Goldsmiths UNIVERSITY OF LONDON

Musical Structure Detection

Michael Casey, Christophe Rhodes (Goldsmiths)
Samer Abdallah, Mark Sandler (Queen Mary)

Work in progress

EPSRC GR/S84750/01

Goldsmiths UNIVERSITY OF LONDON

Outline

- Contrast old and new media
- The metadata problem, or what is structure?
- MPEG-7 Audio Descriptors
- Hierarchical segmentation
- Example
- Discussion

Goldsmiths UNIVERSITY OF LONDON

Introduction

- Old media collections
  - CDs, cassettes: 1000 hours
  - LPs: 3.5 metres
  - Minidiscs: 2
- Library (or shop) collections
  - CDs: 10000 hours
  - LPs: 35 metres
  - Minidiscs: 2

Goldsmiths UNIVERSITY OF LONDON

Introduction

- MP3 players (e.g. iPod)
  - Hard drive capacity: 20GB
  - 4 minute song, 128kbps
  - 5000 songs
- Online retailers (e.g. iTunes)
  - over 10^6 songs (4TB)

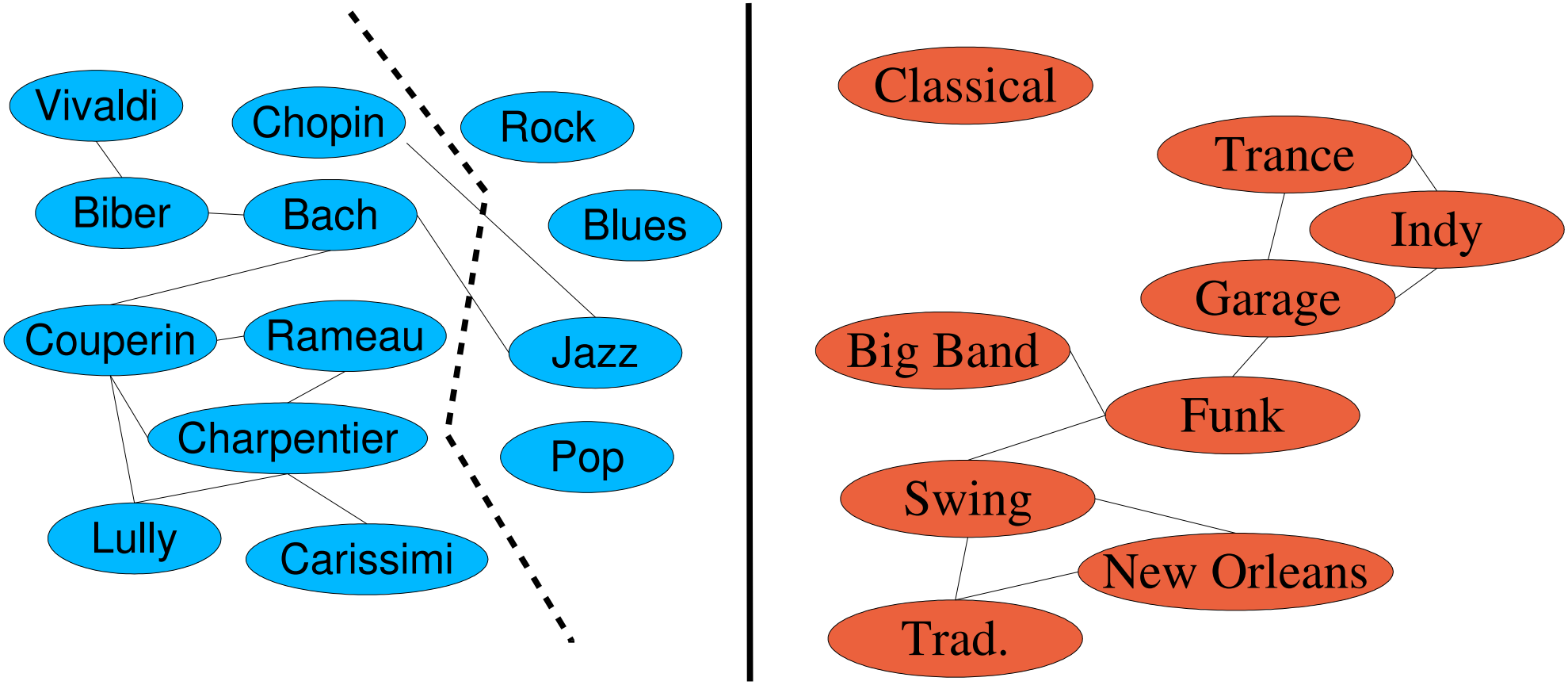
Goldsmiths UNIVERSITY OF LONDON

Metadata

- Dealing with this increase in available media
  - Classification
  - Indexing
  - Librarianship
  - Metadata

Goldsmiths UNIVERSITY OF LONDON

Metadata



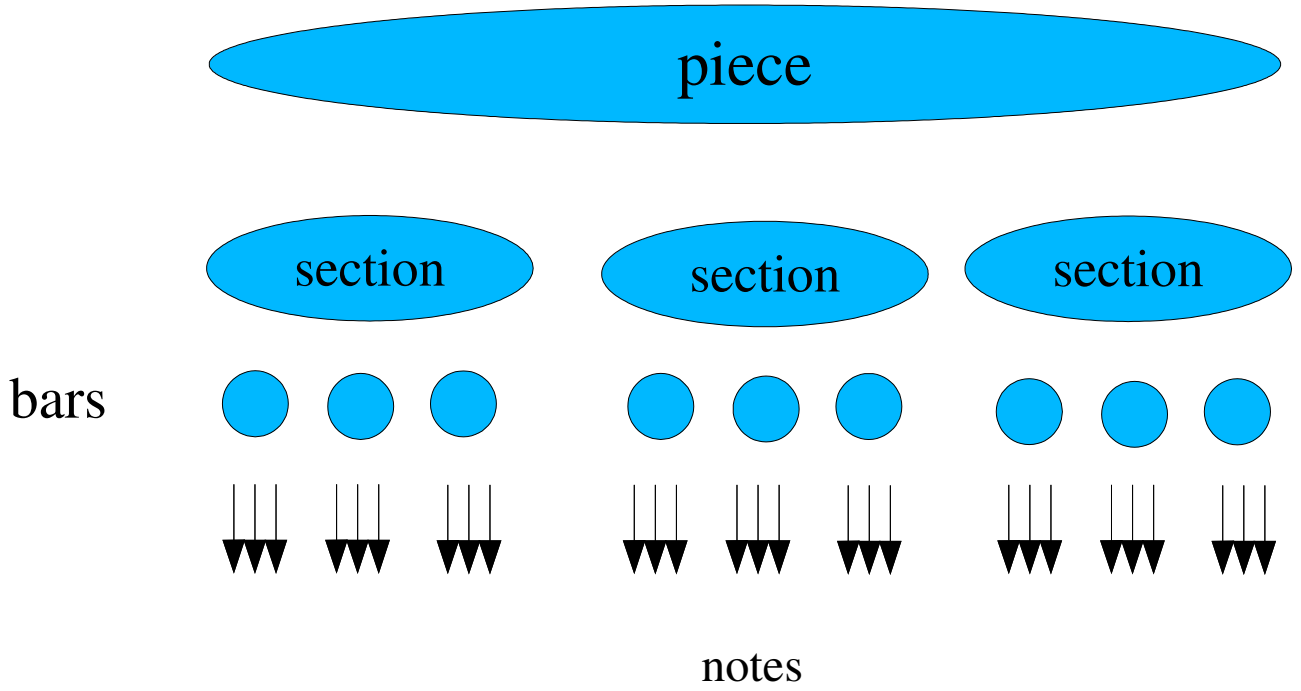
Goldsmiths UNIVERSITY OF LONDON

Project aims

- Automatic structure detection in music
  - working from audio
  - inform indexing, classification
  - commercial applications
- What is structure?
  - cognitive-based research
  - hierarchy

Goldsmiths UNIVERSITY OF LONDON

Structure hierarchy



(not to scale)

Goldsmiths UNIVERSITY OF LONDON

Structure hierarchy

- note: onset detection (DSP work)
- bar: hierarchical agglomeration of notes, given strengths
- phrase, section: perceptual segmentation
  - functional: musicological analysis
  - auditory: acoustic similarity?

Goldsmiths UNIVERSITY OF LONDON

MPEG-7

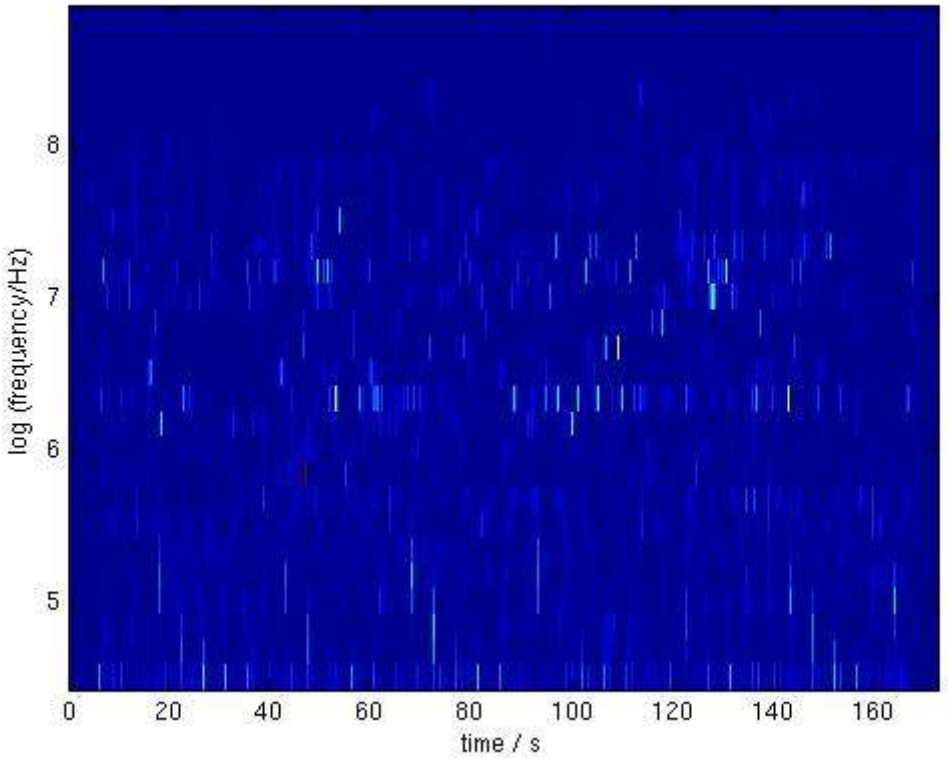
- Standardizing “Content Description”
  - Audio chapter (also image, video, ...)
  - Toolset approach: standardized what people used
  - Variation in semantics of descriptors

```
<complexType name="SeriesOfScalarBinaryType">
  <complexContent>
    <extension base="mpeg7:SeriesOfScalarType">
       ... <!-- ARGH --> </></></></>
```

Goldsmiths UNIVERSITY OF LONDON

MPEG-7

- AudioSpectrumEnvelope
  - Low level descriptor
  - represents power in spectral bands



Goldsmiths UNIVERSITY OF LONDON

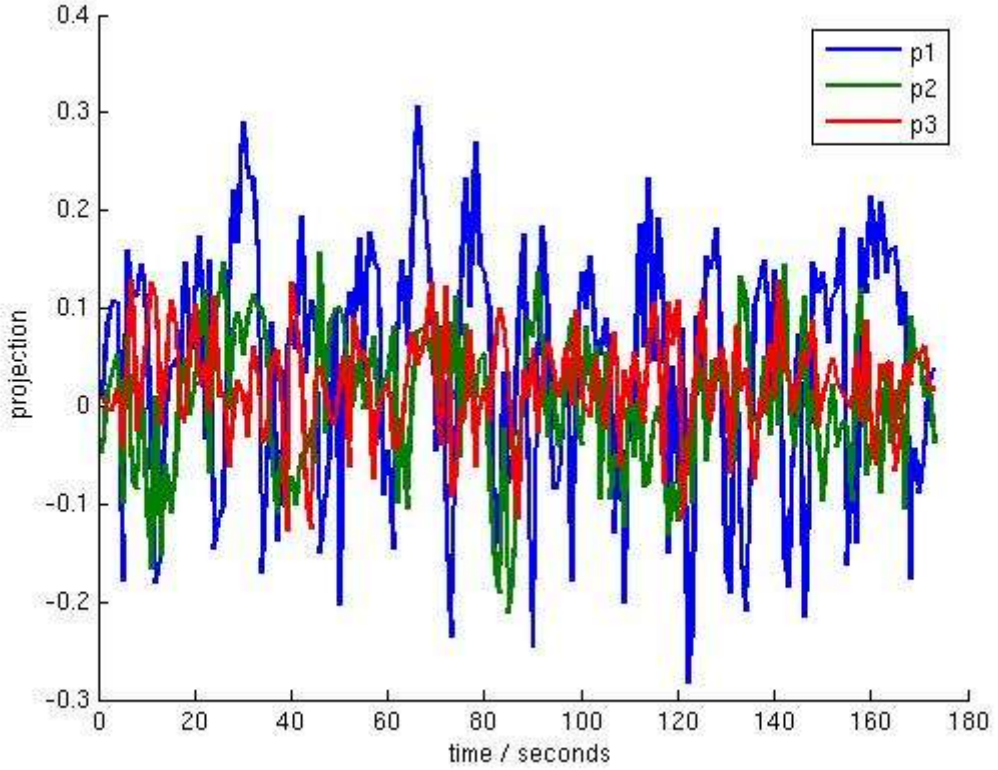
MPEG-7

- AudioSpectrumEnvelope
  - Logarithmic-frequency resampling of short-time Fourier Transform (informative)
  - Mel Frequency Cepstrum Coefficients (almost)
  - Other means of calculation?
    - wavelet decomposition / filter bank

Goldsmiths UNIVERSITY OF LONDON

MPEG-7

- AudioSpectrumProjection
  - Mid-level descriptor
  - Important components of waveform



Goldsmiths UNIVERSITY OF LONDON

MPEG-7

- AudioSpectrumProjection
  - Projection against principal components of logarithmically-scaled logarithmically-resampled frequency spectrum
  - Are principal components the best basis?
    - Independent Component Analysis

Goldsmiths UNIVERSITY OF LONDON

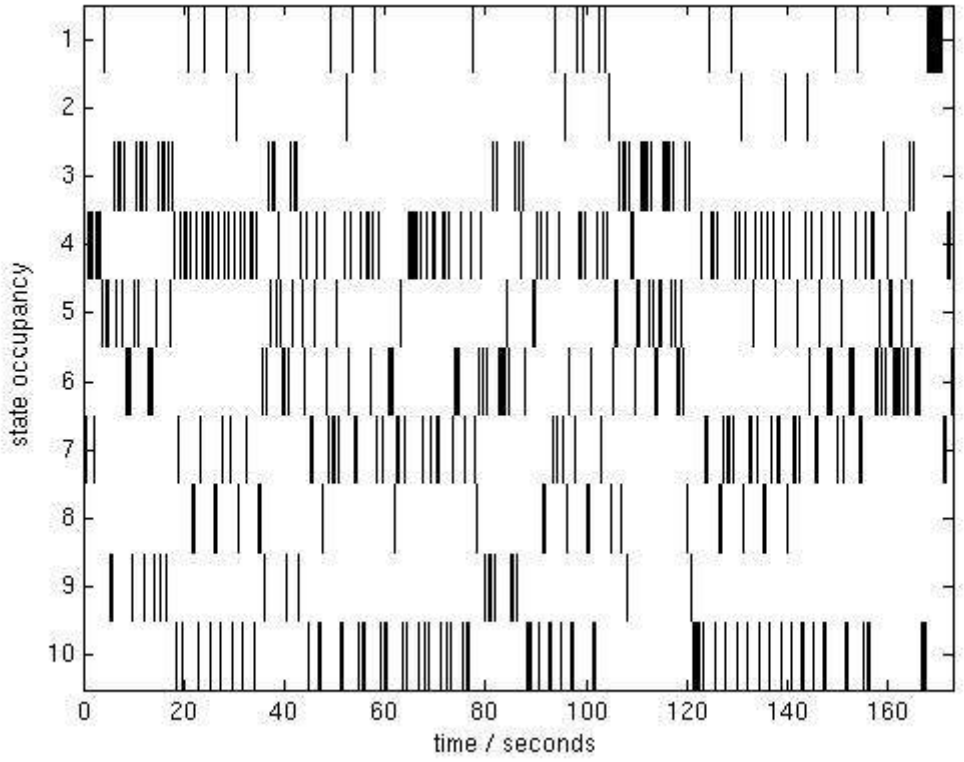
MPEG-7

- Markov Models
  - probability of successive states depend only on previous states
  - toy example: the weather
- Hidden Markov Models
  - latent variable Markov Models
  - we no longer observe states directly
  - toy example: the weather

Goldsmiths UNIVERSITY OF LONDON

MPEG-7

- SoundModelStatePath
  - High-level descriptor
  - Most likely state sequence given input



Goldsmiths UNIVERSITY OF LONDON

Structure Detection

- Raw audio: sequence of numbers, windowed
  - AudioSpectrumEnvelope: sequence of vectors
- State path: sequence of numbers, windowed
  - SoundModelStateHistogram: sequence of vectors

Where do we get off?

Goldsmiths UNIVERSITY OF LONDON

Clustering

- State sequence as cluster assignments?
  - HMM states capture dynamics.
- Many possibilities for clustering
  - K-means
  - pairwise clustering
  - histogram clustering

Goldsmiths UNIVERSITY OF LONDON

Example

- Donation to MPEG-7 team from Sony
  - 22 tracks
    - Alanis Morissette: *Head Over Feet*
    - Britney Spears: *Baby Hit Me One More Time*
    - Spice Girls: *Wannabe*
  - 14 expert segmentations
    - IRCAM (Geoffroy Peeters)
    - experts tend to segment functionally

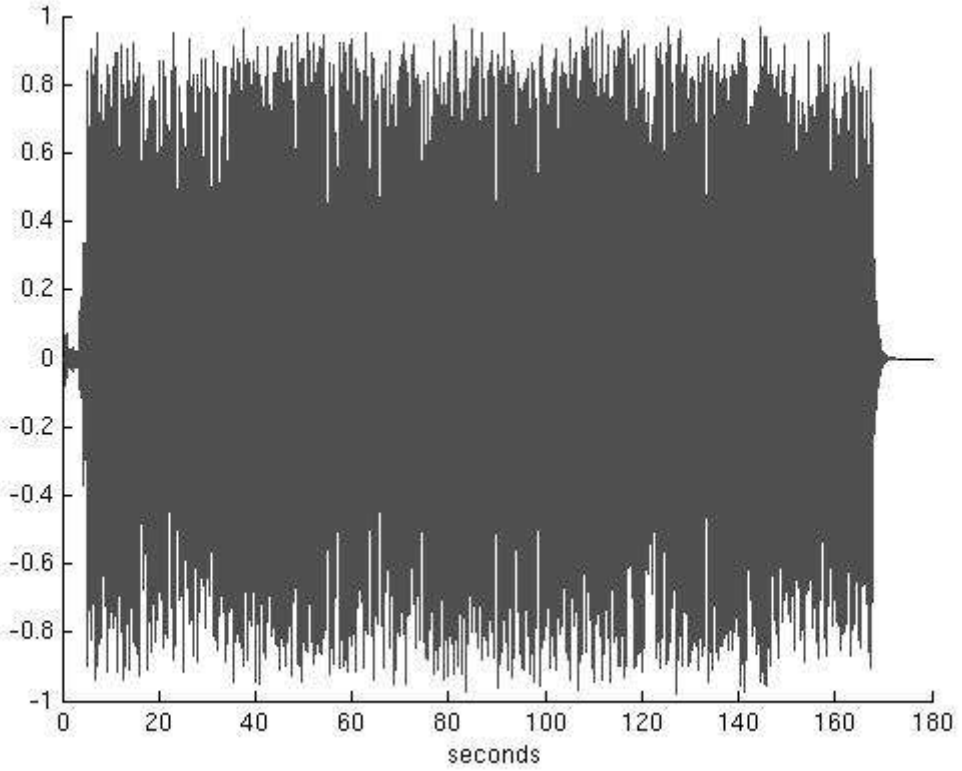
Goldsmiths UNIVERSITY OF LONDON

Example

- Intellectual Property climate different
  - tracks not master quality
  - tracks not even CD quality
    - 11kHz sample rate
    - 16 bits per sample
    - mono
  - do not even know how downsampling was performed
  - potential for experiment!

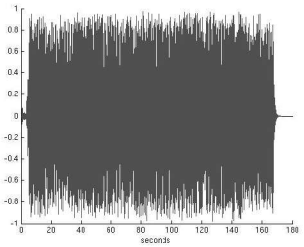
Goldsmiths UNIVERSITY OF LONDON

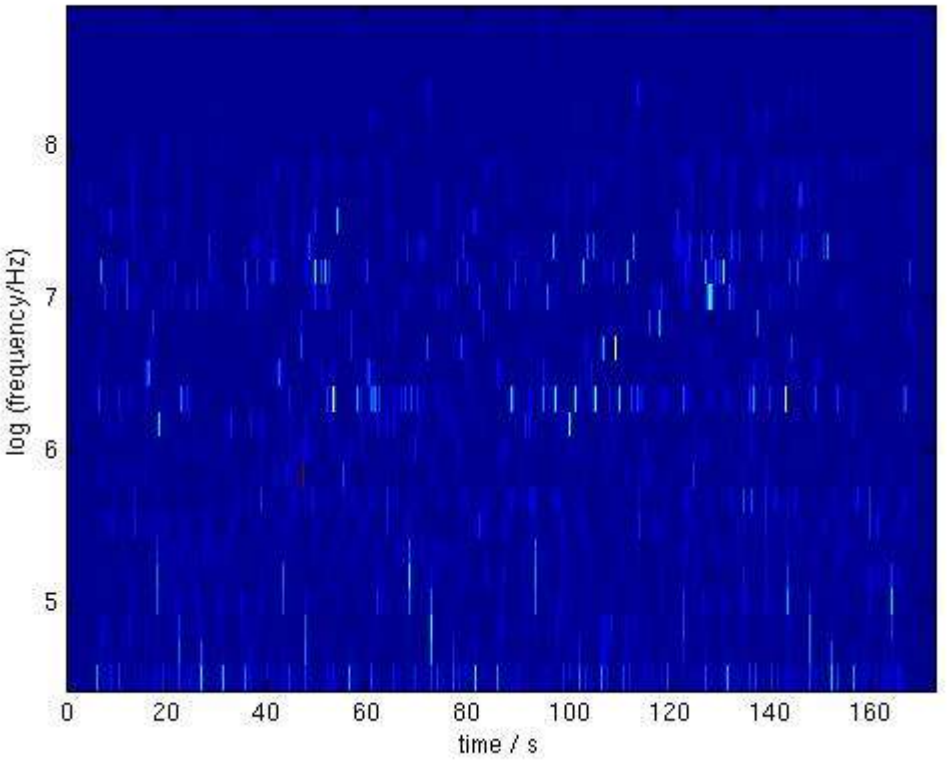
Example: Wannabe



Goldsmiths UNIVERSITY OF LONDON

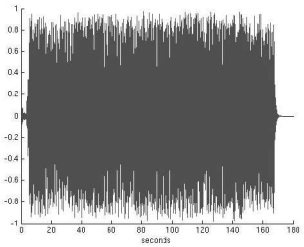
Example: Wannabe

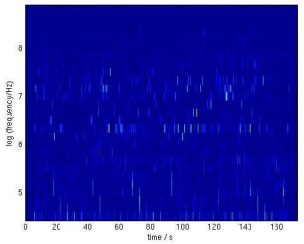


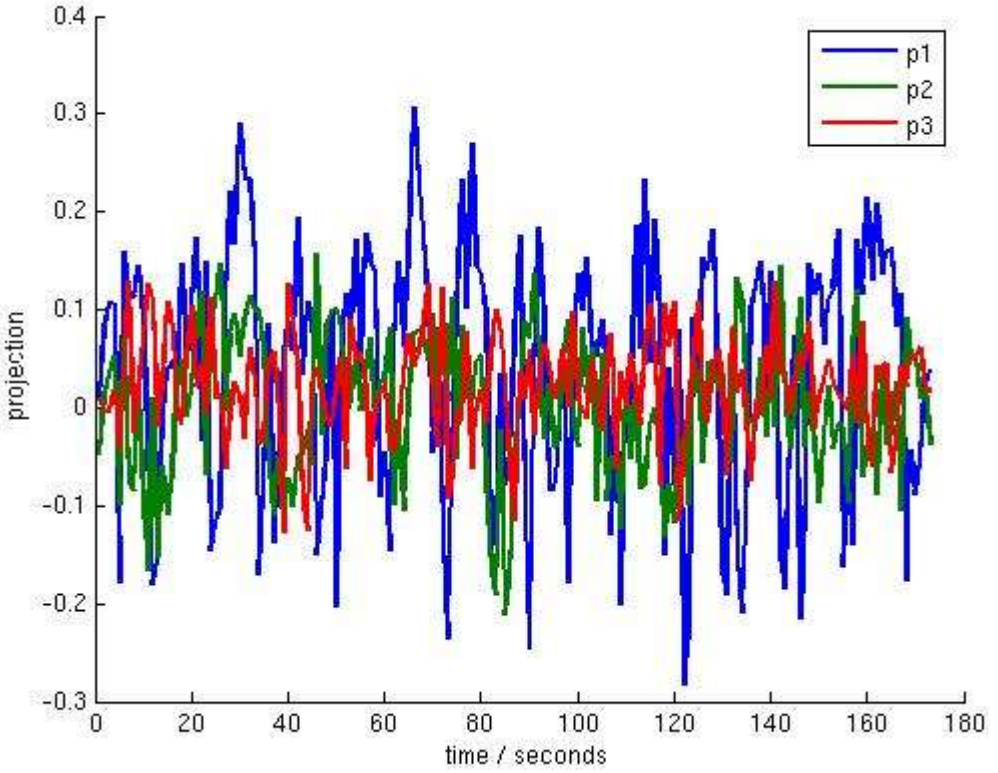


Goldsmiths UNIVERSITY OF LONDON

Example: Wannabe

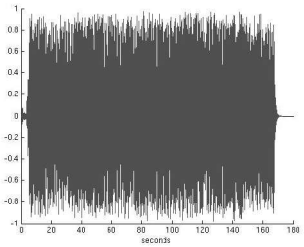


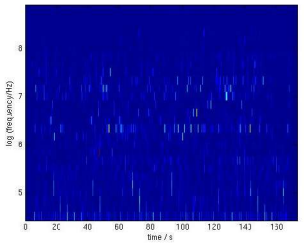


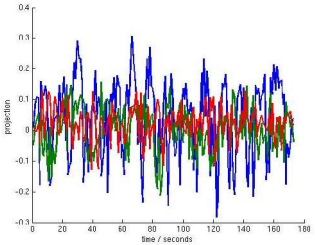


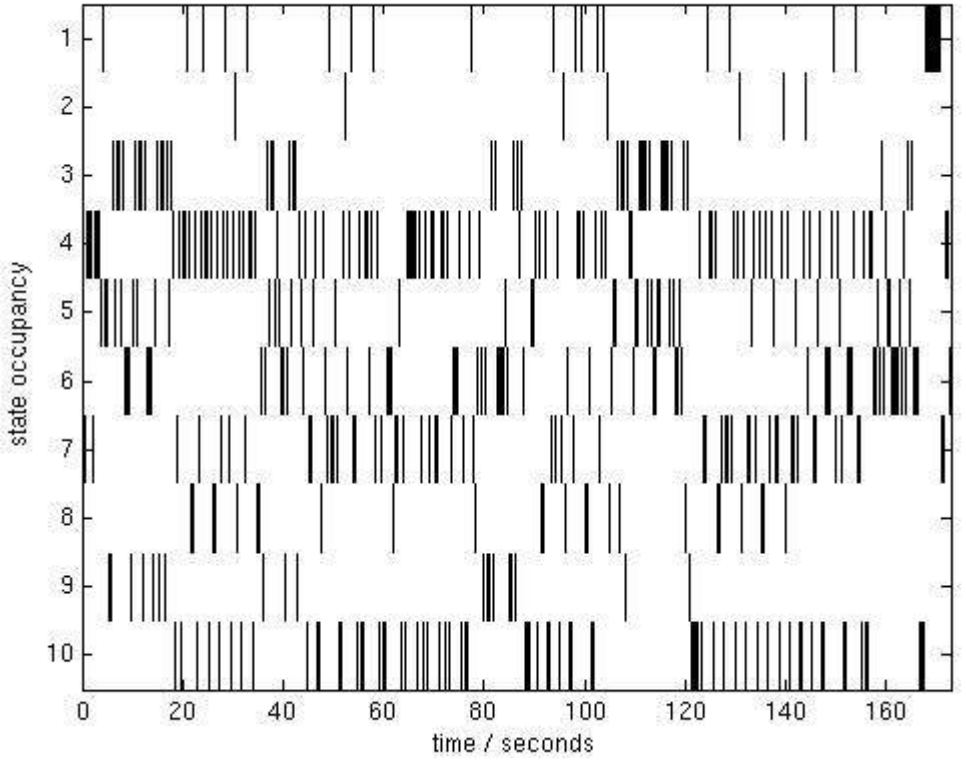
Goldsmiths UNIVERSITY OF LONDON

Example: Wannabe



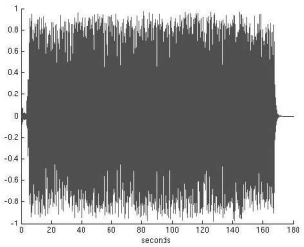


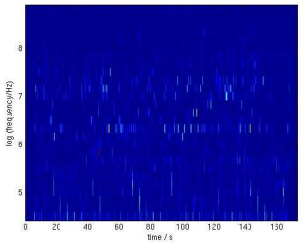


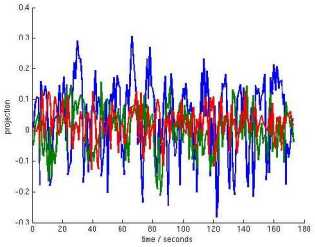


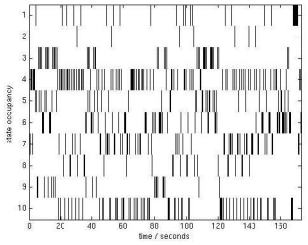
Goldsmiths UNIVERSITY OF LONDON

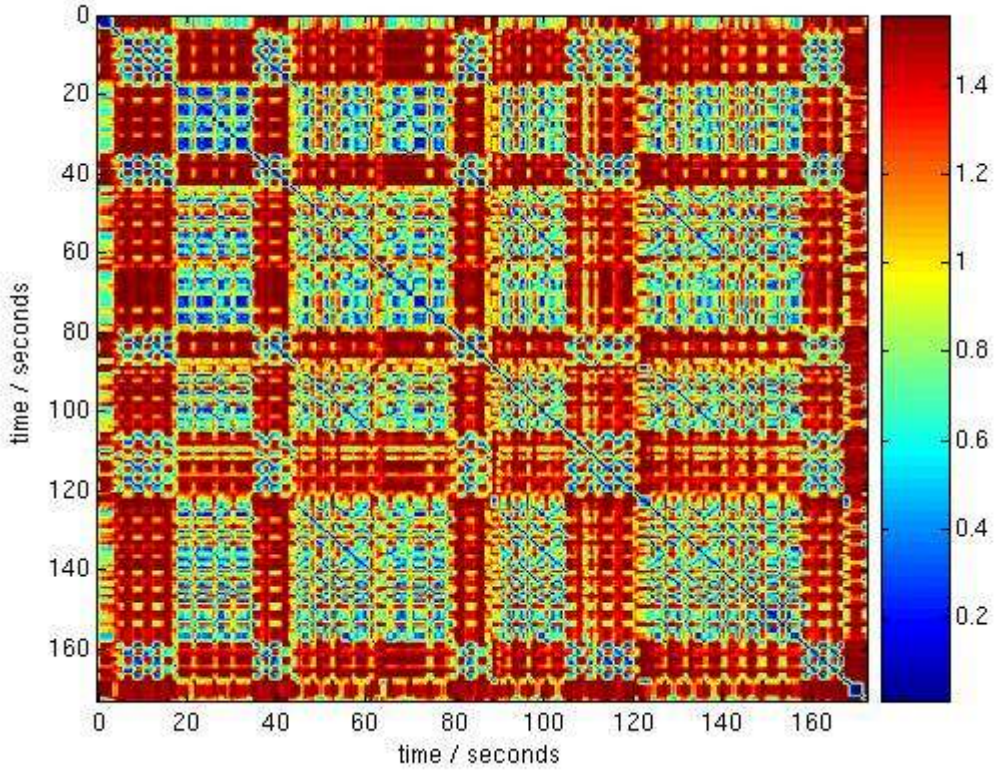
Example: Wannabe





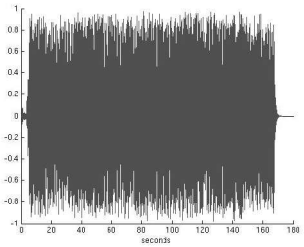


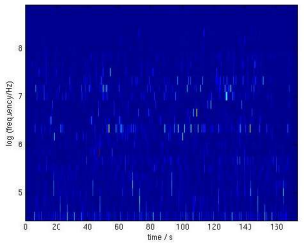


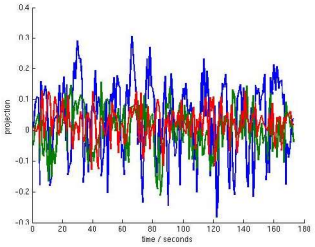


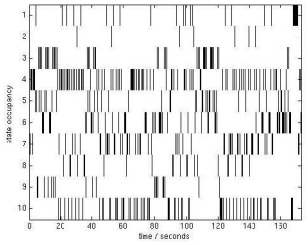
Goldsmiths UNIVERSITY OF LONDON

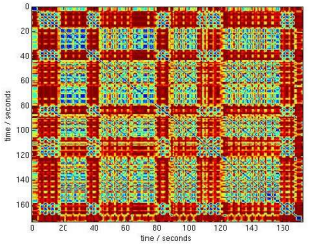
Example: Wannabe

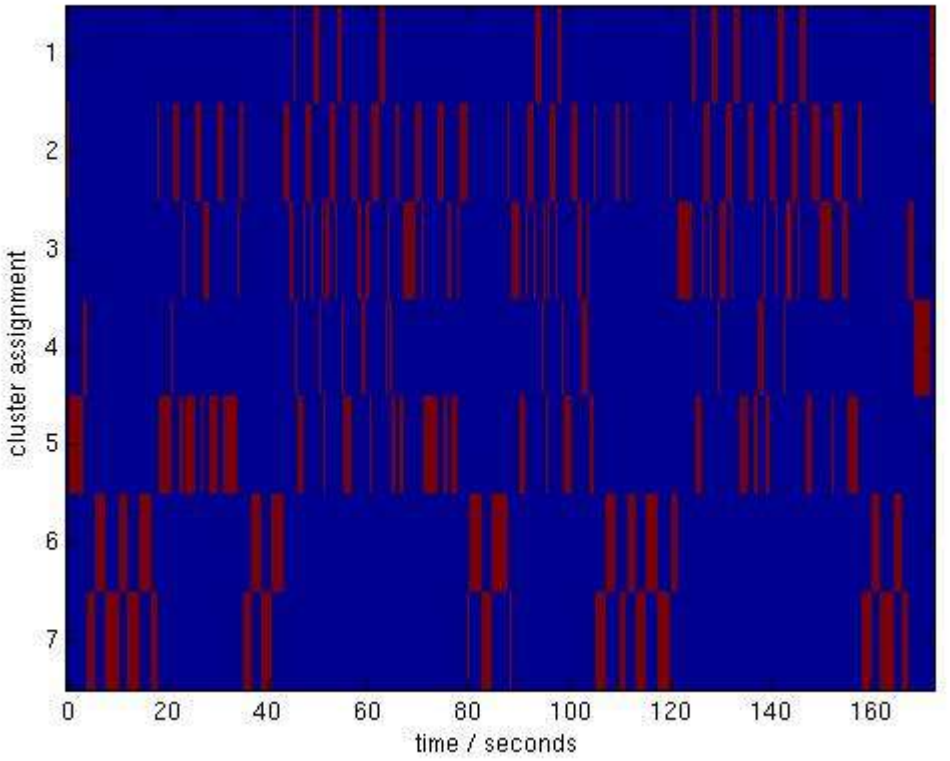








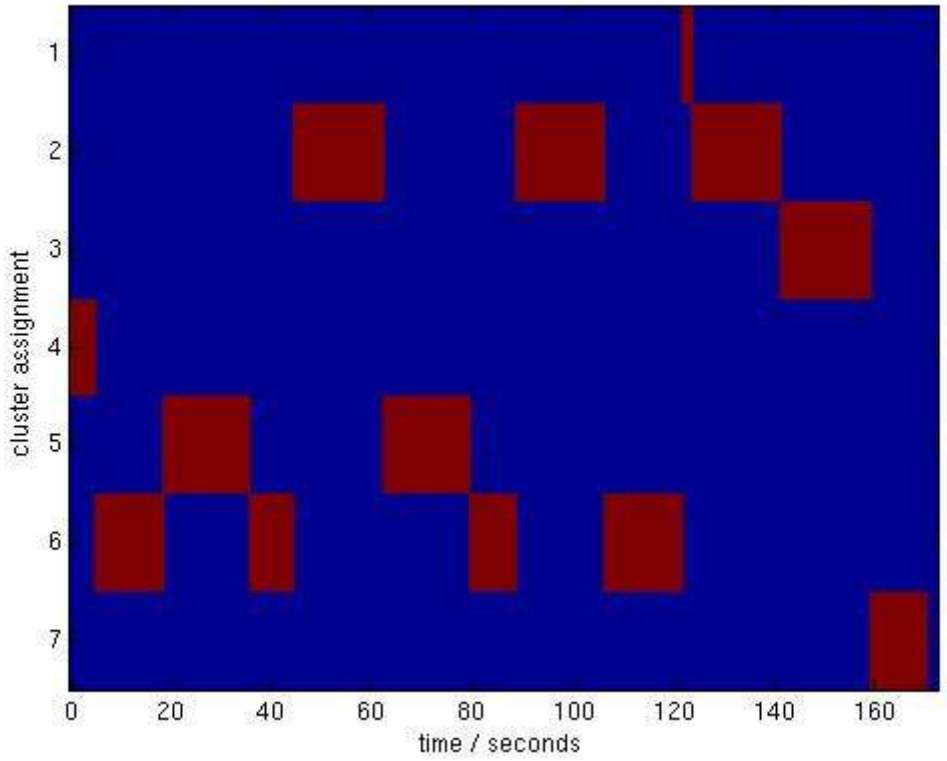


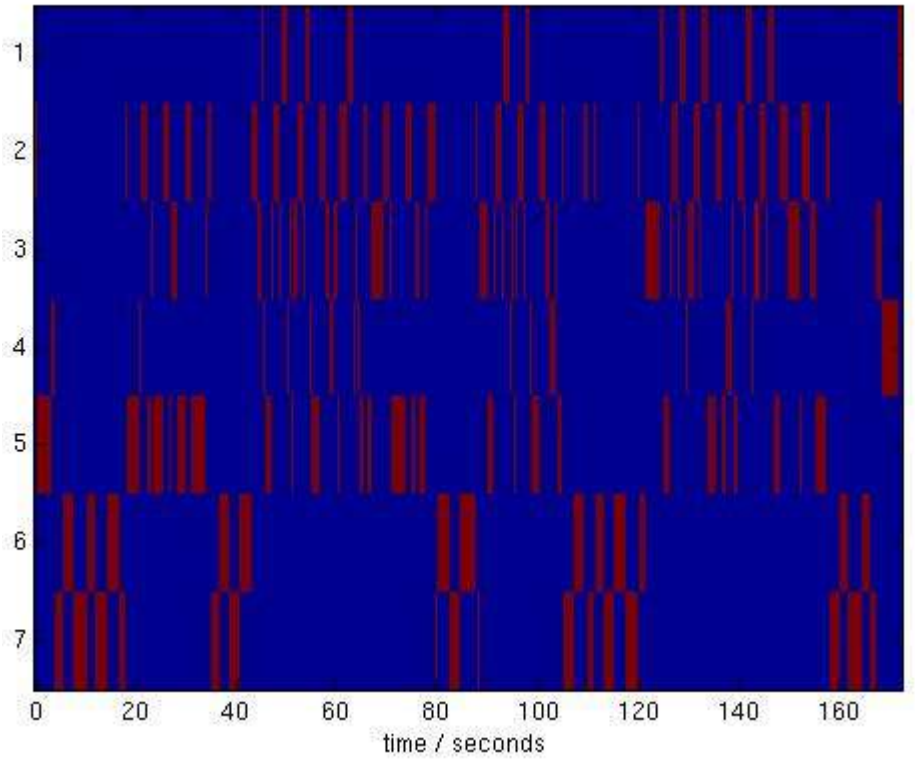


Goldsmiths UNIVERSITY OF LONDON

Example: Wannabe

comparison with expert segmentation

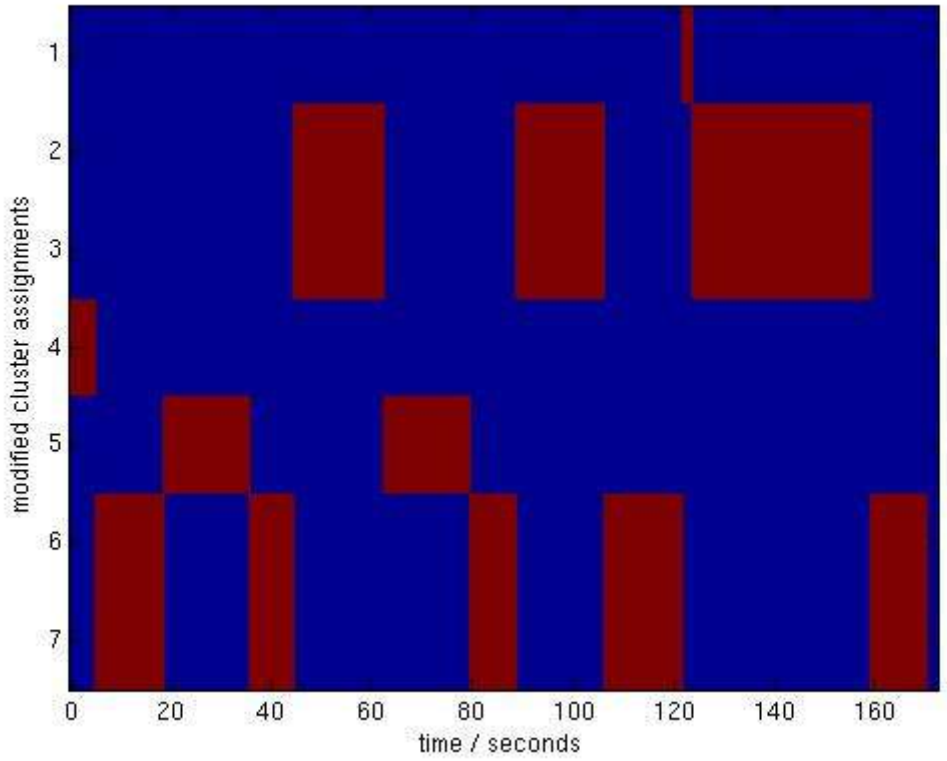


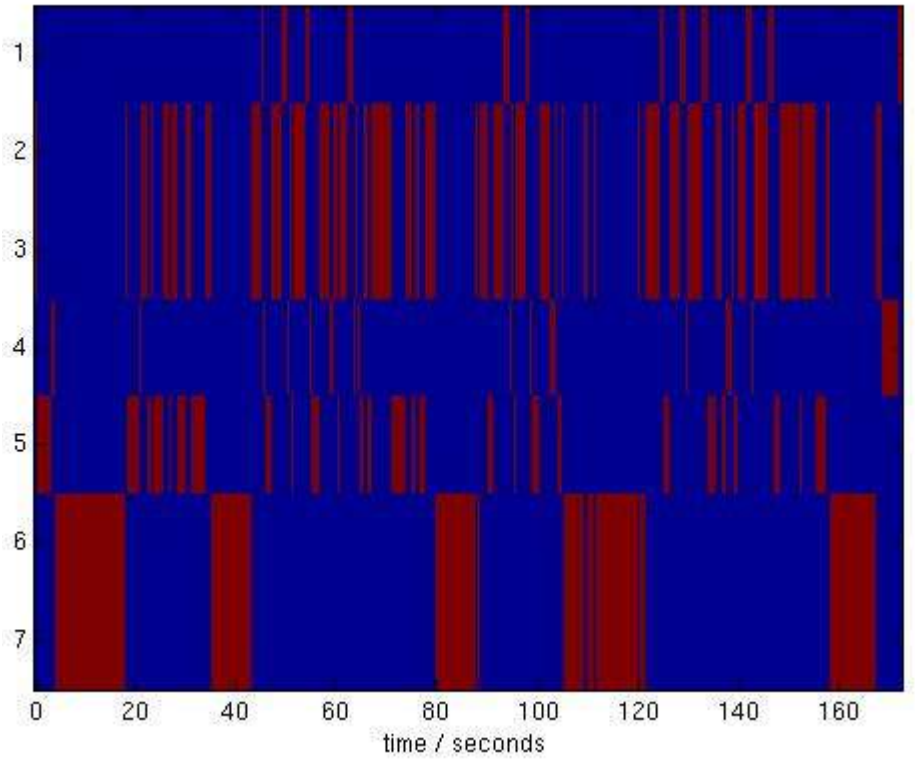


Goldsmiths UNIVERSITY OF LONDON

Example: Wannabe

(informed) cluster agglomeration

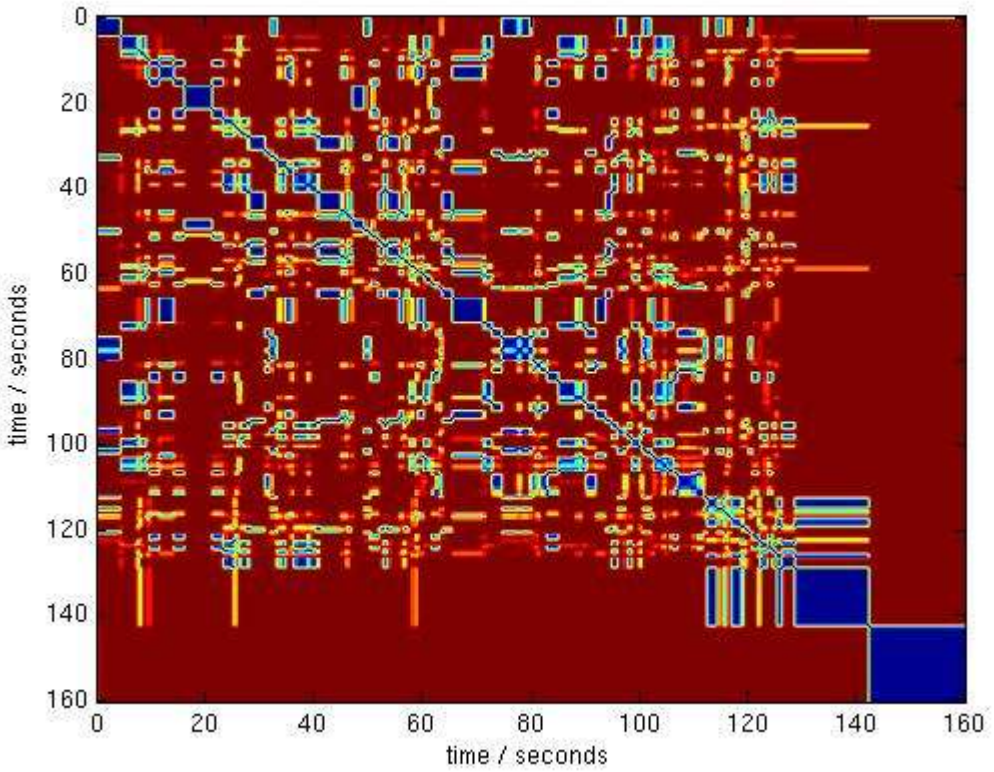




Goldsmiths UNIVERSITY OF LONDON

Example: Gibbons

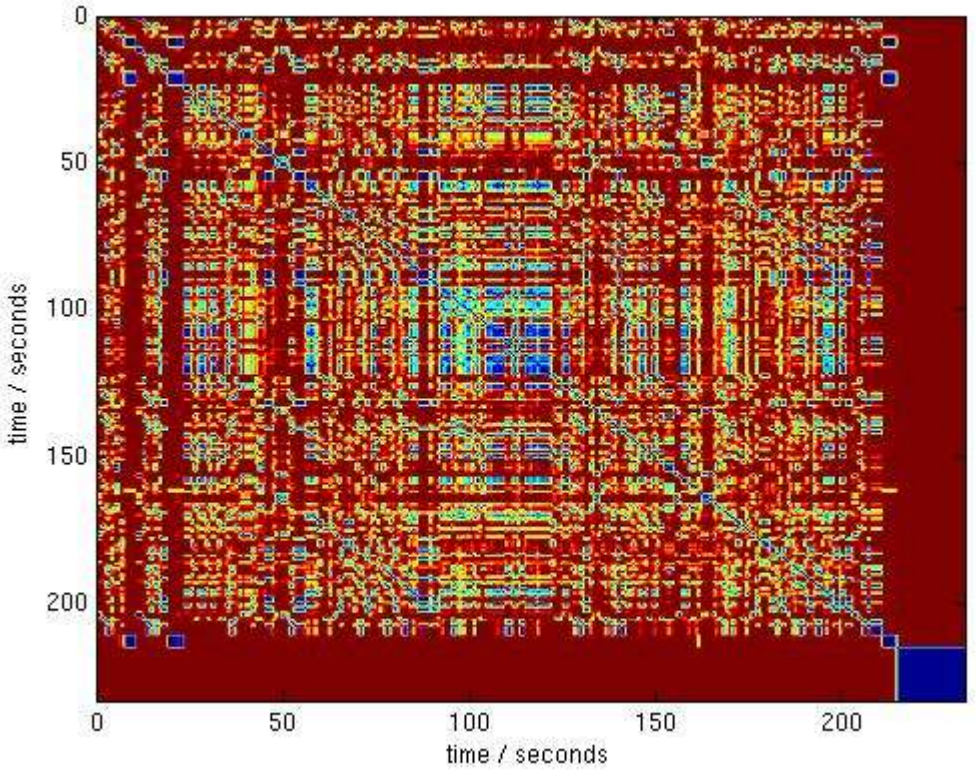
Almighty and everlasting God



Goldsmiths UNIVERSITY OF LONDON

Example: Chopin / Loussier

Impressions on Chopin's Nocturne (Op. 9 No. 1)



Goldsmiths UNIVERSITY OF LONDON

Discussion

- What can we do now?
  - Identify (some) interesting structure...
  - ... in some repertoires
  - building blocks for investigation and applications

Goldsmiths UNIVERSITY OF LONDON

Discussion

- Refinements
  - use segmentations from lower levels (e.g. onset detection)
  - prior on segment length
  - sequence, structural equality
  - full parameter space exploration

Goldsmiths UNIVERSITY OF LONDON

Discussion

- Challenges
  - robust quantification
  - interesting repertoires
    - renaissance vocal music (uniform timbre, no onsets)
    - solo jazz (uniform timbre)
    - poetry (replication of speech recognition)

Goldsmiths UNIVERSITY OF LONDON

Conclusions

- Segmentation is hard
  - highly context-sensitive
  - what is music anyway?
- Can we do it? Yes we can!
  - but only up to a point
  - we want to advance that point further