is currently at track 29 of 54, bar 1 of 211, and page 159 of 285. The score is following on, and the play button is active. The bottom window, titled "AudioViewer", shows a list of tracks for "Beethoven - Piano Sonatas - Alfred Brendel". The list includes tracks 03 through 11, with track 11, "Sonata no.8 in C minor, op.13 'Pathétique' / Rondo (Allegro)", selected. The audio player shows track 11 of 11, and the time is 00:00.00 / 4:30.35. The play button is active.

ScoreViewer
Beethoven - Piano Sonatas-Alfred Brendel
Beethoven - Klaviersonaten Band 1 - Henle
Sonata no.8 in C minor, op.13 "Pathétique" / Rondo (Allegro)

Rondo
Allegro

Track: 29 / 54 Bar: 1 / 211 Page: 159 / 285
Score Following On Play Stop

AudioViewer
Beethoven - Piano Sonatas - Alfred Brendel

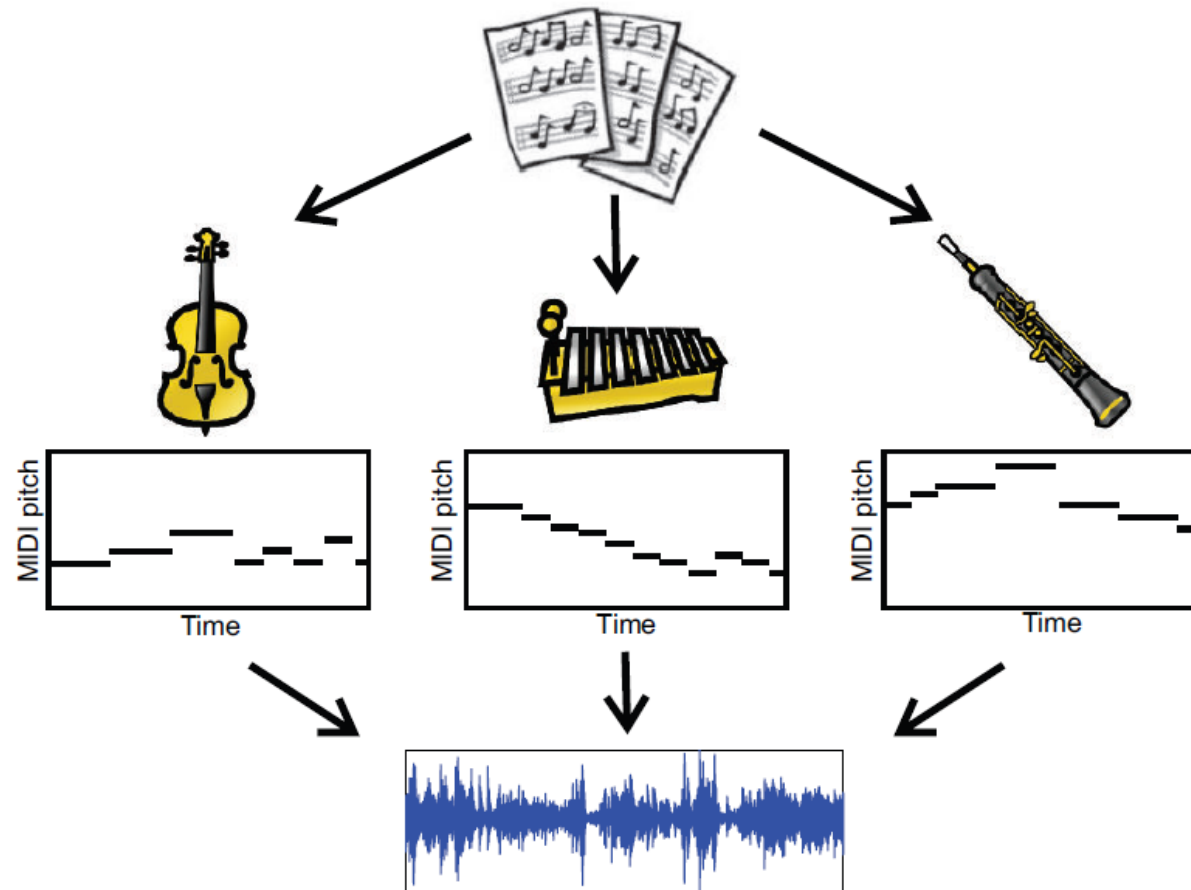
Disc 1

03	Sonata no.1 in F minor, op.2 no.1 / Menuetto (Allegretto)	3:24
04	Sonata no.1 in F minor, op.2 no.1 / Prestissimo	5:32
05	Sonata no.2 in A major, op.2 no.2 / Allegro vivace	7:15
06	Sonata no.2 in A major, op.2 no.2 / Largo appassionato	6:28
07	Sonata no.2 in A major, op.2 no.2 / Scherzo (Allegretto)	3:30
08	Sonata no.2 in A major, op.2 no.2 / Rondo (Grazioso)	7:03
09	Sonata no.8 in C minor, op.13 "Pathétique" / Adagio di molto e con brio	9:40
10	Sonata no.8 in C minor, op.13 "Pathétique" / Adagio cantabile	5:17
11	Sonata no.8 in C minor, op.13 "Pathétique" / Rondo (Allegro)	4:30

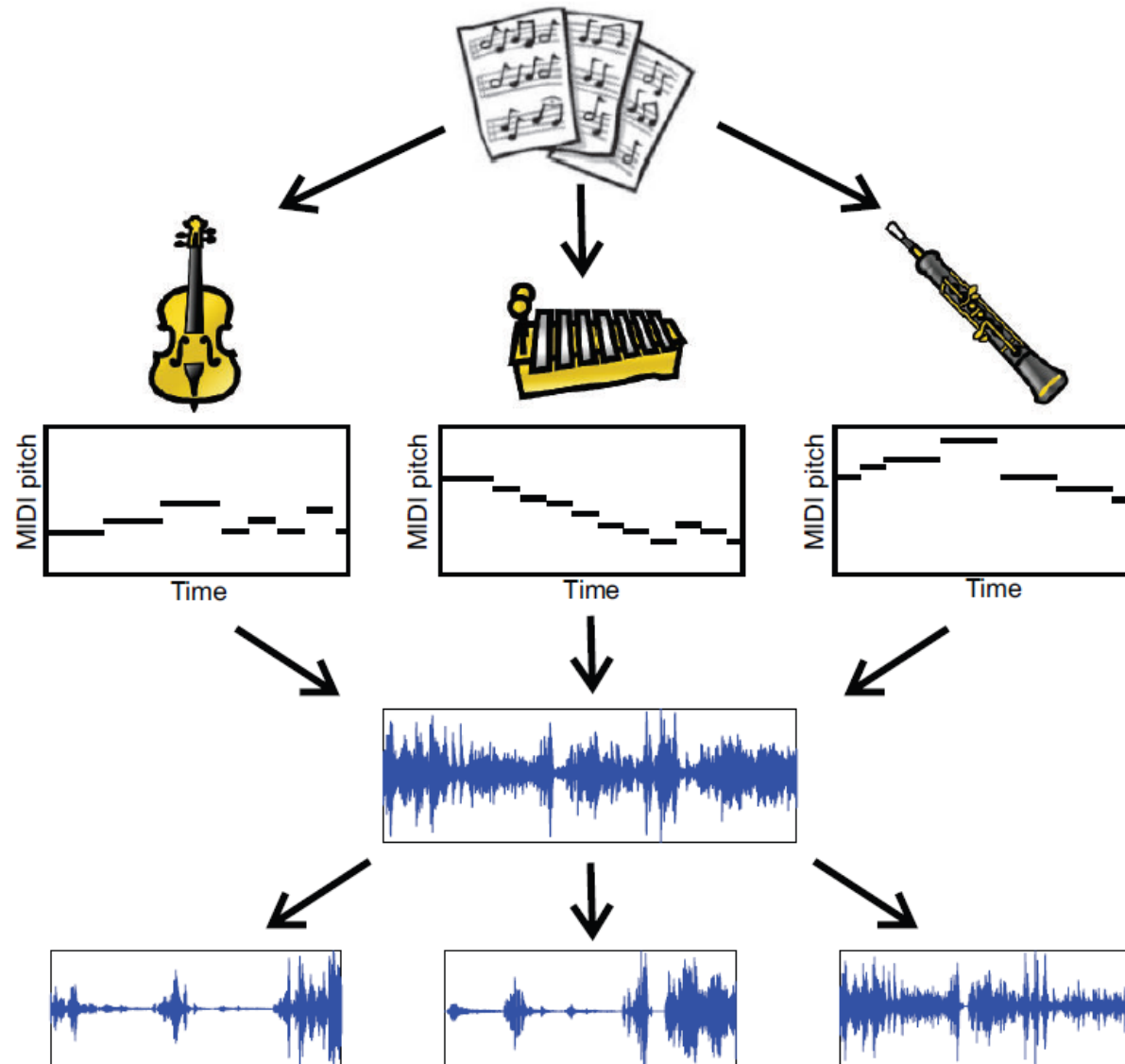
Disc: 1 / 11 Track: 11 / 11 Time: 00:00.00 / 4:30.35
Play Stop



Application: Score-Informed Source Separation

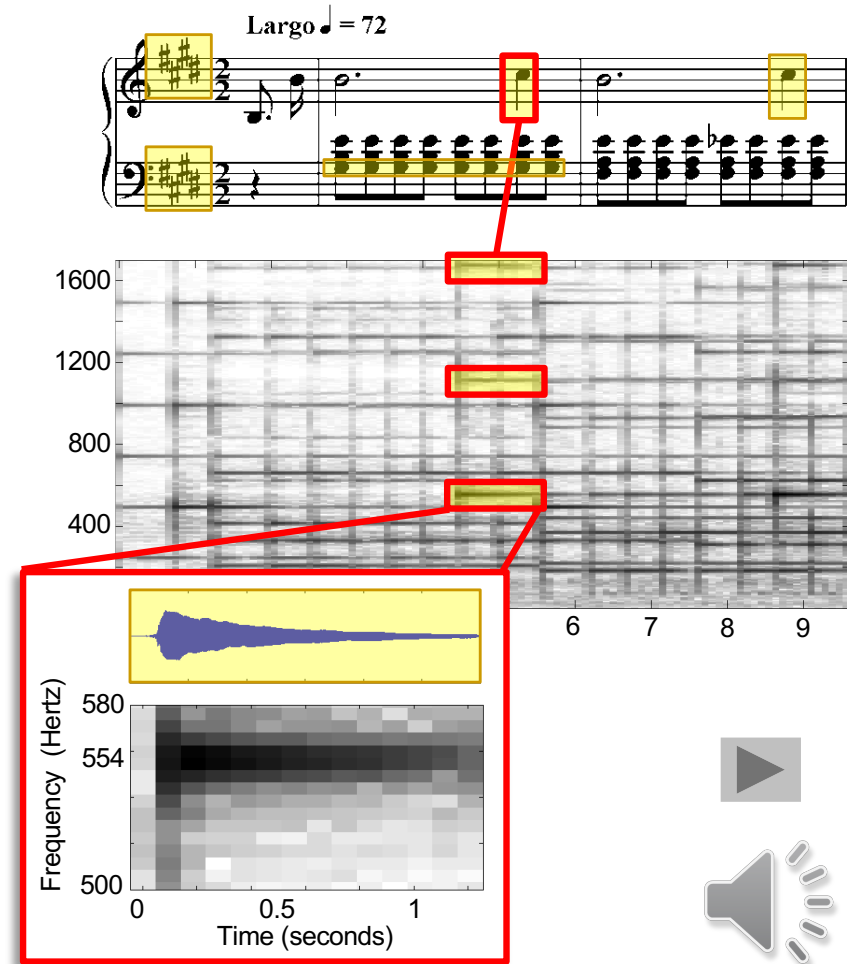
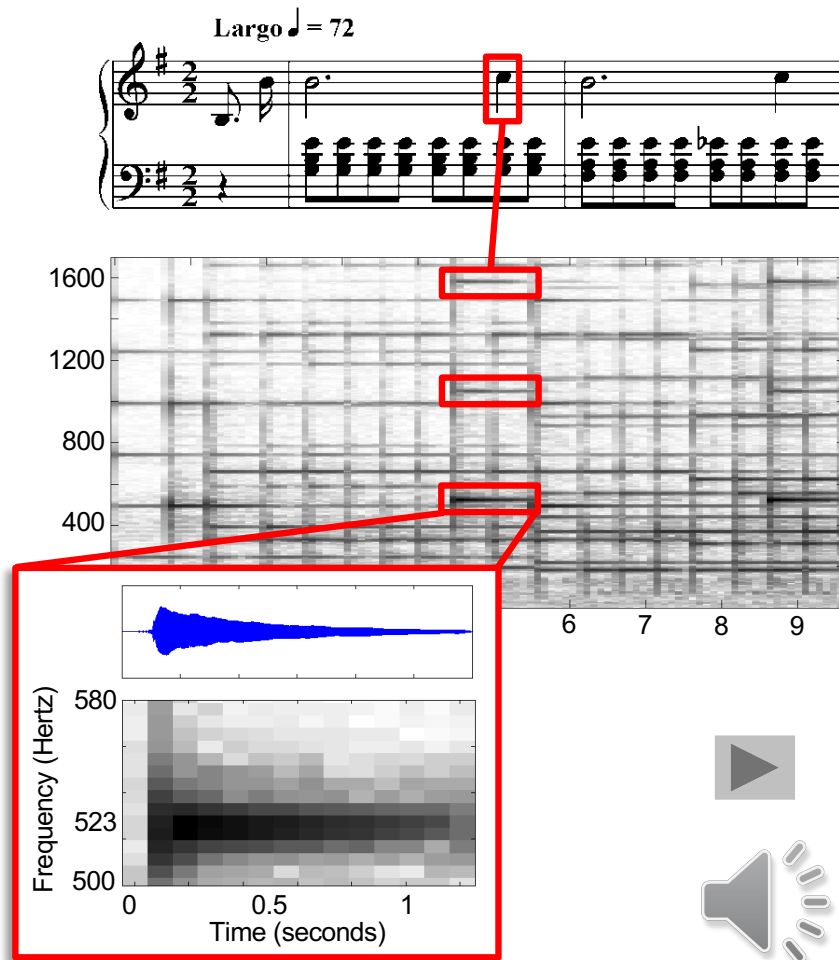


Application: Score-Informed Source Separation



Application: Score-Informed Source Separation

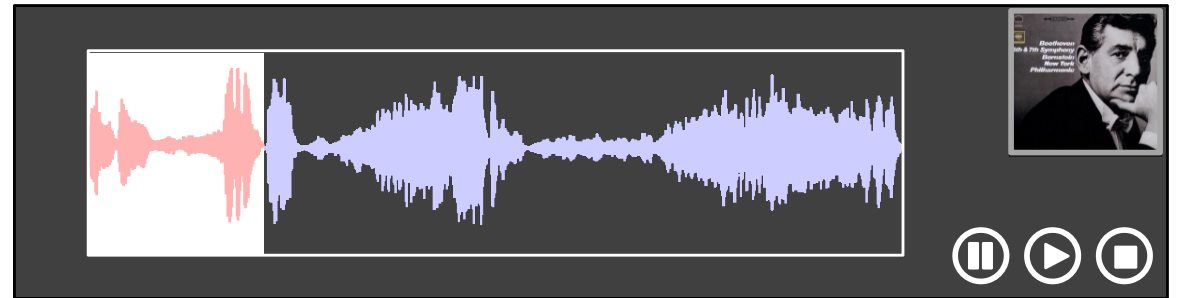
Audio editing



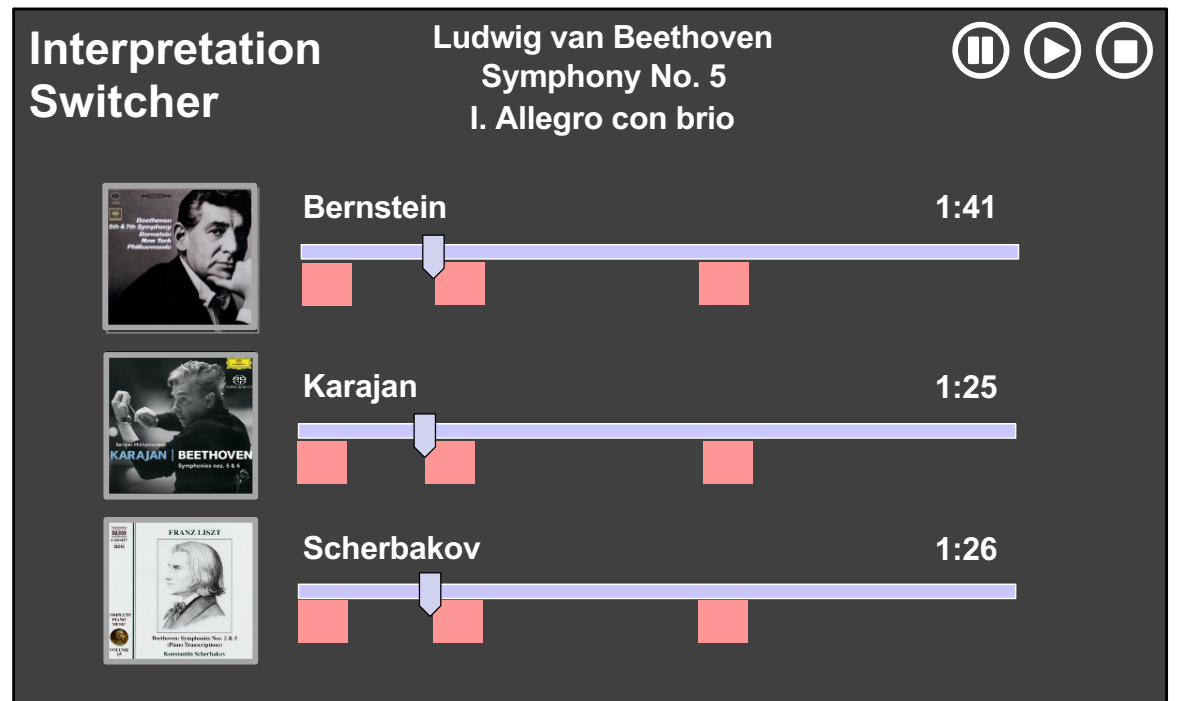
Audio Matching

Task

Query:



Database: Matches

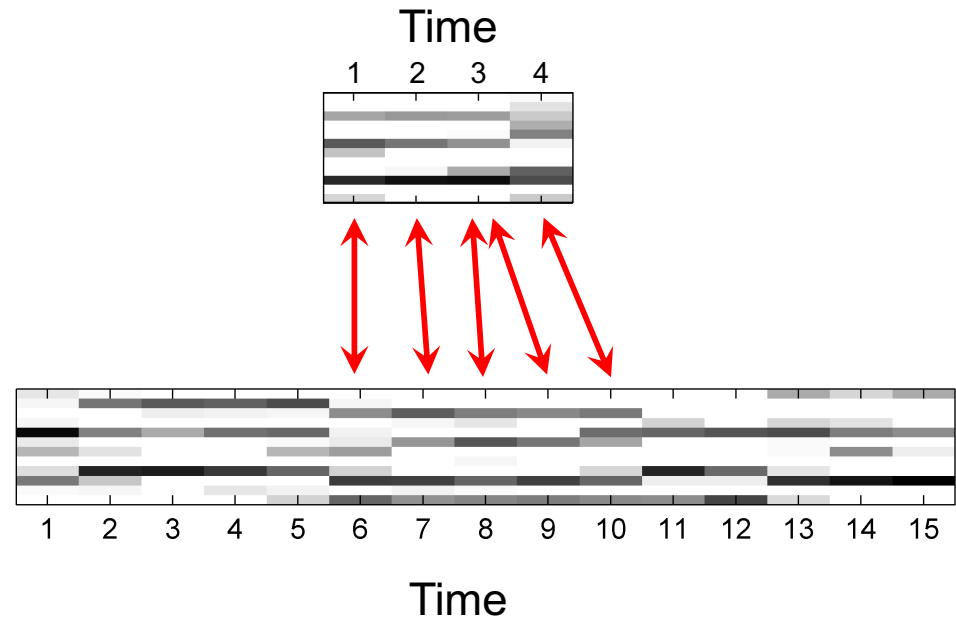


Audio Matching

Task

Query: Sequence X

Database: Sequence Y

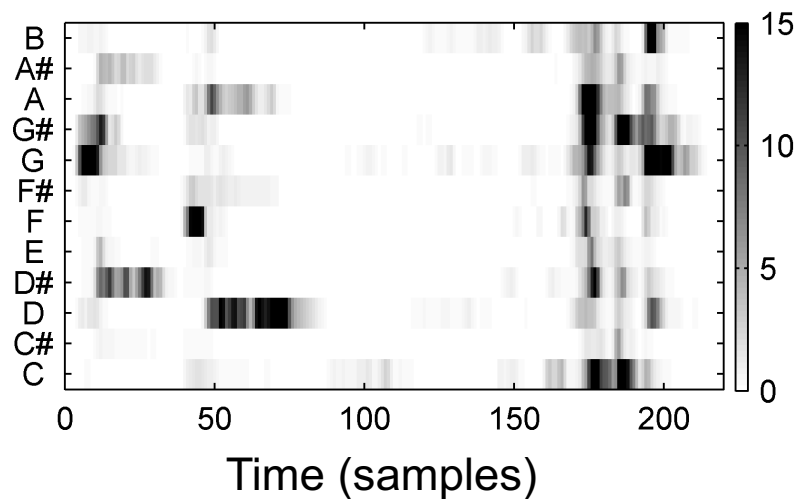


Subsequence matching

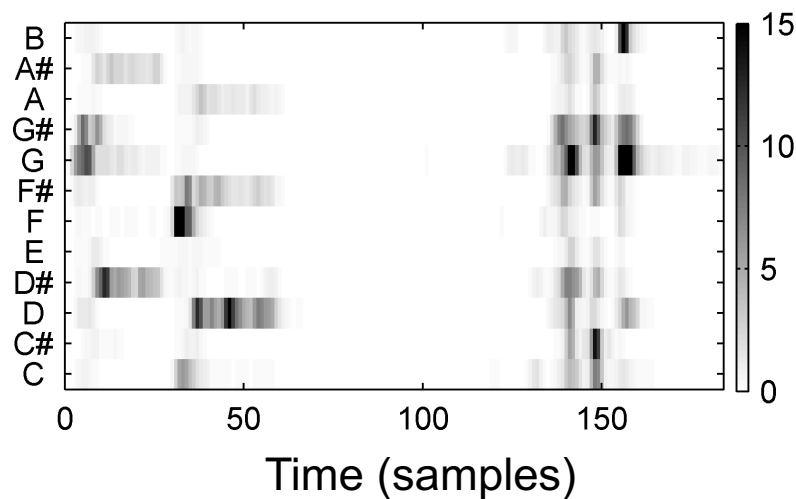
Audio Features

Example: Beethoven's Fifth

Bernstein



Karajan

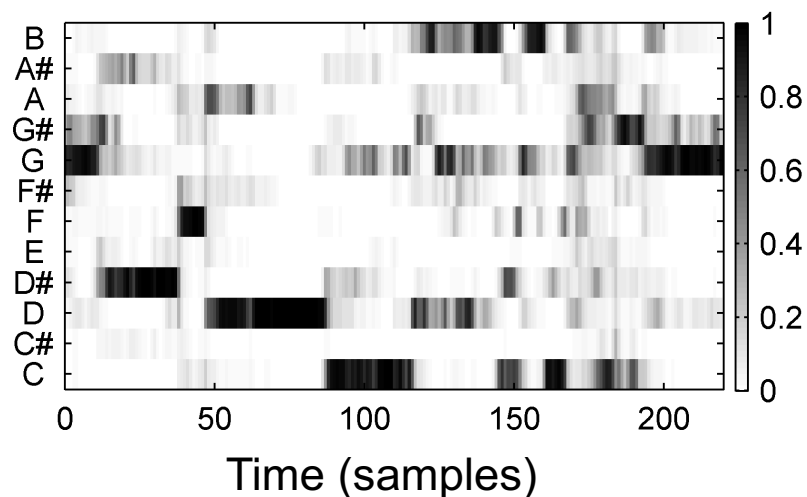


Chroma representation (10 Hz)

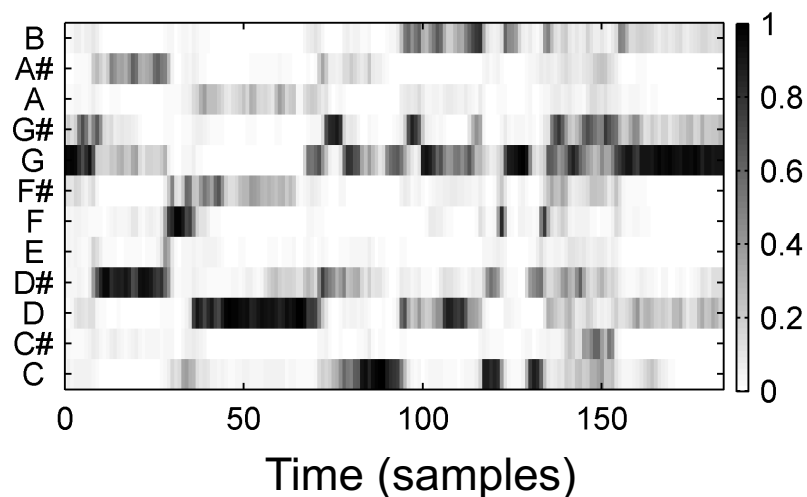
Audio Features

Example: Beethoven's Fifth

Bernstein



Karajan



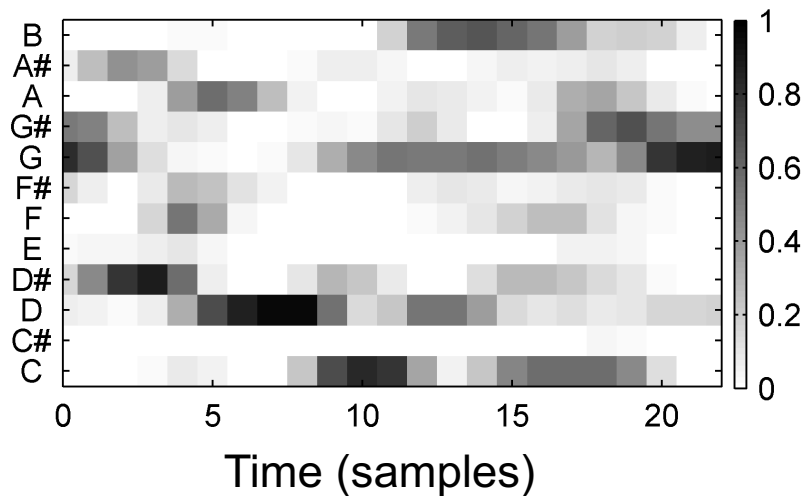
Chroma representation (10 Hz)

- Normalization

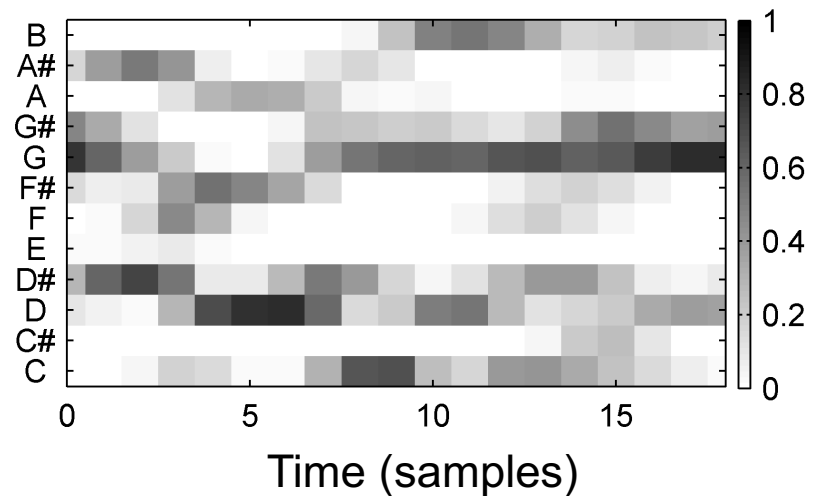
Audio Features

Example: Beethoven's Fifth

Bernstein



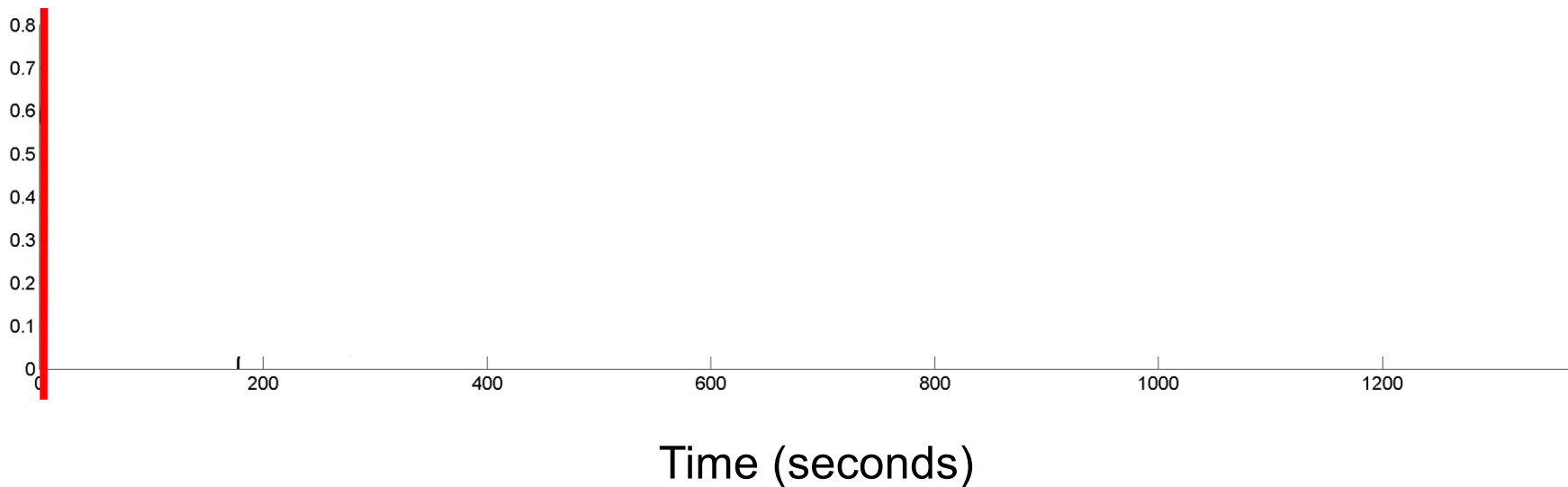
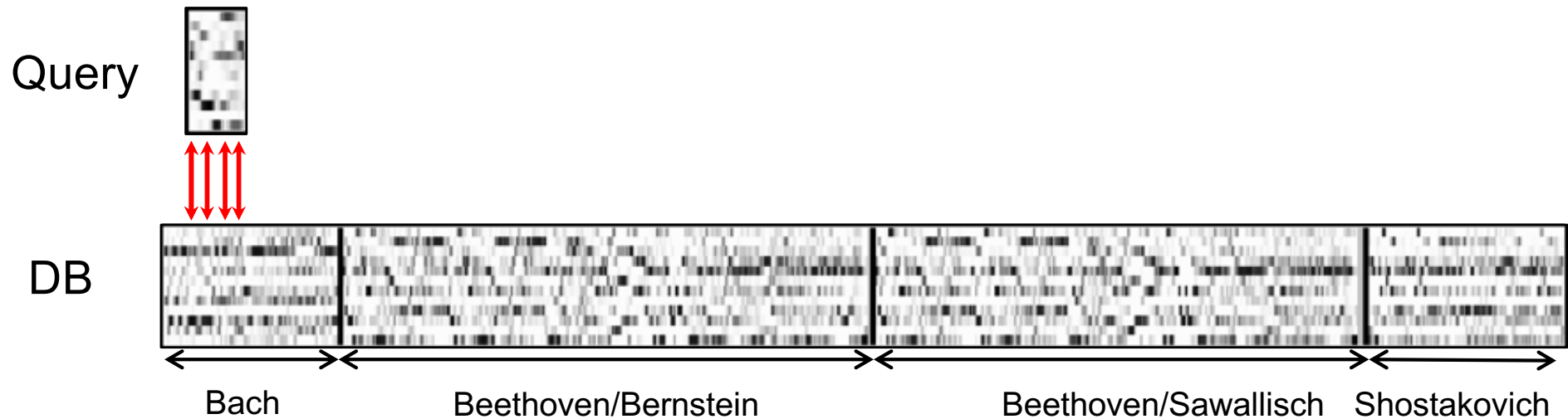
Karajan



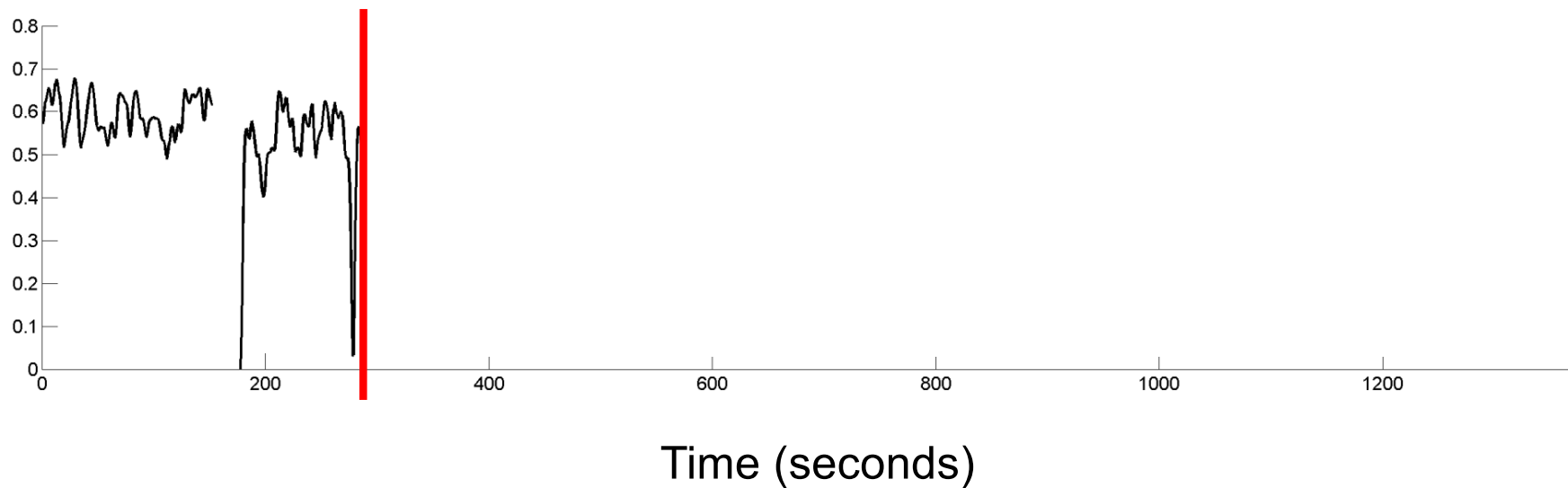
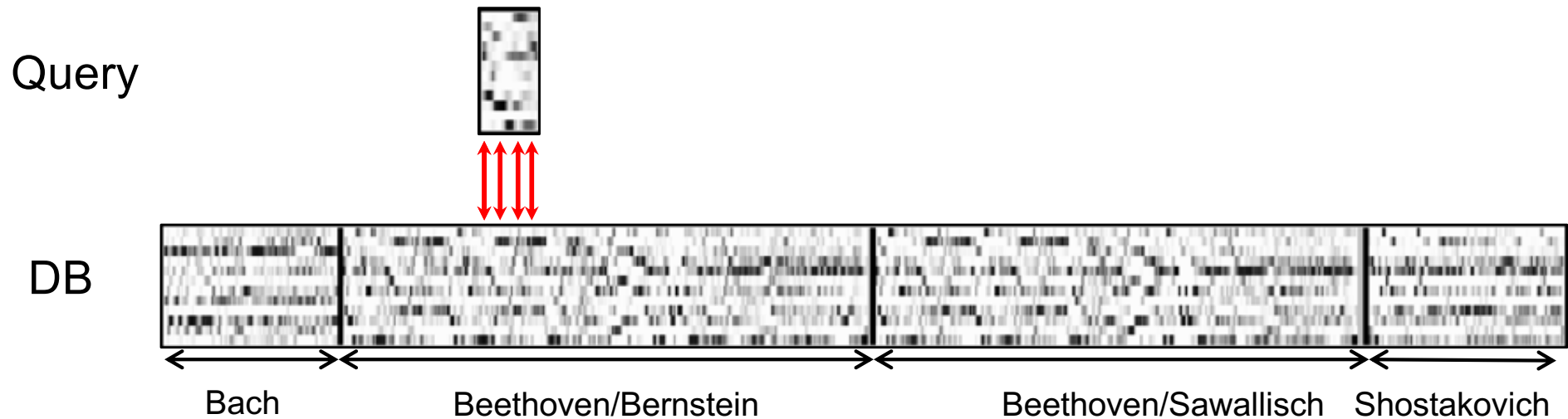
Chroma representation (1 Hz)

- Normalization
- Smoothing & downsampling

Matching Procedure



Matching Procedure

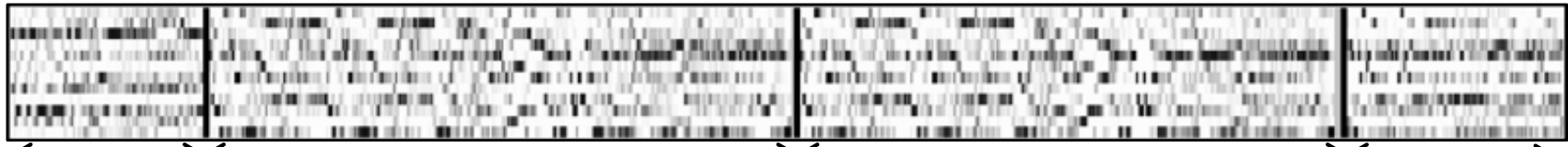


Matching Procedure

Query



DB

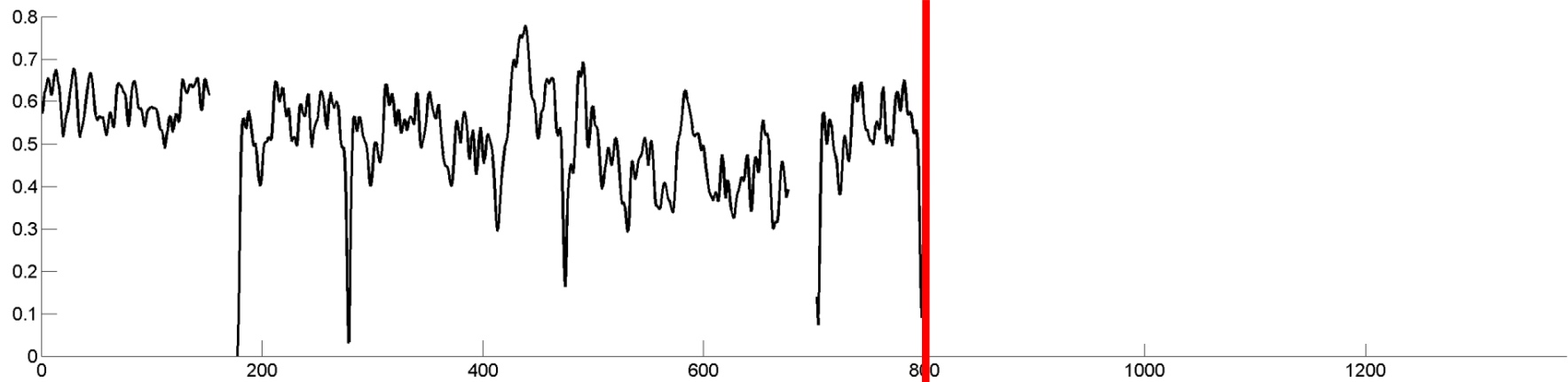


Bach

Beethoven/Bernstein

Beethoven/Sawallisch

Shostakovich



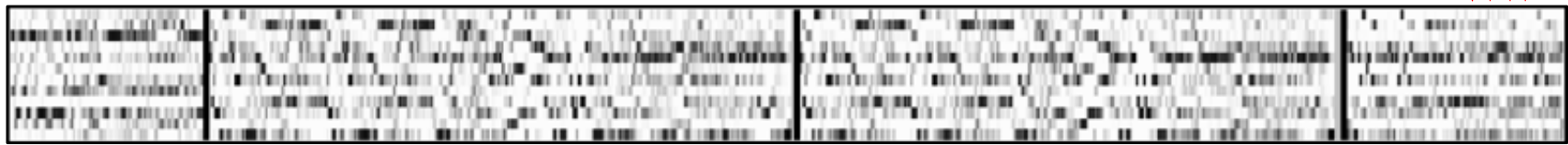
Time (seconds)

Matching Procedure

Query



DB

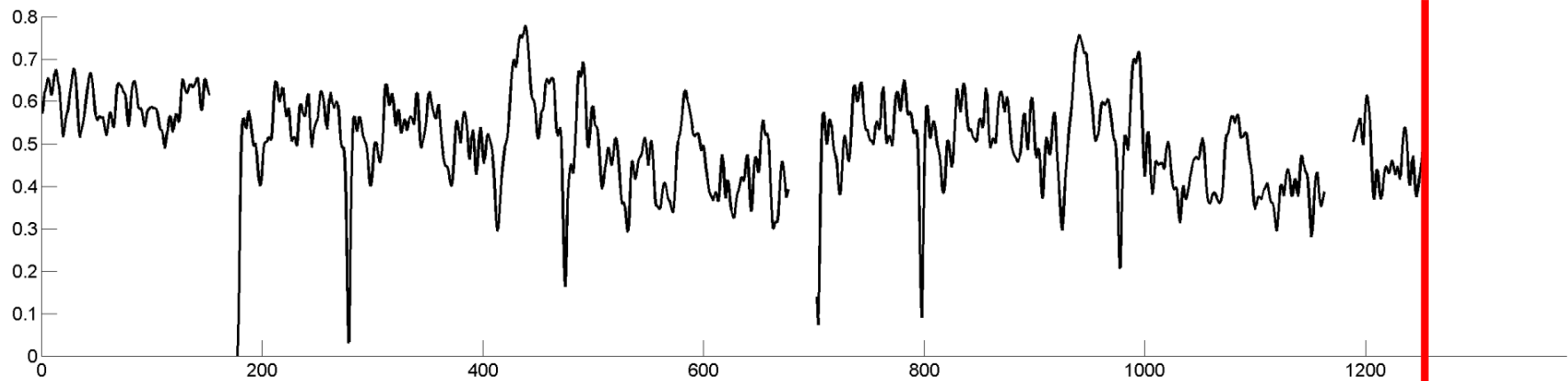


Bach

Beethoven/Bernstein

Beethoven/Sawallisch

Shostakovich

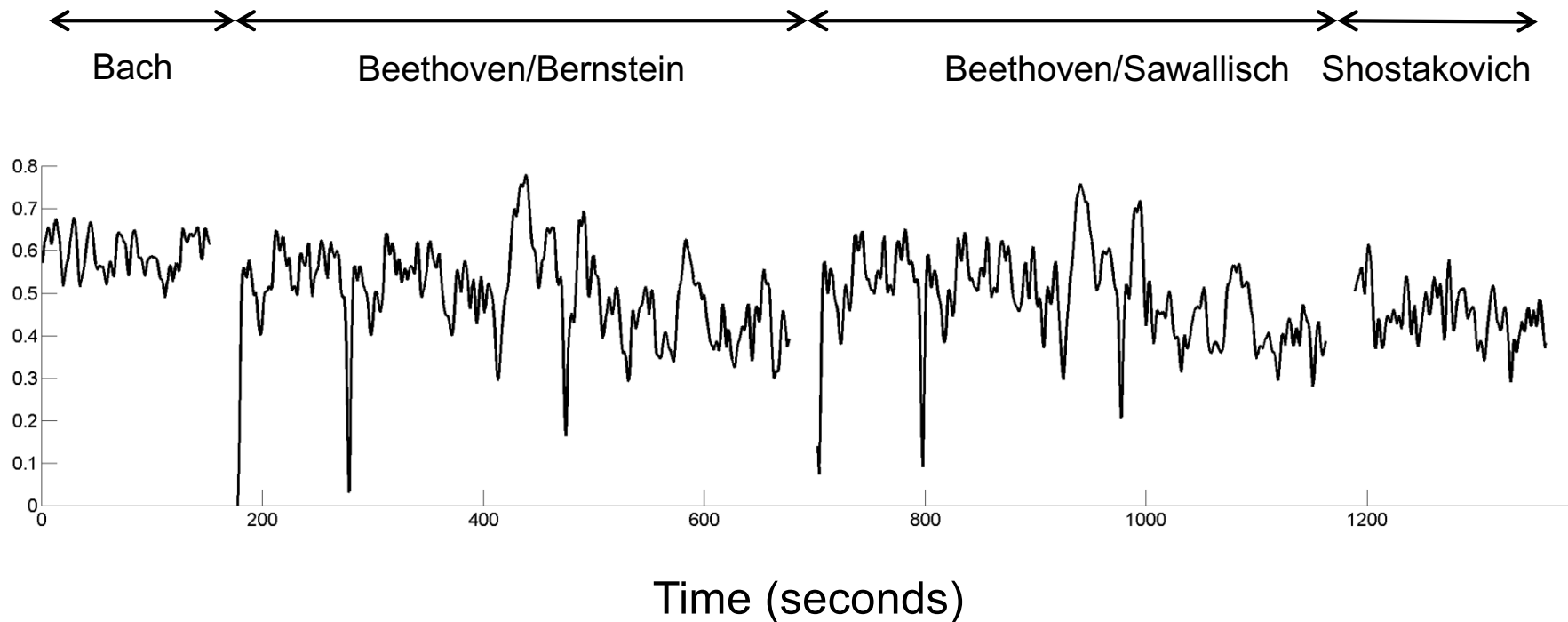


Time (seconds)

Matching Procedure

Matching curve

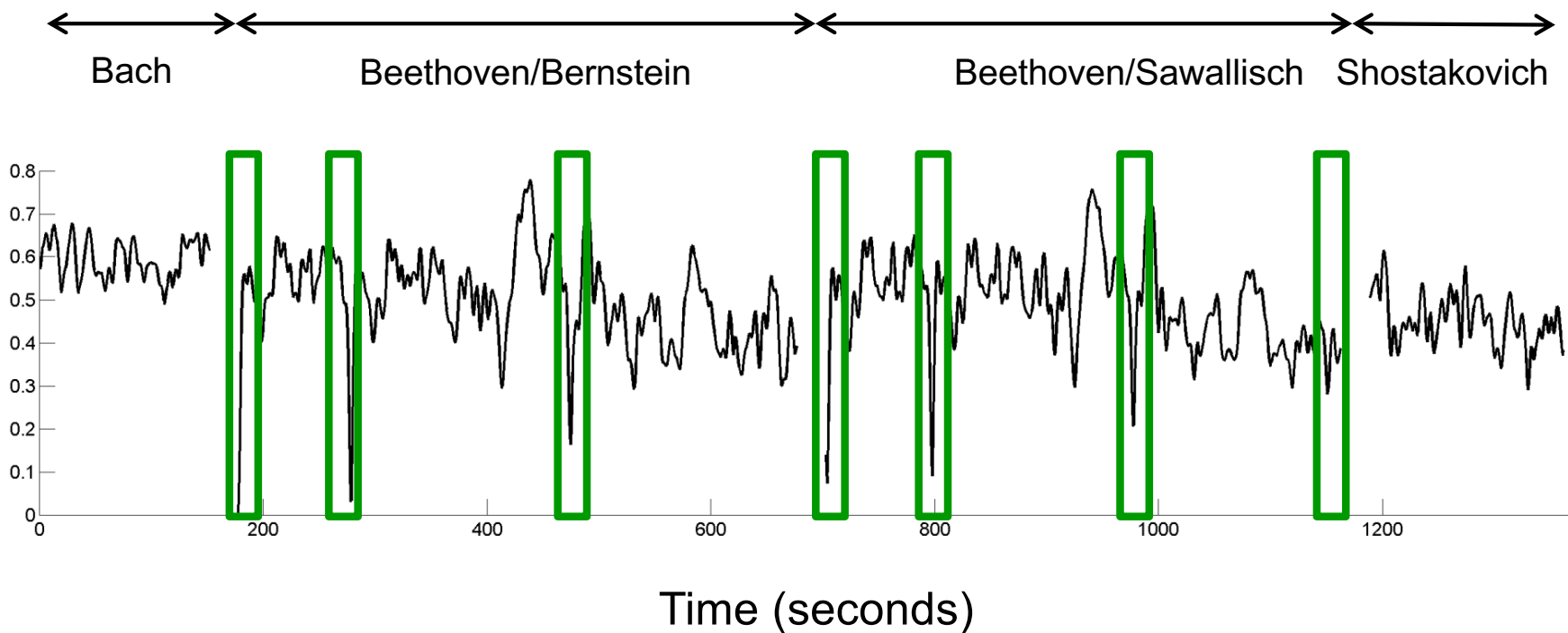
Query: Beethoven's Fifth / Bernstein (first 20 seconds)



Matching Procedure

Matching curve

Query: Beethoven's Fifth / Bernstein (first 20 seconds)

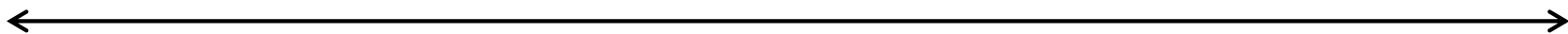
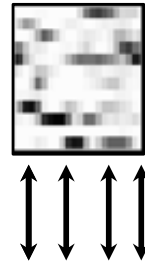


Matches 1 2 5 3 4 6 7

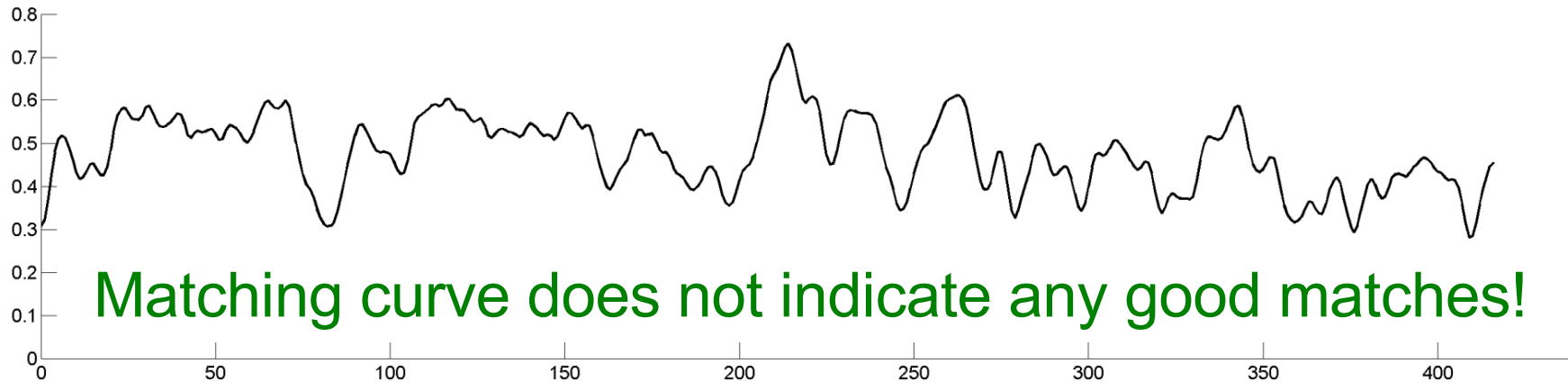
Matching Procedure

Problem: How to deal with tempo differences?

Karajan is much faster than Bernstein!



Beethoven/Karajan



Time (seconds)

Matching Procedure

1. Strategy: Usage of local warping

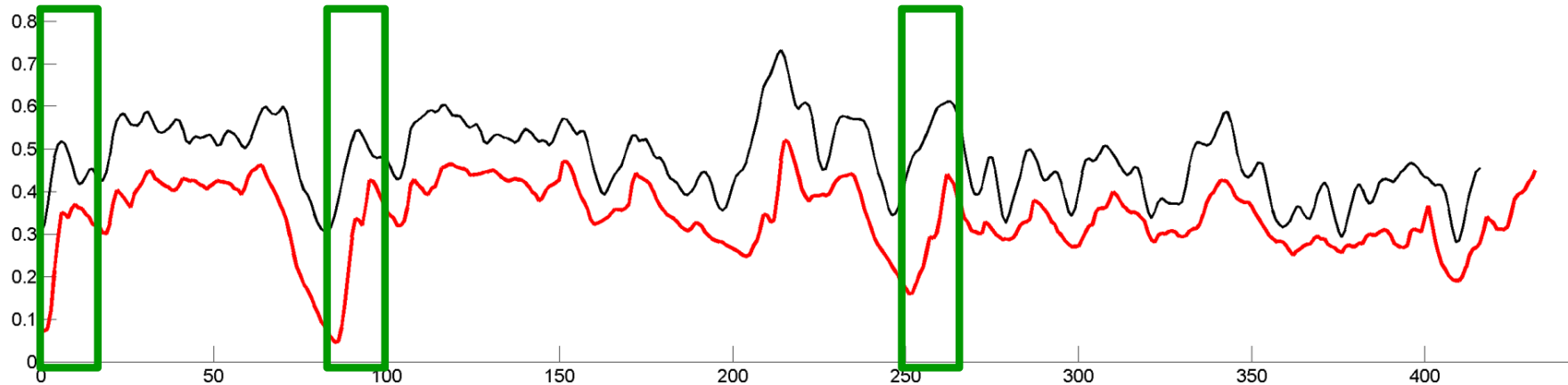
Karajan is much faster than Bernstein!



Warping strategies are computationally expensive and hard for indexing.



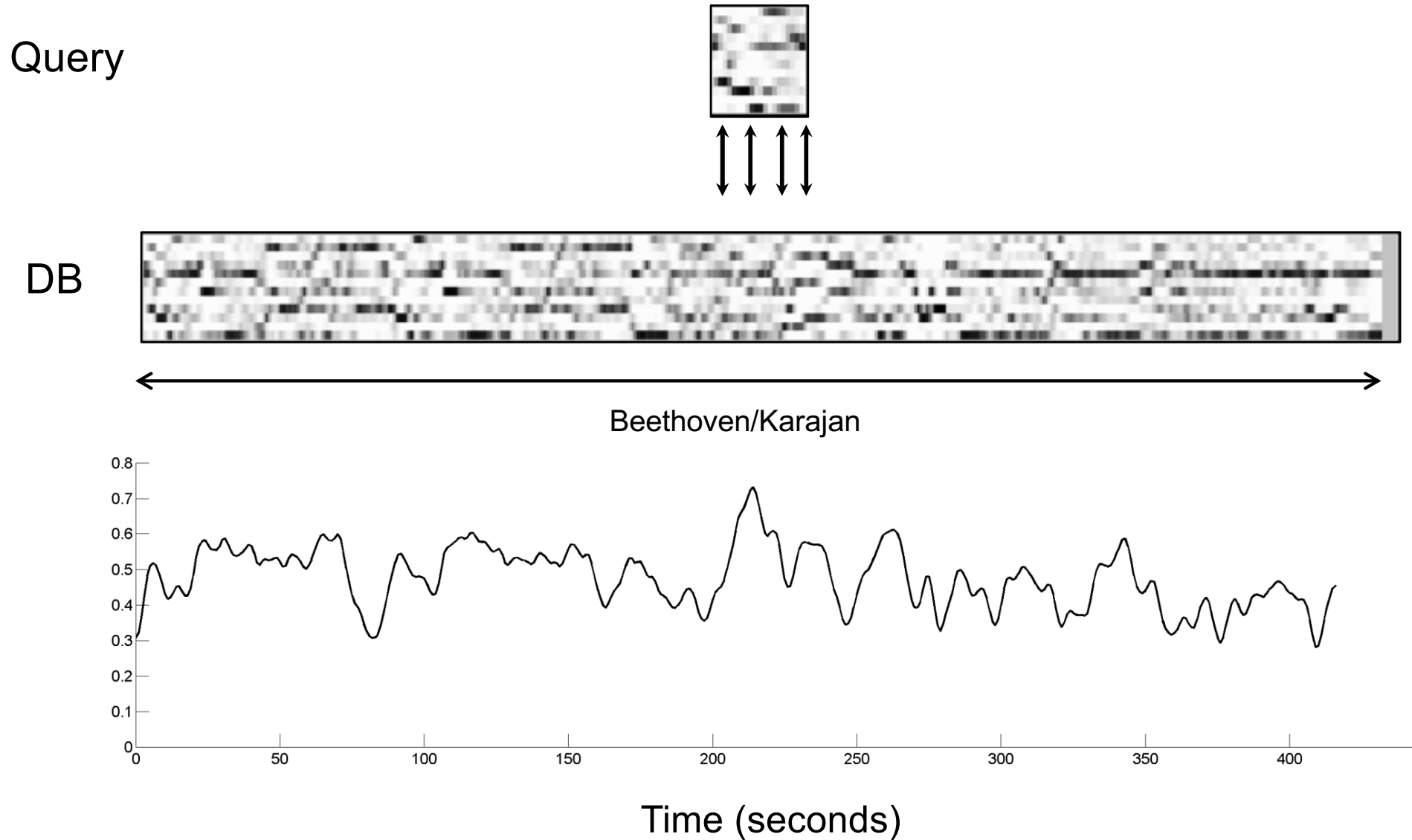
Beethoven/Karajan



Time (seconds)

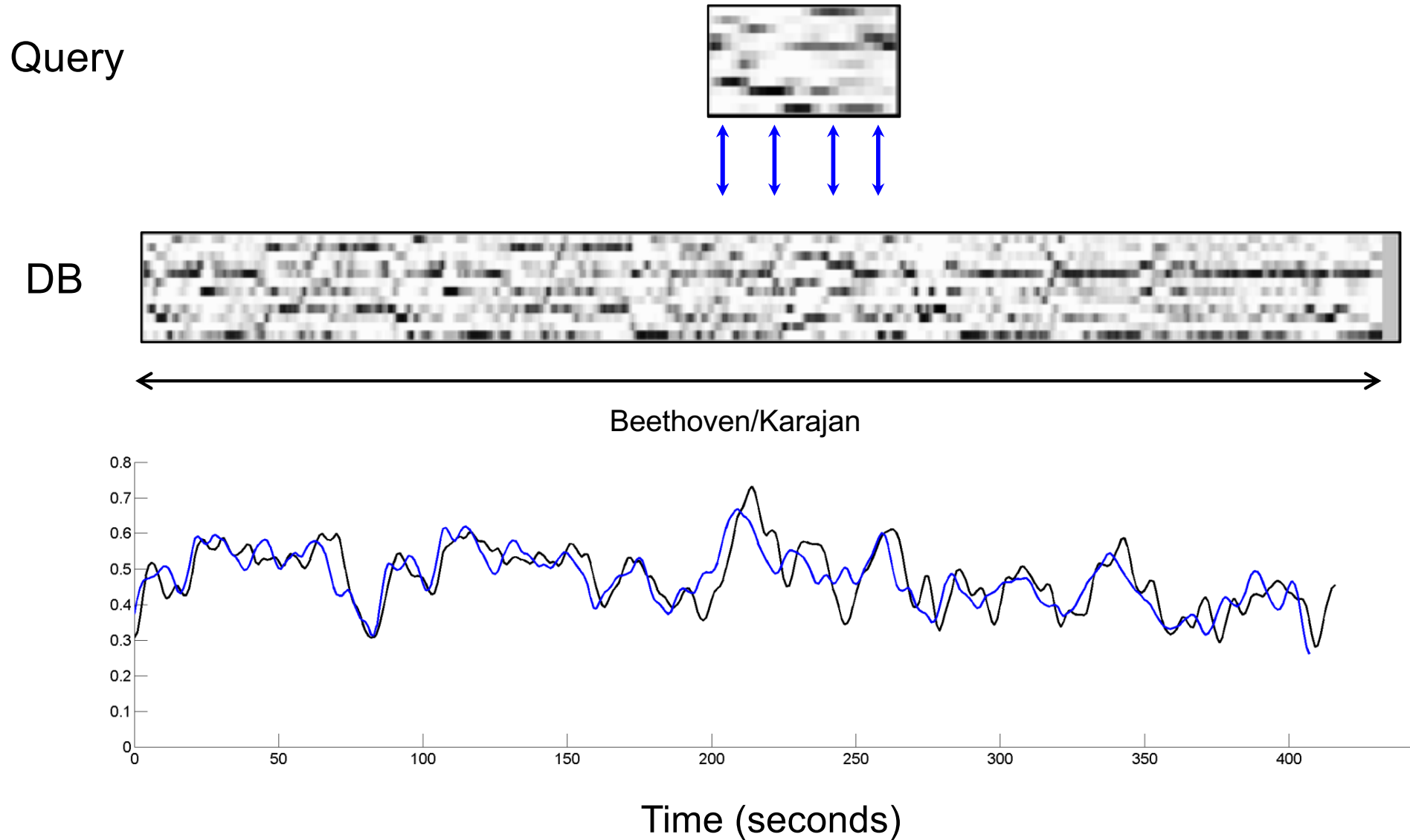
Matching Procedure

2. Strategy: Usage of multiple scaling



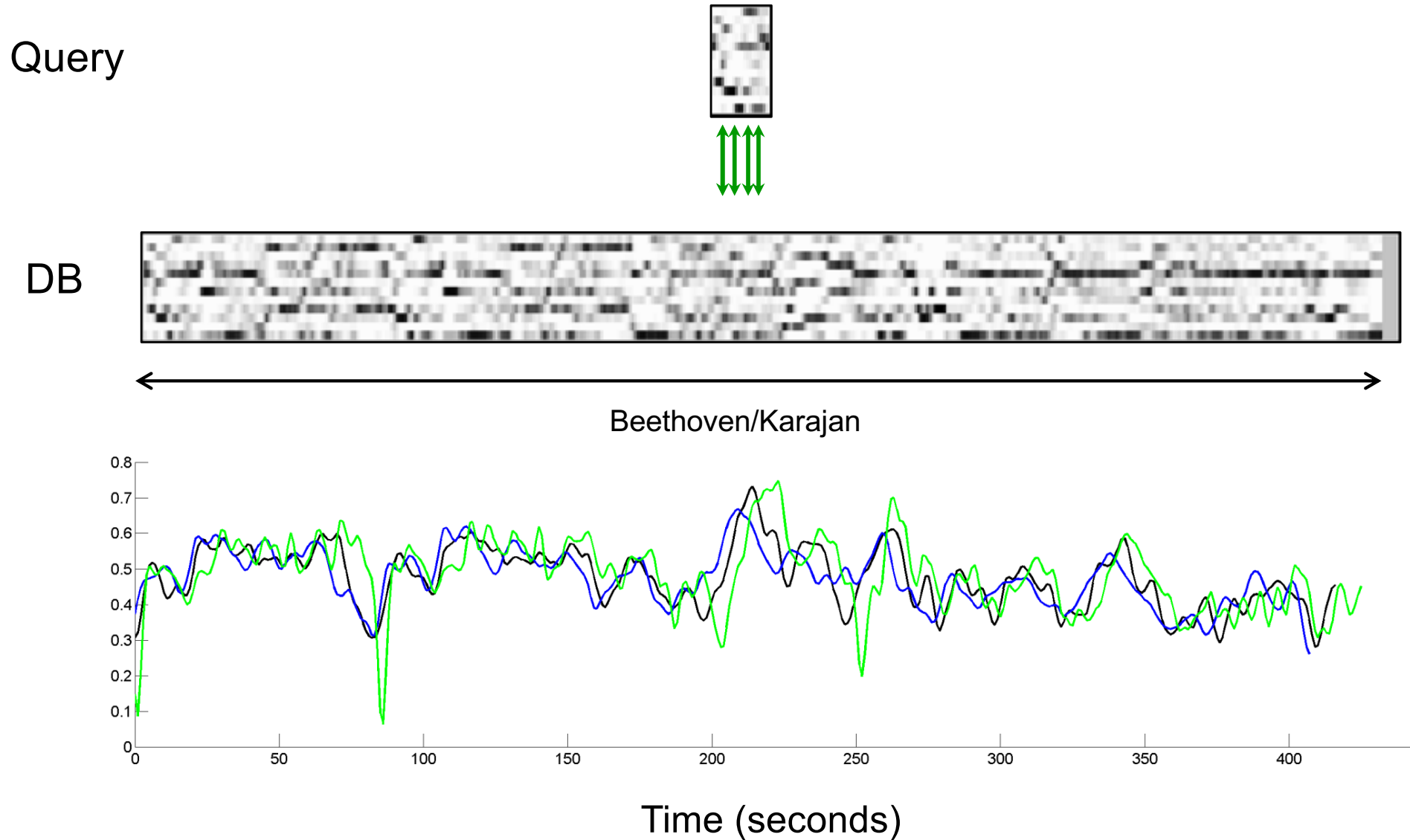
Matching Procedure

2. Strategy: Usage of multiple scaling



Matching Procedure

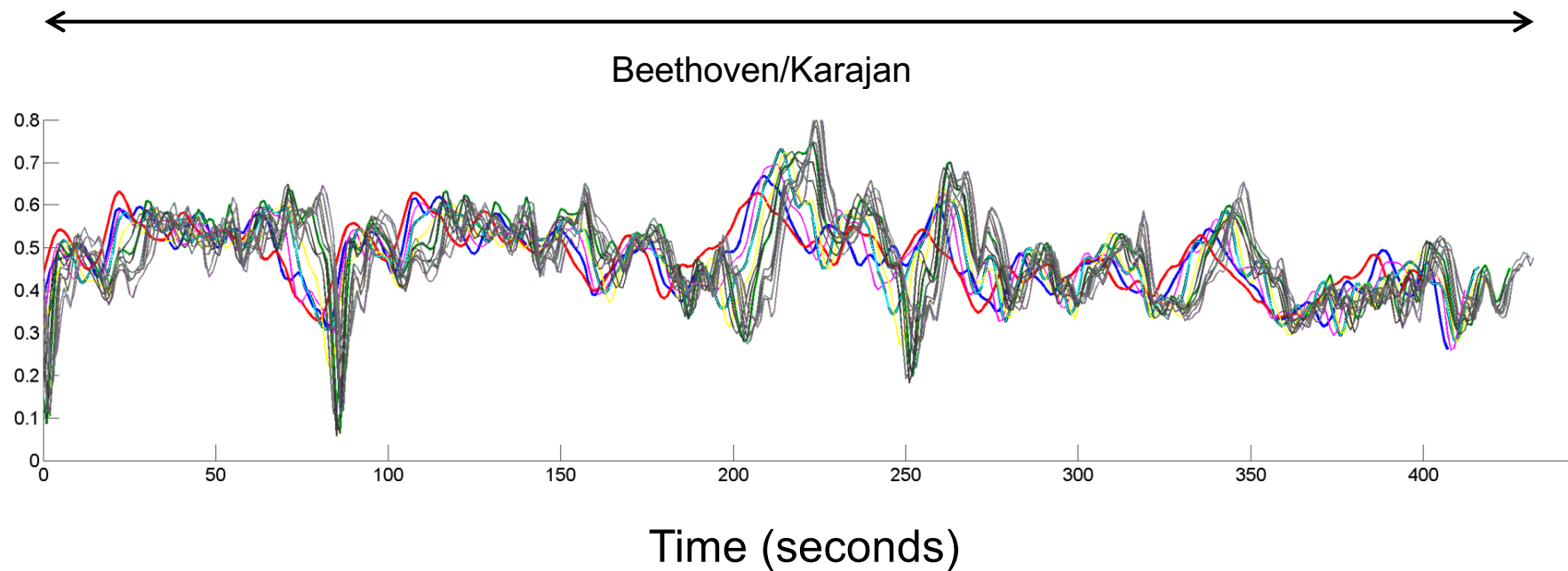
2. Strategy: Usage of multiple scaling



Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

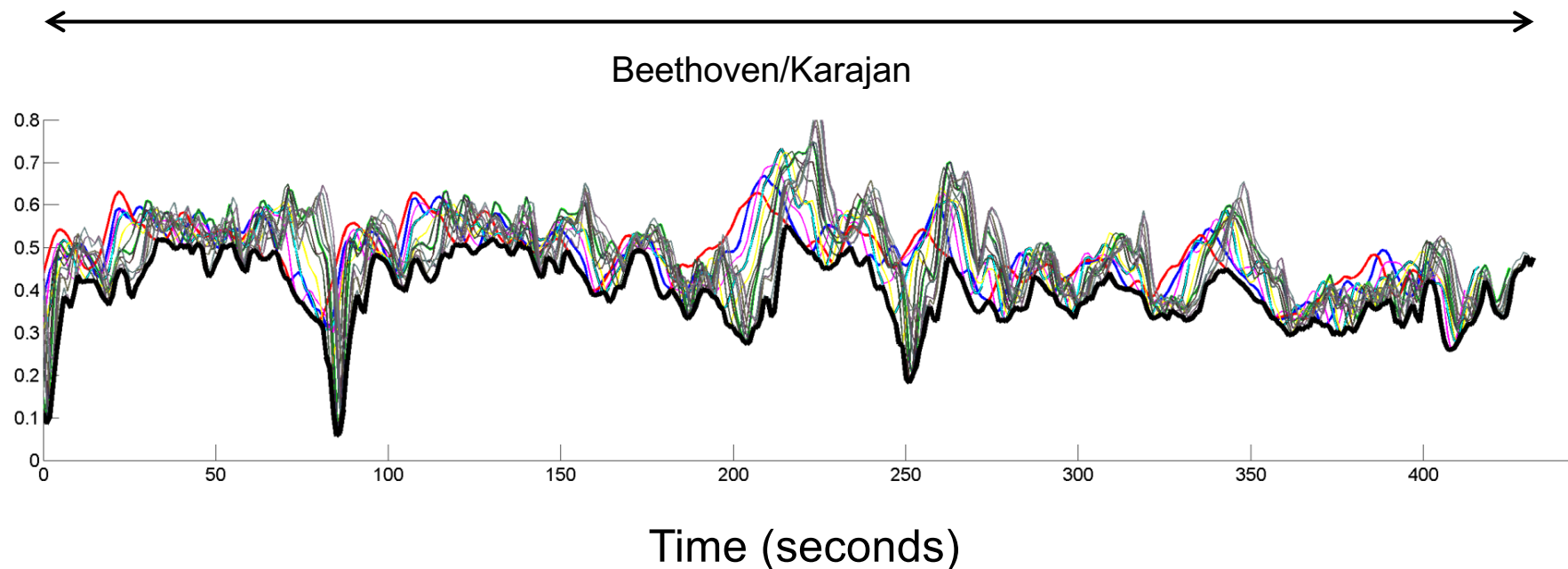


Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

Minimize over all curves



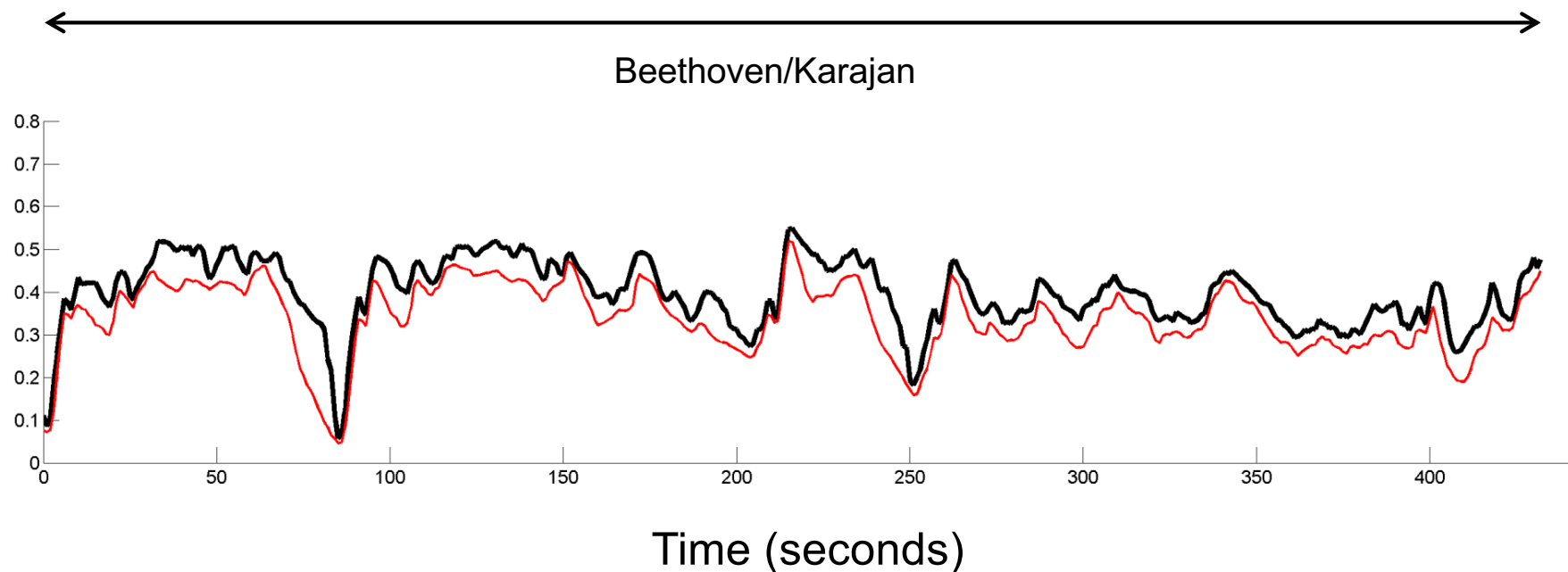
Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

Minimize over all curves

Resulting curve is similar to **warping curve**



Audio Matching

Query: Beethoven's Fifth / Bernstein (first 20 seconds)

Rank	Piece	Position	
1	Beethoven's Fifth/Bernstein	0 - 21	▶
2	Beethoven's Fifth/Bernstein	101- 122	▶
3	Beethoven's Fifth/Karajan	86 - 103	▶
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
10	Beethoven's Fifth/Karajan	252 - 271	▶
11	Beethoven's Fifth/Scherbakov	0 - 19	▶
12	Beethoven's Fifth/Sawallisch	275 - 296	▶
13	Beethoven's Fifth/Scherbakov	86 - 103	▶
14	Schumann Op. 97,1/Levine	28 - 43	▶



Audio Matching: Conclusions

Strategy: Handle variations at various levels

- Chroma → invariance to timbre
- Normalization → invariance to dynamics
- Smoothing → invariance to local time deviations
- Multiple queries → invariance to global tempo

Audio Matching: Conclusions

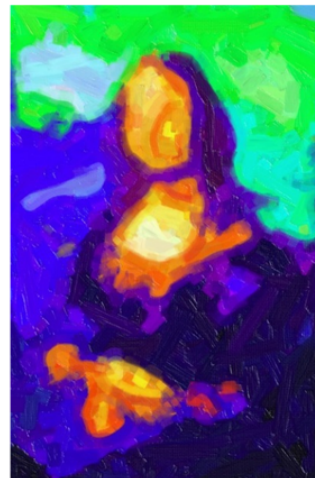
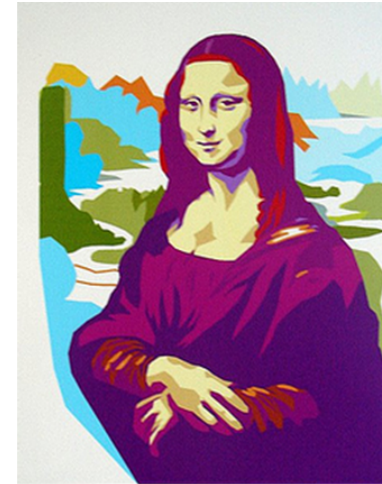
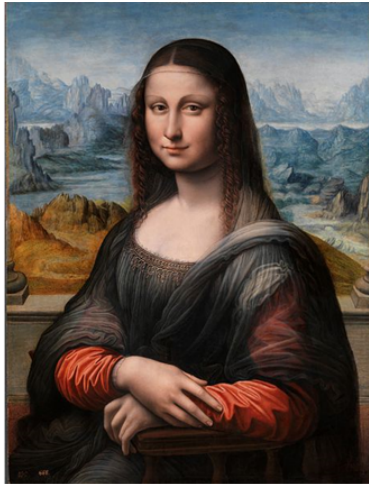
Strategy: Handle variations at various levels

- Chroma → invariance to timbre
- Normalization → invariance to dynamics
- Smoothing → invariance to local time deviations
- Multiple queries → invariance to global tempo

Notes:

- There is no “standard” chroma feature.
→ Variants can make a huge difference!
- Learn invariance from examples
→ “Deep Chroma” [Korzeniowski, Widmer; ISMIR 2016]
- Temporal warping makes problem hard
- Efficiency



















Version (Cover Song) Identification



Version (Cover Song) Identification

Nearly anything can change! But something doesn't change.

Often this is **chord progression** and/or **melody**

 Bob Dylan Knockin' on Heaven's Door 	key 	Avril Lavigne Knockin' on Heaven's Door 
 Metallica Enter Sandman 	timbre 	Apocalyptica Enter Sandman 
 Nirvana Poly [Incesticide Album] 	tempo 	Nirvana Poly [Unplugged] 
 Black Sabbath Paranoid 	lyrics 	Cindy & Bert Der Hund Der Baskerville 
 AC/DC High Voltage 	recording conditions 	AC/DC High Voltage [live] 
	song structure	

Version (Cover Song) Identification

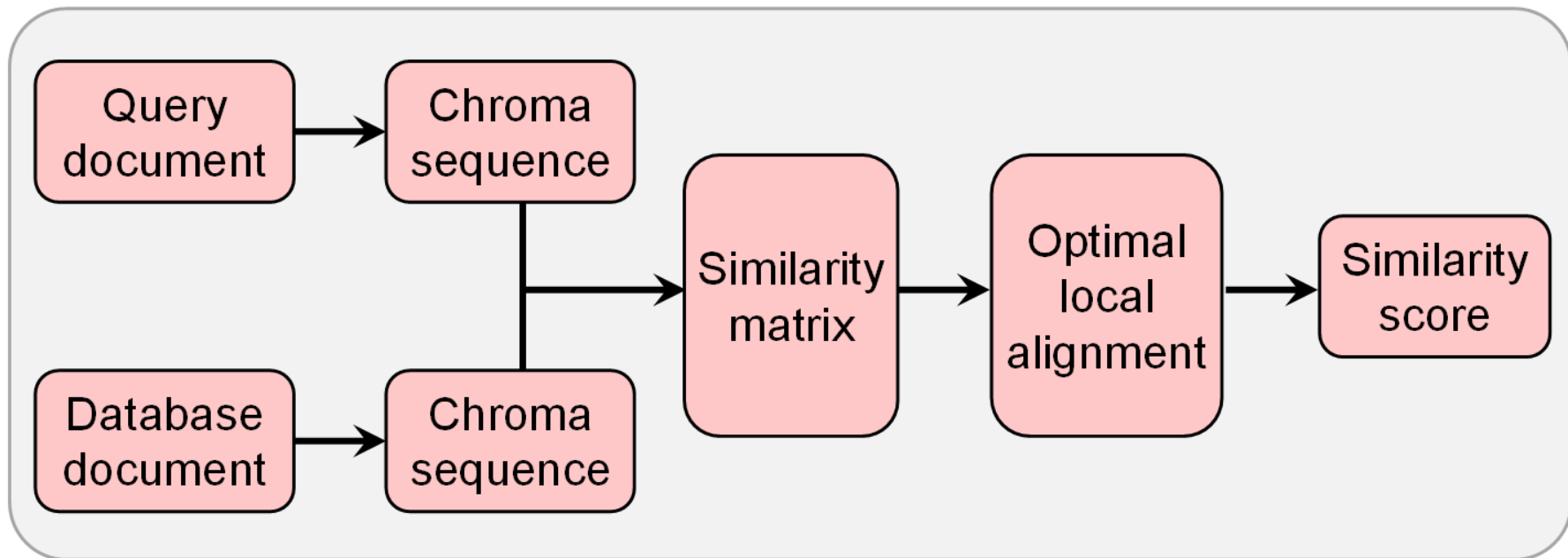
Task

Given a music recording of a song or piece of music as query, find all “similar” music recordings (versions) such as:

- Live versions
- Different interpretations
- Cover songs
- 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

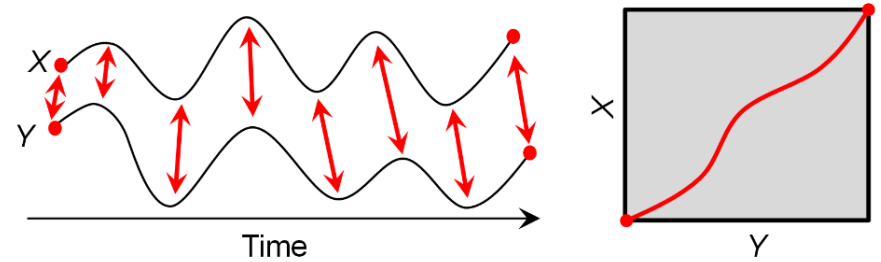
Version (Cover Song) Identification



Alignment Strategies

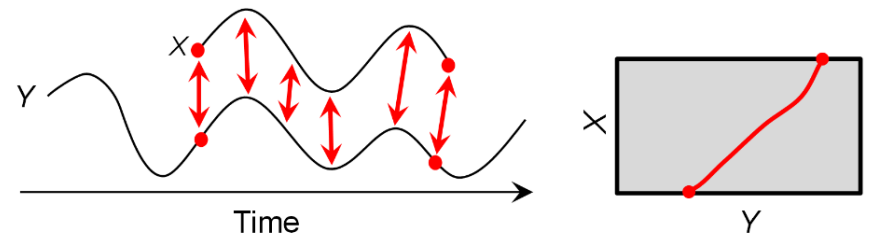
Classical DTW

Music synchronization



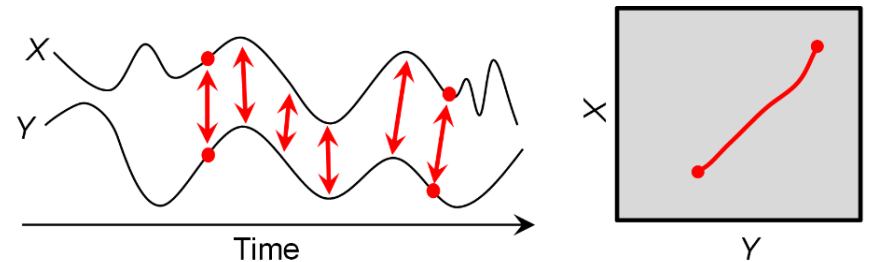
Subsequence DTW

Audio matching

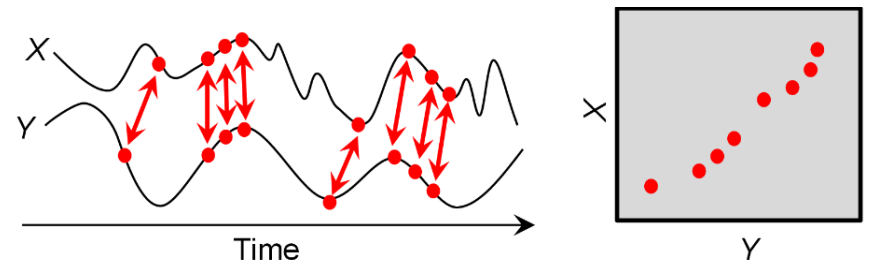


Local alignment

Version (cover song)
identification



Partial alignment



Shingle-Based Retrieval

Idea

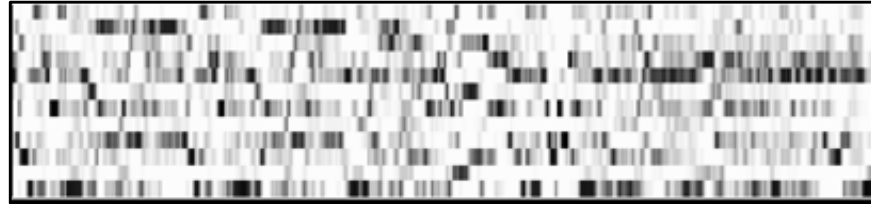
- Query and database are split up into small overlapping shingles that consist of short chroma feature subsequences.
- Shingles can be matched using efficient nearest neighbor retrieval.
- Trade-off:
 - Large shingles have high musical relevance
 - High shingle dimensionality makes indexing difficult

[Casey, Rhodes, Slaney; IEEE TASLP, 2008]

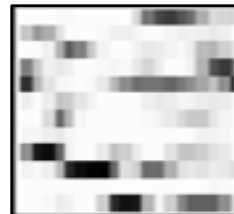
[Grosche, Müller; ICASSP 2012]

Shingle-Based Retrieval

Database
Chroma sequence



Query
Chroma sequence
(ca. 10 to 30 seconds)



Shingle-Based Retrieval

Database
Chroma sequence

Chroma shingles

Retrieval
(index-based)

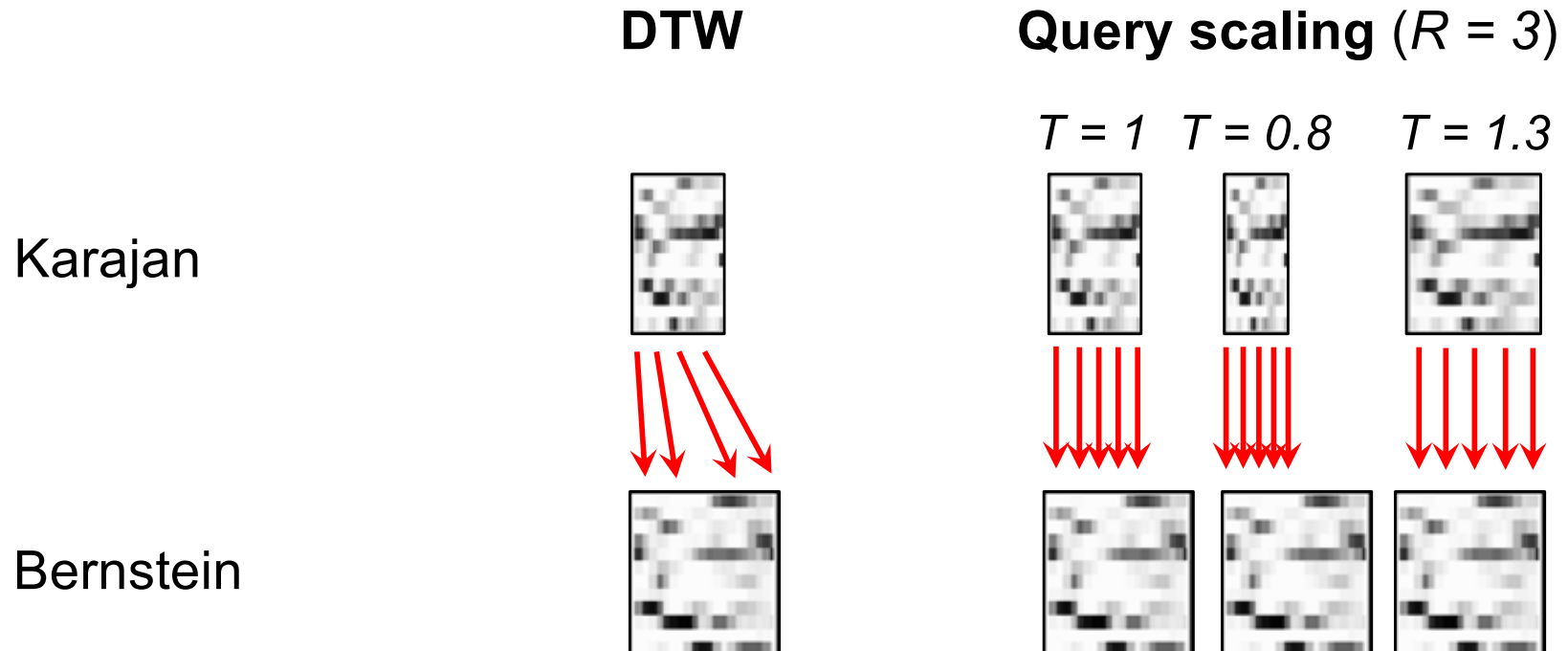
Query
Chroma sequence
(ca. 10 to 30 seconds)



Shingle-Based Retrieval

Tempo-invariant matching

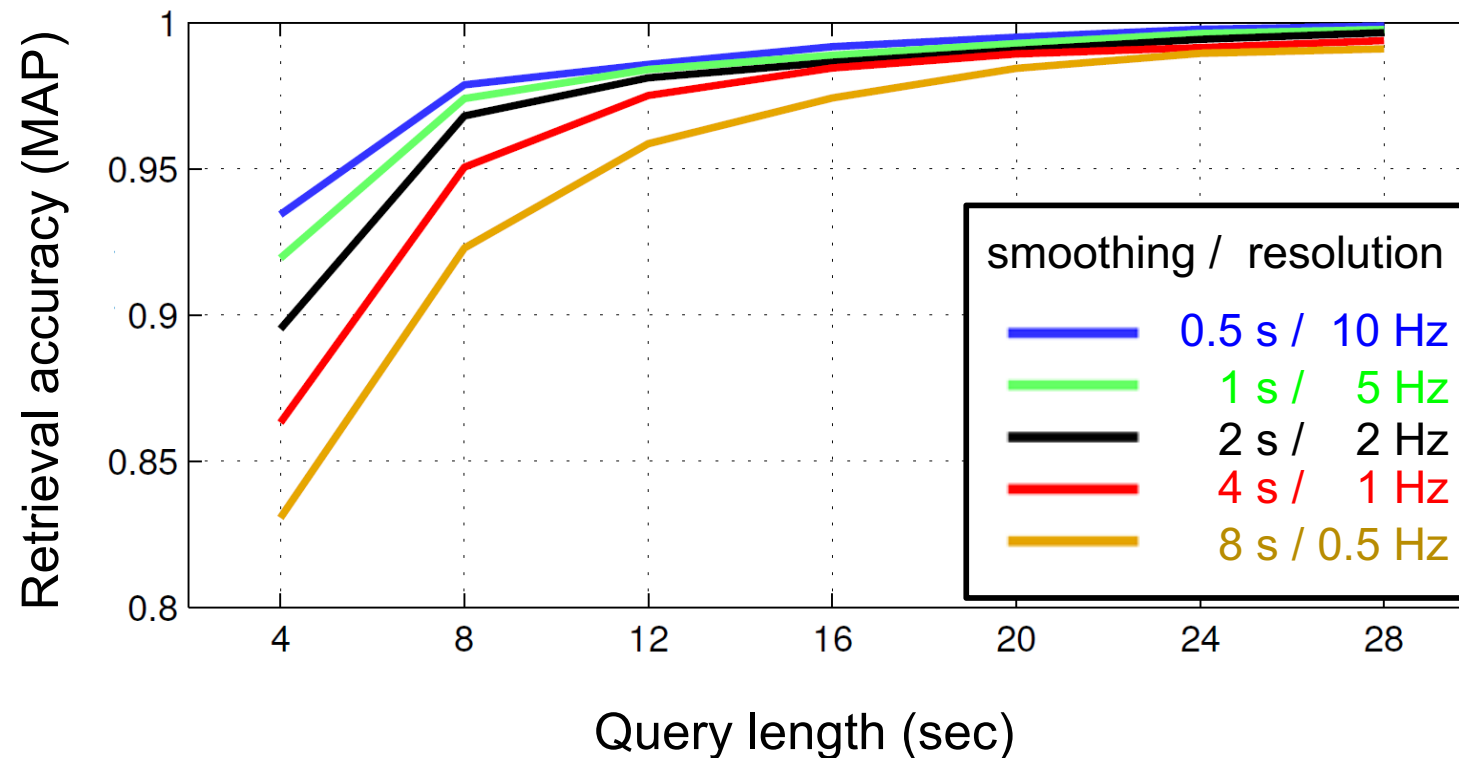
Avoiding expensive temporal warping, tempo differences are handled by creating R scaled variants of the query, each simulating a global change in tempo of up to $\pm 50\%$.



Shingle-Based Retrieval

Query length and feature type

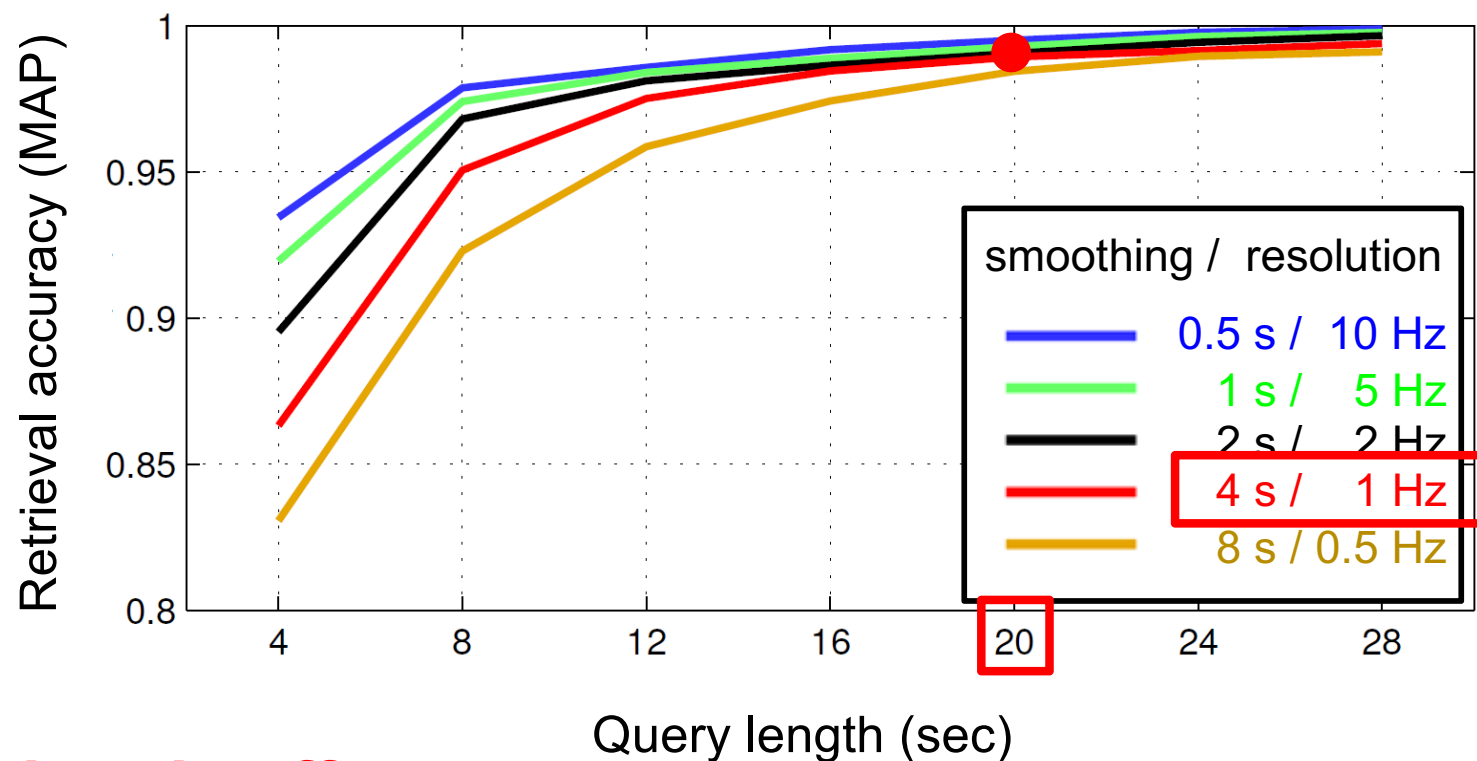
Dataset: 2484 recordings (162 hours)
Per query 10-50 relevant documents



Shingle-Based Retrieval

Query length and feature type

Dataset: 2484 recordings (162 hours)
Per query 10-50 relevant documents



Good trade-off

- Query length = 20 sec
- Feature type: 4 s / 1 Hz

[Grosche, Müller; ICASSP 2012]

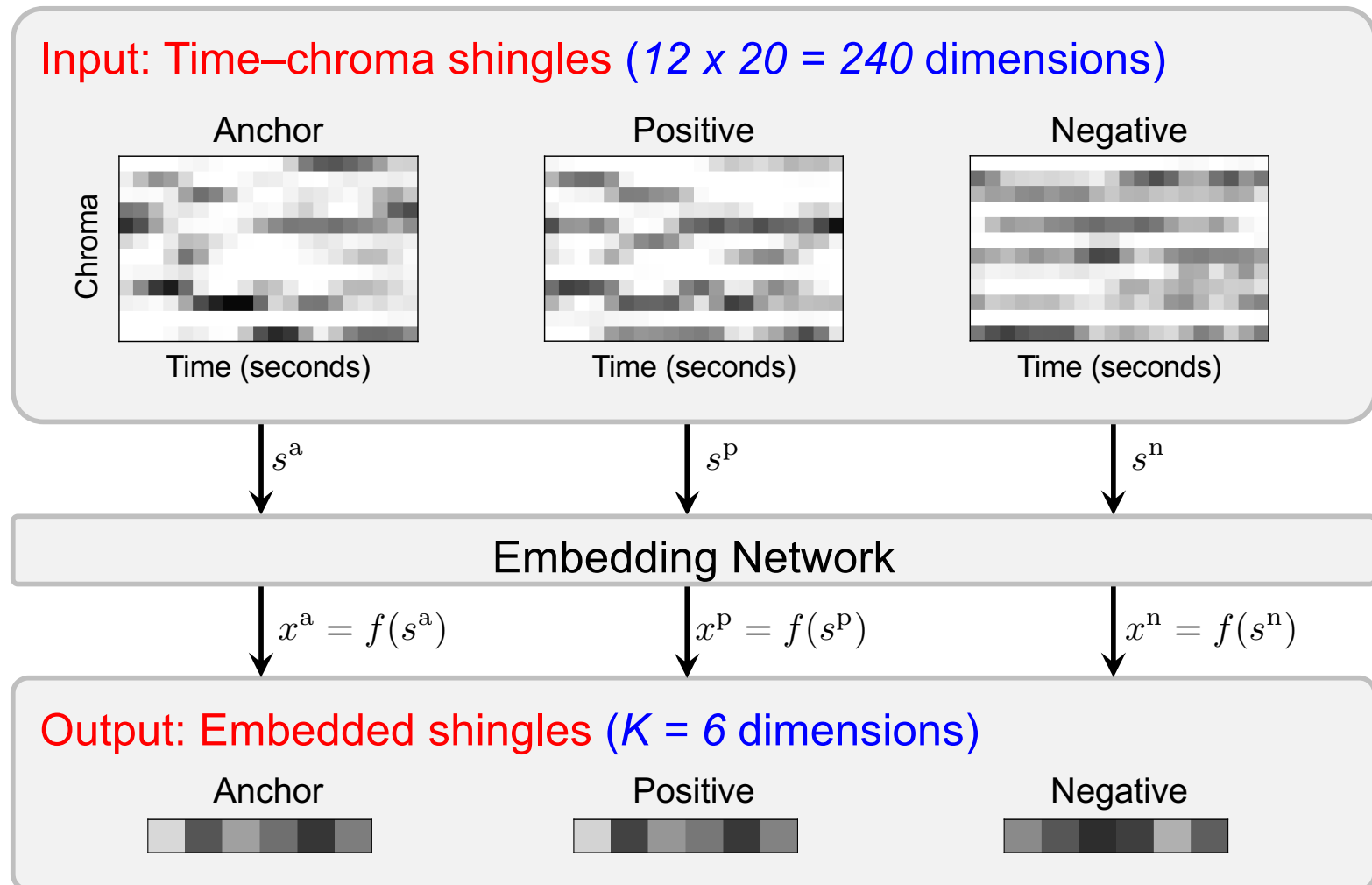
Shingle-Based Retrieval

- Time-chroma shingle: $12 \times 20 = 240$ dimensions
- Indexing via Locality-Sensitive Hashing
 - Speedup factor of 25 with MAP > 0.9
 - Speedup factor of 100 with MAP > 0.8
- Further reduction of shingle dimensionality?
 - Linear embedding using PCA
 - Non-linear embedding using deep learning

Shingle-Based Retrieval

[Schroff et al.; CVPR 2015 (FaceNet)]
[Zalkow, Müller; submitted]

Strategy: Learn embedding using Siamese NN with triplet loss



Shingle-Based Retrieval

[Zalkow, Müller; submitted]

Retrieval accuracy (MAP)

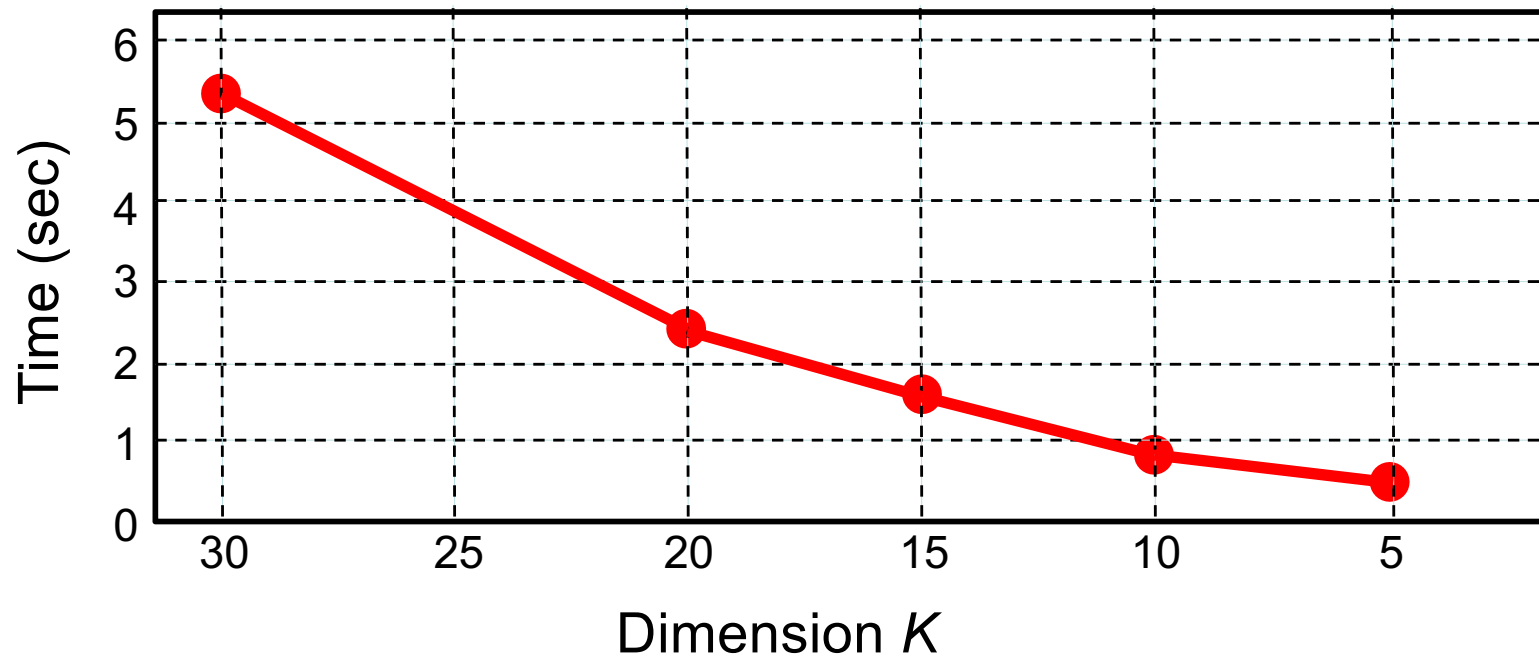
Dimension K	240	30	20	15	10	7	5	3
PCA	0.97	0.95	0.93	0.90	0.87	0.79	0.76	0.58
NN	0.97	0.96	0.94	0.94	0.92	0.87	0.80	0.68

Shingle-Based Retrieval

Retrieval accuracy (MAP)

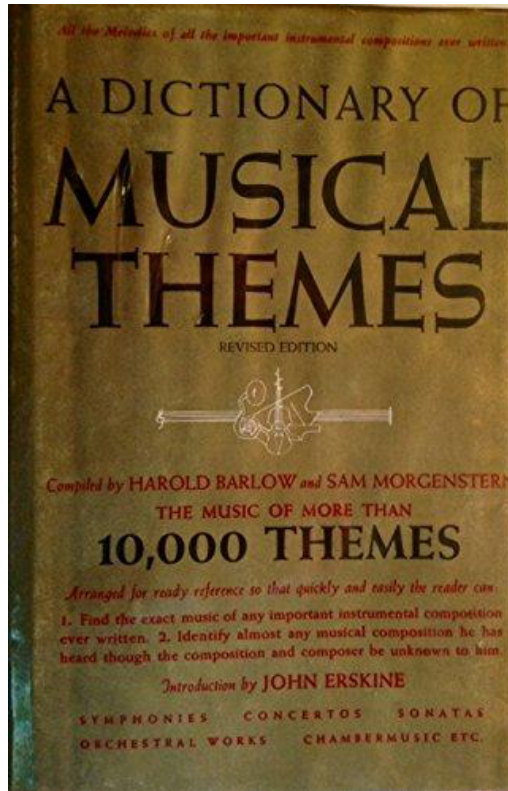
Dimension K	240	30	20	15	10	7	5	3
PCA	0.97	0.95	0.93	0.90	0.87	0.79	0.76	0.58
NN	0.97	0.96	0.94	0.94	0.92	0.87	0.80	0.68

Running time (k-d tree implementation, 3300 queries)



Cross-Modal Music Retrieval

Barlow & Morgenstern (1949): A Dictionary of Musical Themes

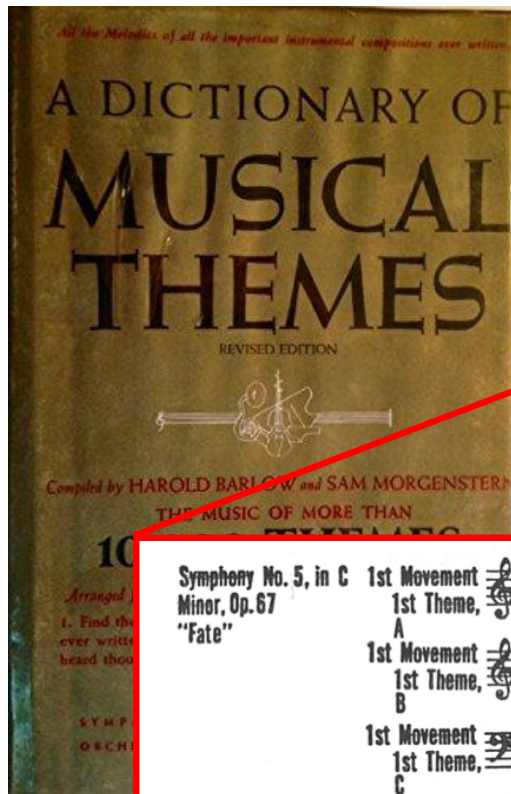


The image shows a page from the book, specifically page 71, which is dedicated to Beethoven's works (B943-B962). The page is filled with musical notation, including staves with notes, rests, and other musical symbols. The notation is arranged in a list format, with each entry consisting of a movement and a theme number, followed by a small musical snippet and a page number. The entries include:

- 3rd Movement 1st Theme 0943
- 3rd Movement 2nd Theme 0944
- 4th Movement 1st Theme 0945
- 4th Movement 2nd Theme 0946
- 4th Movement 2nd Theme 0947
- Symphony No. 5, in C Minor, Op. 67 "Fate" 1st Theme 0948
- 1st Movement 1st Theme, A 0949
- 1st Movement 1st Theme, B 0950
- 1st Movement 1st Theme, C 0951
- 1st Movement 2nd Theme 0952
- 1st Movement 3rd Theme 0953
- 1st Movement 4th Theme 0954
- 2nd Movement 1st Theme 0955
- 2nd Movement 2nd Theme 0956
- 2nd Movement Coda 0957
- 3rd Movement 1st Theme 0958
- 3rd Movement 2nd Theme 0959
- 3rd Movement 3rd Theme 0960
- 4th Movement 1st Theme 0961
- 4th Movement 2nd Theme 0962
- 4th Movement 3rd Theme 0962

Cross-Modal Music Retrieval

Barlow & Morgenstern (1949): A Dictionary of Musical Themes



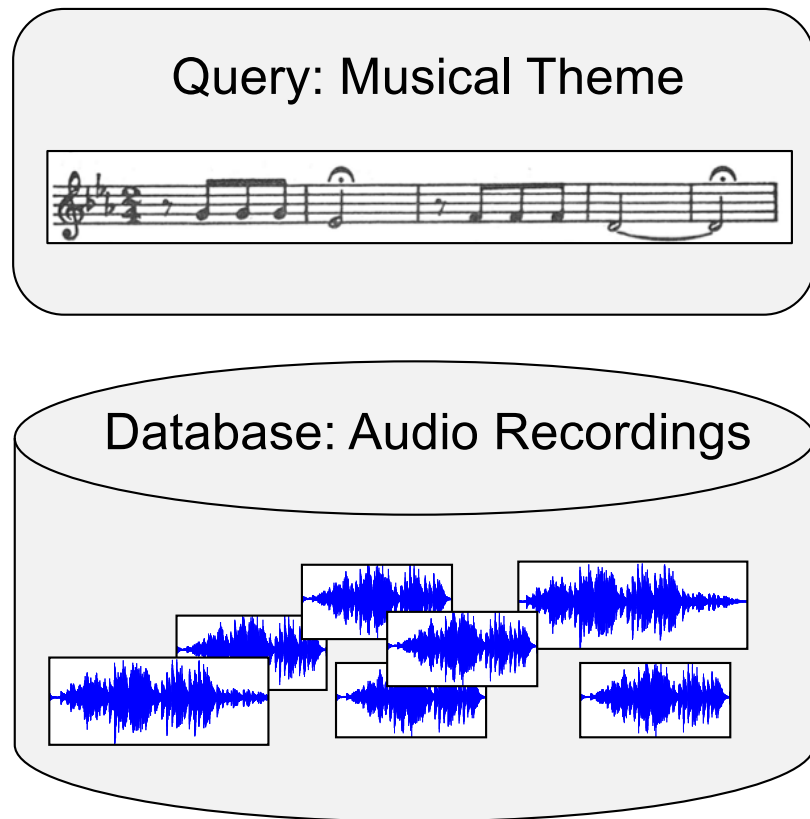
Symphony No. 5, in C Minor, Op. 67 "Fate"

1st Movement	1st Theme	B948
1st Movement	1st Theme, A	B949
1st Movement	1st Theme, B	B950
1st Movement	1st Theme, C	B951
1st Movement	2nd Theme	B951

10,000 Themes from Western classical music

Cross-Modal Music Retrieval

Barlow & Morgenstern (1949): A Dictionary of Musical Themes



Challenges

- **Cross-modality**
Symbolic vs. audio data
- **Tuning**
Deviations from standard tuning
- **Transposition**
Played key vs. written key
- **Tempo**
Local & global tempo deviations
- **Polyphony**
Monophonic query vs. polyphonic audio

Cross-Modal Music Retrieval

Retrieval Experiment

#Queries: 2045 themes

#Database: 1114 recordings (120 hours)

Balke et al.; ICASSP 2016]

	Top-1	Top-20	Top-50
Tuning	18.3	29.2	46.1
Transposition & query length	39.5	66.9	76.1

Cross-Modal Music Retrieval

Retrieval Experiment

#Queries: 2045 themes

#Database: 1114 recordings (120 hours)

[Balke et al.; ICASSP 2016]

	Top-1	Top-20	Top-50
Tuning	18.3	29.2	46.1
Transposition & query length	39.5	66.9	76.1

[Zalkow, Balke, Müller; ICASSP 2019]

Feature Type	Top-1	Top-20	Top-50
Chroma (filter bank, IIS)	47.0	70.0	79.2
Chroma (melody extraction, MEL)	23.1	50.0	59.9
Chroma (saliency, BG1)	75.4	88.5	91.3
Chroma (deep learning, CNN)	69.3	85.3	89.6

<https://www.audiolabs-erlangen.de/resources/MIR/2019-ICASSP-BarlowMorgenstern>

Cross-Modal Music Retrieval



WIKIPEDIA
The Free Encyclopedia

Article Talk

Symphony No. 5 (Beethoven)

From Wikipedia, the free encyclopedia

"Beethoven's Fifth" redirects here. For the movie, see [Beethov](#)

The **Symphony No. 5** in C minor of Ludwig van Beethoven, Op. 67, was first performed at Vienna's Theater an der Wien in 1808, the work achieved its prodigious popularity. Beethoven's Fifth Symphony is in four movements. It begins with a distinctive four-note "short-short-short-long" motif:



YouTube

5:12 / 36:20



Violino I

Violino II



IMSLP
Petrucci Music Library



Symphony no. 5 in C minor, op. 67

~ Symphony

Overview Aliases Tags Details Edit

Recordings

Date	Title	Attributes	Artist
	performance		
1939	Symphony No. 5 in C minor, Op. 67: I. Allegro con brio II. Andante con moto III. Scherzo. Allegretto IV. Allegro		



MusicBrainz