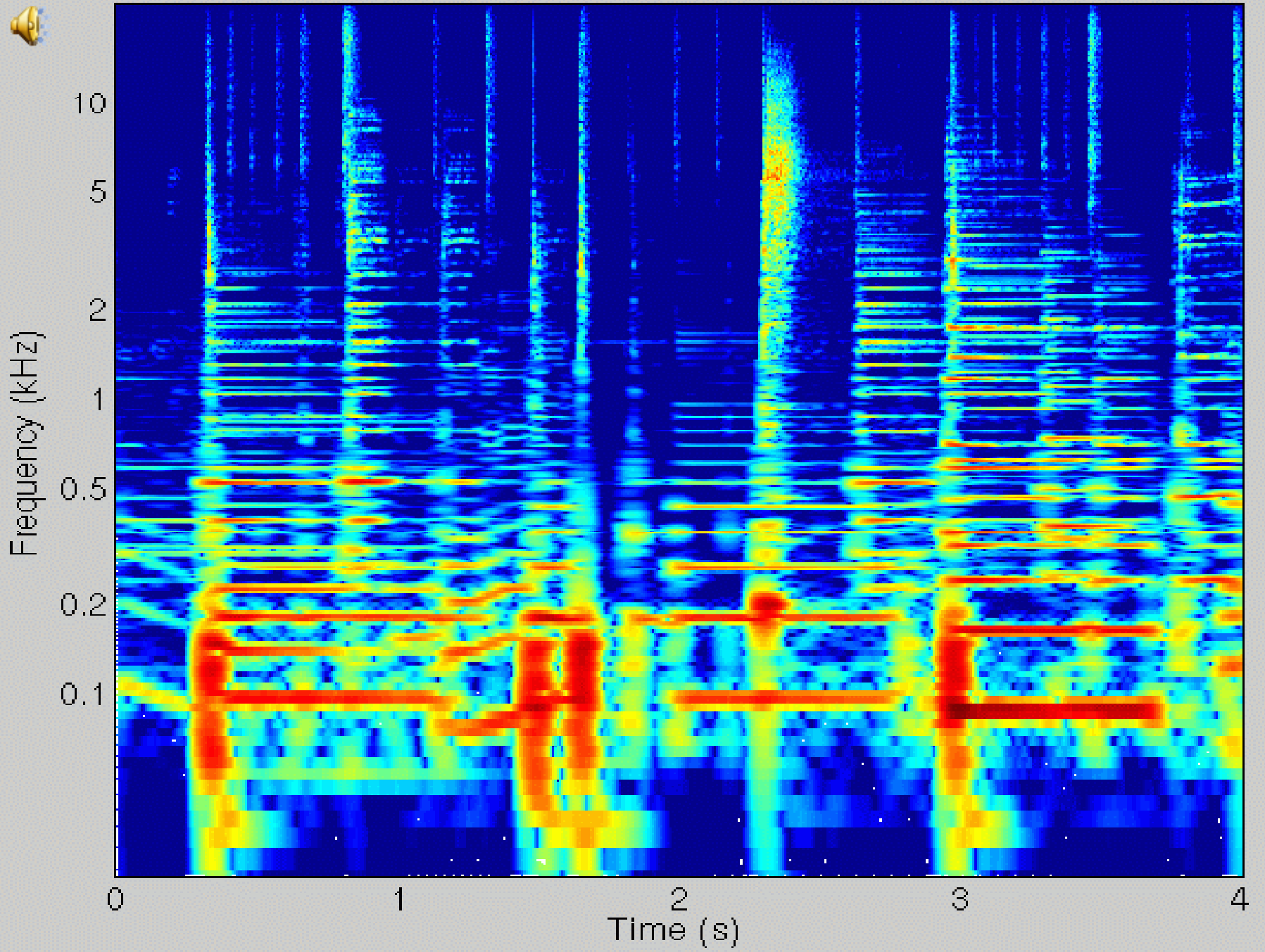


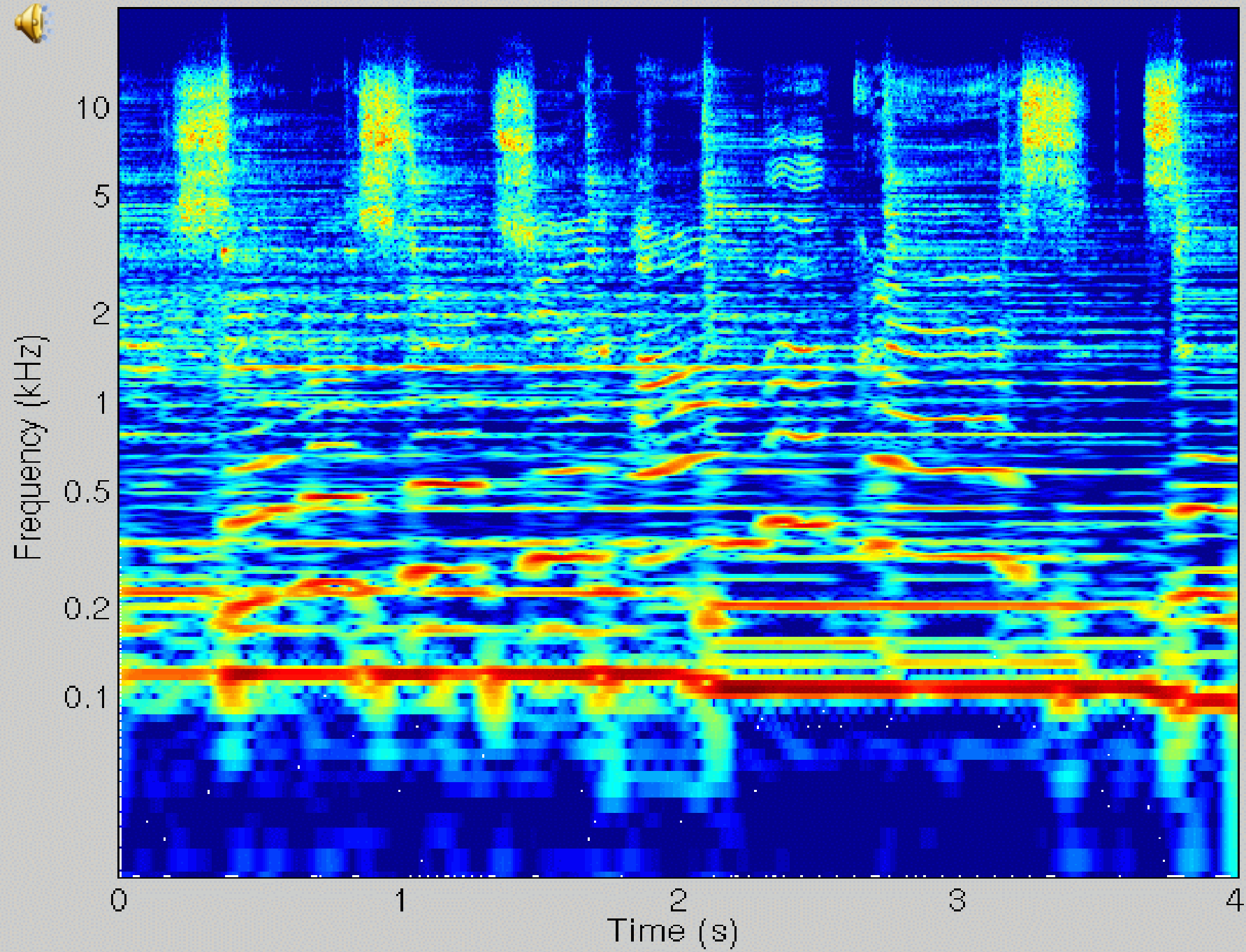
# Extracting meaningful auditory objects from music signals: methods and applications

*Anssi Klapuri (anssi.klapuri@elec.qmul.ac.uk)  
Queen Mary, University of London*

# Breaking up audio signals

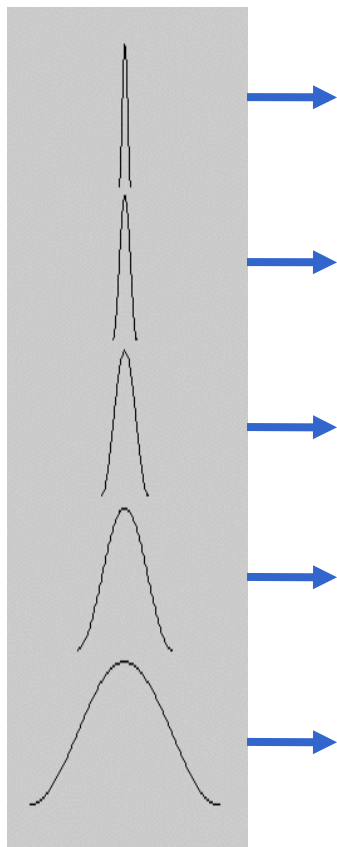
- There are various dimensions along which an audio signal can be decomposed
    - Time (temporal segmenting)
    - Frequency (filtering)
    - Space (angle of arrival)
    - Sinusoids vs. "noise"
    - Sound source separation (various approaches)
- Diagrammatic annotations for the first list item:
- Time (temporal segmenting) } Fundamentals
  - Frequency (filtering) } Fundamentals
  - Space (angle of arrival) } Intermediate difficulty, but "straightforward"
  - Sinusoids vs. "noise" } Intermediate difficulty, but "straightforward"
  - Sound source separation (various approaches) } Ultimate goal, very difficult
- 
- Two different aspects
    - source separation (extracting "layers" of sound)
    - structure analysis (self-similarity analysis, organizing sounds into a structure)



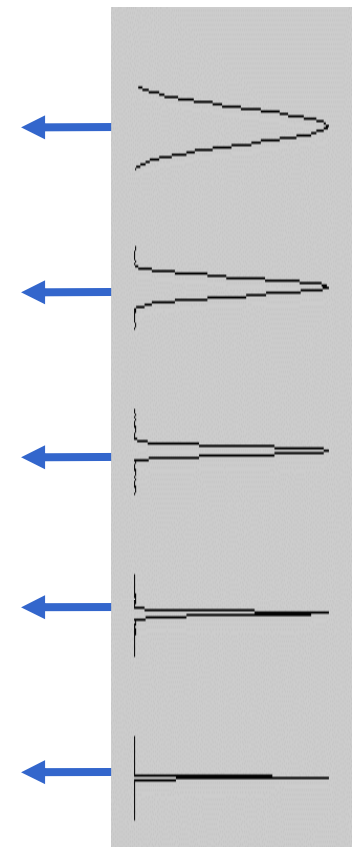
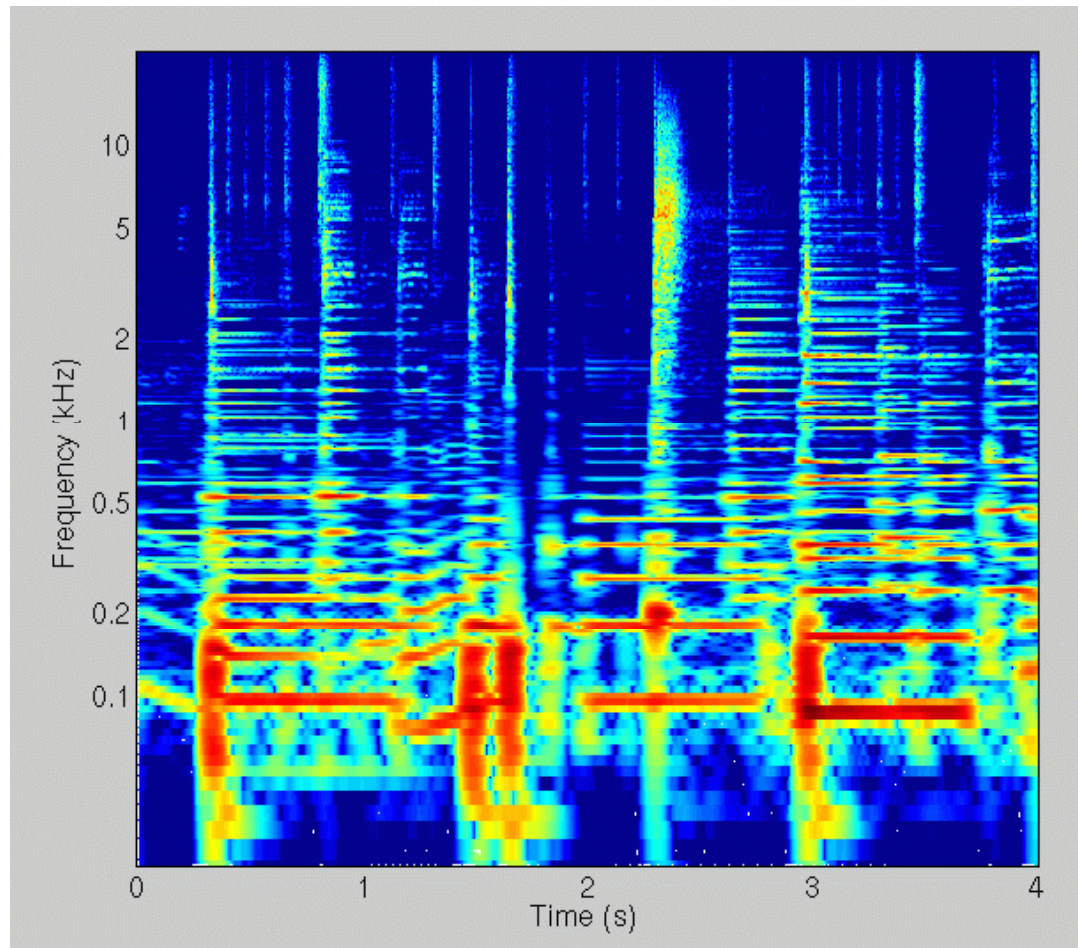


# Constant-Q transform (or bounded Q tr.)

- In humans, time-frequency resolution tradeoff varies with frequency
  - in perceptual audio coding (AAC), best time resolution is 3 ms ( $\rightarrow$  transients)



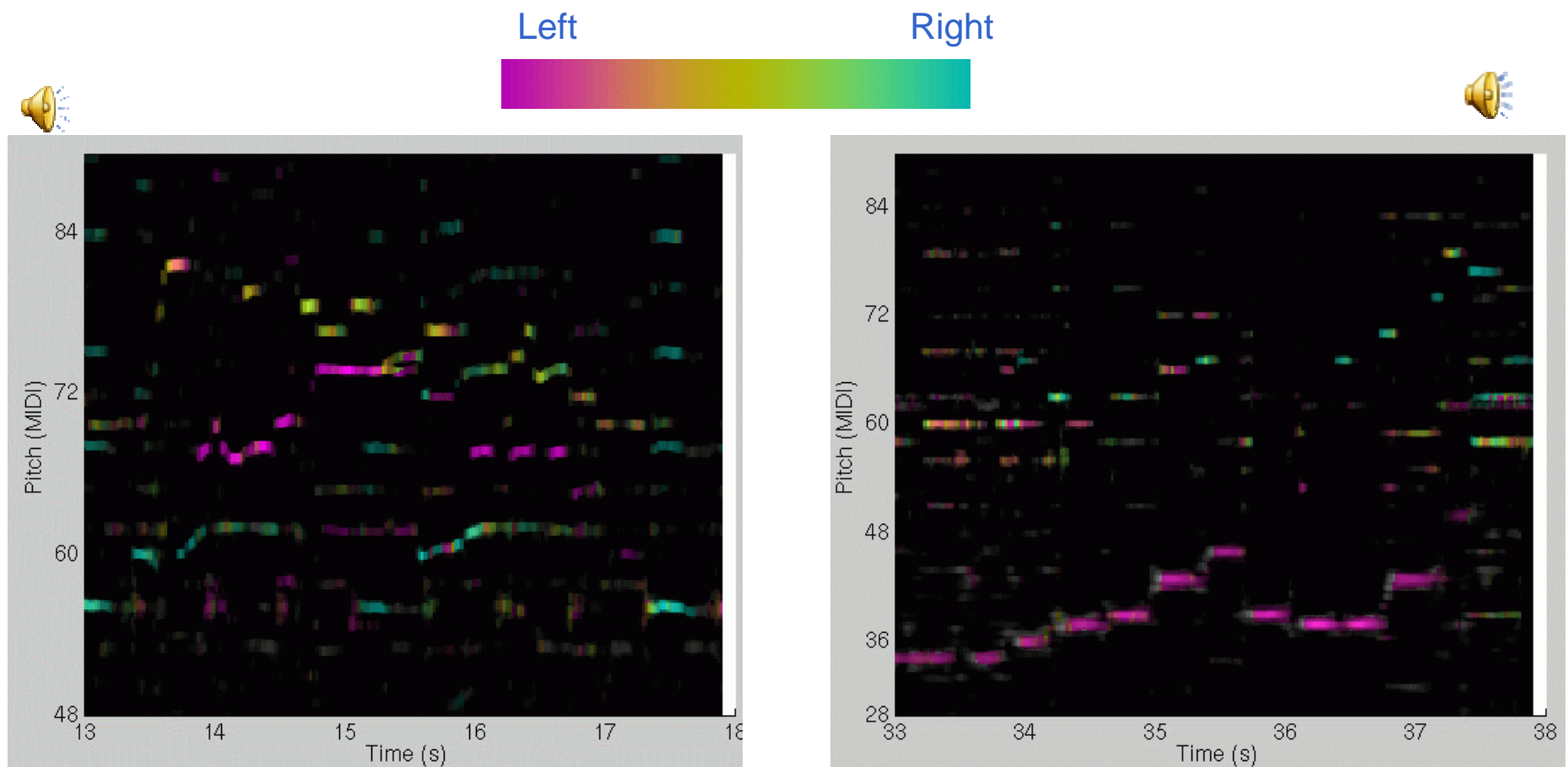
Time-domain window function



Frequency resolution

# Spatial information (angle of arrival)

- Important for human auditory scene analysis (natural environments)
- Usability of spatial information for music analysis depends on genre

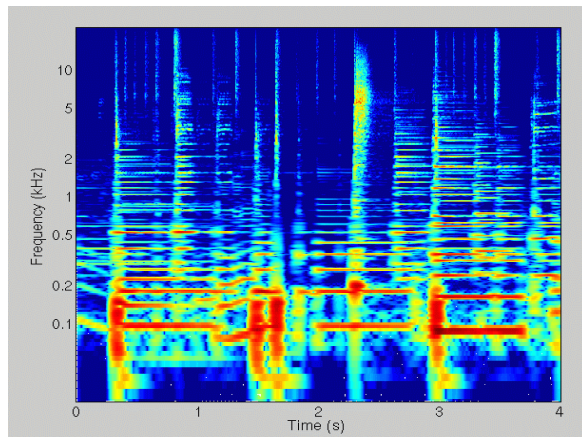


# Sinusoids plus noise model

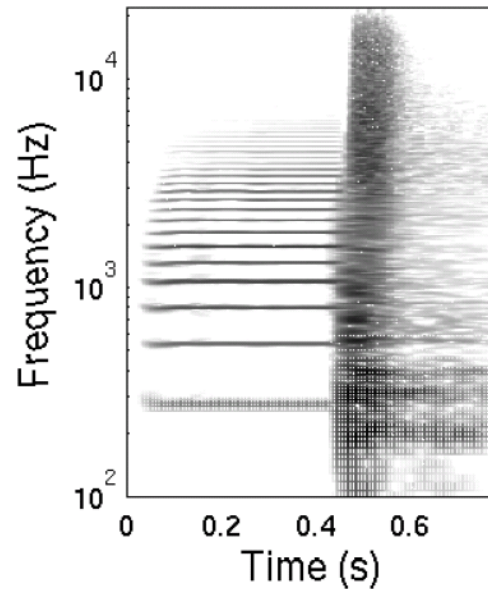
- Harmonic parts and drums (and other "noisy" parts) have so different structure that it makes sense to look at them separately
- Sinusoids + noise model [Serra 1997] is an effective way of separating the two



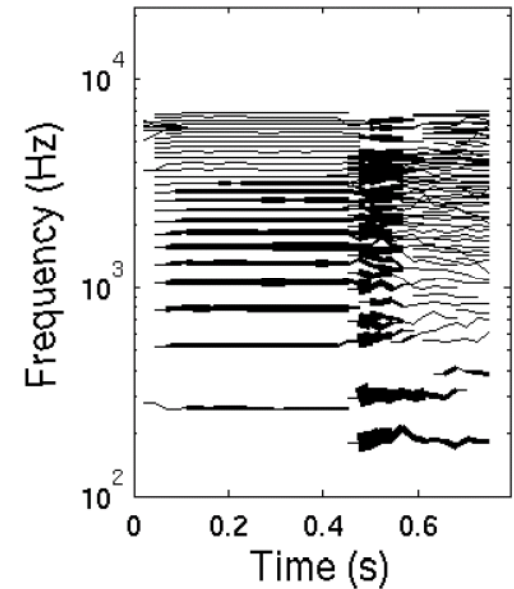
Brentwood jazz quartet



Trumpet tone followed by snare drum hit



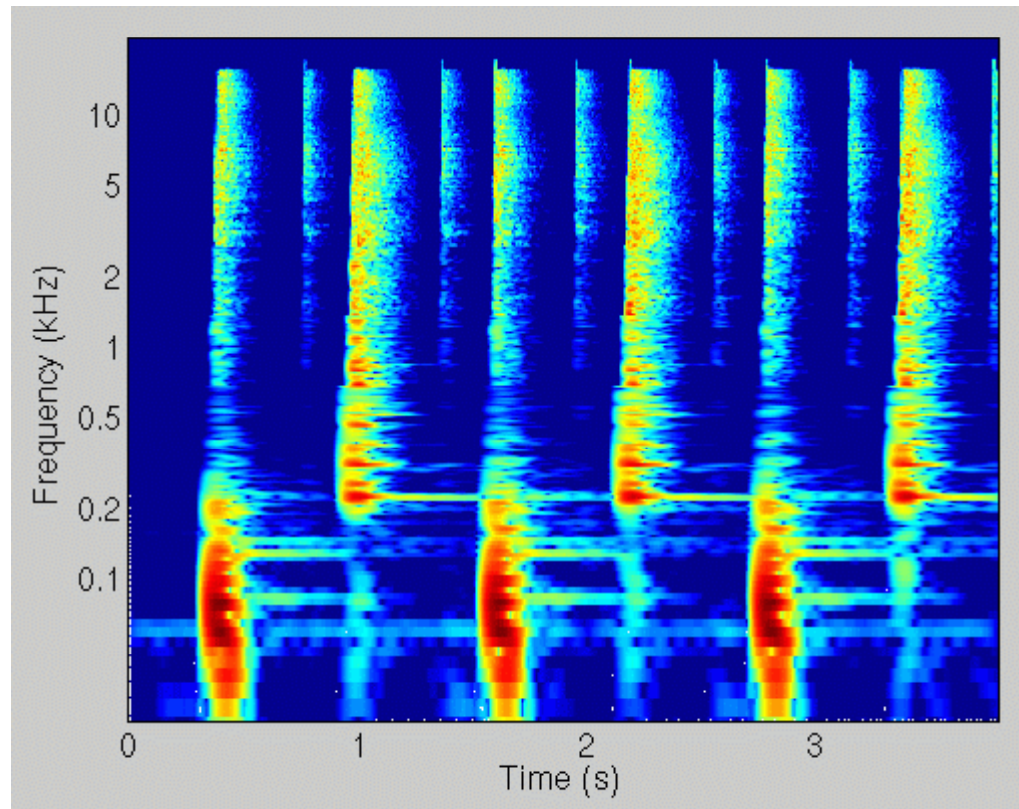
Spectrogram



Sinusoidal model

# Drum track analysis

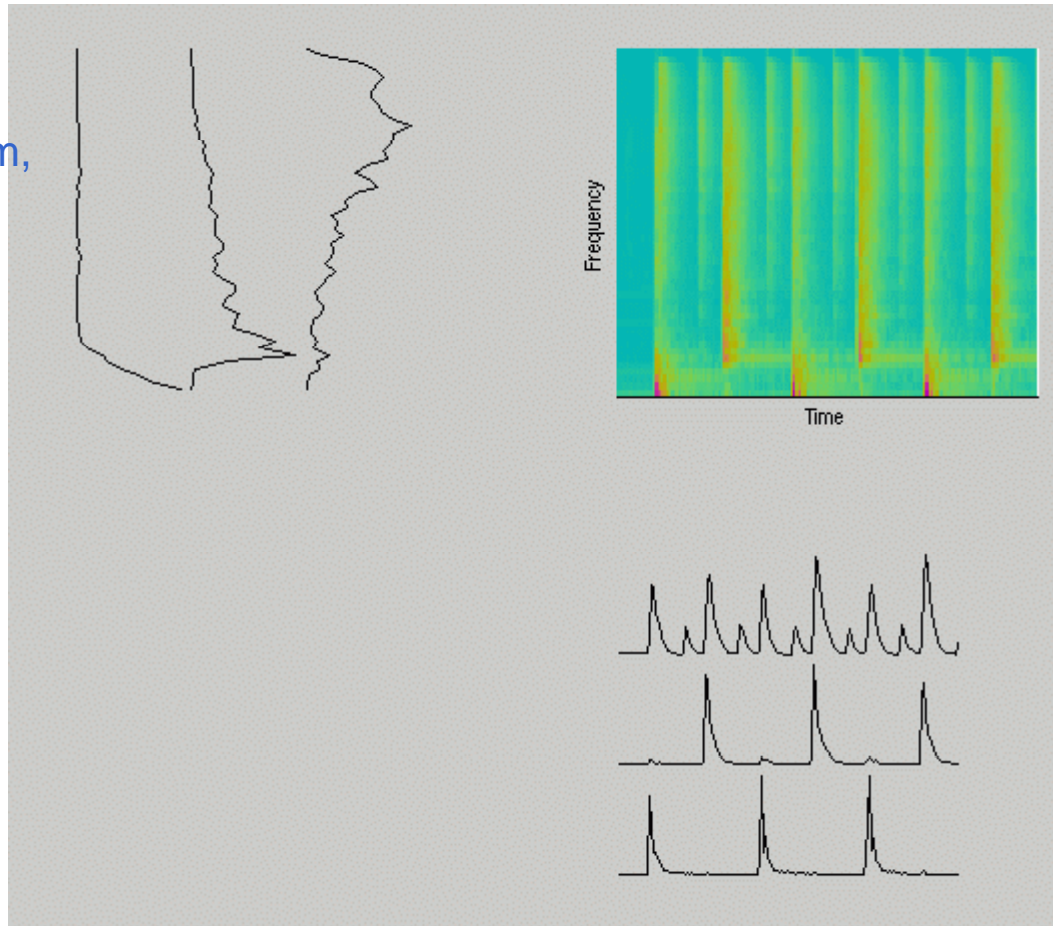
- Characteristic to drum sounds (and sound beginning transients) is that
  - 1) they are short (transient-like)
  - 2) they repeat with relatively little spectral variation (except for playing styles)



# Non-negative matrix factorization (NMF)

- NMF is ideal for analyzing a signal which consists of the superposition of a limited set of spectra. Extremely simple to implement.

Basis functions  
(spectra of bass drum,  
snare drum, hi-hat)

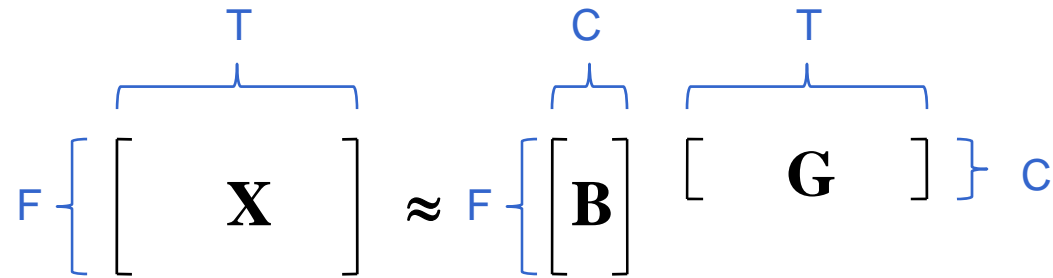


Time-varying  
gains of the  
three spectra

# Drum track analysis with NMF

- Signal model:

$$\mathbf{X} \approx \mathbf{B}\mathbf{G}$$



Magnitude spectrogram

Columns of **B**: basis spectra

Rows of **G**: time-varying gains

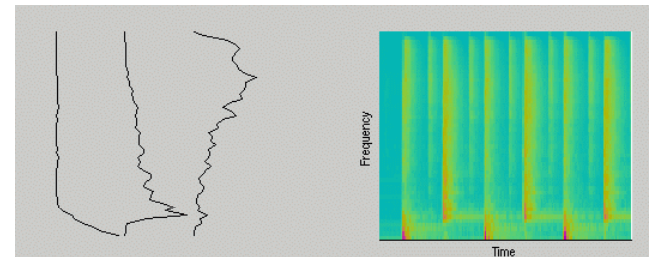
- Algorithm

1. Initialize each entry of **B** and **G** with the absolute values of Gaussian noise
2. Update **G** using the formula below
3. Update **B** using the formula below
4. Repeat (2)--(3) until the values converge

$$\mathbf{B} \leftarrow \mathbf{B} \cdot \frac{(\mathbf{X} ./ \mathbf{B}\mathbf{G})\mathbf{G}^T}{\mathbf{1}\mathbf{G}^T}$$

$$\mathbf{G} \leftarrow \mathbf{G} \cdot \frac{\mathbf{B}^T(\mathbf{X} ./ \mathbf{B}\mathbf{G})}{\mathbf{B}^T\mathbf{1}}$$

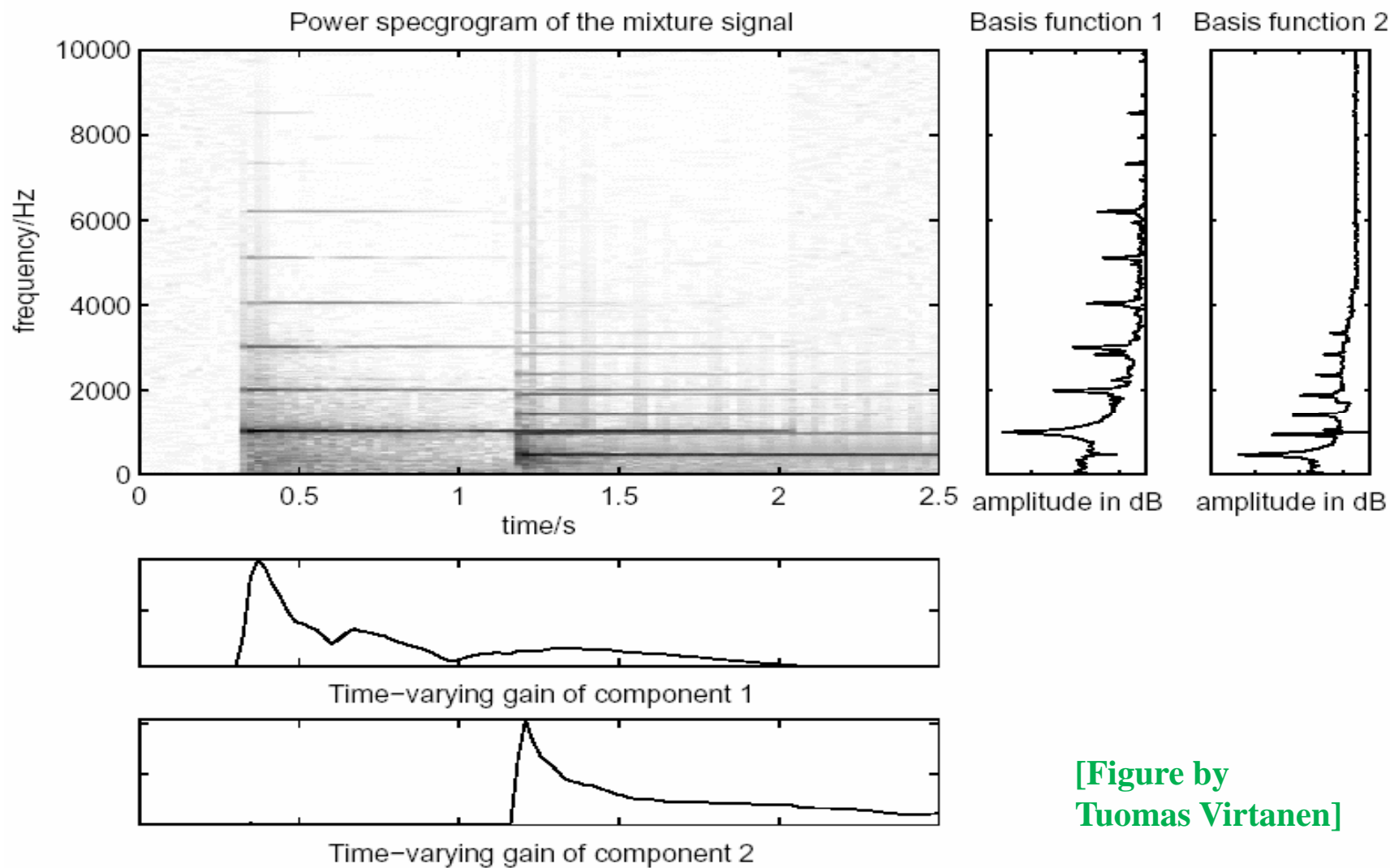
**B:**



**G:**



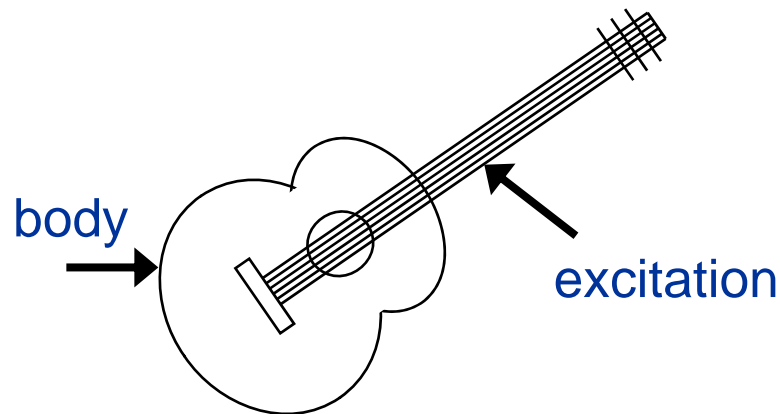
# Harmonic part: NMF...?



[Figure by  
Tuomas Virtanen]

# Excitation-filter model

- "Excitation" represents a vibrating object such as a guitar string and "filter" refers to the resonance structure of the rest of the instrument which colors the produced sound
- Used for decades in speech coding and sound synthesis, but has only recently been used for audio signal analysis



# Sound source separation



Toni Heittola



Tuomas Virtanen

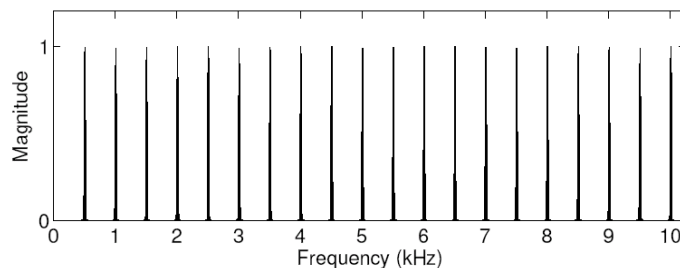
- Combine excitation-filter model, multipitch estimation, and NMF
  - 1) multipitch analysis → excitation spectra
  - 2) augmented NMF → instrument body responses and note gains
  - Details in [Heittola, Klapuri, Virtanen, ISMIR 2009]

$$\hat{x}_t(k) = \sum_{n=1}^N \sum_{i=1}^I g_{n,i,t} e_{n,t}(k) \sum_{j=1}^J c_{i,j} a_j(k)$$

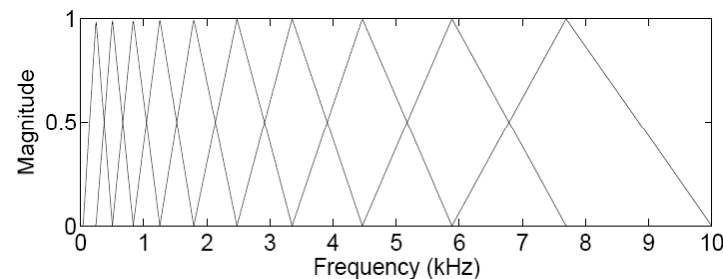
Excitations obtained  
using multipitch estimation

Filters and gains obtained  
using augmented NMF

x gains



x



# Sound source separation

- Audio examples

mixture



source 1



source 2

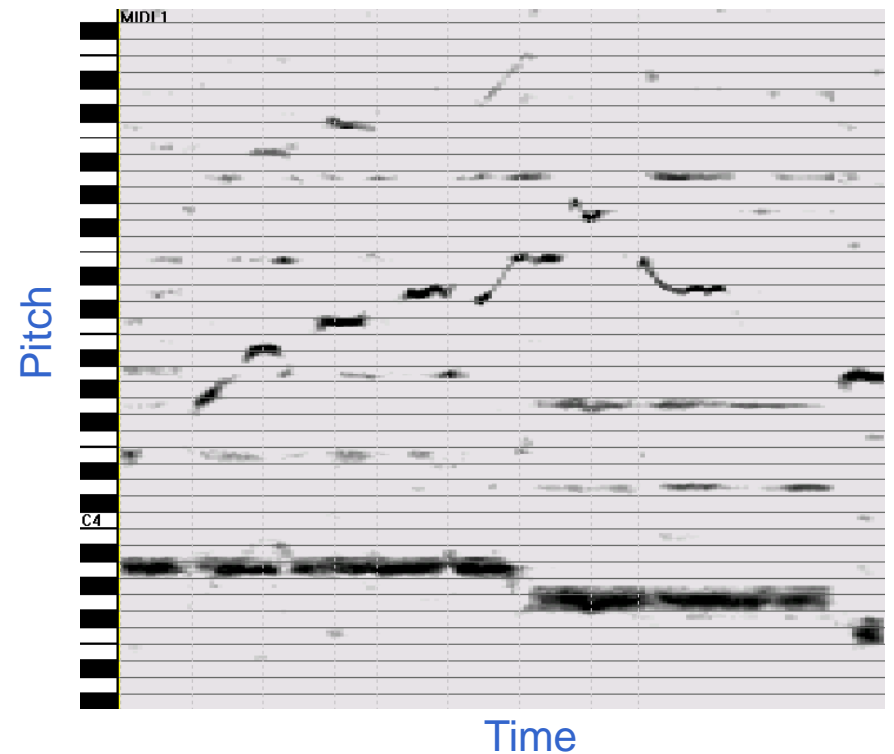
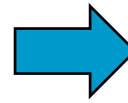
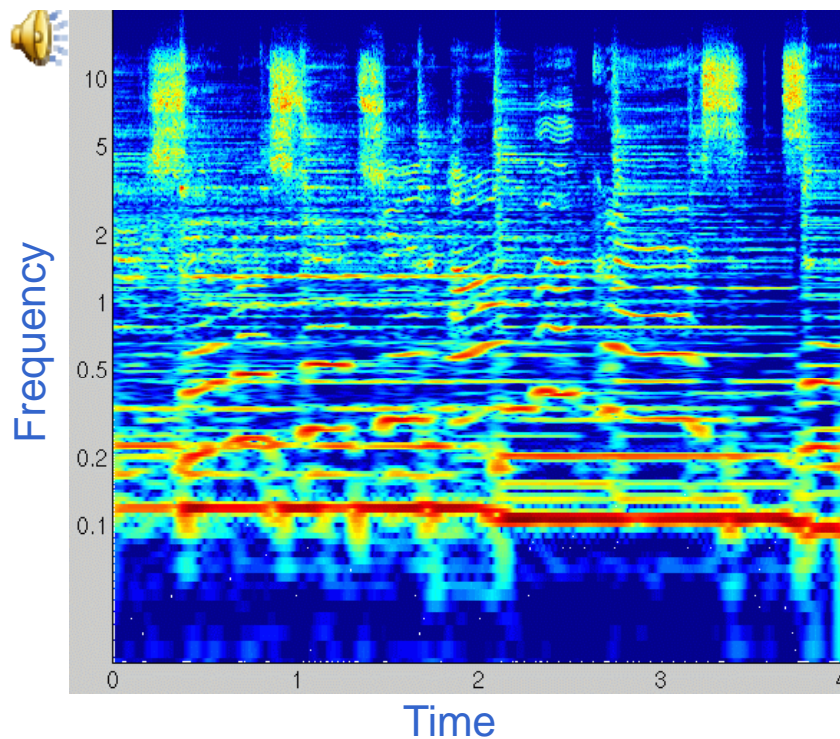


source 3



# Pitch salience – another approach to analyse the harmonic part

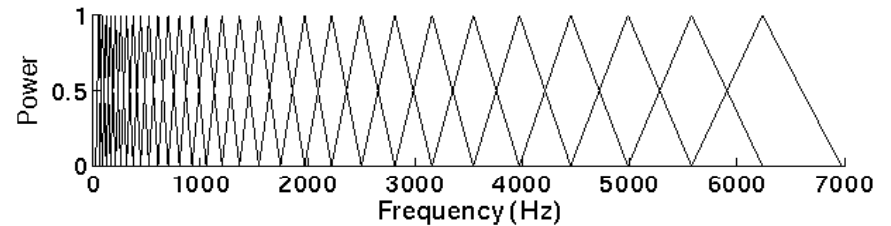
- Pitch perception instead of frequency: auditory system tends to use a single frequency value (= pitch) to summarize certain aspects of sound
- We need mapping from (Time x Frequency)  $\rightarrow$  (Time x Pitch)



# Pitch salience computation

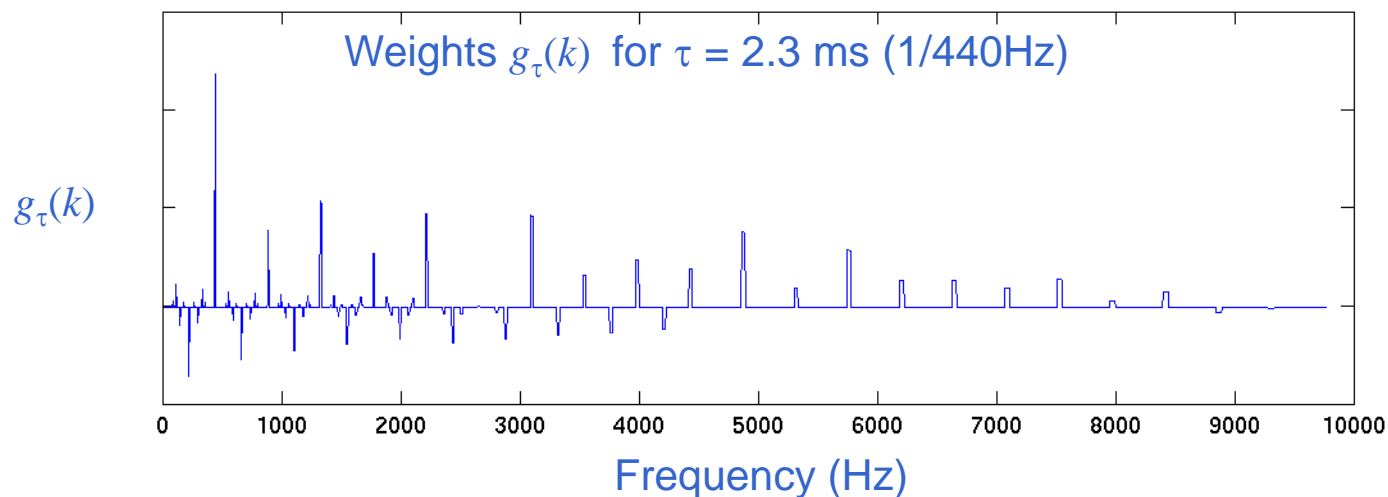
- Spectral flattening (whitening)

- calculate DFT  $X(k)$  within one frame
- weight critical bands inversely proportional to their power
- whitened spectrum  $Y(k)$



- Linear transform from frequency domain to pitch domain [Klapuri-ISMIR-09]

$$s(\tau) = \sum_k g_\tau(k) Y(k)$$

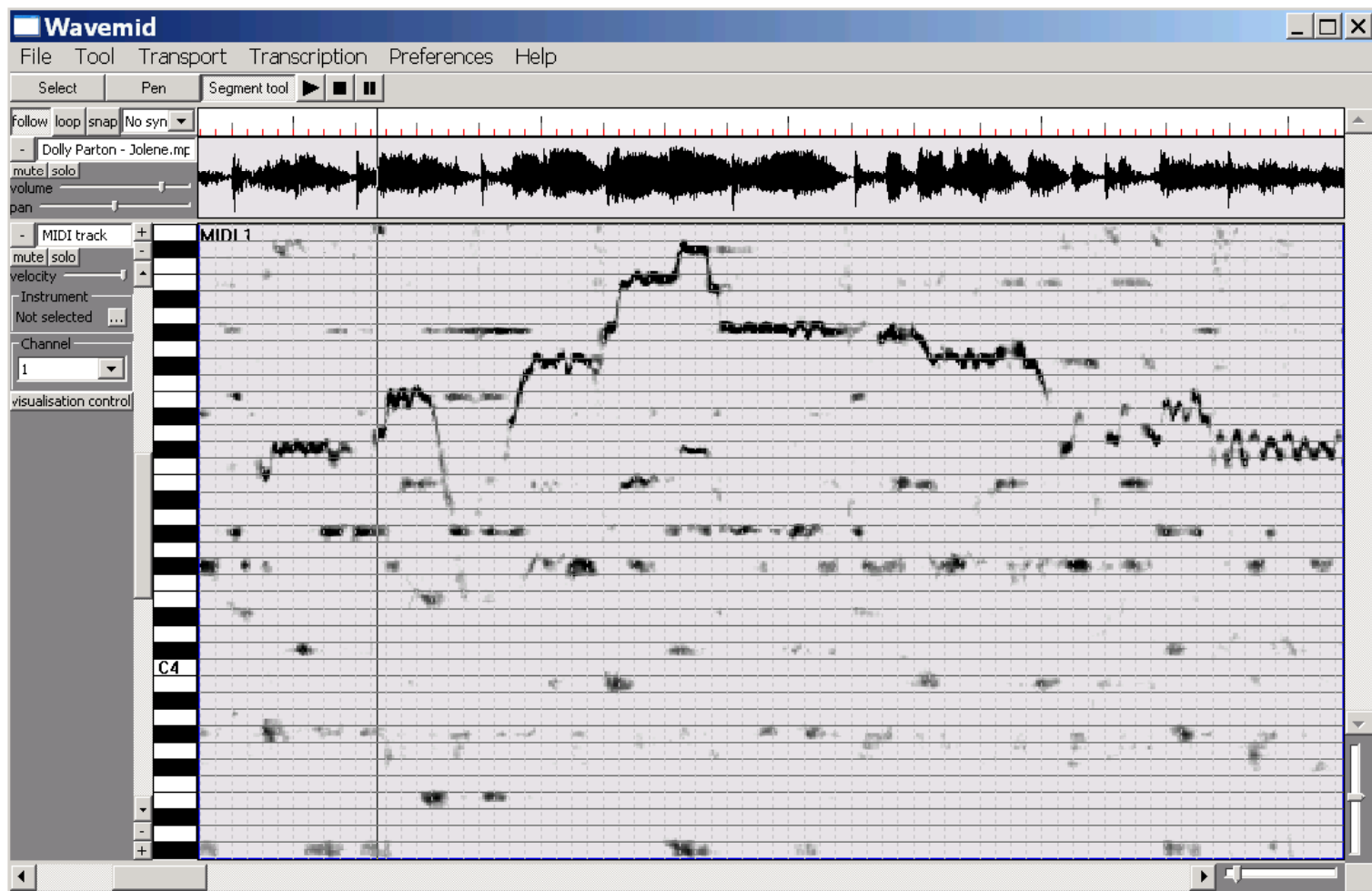


# Pitch salience extraction



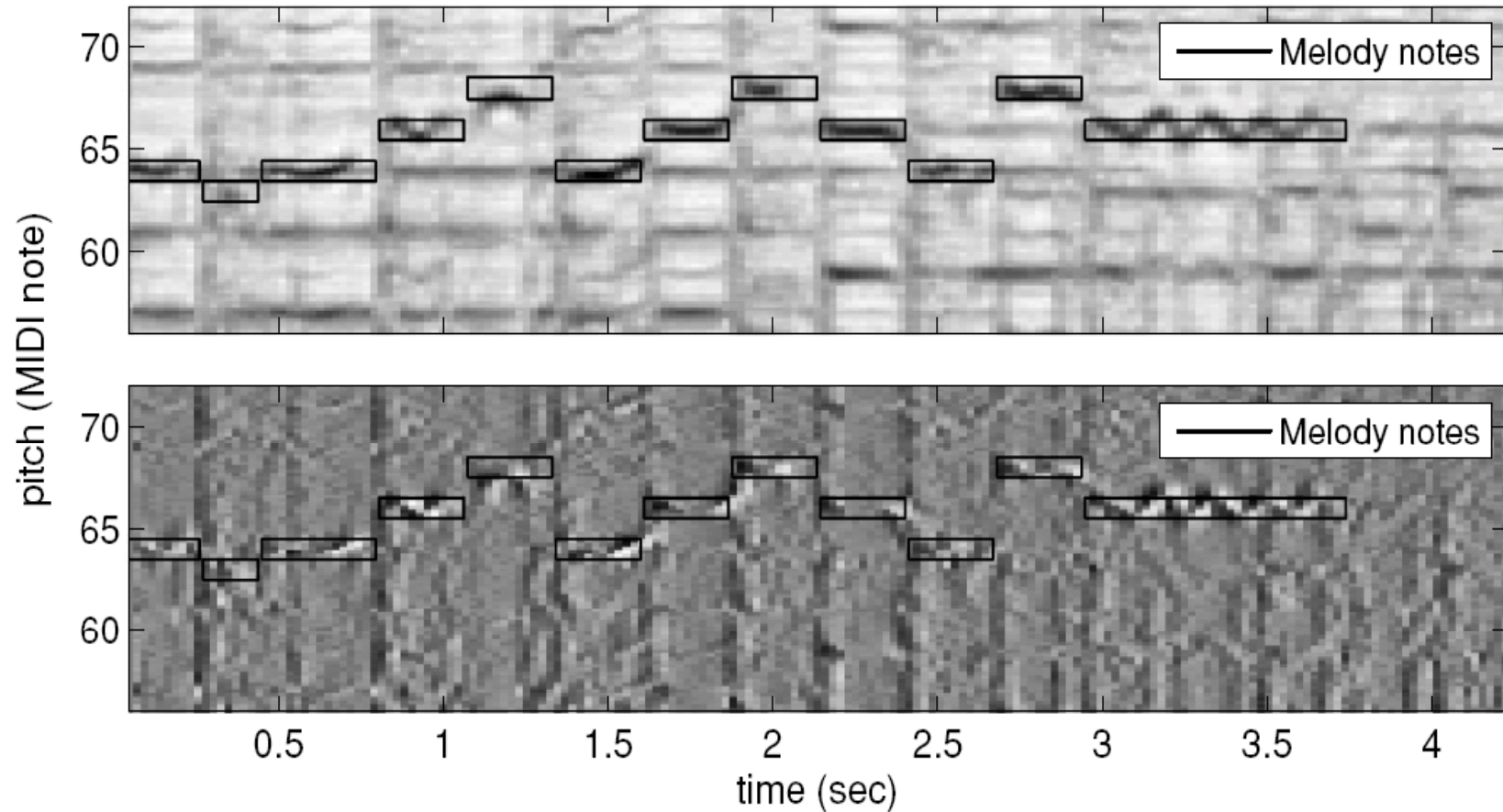
Teemu Karjalainen

- Demo



# Delta salience

- Time differential of the salience reveals onsets of pitched sounds and sounds with vibrato (such as the human voice)



# Vocals separation

- Vocals separation
  - based on melody extraction and unsupervised learning and suppression of the accompaniment
  - [Virtanen, Mesaros, Ryyänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music", ISCA Workshop on Statistical And Perceptual Audition, Brisbane, Australia 2008.]
- Result: SNR of lead vocals improved from  $-4.0$  dB to  $+4.9$  dB on the average
  - realistic test material from Karaoke DVDs
- Audio examples

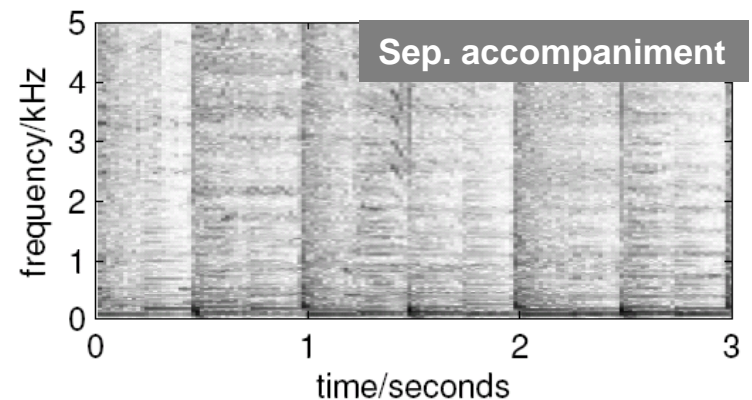
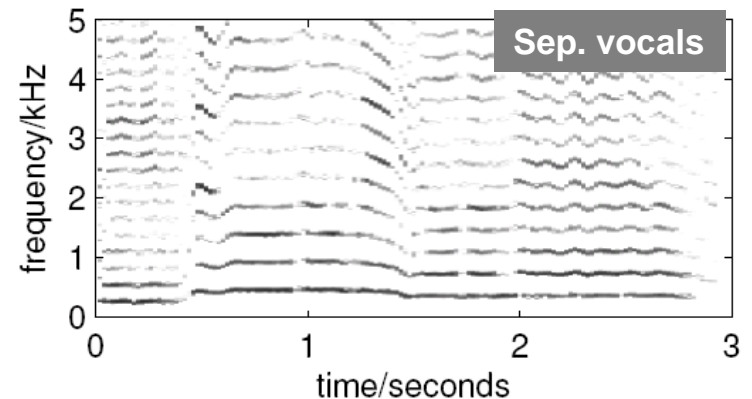
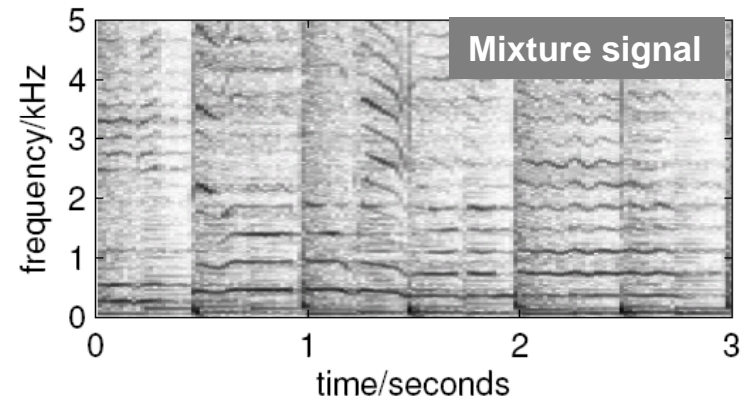
Mixture



Sep. vocals



Sinusoidal separation



# Applications of vocals separation

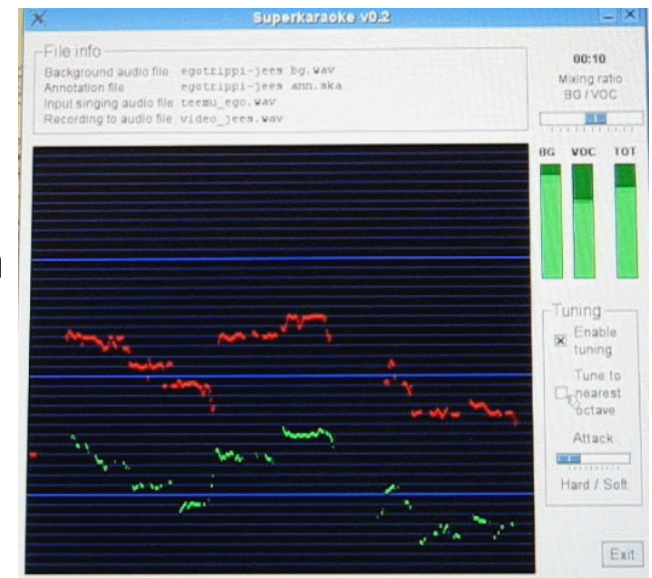



Matti Ryyänen

- Produce karaoke versions automatically



- Karaoke application: include pitch correction
  - automatically remove lyrics from a piece and replace them with user's singing
  - [Ryyänen-Virtanen-Paulus-Klapuri-ICME-2008]
- Replace lyrics with your own



- Query by humming
  - query  retrieval results #1  #2  #3 

# Further applications of vocals and lyrics analysis



*Annamaria Mesaros*

- Singer identification and vocals similarity
- Alignment (synchronization) of given lyrics to musical audio
- Keyword spotting in singing



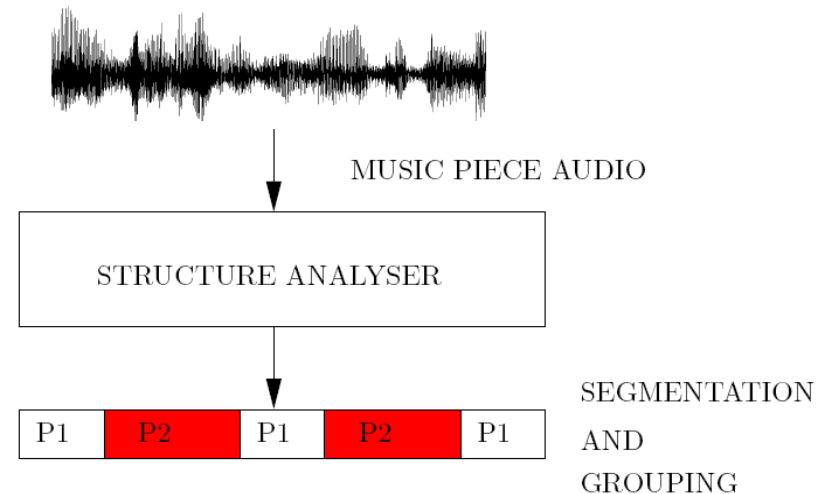
Jouni Paulus

# Music structure analysis

- Segment a music piece in to parts such as "verse", "chorus", "bridge", ...

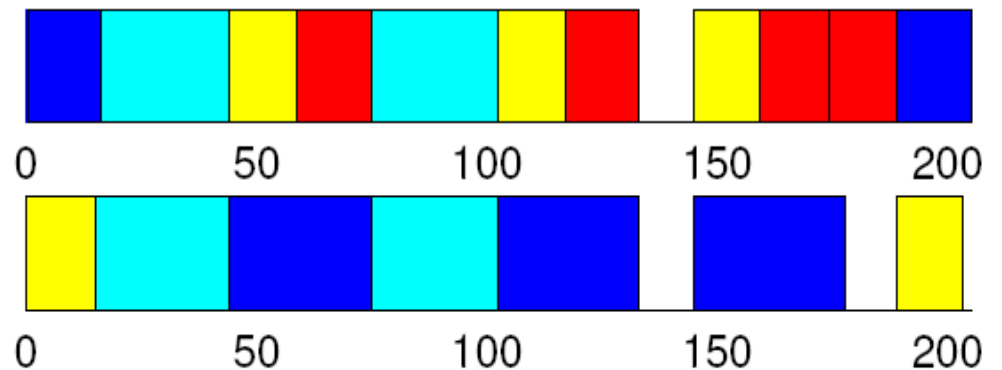
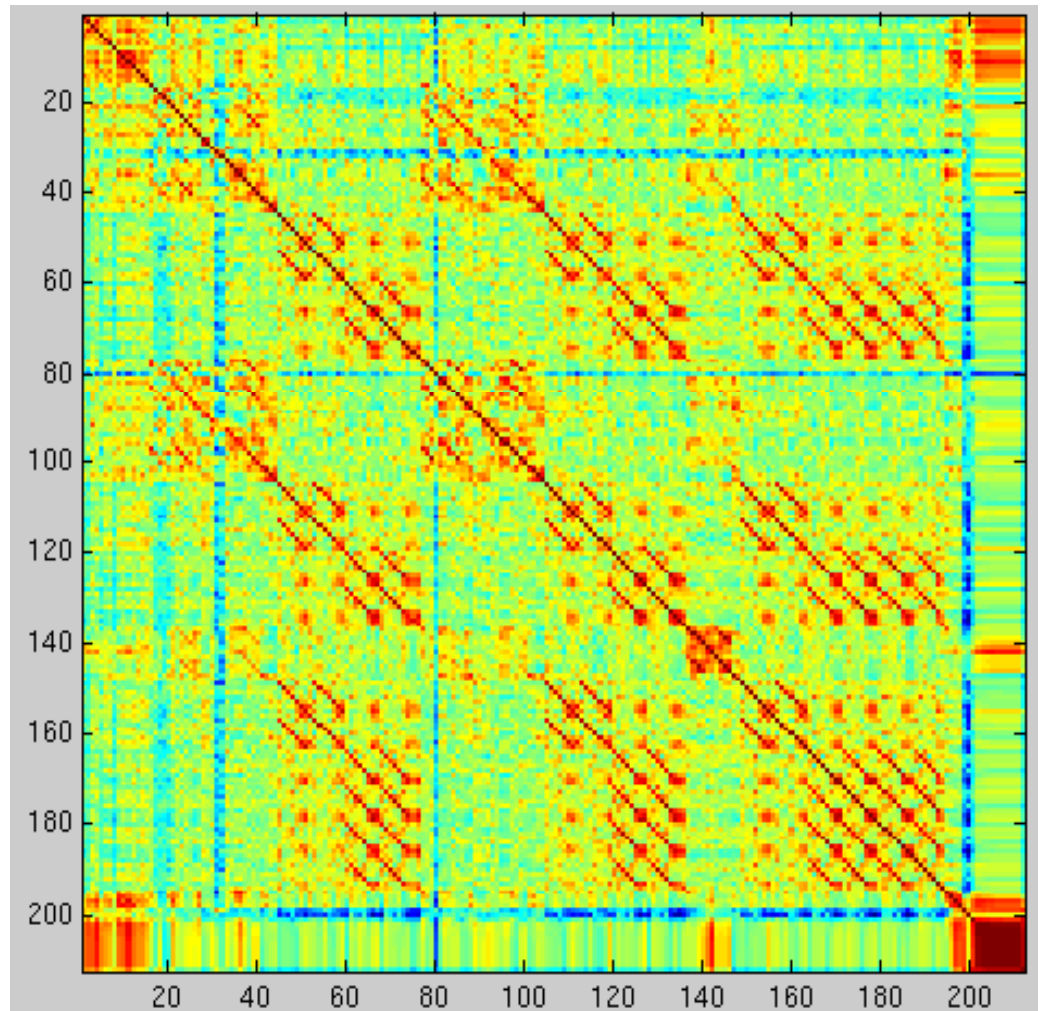
- Approach

1. **Probabilistic fitness function** that defines the probability of structure segmentations, given the observations
2. **Search algorithm** to maximize the fitness



# Music structure analysis

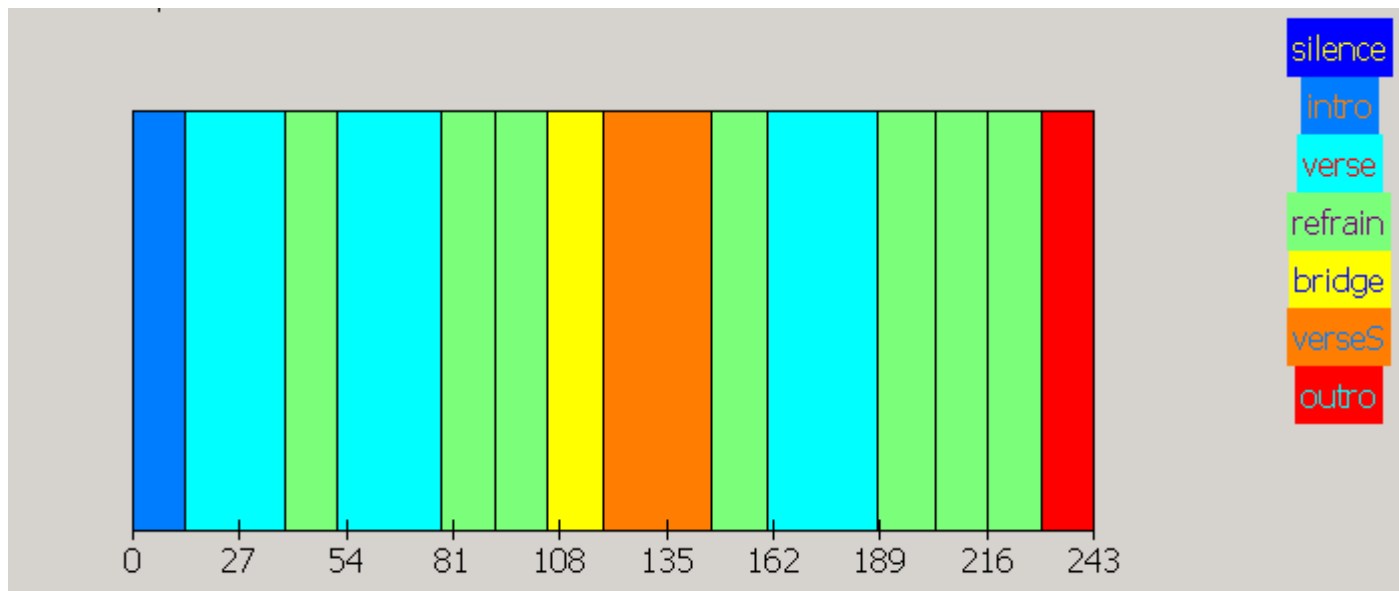
- Repeated parts can be found based on self-similarity
- Example  
Abba: SOS
  - chroma-based similarity matrix
  - reference structure (up) and analysis result (low)
- Probabilistic approach based on fitness function
  - [Paulus-Klapuri-TASLP-2009]



# Music structure analysis

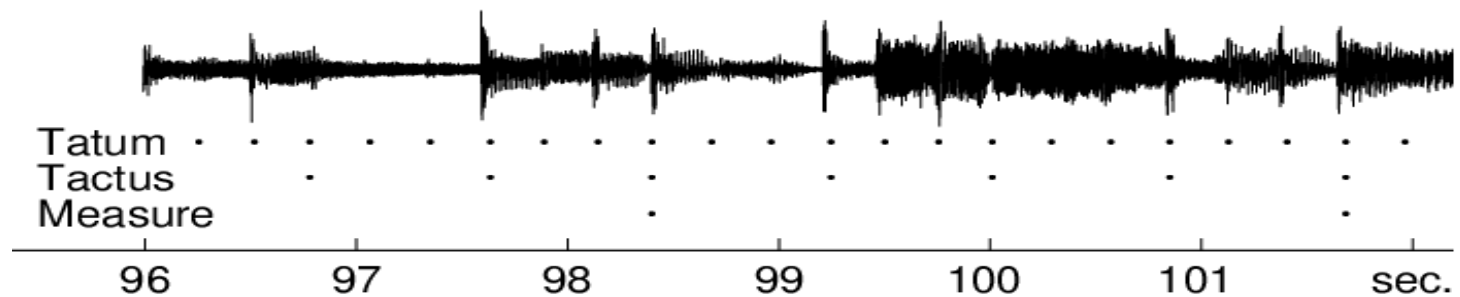
- Structure can be used for intelligent music processing
  - for example parts can be re-organized or copied or deleted
- Demo

Structure of "Let It Be" by The Beatles



# Musical meter analysis

- Meter analysis: tracking the beat, bar lines, and tatum grid



- Demo: switching the first and second half of each bar

– original






shuffled



# Music retrieval and music "matching"




- Widely used "standard" acoustic features
  - Mel-frequency cepstral coefficients (MFCCs) → timbre/instrumentation
  - chroma [Bartsch-2001] → harmonic content
  - rhythmogram (or, fluctuation patterns) [Dixon-2003, Jensen-2007] → rhythm




- MFCC-based retrieval: query      retrieved      mix
-             

- retrieve segment  $t_0$  from song  $i$  that best matches the query:




$$(i, t_0) = \arg \min_{i, t_0} \sum_{t=1}^T (\mathbf{q}_t - \mathbf{s}_{t_0+t}^i)^2$$

$\mathbf{q}_t$  feature vector  $t$  of the query  
 $\mathbf{s}_t^i$  feature vector  $t$  of target song  $i$

- Rhythmogram-based retrieval: query      retrieved      mix
-             

- Chord transcription based retrieval: query      retrieved      mix
-             

# Applications of music matching

- Narrowing down "a perfect piece" by making multiple specific queries
  - enabled when huge amounts of music is available
- DJ: match tempo and musical key
- Mix pieces that match harmonically / rhythmically
- Find piece whose accompaniment matches a given melody
- While playing music, show pieces that locally match the played piece in some respect
- ...
- Sound source separation allows more specific music retrieval (such as searching for similar vocal characteristics)
- Features define similarity
  - BR by Queen  BR by London SO  A Kind of Magic by Queen 

# Summary

- Some ways of extracting semantically meaningful parts of music signals were introduced, and applications were discussed
- There are more problems and methods than fit into a single talk

Thanks!