Using the Soundtrack to Classify Videos

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/

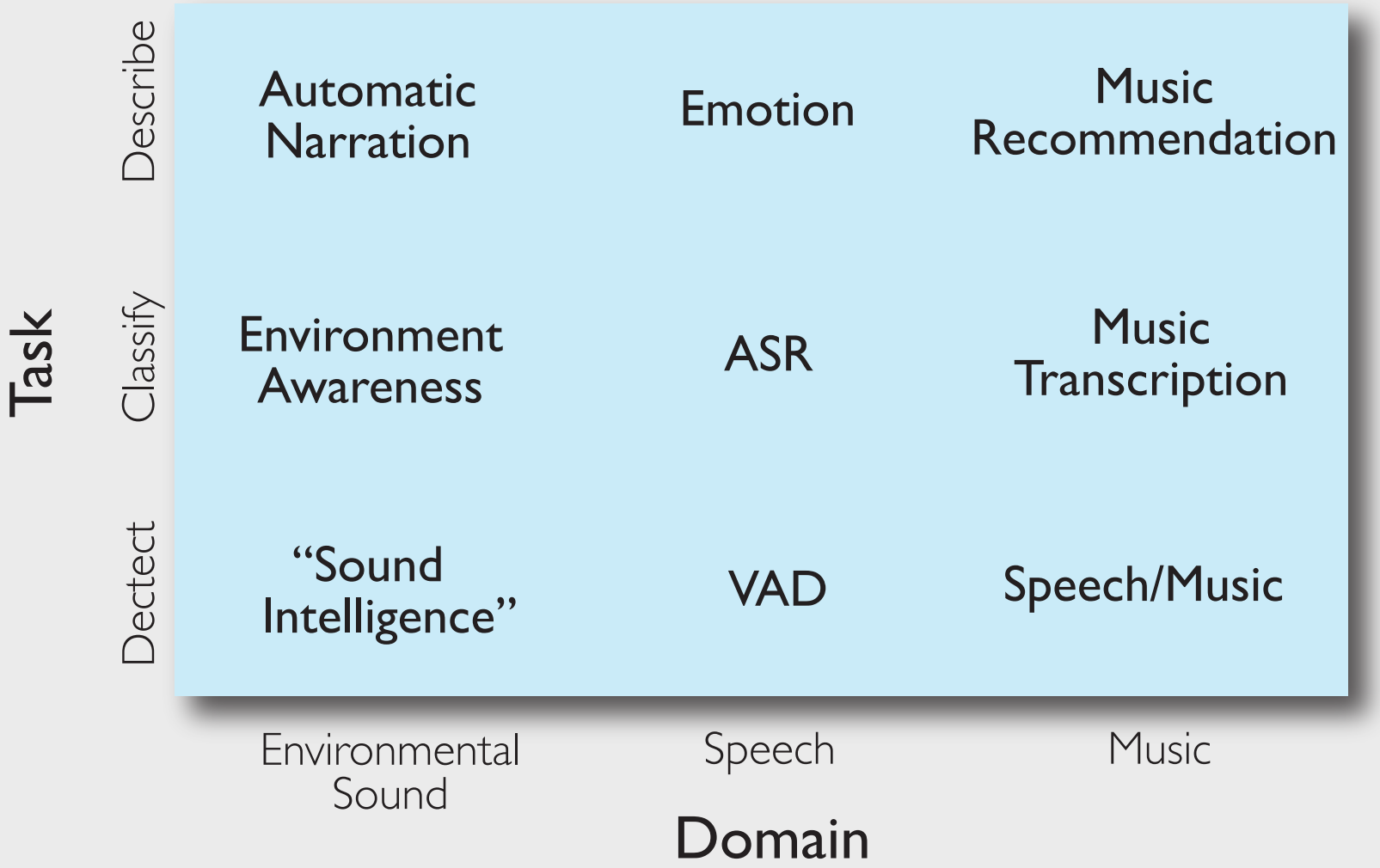
1. Machine Listening
2. Global Classification
3. Foreground Classification
4. Outstanding Issues

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



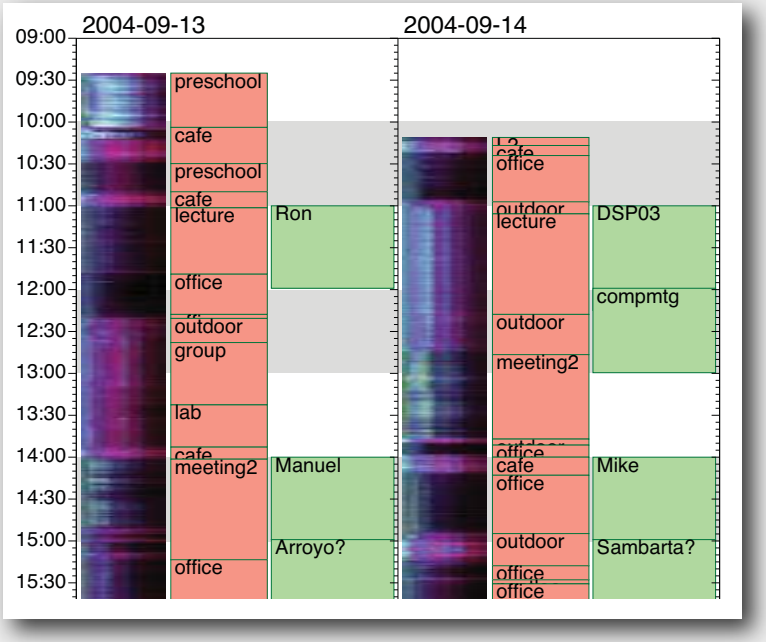
The Information in Audio

- **Environmental recordings contain info on:**
  - location – type (restaurant, street, ...) and specifics
  - activity – talking, walking, typing, ...
  - people – generic (2 males), specific (Chuck & John)
  - spoken content (sometimes)

- **but not:**
  - what people and things "looked like"
  - day/night ...
  - ... except when correlated with audible features

Applications

- Audio Lifelog Diarization



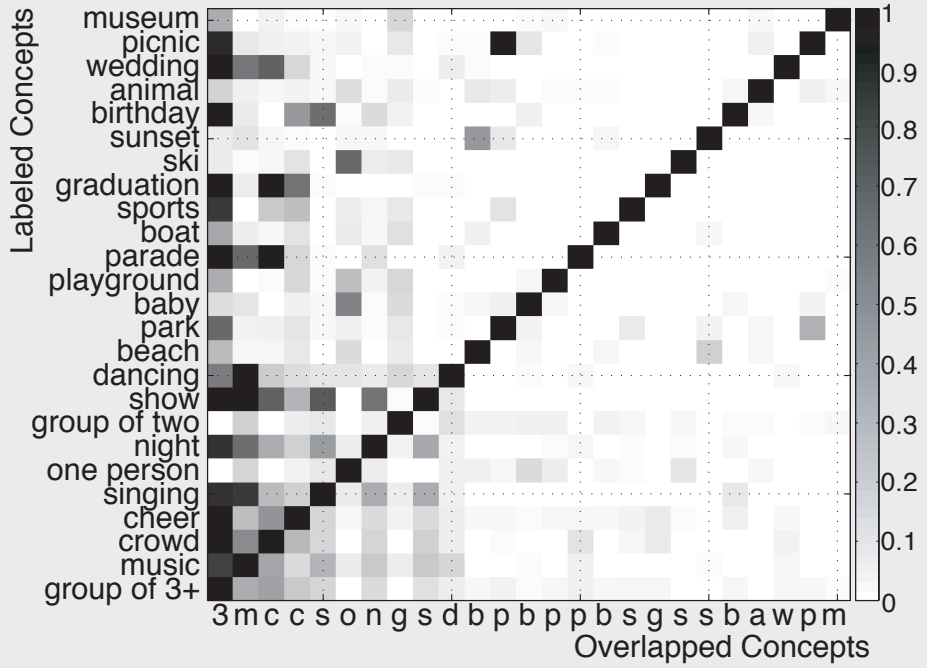
- Consumer Video Classification

Consumer Video Dataset

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

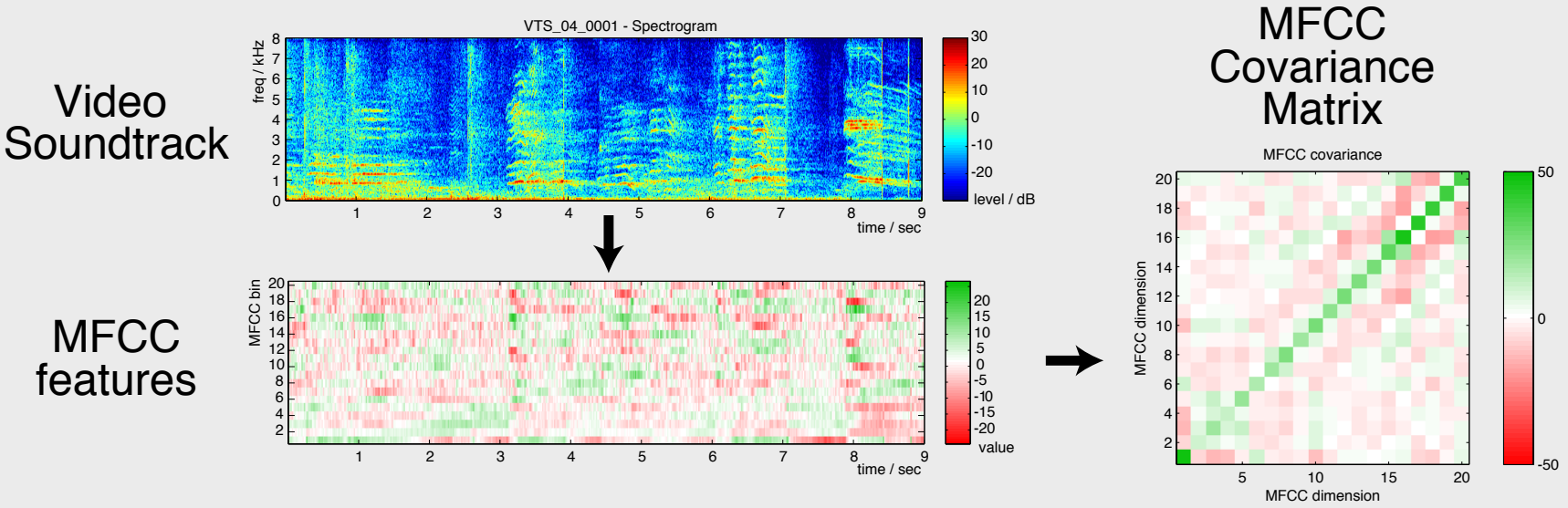
- Grab top 200 videos from YouTube search
  - filter for raw, unedited = 1873 videos
  - manually relabel with concepts

- 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"



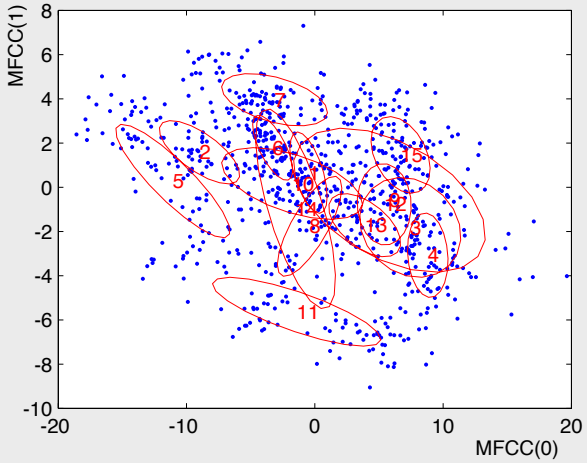
- Classify by e.g. Mahalanobis distance + SVM

Codebook Histograms

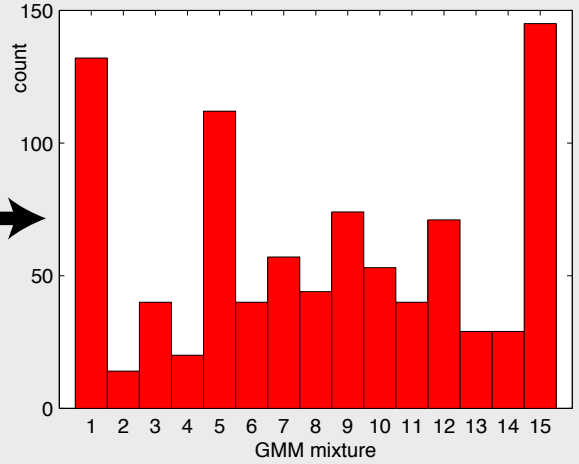
- Convert nonplanar distributions to multinomial

MFCC features →

Global Gaussian Mixture Model



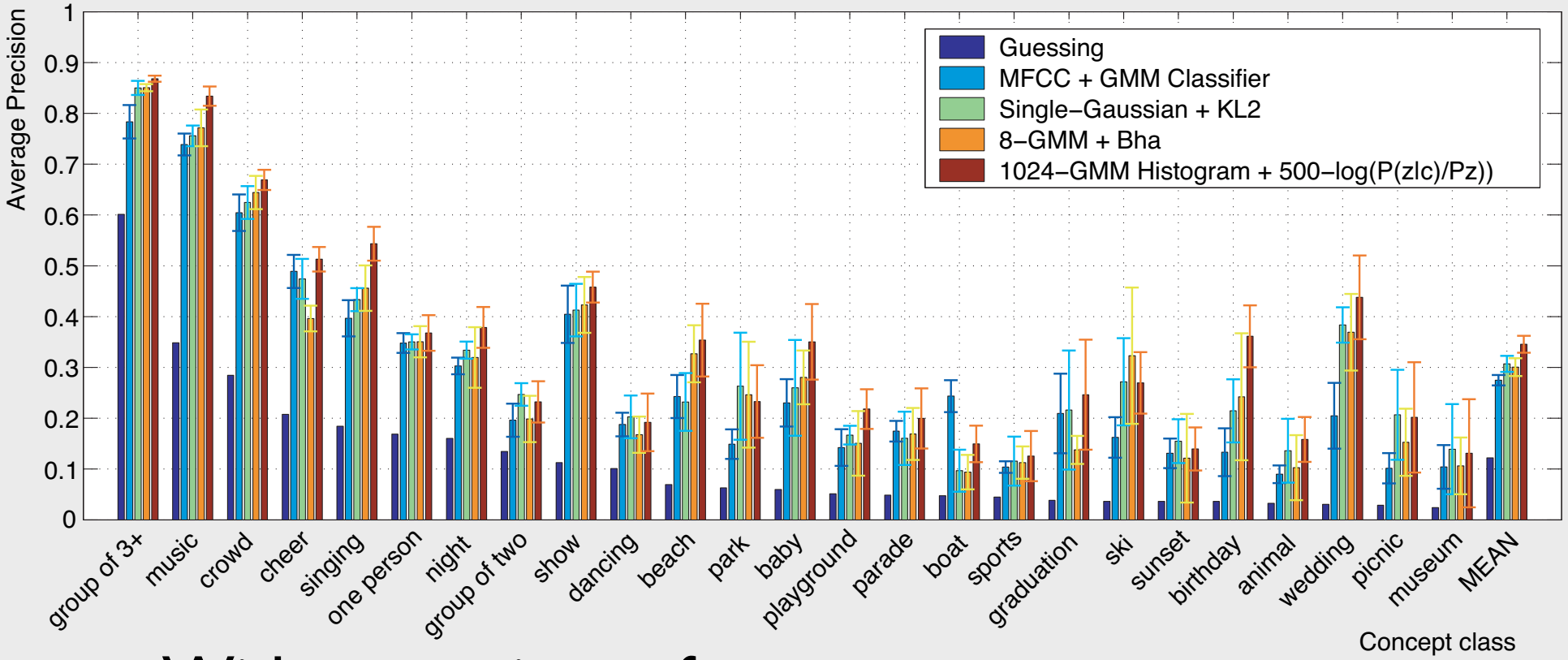
→ **Per-Category** Mixture Component Histogram



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

Global Classification Results

Lee & Ellis '10

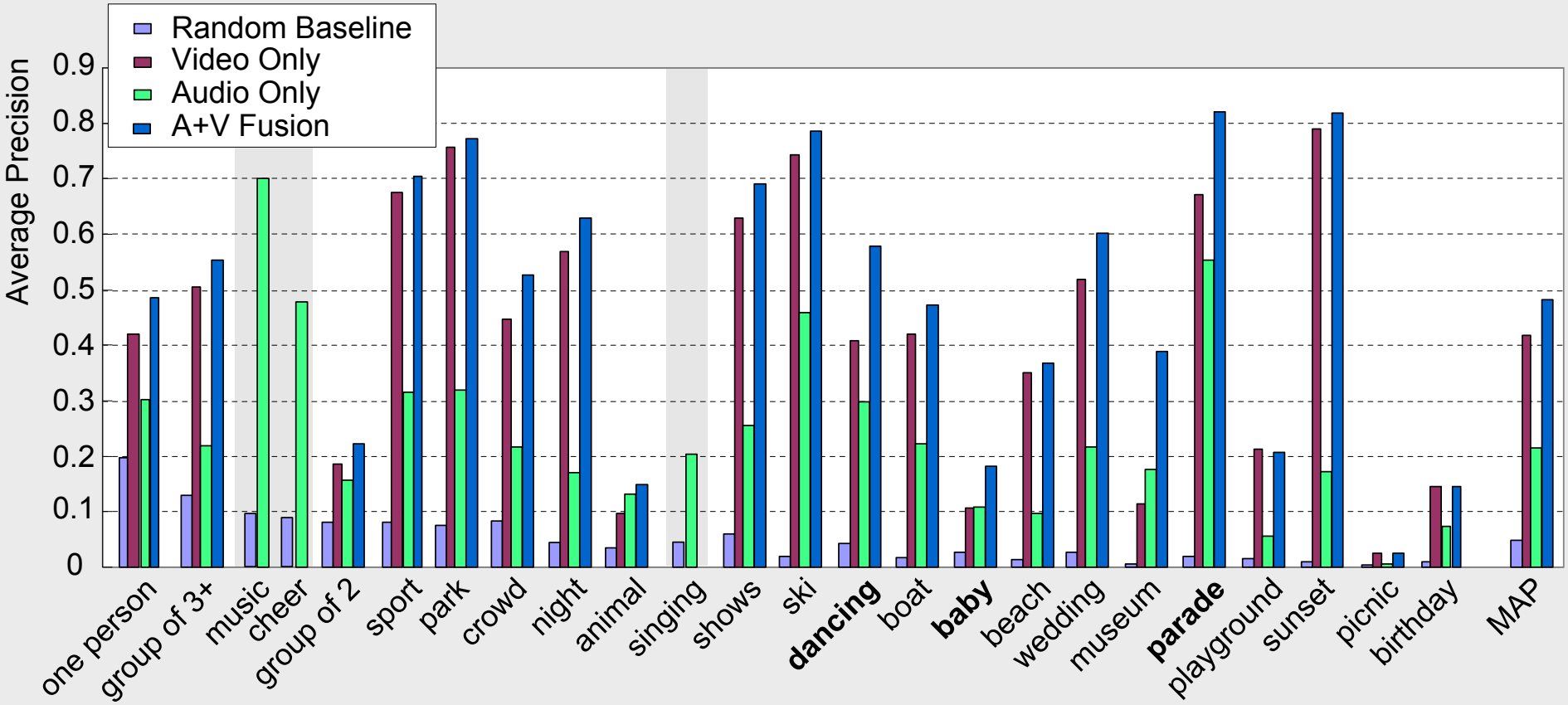


- Wide range in performance
  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

Combining with Video

Chang et al. '07

- Video classification by SIFT codebooks



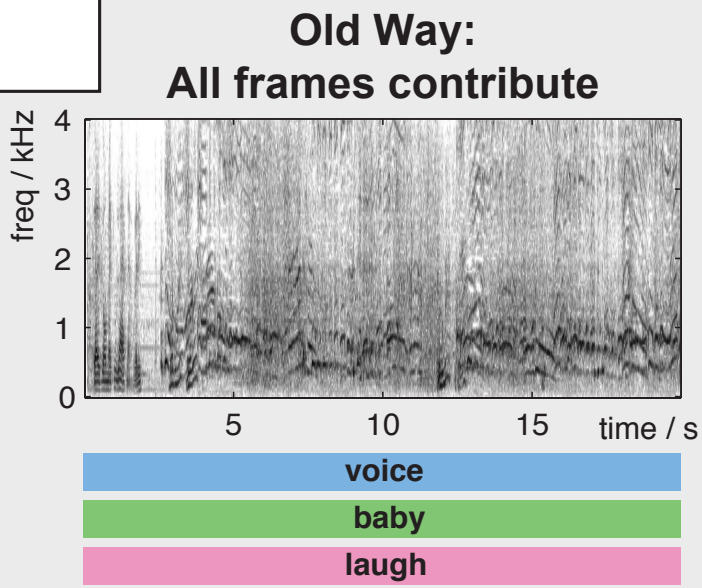
- Audio adds most for dancing, baby, museum ...

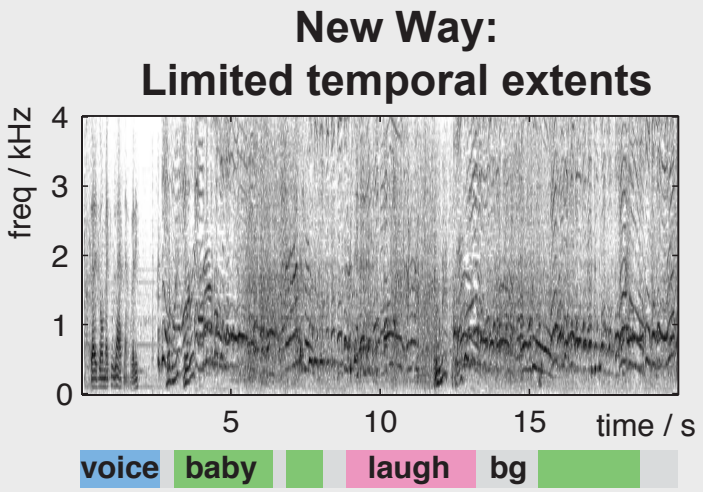
3. Foreground Classification

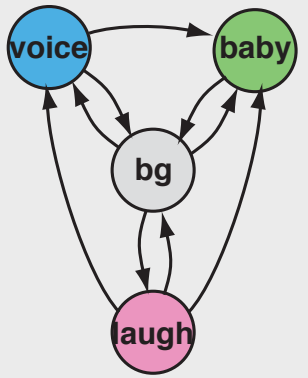
Lee, Ellis & Loui '10

- 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



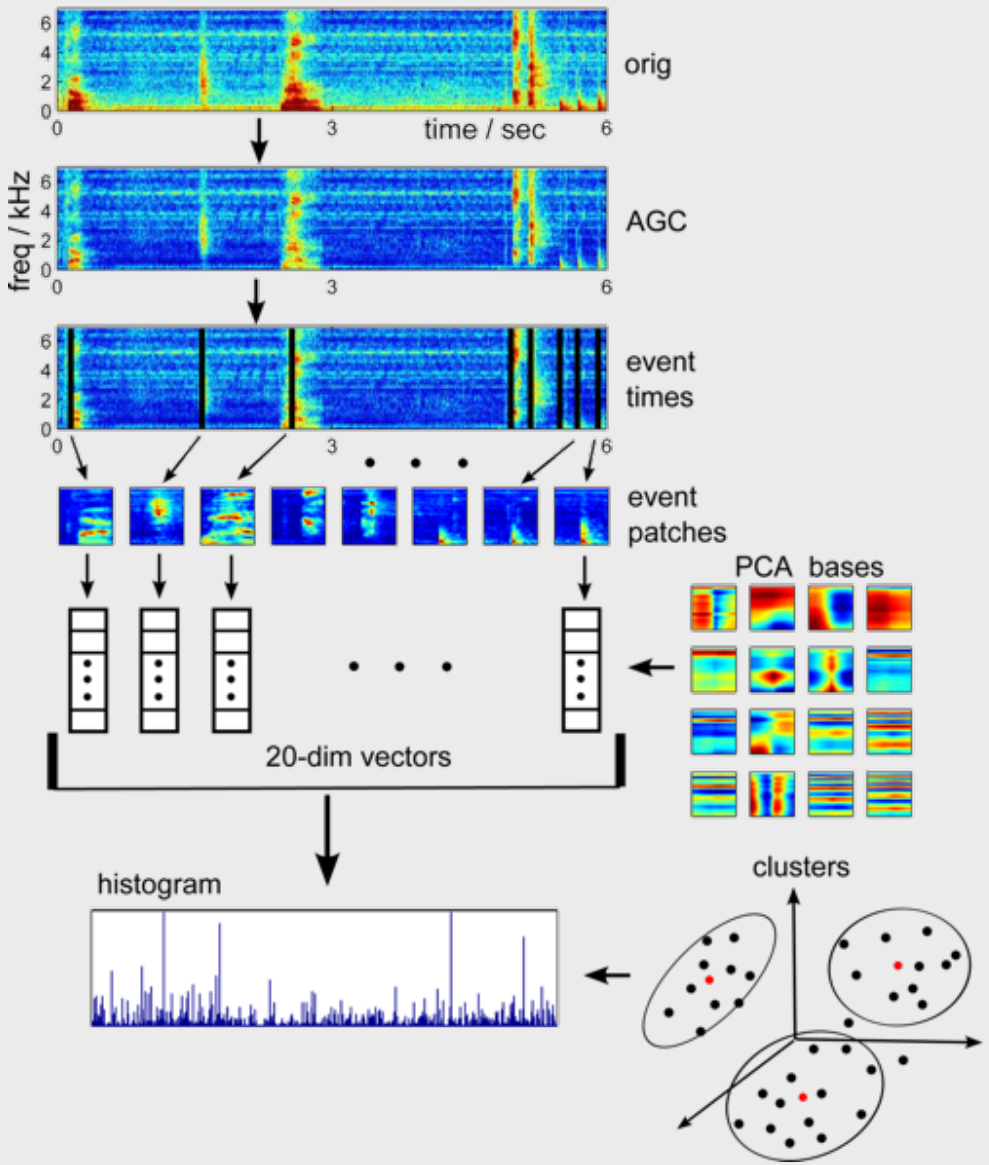




  - "background" model shared by all clips

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"
- Object-related...

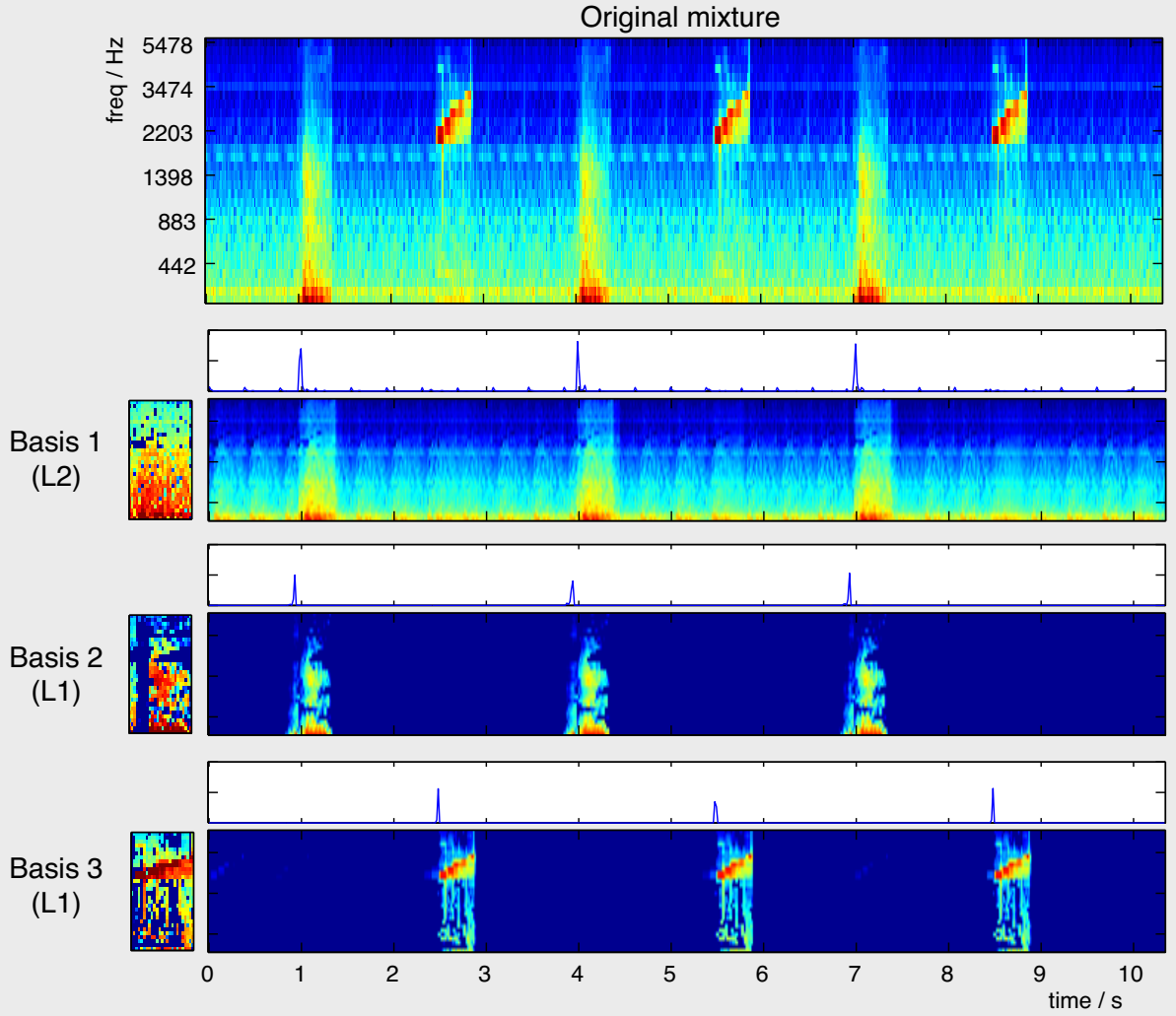
Nonnegative Matrix Factorization

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...

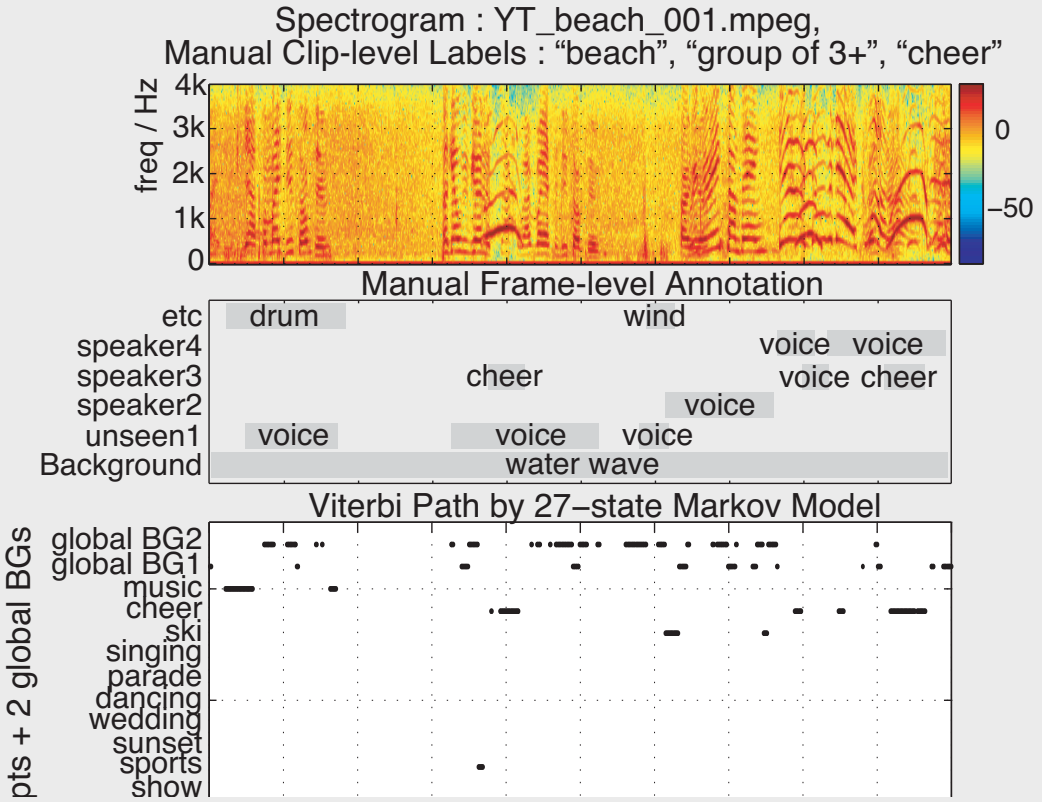


4. Outstanding Issues

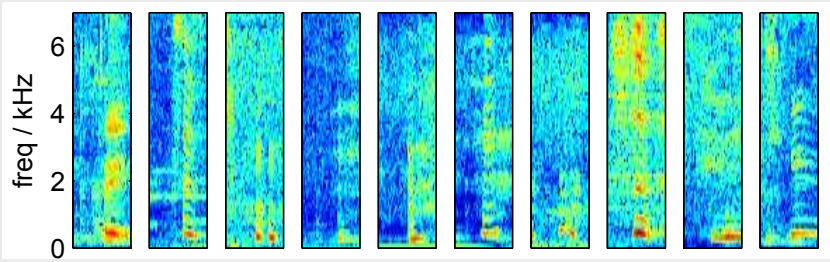
- How to separate foreground & background?
- How to exploit prior knowledge of sounds?
- How to make classification credible?

Classifier Insight

- How can we understand classification results?
  - .. to inspire confidence & enrich results
- Temporal breakdown



- Cluster examples



cluster 123 (baby)

Summary

- Machine Listening:
  Getting useful information from sound

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

- Transients energy peaks
  ... separate foreground background

- Classifier insight
  ... for confidence & analysis

References

- [Chang et al. 2007] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, & J. Luo, "Large-scale multimodal semantic concept detection for consumer video," Proc. ACM MIR, 255-264, 2007.
- [Lee & Ellis 2010] K. Lee & D. Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," IEEE Tr. Audio, Speech, Lang. Process. 18 (6): 1406-1416, Aug. 2010.
- [Lee, Ellis & Loui 2010] K. Lee, D. Ellis, & A. Loui, "Detecting Local Semantic Concepts in Environmental Sounds using Markov Model based Clustering," Proc. ICASSP, 2278-2281, Dallas, 2010.
- [Cotton, Ellis & Loui 2011] C. Cotton, D. Ellis, & A. Loui, "Soundtrack classification by transient events," Proc. ICASSP, to appear, Prague, 2011.
- [Ellis, Zhang & McDermott 2011] D. Ellis, X. Zheng, & J. McDermott, "Classifying soundtracks with audio texture features," Proc. ICASSP, to appear, Prague, 2011.