Mining Audio

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/

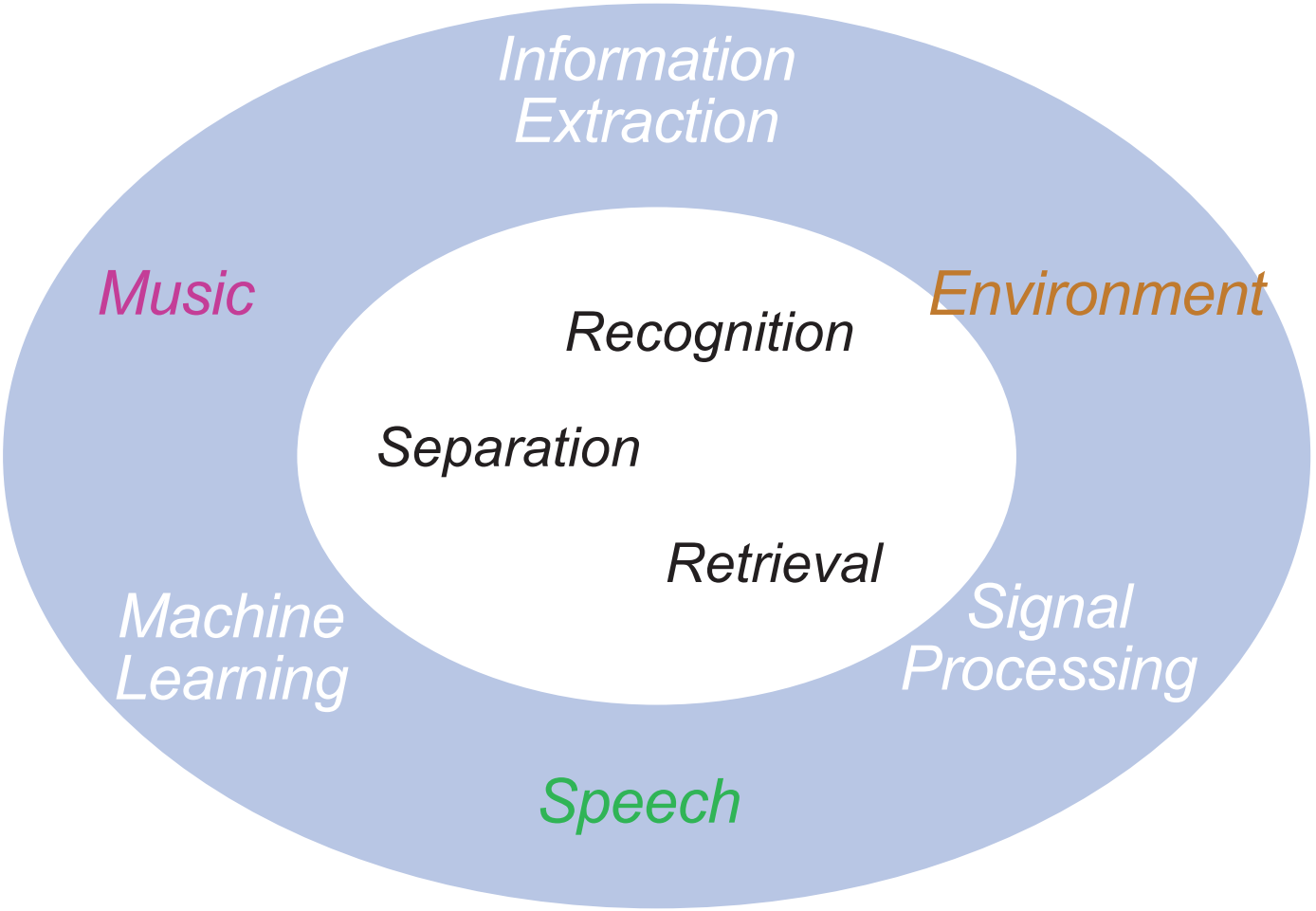
1. Real-World Sound
2. Music Audio
3. Environmental Audio
4. Outstanding Issues

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

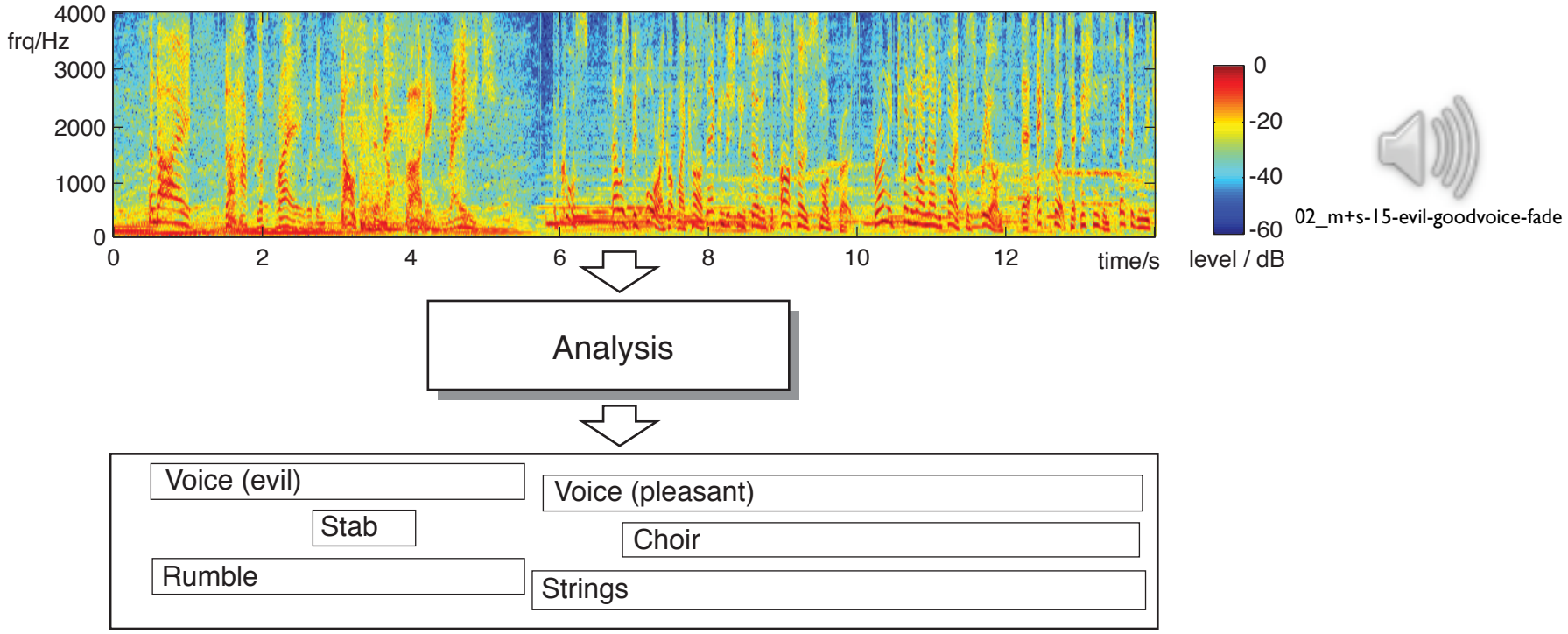
Columbia University
In the City of New York

LabROSA Overview

- Getting information from sound

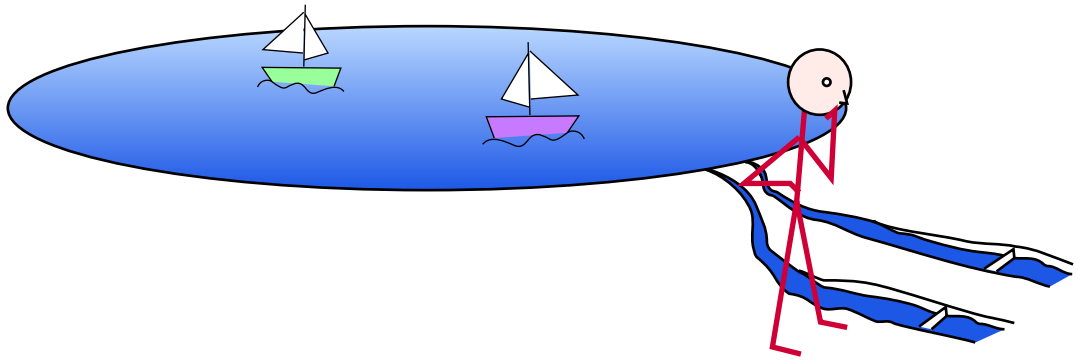


1. Real-World Sound



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

Auditory Scene Analysis

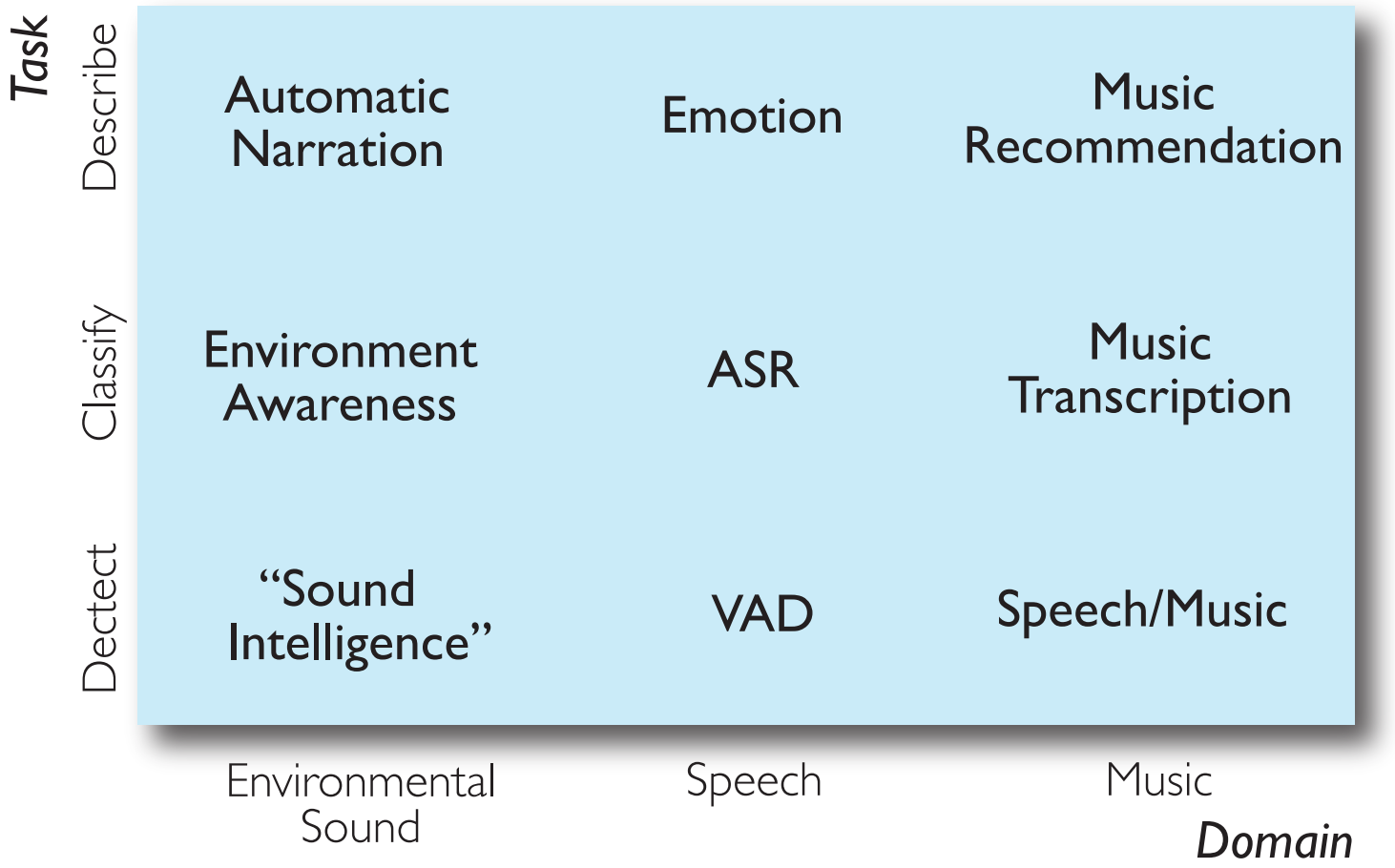


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

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

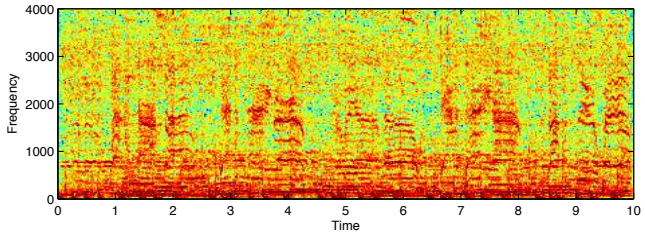
Machine Listening

- Extracting useful information from sound
  - ... like animals do



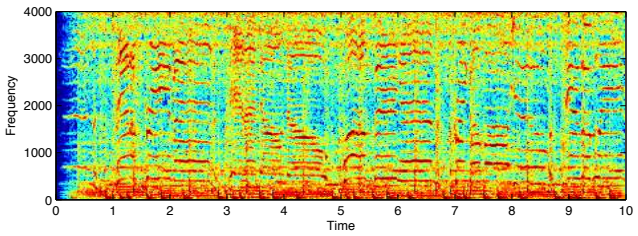
2. Mining Music Audio

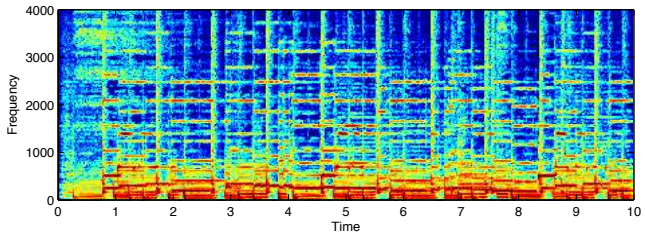
Query track

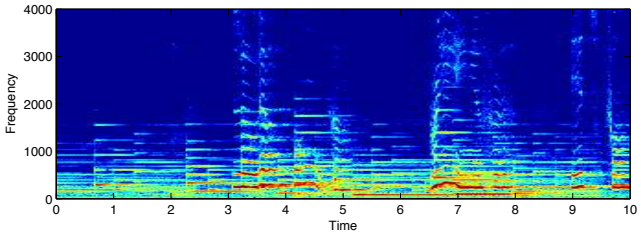


One Million Songs

"Similar" tracks

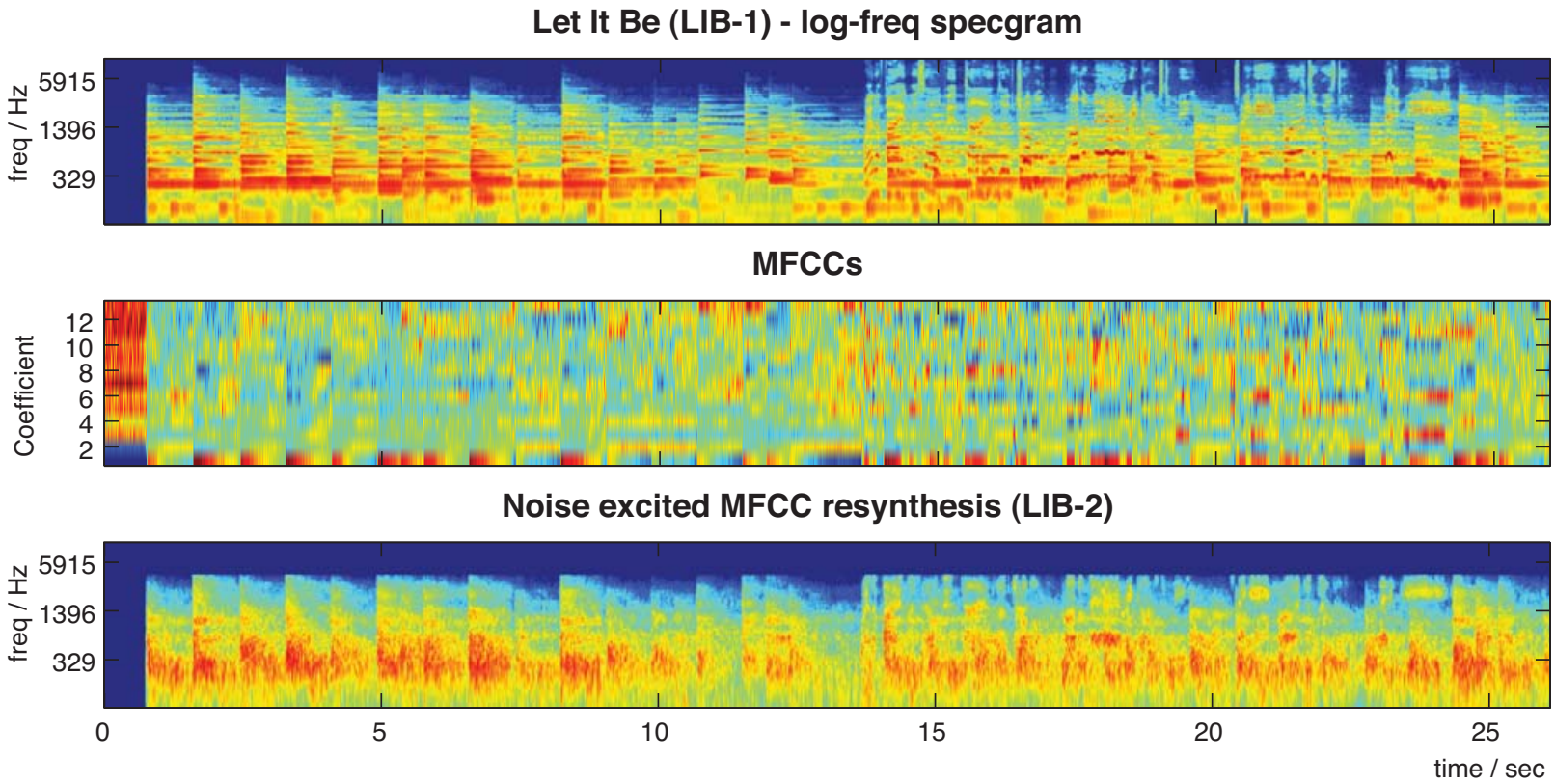




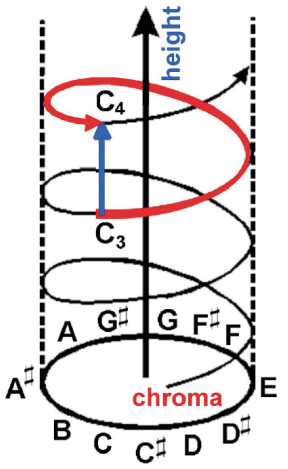


Music Audio Representations

- We need a description of the audio that contains "suitable" detail
- Speech Recognition uses Mel-Frequency Cepstral Coefficients (MFCCs)

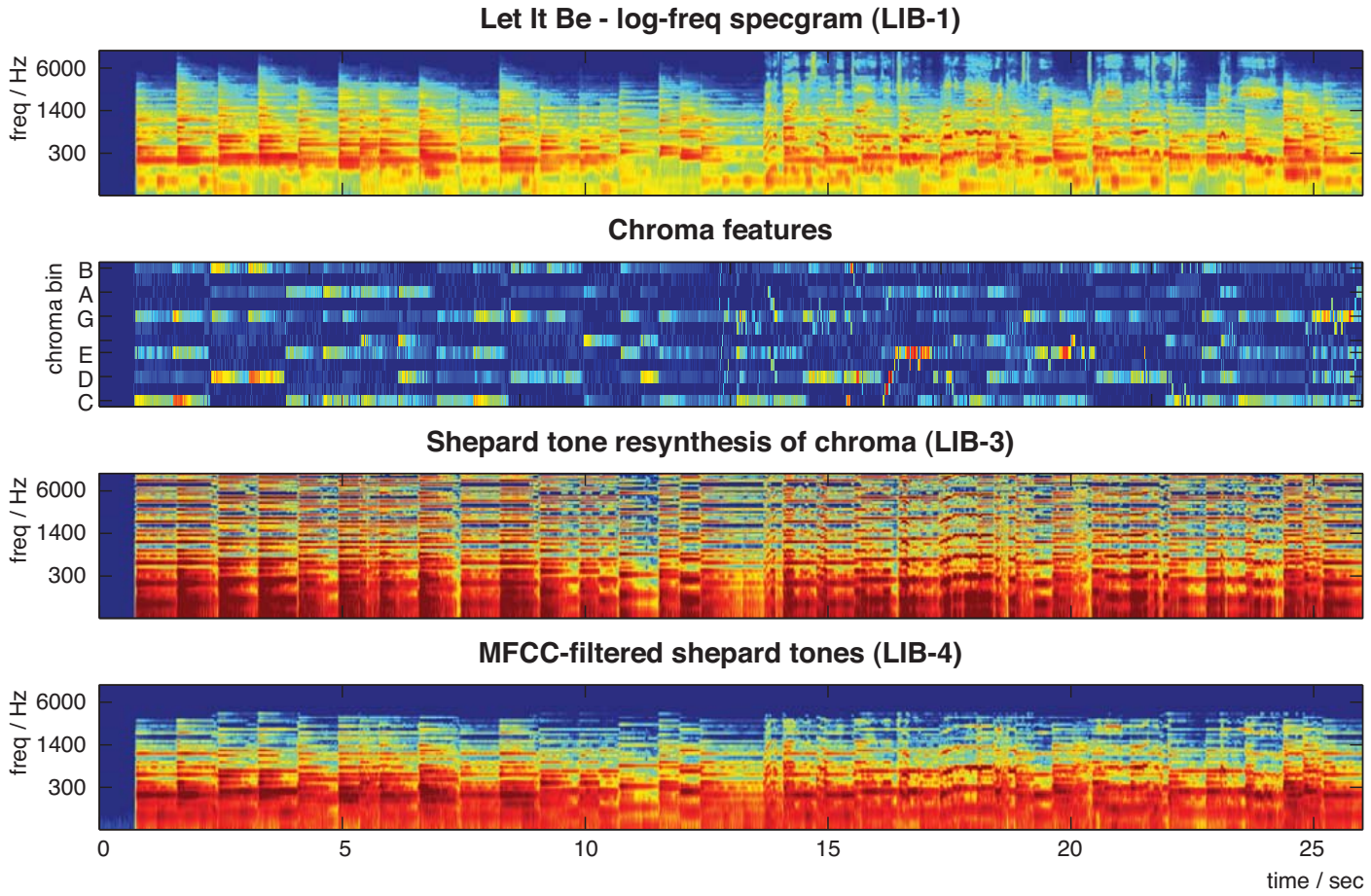


Chroma Features



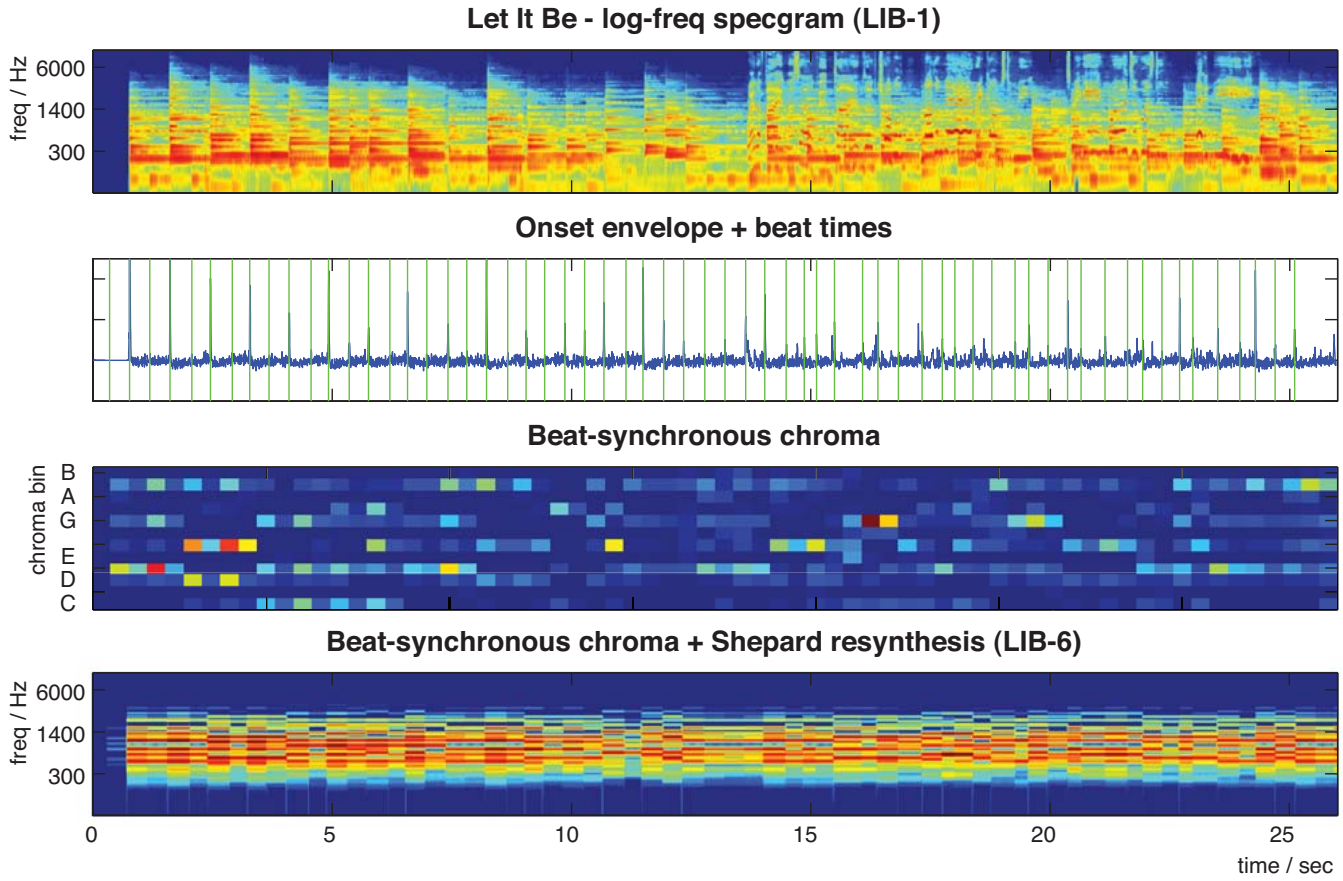
Warren et al. 2003

- We'd like to preserve the notes
  - at least within one octave



Beat-Synchronous Chroma

- Perform Beat Tracking
  .. and record one chroma vector per beat
  - compact representation of harmonies



Million Song Dataset (MSD)

Thierry Bertin-Mahieux

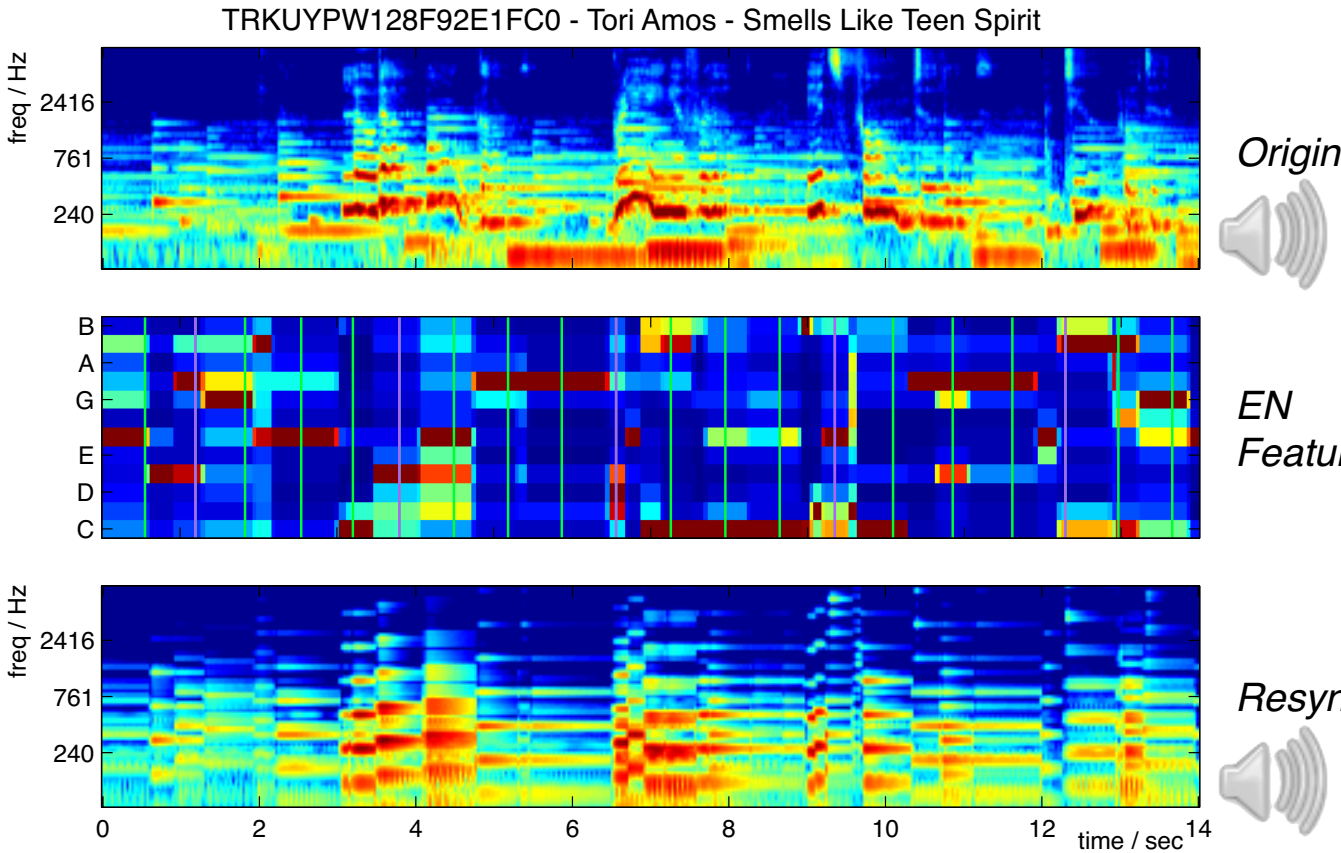
- **Commercial-scale dataset available to MIR researchers**
  - 1M pop songs
  - 250 GB of features
  - (6 years of listening)

- **Many facets:**
  - Features,
    Lyrics,
    Tags,
    Covers,
    Listeners ...

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

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

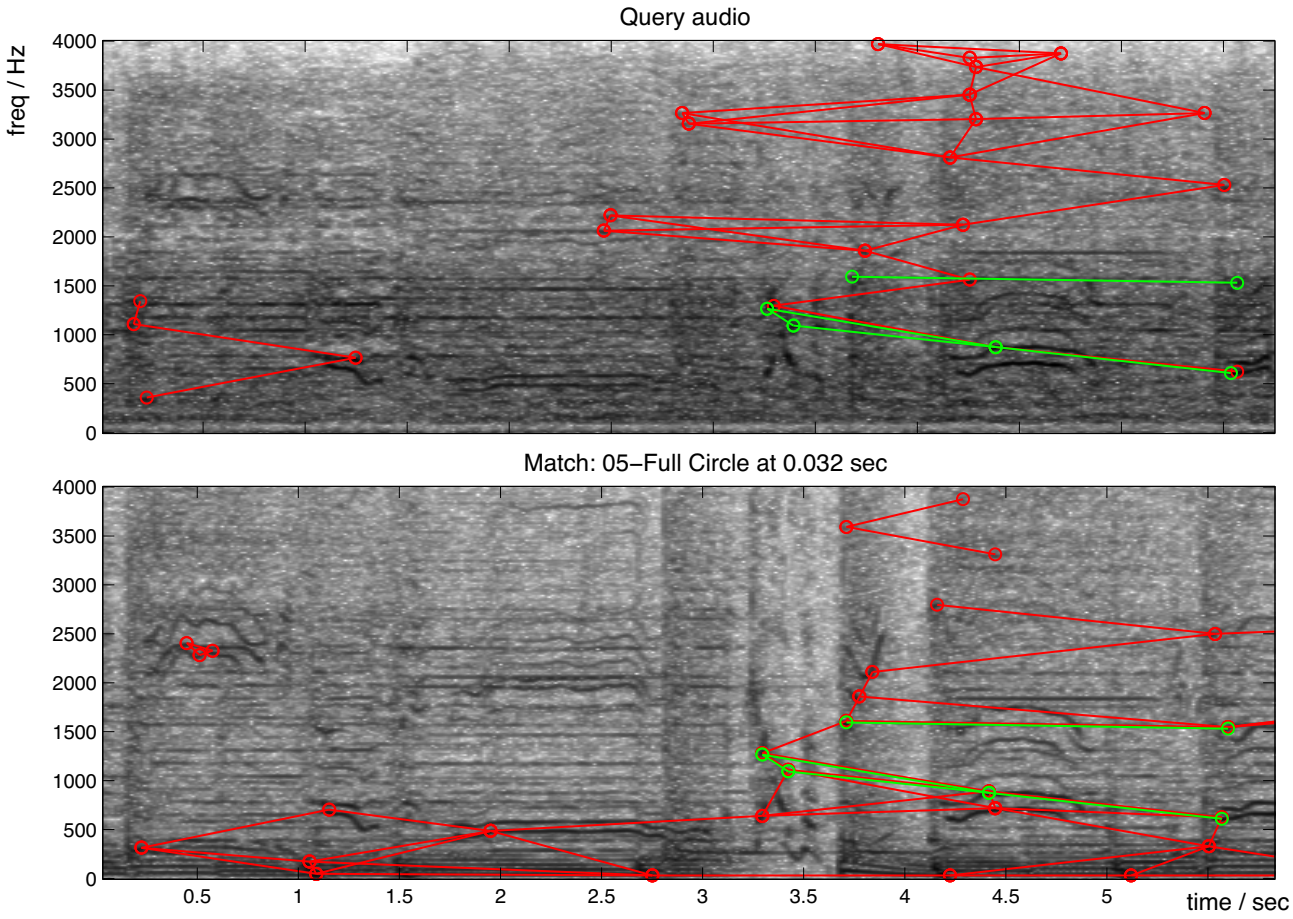
```
12 hello      6 here       3 is
11 i          6 us         3 with
10 a          6 entertain  3 oh
 9 and        4 the        3 out
 7 it         4 feel       3 an
 6 are        4 yeah       3 light
 6 we         3 to         3 less
 6 now        3 my         3 danger
```

Finding Similar Items

Avery Wang '03

- If it's really the same, we can use fingerprinting (e.g. Shazam):

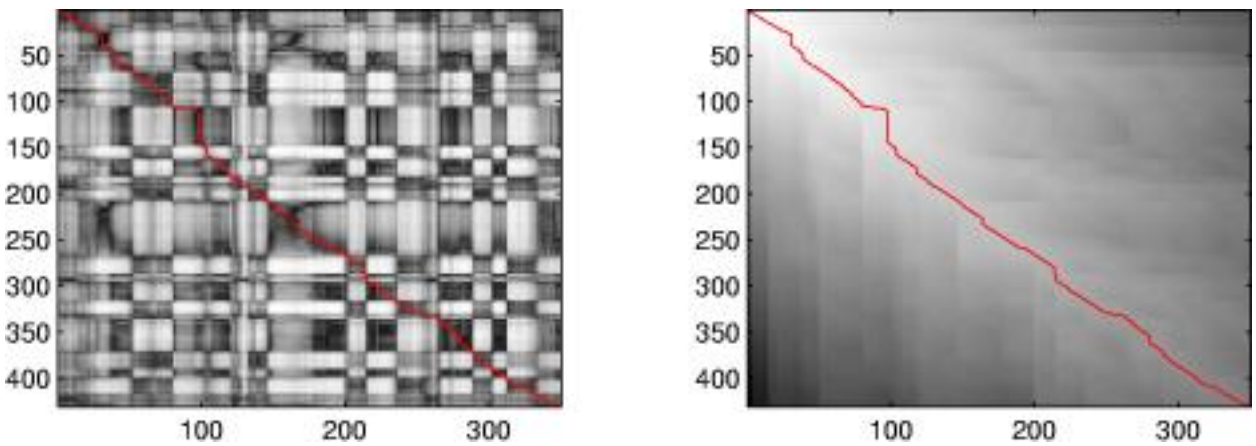
  - find local "landmarks"
  - quantize by pairs
  - inverted index



Dynamic Programming Alignment

Serra, Gomez et al. '08

- We can match the chroma representations by DP alignment

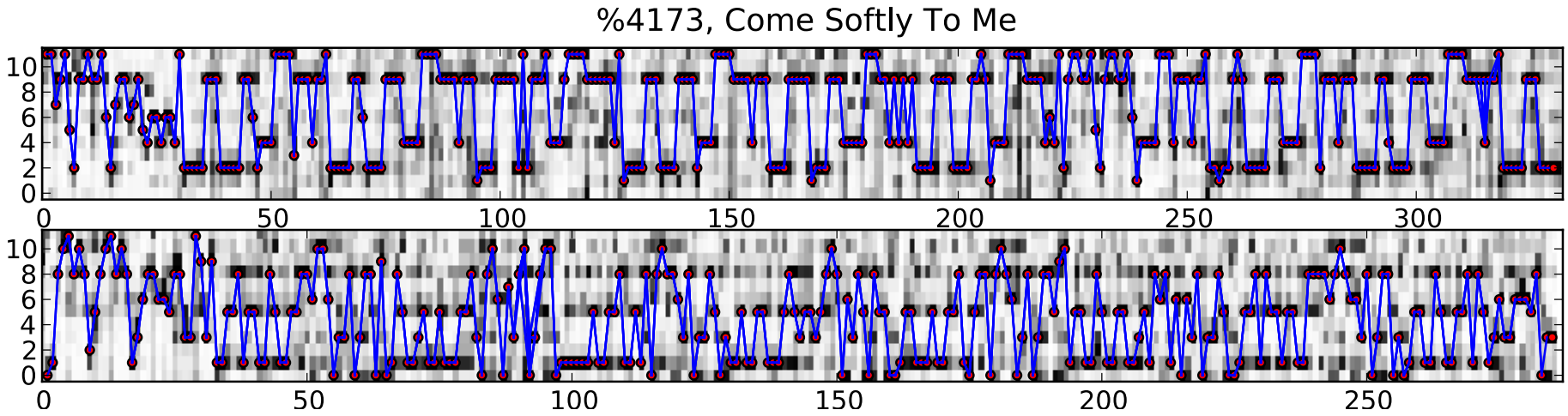


- robust to timing changes
- quite efficient
- best performer in MIREX Cover Song evaluations

Large-Scale Cover Matching

Bertin-Mahieux & Ellis '11

- How can we find covers in 1M songs?
  - @ 1 sec / comparison, one search = 11.5 CPU-days
  - full N^2 mining = 16,000 CPU-years
- Hashing?
  - landmarks from chroma patches (like Shazam)



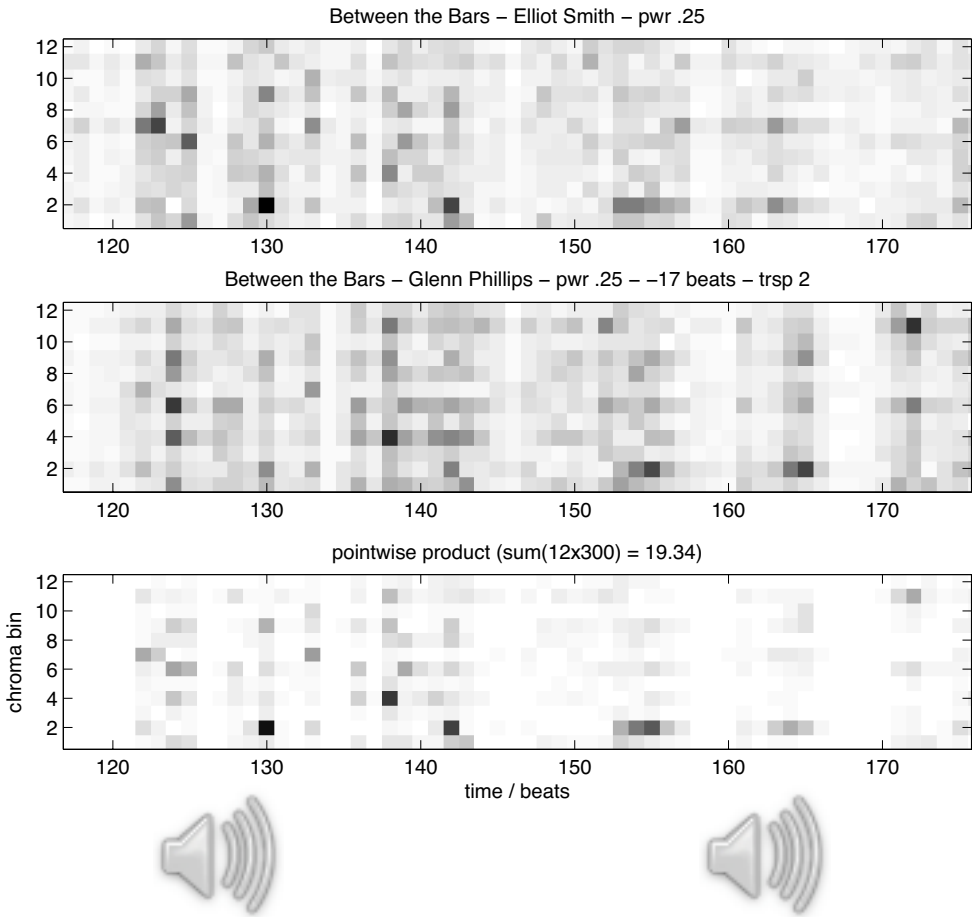
  - represent item by distribution of contour fragments
  - one stage in a filtering hierarchy...

Euclidean Cover Matching

Ellis & Poliner '08
Bertin-Mahieux & Ellis '12

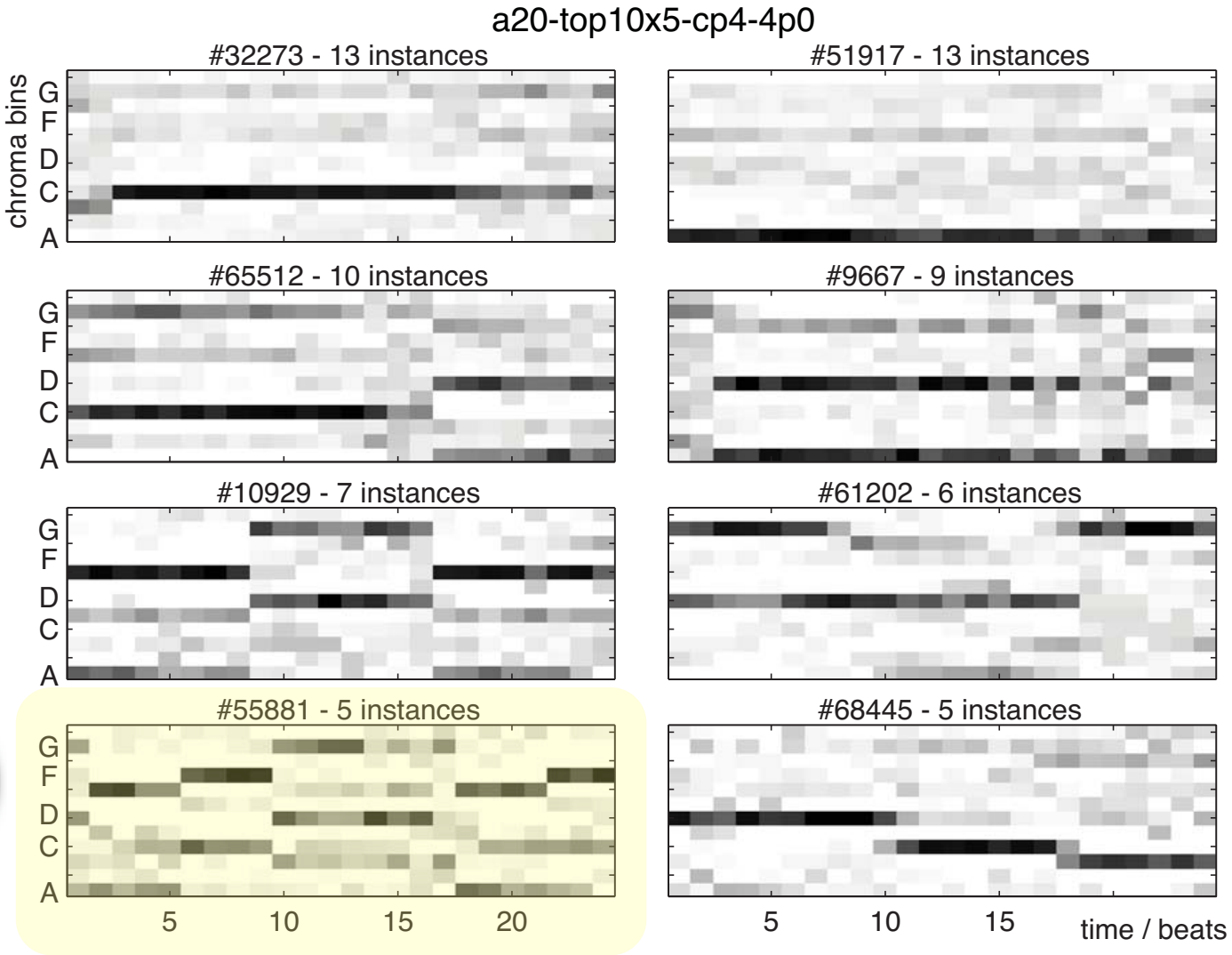
- Reduce cover matching to nearest-neighbor search in fixed-size beat-chroma patches

- Right "distance"?
  - principal components
  - learned weightings
- Alignment/ segmentation?
  - music segmentation
  - multiple probes
  - framing-insensitive representation



Pattern Mining

- Cluster beat-synchronous chroma patches



Other Work: MSD Challenge

with Brian McFee

- Using "Taste Profile" data
  - User-Song-playcount

```
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SOAKIMP12A8C130995    1
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SOAPDEY12A81C210A9    1
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SOBBMDR12A8C13253B    2
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SOCNMUH12A6D4F6E6D    1
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SODACBL12A8C13C273    1
b80344d063b5ccb3212f76538f3d9e43d87dca9e    SODDNQT12A6D4F5F7E    5
```

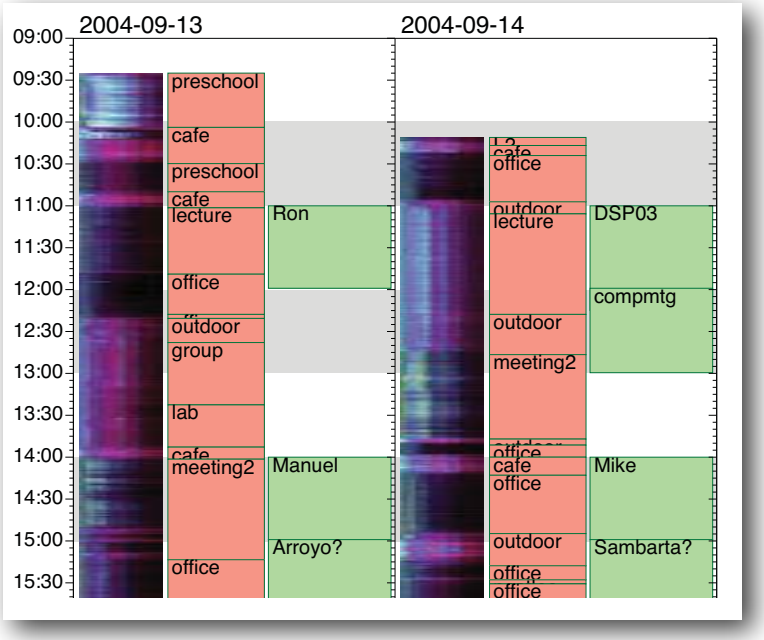
- Challenge:
  Rank 1M songs to complete half a playlist
  - using audio, metadata, lyrics, whatever
- Kaggle.com based contest
  - self-service evaluation, leader board
- Results at MIREX, AdMIRe

Music: Summary

- Finding Musical Similarity at Large Scale
- Beat Chroma Representation
- Million Song Dataset
- MSD Challenge

3. Environmental Audio

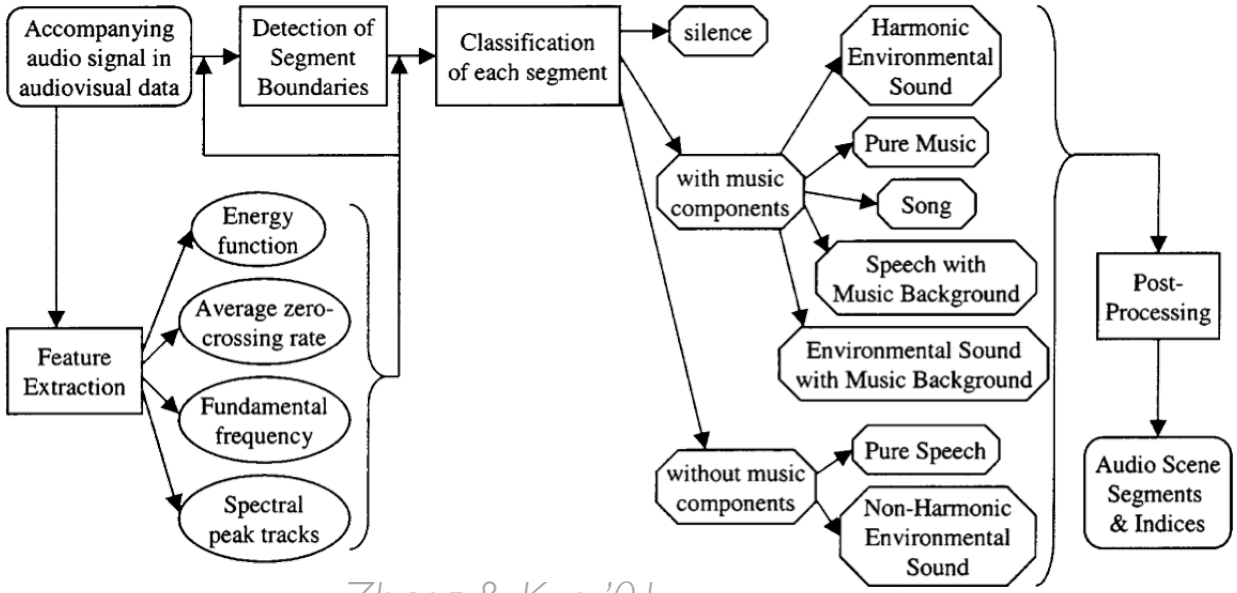
- Audio Lifelog Diarization



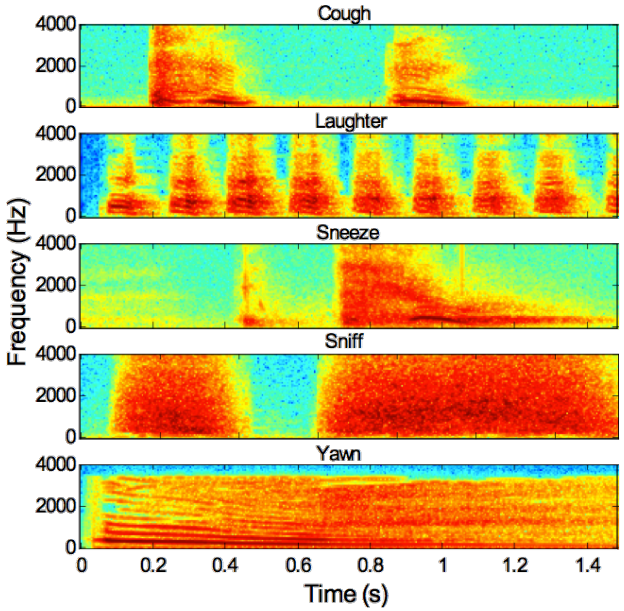
- Consumer Video Classification & Search

Prior Work

- Environment Classification
  - speech/music/ silent/machine



Zhang & Kuo '01

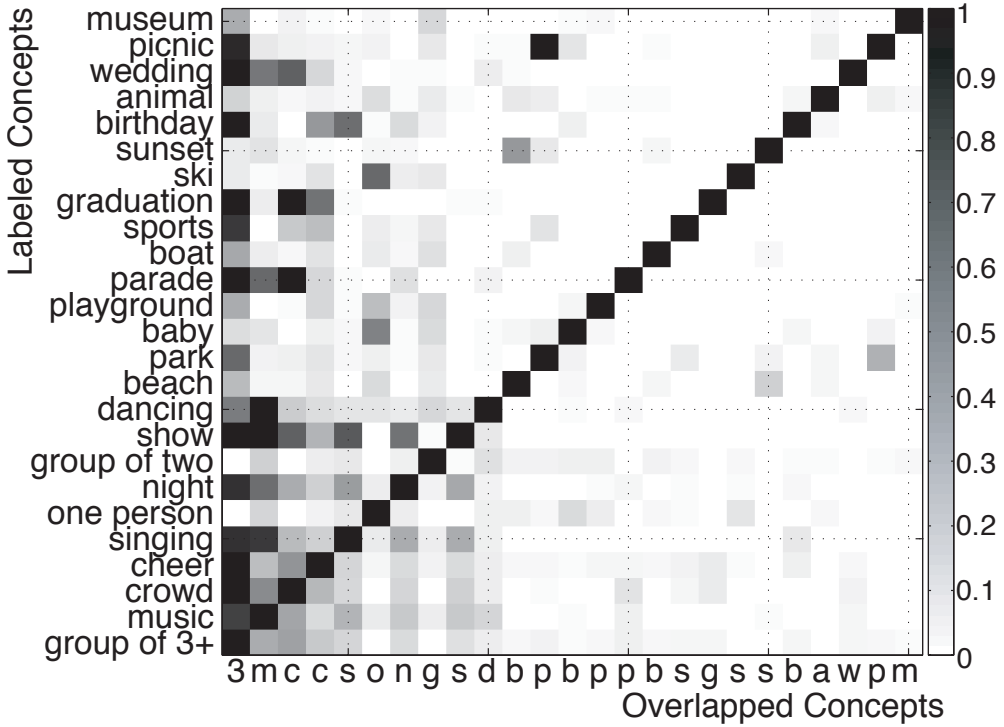


Temko & Nadeu '06

- Nonspeech Sound Recognition
  - Meeting room Audio Event Classification
  - sports events - cheers, bat/ball sounds, ...

Consumer Video Dataset

- 25 "concepts" from Kodak user study
  - boat, crowd, cheer, dance, ...

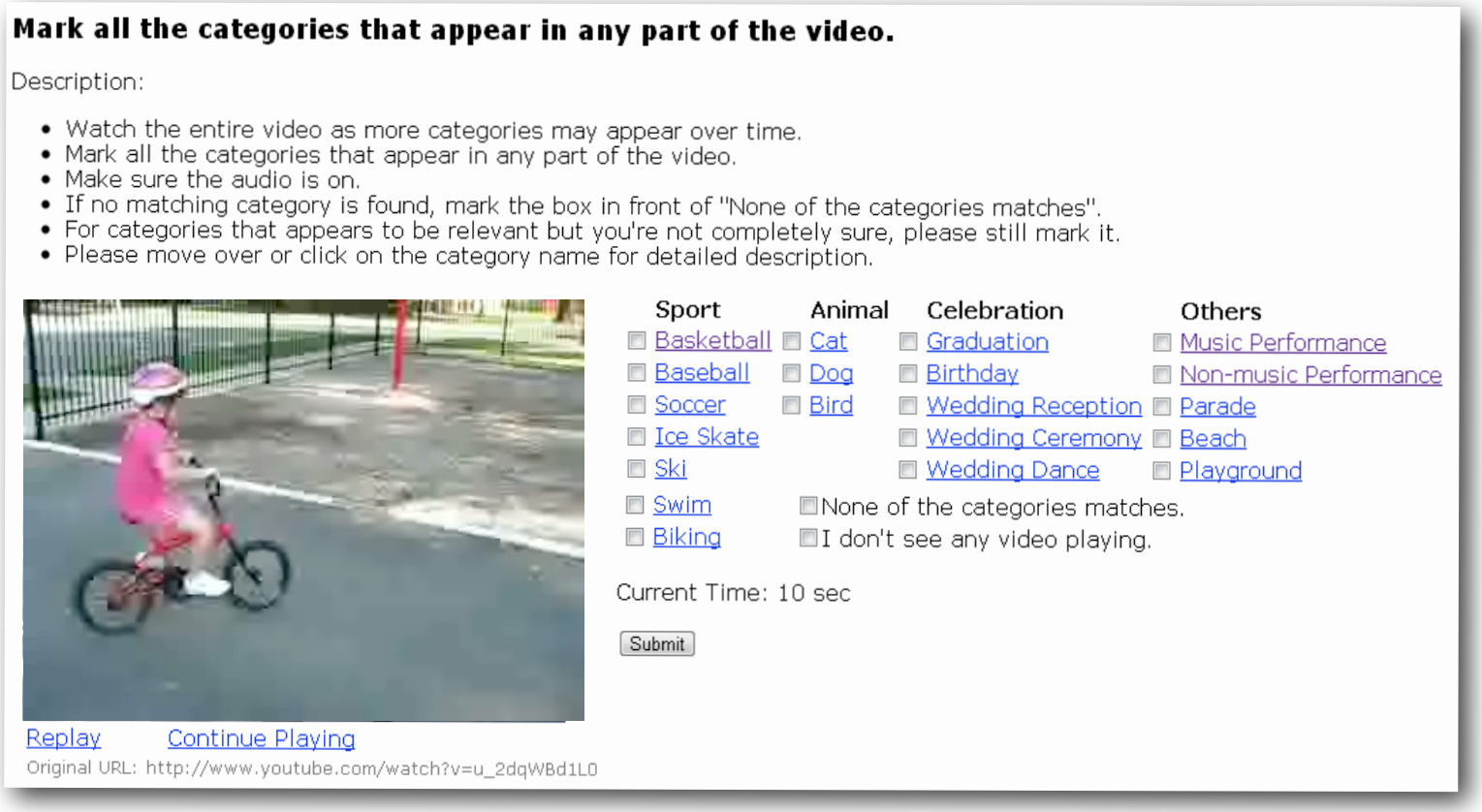


- Grab top 200 videos from YouTube search
  - then filter for quality, unedited = 1873 videos
  - manually relabel with concepts

Obtaining Labeled Data

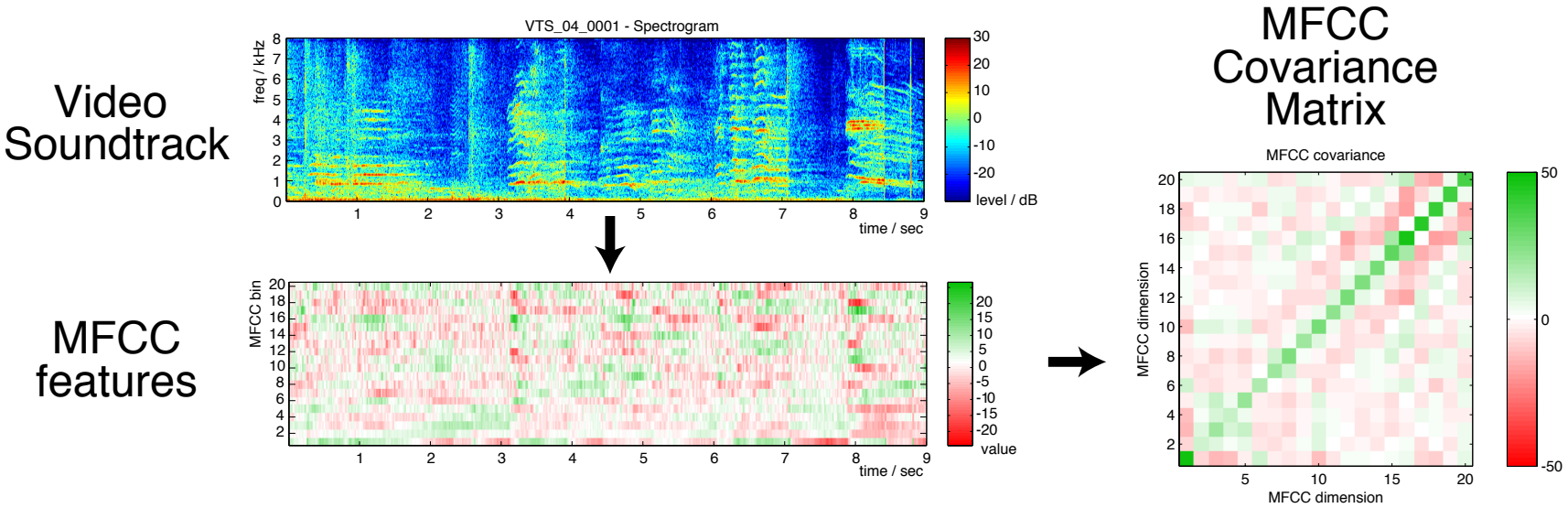
Y-G Jiang et al. 2011

- Amazon Mechanical Turk
  - 10s clips
  - 9,641 videos in 4 weeks



2. Background Classification

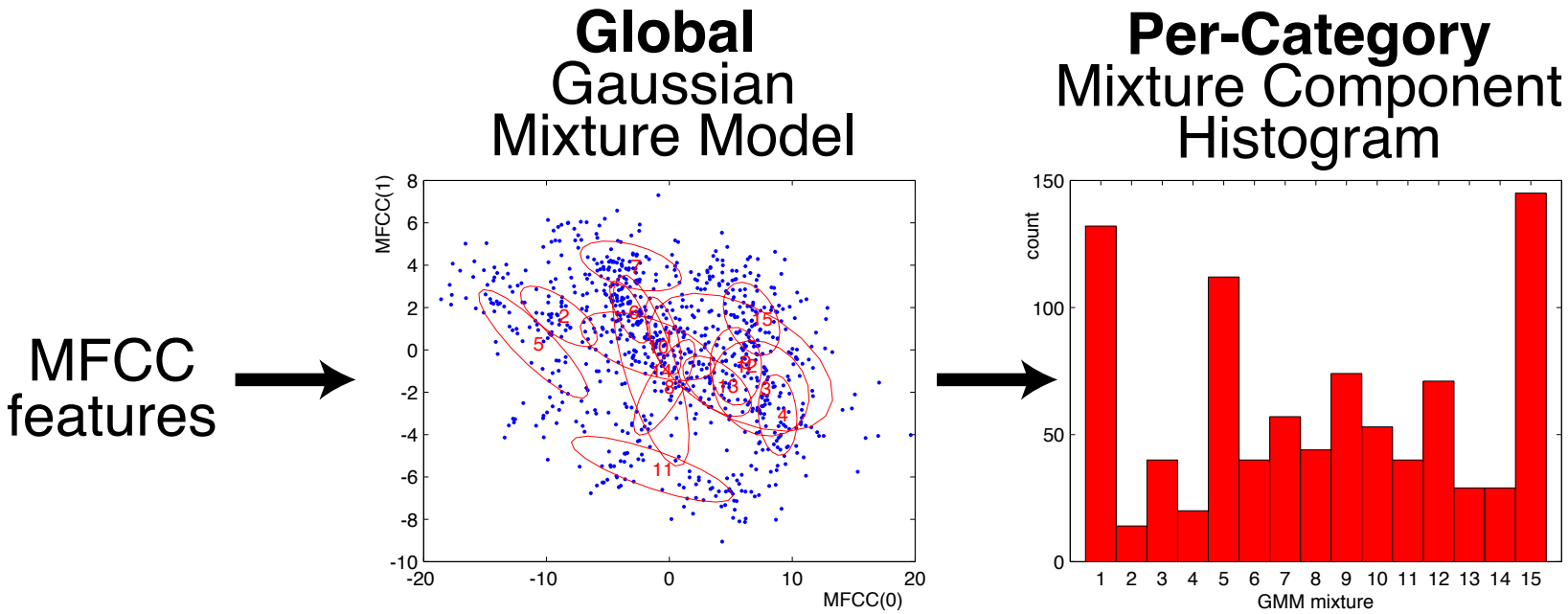
- Baseline for soundtrack classification
  - divide sound into short frames (e.g. 30 ms)
  - calculate features (e.g. MFCC) for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = "bag of features"



- Classify by e.g. KL distance + SVM

Codebook Histograms

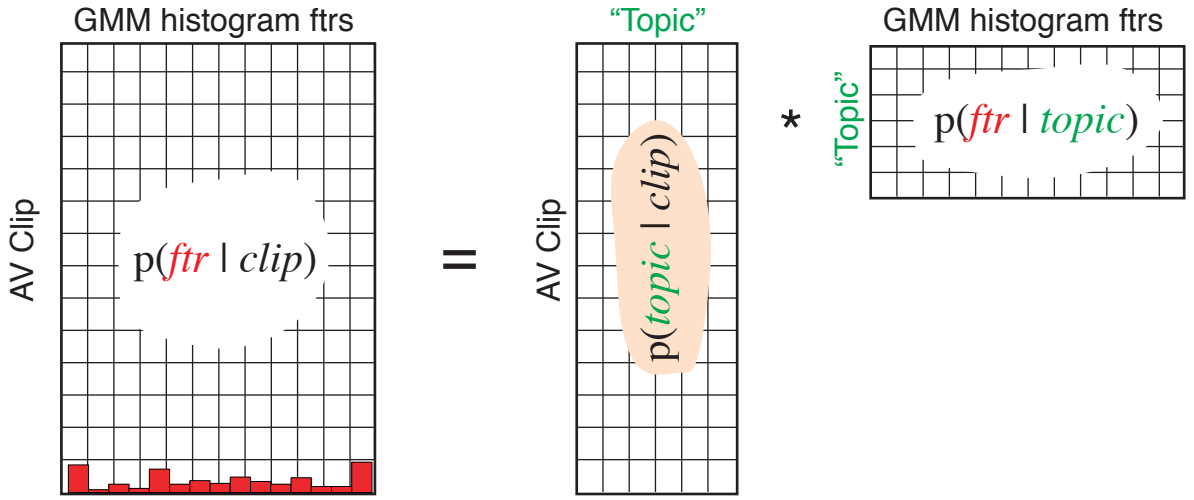
- Convert high-dim. distributions to multinomial



- Classify by distance on histograms
  - KL, Chi-squared
  - + SVM

Latent Semantic Analysis (LSA)

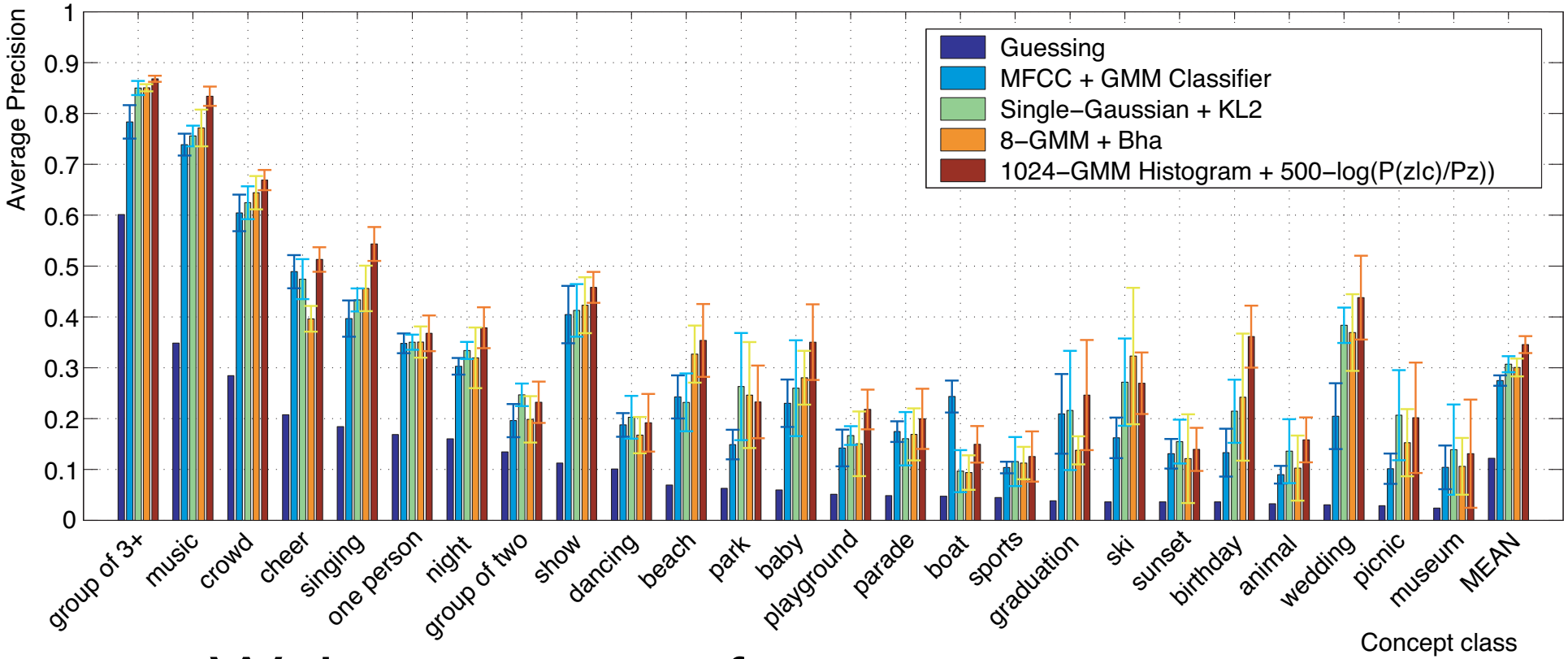
- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - $p(ftr \mid clip) = \sum_{topics} p(ftr \mid topic)\, p(topic \mid clip)$



  - use (normalized?) $p(topic \mid clip)$ as per-clip features

Background Classification Results

K Lee & Ellis '10

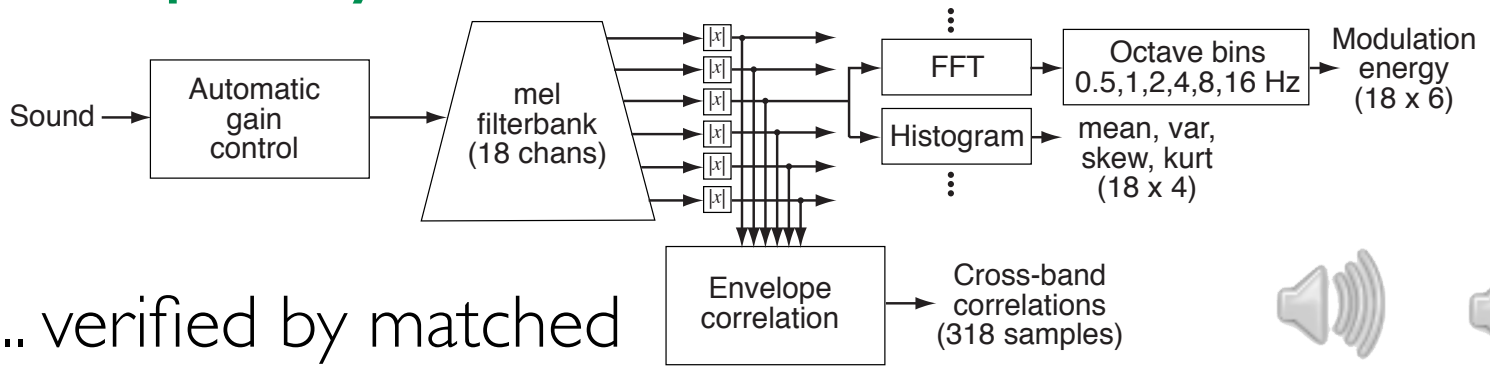


- Wide range in performance
  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

Sound Texture Features

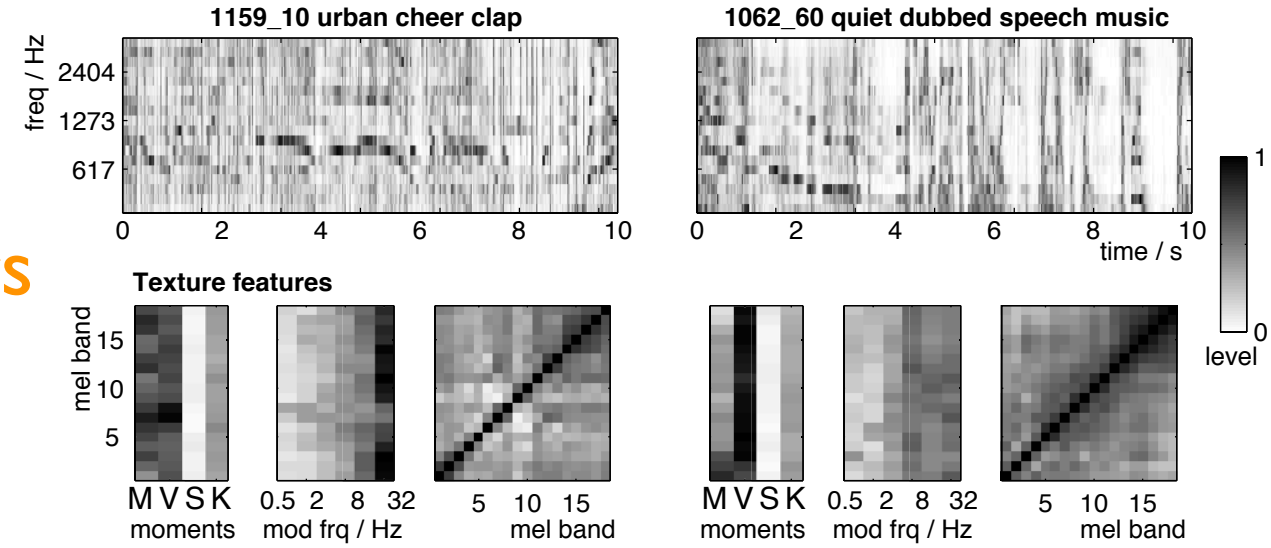
McDermott Simoncelli '09
Ellis, Zheng, McDermott '11

- Characterize sounds by perceptually-sufficient statistics



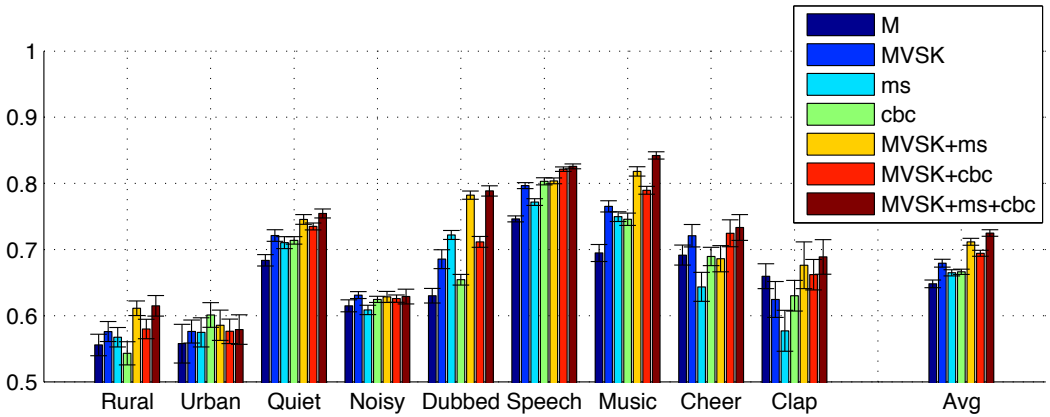
  - .. verified by matched resynthesis

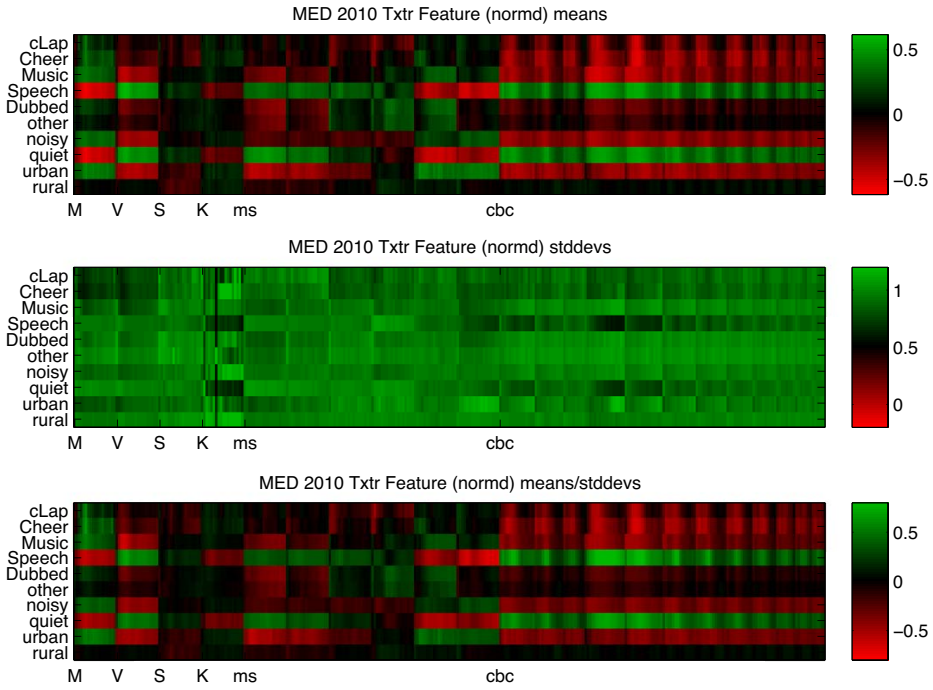
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



Sound Texture Features

- Test on MED 2010 development data
  - 10 specially-collected manual labels



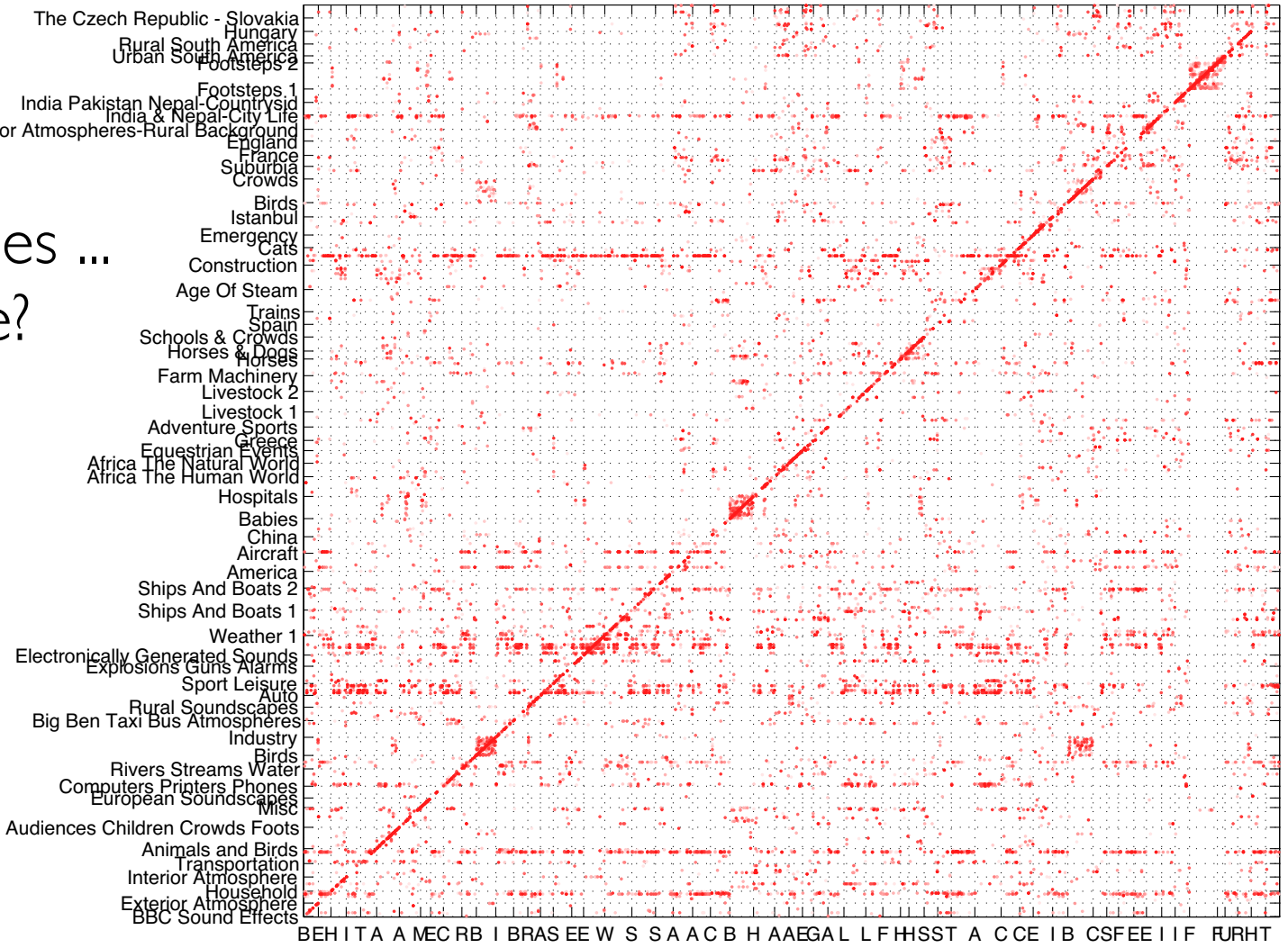


- Contrasts in feature sets
  - correlation of labels...
- Perform ~ same as MFCCs
  - combine well

Real-World Dictionary

- BBC Sound Effects as reference library

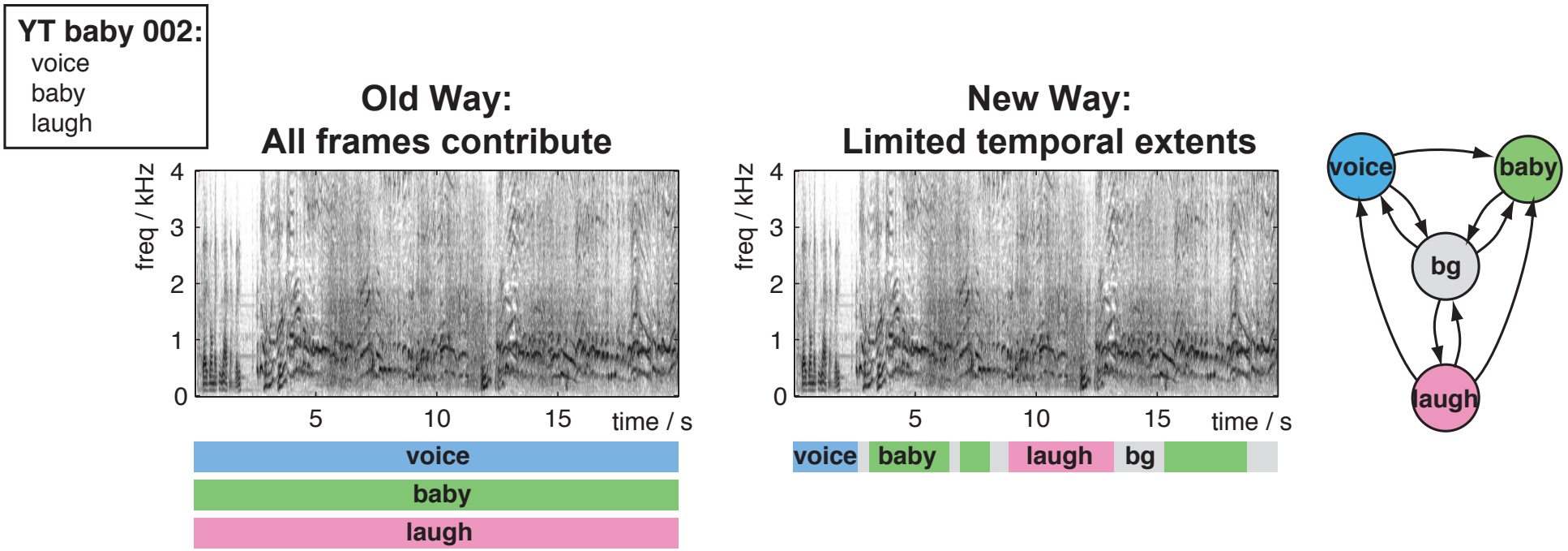
- 1000+ examples ... comprehensive?
- similarity via normalized textures (over 10s chunks)



3. Foreground Event Recognition

K Lee, Ellis, Loui '10

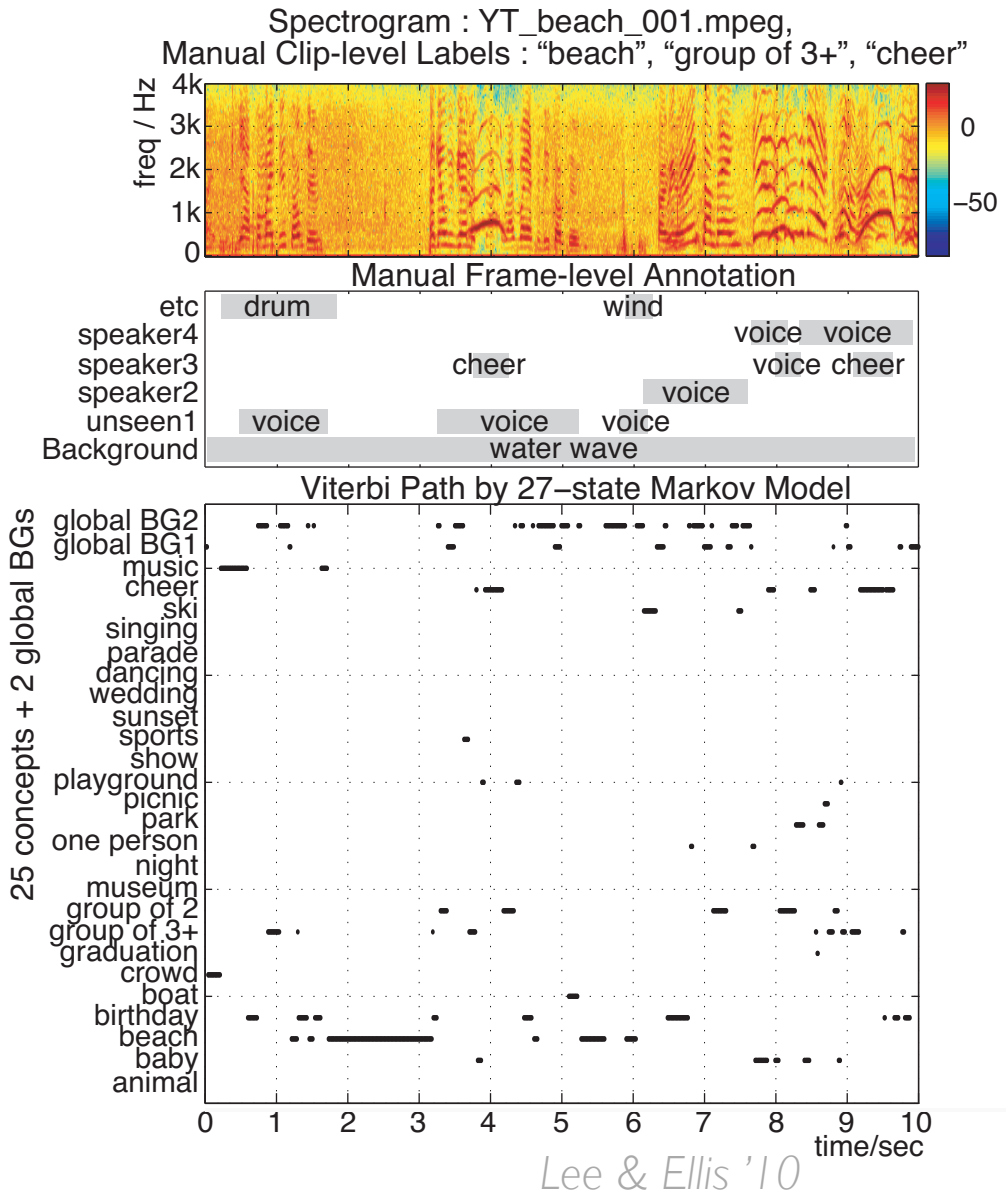
- Global vs. local class models
  - tell-tale acoustics may be 'washed out' in statistics
  - try iterative realignment of HMMs:



  - "background" model shared by all clips

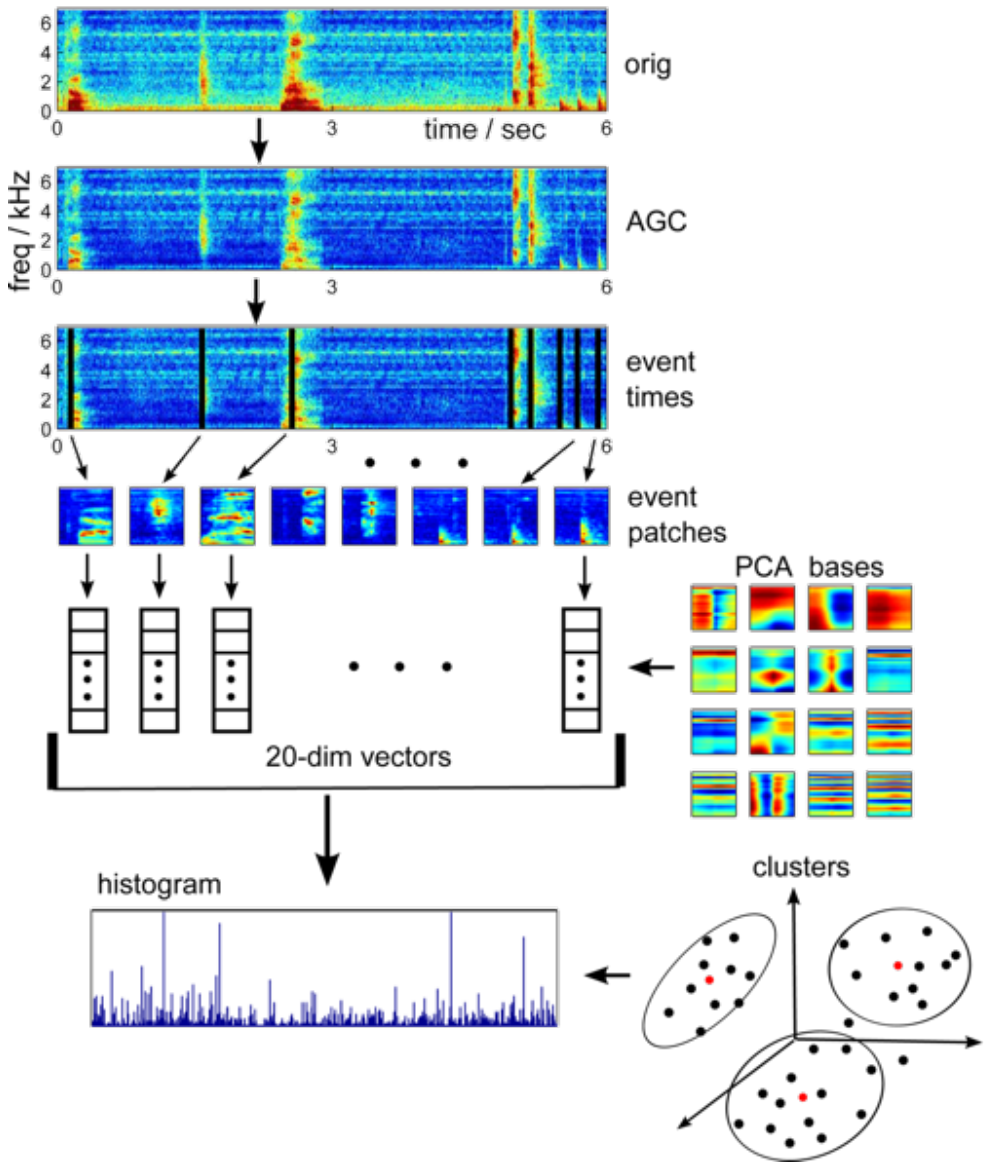
Foreground Event HMMs

- Training labels only at clip-level
- Refine models by EM realignment
- Use for classifying entire video...
  - or seeking to relevant part



Lee & Ellis '10

Transient Features



Cotton, Ellis, Loui '11

- Transients = foreground events?
- Onset detector finds energy bursts
  - best SNR
- PCA basis to represent each
  - 300 ms x auditory freq
- "bag of transients"

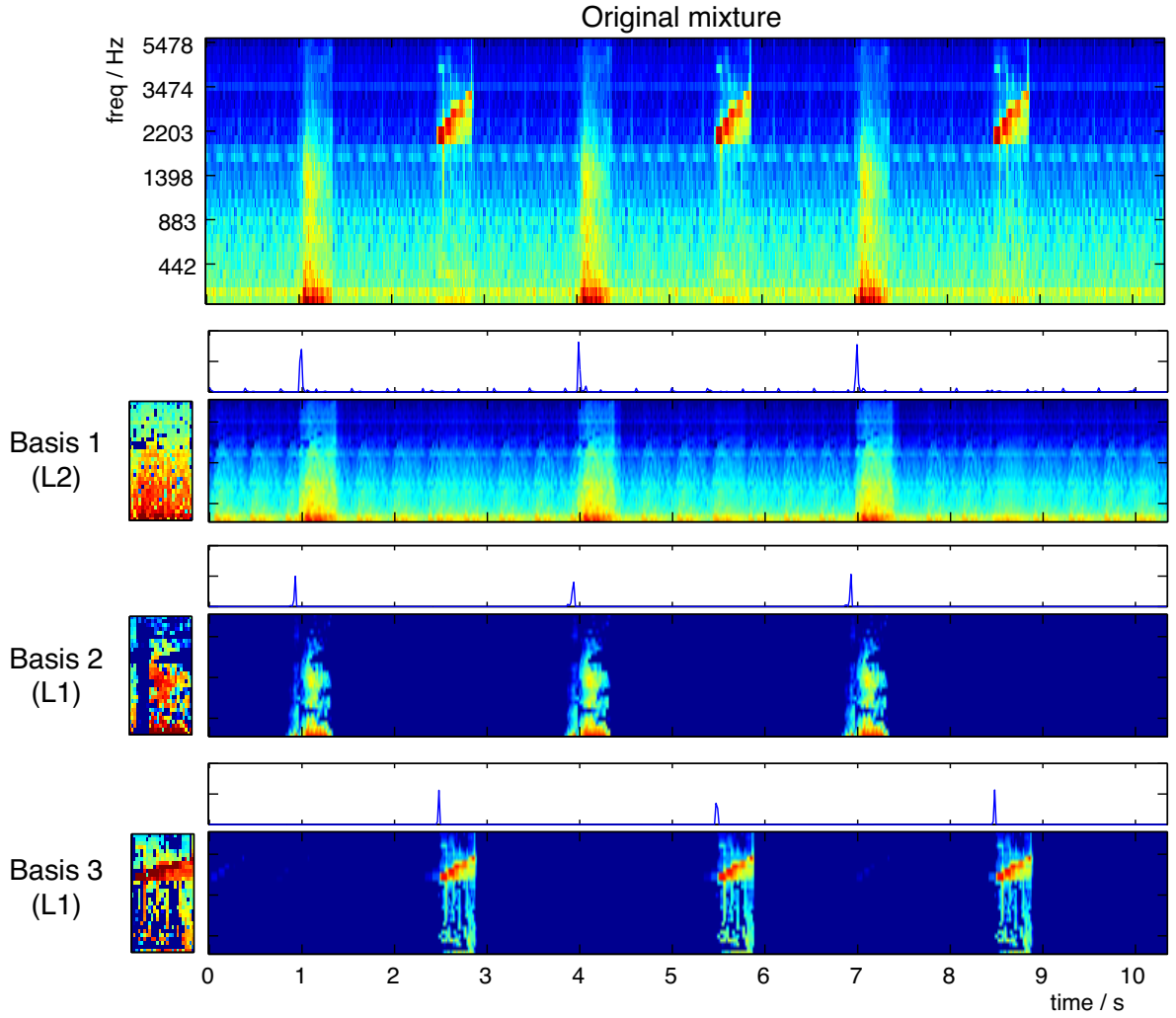
Nonnegative Matrix Factorization

Smaragdis Brown '03
Abdallah Plumbley '04
Virtanen '07

- Decompose spectrograms into templates + activation

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

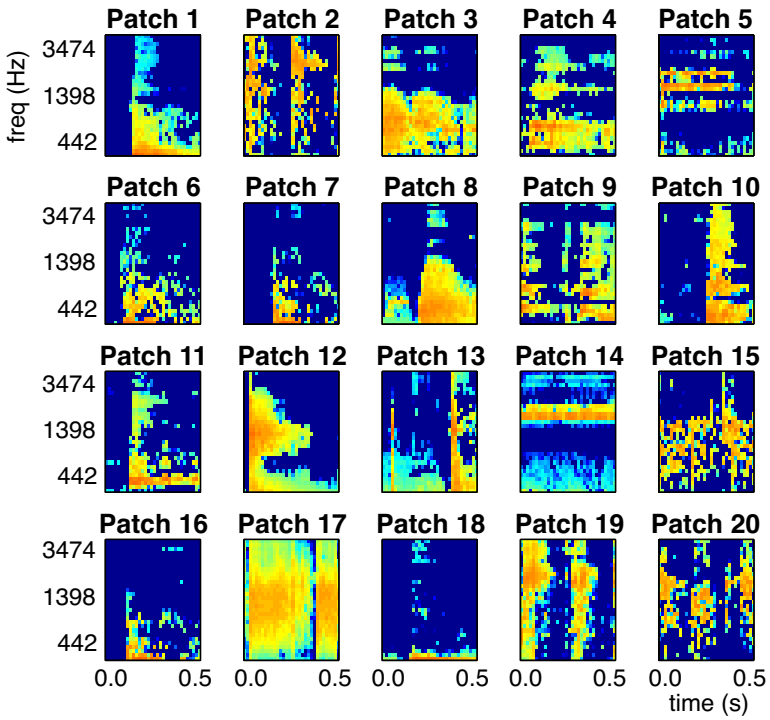
  - fast forgiving gradient descent algorithm
  - 2D patches
  - sparsity control...

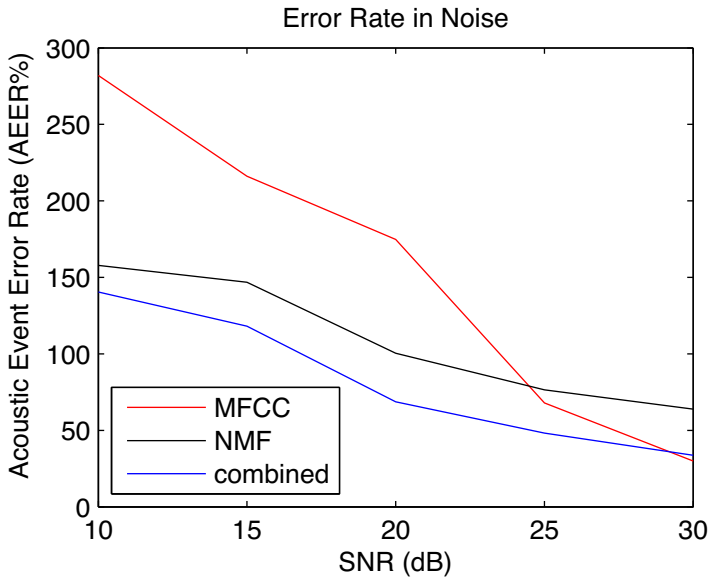


NMF Transient Features

- Learn 20 patches from Meeting Room Acoustic Event data
- Compare to MFCC-HMM detector

Cotton, Ellis '11



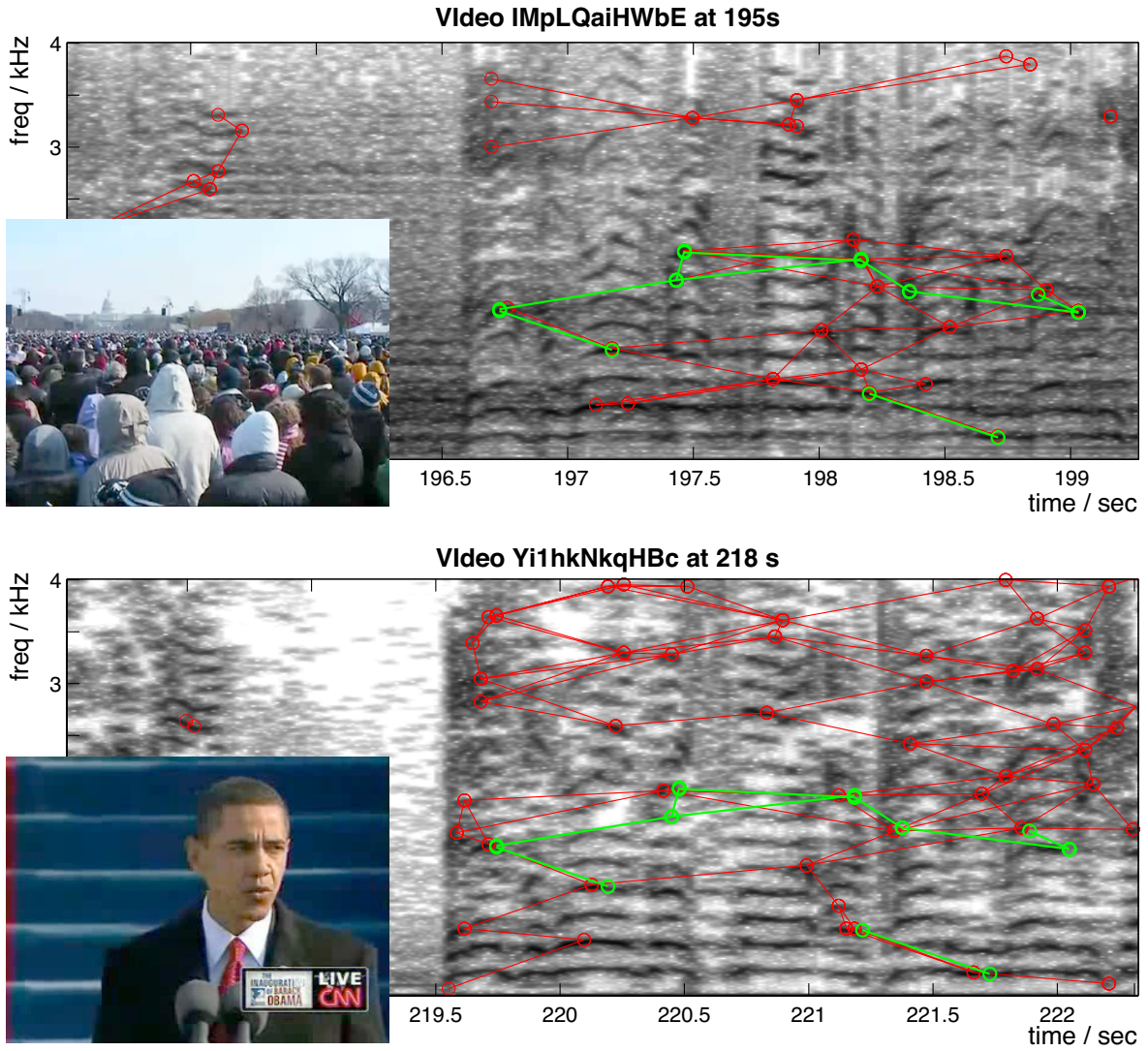


- NMF more noise-robust
  - combines well ...

Matching Videos via Fingerprints

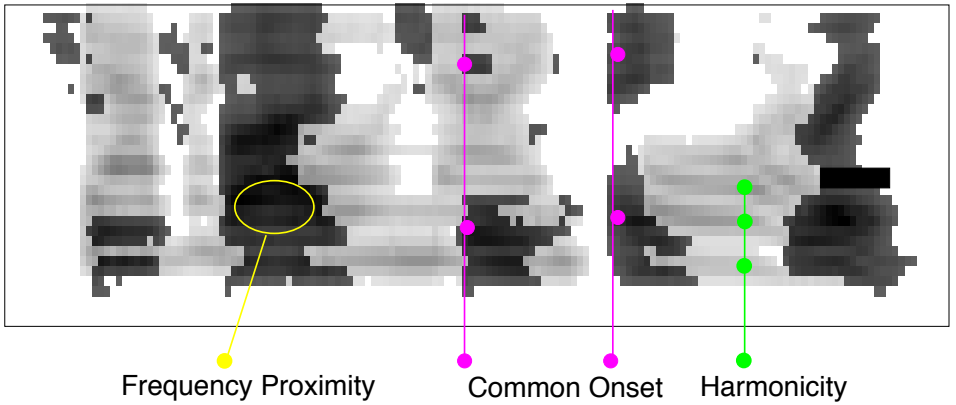
Cotton & Ellis '10

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



4. Outstanding Issues

- Better object/event separation
  - parametric models
  - spatial information?
  - computational auditory scene analysis...



Barker et al. '05

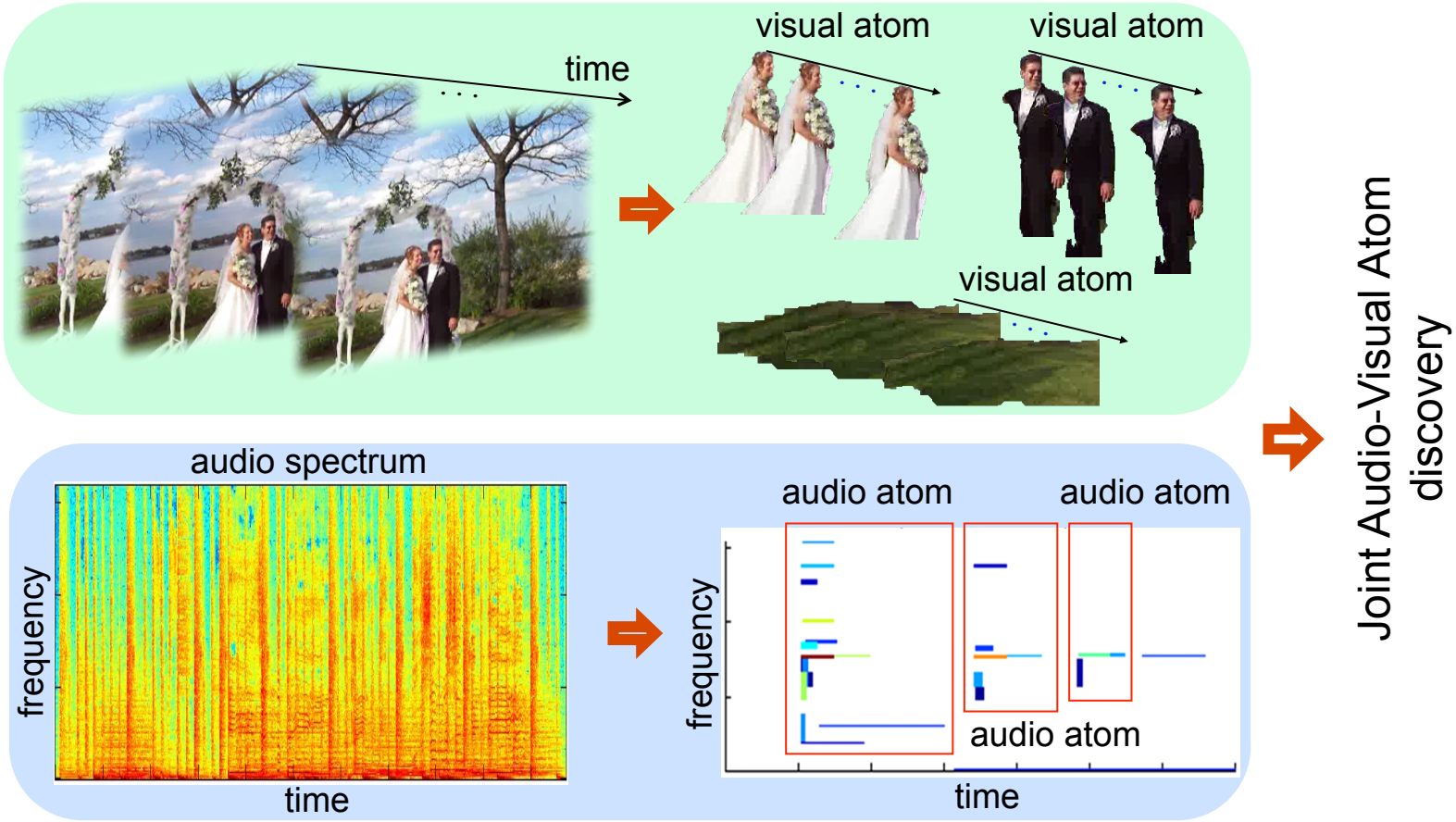
- Large-scale analysis

- Integration with video

Audio-Visual Atoms

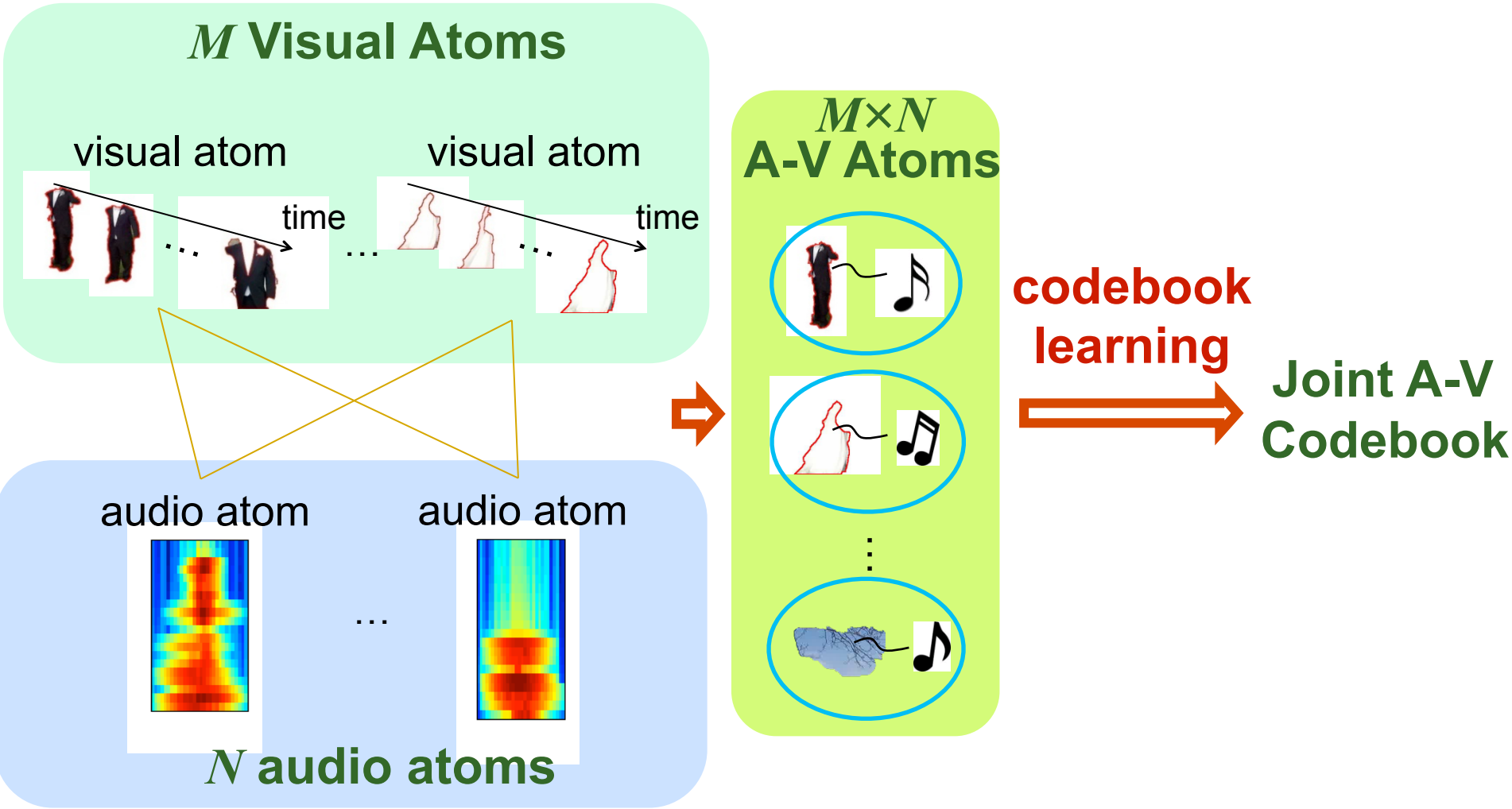
Jiang et al. '09

- Object-related features
  from both audio (transients) & video (patches)



Audio-Visual Atoms

- Multi-instance learning of A-V co-occurrences



Audio-Visual Atoms

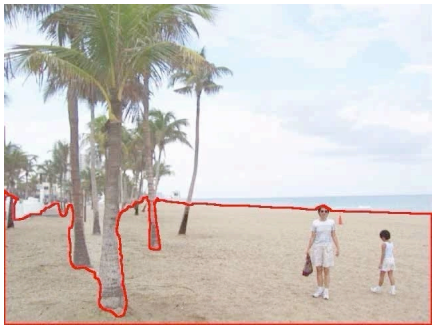
black suit
+ romantic
music

Wedding

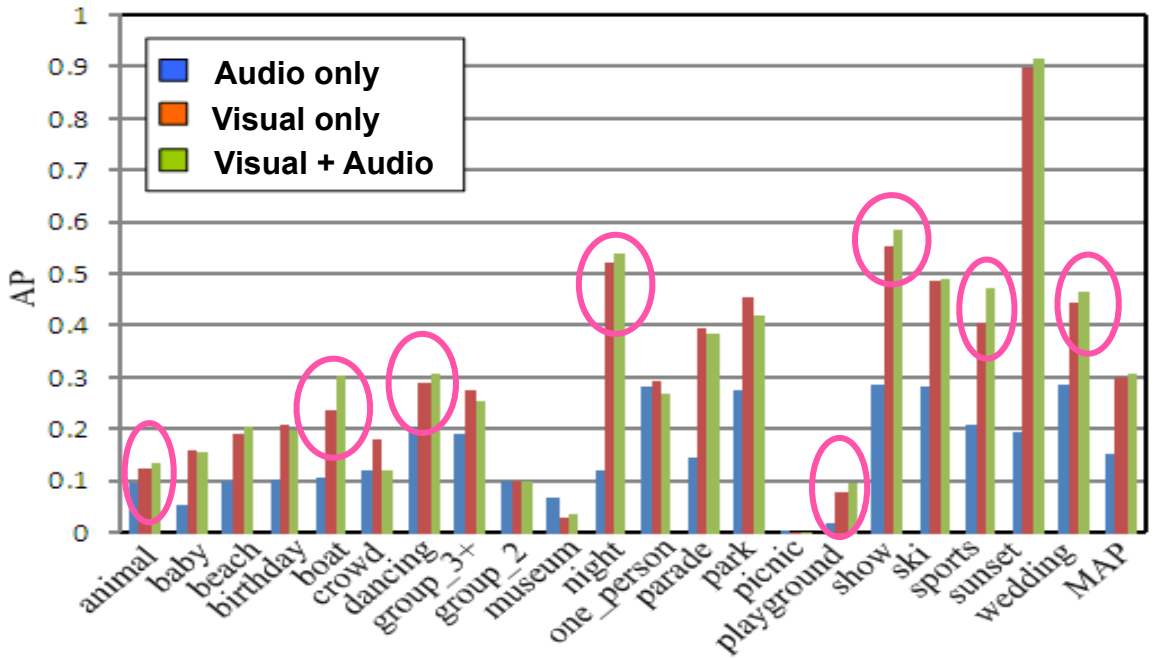
marching
people
+ parade
sound

Parade

sand
+ beach
sounds



Beach



Summary

- Machine Listening:
  Getting useful information from sound
- Background sound classification
  ... from whole-clip statistics?
- Foreground event recognition
  ... by focusing on peak energy patches
- Speech content is very important
  ... separate with pitch, models, ...