Extracting Information from Sound

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/

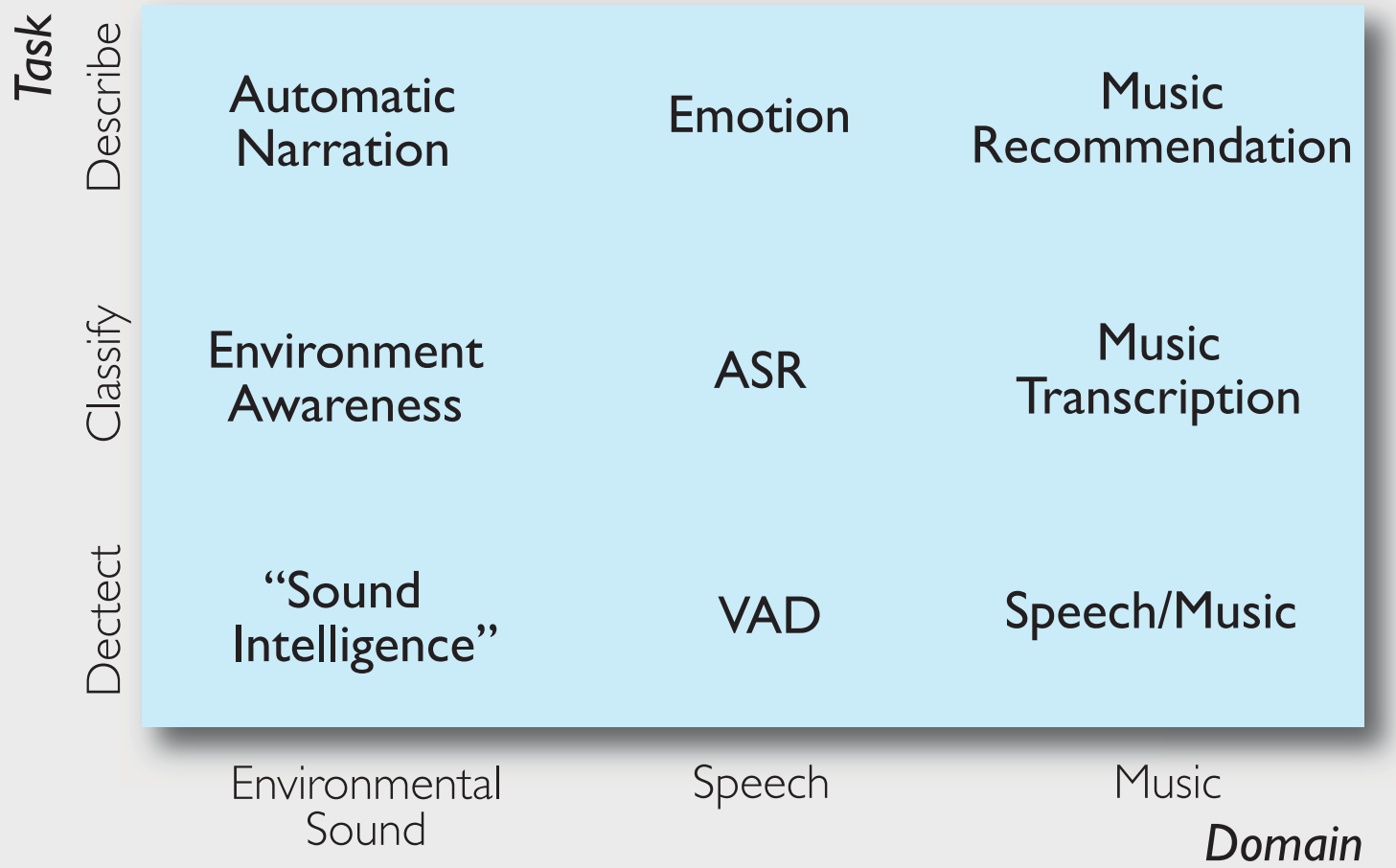
1. Machine Listening
2. Global Classification
3. Foreground & Transients
4. Outstanding Issues

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

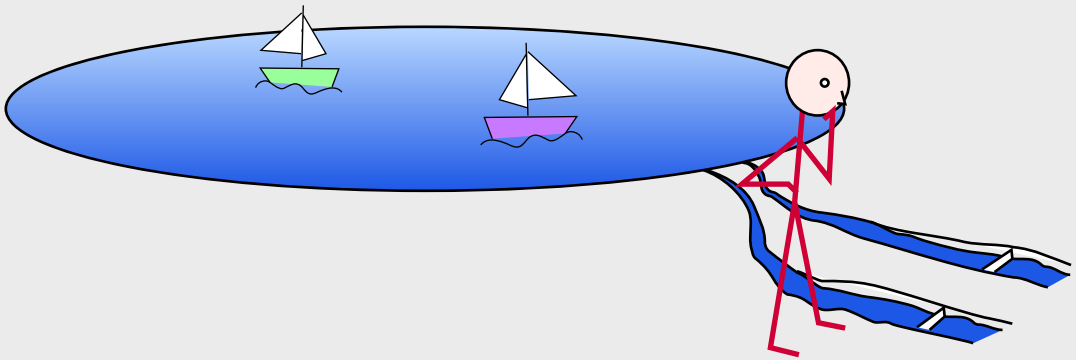
Columbia University
in the City of New York

1. Machine Listening

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



Listening to Mixtures

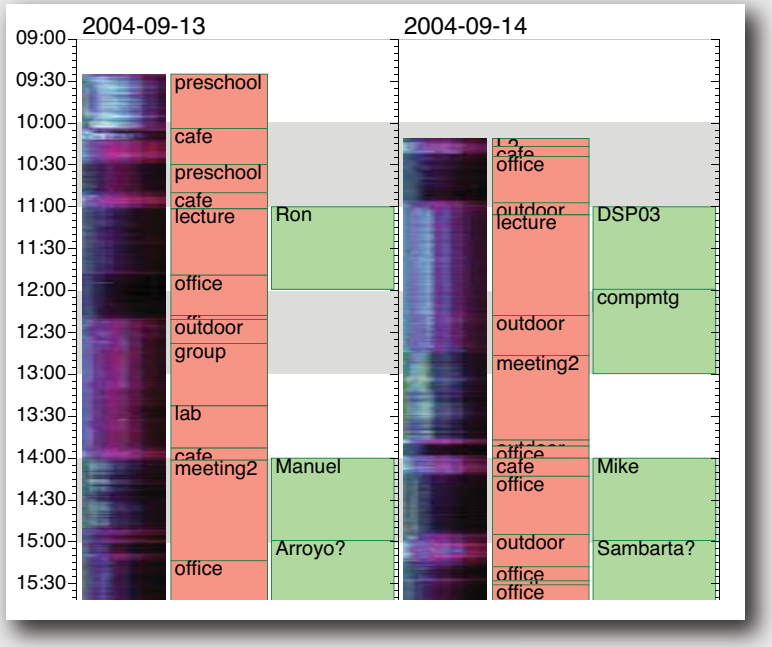


- The world is cluttered
  & sound is transparent
  - mixtures are inevitable

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

Applications

- Audio Lifelog Diarization



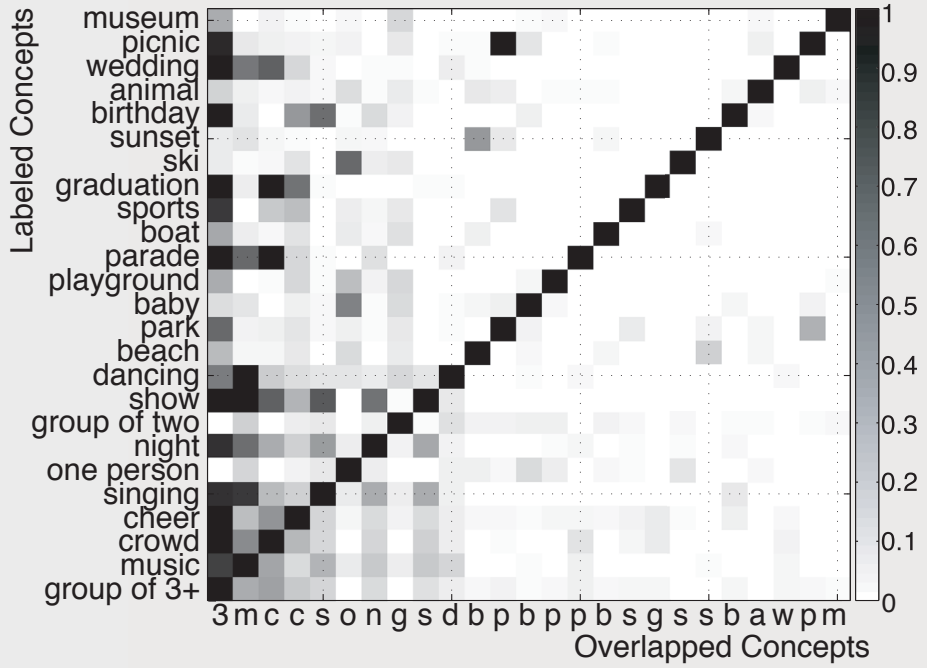
- Consumer Video Classification

Consumer Video Dataset

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

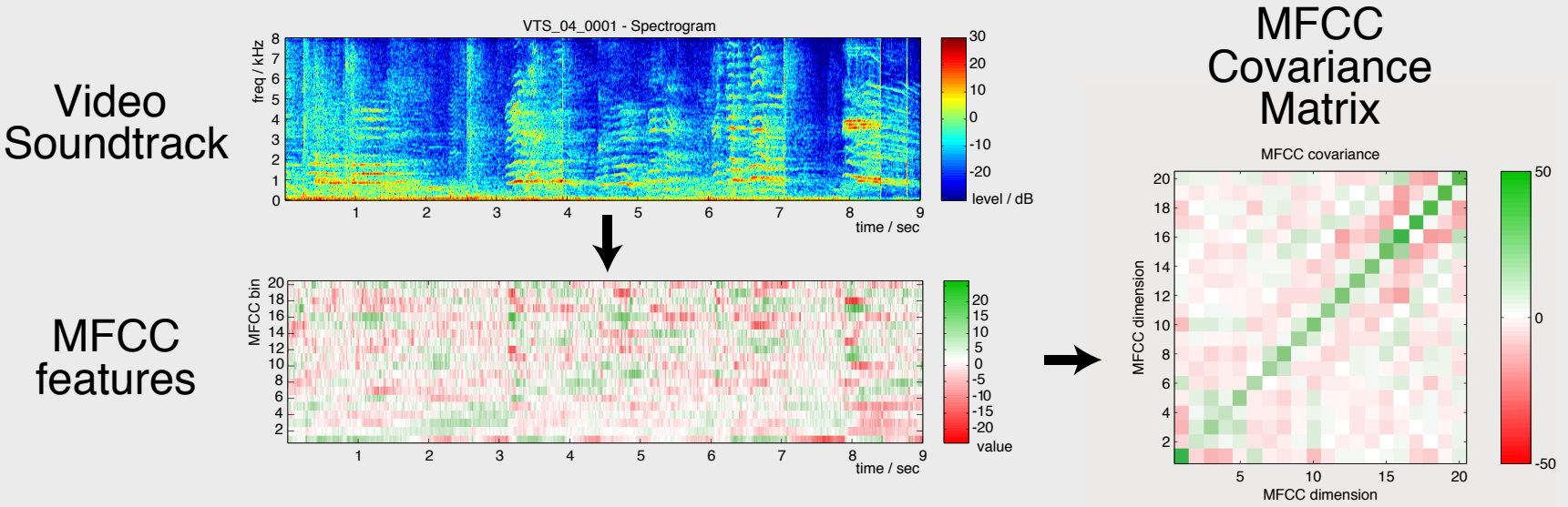
- Grab top 200 videos from YouTube search
  - then filter for quality, unedited = 1873 videos
  - manually relabel with concepts

- Concept overlap:



2. Global 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"



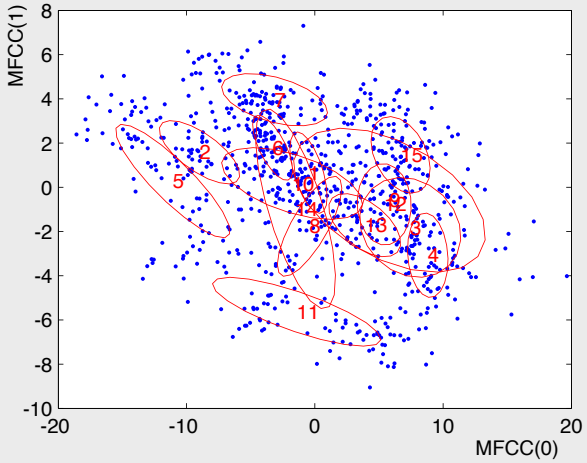
- Classify by e.g. Mahalanobis distance + SVM

Codebook Histograms

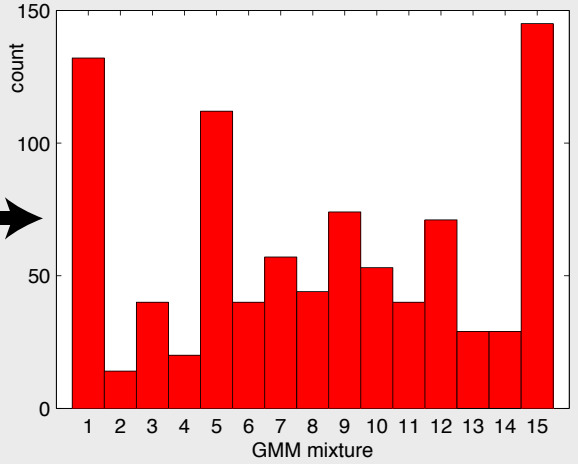
- Convert nonplanar distributions to multinomial

MFCC features →

Global Gaussian Mixture Model



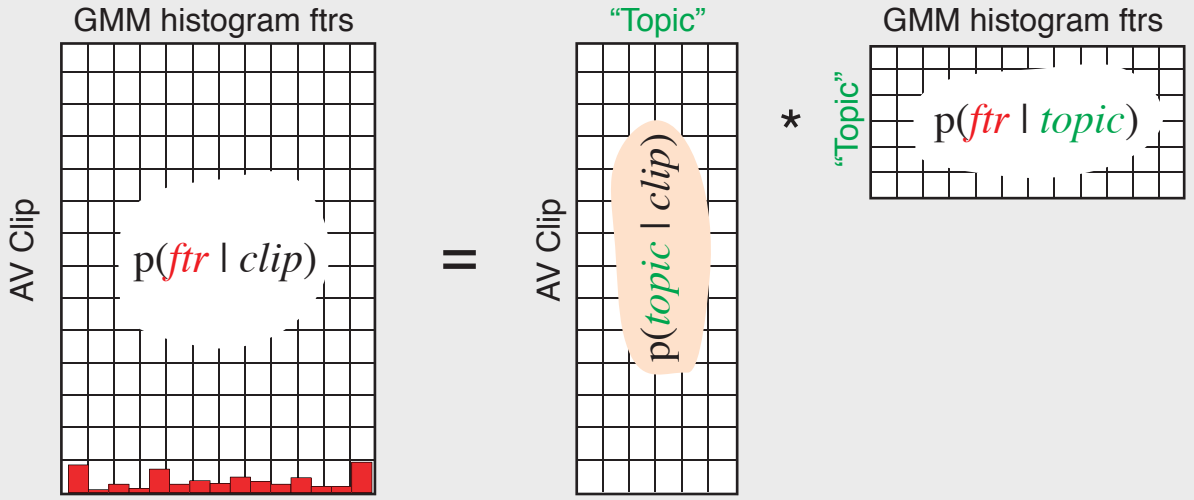
→ **Per-Category** Mixture Component Histogram



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

Latent Semantic Analysis (LSA)

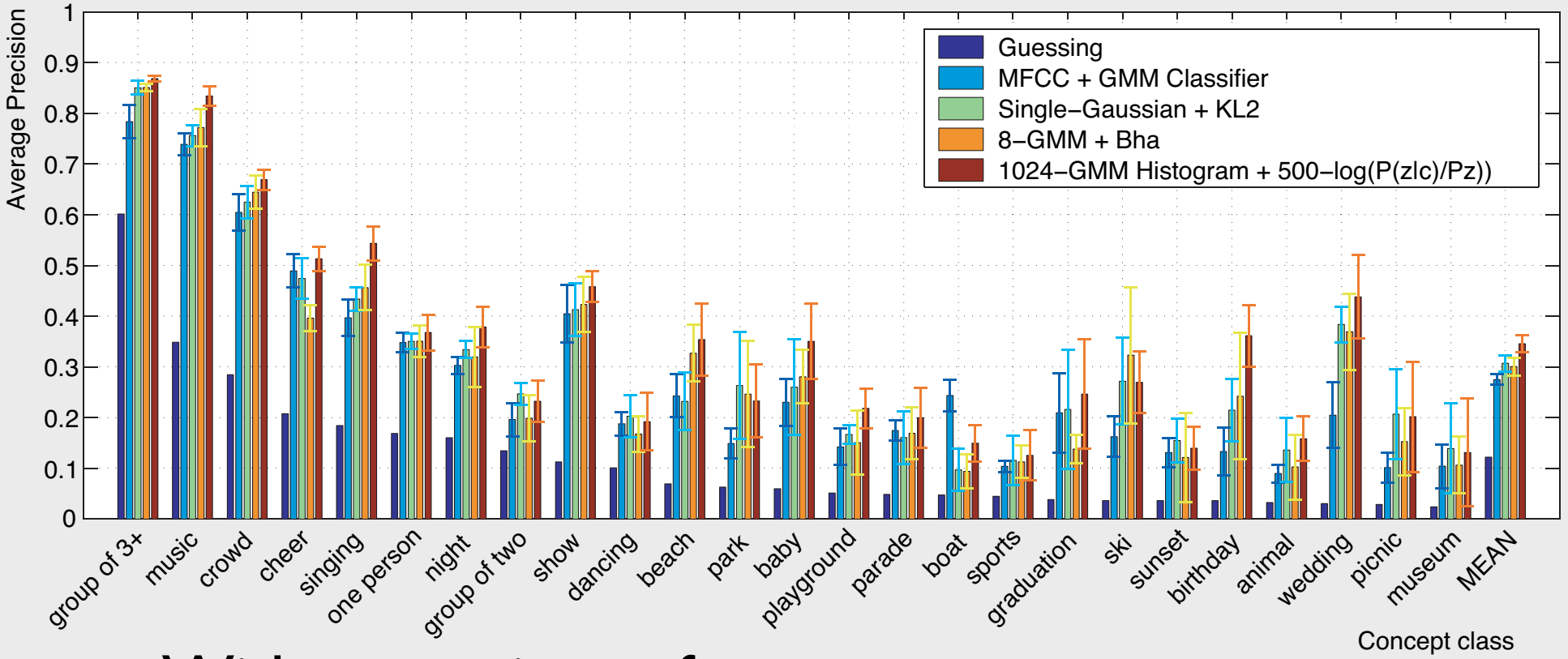
- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - $p(ftr \mid clip) = \sum_{topics} p(ftr \mid topic)\, p(topic \mid clip)$



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

Global Classification Results

Lee & Ellis '10

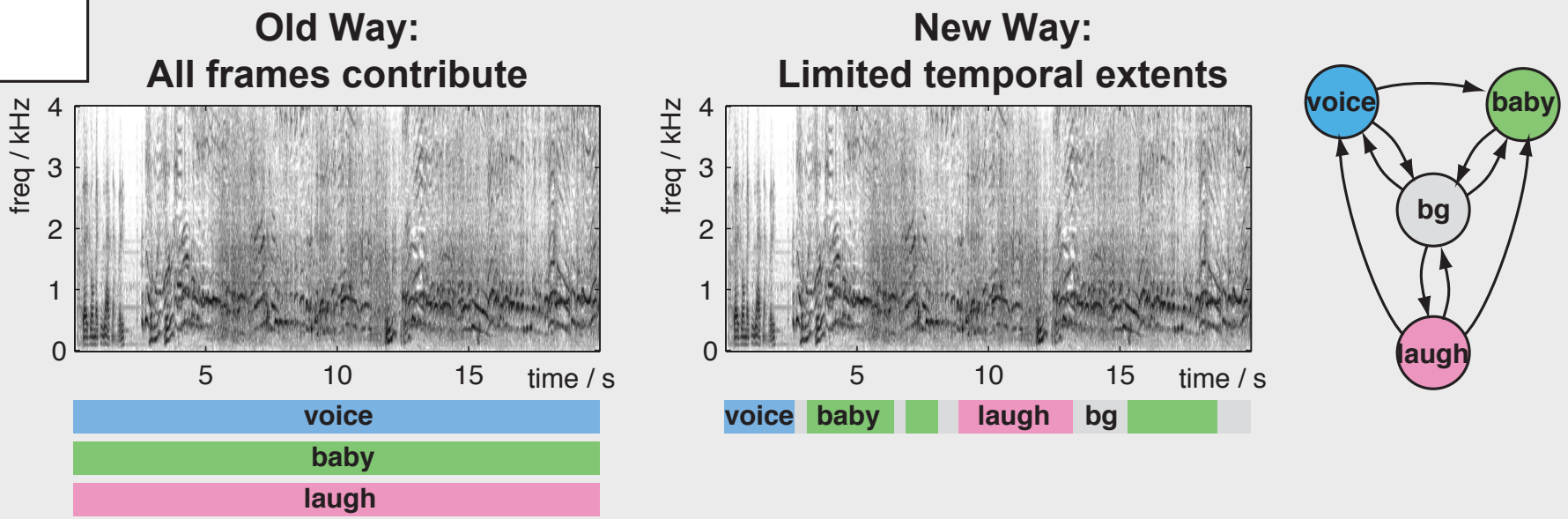


- Wide range in performance
  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

3. Foreground & Transients

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

YT baby 002:
voice
baby
laugh

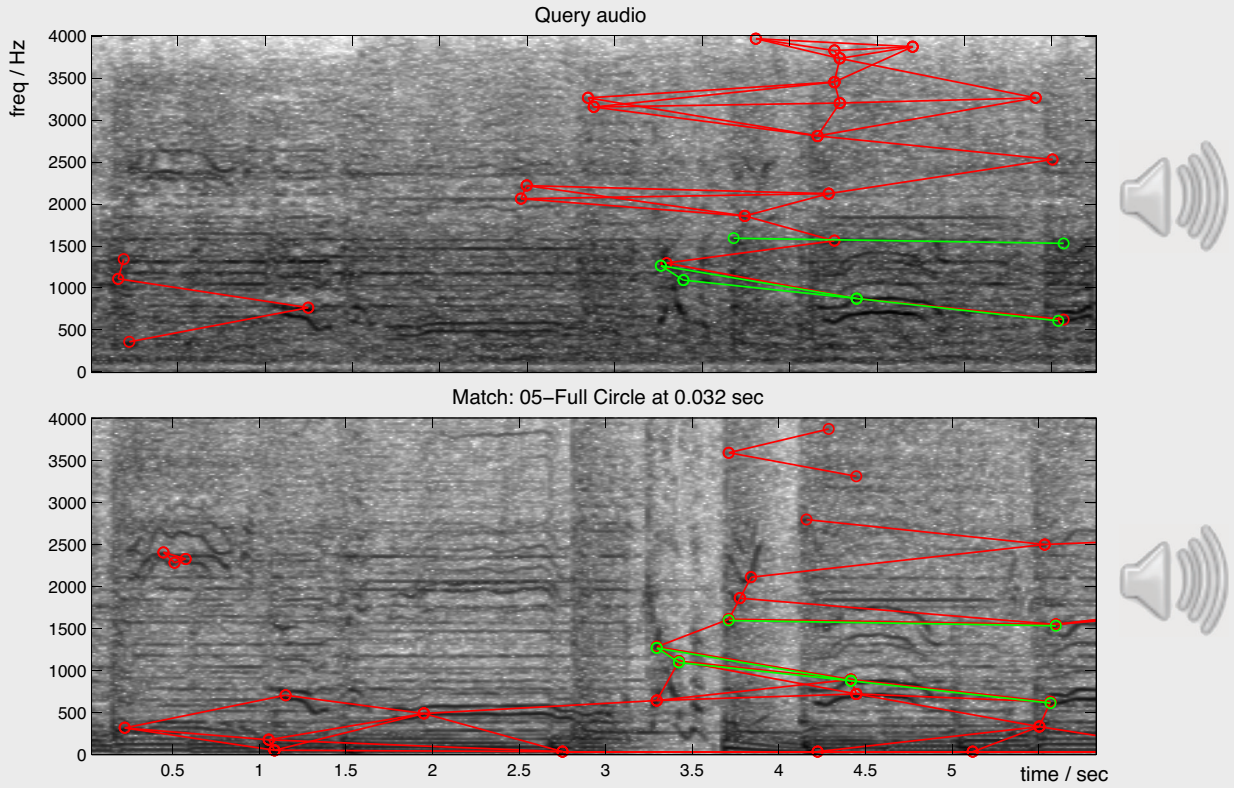


  - "background" model shared by all clips

Landmark-based Fingerprints

Shazam '03

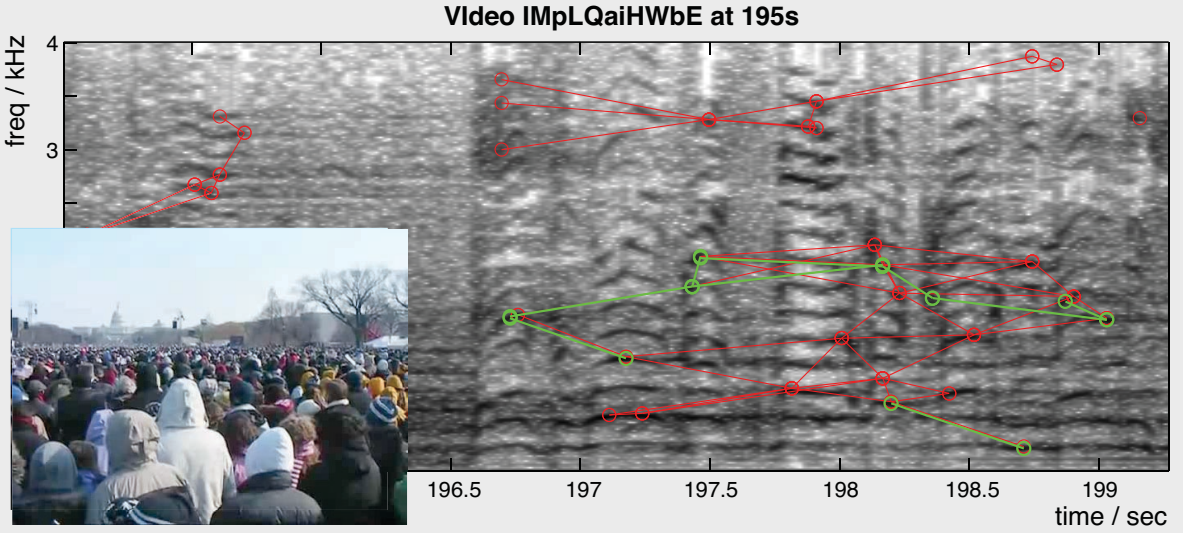
- Sound characterized by time-frequency peaks
  - robust to channel, background
  - relies on precise timing

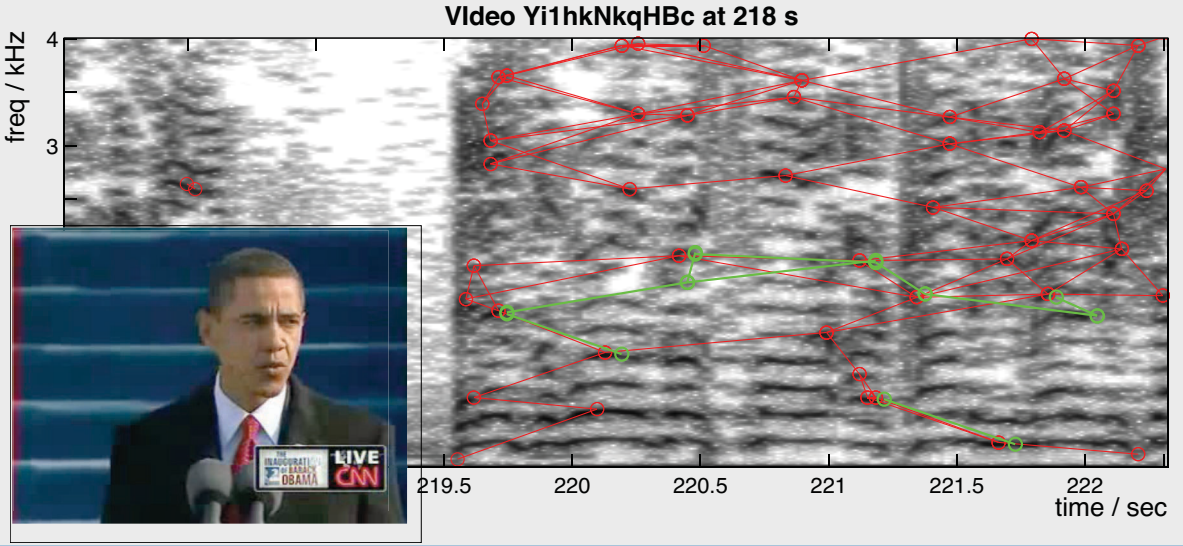


Soundtrack Fingerprint Matching

Cotton & Ellis '10

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

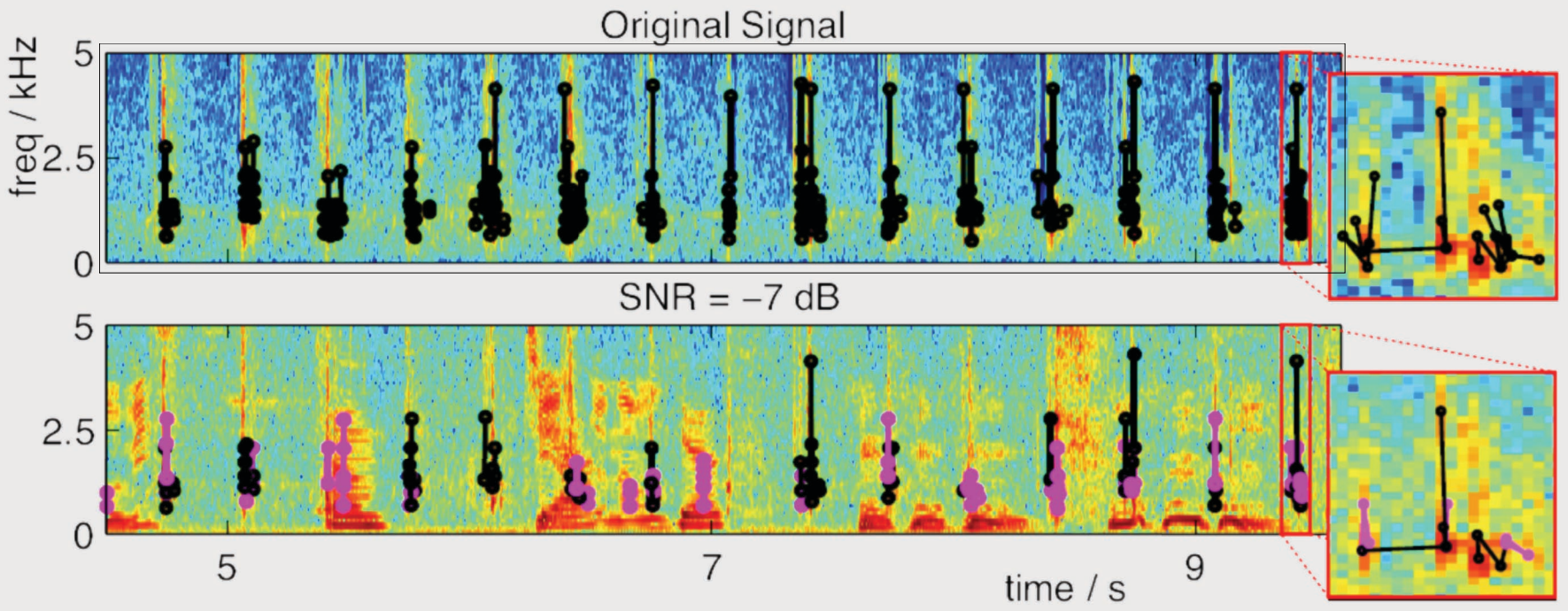




Event Landmark Signatures

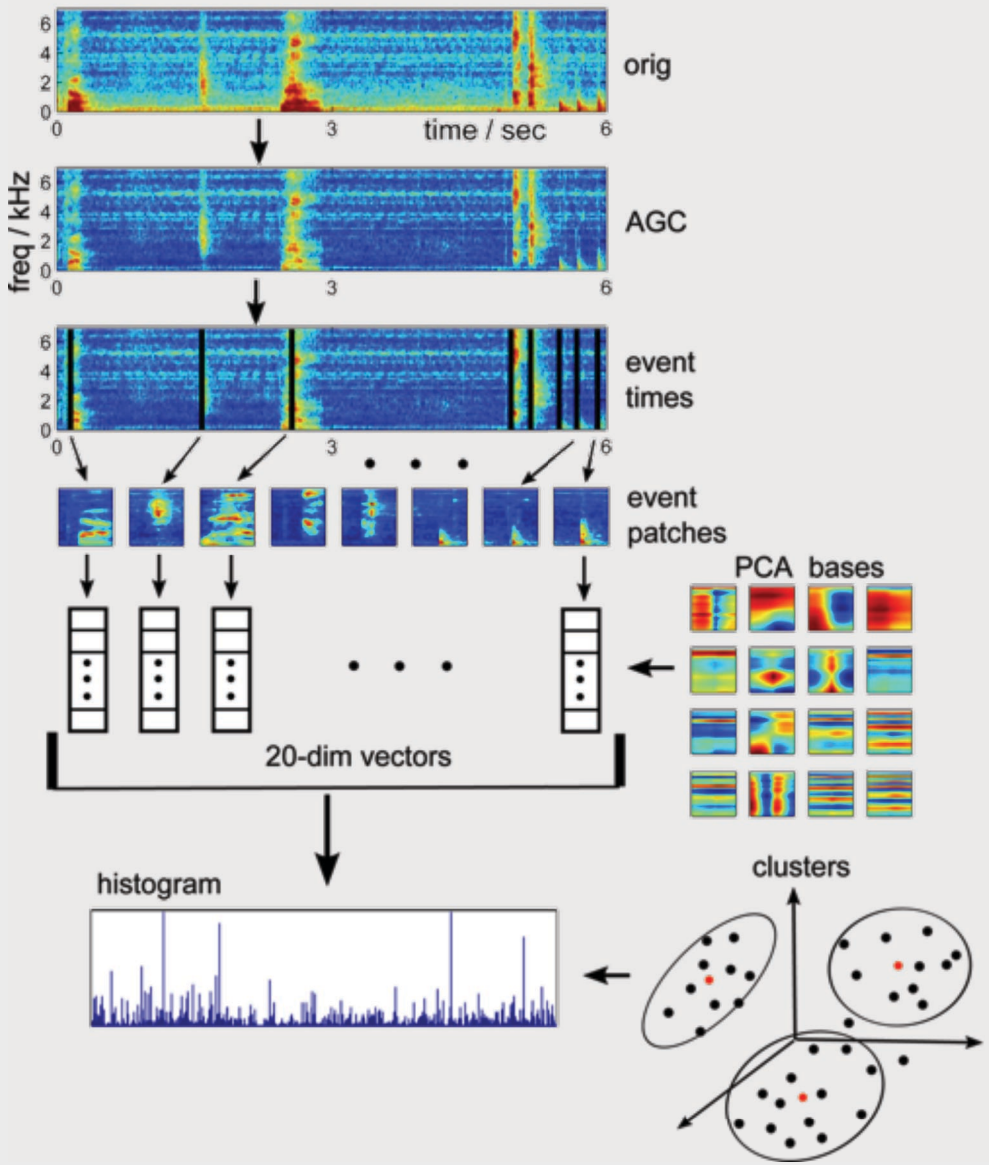
Cotton & Ellis '09

- Build index of Gabor neighbor pairs
  - recognize repeated events with similar pairs



Transient Features

Cotton, Ellis, Loui '11



- Onset detector finds energy bursts
  - best SNR
- PCA basis to represent each
  - 300 ms x auditory frq
- "bag of transients"

4. Outstanding Issues

- How to define "transients"?
- How to separate foreground & background?
- How to exploit prior knowledge of sounds?
- How to make classification discriminative?
- Large-scale soundtrack classification

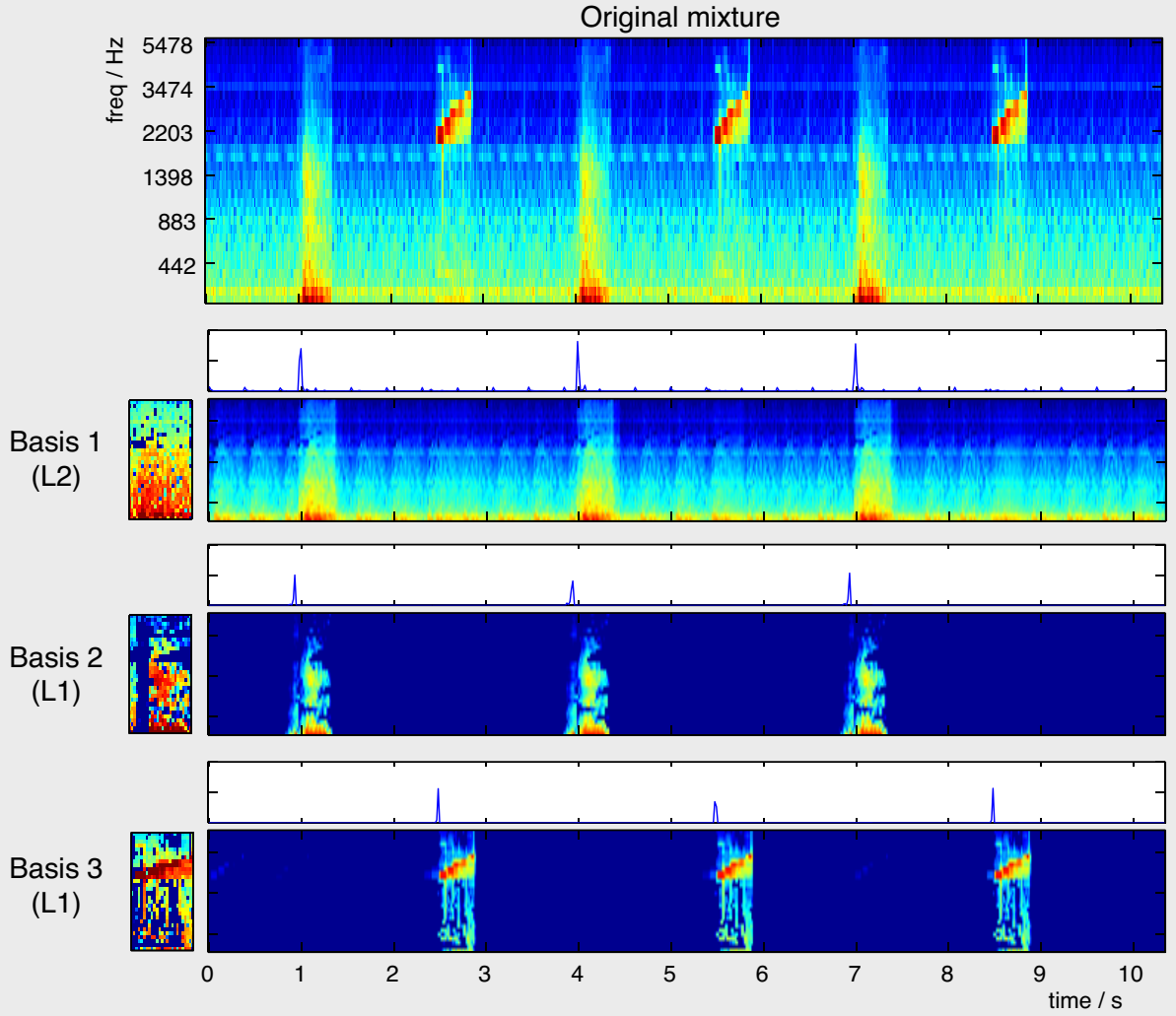
Nonnegative Matrix Factorization

Smaragdis & Brown '03
Abdallah & Plumbley '04
Virtanen '07

- Decompose spectrograms into templates + activation

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

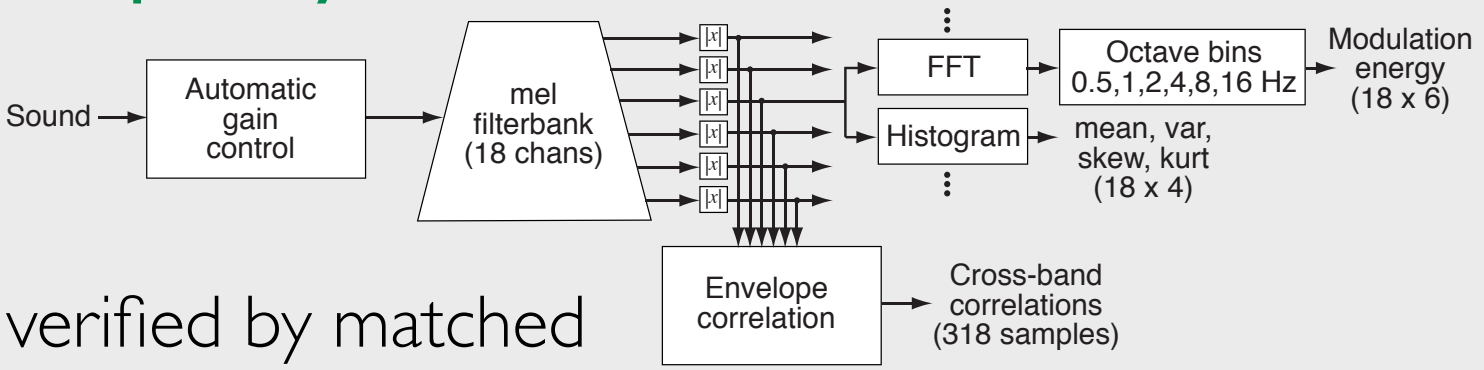
  - fast & forgiving gradient descent algorithm
  - 2D patches
  - sparsity control...



Sound Textures

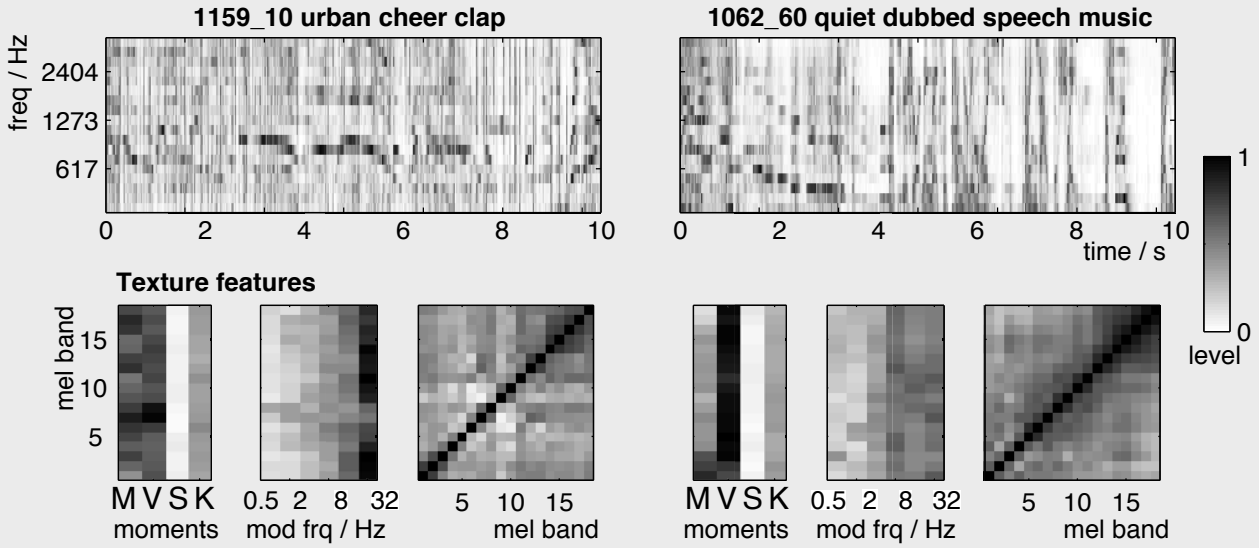
McDermott & Simoncelli '09
Ellis, Zhang, McDermott '11

- Characterize sounds by perceptually-sufficient statistics



  - .. verified by matched resynthesis

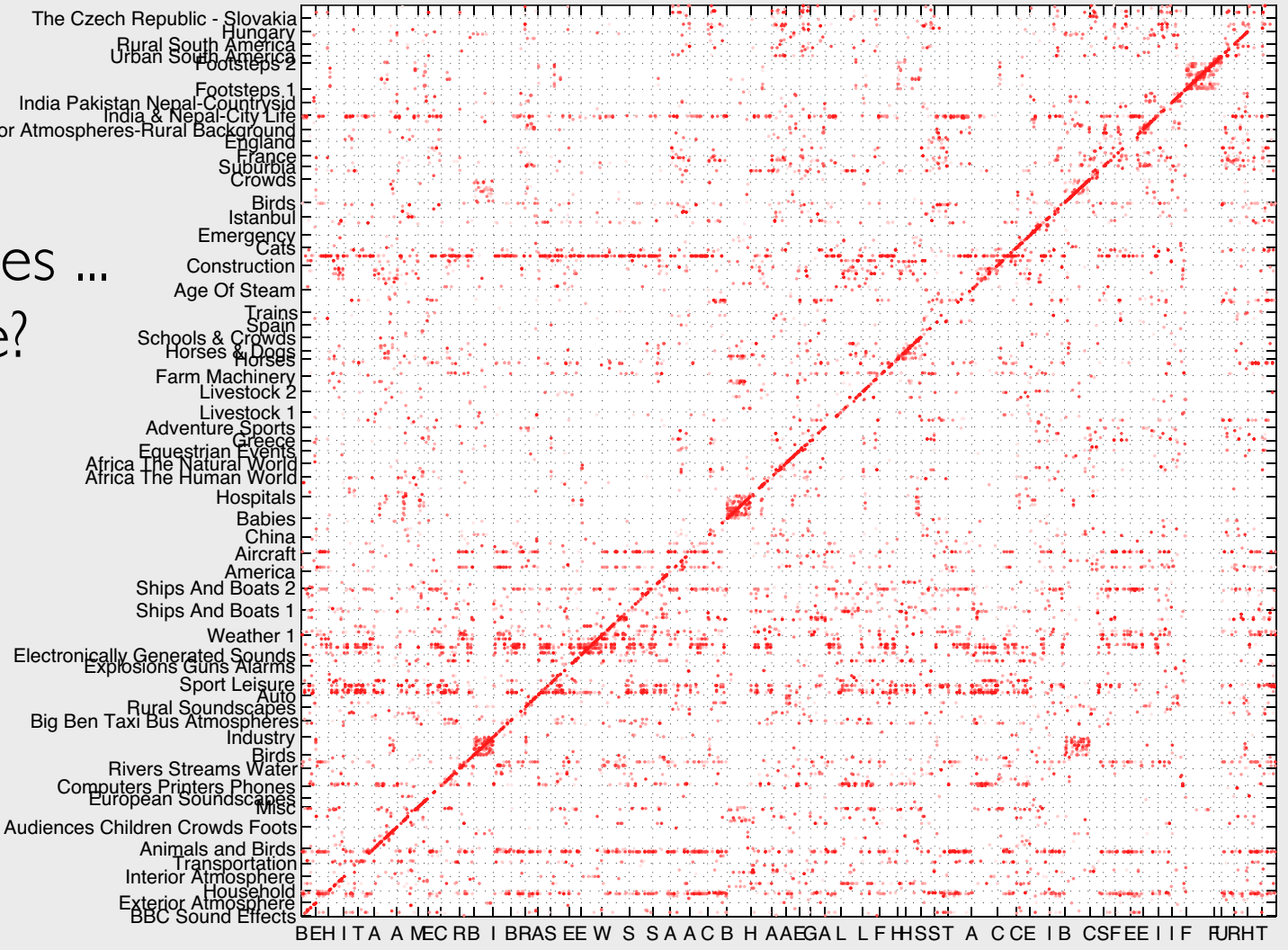
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



Real-World Dictionary

- BBC Sound Effects as reference library

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



Summary

- Machine Listening:
  Getting useful information from sound

- Environmental sound classification
  ... from whole-clip statistics?

- Transients & energy peaks
  ... separate foreground & background

- Useful classification of unconstrained audio
  ... to combine with video analysis