

# Music Information Retrieval for Jazz

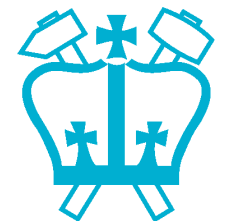
Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio  
Dept. Electrical Eng., Columbia Univ., NY USA

{dpwe,thierry}@ee.columbia.edu

<http://labrosa.ee.columbia.edu/>

1. Music Information Retrieval
2. Automatic Tagging
3. Musical Content
4. Future Work



COLUMBIA UNIVERSITY  
IN THE CITY OF NEW YORK

# Machine Listening

- Extracting **useful information** from sound
  - ... like (we) animals do

|      |          |                       |         |                      |
|------|----------|-----------------------|---------|----------------------|
| Task | Describe | Automatic Narration   | Emotion | Music Recommendation |
|      | Classify | Environment Awareness | ASR     | Music Transcription  |
|      | Detect   | “Sound Intelligence”  | VAD     | Speech/Music         |
|      |          | Environmental Sound   | Speech  | Music                |
|      |          |                       |         | <i>Domain</i>        |

# I. The Problem

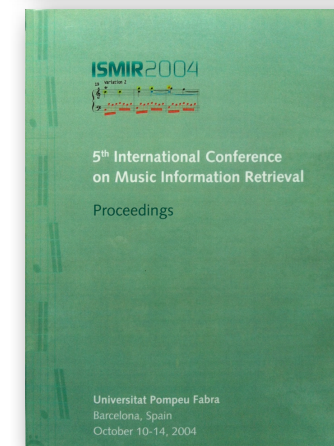
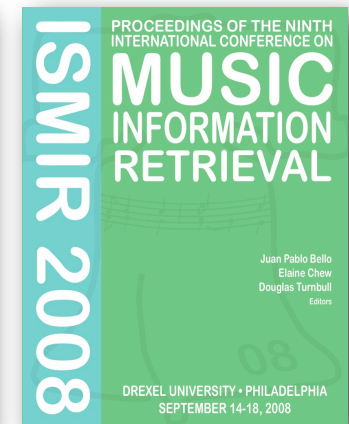
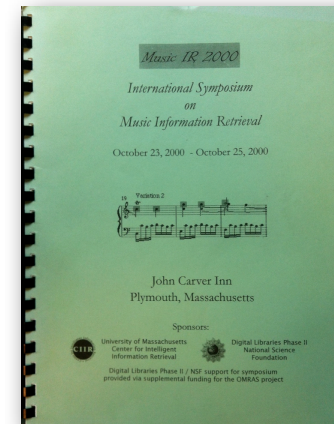
- We have a lot of music Can computers help?



- Applications
  - archive organization
  - music recommendation
  - musicological insight?

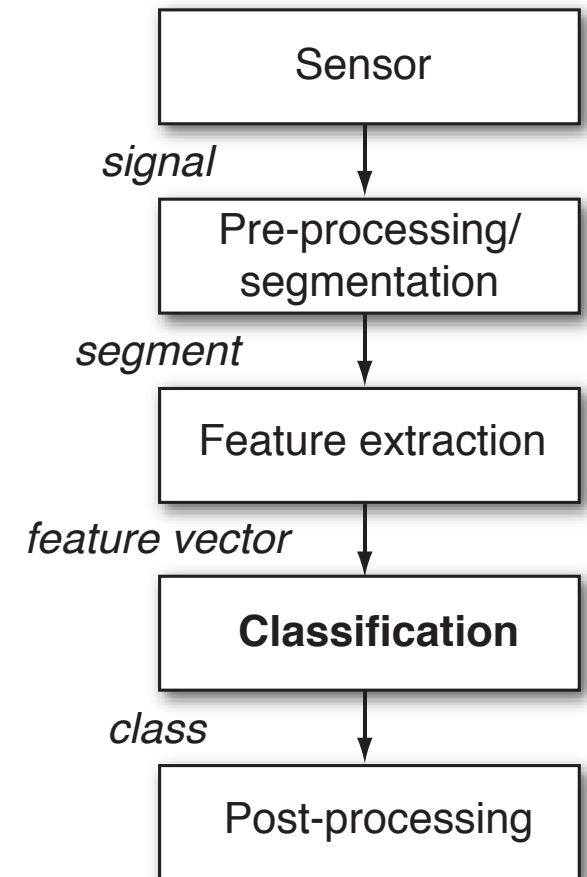
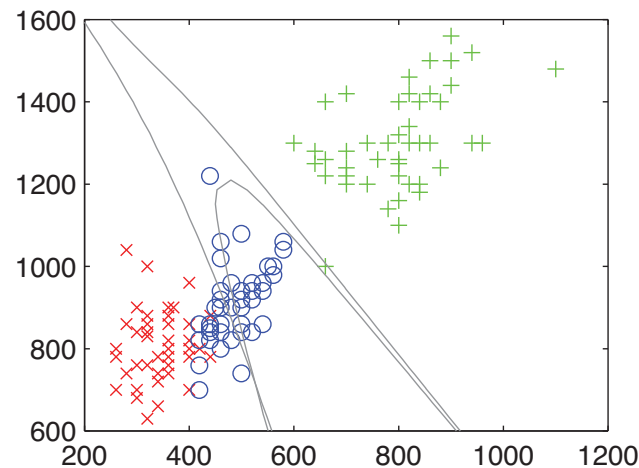
# Music Information Retrieval (MIR)

- Small field that has grown since ~2000
  - musicologists, engineers, librarians
  - significant commercial interest
- MIR as musical analog of text IR
  - find stuff in large archives
- Popular tasks
  - genre classification
  - chord, melody, full transcription
  - music recommendation
- Annual evaluations
  - “Standard” test corpora - pop



# 2. Automatic Tagging

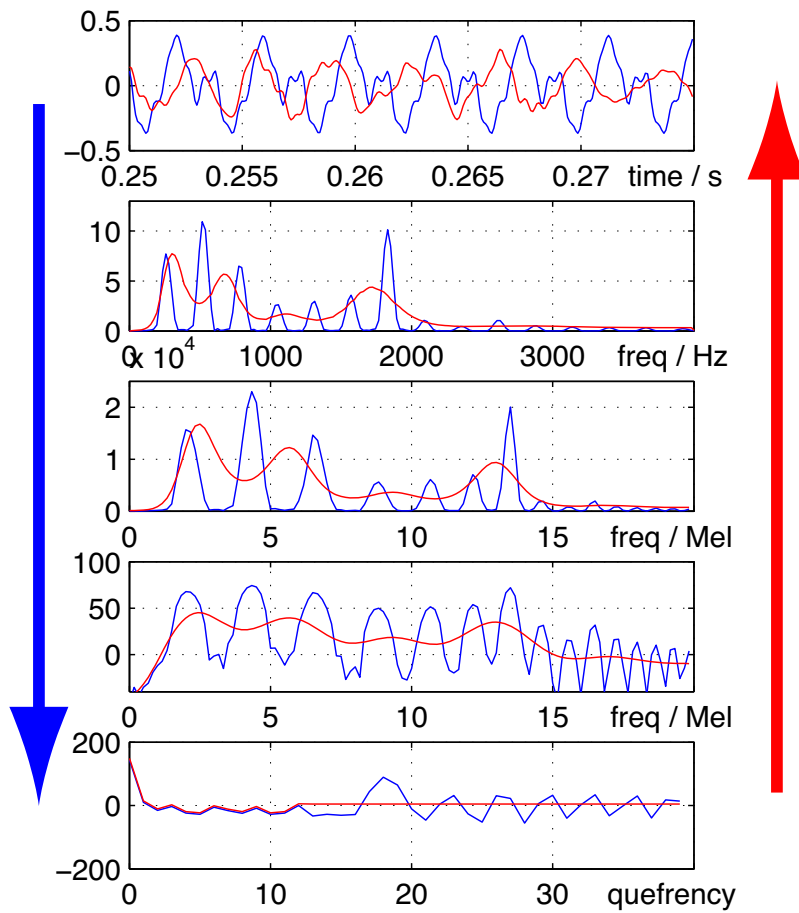
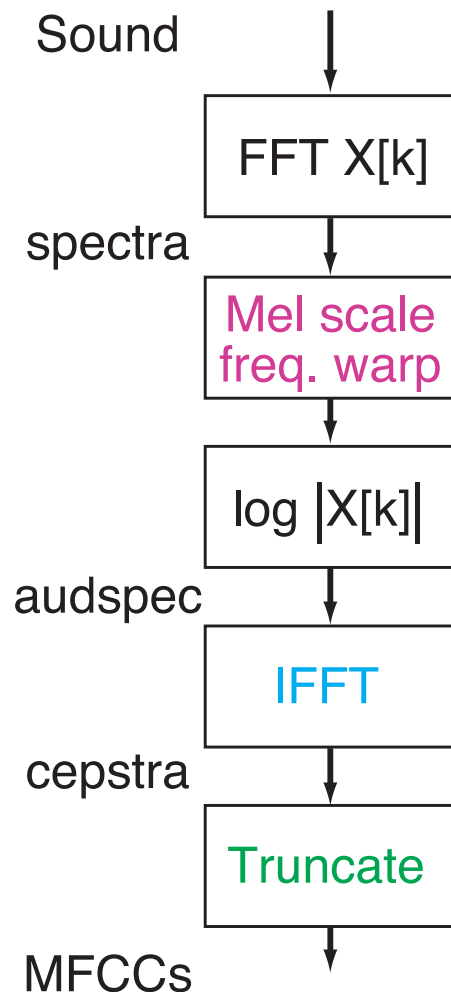
- **Statistical Pattern Recognition:**  
Finding matches  
to training examples



- **Need:**
  - Feature design
  - Labeled training examples
- **Applications**
  - Genre
  - Instrumentation
  - Artist
  - Studio ...

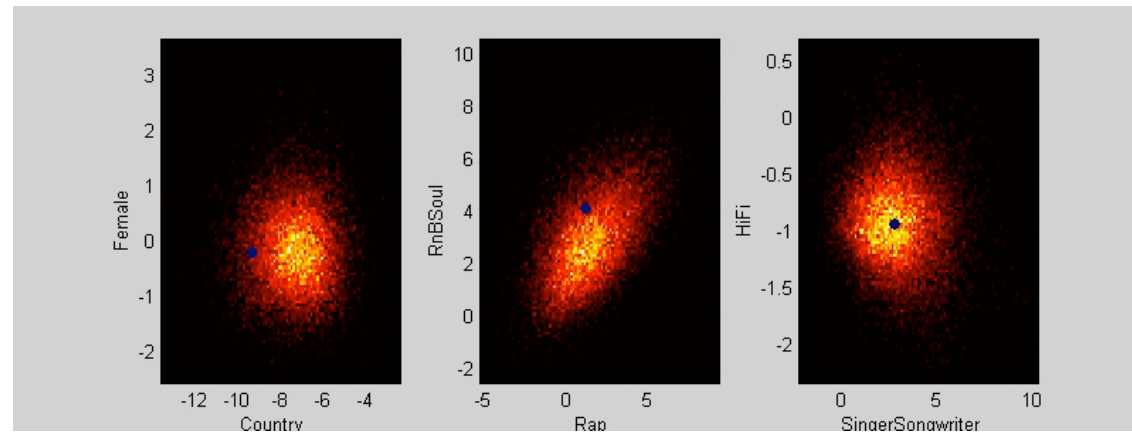
# Features: MFCC

- Mel-Frequency Cepstral Coefficients
  - the standard features from speech recognition

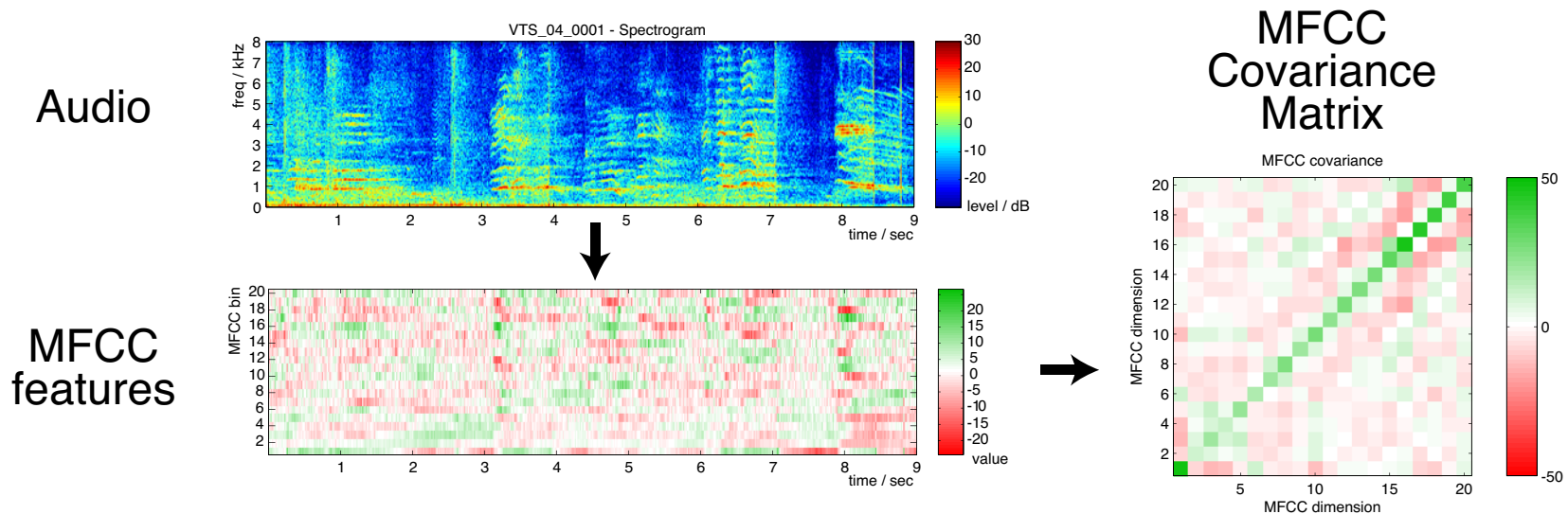


# Representing Audio

- MFCCs are short-time features (25 ms)
- Sound is a “trajectory” in MFCC space

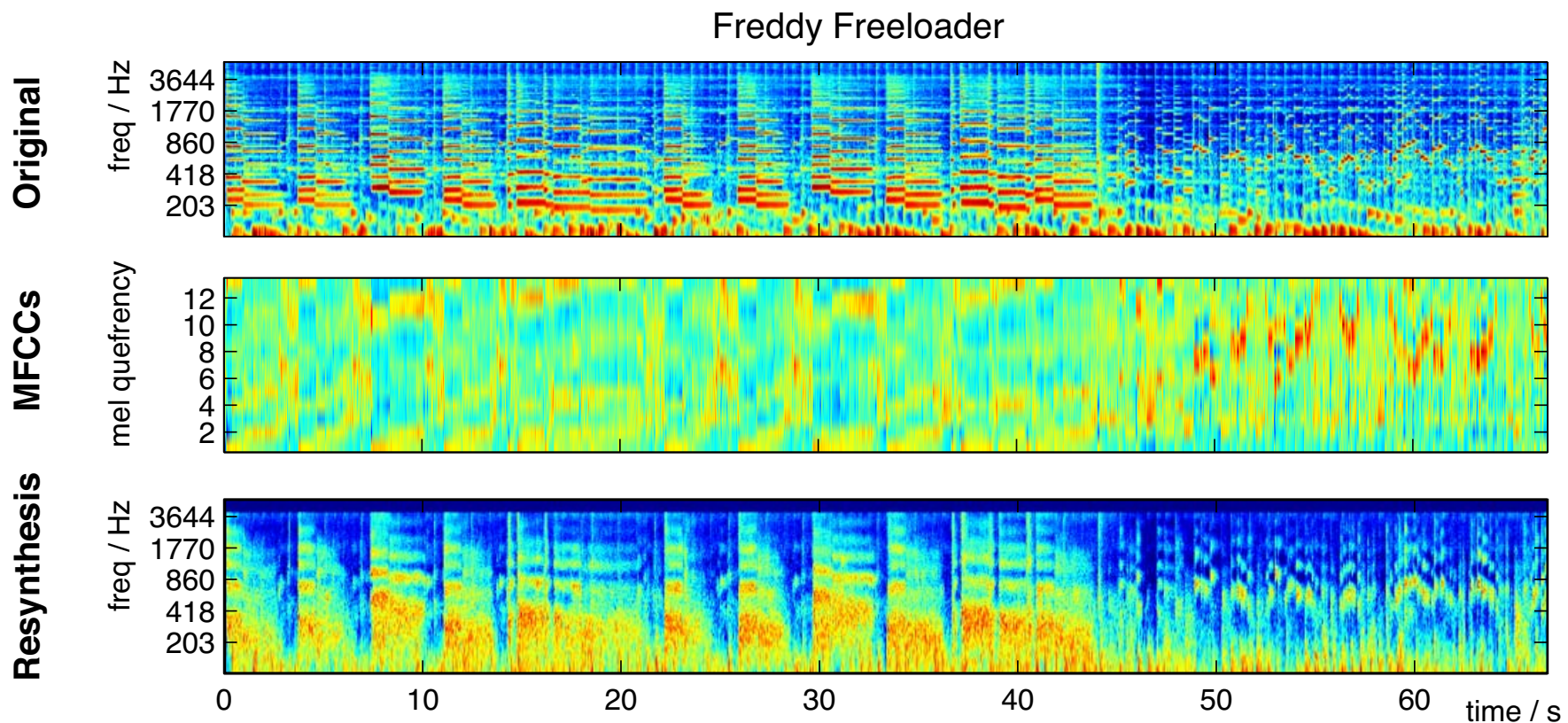


- Describe whole track by its statistics



# MFCCs for Music

- Can resynthesize MFCCs by shaping noise
  - gives an idea about the information retained





# Ground Truth

Mandel & Ellis '08

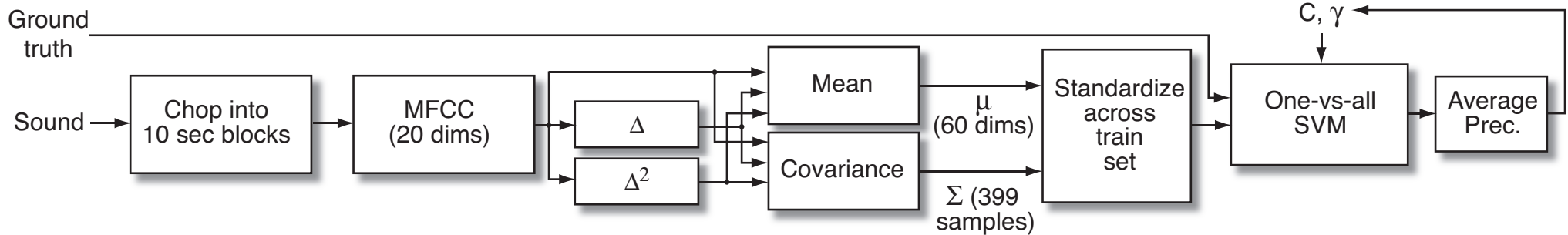
- **Major Miner**: Free-text tags for 10s clips
  - 400 users, 7500 unique tags, 70,000 taggings



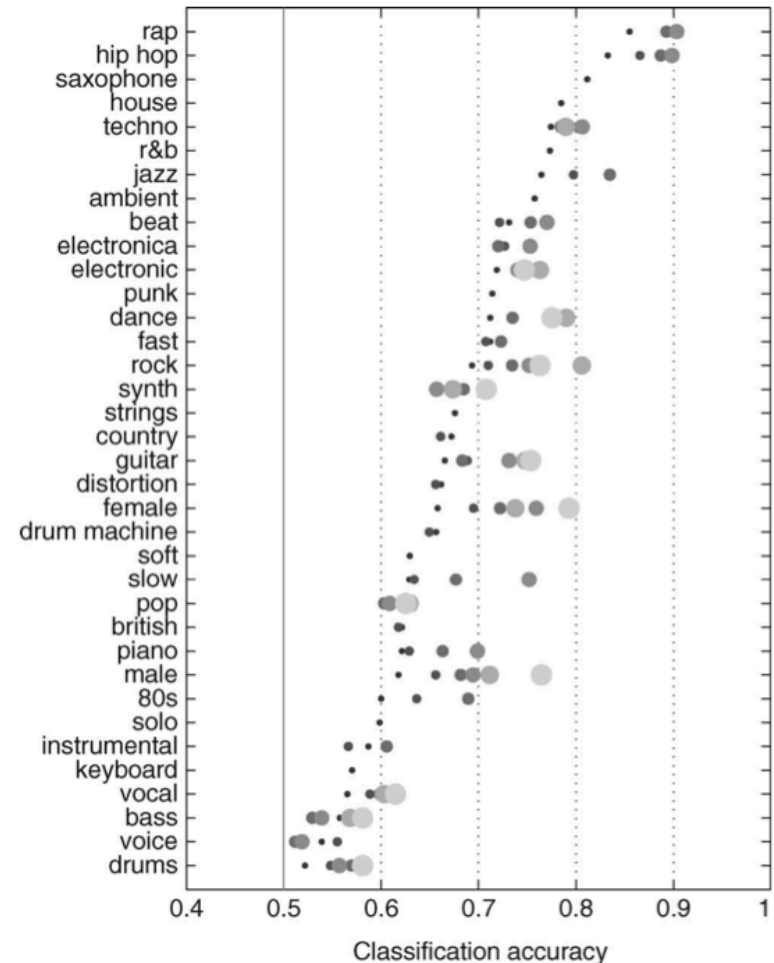
- **Example**: drum, bass, piano, jazz, slow, instrumental, saxophone, soft, quiet, club, ballad, smooth, soulful, easy\_listening, swing, improvisation, 60s, cool, light



# Classification



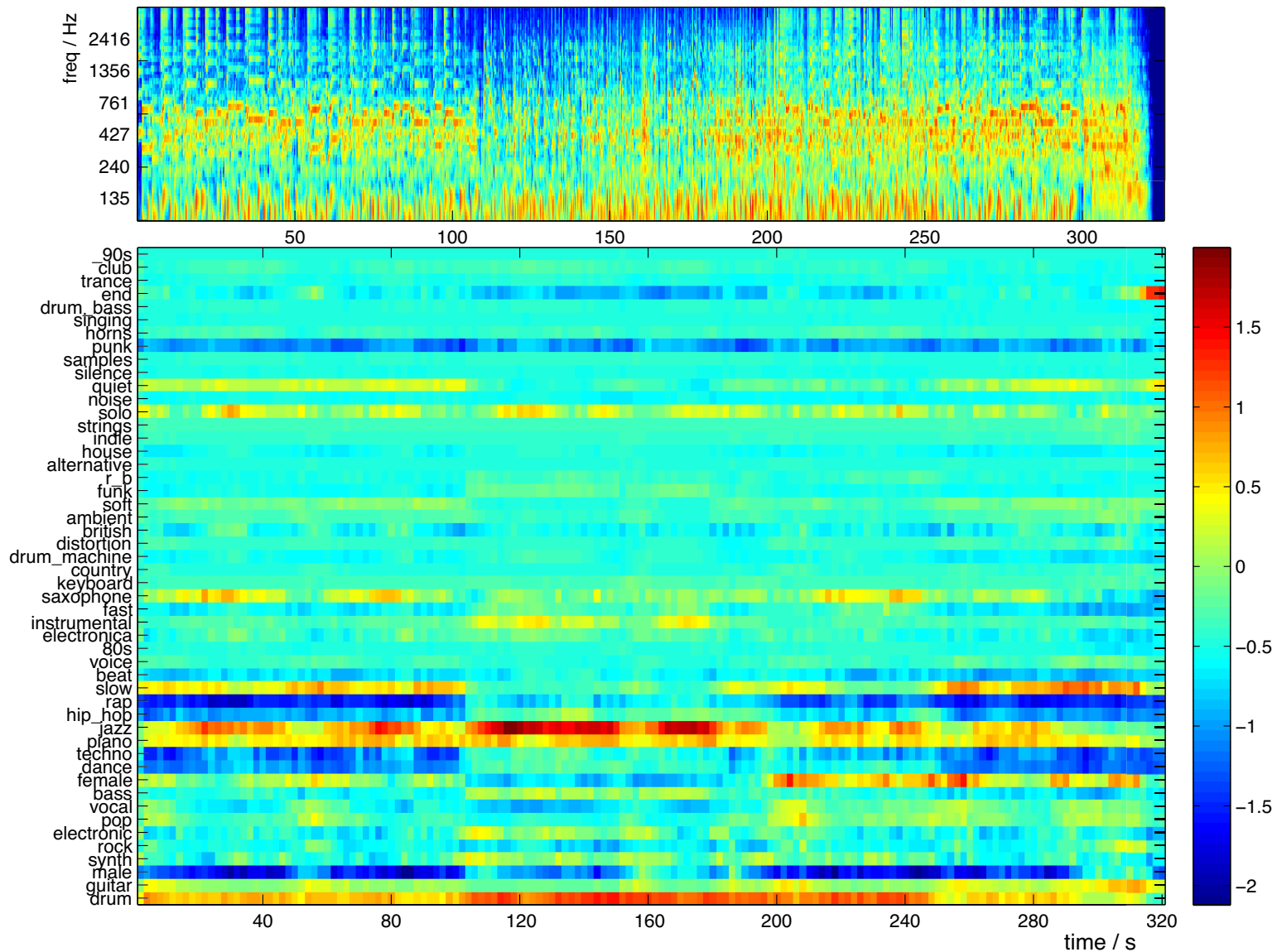
- MFCC features
- + human ground truth
- + standard machine learning tools



# Classification Results

- Classifiers trained from top 50 tags

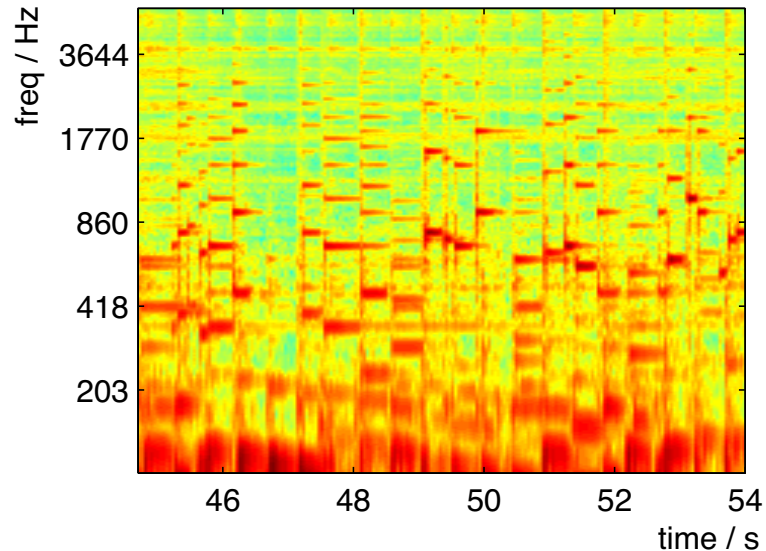
01 Soul Eyes



# 3. Musical Content

- MFCCs (and speech recognizers) don't respect pitch

- pitch is important
- visible in spectrogram



- Pitch-related tasks

- note transcription
- chord transcription
- matching by musical content (“cover songs”)

# Note Transcription

Poliner & Ellis  
'05,'06,'07

## Training data and features:

- MIDI, multi-track recordings, playback piano, & resampled audio (less than 28 mins of train audio).
- Normalized magnitude STFT.



## Classification:

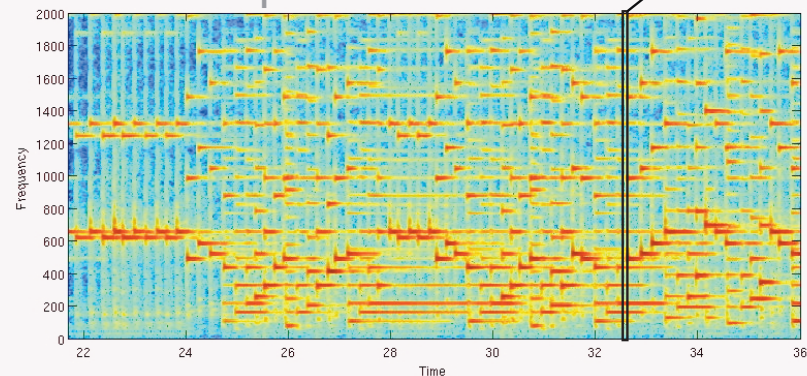
- N-binary SVMs (one for ea. note).
- Independent frame-level classification on 10 ms grid.
- Dist. to class body as posterior.



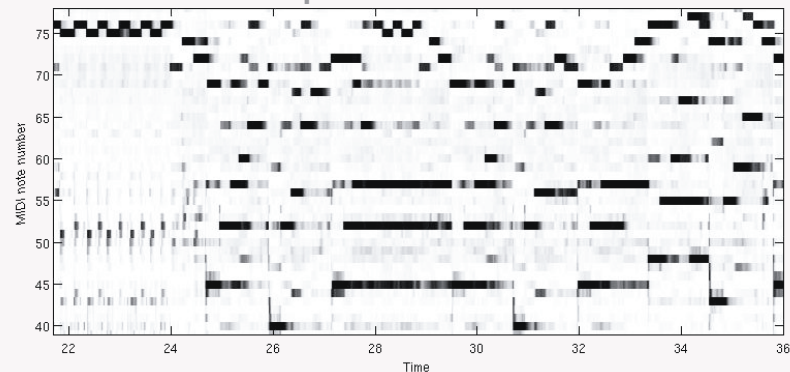
## Temporal Smoothing:

- Two state (on/off) independent HMM for ea. note. Parameters learned from training data.
- Find Viterbi sequence for ea. note.

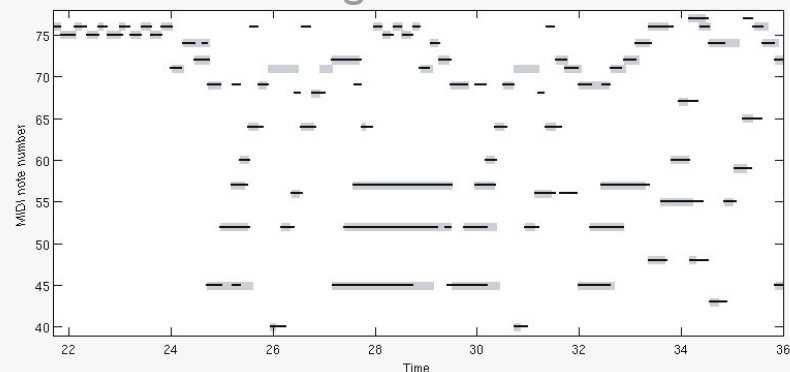
feature representation



classification posteriors



hmm smoothing



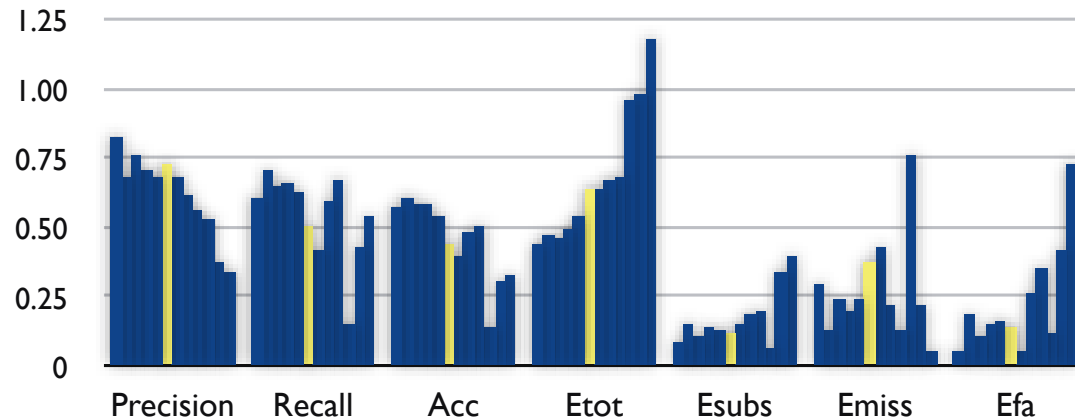
# Polyphonic Transcription

MIREX 2007

- Real music excerpts + ground truth

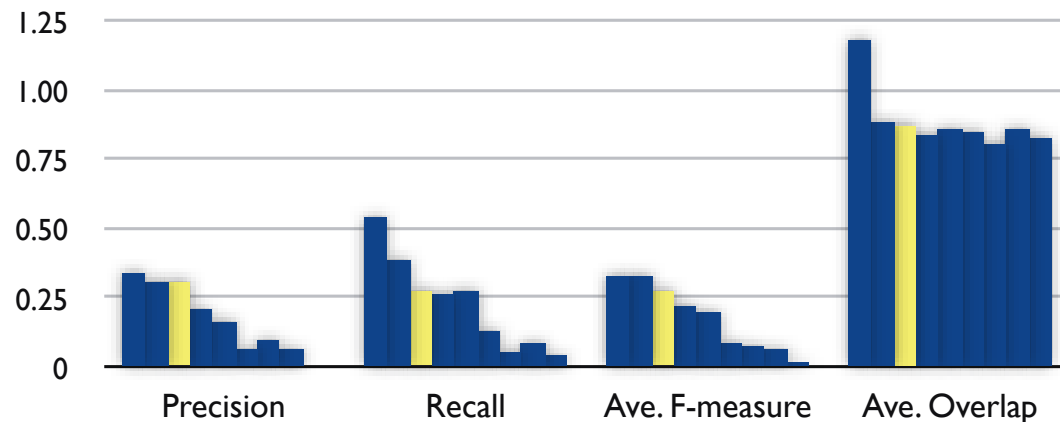
## Frame-level transcription

Estimate the fundamental frequency of all notes present on a 10 ms grid



## Note-level transcription

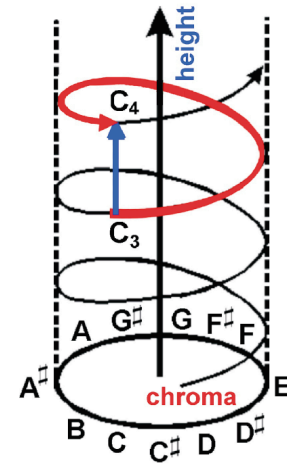
Group frame-level predictions into note-level transcriptions by estimating onset/offset



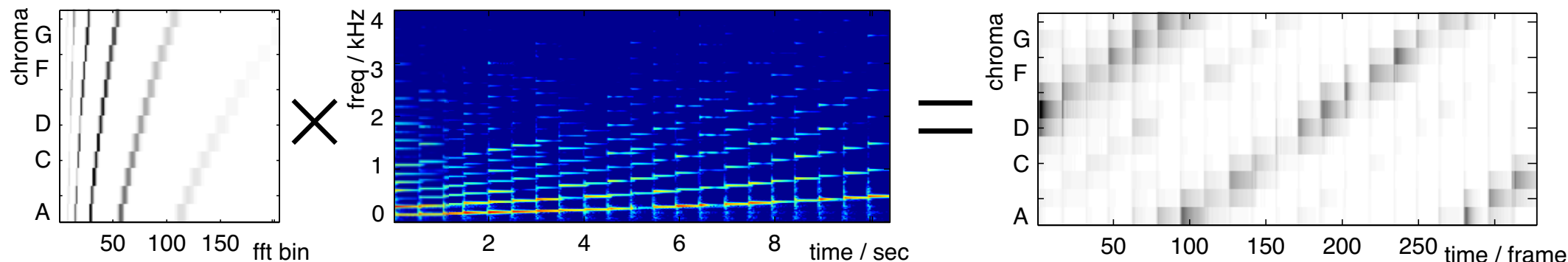
# Chroma Features

Fujishima 1999

- Idea: Project onto **12 semitones** regardless of **octave**
  - maintains main “musical” distinction
  - **invariant** to musical equivalence
  - no need to worry about **harmonics**?



Warren et al. 2003



$$C(b) = \sum_{k=0}^{N_M} B(12 \log_2(k/k_0) - b) W(k) |X[k]|$$

- $W(k)$  is weighting,  $B(b)$  selects every  $\sim \text{mod } 12$

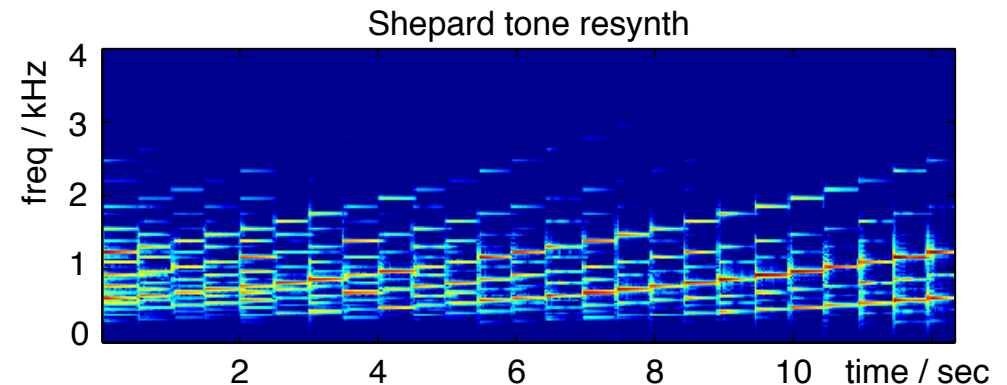
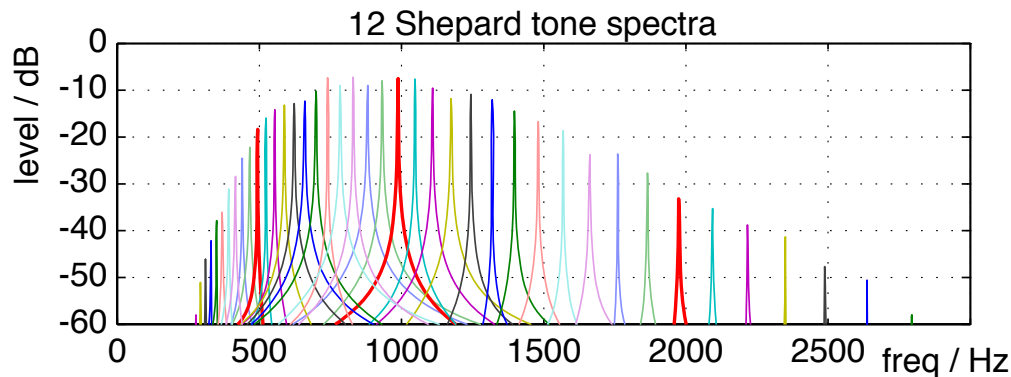


# Chroma Resynthesis

Ellis & Poliner 2007

- Chroma describes the notes in an octave
  - ... but not the octave
- Can **resynthesize** by presenting **all octaves**
  - ... with a smooth envelope
  - “Shepard tones” - octave is ambiguous

$$y_b(t) = \sum_{o=1}^M W\left(o + \frac{b}{12}\right) \cos 2^{o + \frac{b}{12}} w_0 t$$



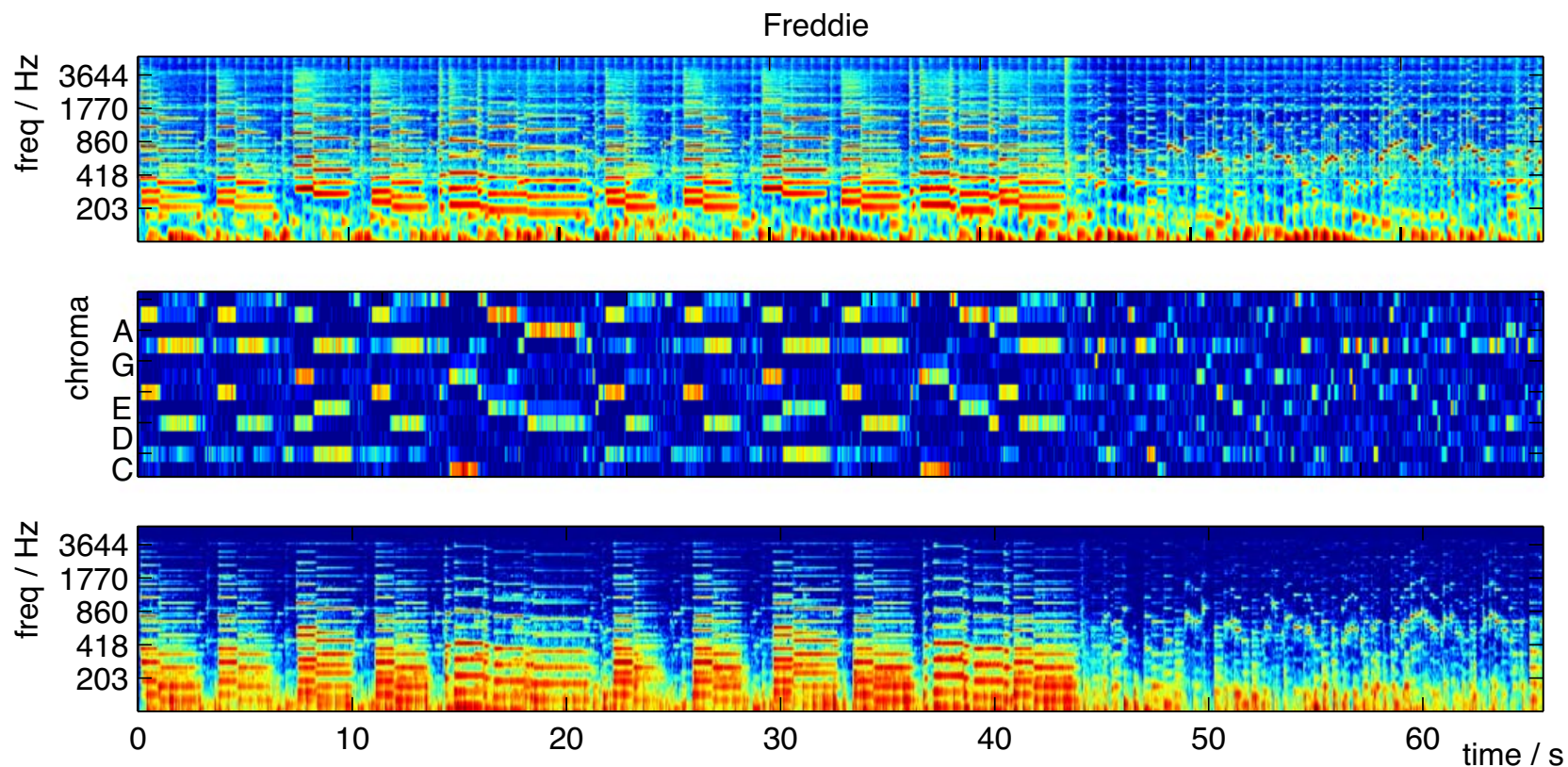
- endless sequence illusion





# Chroma Example

- Simple **Shepard tone** resynthesis
  - can also reimpose **broad spectrum** from MFCCs



# Onset detection

Bello et al. 2005

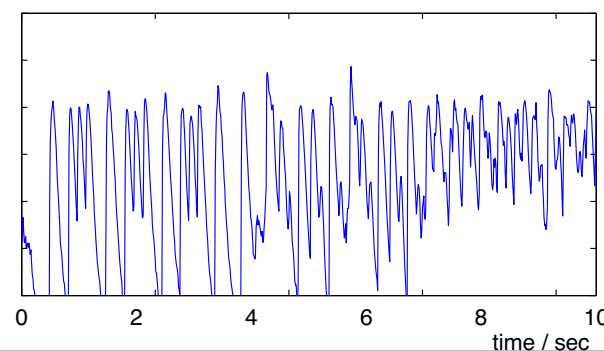
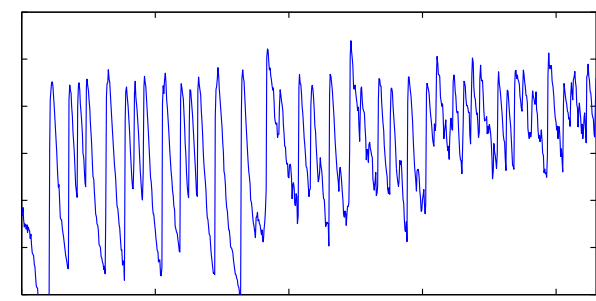
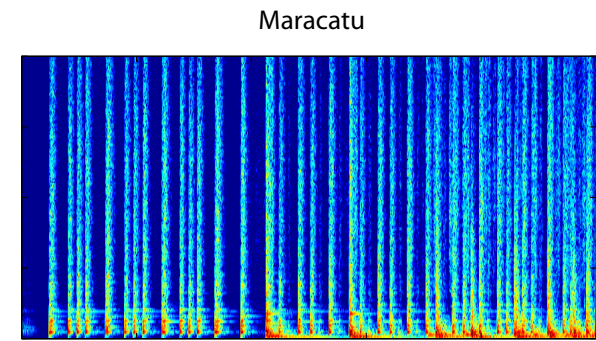
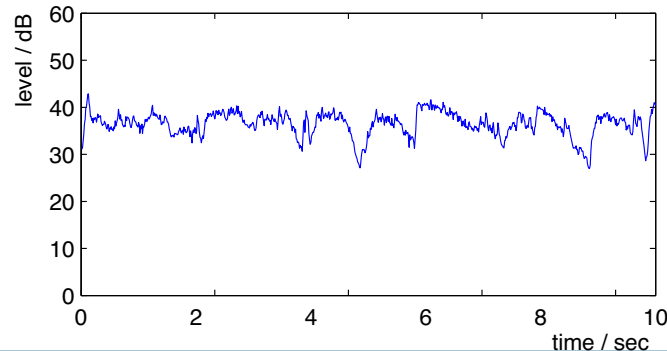
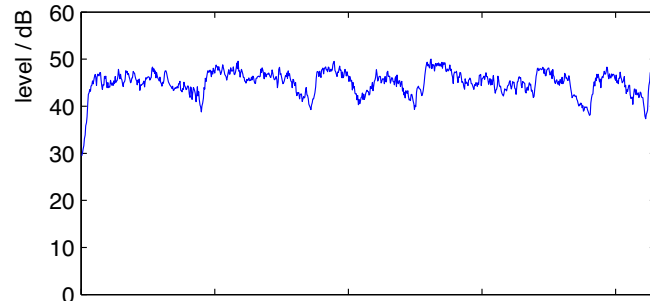
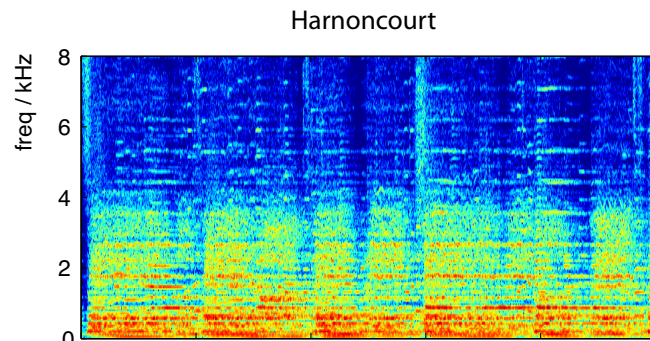
- Simplest thing is **energy envelope**

$$e(n_0) = \sum_{n=-W/2}^{W/2} w[n] |x(n + n_0)|^2$$

- emphasis on high frequencies?

$$\sum_f |X(f, t)|$$

$$\sum_f f \cdot |X(f, t)|$$

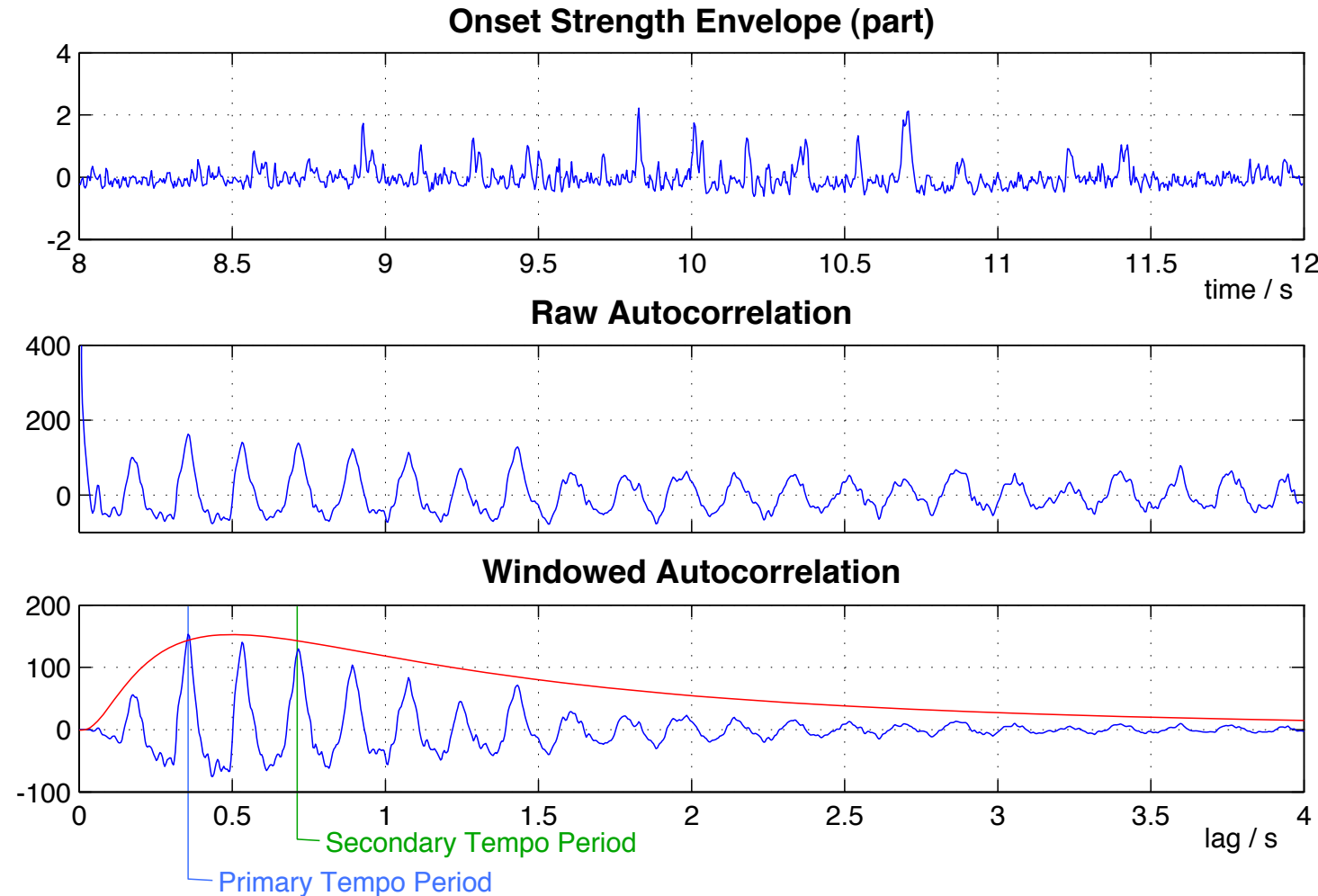


# Tempo Estimation

- Beat tracking (may) need global tempo period  $\tau$ 
  - otherwise problem is not “optimal substructure”

- Pick peak in onset envelope autocorrelation

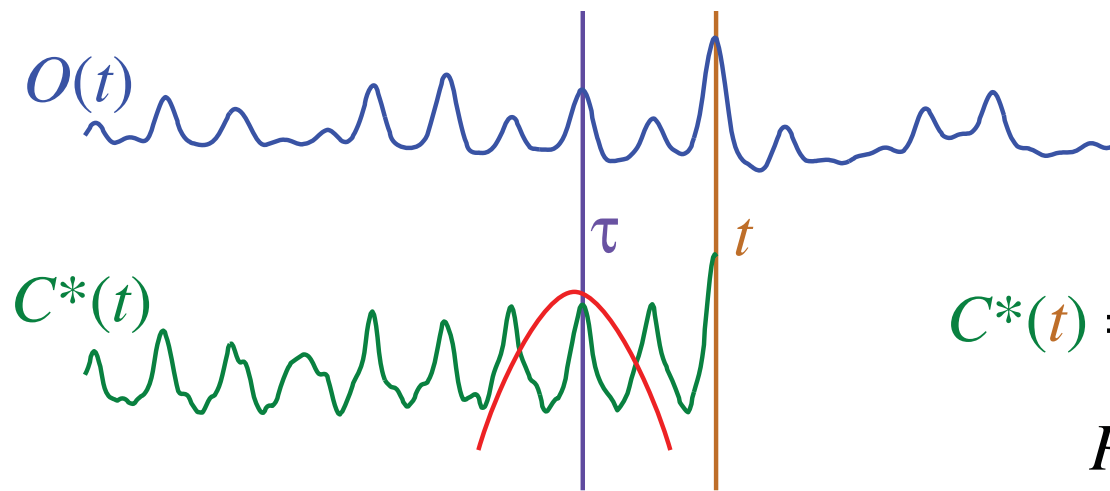
- after applying “human preference” window
- check for subbeat



# Beat Tracking by Dynamic Programming

- To optimize  $C(\{t_i\}) = \sum_{i=1}^N O(t_i) + \alpha \sum_{i=2}^N F(t_i - t_{i-1}, \tau_p)$

- define  $C^*(t)$  as best score up to time  $t$
- then build up recursively (with traceback  $P(t)$ )



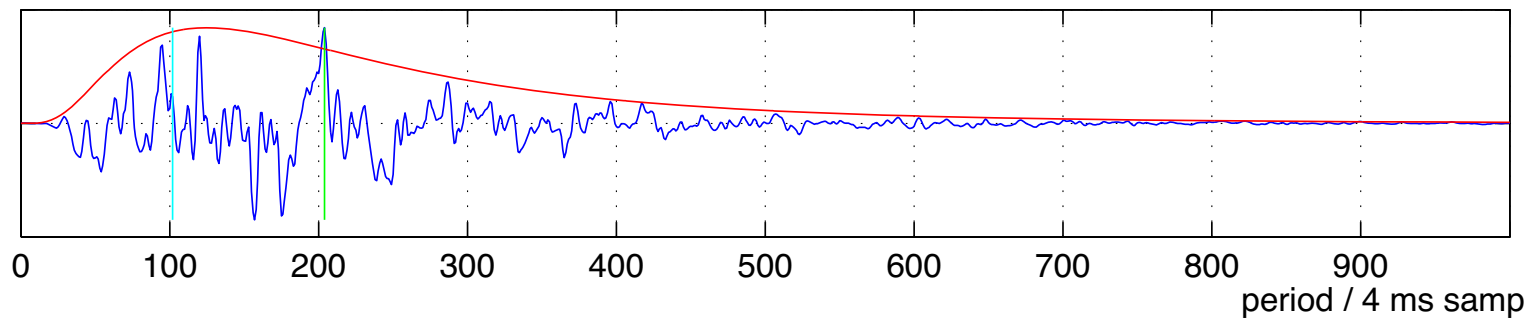
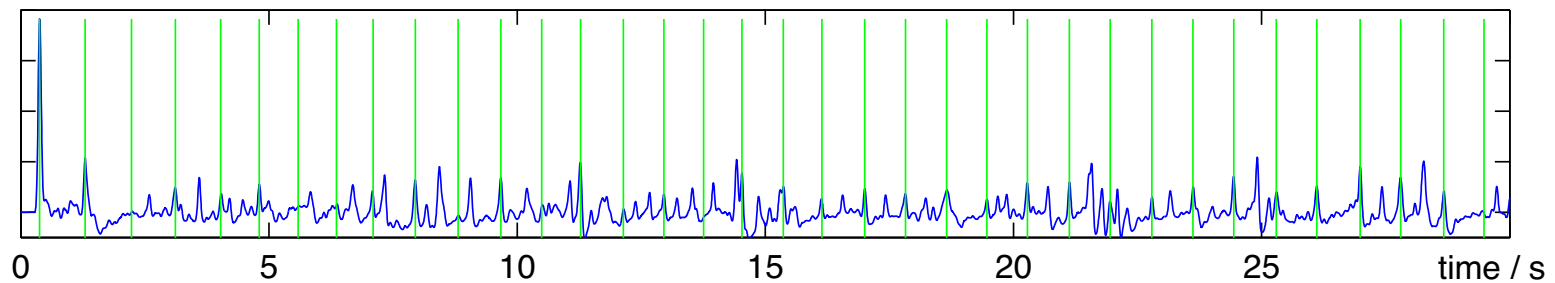
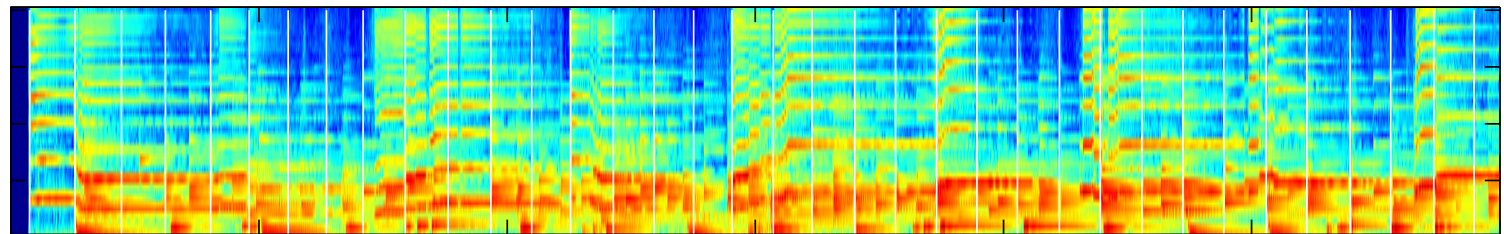
$$C^*(t) = O(t) + \max_{\tau} \{ \alpha F(t - \tau, \tau_p) + C^*(\tau) \}$$
$$P(t) = \operatorname{argmax}_{\tau} \{ \alpha F(t - \tau, \tau_p) + C^*(\tau) \}$$

- final beat sequence  $\{t_i\}$  is best  $C^*$  + back-trace

# Beat Tracking Results

- Prefers drums & steady tempo

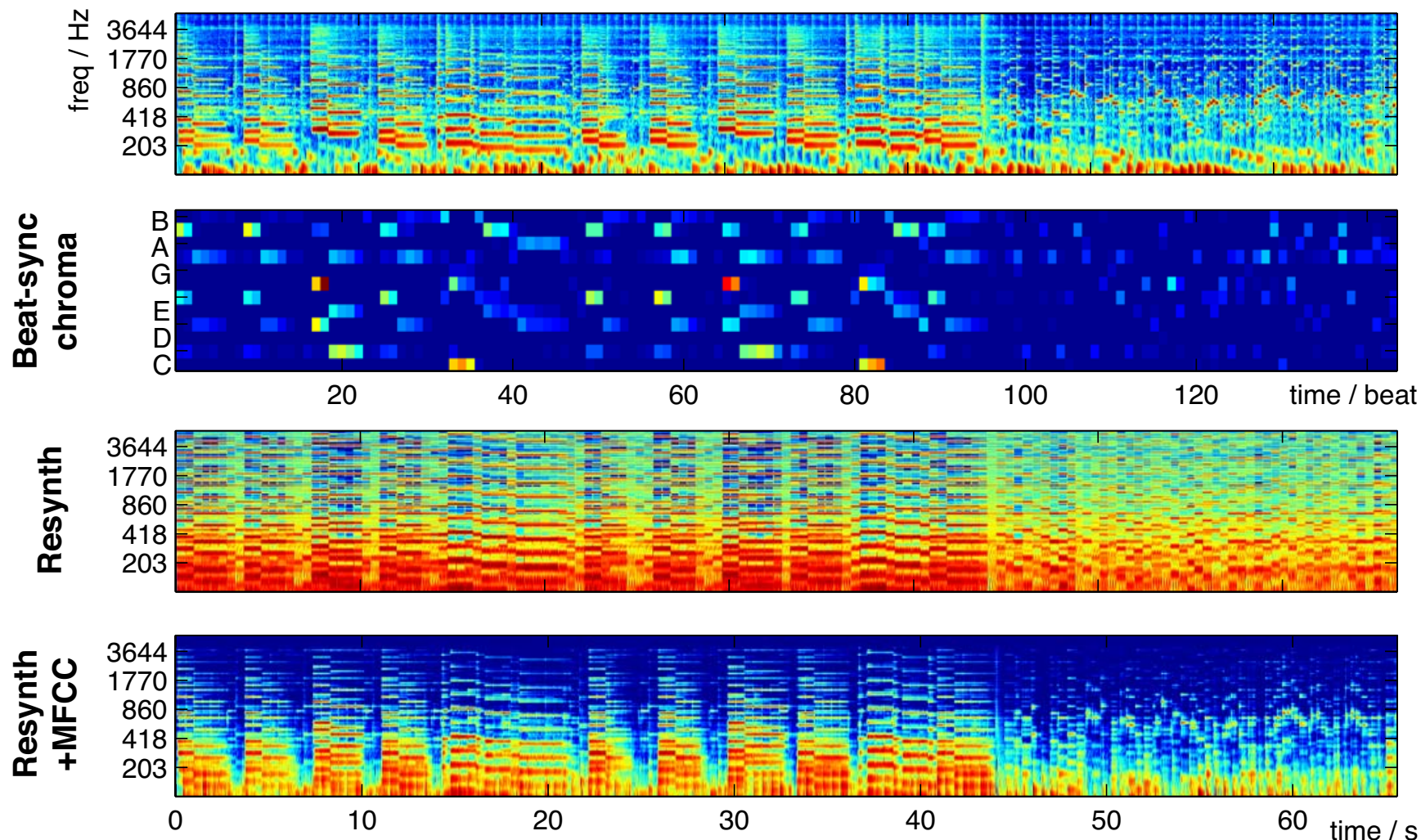
Soul Eyes



# Beat-Synchronous Chroma

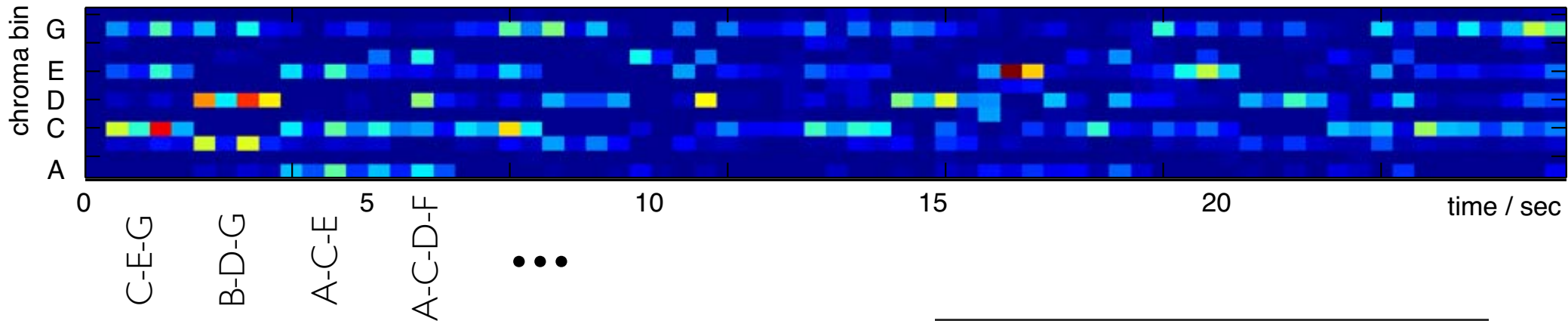
- Record one chroma vector per beat
  - compact representation of harmonies

Freddy



# Chord Recognition

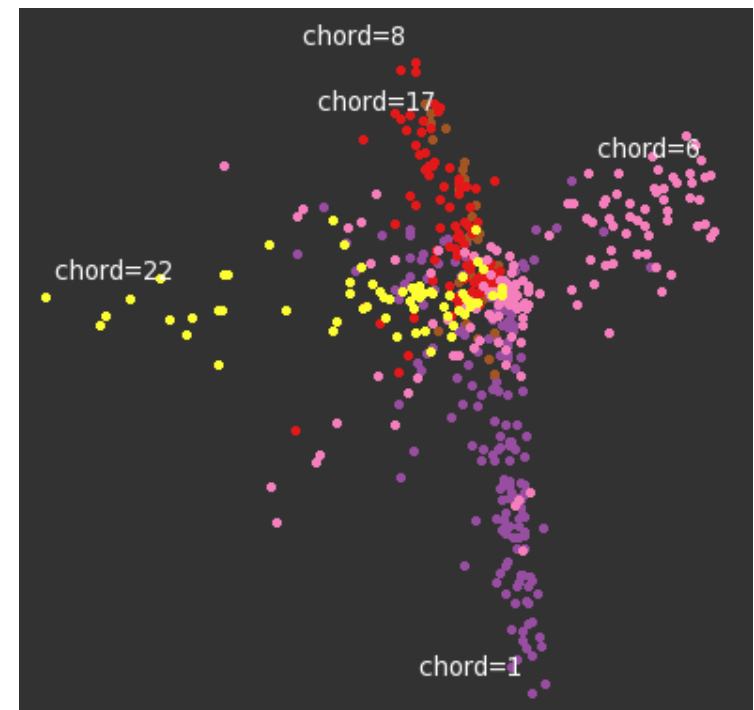
- Beat synchronous chroma look like **chords**



- can we transcribe them?

- **Two approaches**

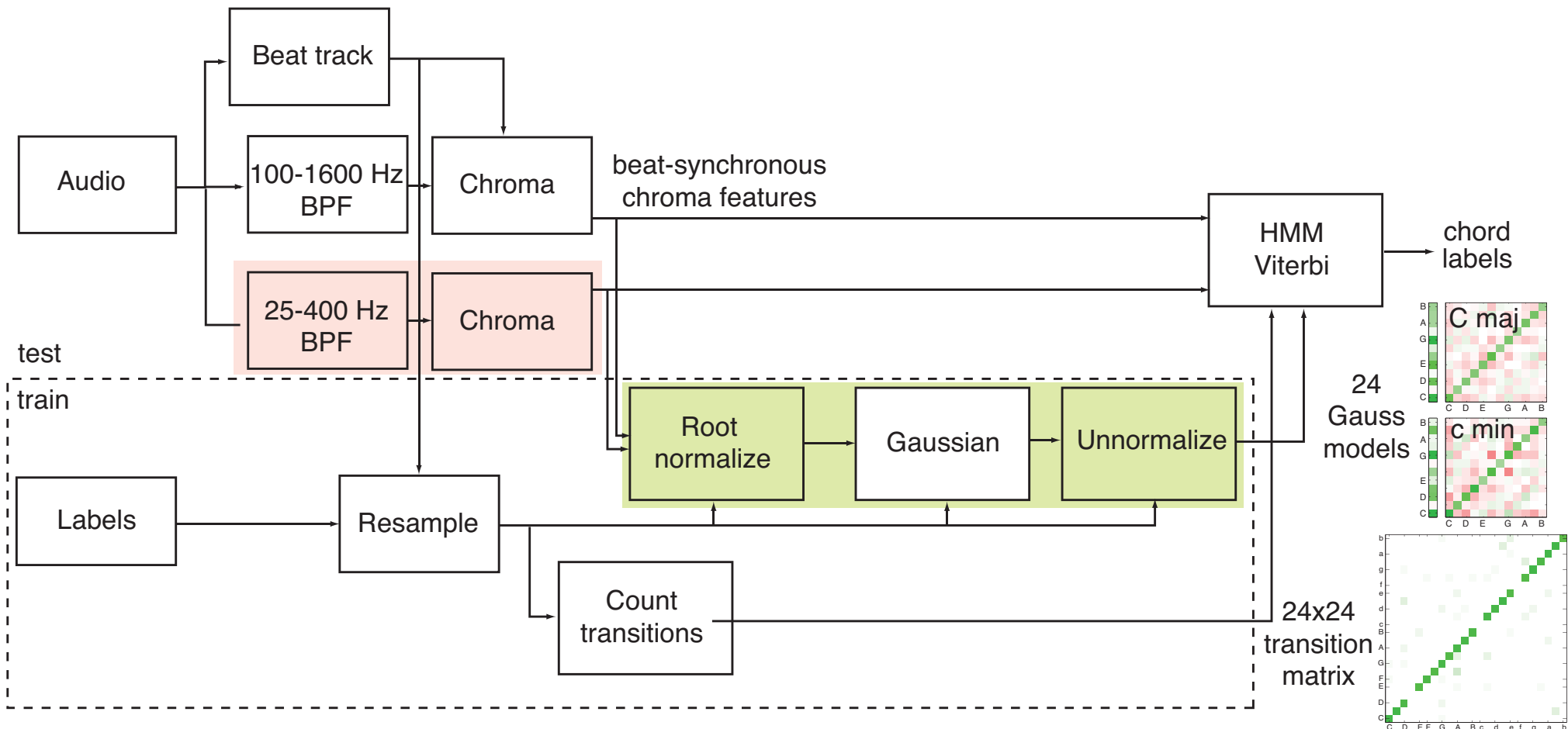
- **manual templates**  
(prior knowledge)
- **learned models**  
(from training data)



# Chord Recognition System

Sheh & Ellis 2003

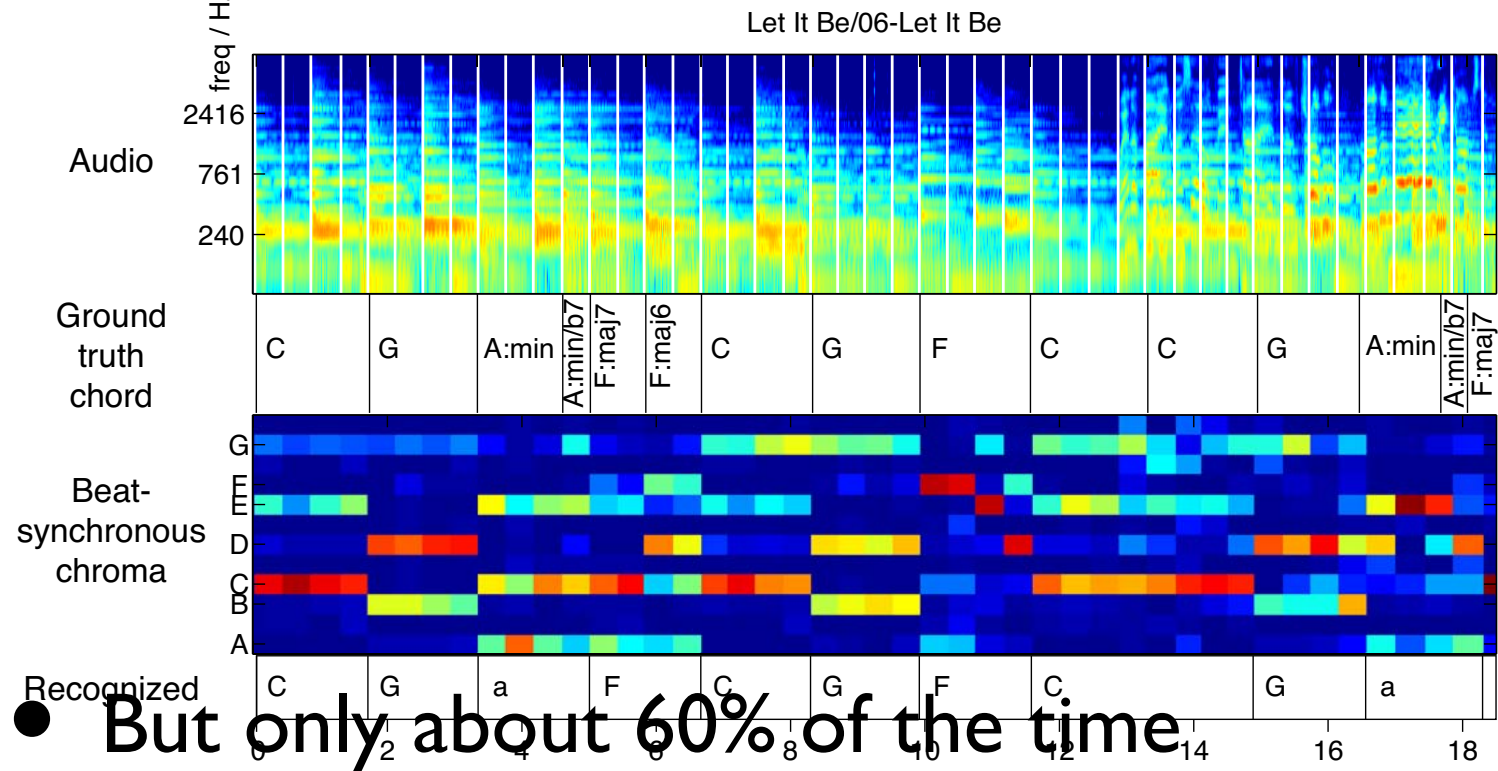
- Analogous to **speech recognition**
  - **Gaussian models** of features for each chord
  - **Hidden Markov Models** for chord transitions





# Chord Recognition

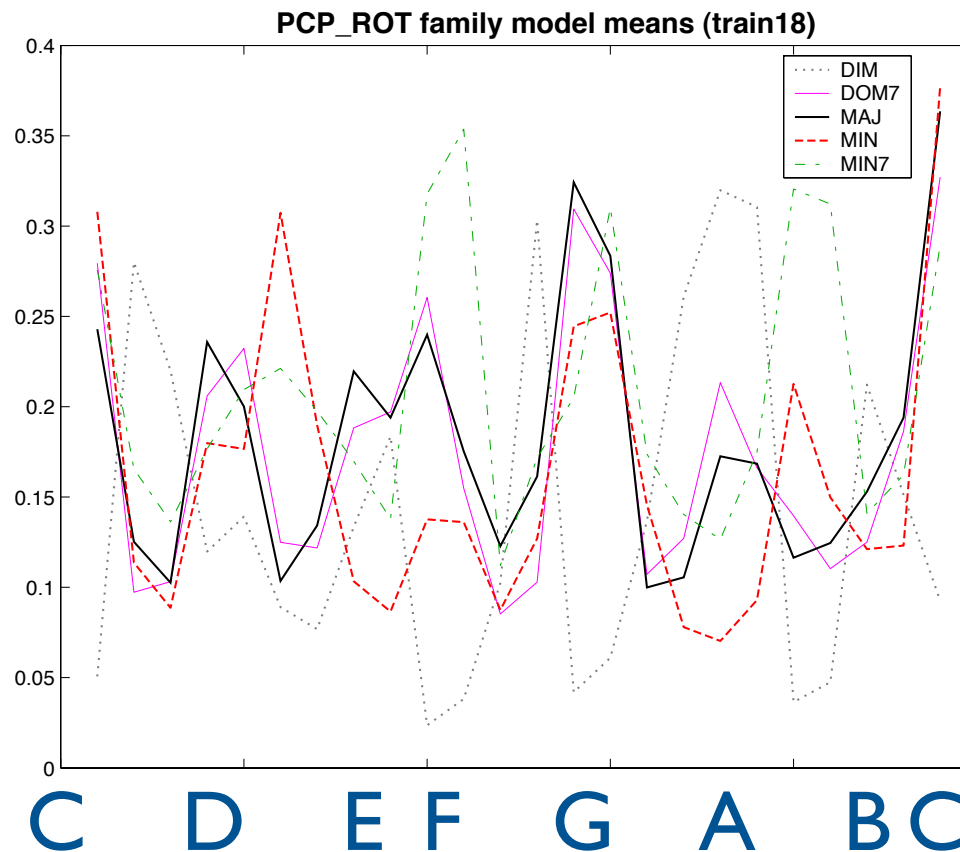
- Often works:



|               | 12 chroma | +bass |
|---------------|-----------|-------|
| indep. models | 0.539     | 0.552 |
| pooled models | 0.556     | 0.578 |

# What did the models learn?

- Chord model centers (**means**) indicate chord **'templates'**:

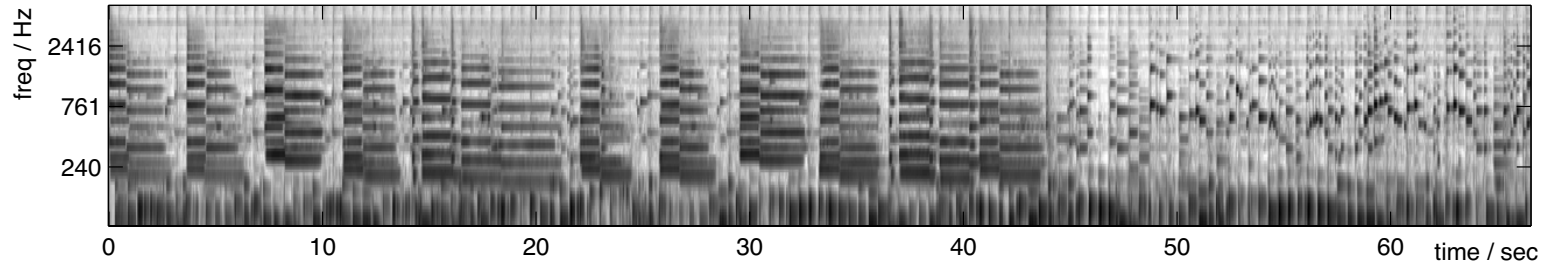


(for C-root chords)

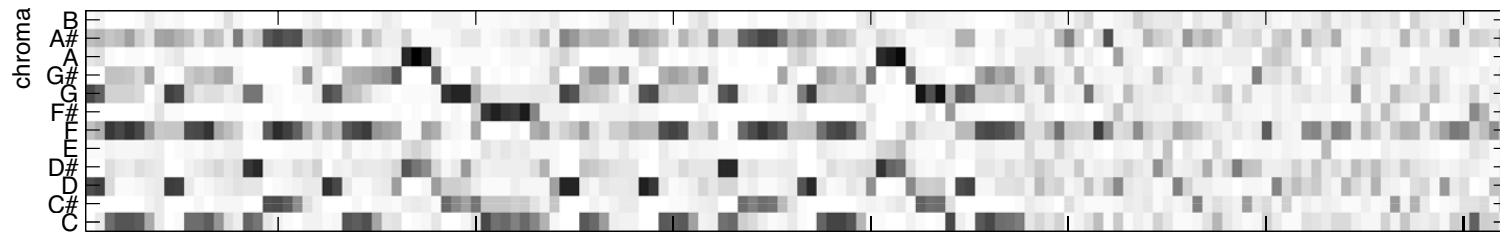
# Chords for Jazz

- *How many types?*

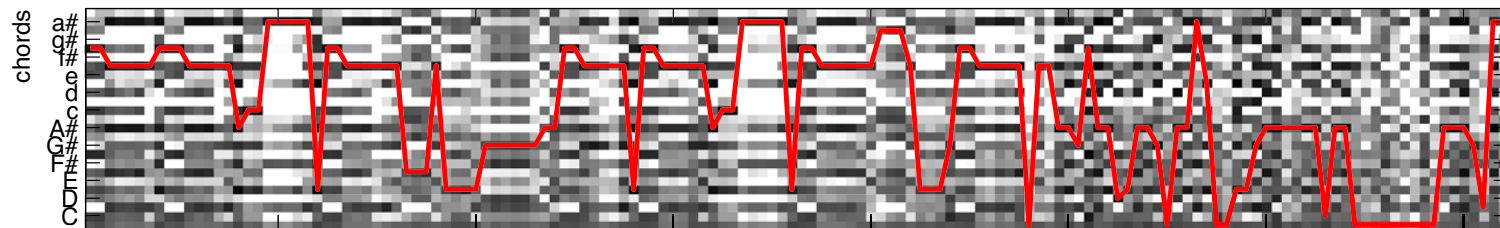
Freddy – logf sgram



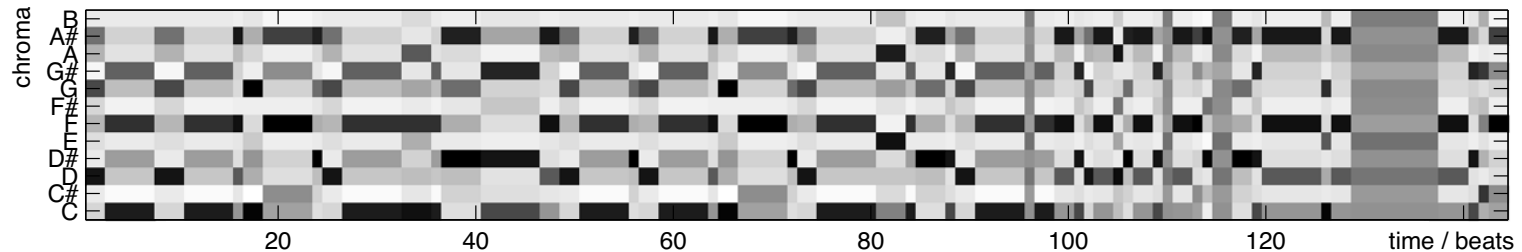
Freddy – beat-sync chroma



Freddy – chord likelihoods + Viterbi path

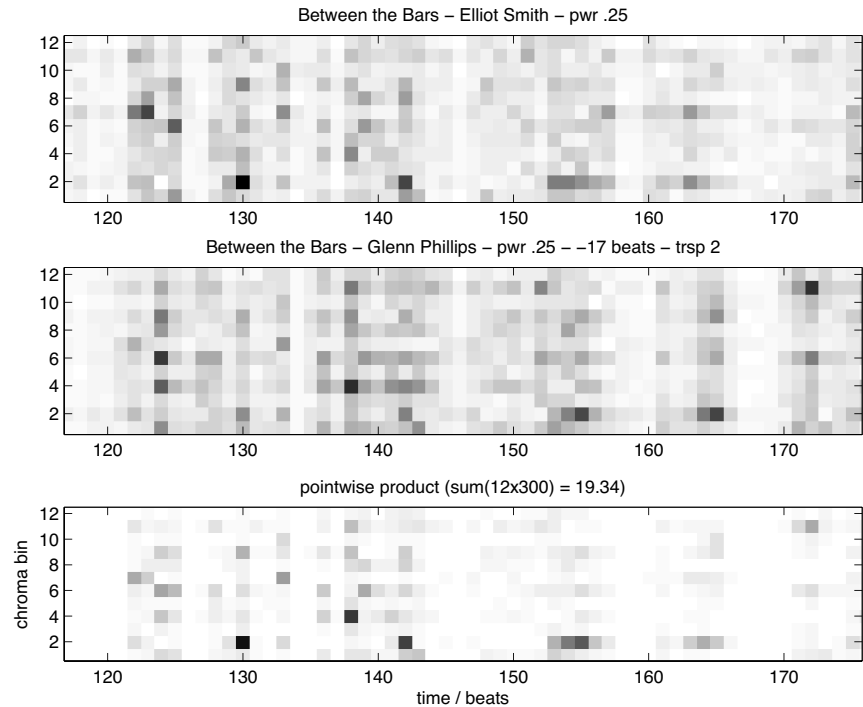


Freddy – chord-based chroma reconstruction



# Future Work

- **Matching items**
  - cover songs / standards
  - similar instruments, styles



- **Analyzing musical content**
  - solo transcription & modeling
  - musical structure
- **And so much more...**

# Summary

- Finding Musical Similarity at Large Scale

