LabROSA
Research Overview

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

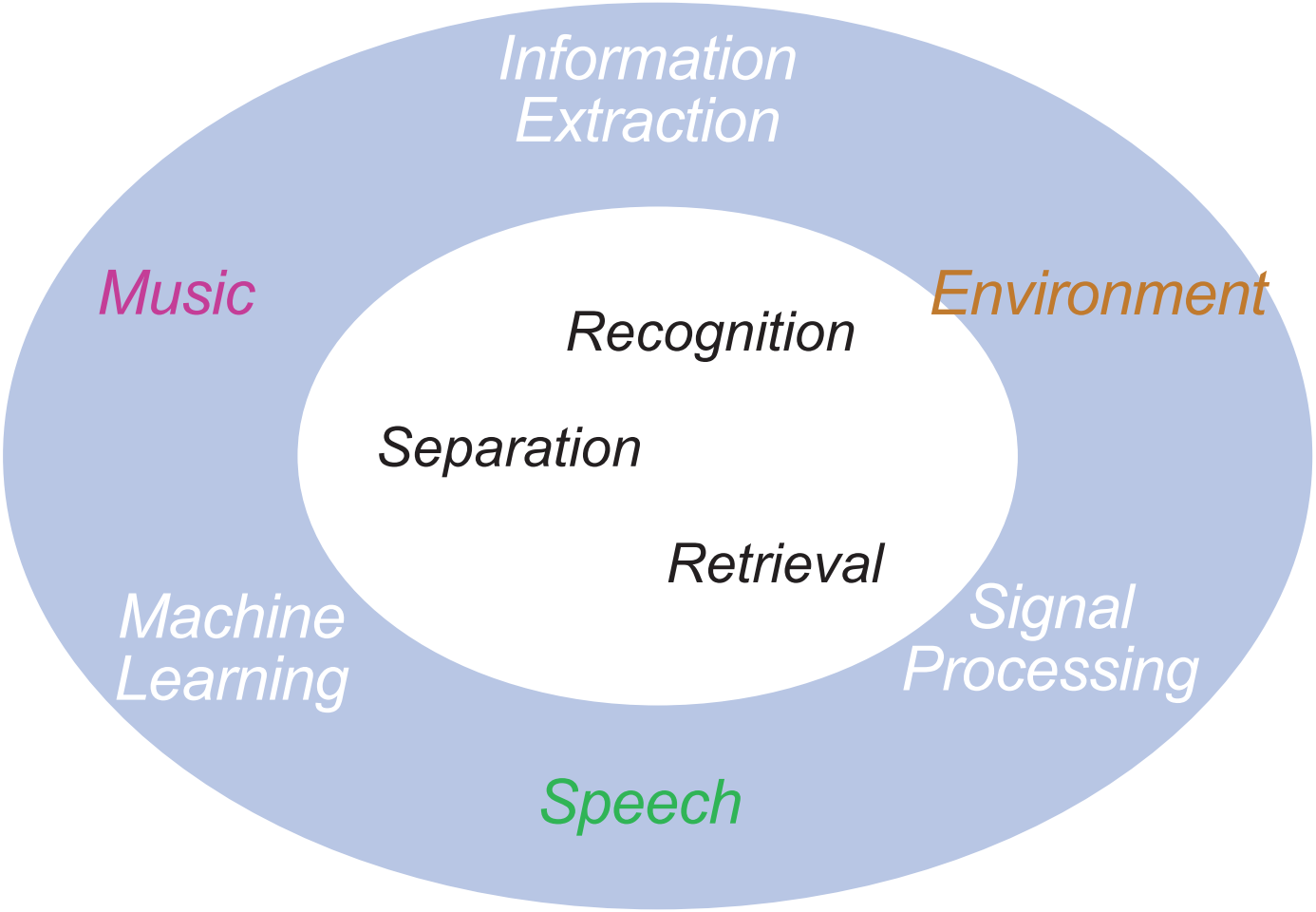
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Real-World Sound
2. Speech Separation
3. Environmental Audio Classification
4. Music Audio Analysis

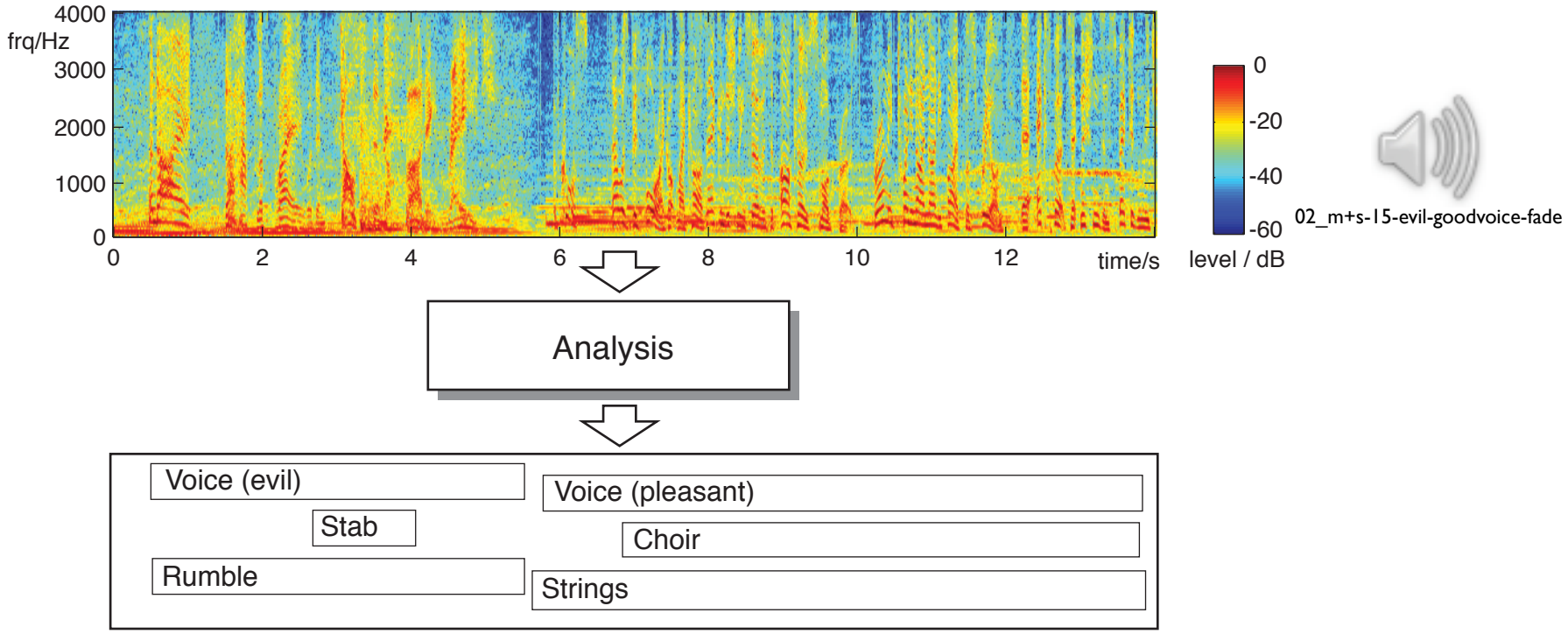
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University
In the City of New York

LabROSA Overview

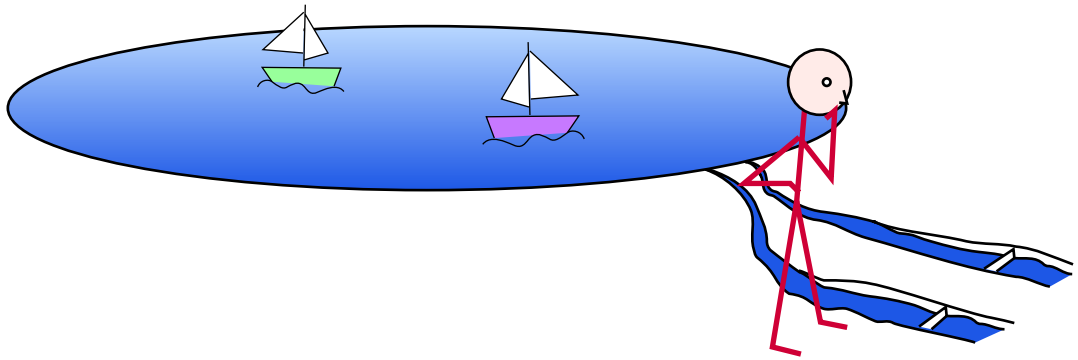


1. Real-World Sound



- Sounds rarely occur in isolation
  - .. so analyzing mixtures ("scenes") is a problem
  - .. for humans and machines

Auditory Scene Analysis



"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- Received waveform is a mixture
  - 2 sensors, N sources - underconstrained
- Use prior knowledge (models) to constrain

2. Speech Separation

Roweis '01, '03
Kristjannson '04, '06

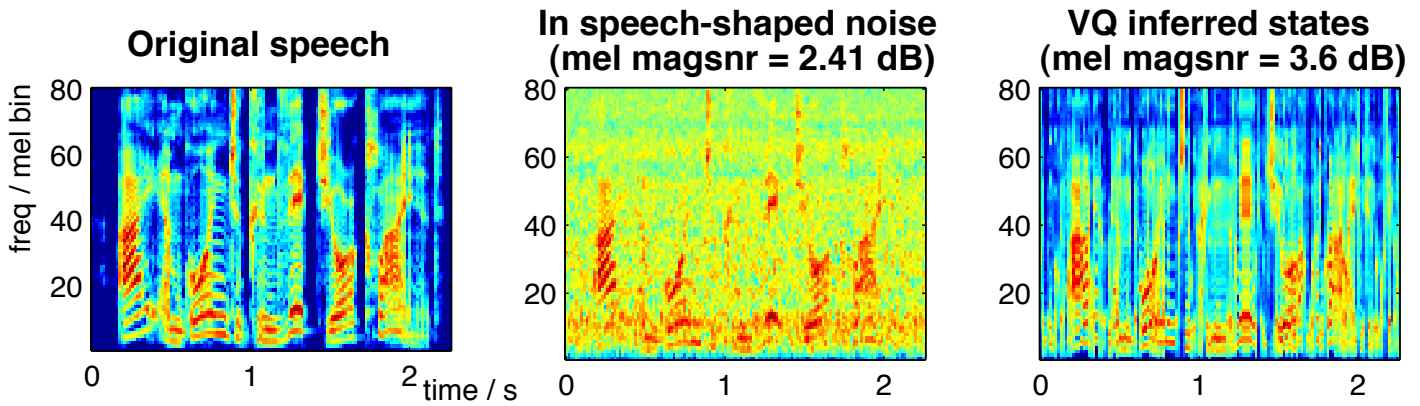
- Given models for sources,
  find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1,i_2) = \mathcal{N}(\mathbf{x};\mathbf{c}_{i1}+\mathbf{c}_{i2},\Sigma)$$ *combination model*

$$\{i_1(t),i_2(t)\} = argmax_{i_1,i_2} p(\mathbf{x}(t)|i_1,i_2)$$ *inference of source state*

  - can include sequential constraints...

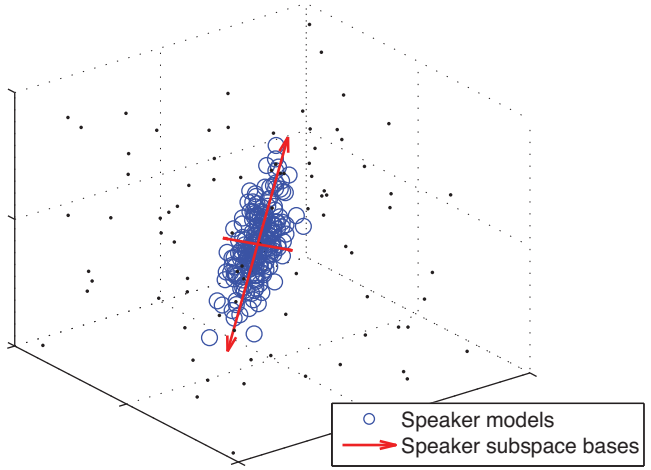
- E.g. stationary noise:



Eigenvoices

Weiss & Ellis '09, '10

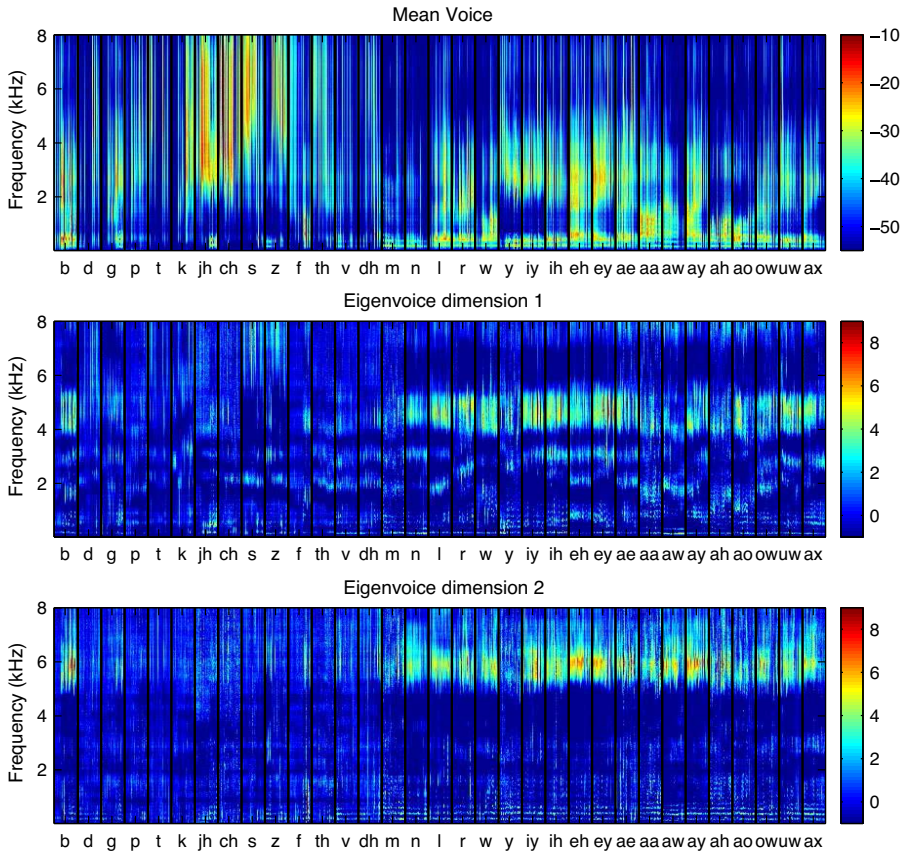
- Idea: Find speaker model parameter space
  - generalize without losing detail?



- Eigenvoice model:
  - 280 states x 320 bins = 89,600 dimensions
  - 10-30 dimensions

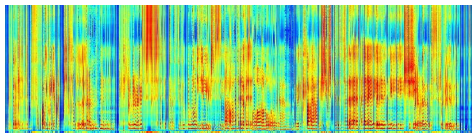
$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} + U\,\mathbf{w}$$

adapted model | mean voice | eigenvoice bases | weights

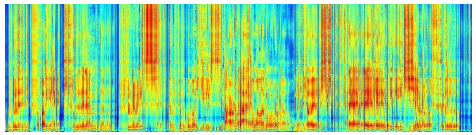


Speaker-Adapted Separation

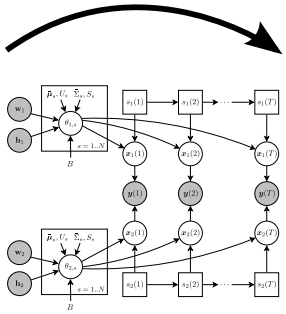
$$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$$

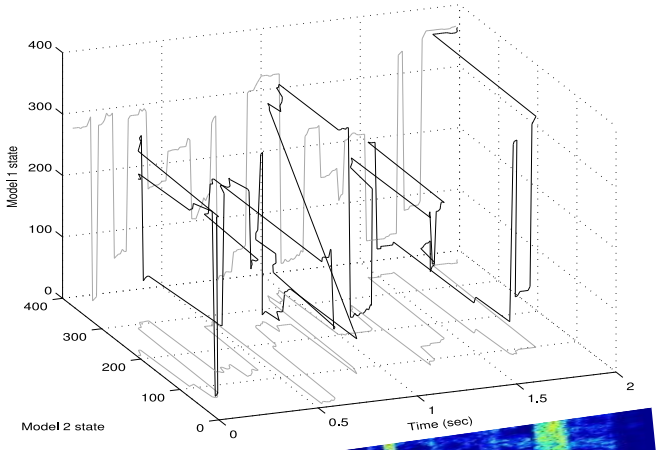


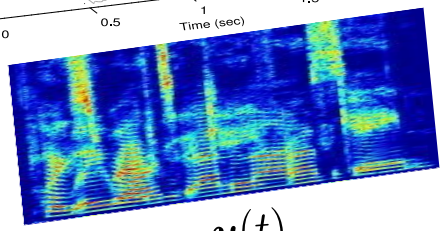
$$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$$



Find Viterbi path

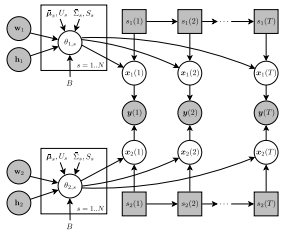




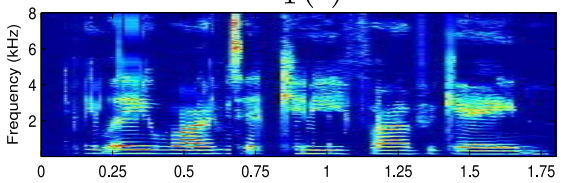


$\boldsymbol{y}(t)$

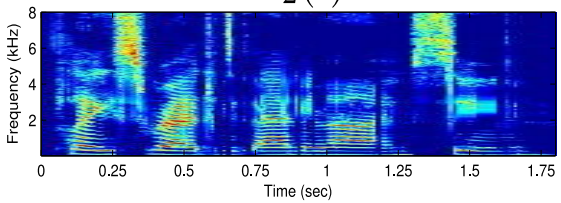
Estimate source signals



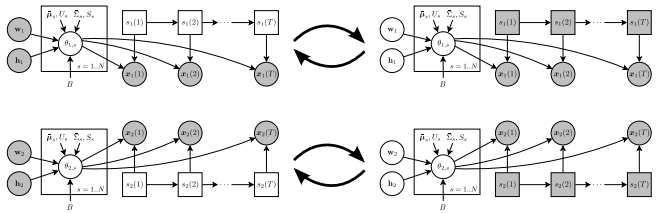
$\hat{\boldsymbol{x}}_1(t)$



$\hat{\boldsymbol{x}}_2(t)$

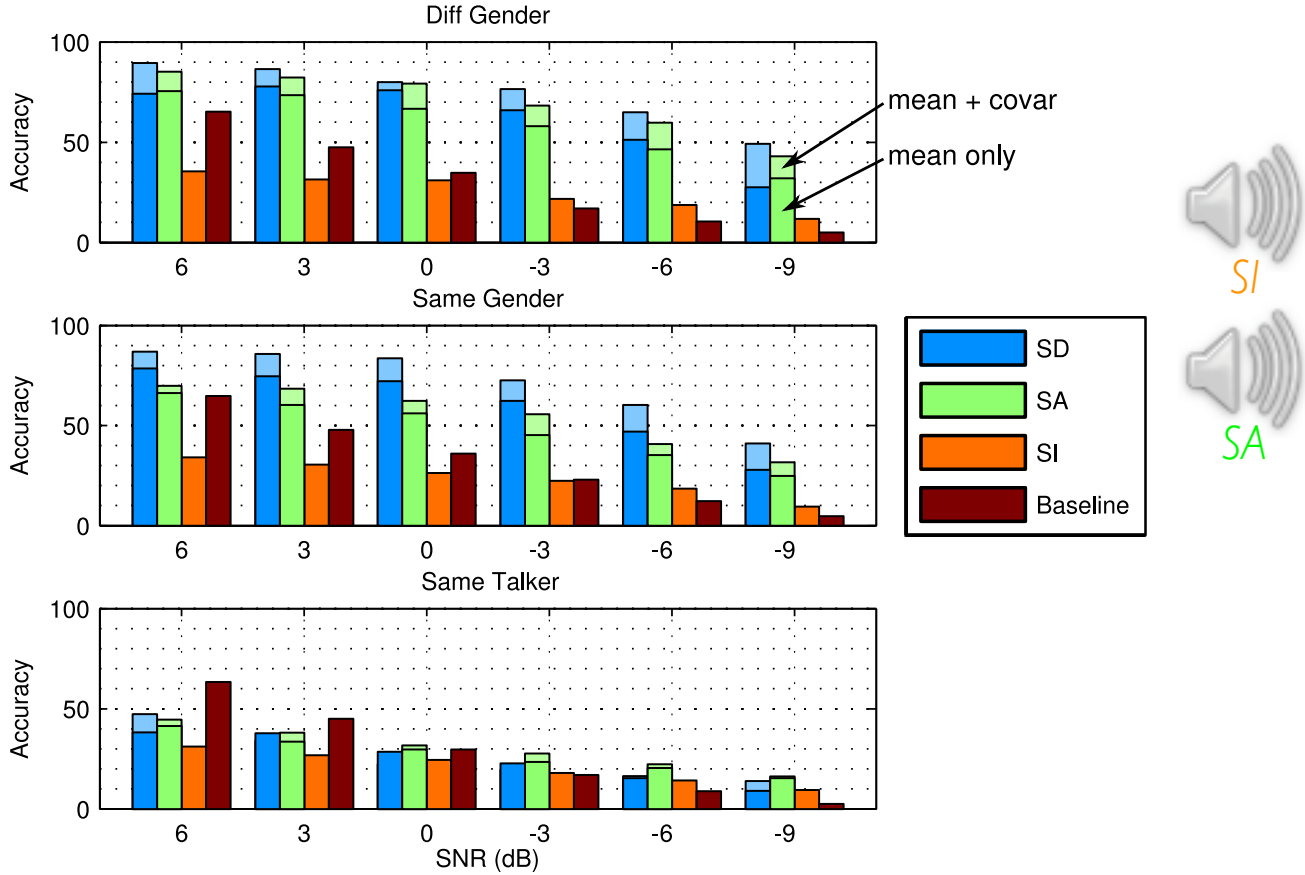


Update model parameters using EM algorithm from Kuhn et al., (2000)



Speaker-Adapted Separation

- Eigenvoices for Speech Separation task
  - speaker adapted (SA) performs midway between speaker-dependent (SD) & speaker-indep (SI)



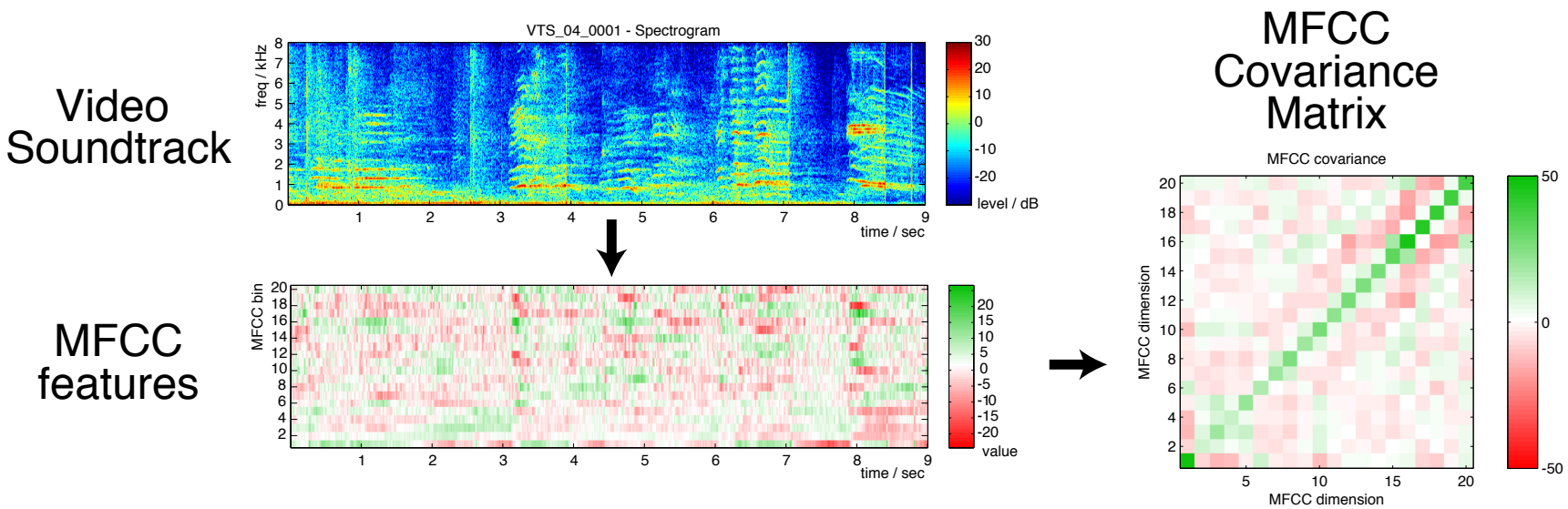
3. Soundtrack Classification

- Short video clips as the evolution of snapshots
  - 10-100 sec, one location, no editing
  - browsing?

- Need information for indexing...
  - video + audio
  - foreground + background

MFCC Covariance Representation

- Each clip/segment → fixed-size statistics
  - similar to speaker ID and music genre classification
- Full Covariance matrix of MFCCs
  - maps the kinds of spectral shapes present

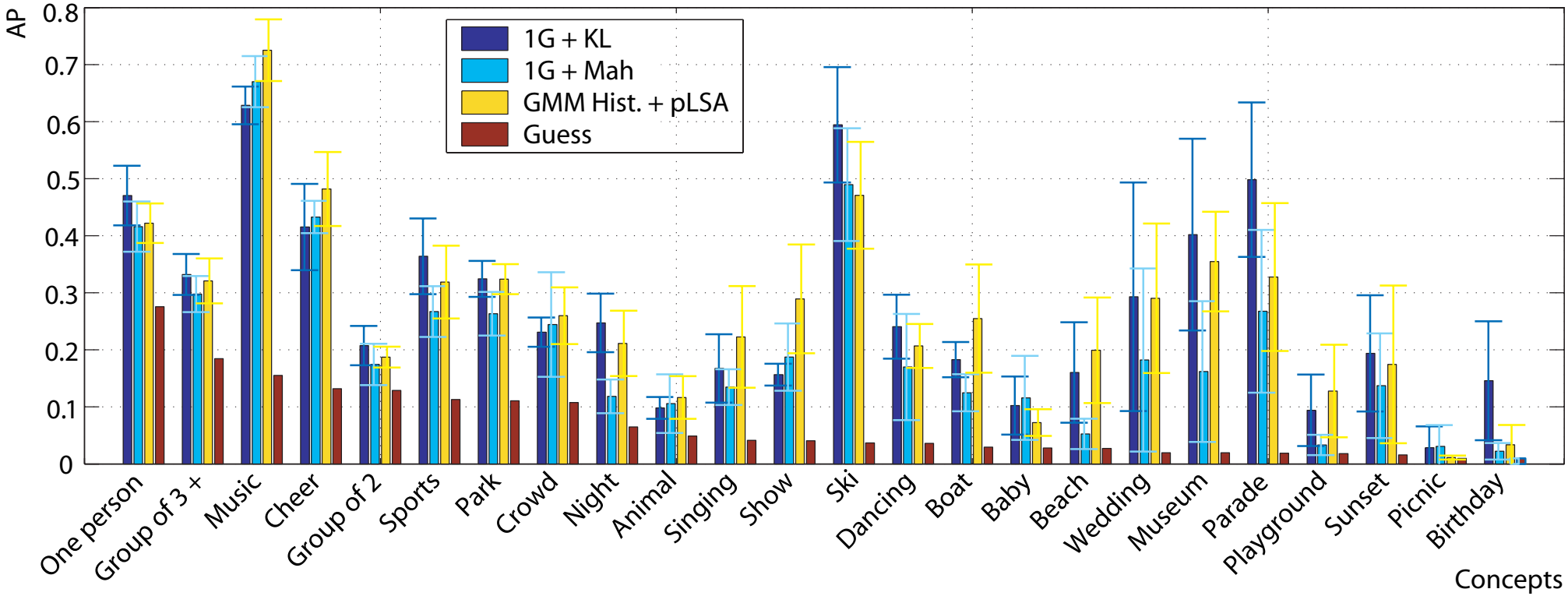


- Clip-to-clip distances for SVM classifier
  - by KL or 2nd Gaussian model

Classification Results

Chang, Ellis et al. '07
Lee & Ellis '10

- Wide range:

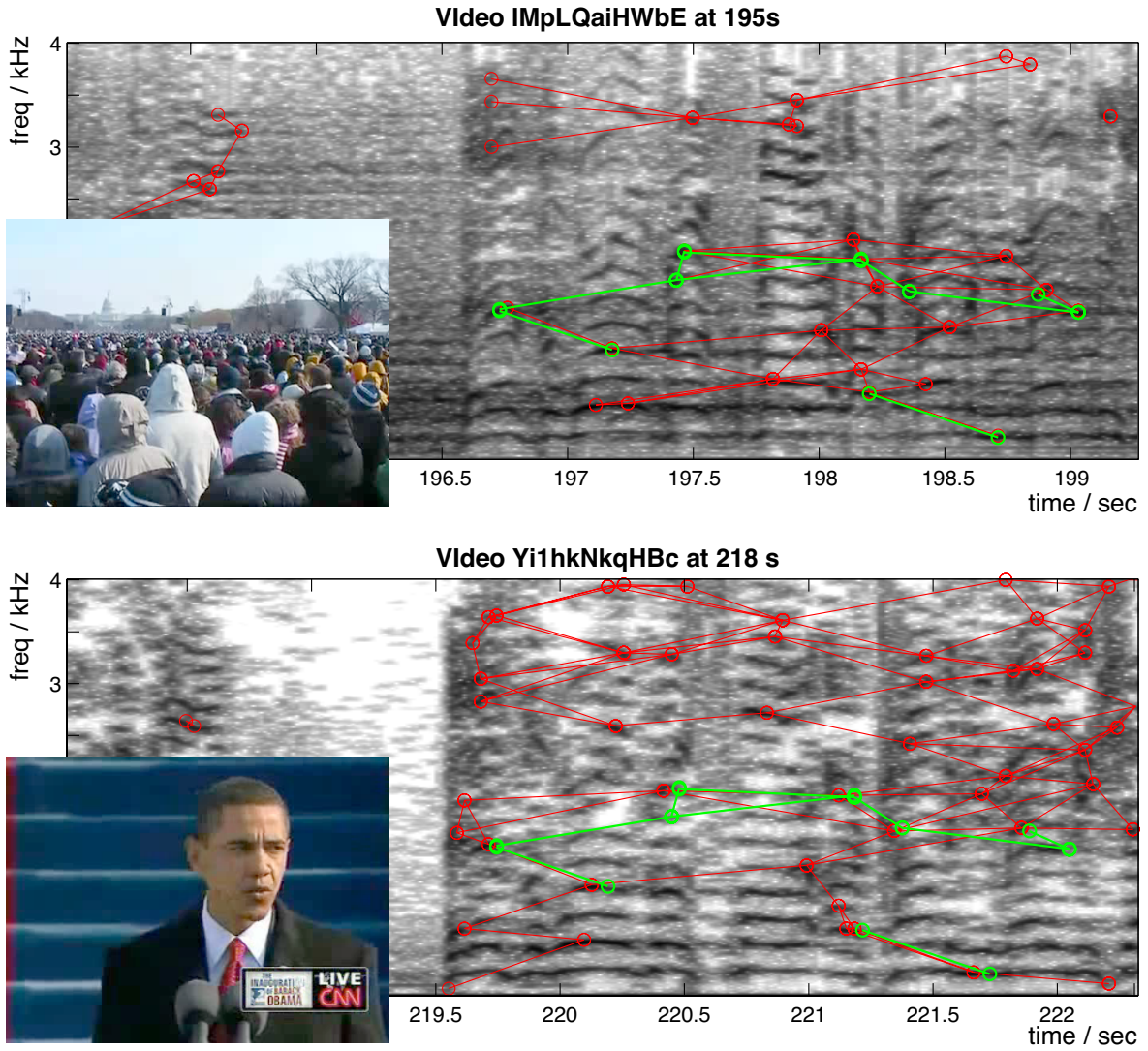


- audio (music, ski) vs. non-audio (group, night)
- large AP uncertainty on infrequent classes

Matching Videos via Fingerprints

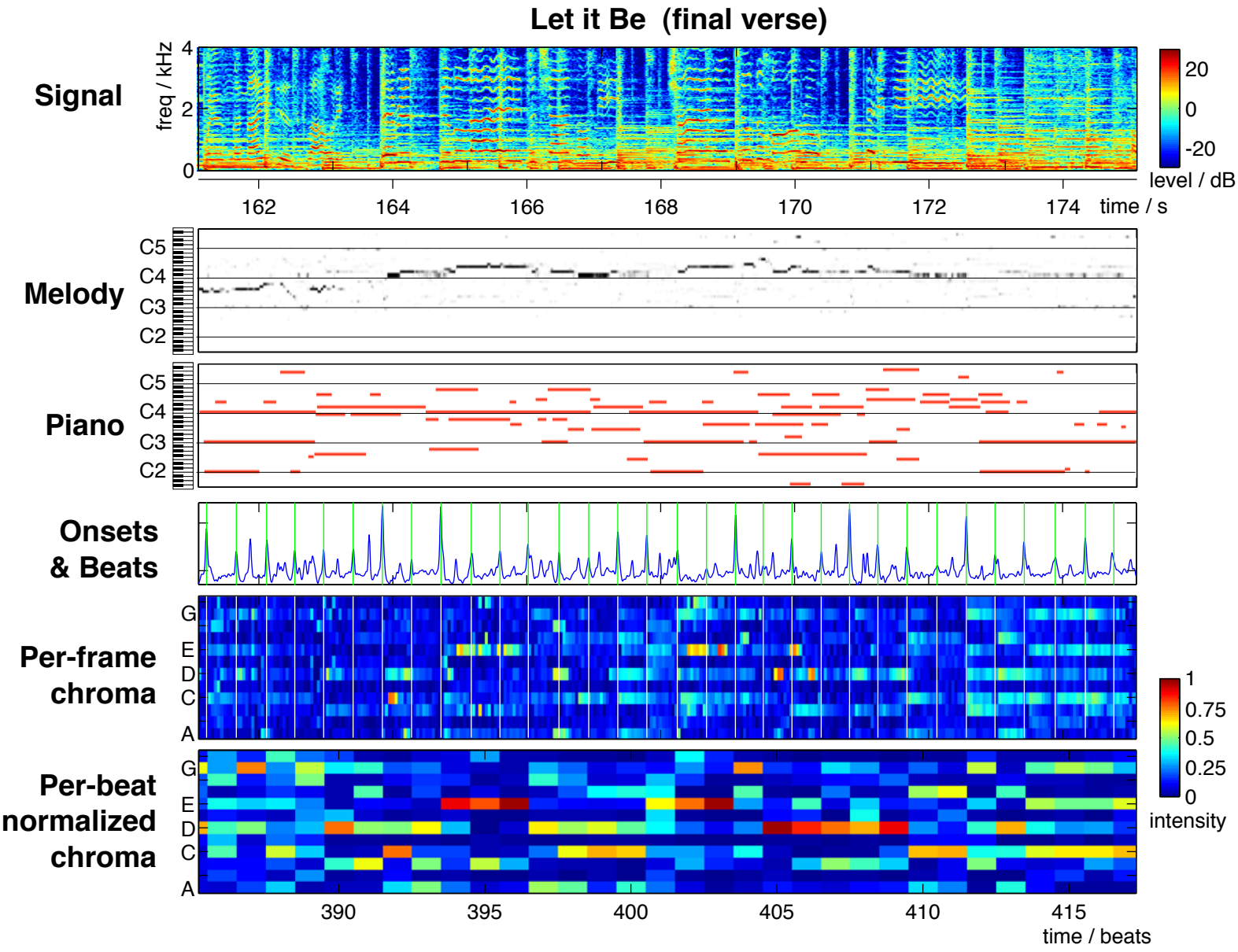
Cotton & Ellis '10

- Landmark pairs are a noise-robust fingerprint
- Use to match distinct videos with same sound ambience



4. Music Audio Analysis

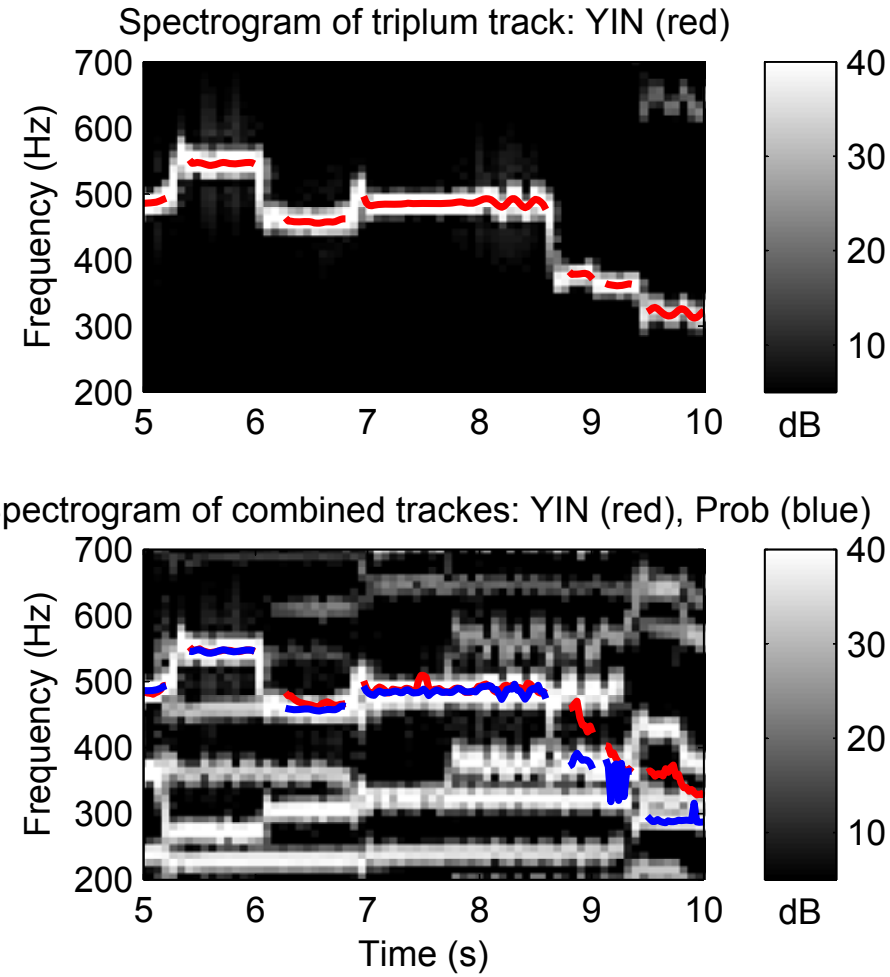
- ... at all levels from notes to genres



High-Resolution Transcription

Smit & Ellis '09

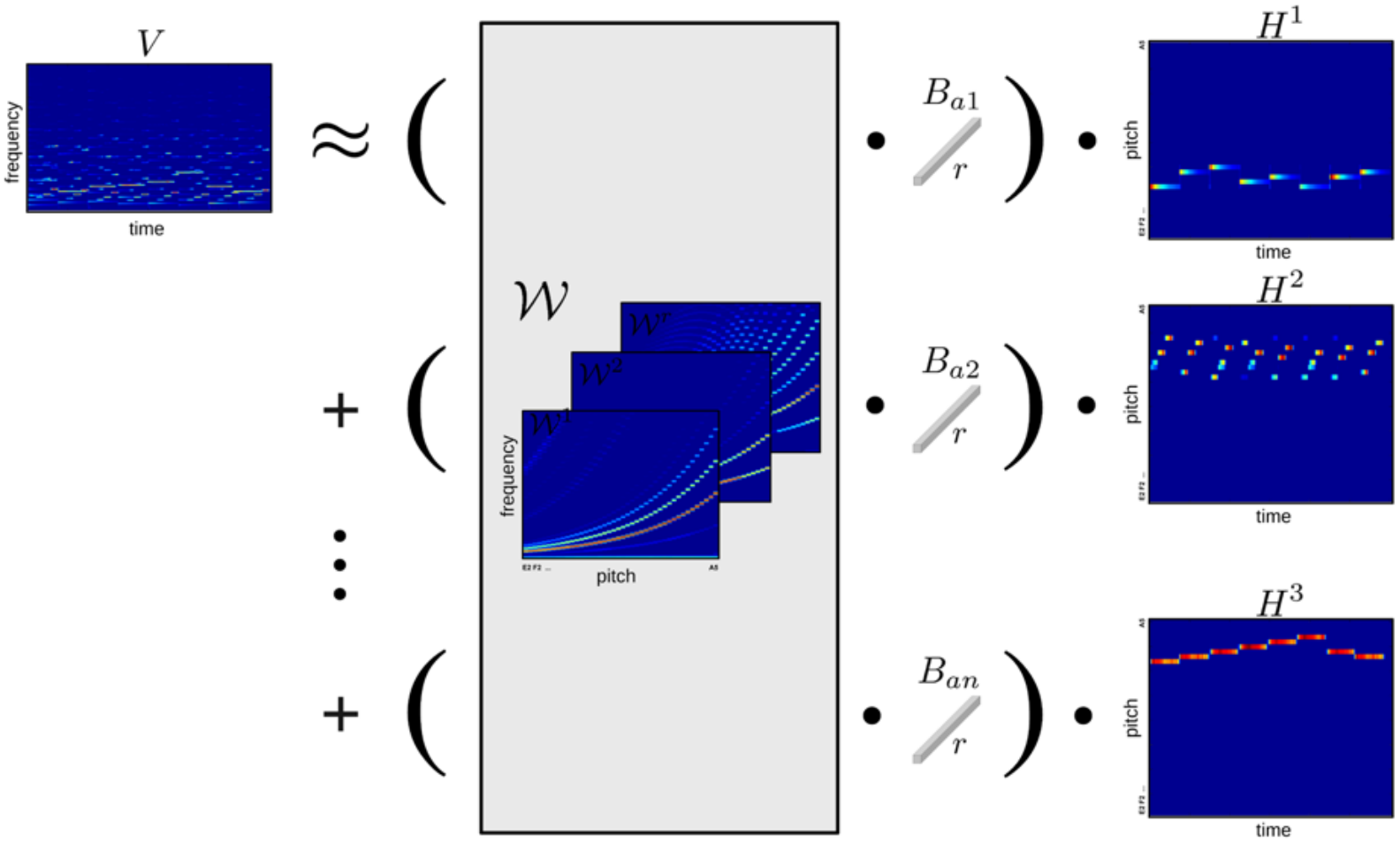
- Extract notes from audio
  - despite overlapping voices
  - with precise tuning, timing
- Use musical score
  - time alignment & approximate match
- Find most likely parameters
  - fundamental, amplitudes of all harmonics



Polyphonic Transcription

Grindlay & Ellis '09

- Apply the Eigenvoice idea to music
  - eigeninstruments?
  - Subspace NMF



Chord Recognition

Weller, Ellis, Jebara '09

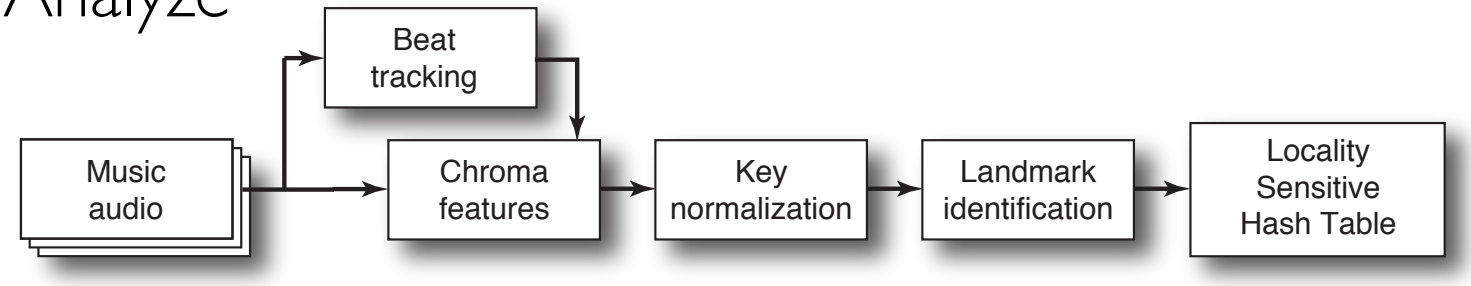
- StructSVM vs. HMM: better results

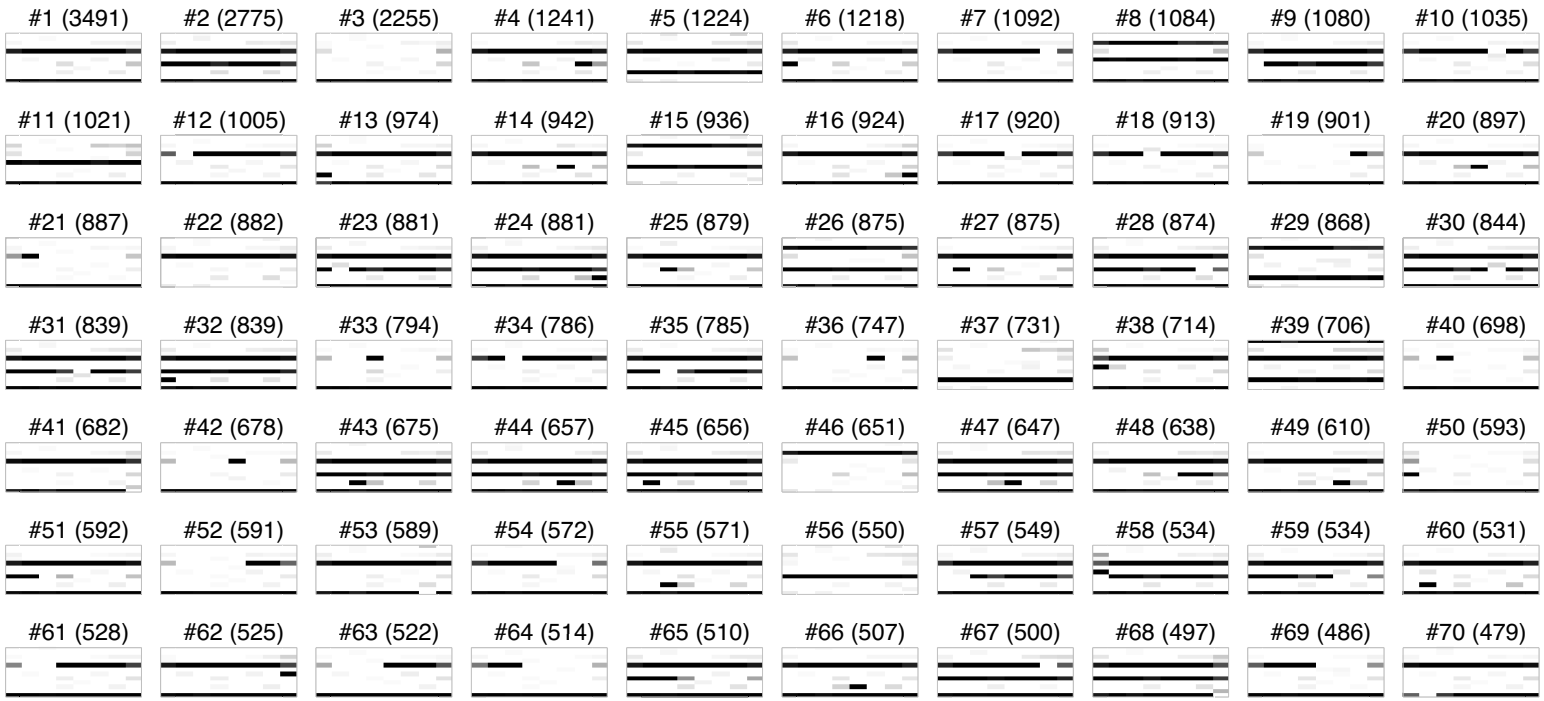
	RUSUSL	WEJ1	WEJ2	WEJ3	WEJ4
Wt. Ave. Overlap Score	0.701	0.704	0.723	0.723	0.742
Wt. Ave. Overlap Score (Merged maj/min)	0.760	0.743	0.762	0.760	0.777

Melodic-Harmonic Mining

Bertin-Mahieux et al.'10

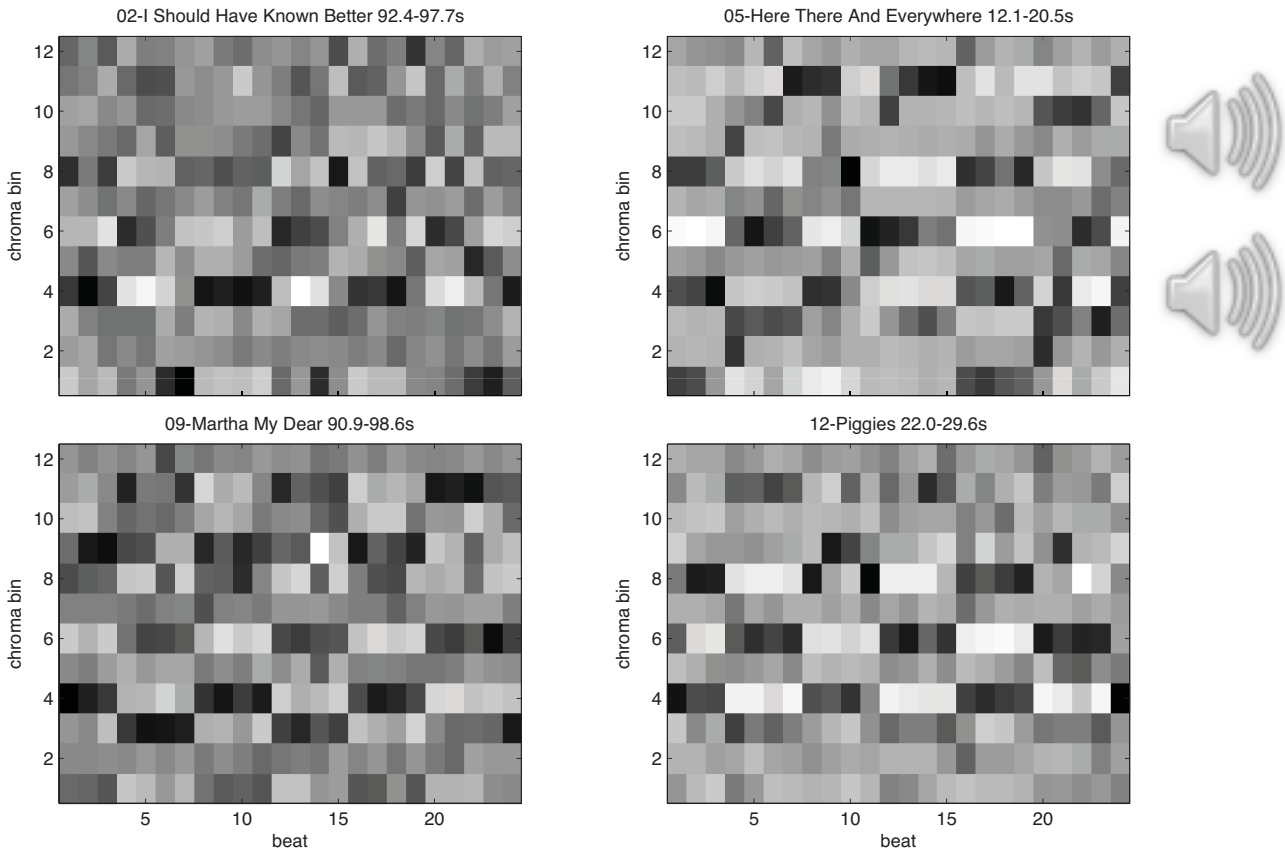
- 100,000 tracks from morecowbell.dj
  - as Echo Nest Analyze
- Frequent clusters of 12 x 8 binarized event-chroma





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



Summary

- LabROSA : getting information from sound
- Speech
  - monaural separation using eigenvoices
  - binaural + reverb using MESSL
- Environmental
  - classification of consumer video
  - landmark-based events and matching
- Music
  - transcription of notes, chords, ...
  - large corpus mining