Research in Sound Analysis

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

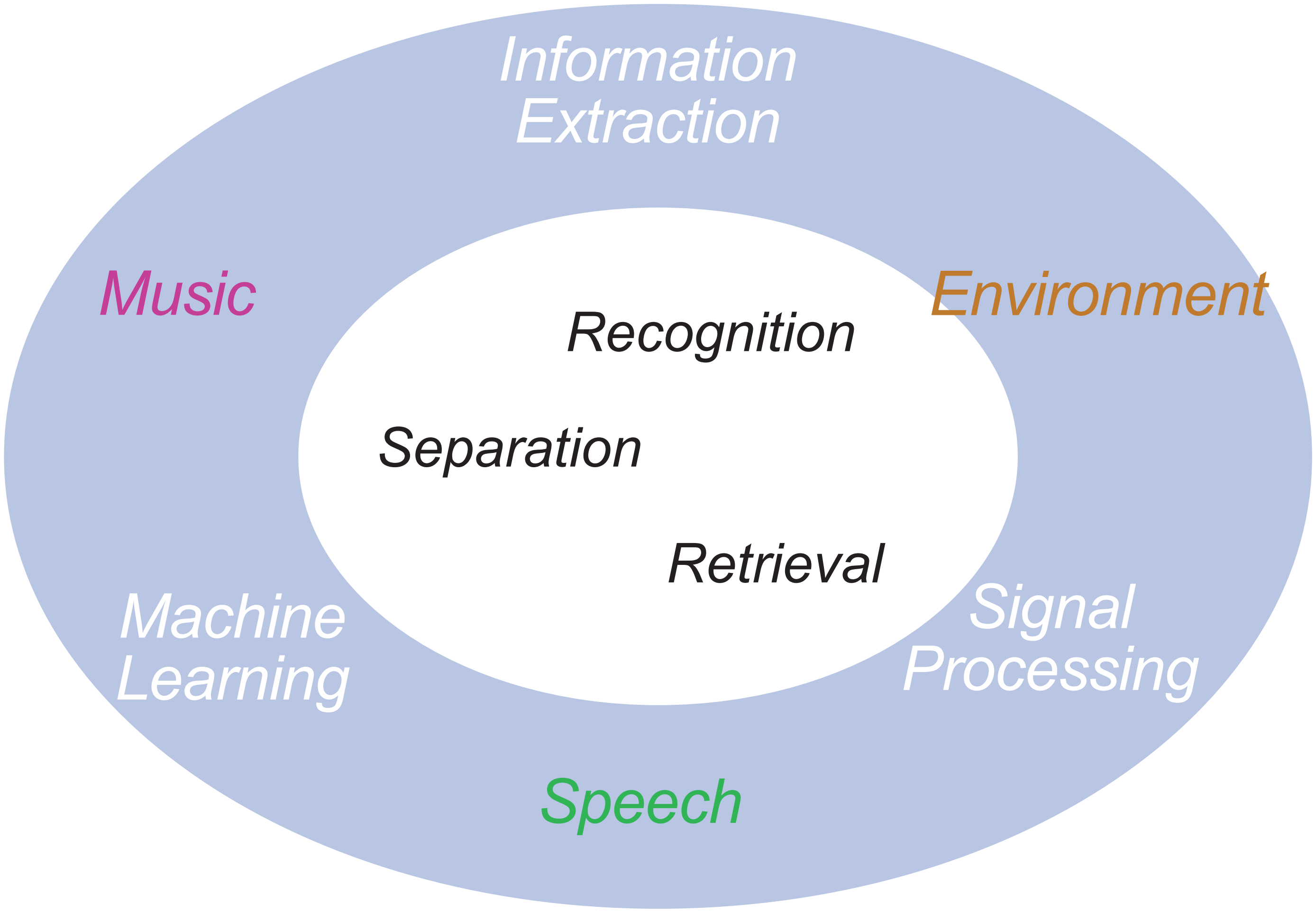
dpwe@ee.columbia.edu

1. Personal and Consumer Audio
2. Musical Cover Song Detection
3. Binaural Source Separation

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

Columbia University
in the City of New York

LabROSA Overview



Information Extraction

Music

Environment

Recognition

Separation

Retrieval

Machine Learning

Signal Processing

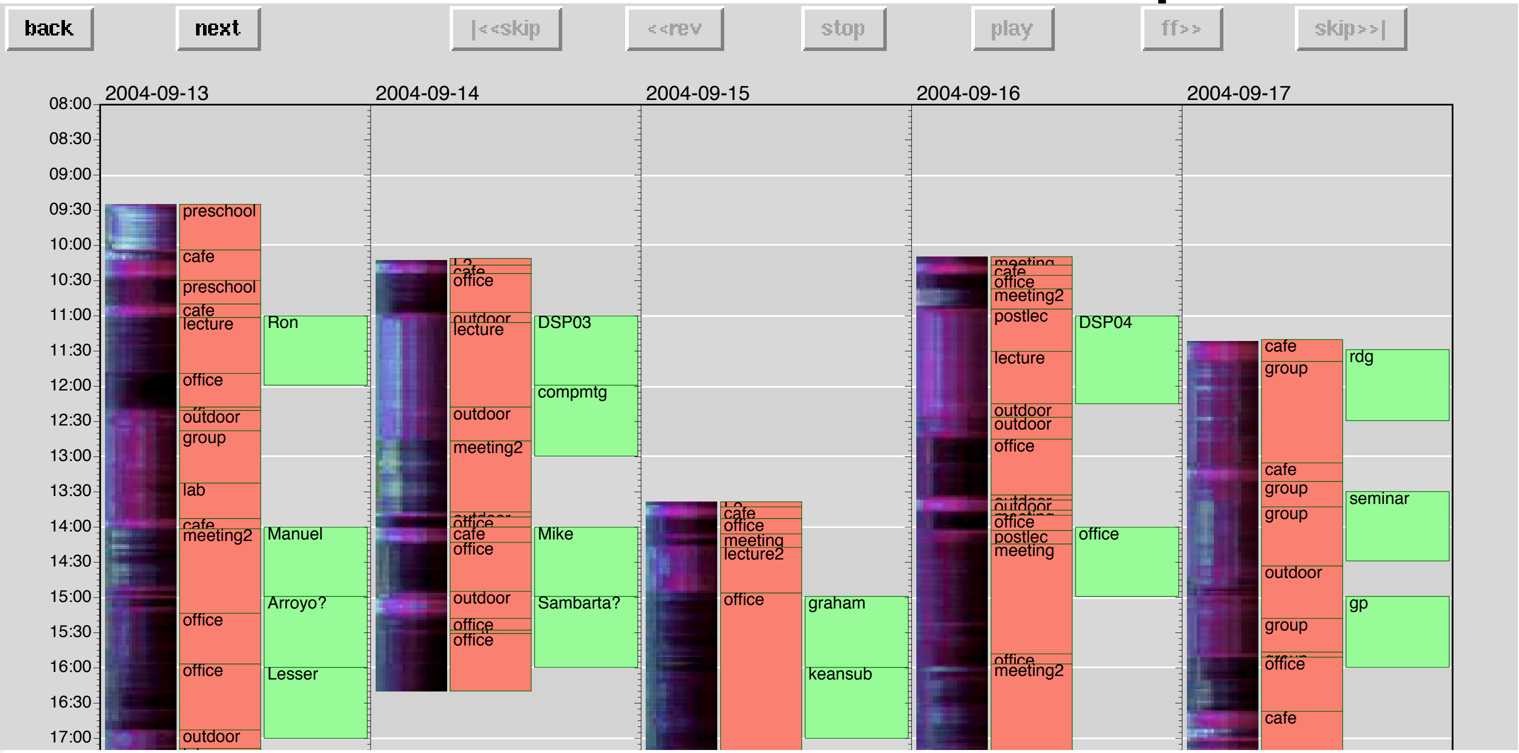
Speech

1. Personal Audio Archives

- Easy to record everything you hear
  - <2GB / week @ 64 kbps
- Hard to find anything
  - how to scan?
  - how to visualize?
  - how to index?
- Need automatic analysis
- Need minimal impact

Browsing Interface

- Browsing / Diary interface
  - links to other information (diary, email, photos)
  - synchronize with note taking? *(Stifelman & Arons)*
  - audio thumbnails
- Release Tools + "how to" for capture



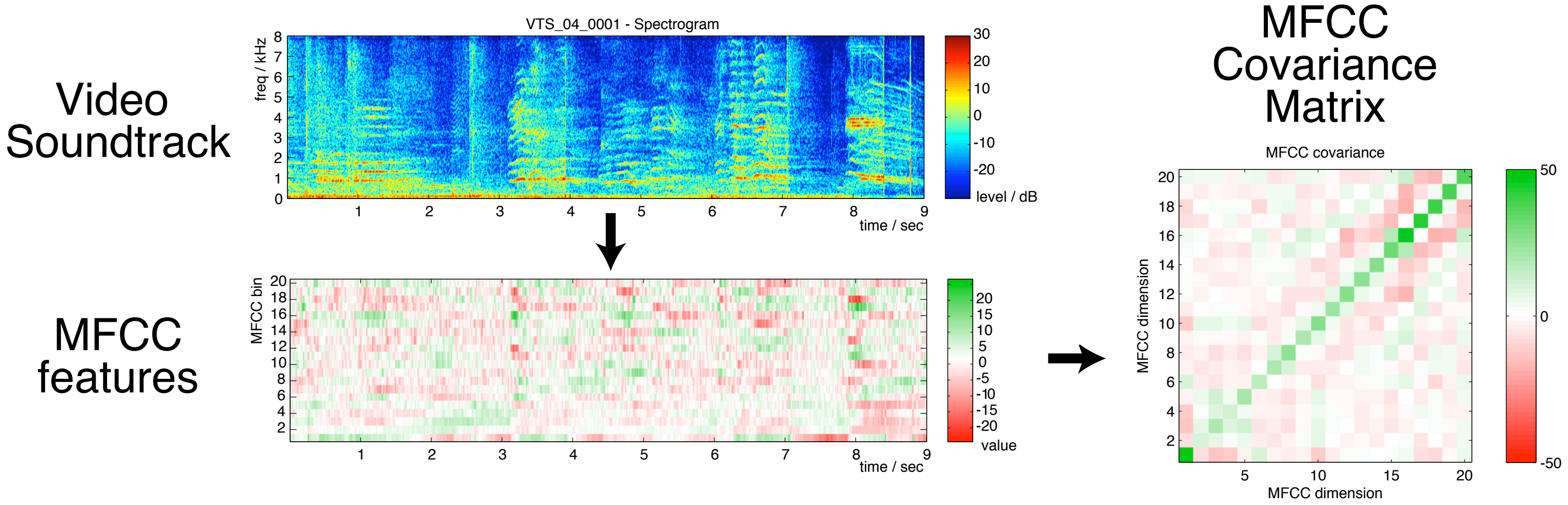
Consumer Video

- Short video clips as the evolution of snapshots
  - 10-60 sec, one location, no editing
  - browsing?

- More 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


Video Soundtrack

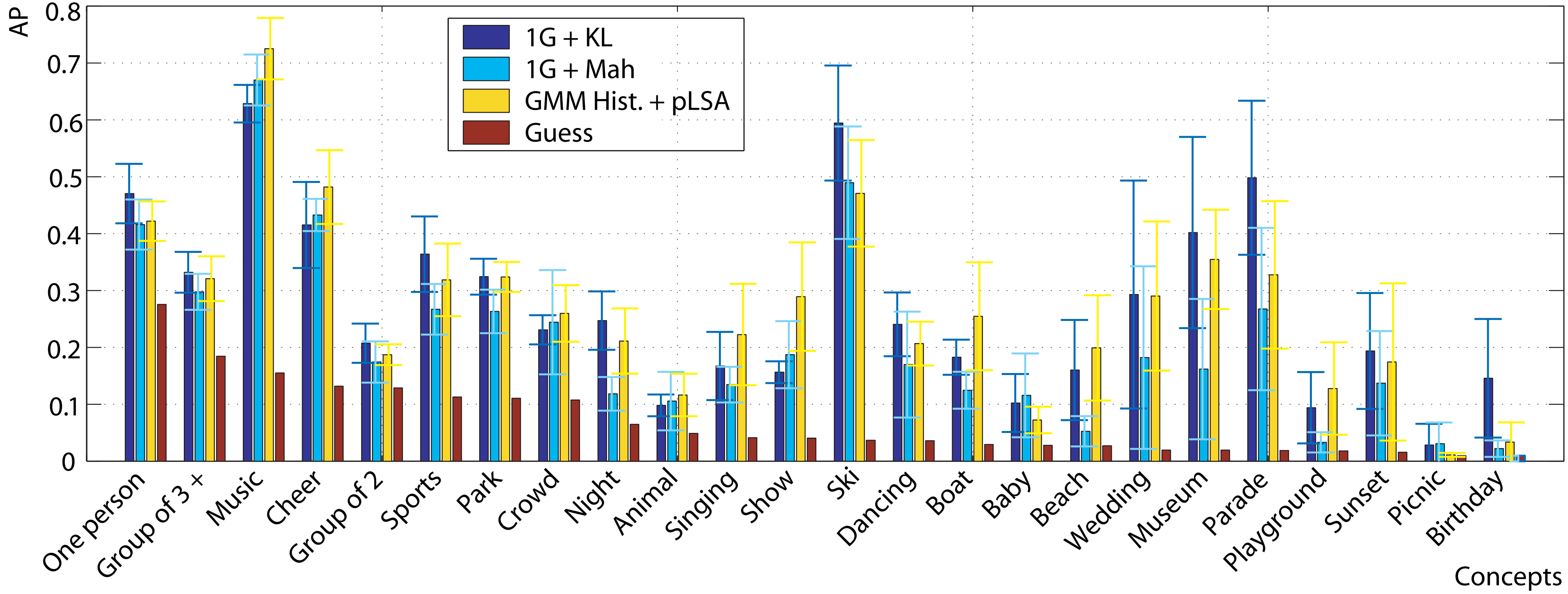
MFCC features

MFCC Covariance Matrix

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

Audio-Only Results

- Wide range of results:

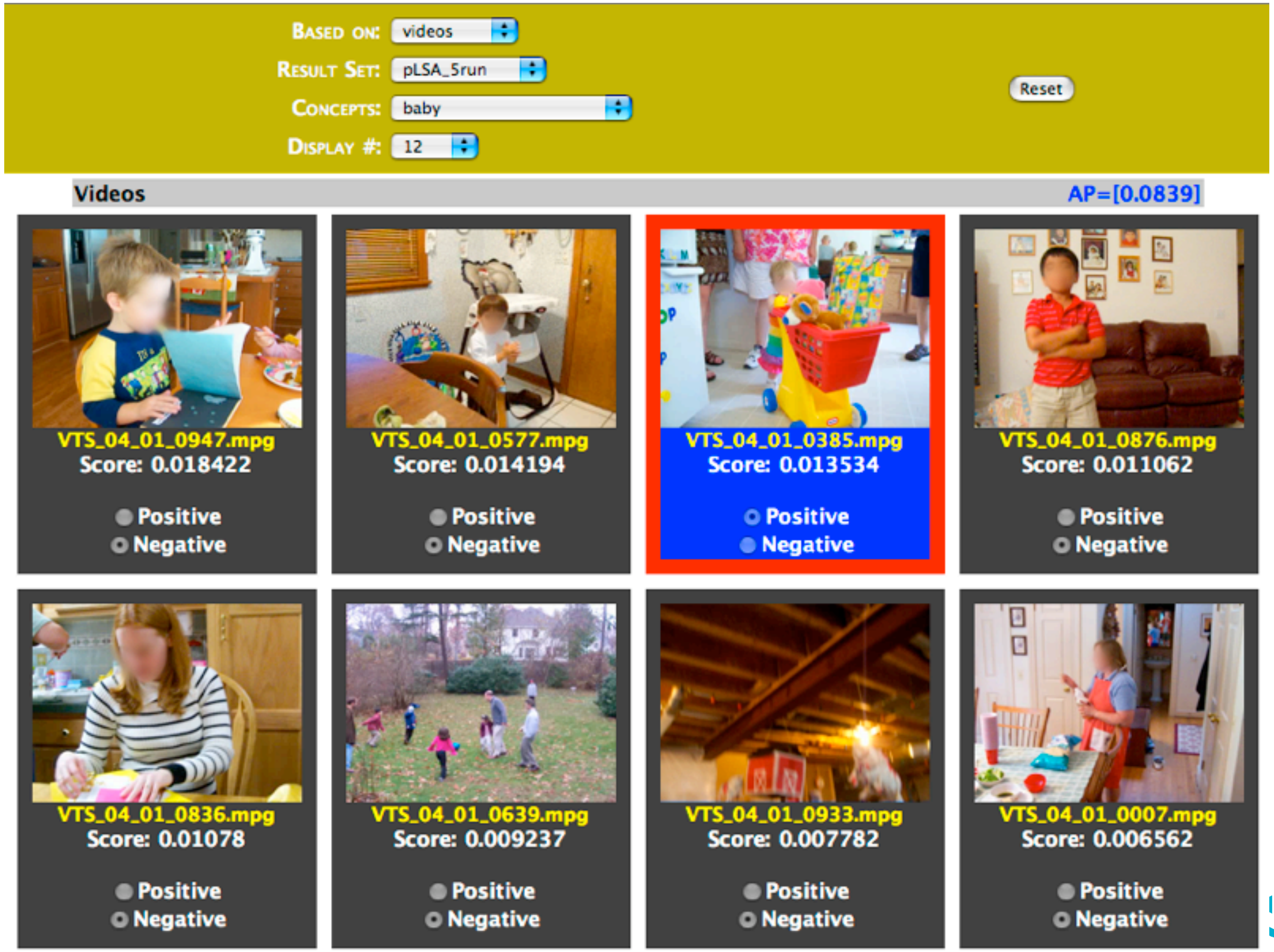


  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

How does it 'feel'?

- Browser impressions: How wrong is wrong?

Top 8 hits for "Baby"



Based on: videos

Result Set: pLSA_5run

Concepts: baby

Display #: 12

Reset

Videos AP=[0.0839]

VTS_04_01_0947.mpg
Score: 0.018422
○ Positive
○ Negative

VTS_04_01_0577.mpg
Score: 0.014194
○ Positive
○ Negative

VTS_04_01_0385.mpg
Score: 0.013534
○ Positive
● Negative

VTS_04_01_0876.mpg
Score: 0.011062
○ Positive
○ Negative

VTS_04_01_0836.mpg
Score: 0.01078
○ Positive
○ Negative

VTS_04_01_0639.mpg
Score: 0.009237
○ Positive
○ Negative

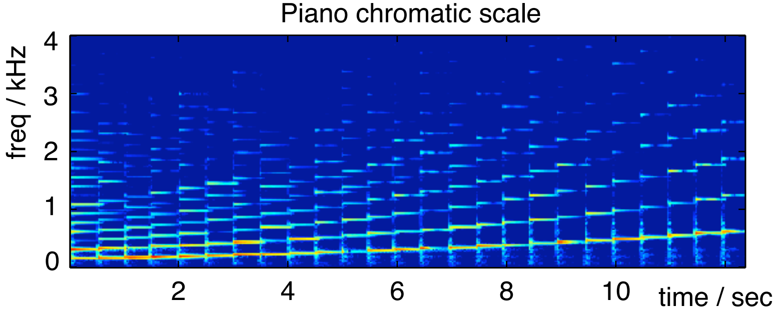
VTS_04_01_0933.mpg
Score: 0.007782
○ Positive
○ Negative

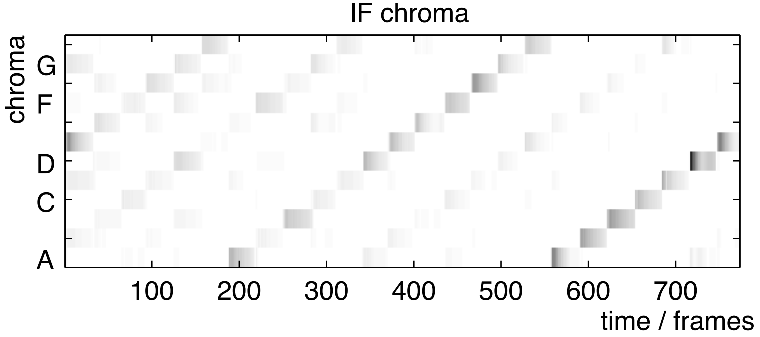
VTS_04_01_0007.mpg
Score: 0.006562
○ Positive
○ Negative

2. Cover Song Detection: Chroma

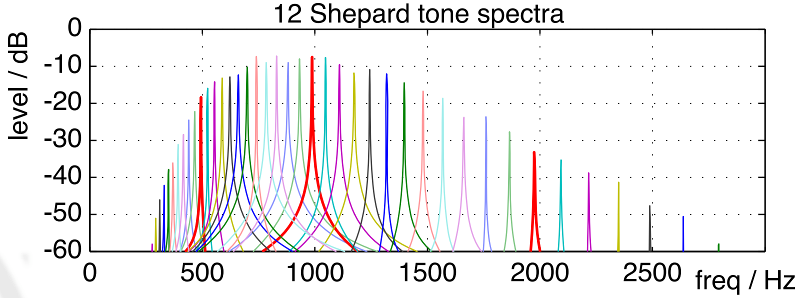
- Chroma features map spectral energy into one canonical octave
  - i.e. 12 semitone bins

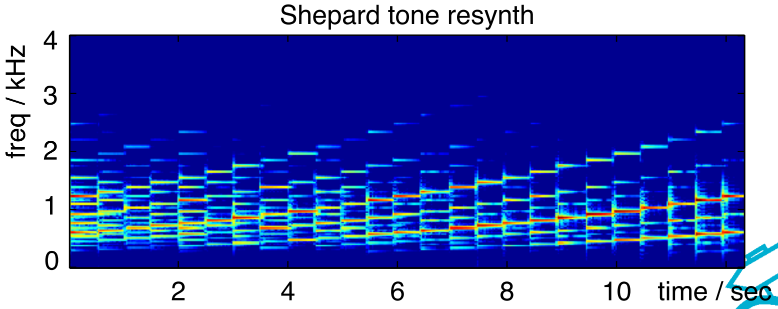
Piano scale


Piano chromatic scale


IF chroma

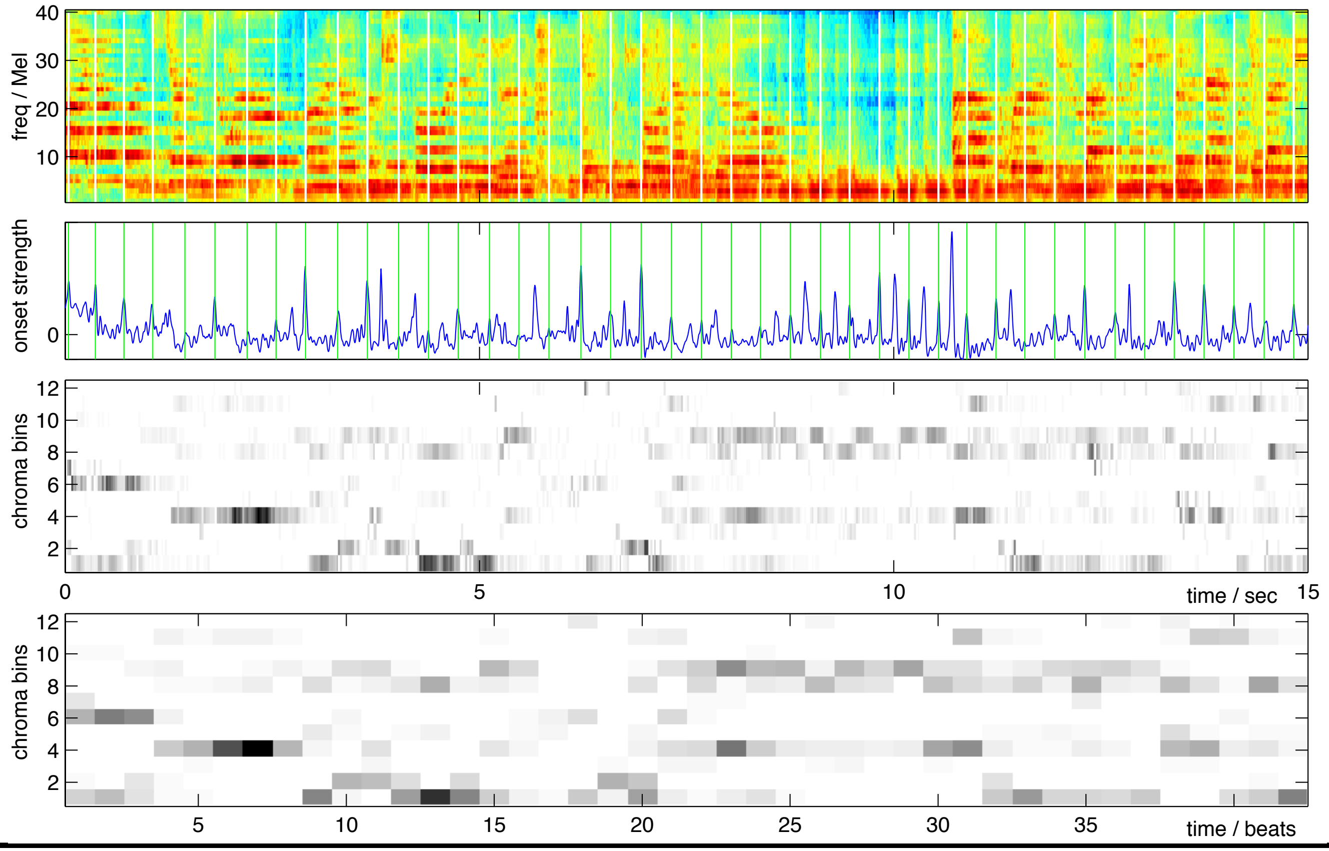
- Can resynthesize as "Shepard Tones"
  - all octaves at once


12 Shepard tone spectra


Shepard tone resynth

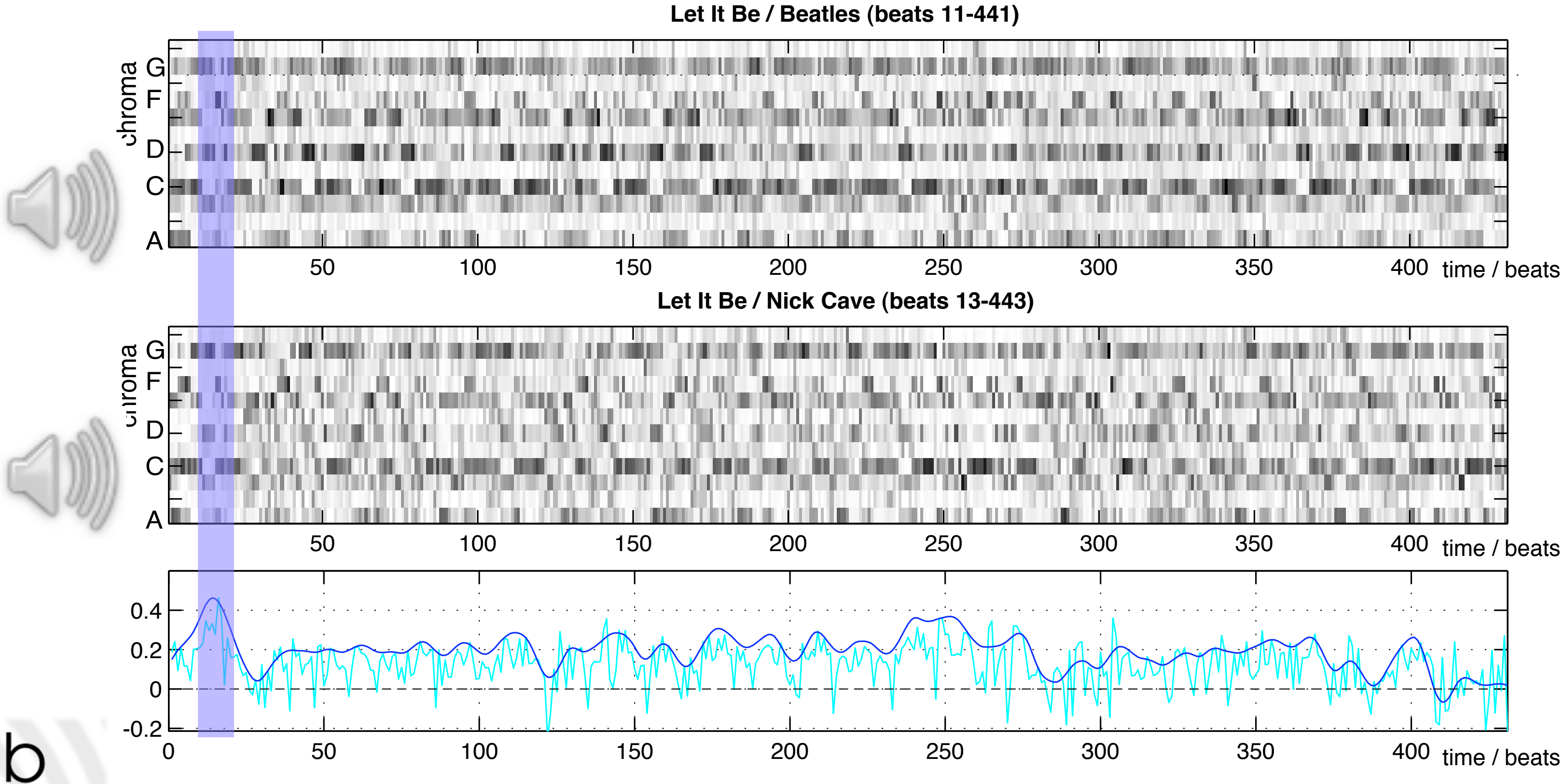
Beat-Synchronous Chroma Features

- Beat + chroma features / 30ms frames
  → average chroma within each beat
  - compact; sufficient?



Cross-Correlation Matching

- Look inside global cross-correlation to find matching fragments...
  - $\text{xcorr} = \Sigma_t \, \Sigma_f \left(C_1(t,f) \cdot C_2(t,f) \right)$ - view along time



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

Columbia University in the city of New York

“The Meaning of Music”

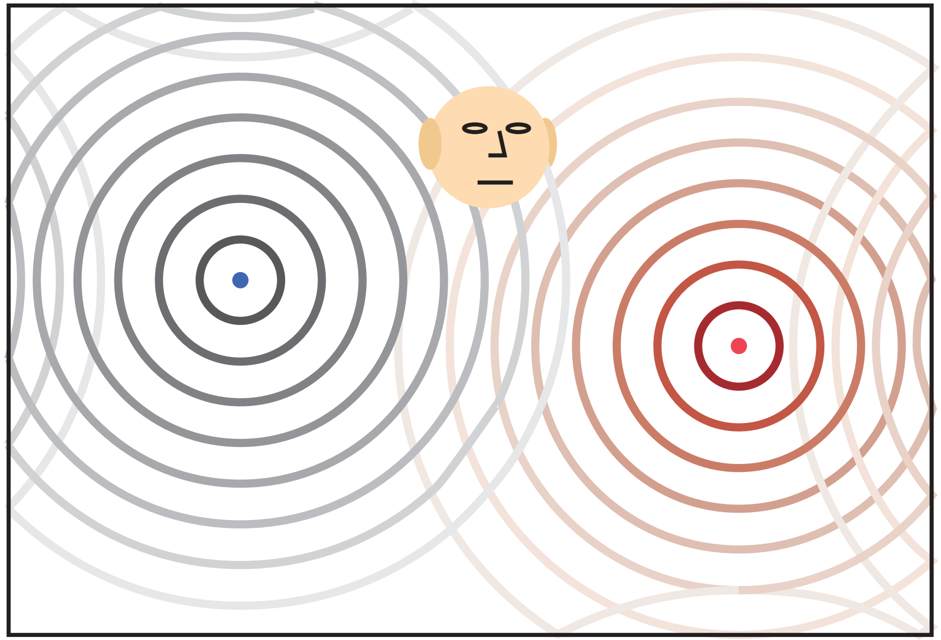
The ultimate goal of this research...

- What does music evoke in a listener’s mind?
  - i.e. “what does it all mean?” (metaphysics?)
  - study with subjective experiments
  - (then build detectors for specific responses ...?)

- What phenomena are denoted by “music”?
  - i.e. delineate the “set of all music”
  - (the ultimate music/nonmusic classifier?)

3. Binaural Source Separation

- 2 or 3 sources in reverberation
  - assume just 2 'ears'



- Tasks:
  - identify positions of sources (and number?)
  - recover source signals

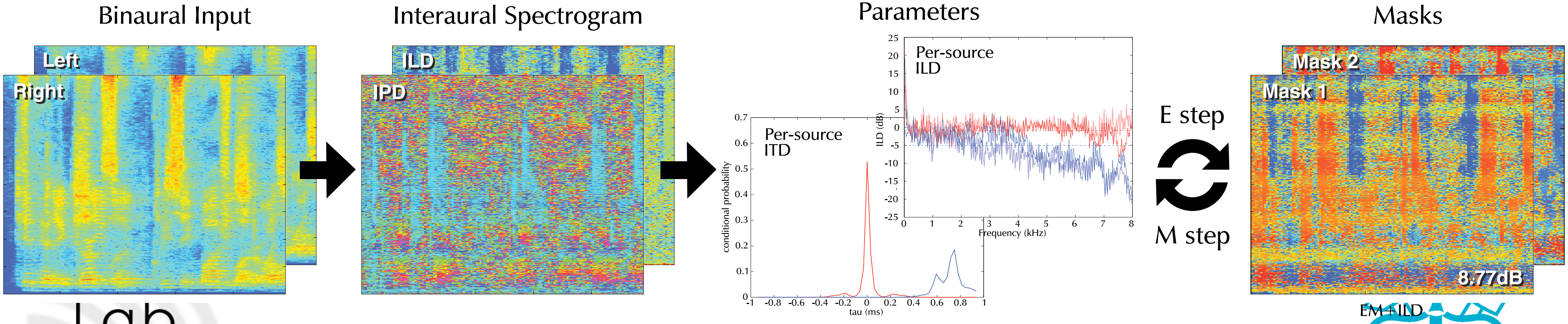
Spatial Estimation in Reverb

Mandel & Ellis '07

- Model interaural spectrum of each source as stationary level and time differences:

$$\frac{L(\omega,t)}{R(\omega,t)} = a(\omega)e^{j\omega\tau}N(\omega,t)$$

  - converge via EM to $a()$, τ for each source
  - mask is $\Pr(X(t,\omega)$ dominated by source $i)$


Binaural Input

Interaural Spectrogram

Parameters

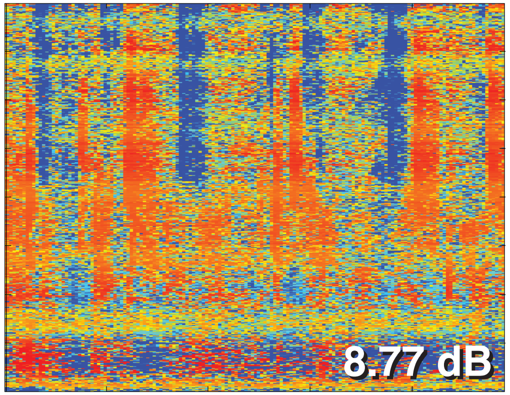
E step

M step

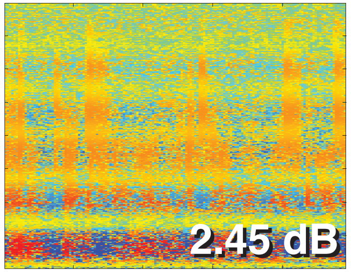
Masks

Spatial Estimation Results

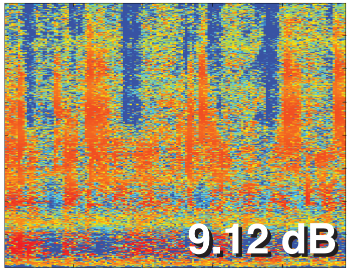
- Modeling uncertainty improves results
    - tradeoff between constraints & noisiness



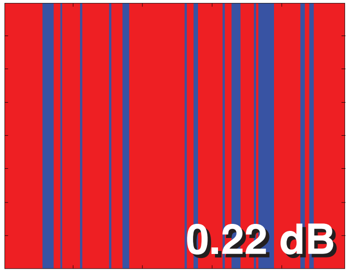
EM+ILD



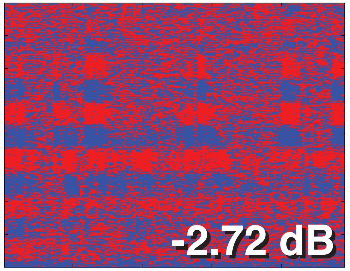
EM-ILD (only IPD)



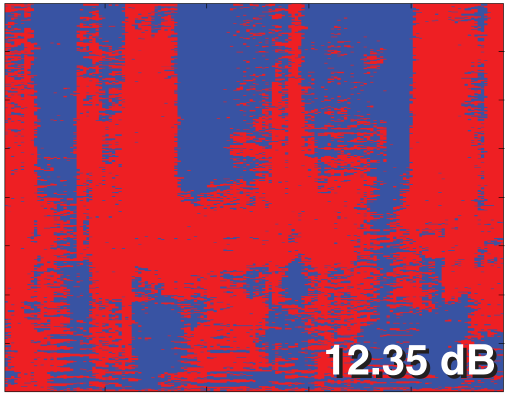
EM+1ILD (tied means)



PHAT-histogram



DUET



Ground Truth

Conclusions

- LabROSA
  - information from sound ...
  - ... via signal processing and machine learning
- Environmental Sound
- Music Audio
- Source Separation
- Speech, models, dolphins...