Perceptually-Inspired Music Audio Analysis

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

dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Perceptually-Inspired Analysis
2. The Acoustic Structure of Music
3. Music Scene Analysis
4. Large Music Collections
5. Open Issues

Laboratory for the Recognition and Organization of Speech and Audio

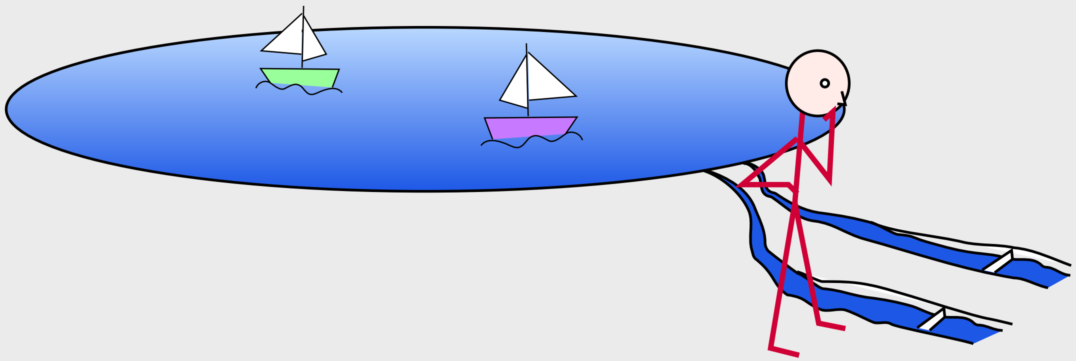
1. Perceptually-Inspired Analysis

- Machine Listening:
  Extracting useful information from sound

Task	Environmental Sound	Speech	Music
Describe	Automatic Narration	Emotion	Music Recommendation
Classify	Environment Awareness	ASR	Music Transcription
Dectect	"Sound Intelligence"	VAD	Speech/Music
			Domain

Listening to Mixtures

Bregman '90

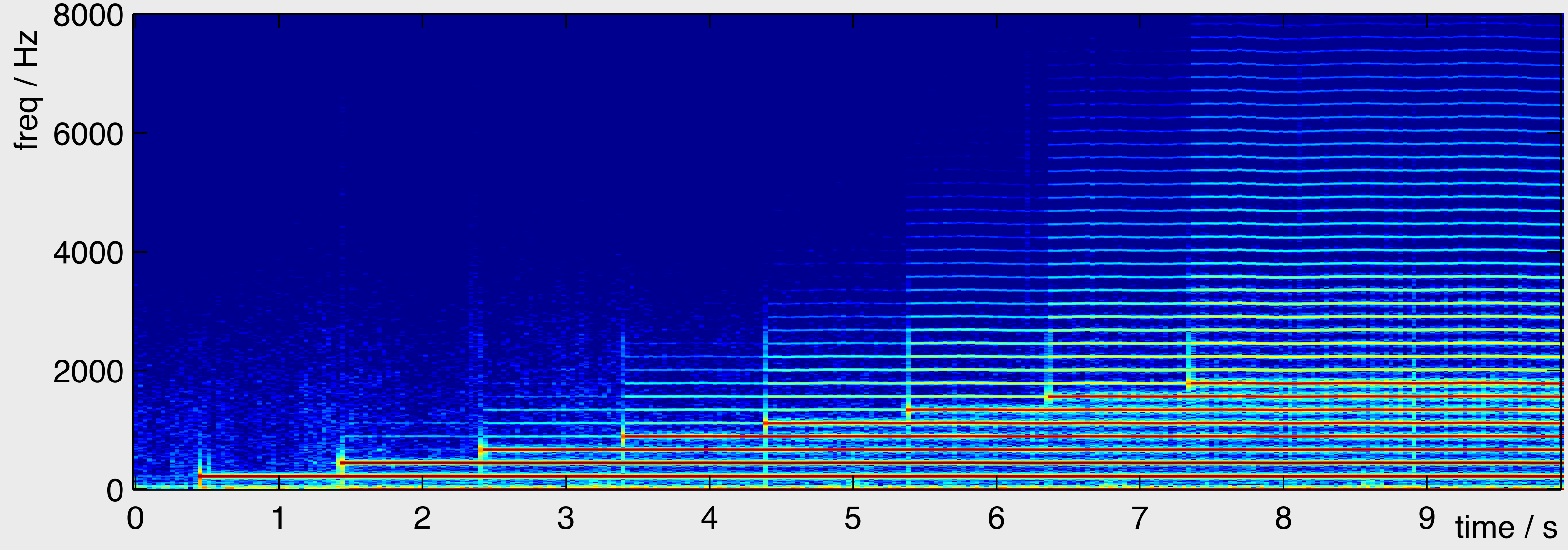


- The world is cluttered
  sound is transparent
  - mixtures are inevitable

- Useful information is structured by 'sources'
  - specific definition of a 'source':
    intentional independence

Scene Analysis

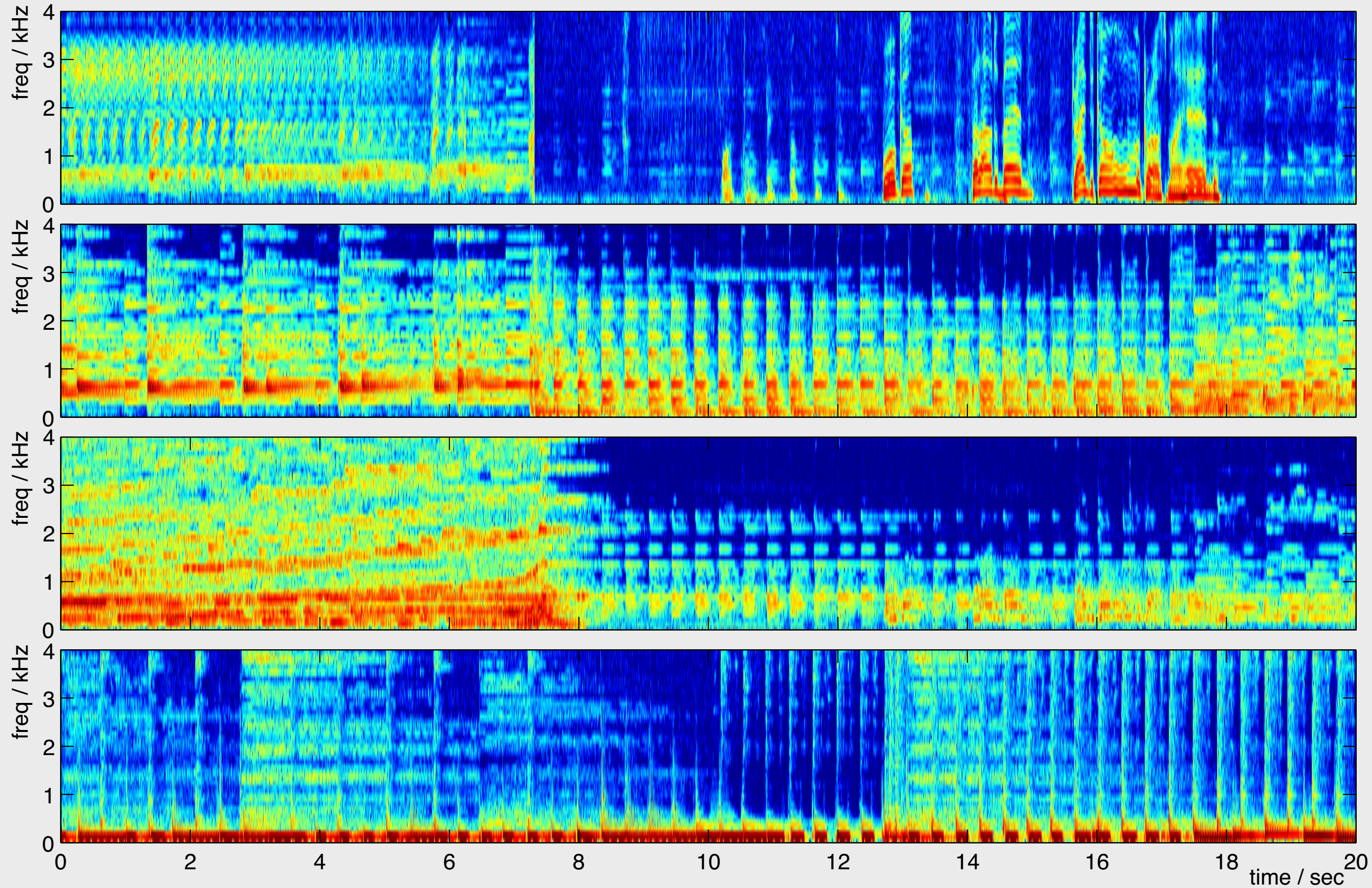
- Detect separate events
  - common onset
  - common harmonicity



Pierce '83

  - instruments & timbre

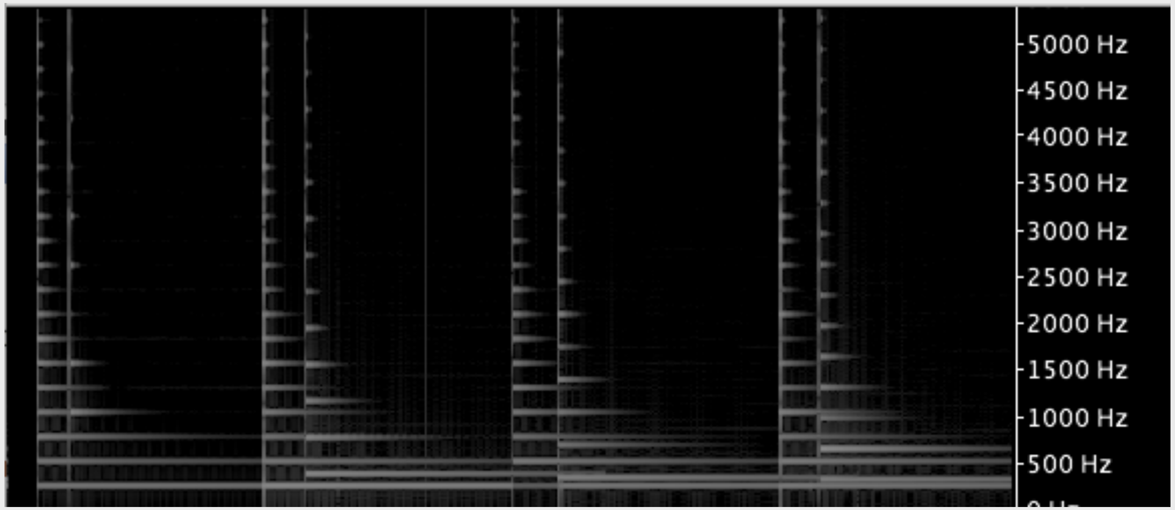
2. The Acoustic Structure of Music



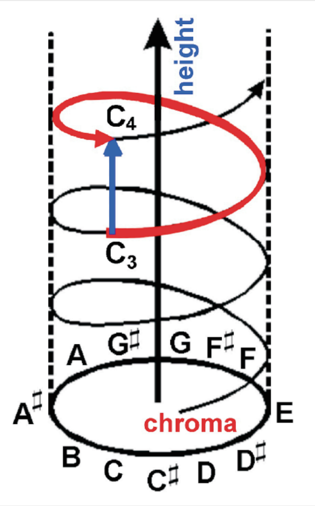
- Pitches, Voices, Rhythm

Pitch, Harmony, Consonance

- Musical intervals relate to harmonic proximity

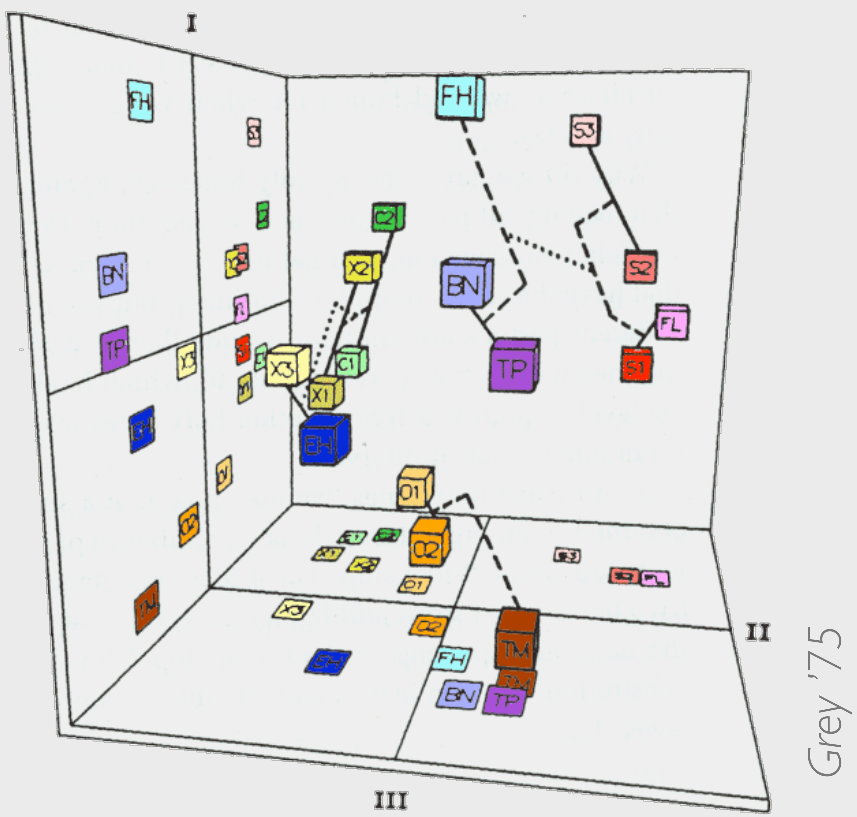


- Pitch Helix



Timbre

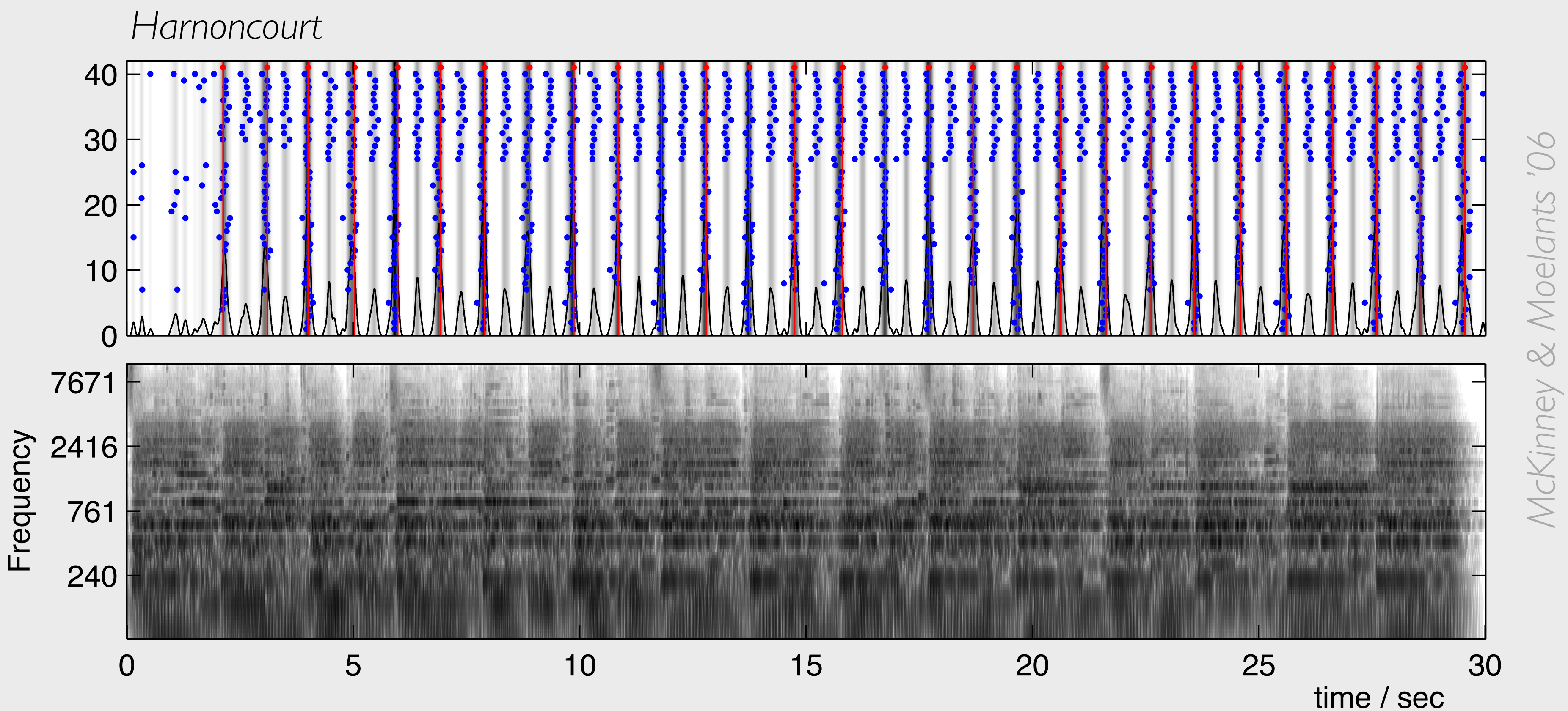
- The property that distinguishes instruments



  - spectrum, noise, onset, dynamics, ...

Rhythm

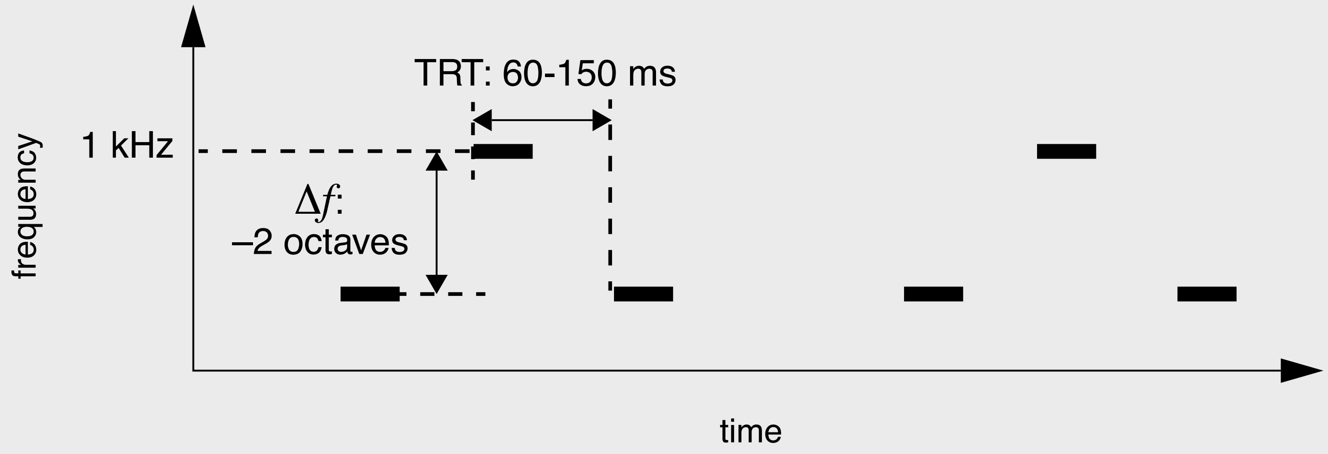
- Periodic "events" perceived as a structure



- hierarchy, swing

Sequences & Streaming

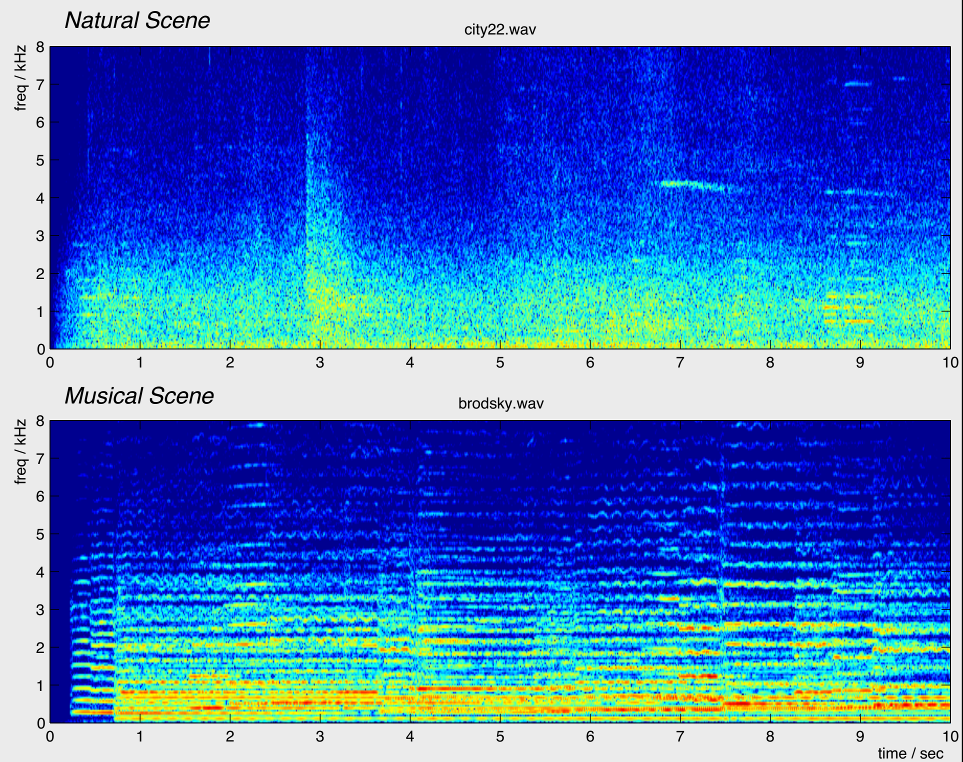
- Perceptual effects of sequences
  - e.g. streaming



- Music is built of sequences
  - at many different levels

Music and Scene Analysis

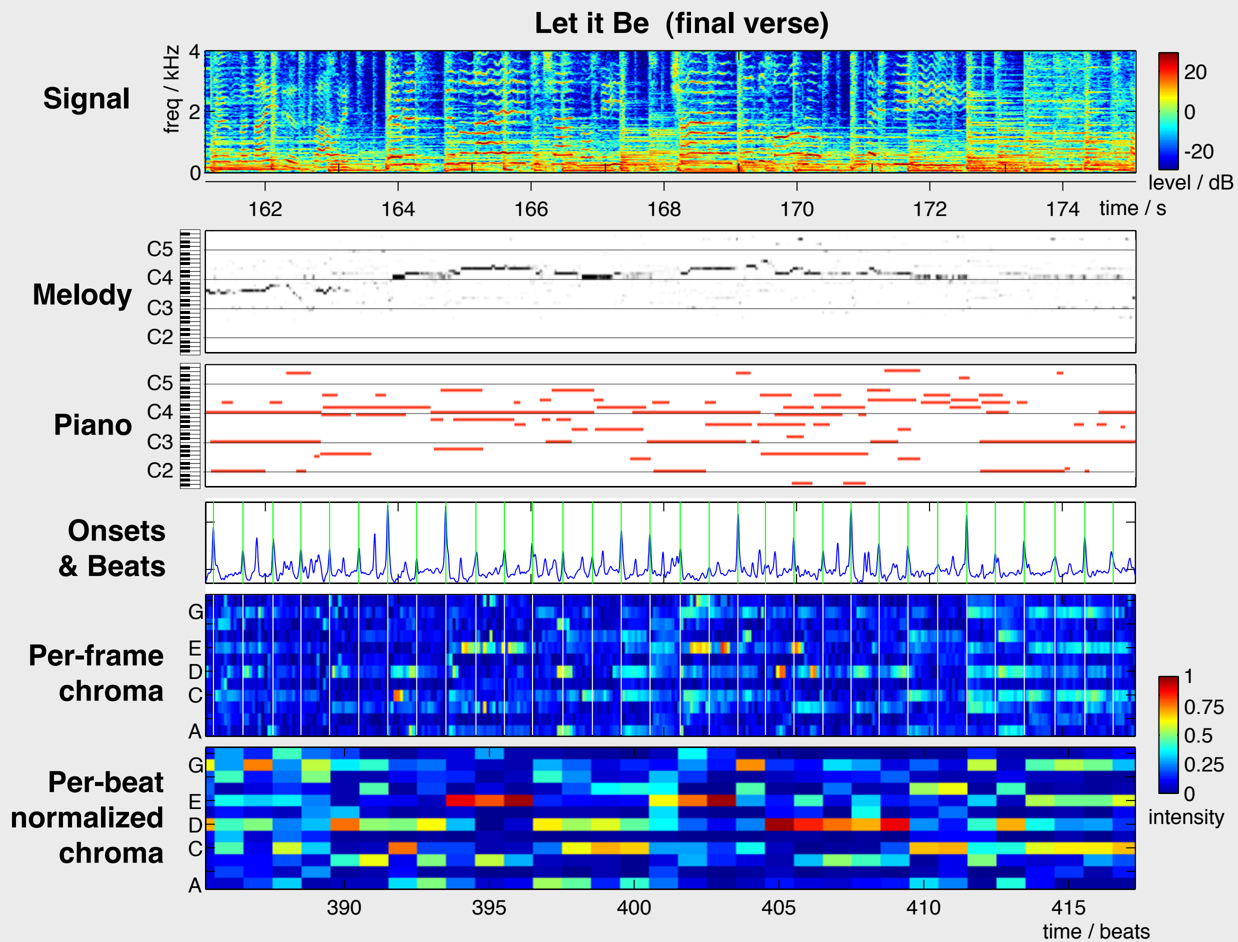
- Music appears designed to "defeat" auditory scene analysis
  - harmonic relations
    → overlapped harmonics
  - rhythmic playing
    → synchronized onsets
  - co-ordinated ensembles
    → mutual dependence of sources



- Maybe that's why we like it!

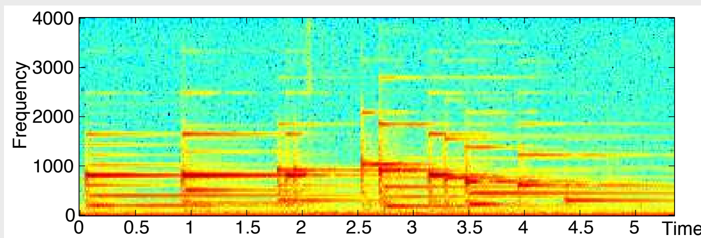
3. Music Scene Analysis

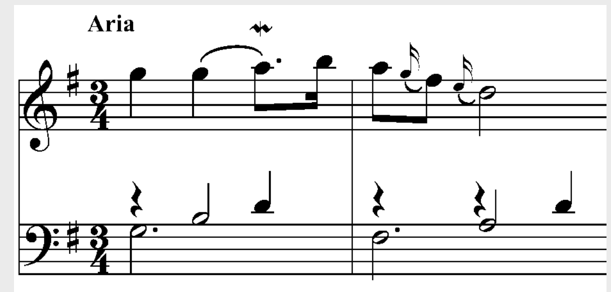
- Interesting music audio analysis tasks are perceptually defined
    - what do listeners hear?



Note Transcription

- Goal: Recover the score (notes, timing, voices)

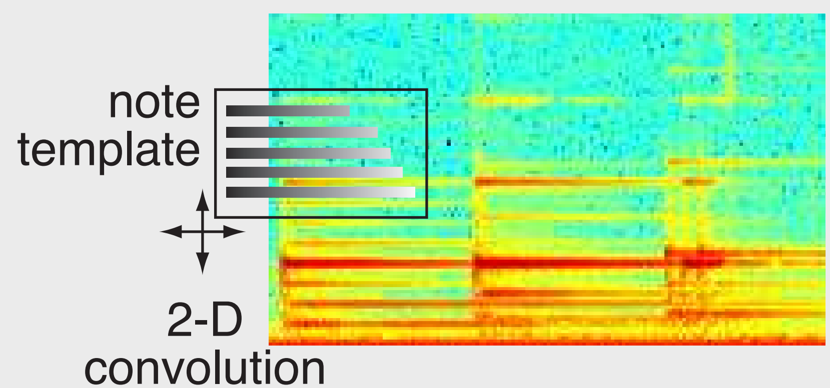




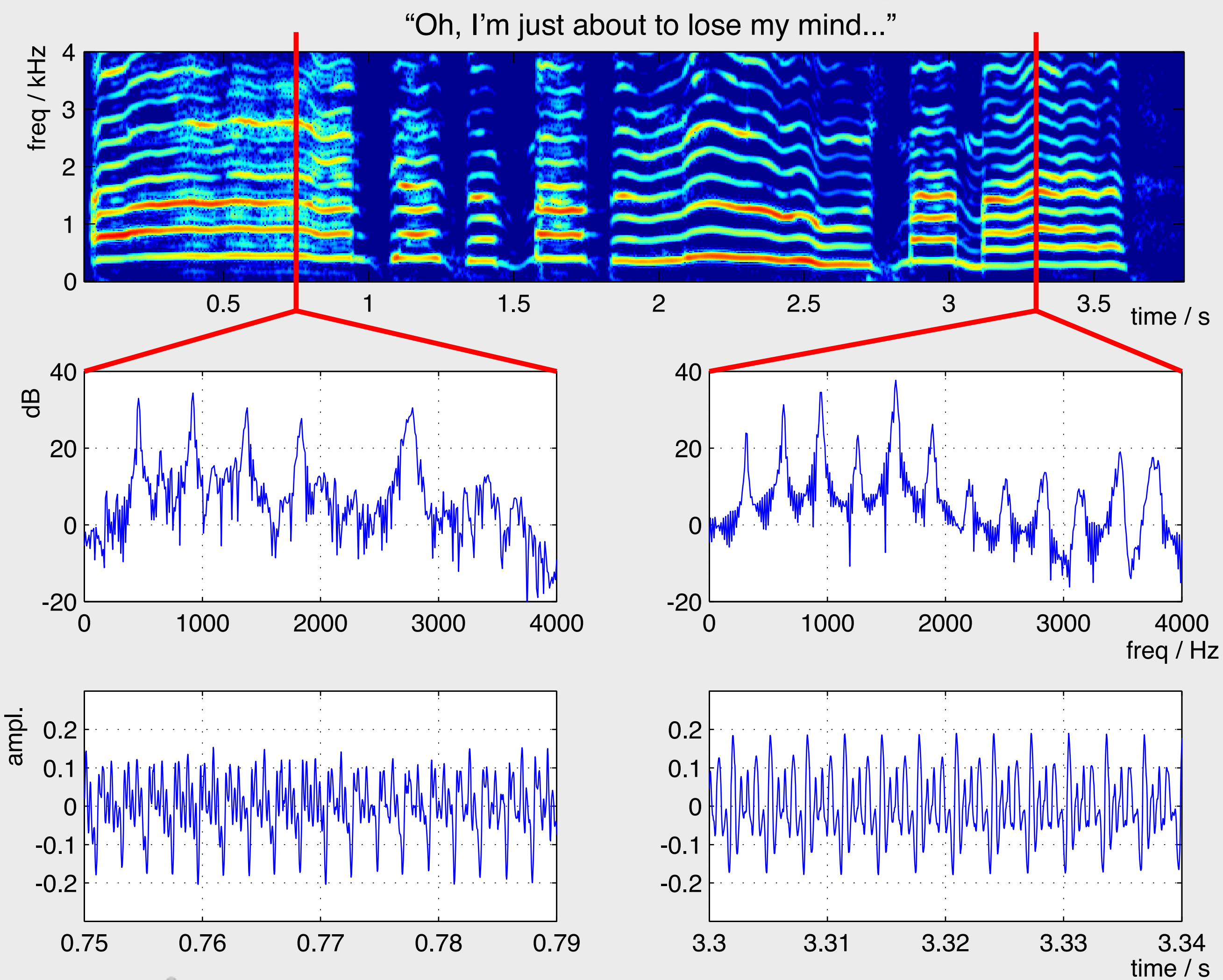
  - musicians can (be trained to) do it

- **Framework:**
  - find the best-matching synthesis parameters?

Note events $\{t_k, p_k, i_k\}$ $\xrightarrow[\textit{synthesis}]{}$? Observations $X[k,n]$



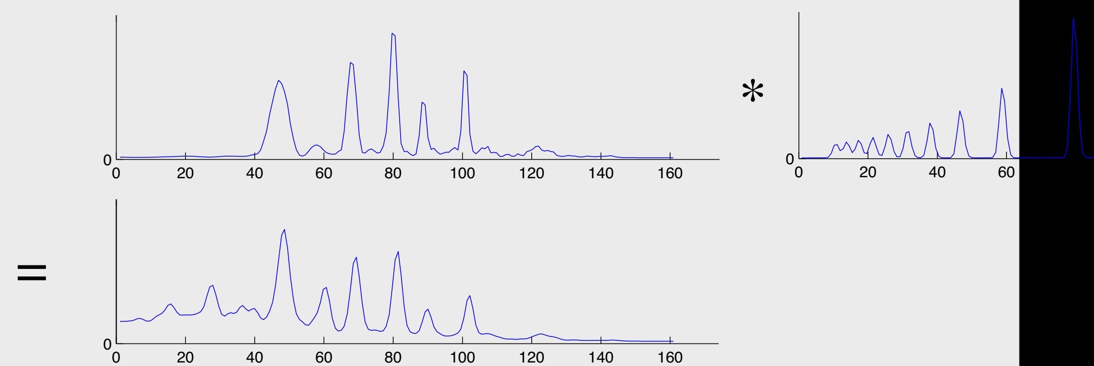
Note Transcription Problems

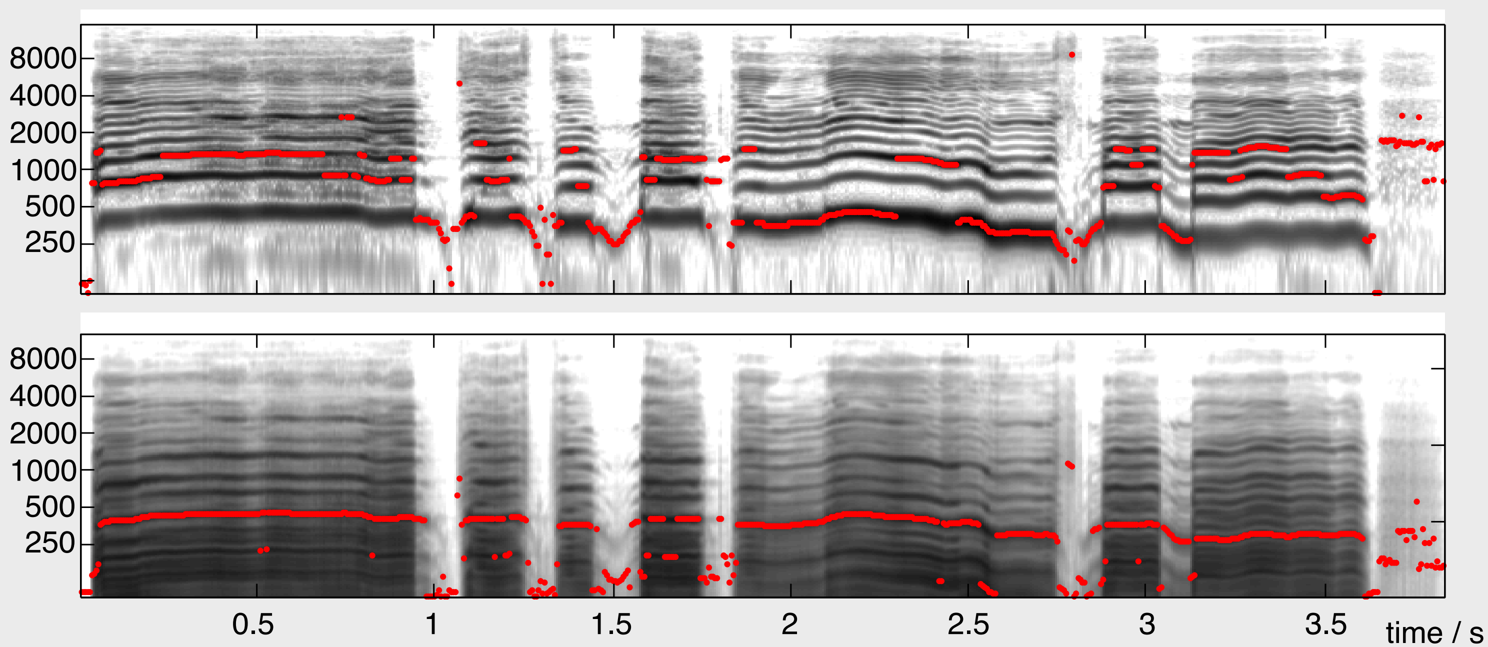


- noise / multiple f_0s
- Voice Activity Detection
- unclear f_0
- note segmentation

Pitch Templates

- Harmonic series as patterns on log-frequency spectrograms
  - look for largest peak?
  - matched filters can enhance fundamental

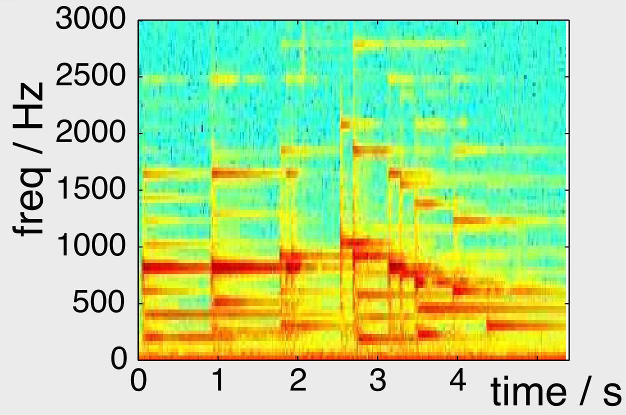


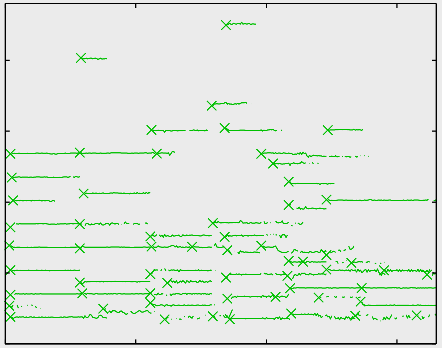


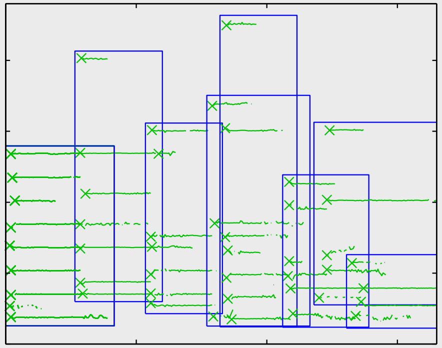
Sinusoid Tracks

Maher & Beauchamp '94

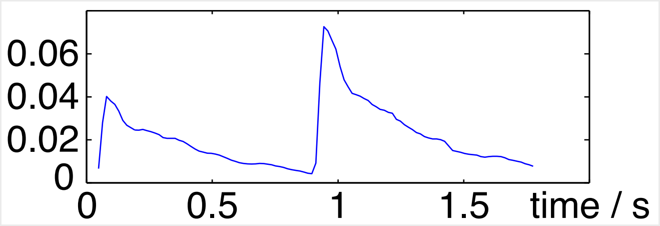
- **Notes generate multiple harmonics in sinusoid analysis**
  - find pitches by grouping them?







- **Problems**
  - when to "break tracks"

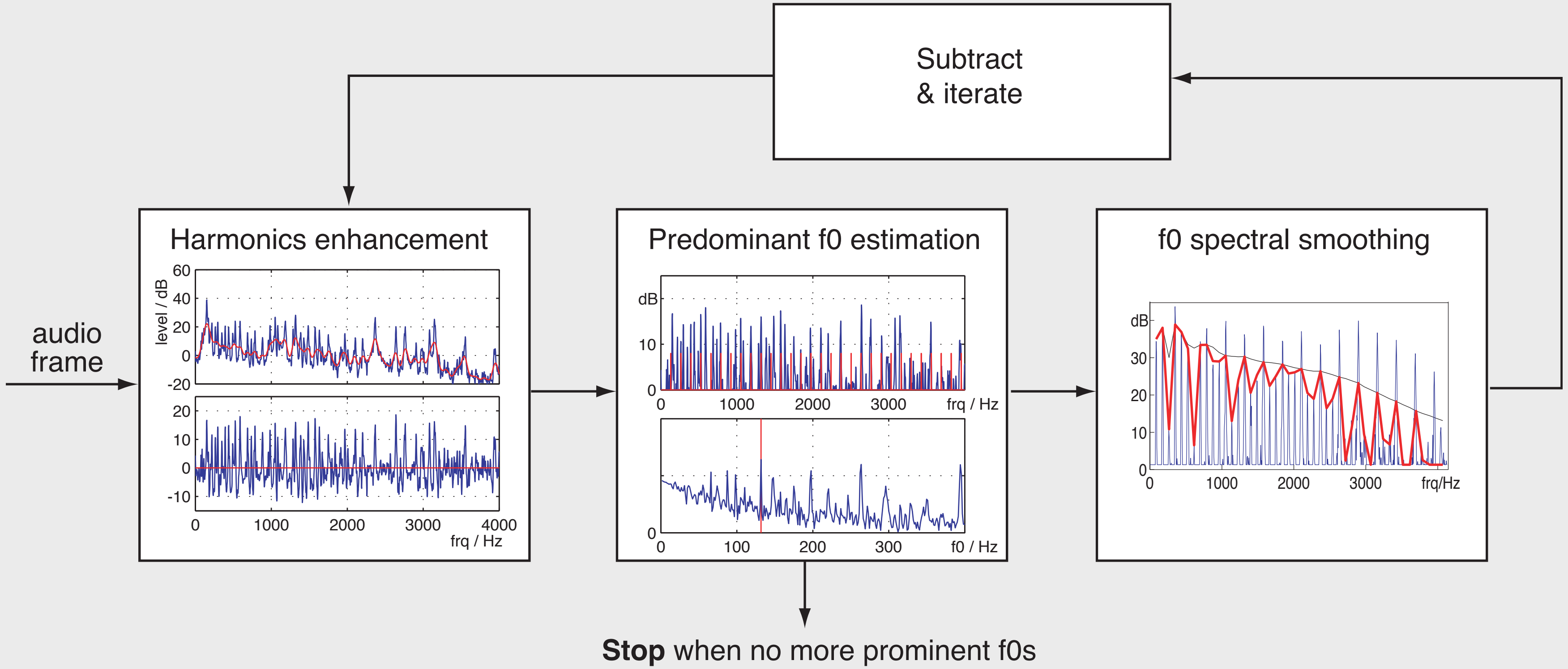


  - how to group (harmonicity, onset)

Iterative Removal

Klapuri '01, '06

- At each frame:
  - estimate dominant f_0 by checking for harmonics
  - cancel it from spectrum
  - repeat until no f_0 is prominent



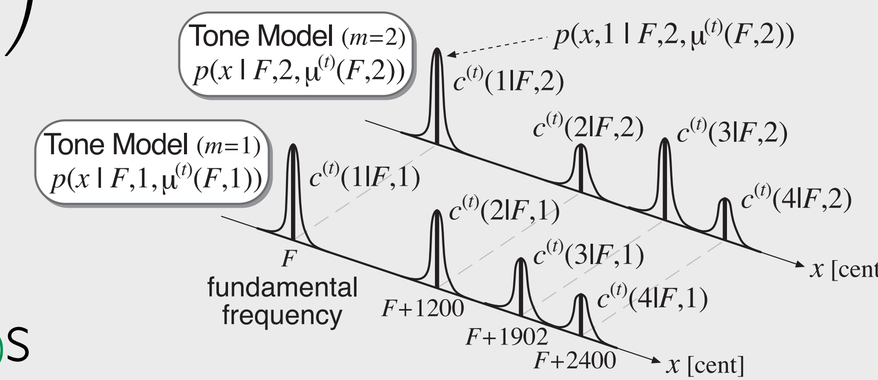
Probabilistic Model

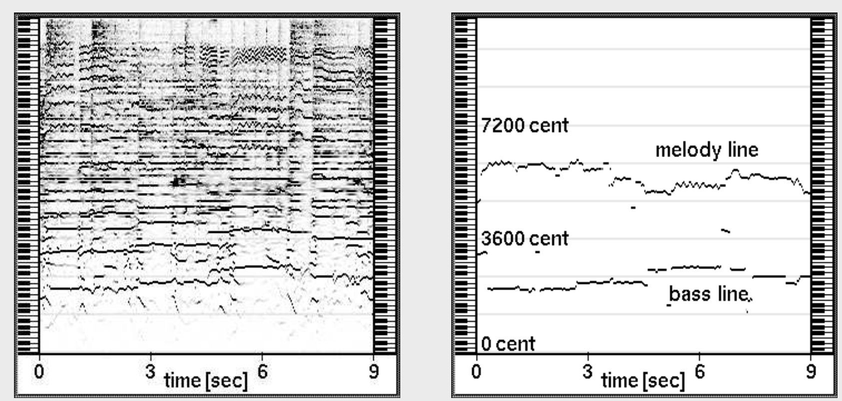
Goto '97, '00

- **Generative model:**

$$p(x(f)) = \int \left(\sum_m w(F,m)\, p(x(f)|F,m) \right) dF$$

  - spectrum
    = weighted combination
    of tone models at specific f_0s
  - 'knowledge' in models
    & prior distributions for f_0





- **Is it perceptually relevant?**

Trained Pitch Classifier

Poliner & Ellis ‘05,’06,’07

- Exchange signal models for data
  - transcription as pure classification problem

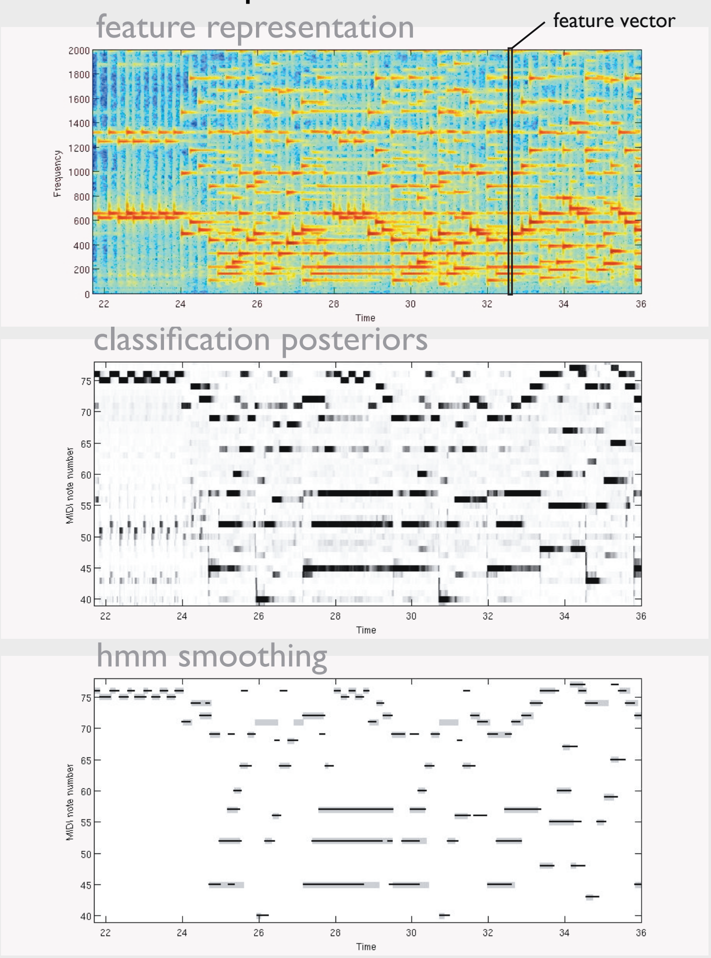
Training data and features:
- MIDI, multi-track recordings, playback piano, & resampled audio (less than 28 mins of train audio).
- Normalized magnitude STFT.

↓

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

↓

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



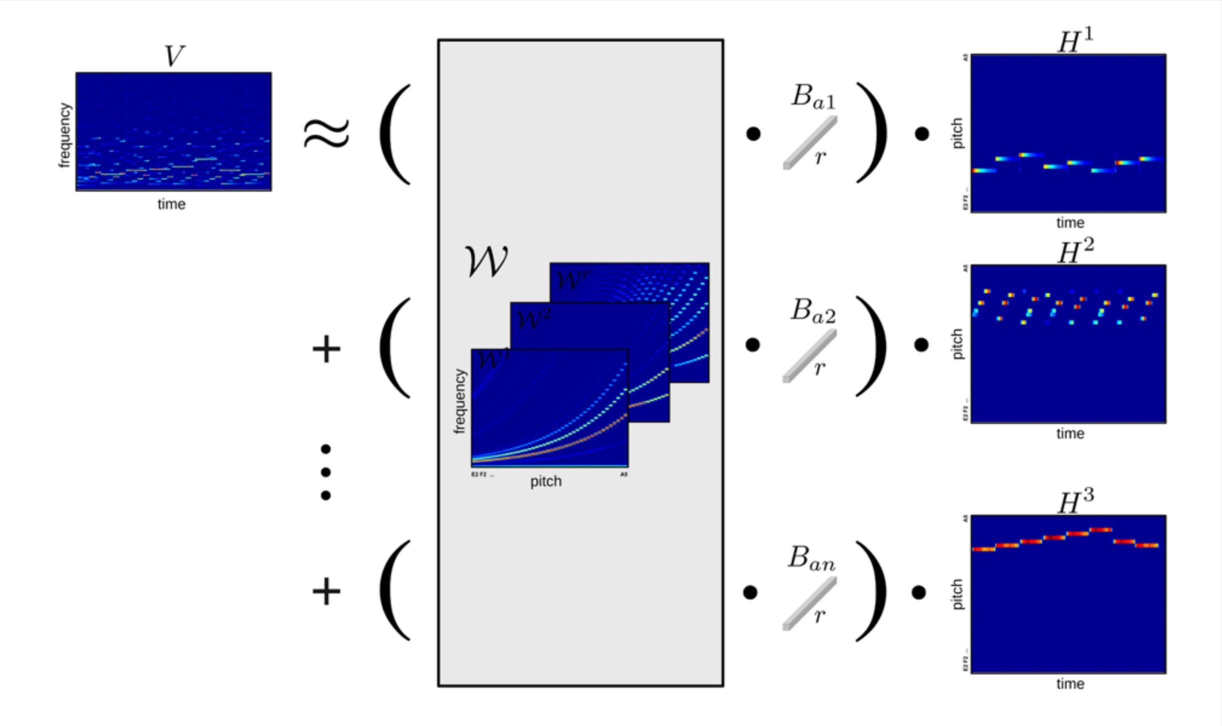
Instrument Modeling

Grindlay & Ellis '09, '11

- Use NMF to model spectrum as templates + activation

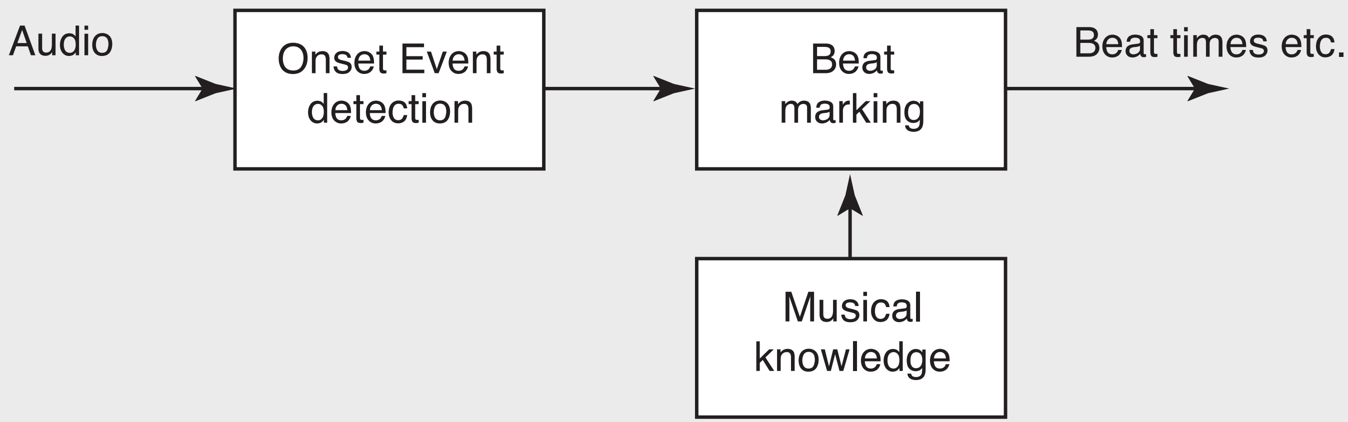
$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$$

- Eigeninstrument bases constrain instrument spectra



Rhythm Tracking

- Rhythm/Beat tracking has 2 main components:
  - front end: extract 'events' from audio
  - back end: find plausible beat sequence to match



- Other outputs
  - tempo
  - time signature
  - metrical level(s)

Onset Detection

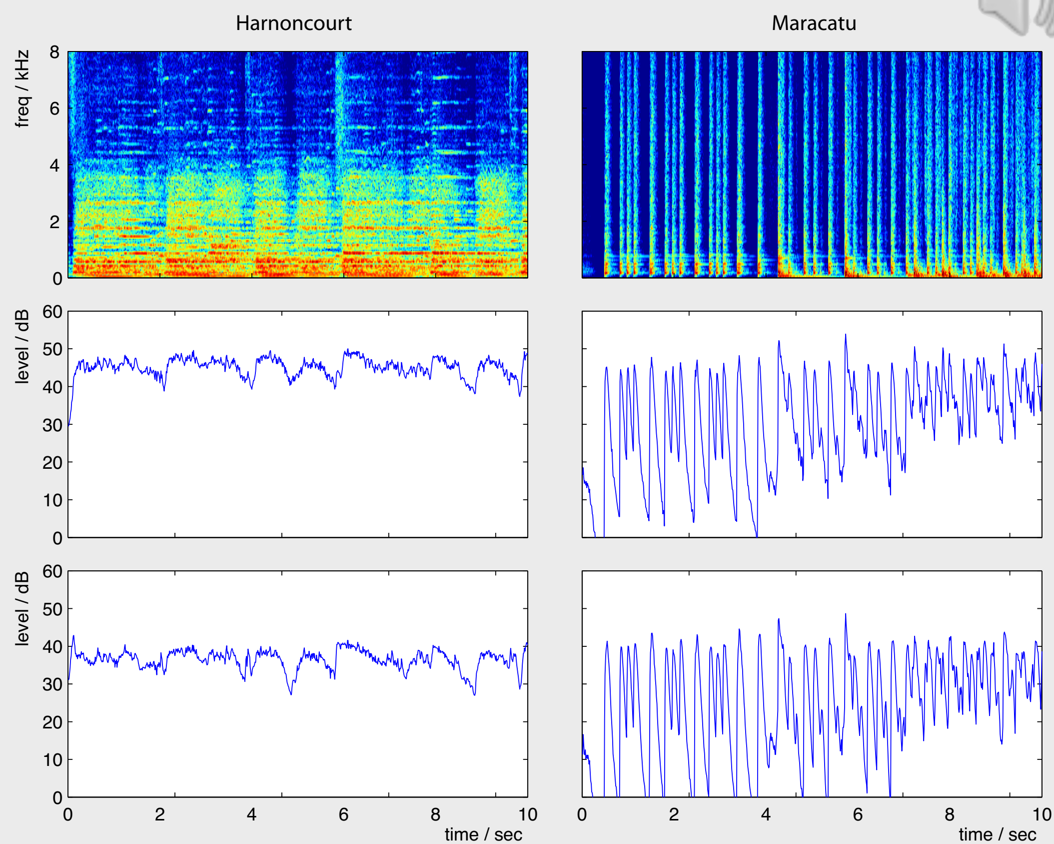
- Simplest thing is energy envelope

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

  - emphasis on high frequencies?

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

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



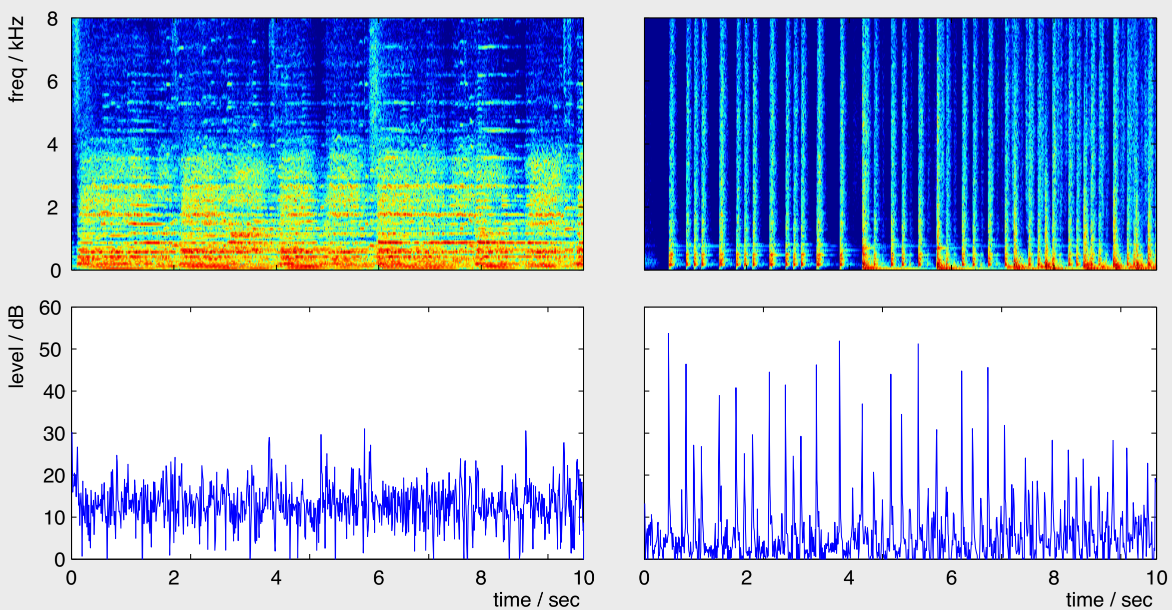
Multiband Derivatives

Puckette et. al '98

- Sometimes energy just "shifts"
  - calculate & sum onset in multiple bands
  - use ratio instead of difference - normalize energy

$$o(t) = \sum_f W(f) \max(0, \frac{|X(f,t)|}{|X(f,t-1)|} - 1)$$

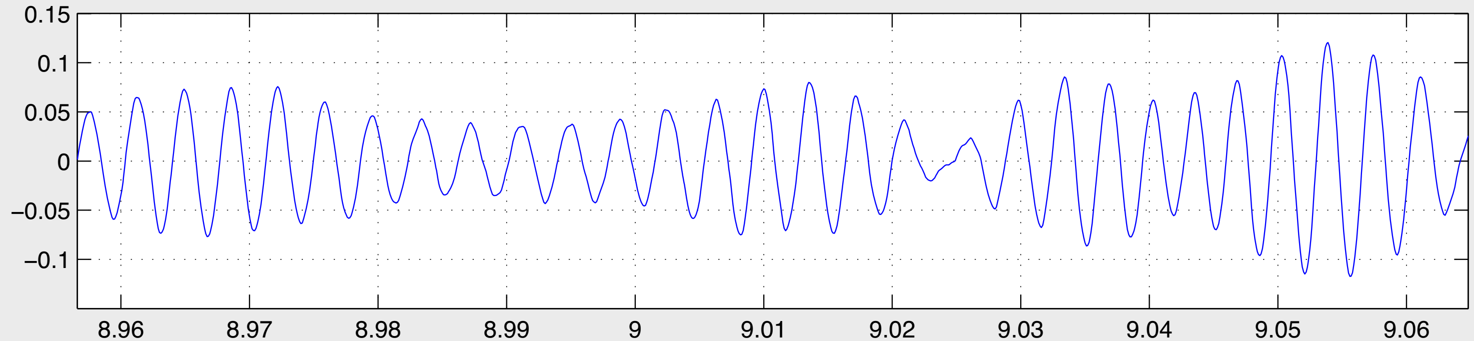
bonk~



Phase Deviation

Bello et al. '05

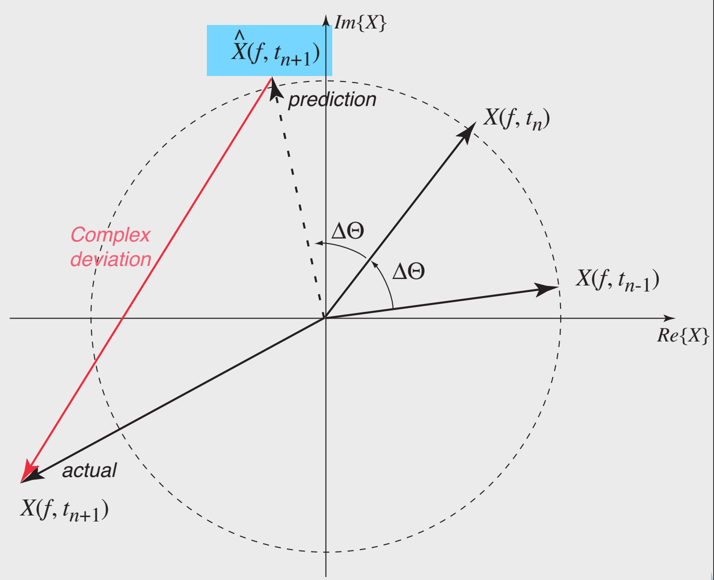
- When amplitudes don't change much, phase discontinuity may signal new note



- Can detect by comparing actual phase with extrapolation from past

$$\hat{X}(f, t_{n+1}) = X(f, t_n) \frac{X(f, t_n)}{X(f, t_{n-1})}$$

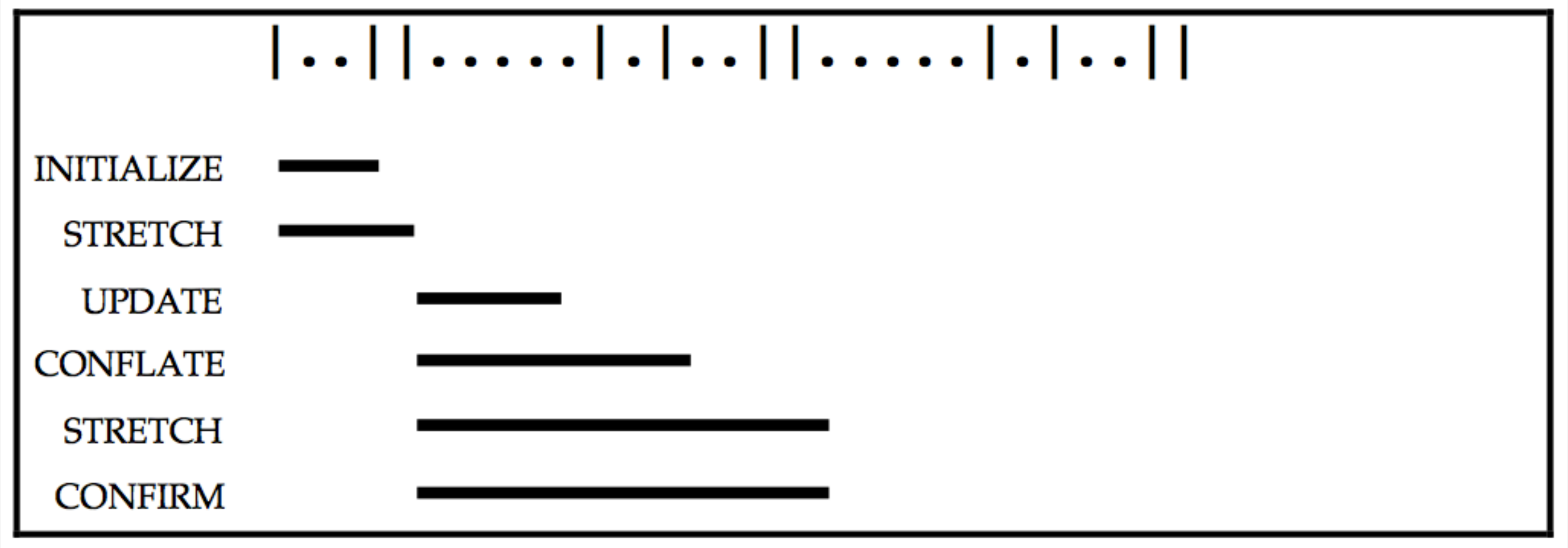
  - combine with amplitude...



Rhythm Tracking

Desain & Honing 1999

- **Earliest systems were rule based**
  - based on musicology *Longuet-Higgins and Lee, 1982*
  - inspired by linguistic grammars - Chomsky

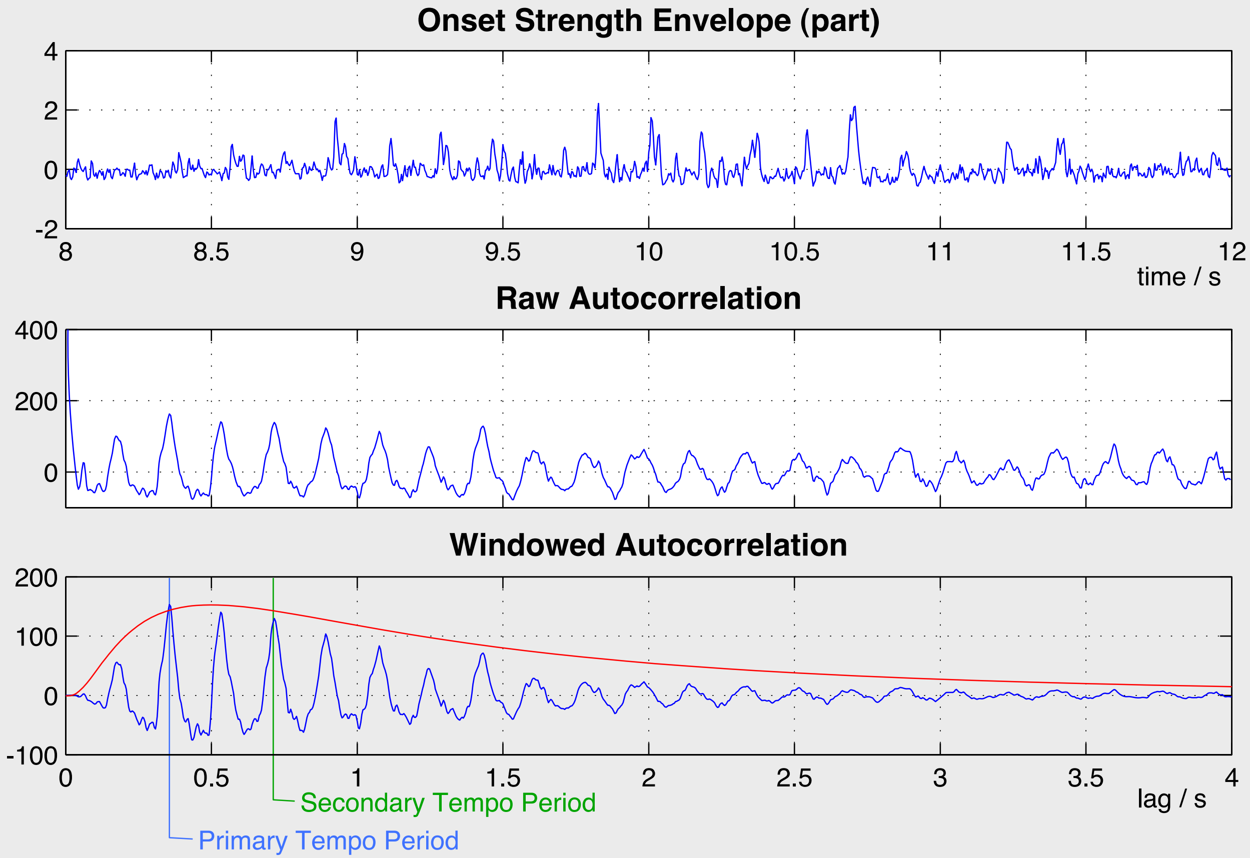


  - input: event sequence (MIDI)
  - output: quarter notes, downbeats

Tempo Estimation

- Perception of beat comes from regular spacing
  - .. the kind of thing we detect with autocorrelation

- Pick peak in onset envelope autocorrelation
  - after applying "human preference" window
  - check for subbeat



Resonators for Beat Tracking

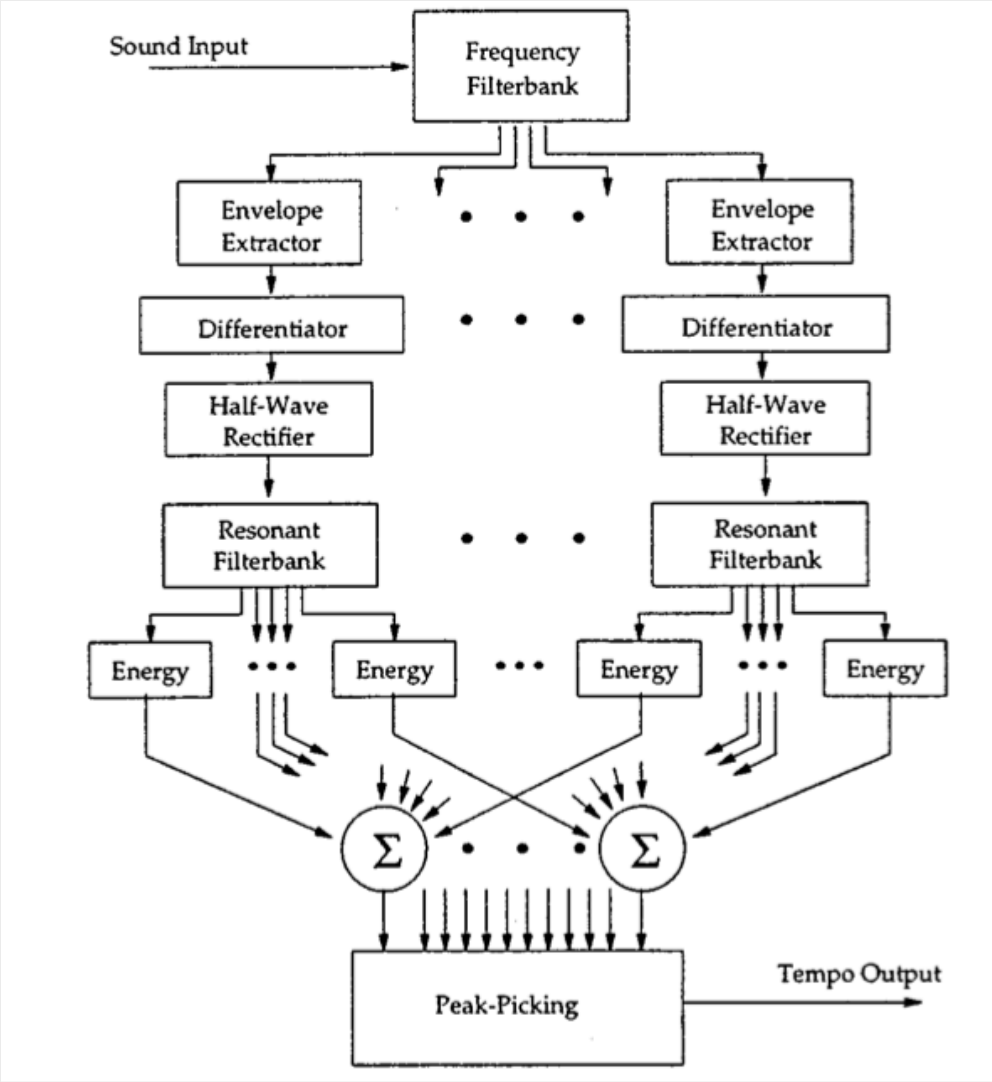
- How to address:
  - build-up of rhythmic evidence
  - "ghost events"
  - (audio input)

- Reminiscent of a comb filter...
  - resonant filterbank of

$$y(t) = \alpha y(t - T) + (1 - \alpha)x(t)$$

for all possible T

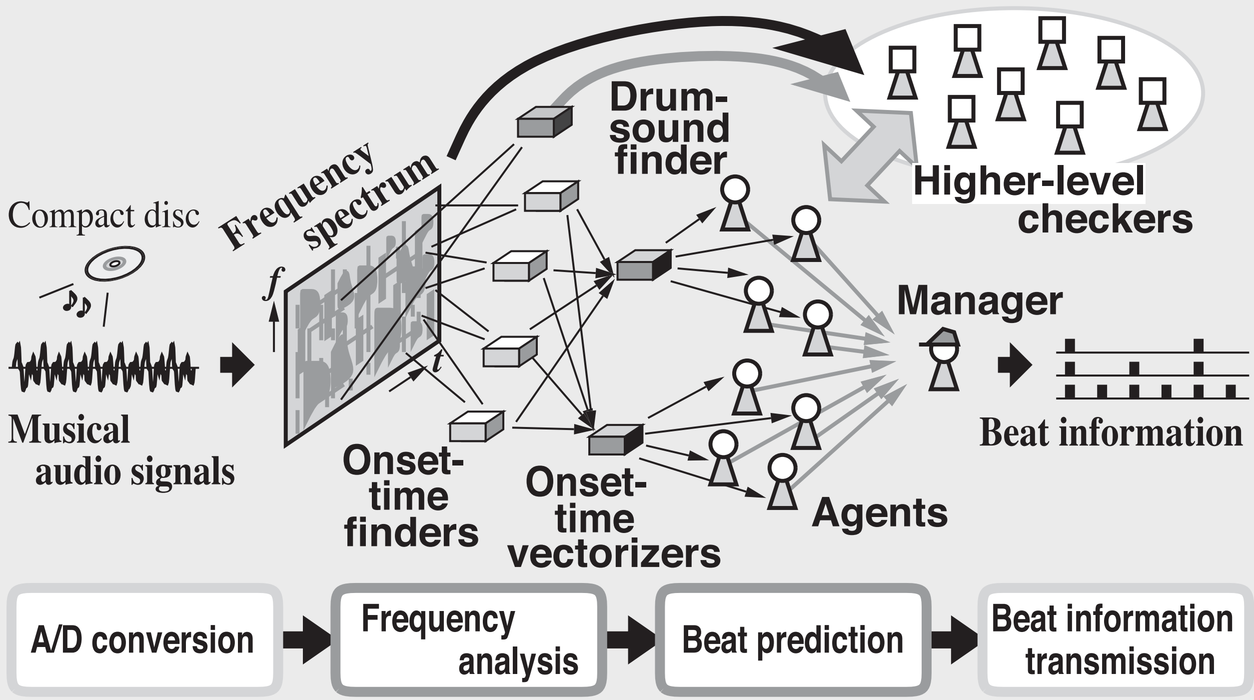
Scheirer '98



Multi-Hypothesis Systems

Goto & Muraoka 1994
Goto 2001
Dixon 2001

- Beat is ambiguous
  → develop several alternatives



- inputs: music audio
- outputs: beat times, downbeats, BD/SD patterns...

Objective Function Optimization

Ellis 2007

- Re-cast beat tracking as optimization:
  Find beat times $\{t_i\}$ to maximize

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

  - $O(t)$ is onset strength function
  - $F(\Delta t, \tau)$ is tempo consistency score e.g.

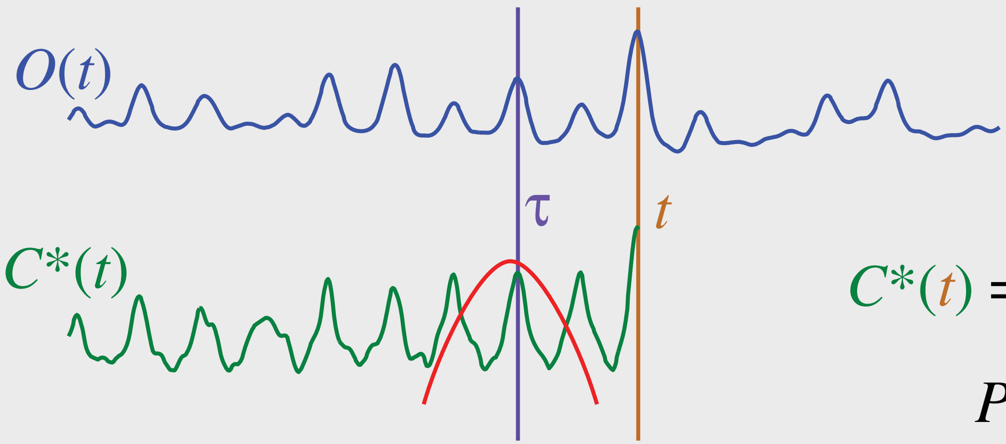
$$F(\Delta t, \tau) = -\left(\log \frac{\Delta t}{\tau}\right)^2$$

  - (needs tempo for τ)

- Looks like an exponential search over all $\{t_i\}$
  - ... but Dynamic Programming saves us

Beat Tracking by DP

- 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

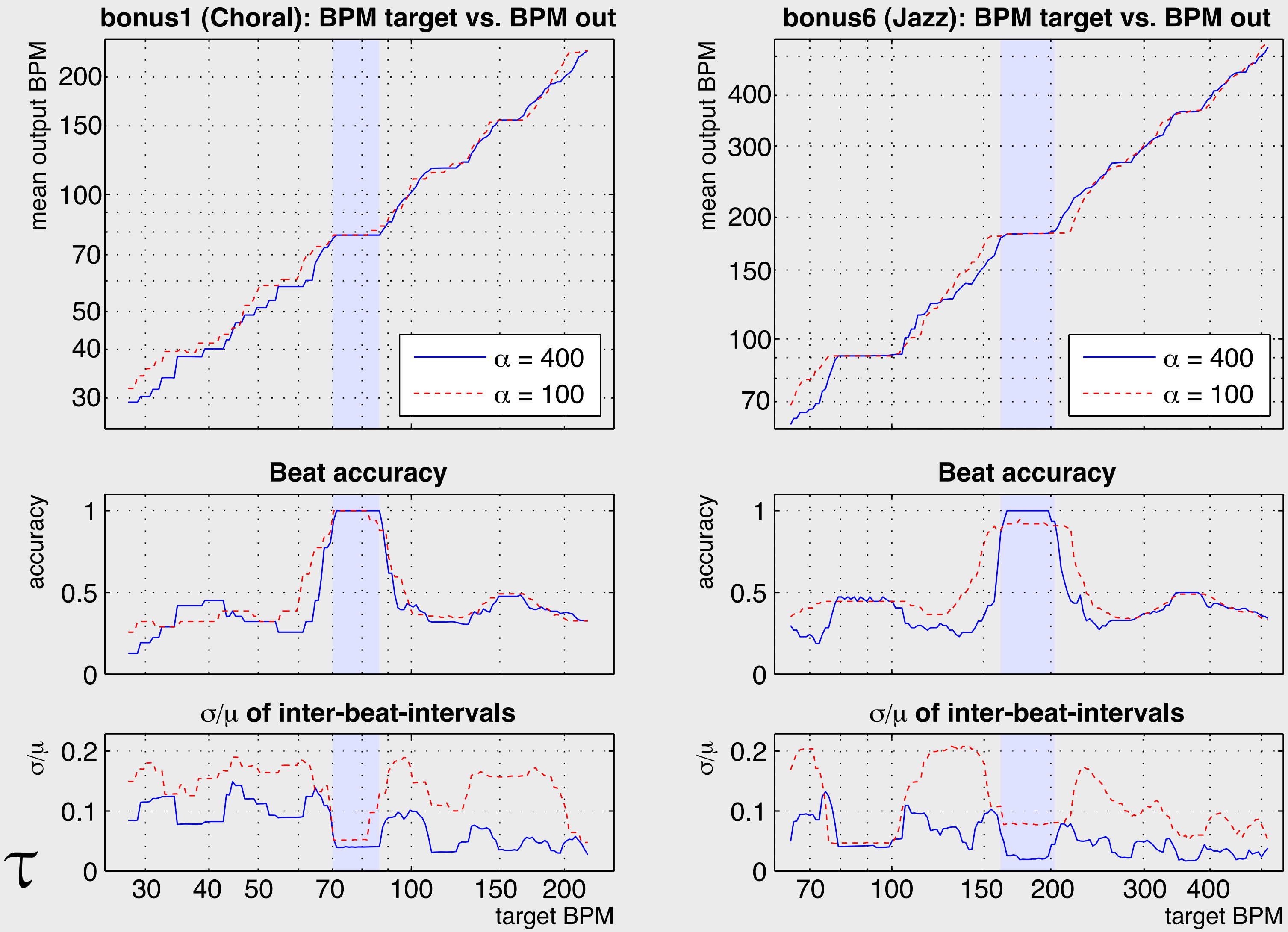
beatsimple

- Beat tracking in 15 lines of Matlab

```
function beats = beatsimple(localscore, period, alpha)
% beats = beatsimple(localscore, period, alpha)
%   Core of the DP–based beat tracker
%   <localscore> is the onset strength envelope
%   <period> is the target tempo period (in samples)
%   <alpha> is weight applied to transition cost
%   <beats> returns the chosen beat sample times.
% 2007–06–19 Dan Ellis dpwe@ee.columbia.edu

% backlink(time) is best predecessor for this point
% cumscore(time) is total cumulated score to this point
backlink = –ones(1,length(localscore));
cumscore = localscore;

% Search range for previous beat
prange = round(–2*period):–round(period/2);
% Log–gaussian window over that range
txcost= (–alpha*abs((log(prange/–period)).^2));

for i = max(–prange + 1):length(localscore)

  timerange = i + prange;

  % Search over all possible predecessors
  % and apply transition weighting
  scorecands = txcost    + cumscore(timerange);
  % Find best predecessor beat
  [vv,xx] = max(scorecands);
  % Add on local score
  cumscore(i) = vv + localscore(i);
  % Store backtrace
  backlink(i) = timerange(xx);

end

% Start backtrace from best cumulated score
[vv,beats] = max(cumscore);
% .. then find all its predecessors
while backlink(beats(1)) > 0
  beats = [backlink(beats(1)),beats];
end
```

Results

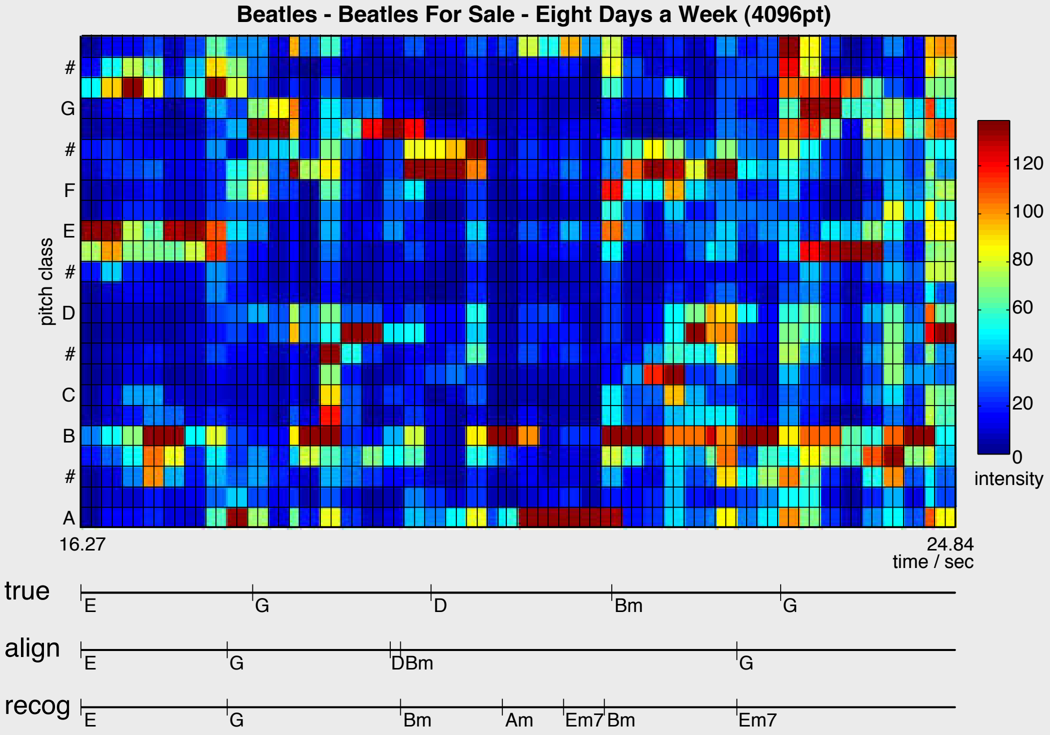
- Verify against human tapping data
  - vary tradeoff weight α
  - vary tempo estimate τ



Chord Recognition

- Do people hear simultaneous notes or do they learn the sound of chords?
  - music limits the likely combinations
  - chords have a definite "color"

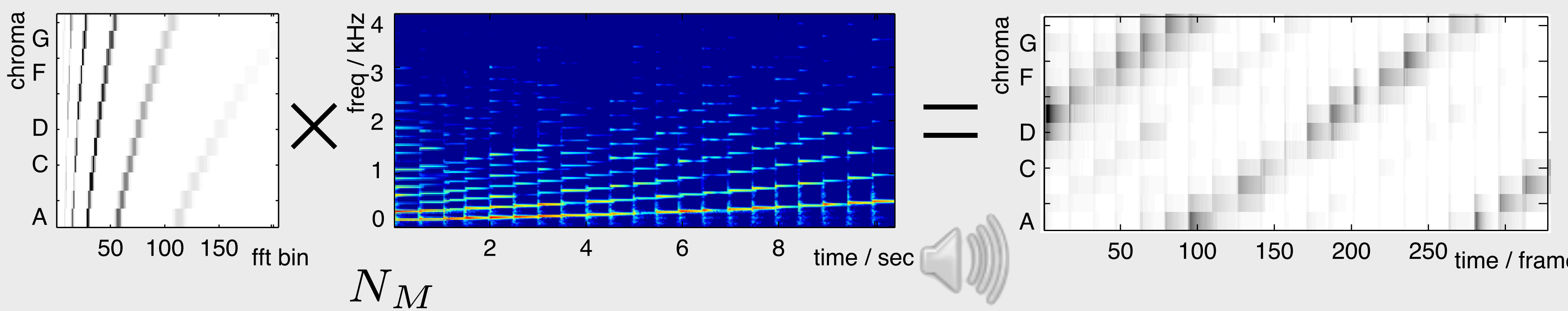
- Recognize chords instead of notes?
  - labeled data available
  - analogous to speech recognition



Chord Features: Chroma

Fujishima 1999

- Idea: Project all energy onto 12 semitones regardless of octave
  - maintains main "musical" distinction
  - invariant to musical equivalence
  - no need to worry about harmonics?



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

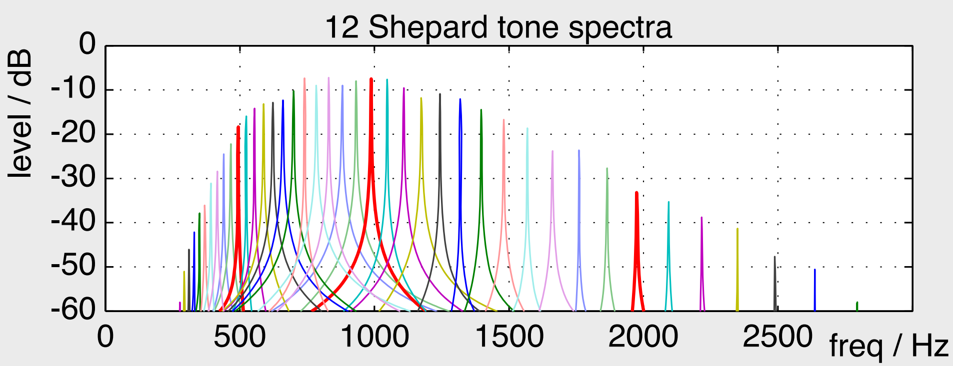
  - $W(k)$ is weighting, $B(b)$ selects every ~ mod12

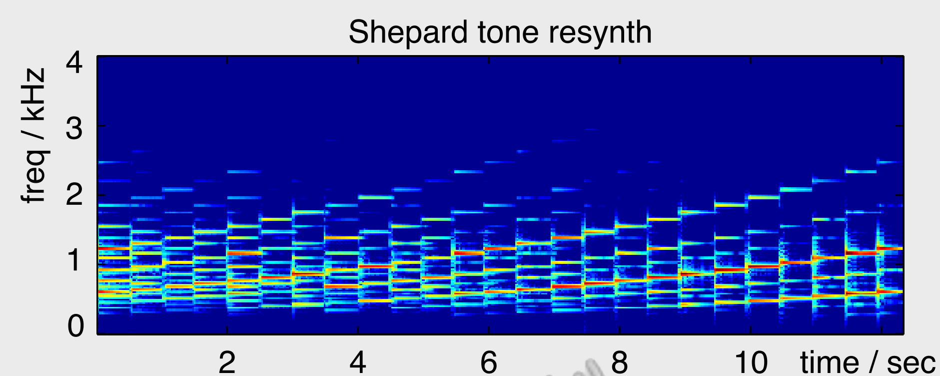
Chroma Resynthesis

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

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

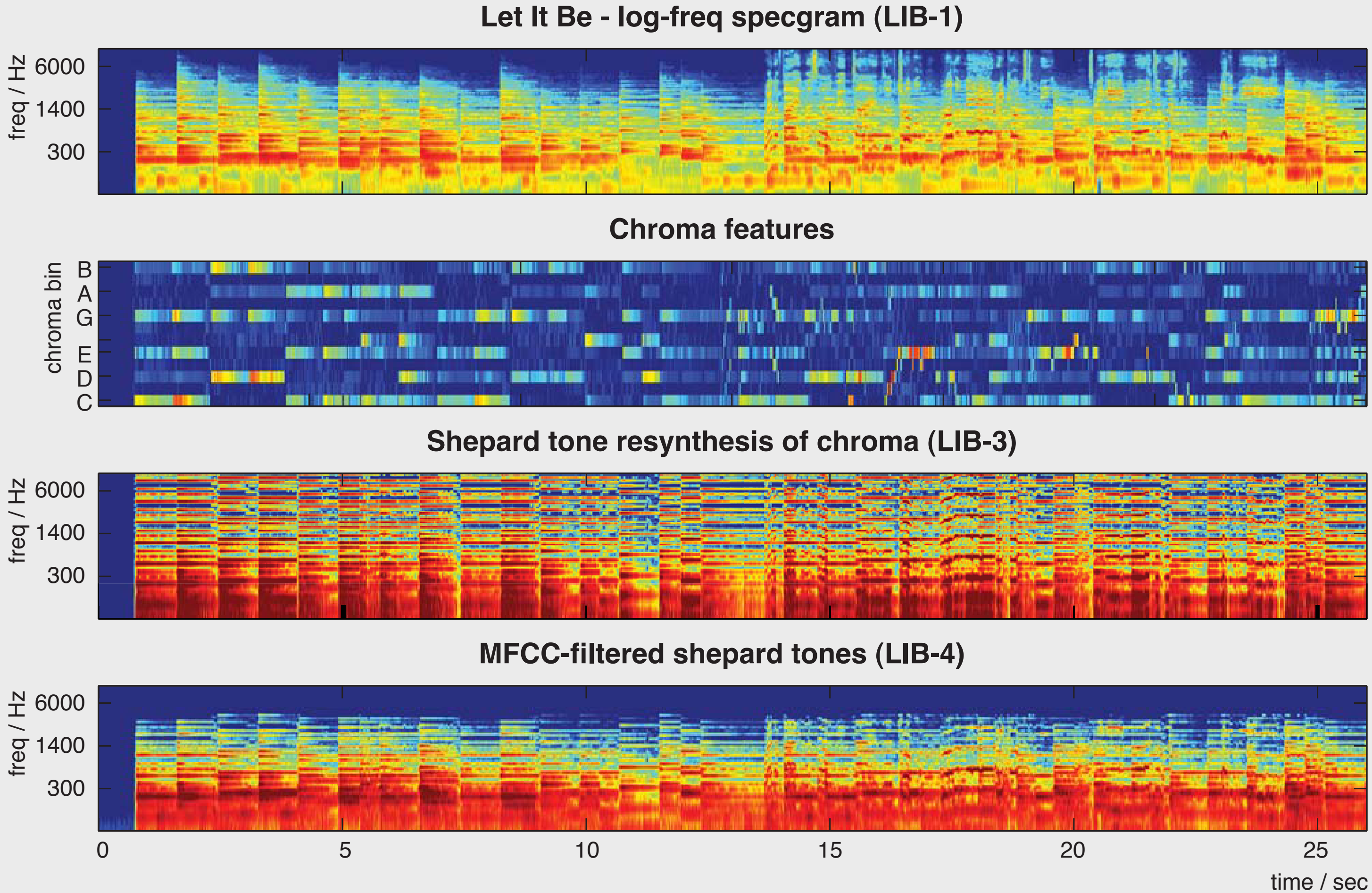




  - endless sequence illusion

Chroma Resynthesis

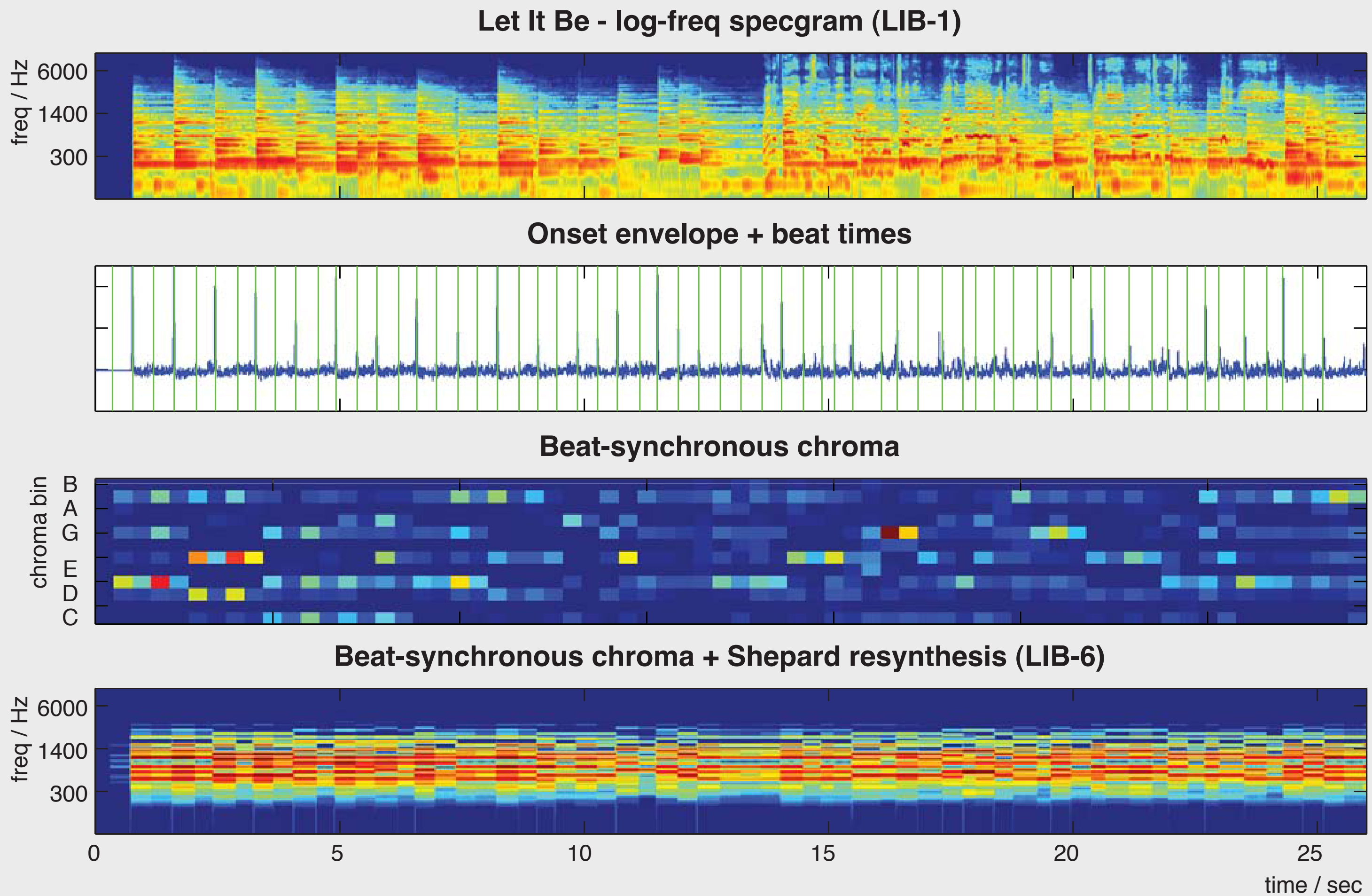
- Resynthesis illustrates what has been captured
  - can combine with MFCC features for coarse spectrum



Beat-Synchronous Chroma

Bartsch & Wakefield '01

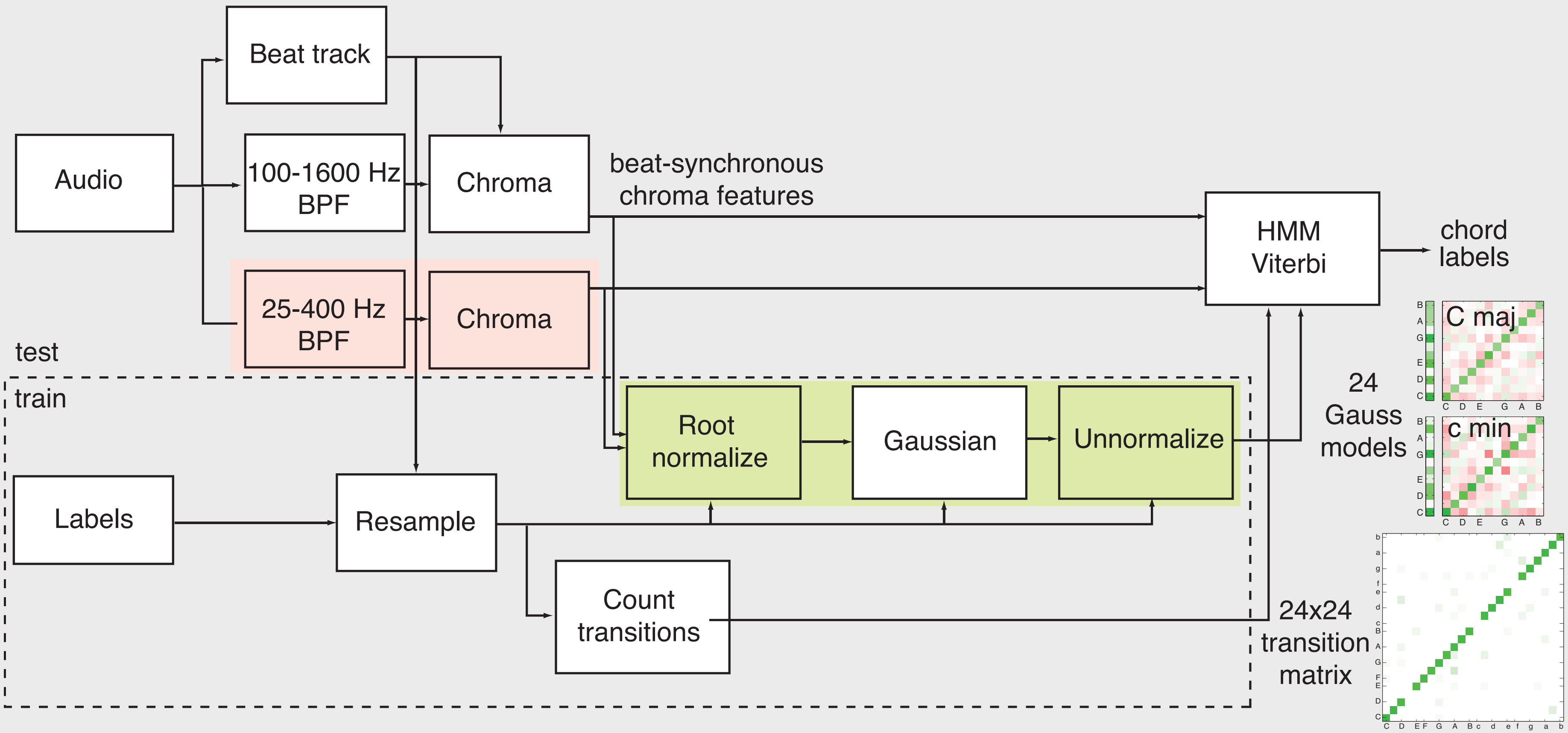
- Store just one chroma frame per beat
  - a compact representation of musical content



Chord Recognition System

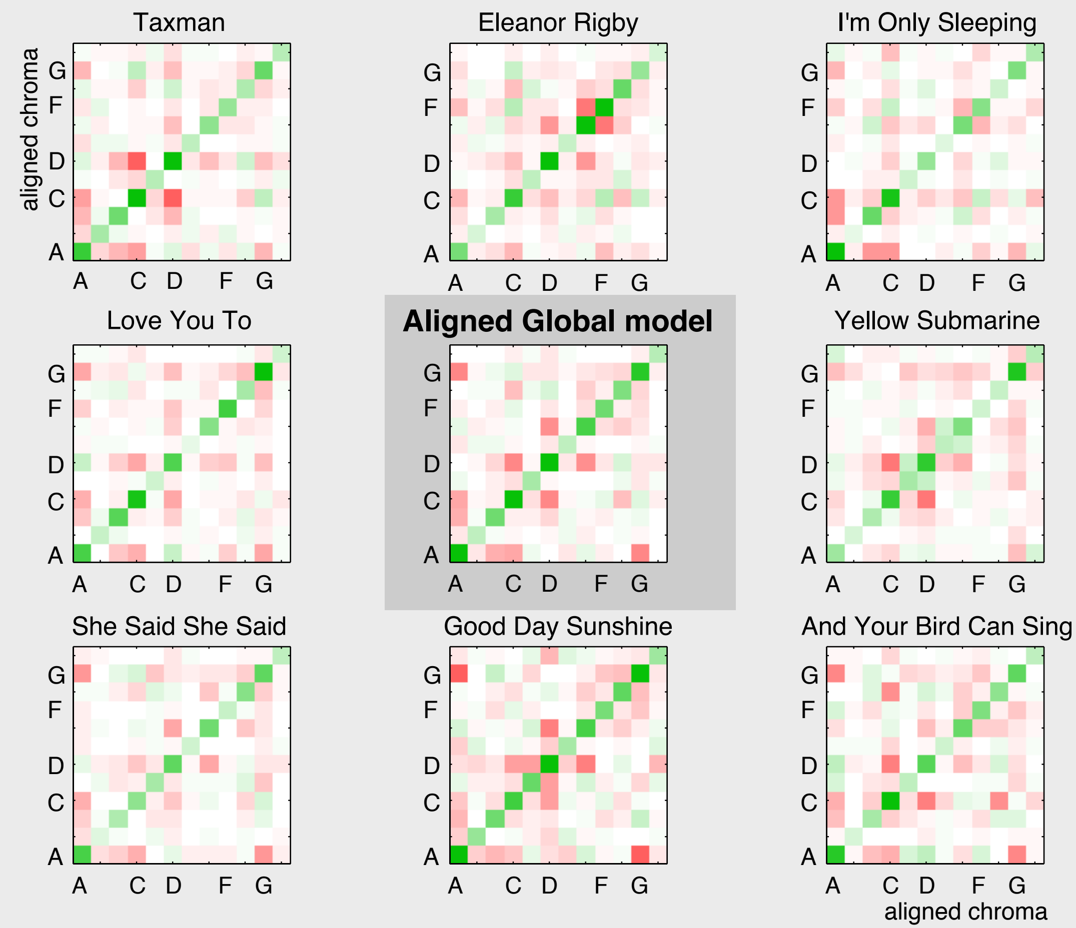
Sheh & Ellis '03

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



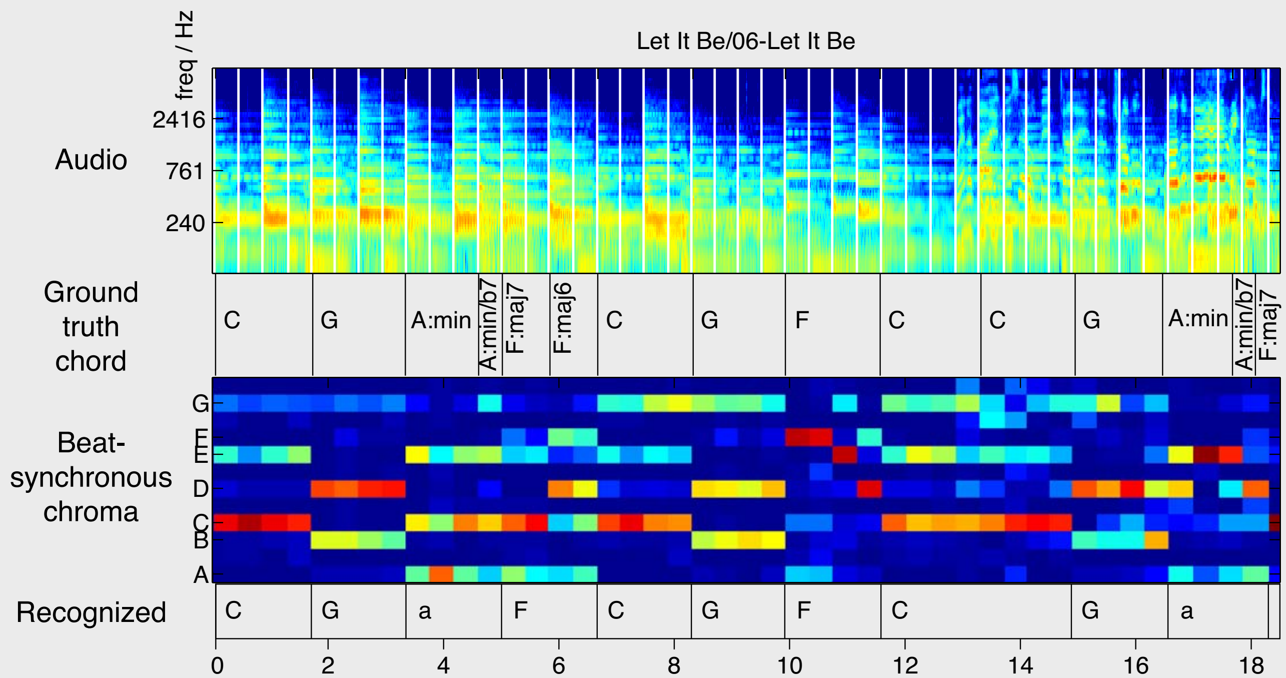
Key Normalization

- Chord transitions depend on key of piece
  - dominant, relative minor, etc...
- Chord transition probabilities should be key-relative
  - estimate main key of piece
  - rotate all chroma features
  - learn models



Chord Recognition

- Often works:



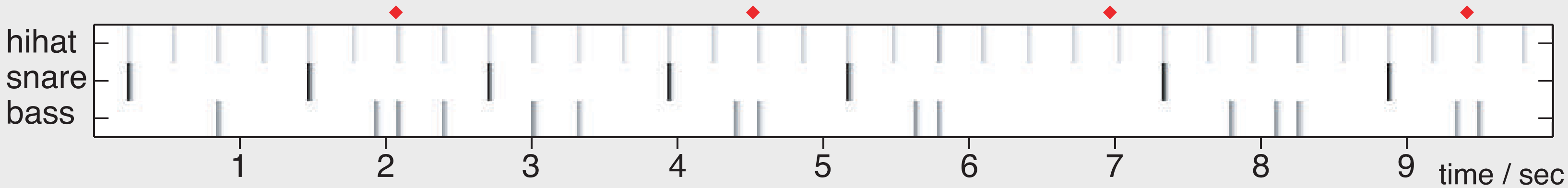
- But not always:

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

Eigenrhythms: Drum Pattern Space

Ellis & Arroyo '04

- Pop songs built on repeating "drum loop"
  - variations on a few bass, snare, hi-hat patterns



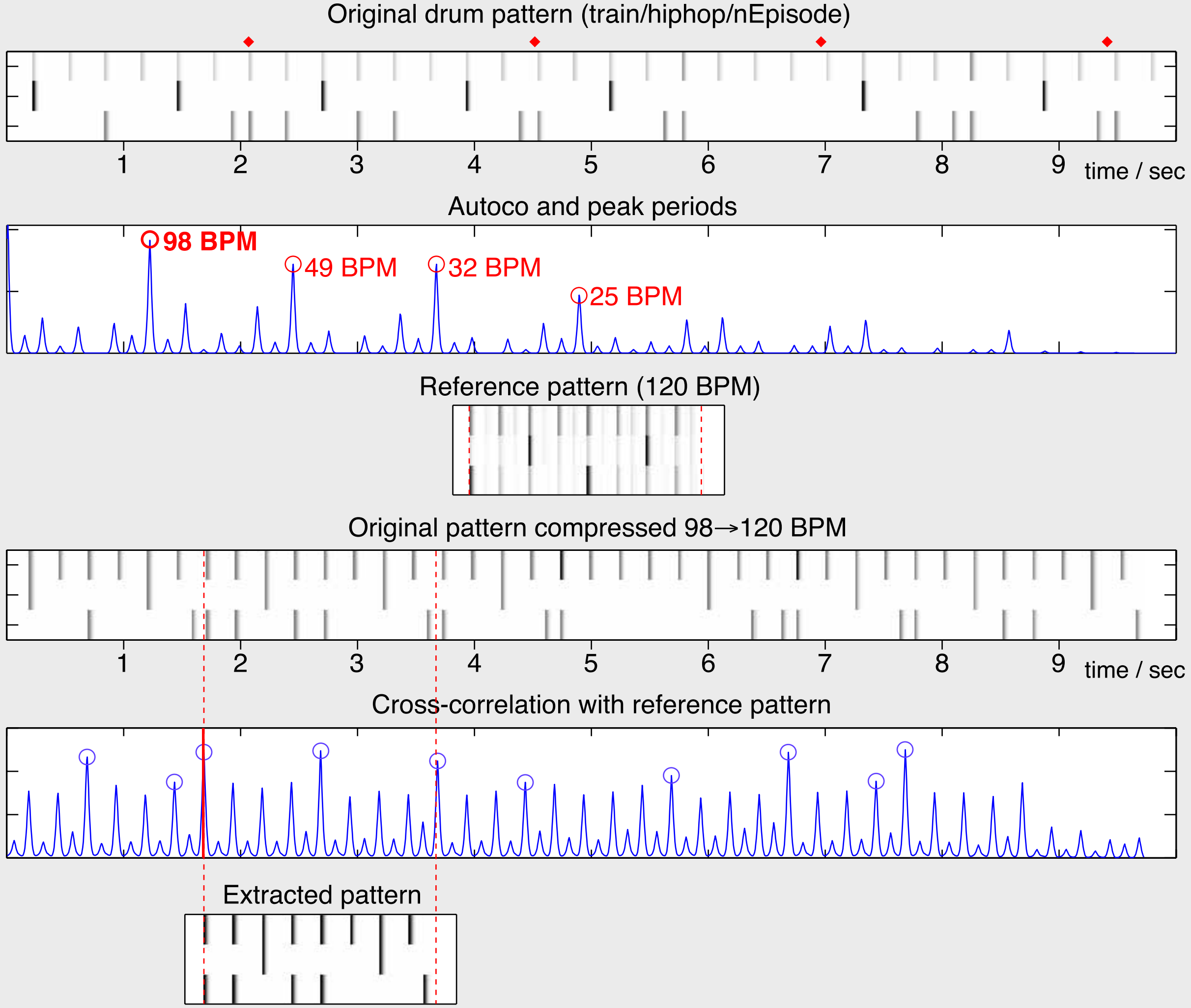
- Eigen-analysis (or ...) to capture variations?
  - by analyzing lots of (MIDI) data, or from audio
- Applications
  - music categorization
  - "beat box" synthesis
  - insight

Aligning the Data

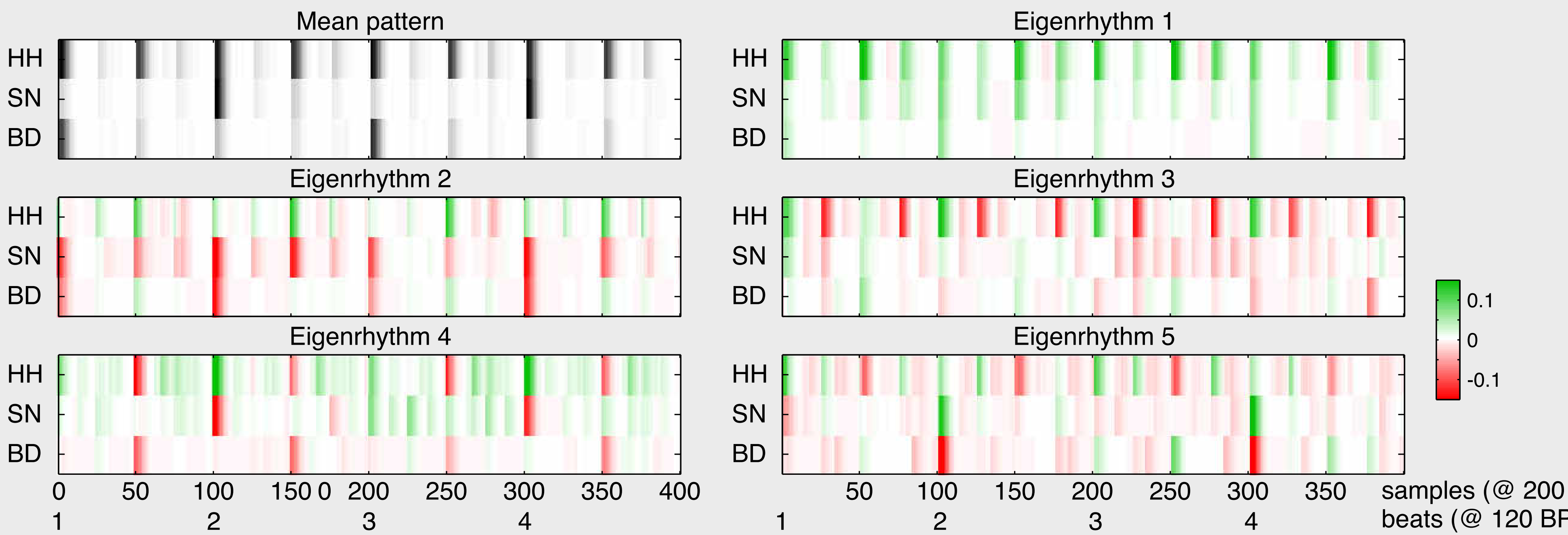
- Need to align patterns prior to modeling...

tempo (stretch): by inferring BPM & normalizing

downbeat (shift): correlate against 'mean' template

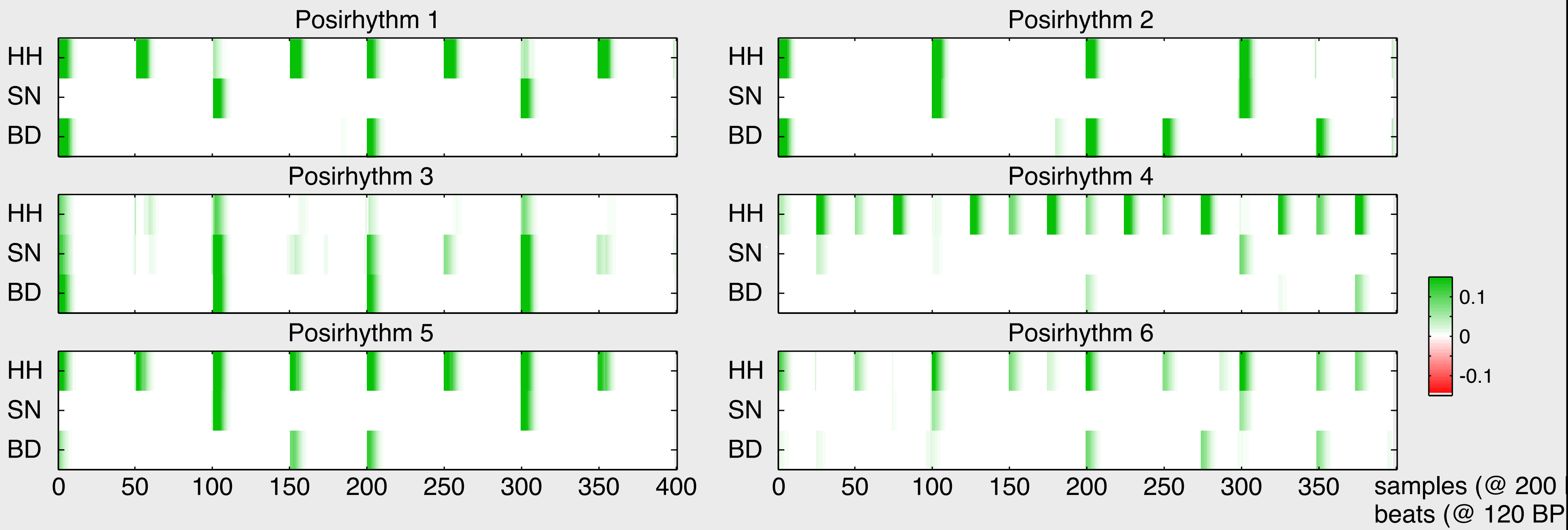


Eigenrhythms (PCA)



- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Eigenrhythms both add and subtract

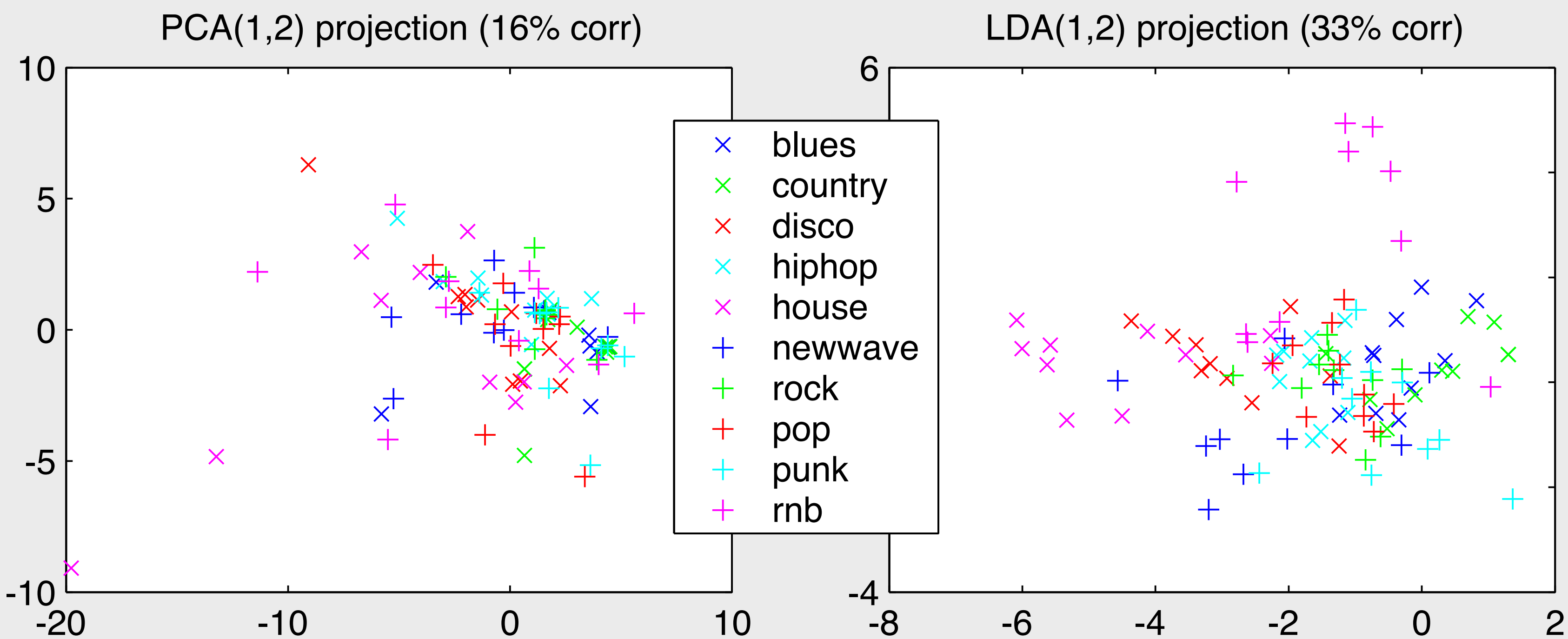
Posirhythms (NMF)



- Nonnegative: only adds beat-weight
- Capturing some structure...

Eigenrhythms for Classification

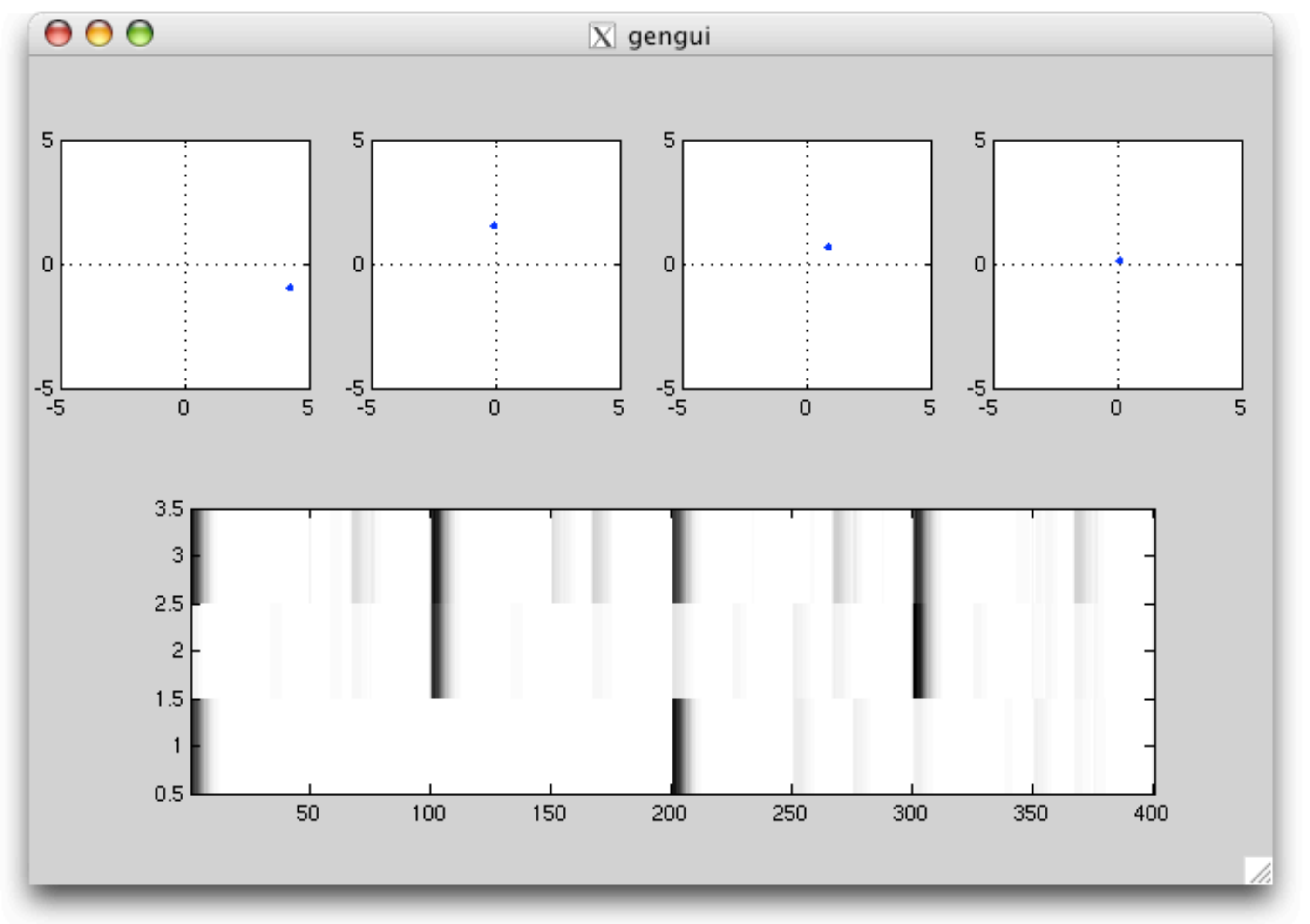
- Projections in Eigenspace / LDA space



- 10-way Genre classification (nearest nbr):
  - PCA3: 20% correct
  - LDA4: 36% correct

Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space



4. Large Music Audio Datasets

Bertin-Mahieux et al '11

- Music Information Retrieval (MIR) is a vibrant new field
  - many commercial opportunities
- But: music audio is hard to share
  - copyright owners have been burned
  - researchers use personal collections...
- Idea: Million Song Dataset (MSD)
  - commercial scale
  - available to all
  - many different "facets"

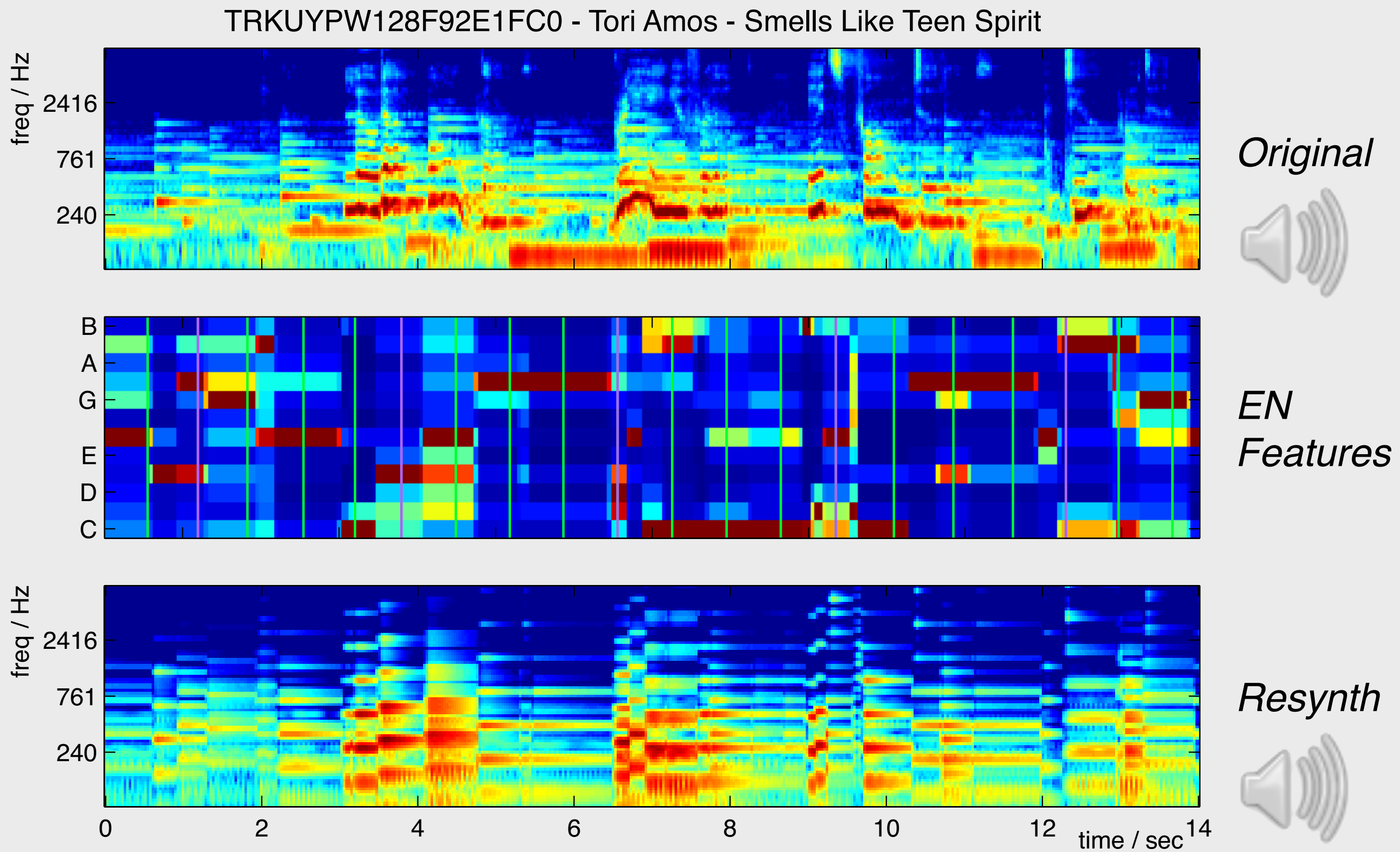
  - http://labrosa.ee.columbia.edu/millionsong

MSD Facets

- Features, Lyrics, Tags, Covers, Listeners ...

MSD Audio Features

- Use Echo Nest "Analyze" features
  - segment audio into variable-length "events"
  - represent by 12 chroma + 12 "timbre"
  - supports a crude resynthesis:



MSD Metadata

EN Metadata

```
         artist: 'Tori Amos'
        release: 'LIVE AT MONTREUX'
          title: 'Smells Like Teen Spirit'
             id: 'TRKUYPW128F92E1FC0'
            key: 5
           mode: 0
       loudness: -16.6780
          tempo: 87.2330
 time_signature: 4
       duration: 216.4502
    sample_rate: 22050
      audio_md5: '8'
     7digitalid: 5764727
    familiarity: 0.8500
           year: 1992
```

Last.fm Tags

```
100.0 — cover                   5.0 — cover songs
57.0 — covers                   4.0 — soft rock
43.0 — female vocalists         4.0 — nirvana cover
42.0 — piano                    4.0 — Mellow
34.0 — alternative              4.0 — alternative rock
14.0 — singer-songwriter        3.0 — chick rock
11.0 — acoustic                 3.0 — Ballad
8.0 — tori amos                 3.0 — Awesome Covers
7.0 — beautiful                 2.0 — melancholic
6.0 — rock                      2.0 — k00l ch1x
6.0 — pop                       2.0 — indie
6.0 — Nirvana                   2.0 — female vocalistist
6.0 — female vocalist           2.0 — female
6.0 — 90s                       2.0 — cover song
5.0 — out of genre covers       2.0 — american
```

SHS Covers

```
%5489,4468, Smells Like Teen Spirit
TRTUOVJ128E078EE10 Nirvana
TRFZJOZ128F4263BE3 Weird Al Yankovic
TRJHCKN12903CDD274 Pleasure Beach
TRELTOJ128F42748B7 The Flying Pickets
TRJKBXL128F92F994D Rhythms Del Mundo feat. Shanade
TRIHLAW128F429BBF8 The Bad Plus
TRKUYPW128F92E1FC0 Tori Amos
```

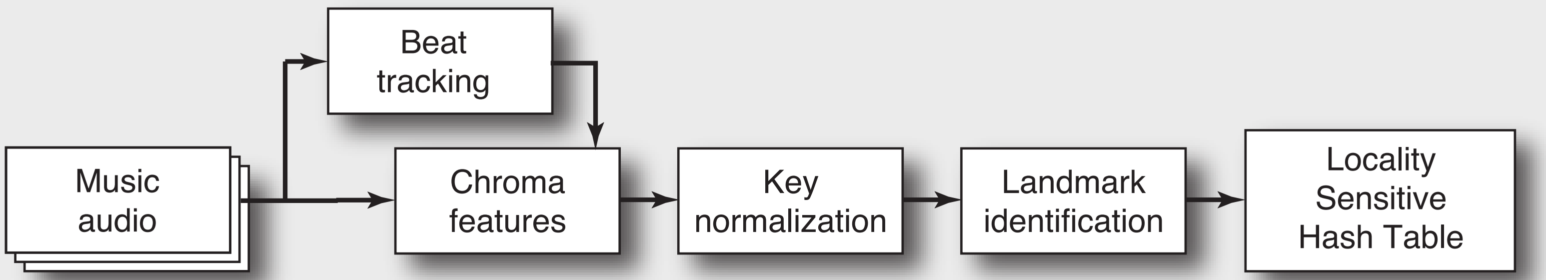
MxM Lyric Bag-of-Words

```
12 hello          6 here          3 is
11 i              6 us            3 with
10 a              6 entertain     3 oh
 9 and            4 the           3 out
 7 it             4 feel          3 an
 6 are            4 yeah          3 light
 6 we             3 to            3 less
 6 now            3 my            3 danger
```

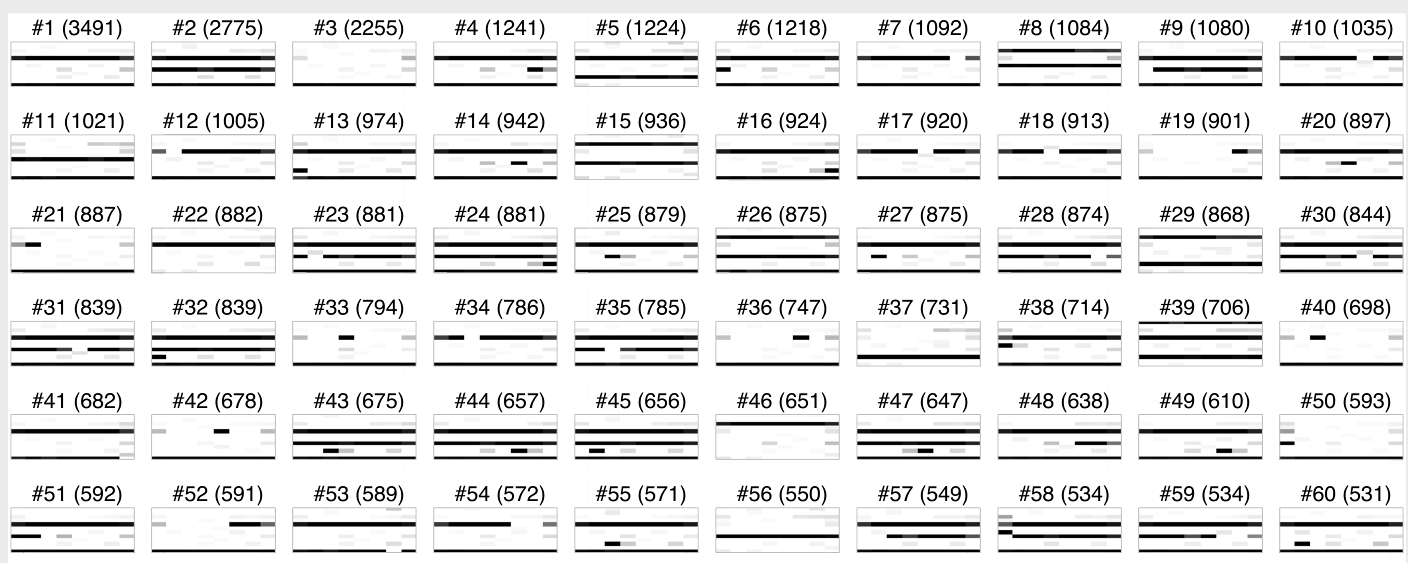
Melodic-Harmonic Mining

Bertin-Mahieux et al. '10

- What can you find in a million songs?
  - what characterizes the content?

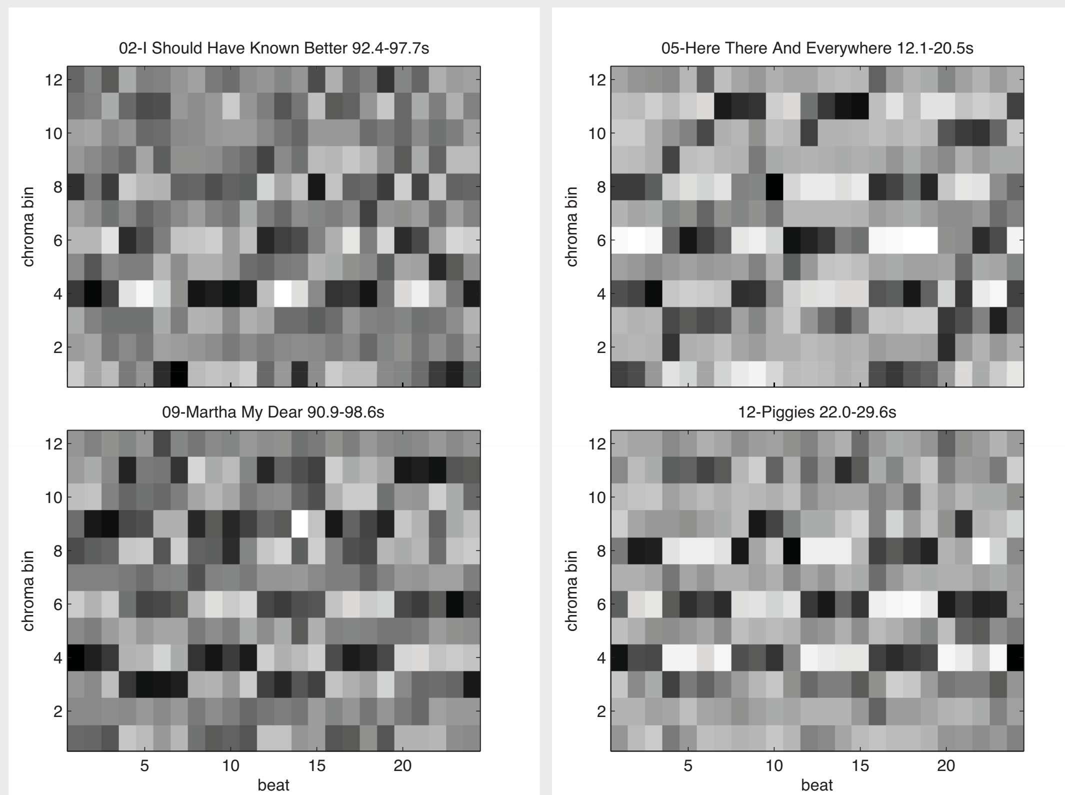


- Frequent clusters of 12 x 8 binarized event-chroma



Results - Beatles

- Over 86 Beatles tracks
- All beat offsets = 41,705 patches
  - LSH takes 300 sec - approx NlogN in patches?
- High-pass along time
  - to avoid sustained notes
- Song filter
  - remove hits in same track



5. Outstanding Issues

- Perceptually Inspired?
  - Music Perception is complex:
    Structure
    Expectation
    Memory
    Enjoyment

- **Many problems still to solve**
  - structure
  - metrical hierarchy
  - music similarity & preference

Summary

- Machine Listening:
  Getting useful information from sound
- Musical sound
  ... constructed to confound scene analysis?
- Transcription tasks
  ... recover notes, beats, chords etc.
- Million Song Dataset for research
  ... large-scale, multiple facets

References

- M. A. Bartsch & G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," *Proc. IEEE WASPAA*, 15-18, Mohonk, 2001.
- J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. B. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Tr. Speech and Audio Proc.*, 13(5):1035-1047, 2005.
- T. Bertin-Mahieux, D. Ellis, B. Whitman, & P. Lamere, "The Million Song Dataset," *Int. Symp. Music Inf. Retrieval ISMIR*, Miami, 2011. A. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.
- P. Desain & H. Honing, "Computational models of beat induction: The rule-based approach," *J. New Music Research*, 28(1):29-42, 1999. J. M. Grey, *An Exploration of Musical Timbre*, Ph.D. Dissertation, Stanford CCRMA, No. STAN-M-2, 1975.
- S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Research*, 30(1):39-58, 2001.
- D. Ellis, "Beat Tracking by Dynamic Programming," *J. New Music Research*, 36(1):51-60, 2007.
- D. Ellis & J. Arroyo, "Eigenrhythms: Drum pattern basis sets for classification and generation," *Int. Symp. Music Inf. Retrieval ISMIR*, 101-106, Barcelona, 2004.
- D. Ellis & G. Poliner, "Identifying Cover Songs With Chroma Features and Dynamic Programming Beat Tracking," *Proc. ICASSP*, IV-1429-1432, Hawai'i, 2007.
- T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," *Proc. Int. Comp. Music Conf.*, 464–467, Beijing, 1999.
- M. Goto, "A Robust Predominant-F0 Estimation Method for Real-time Detection of Melody and Bass Lines in CD Recordings," *Proc. ICASSP,* II-757-760, 2000.
- M. Goto, "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds," *J. New Music Research*, 30(2):159-171, 2001.

References

- G. Grindlay & D. Ellis, "Transcribing Multi-instrument Polyphonic Music with Hierarchical Eigeninstruments", *IEEE J. Sel. Topics in Sig. Process.*, 5(6):1159-1169, 2011.
- A. Klapuri, "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes", *Proc. Int. Symp. Music IR*, 216-221, 2006.
- R. Maher & J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Am*. 95(4):2254-2263, 1994.
- M. F. McKinney & D. Moelants, "Ambiguity in tempo perception: What draws listeners to different metrical levels?" *Music Perception*, 24(2):155–166, 2006.
- J. R. Pierce, *The Science of Musical Sound*, Scientific American Press, 1983.
- G. Poliner & D. Ellis, "A Discriminative Model for Polyphonic Piano Transcription", *Eurasip Journal of Advances in Signal Processing*, article id 48317, 2007.
- M. Puckette, T. Apel, D. Zicarelli, "Real-time audio analysis tools for Pd and MSP," *Proc. Int. Comp. Music Conf.*, Ann Arbor, 109–112, 1998.
- E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, 103:588-601, 1998.
- A. Sheh & D. Ellis, "Chord Segmentation and Recognition using EM-Trained Hidden Markov Models," *Int. Symp. Music Inf. Retrieval ISMIR*, 185-191, Baltimore, 2003.