# Recognizing and Classifying Environmental Sounds

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/

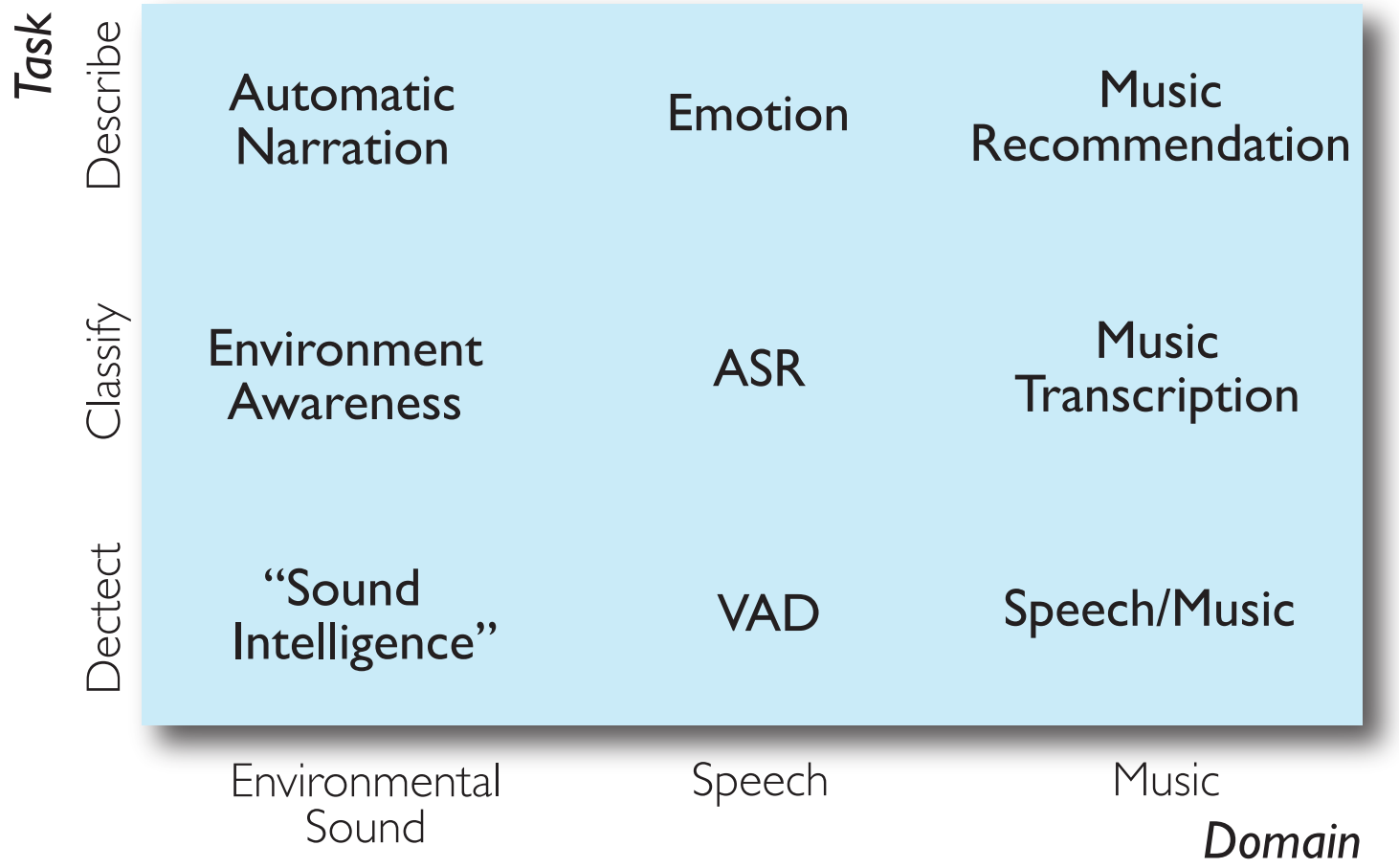
1. Machine Listening
2. Background Classification
3. Foreground Event Recognition
4. Open Issues

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

Columbia University
In the City of New York

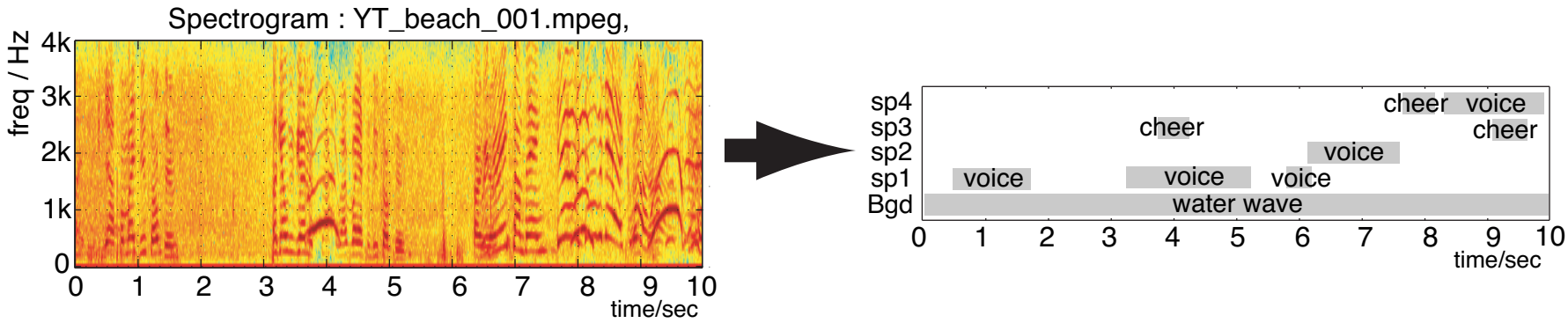
# 1. Machine Listening

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



# Environmental Sound Recognition

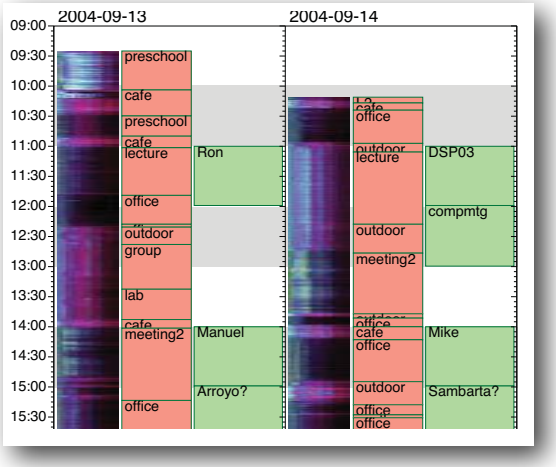
- Goal: Describe soundtracks with a vocabulary of user-relevant acoustic events/sources



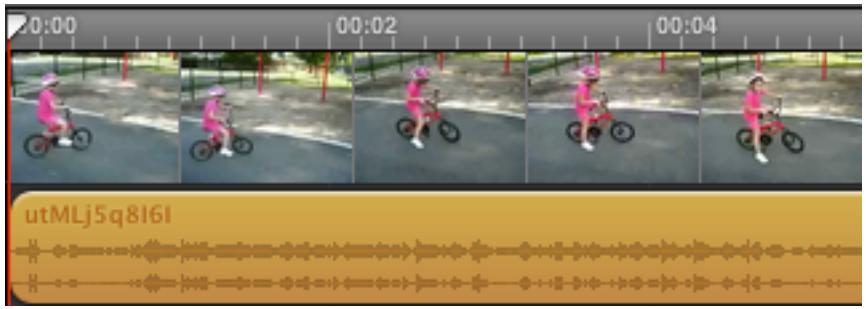

- Challenges:
  - Defining acoustic event vocabulary
  - Overlapping sounds
  - Ground-truth training data
  - Classifier accuracy

# Environmental Sound Applications

- Audio Lifelog
  Diarization




- Consumer Video
  Classification & Search

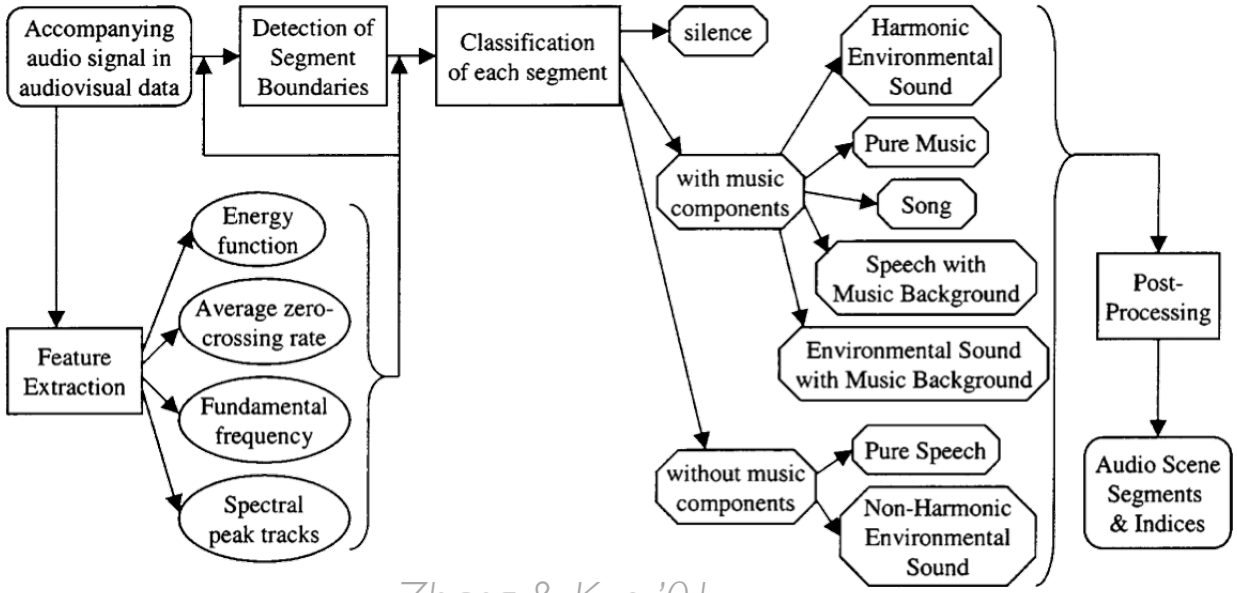



- Live hearing prosthesis app

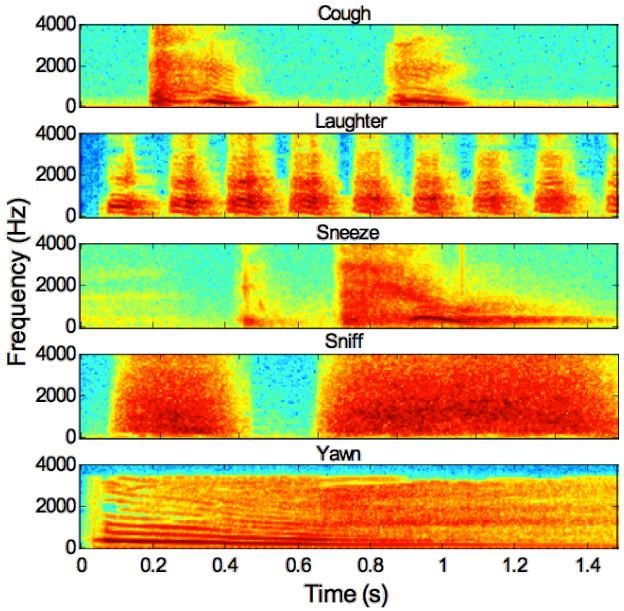
- Robot environment sensitivity

# Prior Work

- Environment Classification
  - speech/music/ silent/machine




*Zhang & Kuo '01*

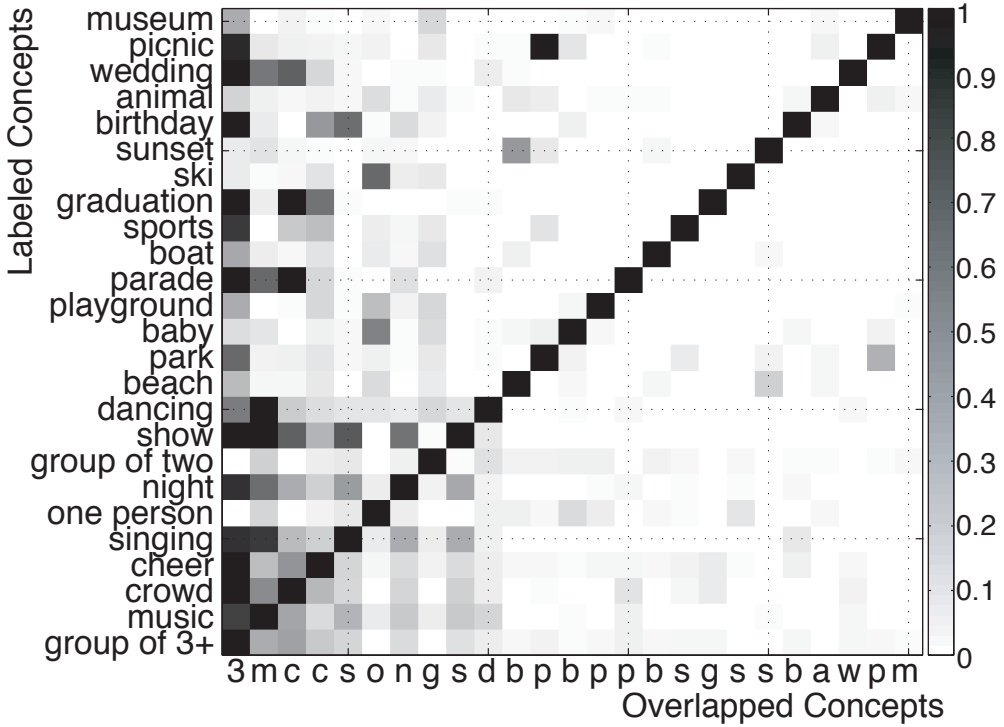



*Temko & Nadeu '06*

- 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

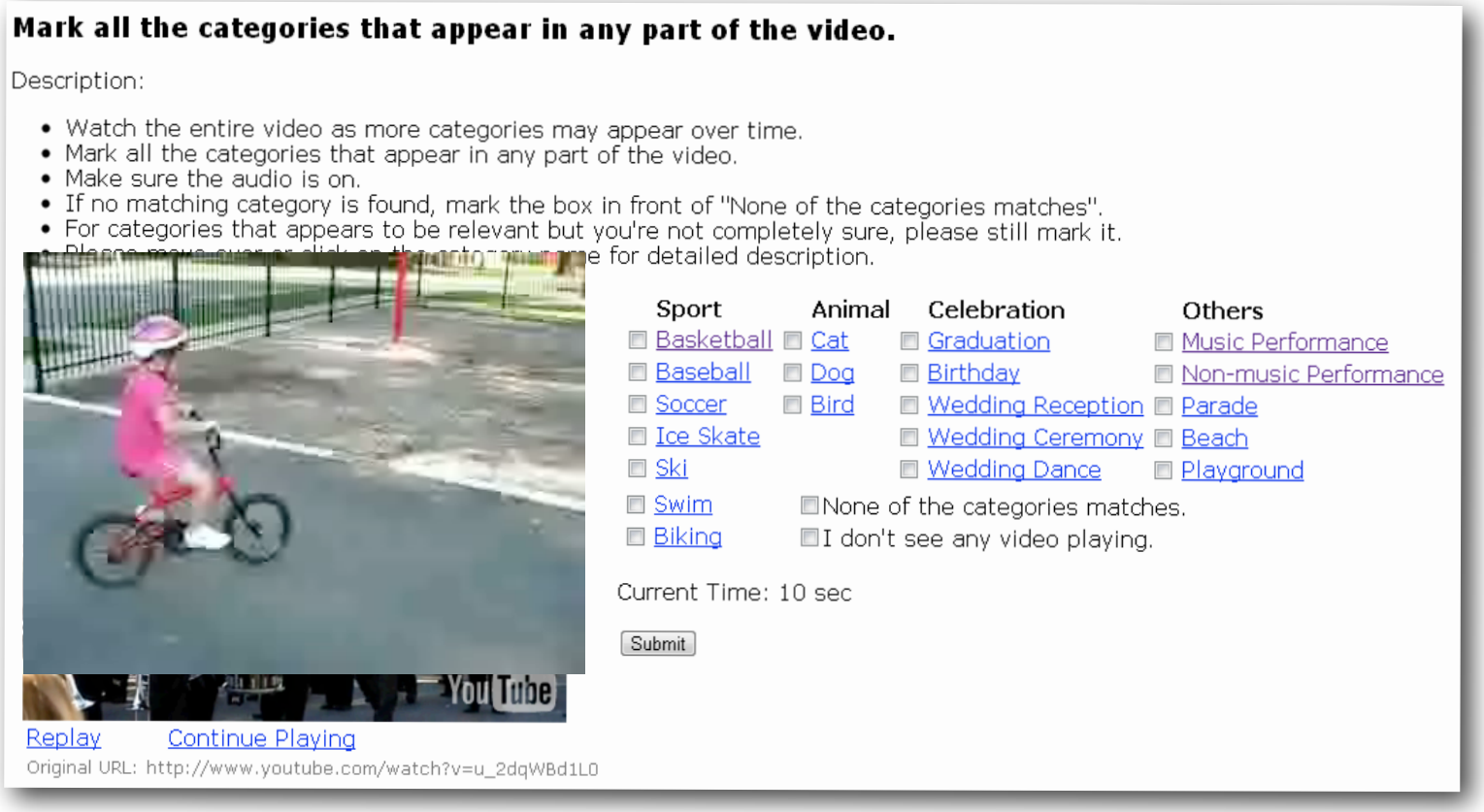
# Obtaining Labeled Data

- Amazon Mechanical Turk

  *Y-G Jiang et al. 2011*

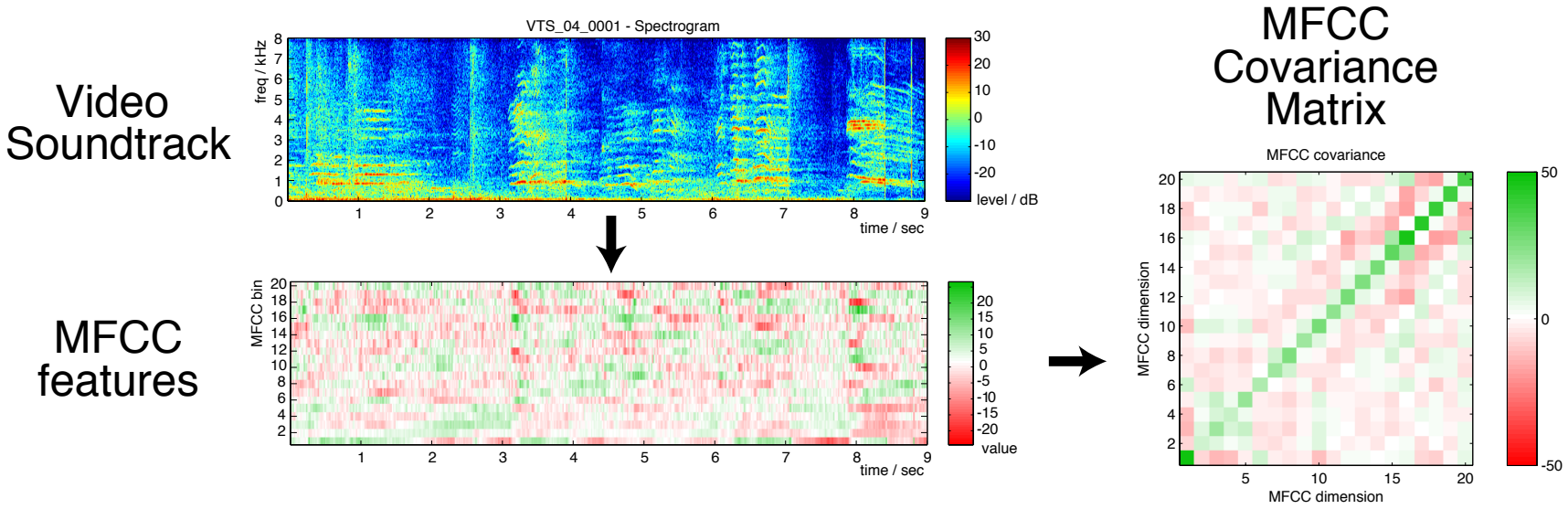
  - 10s clips
  - 9,641 videos in 4 weeks

    Columbia Consumer Video (CCV) set



# 2. Background Classification

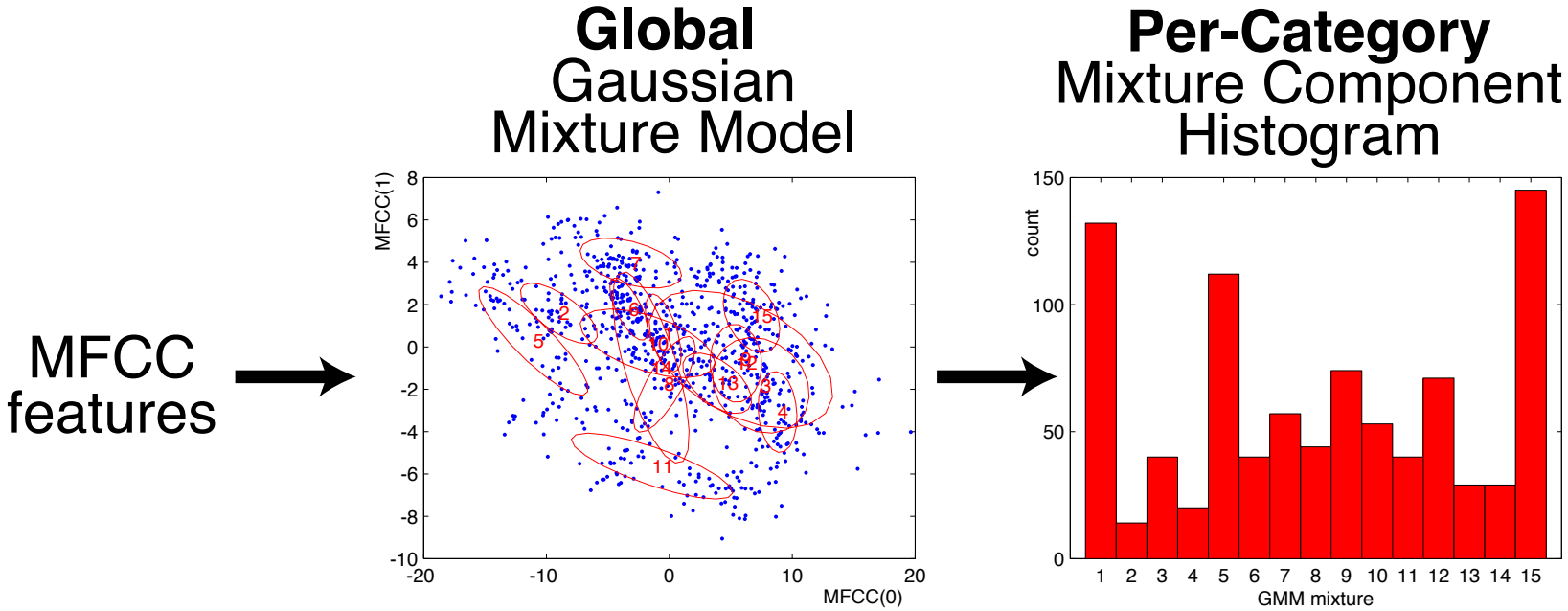
- Baseline for soundtrack classification
  - divide sound into short frames (e.g. 30 ms)
  - calculate features (e.g. MFCC) for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = "bag of features"




- Classify by e.g. Mahalanobis distance + SVM

# Codebook Histograms

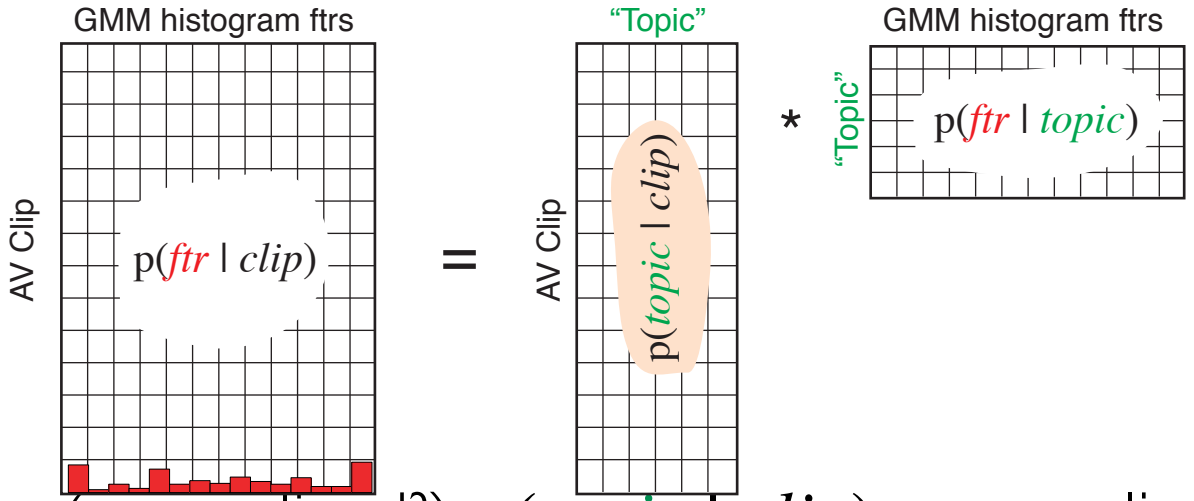
- Convert high-dim. distributions to multinomial




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

# Latent Semantic Analysis (LSA)

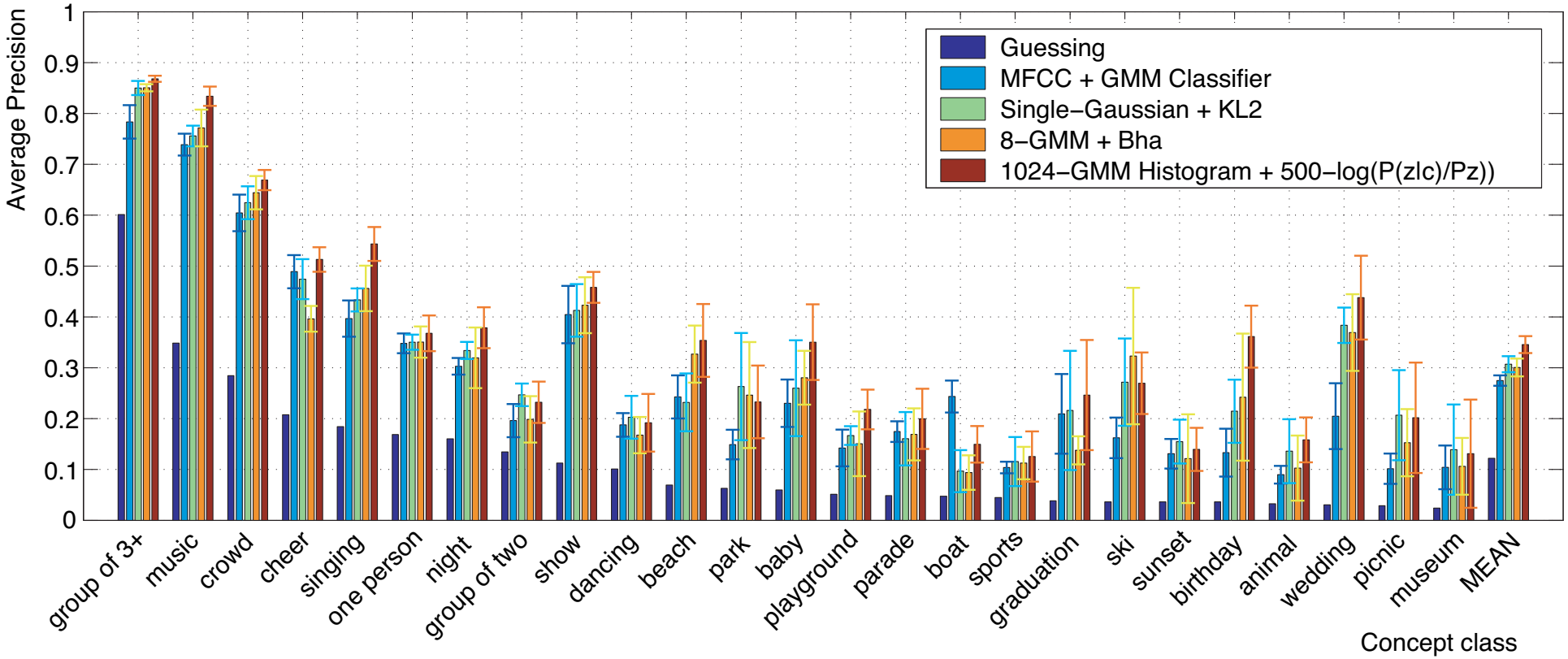
- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - $p(ftr \mid clip) = \sum_{topics} p(ftr \mid topic)\, p(topic \mid clip)$




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

# Background Classification Results

*K Lee & Ellis '10*

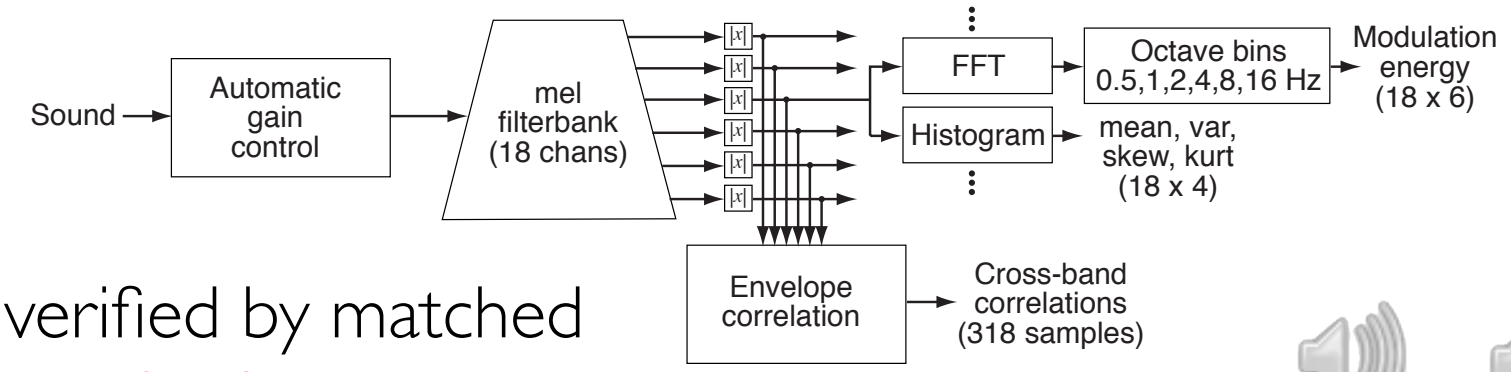



- Wide range in performance
  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

# Sound Texture Features

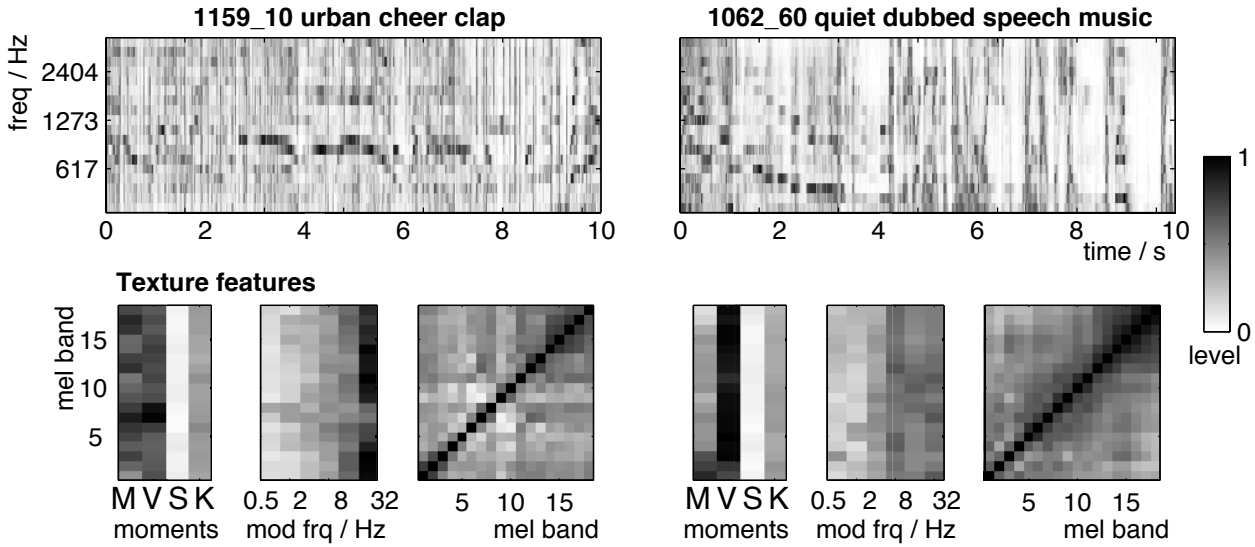
- Characterize sounds by perceptually-sufficient statistics

*McDermott Simoncelli '09*
*Ellis, Zheng, McDermott '11*



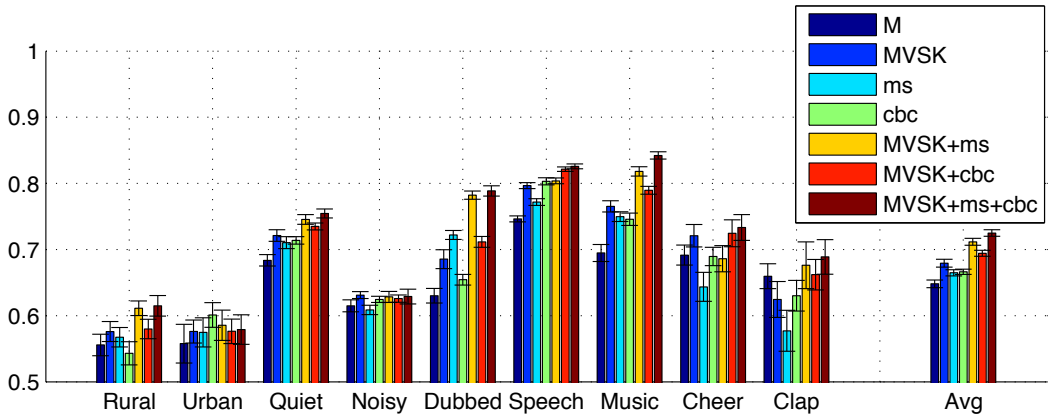

- .. verified by matched resynthesis

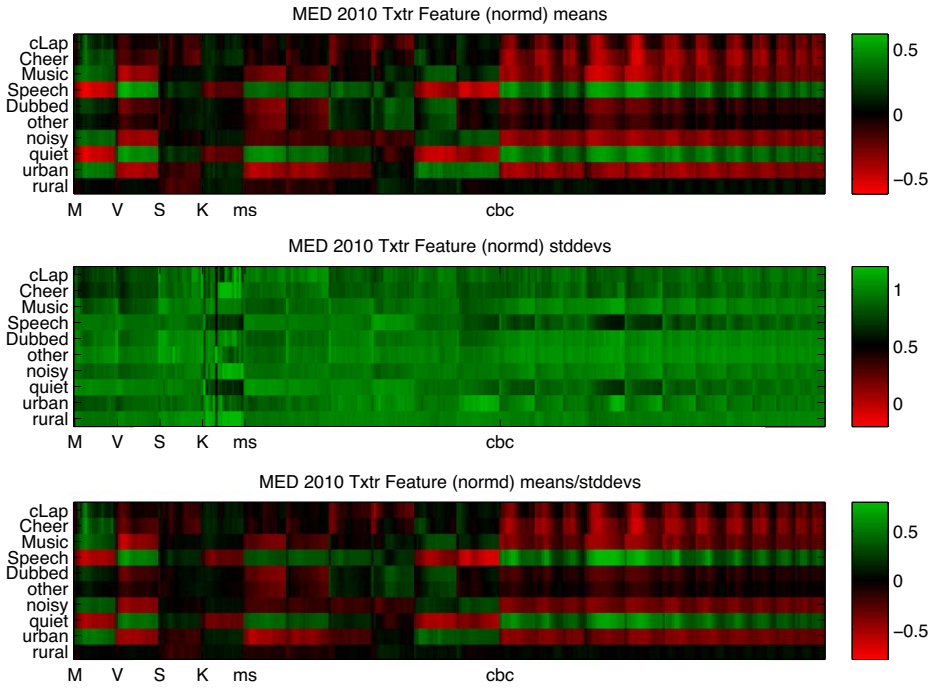
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



# Sound Texture Features

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

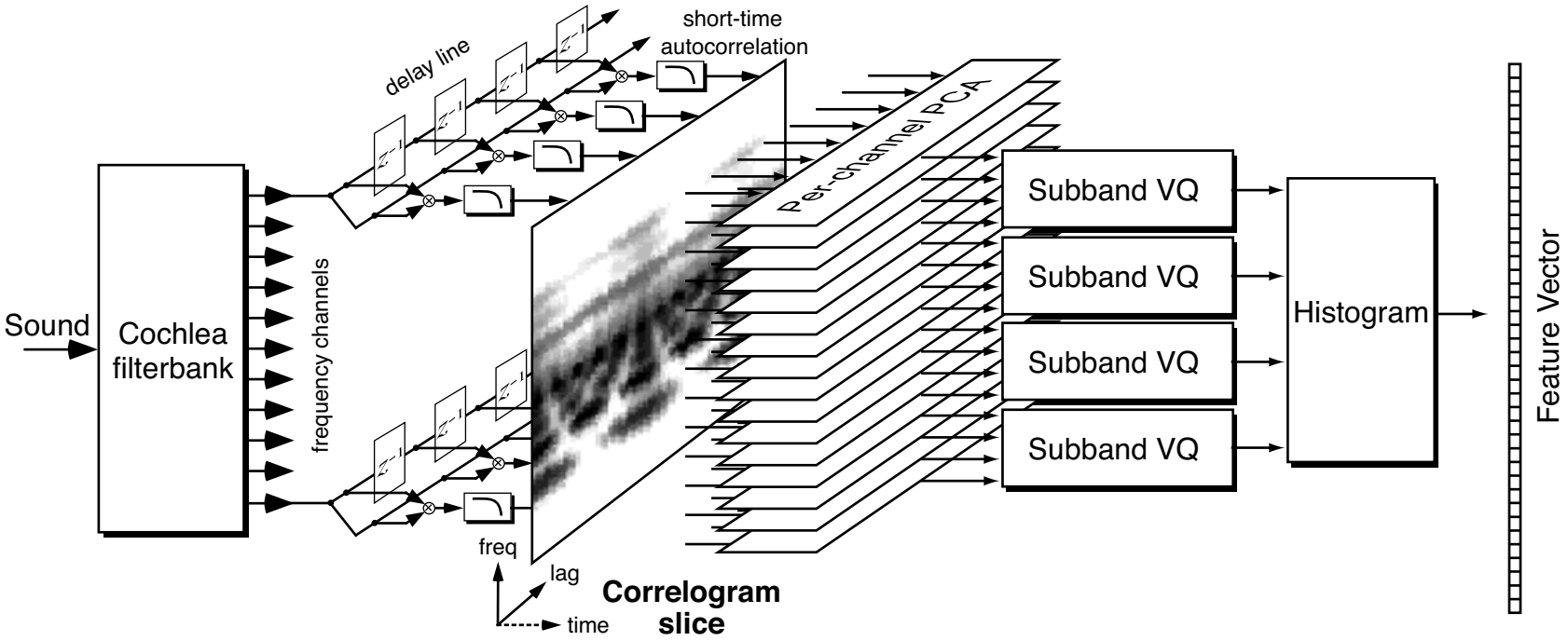



- Contrasts in feature sets
  - correlation of labels...
- Perform ~ same as MFCCs
  - combine well

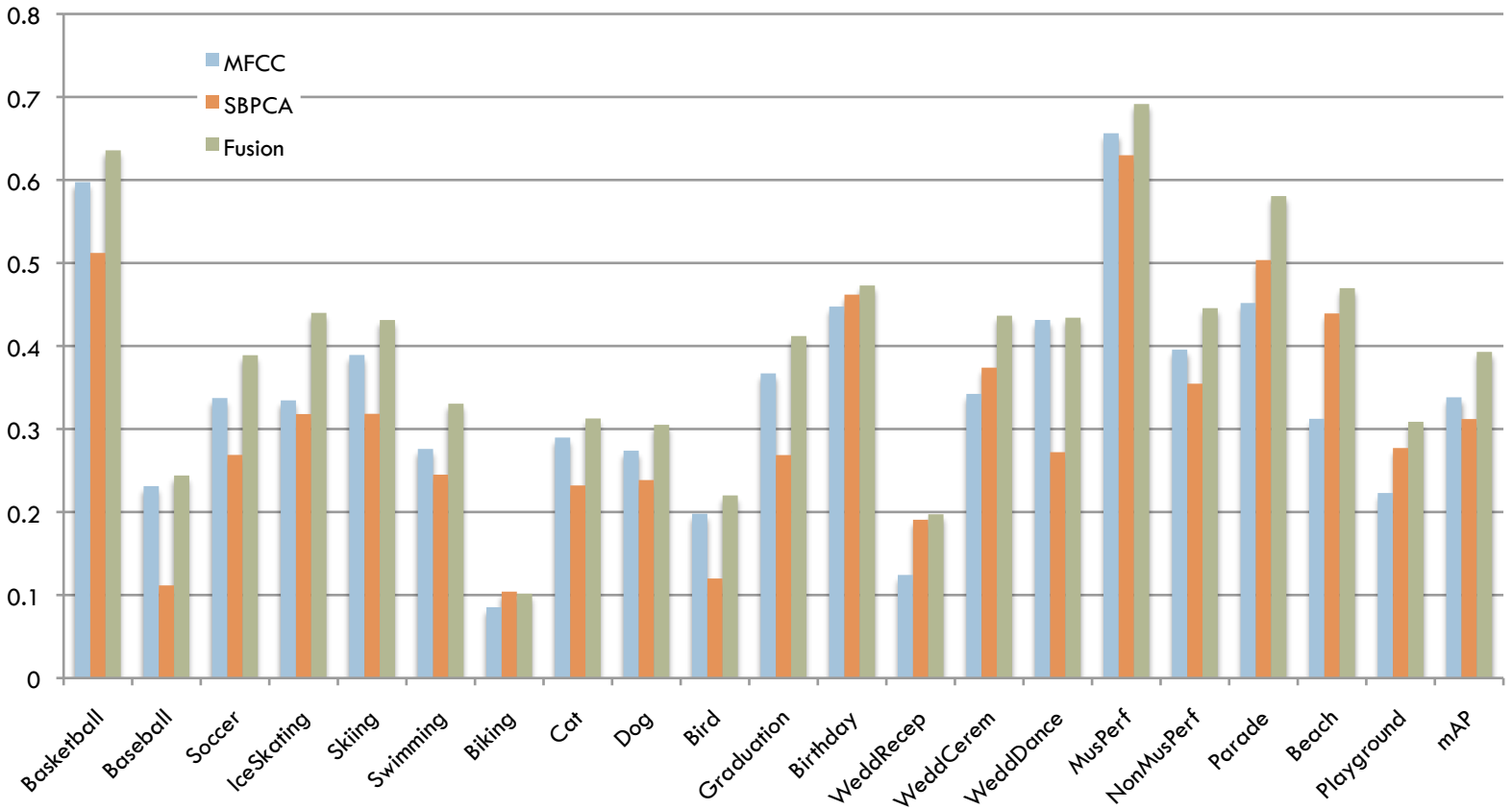
# Auditory Model Features

- **Based on Lyon/Patterson Auditory Image Model**
  - Simplified version is 10x faster (RTx5 → RT/2)
- **Captures fine time structure in multiple bands**
  - ..The information that is lost in MFCC features



# Auditory Model Feature Results

- Results vary with class, but fusion helps 15%

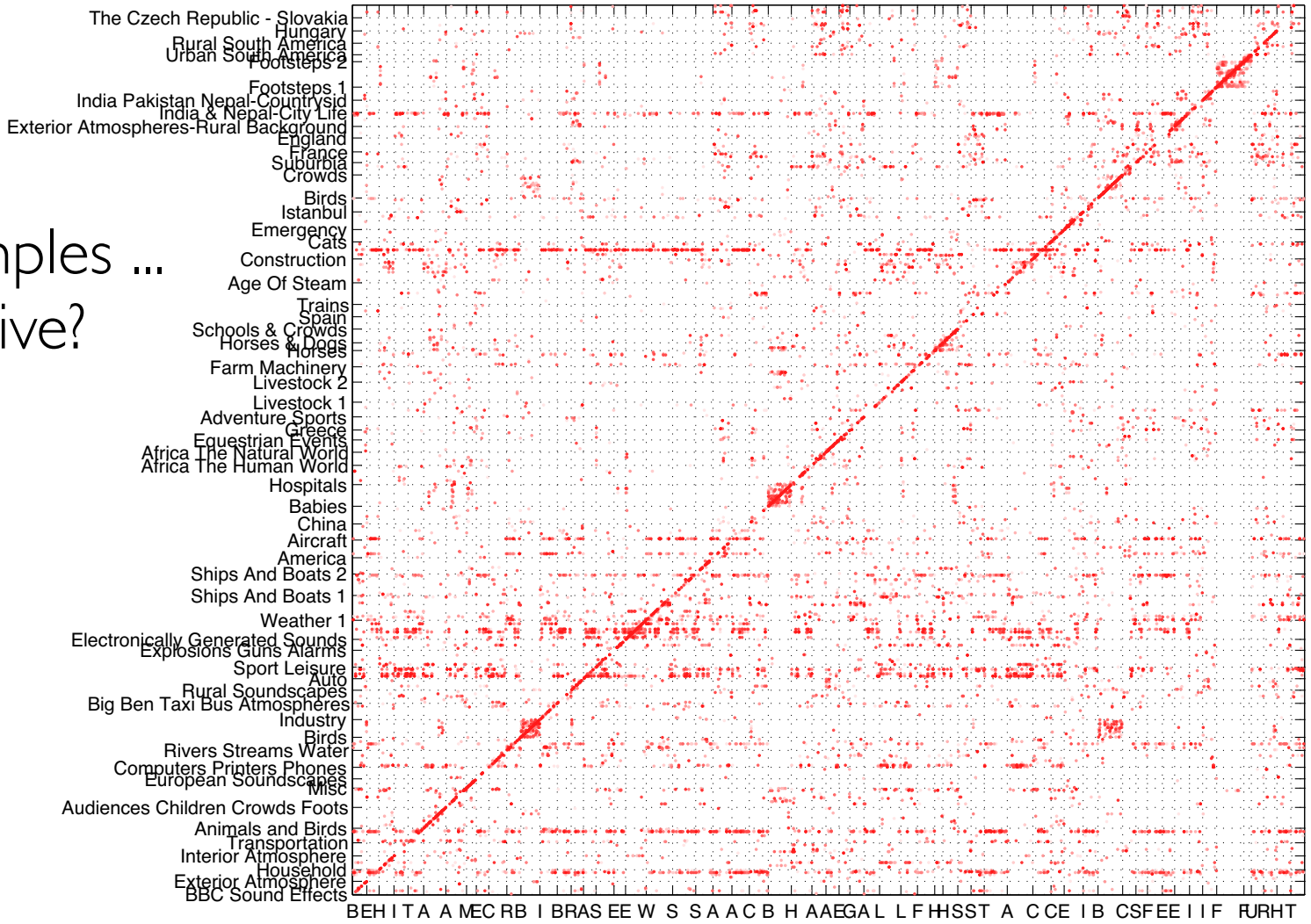



- Results for CCV (9317 videos)

# Real-World Dictionary

- BBC Sound Effects as reference library

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



# BBC Audio Semantic Classes

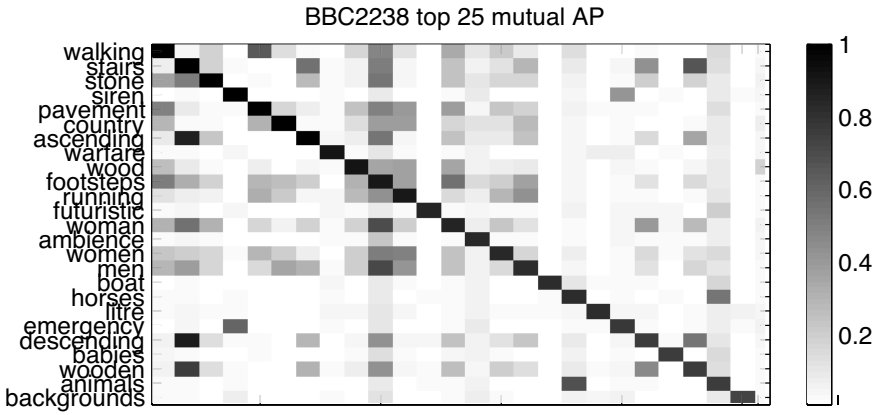
- Use BBC Sound Effects Library
  - 2238 sound files
  - Short keyword descriptions

```
SFX001-04-01    Wood Fire Inside Stove  5:07
SFX001-05-01    City Skyline City Skyline    9:46
SFX001-06-01    High Street With Traffic, Footsteps
SFX001-07-01    Car Wash Automatic, Wash Phase Inside R
SFX001-08-01    Motor Cycle Yamaha Rd 350: Motor Cycle
SFX001-09-01    Motor Cycle Yamaha Rd 350, Rider Runs U
```

- Train classifiers for 100 most common keywords
  - Select 45 best-performing

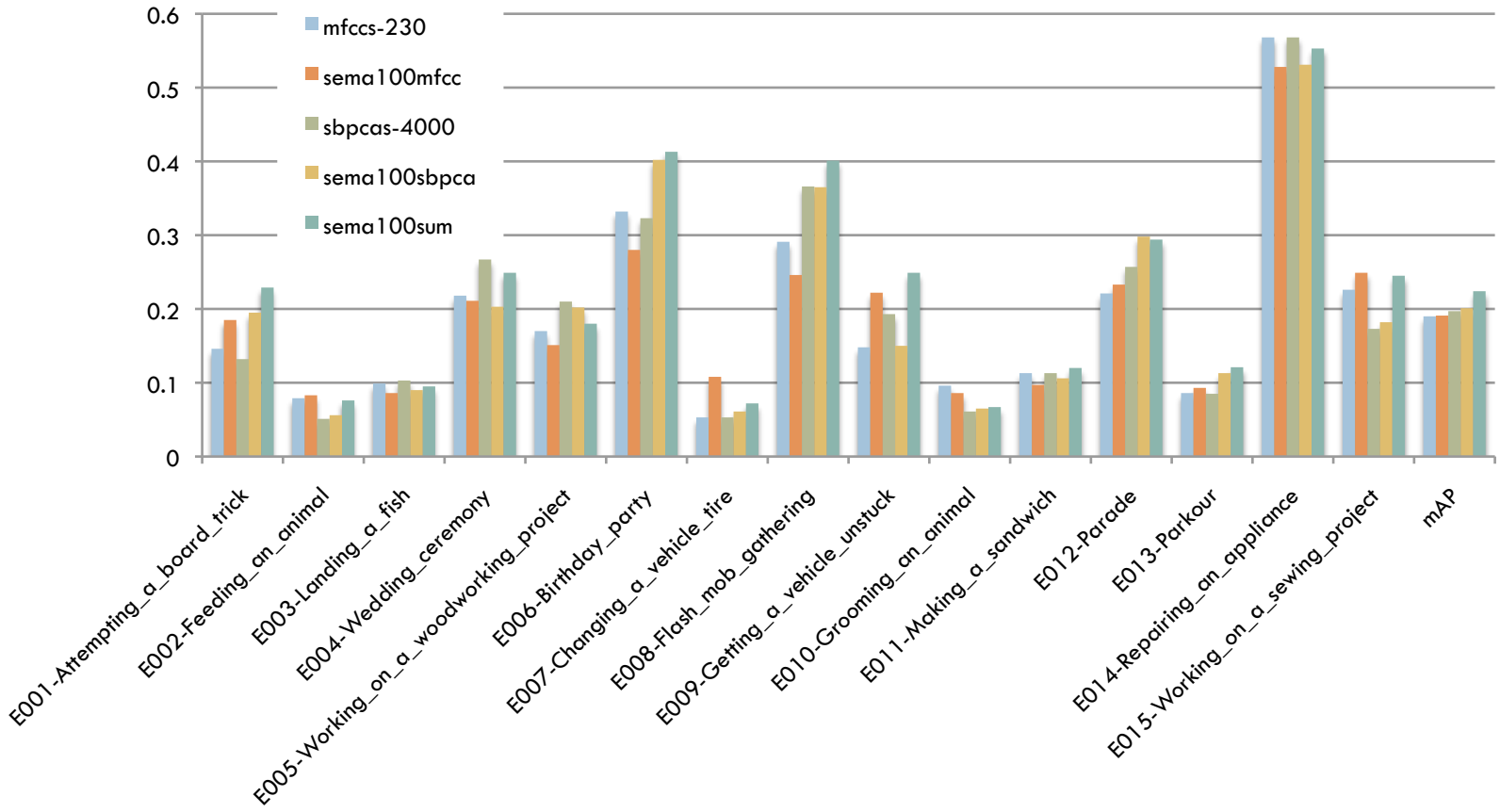
```
290 footsteps       59 men
267 on              59 general
240 animals         55 switch
197 ambience        53 starts
193 interior        53 crowds
...                 ...
```

- Added as "semantic units"
  - .. Useful for description?



# Audio Semantic Results

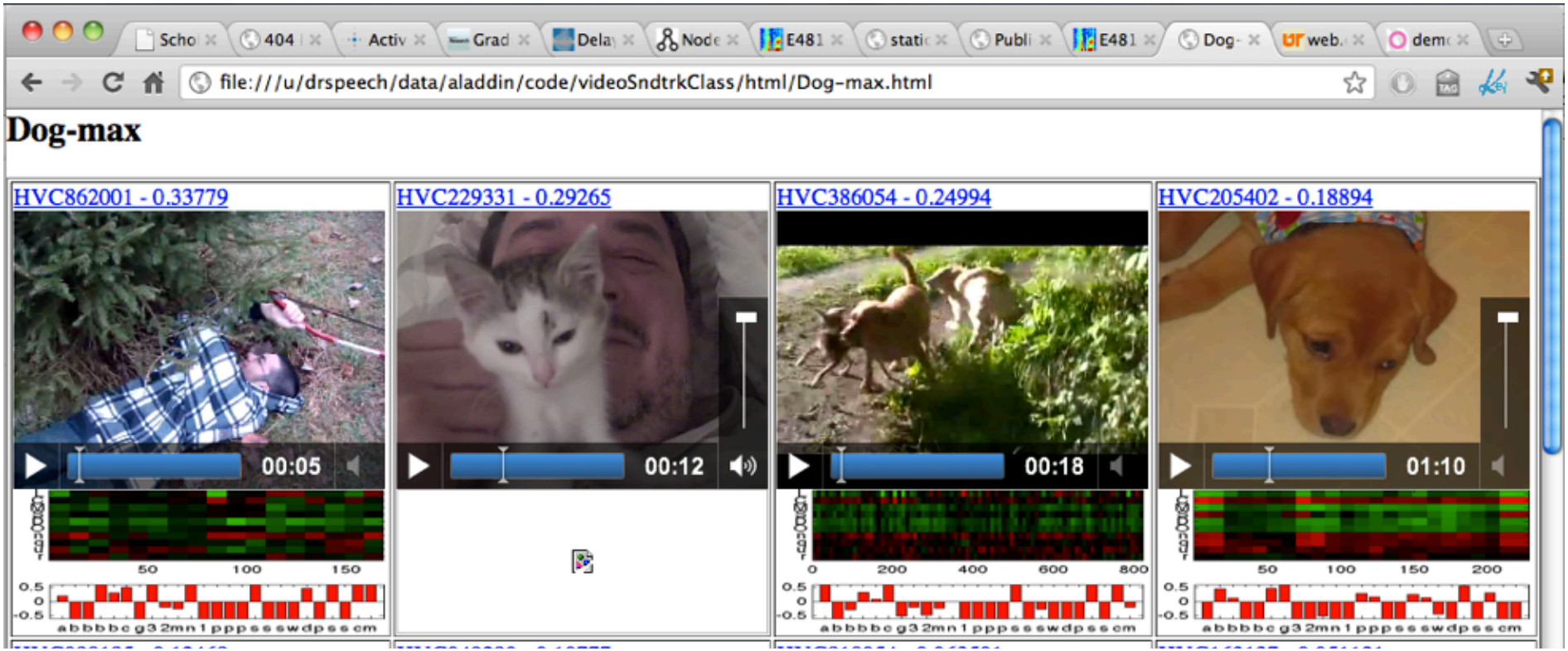
- Semantic-level feature fusion helps ~15% relative



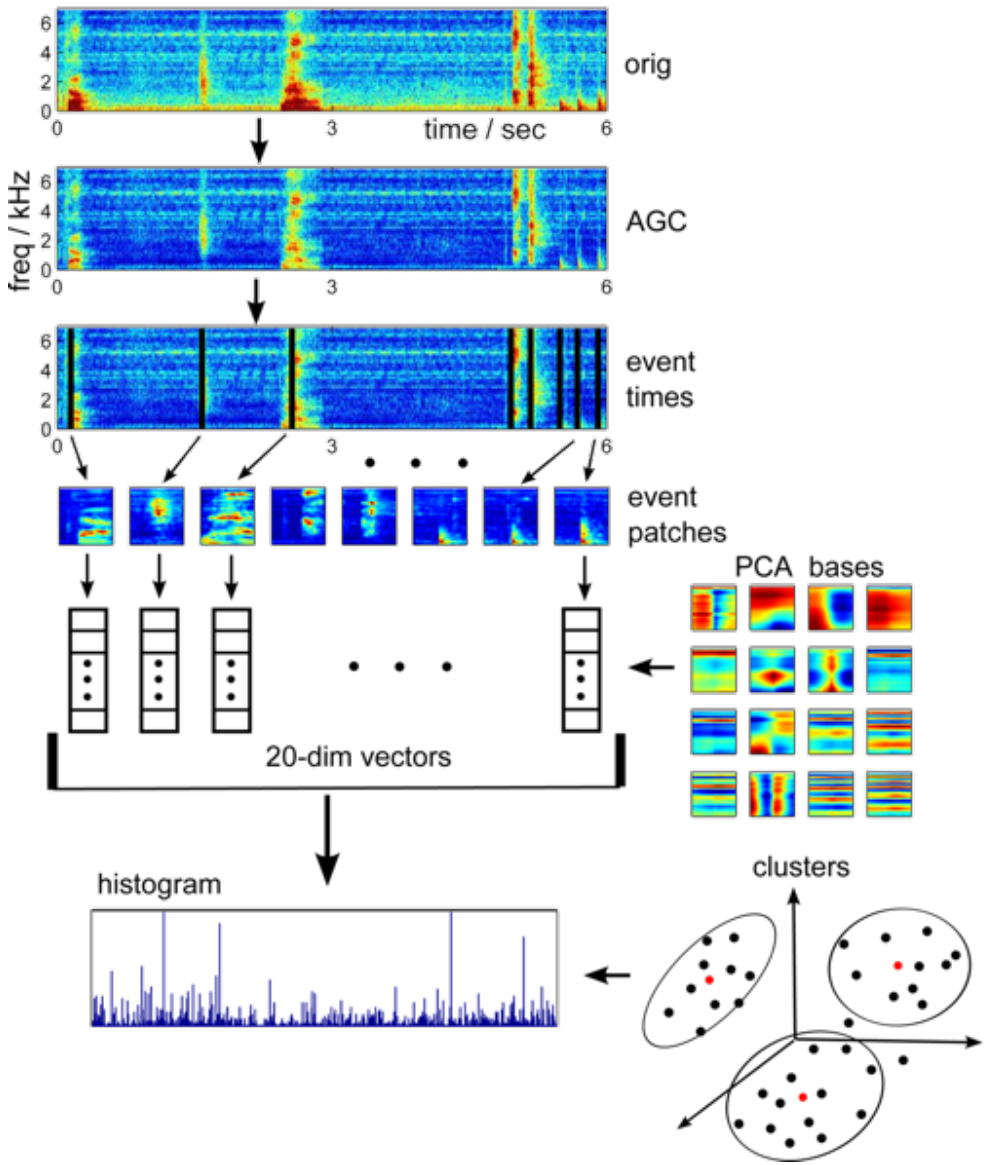

- Results on TREC MED 2011 DEVT1 (6314 videos)

# Audio Semantic Results

- Browsing results is surprisingly good (sometimes)



# 3. Foreground: 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"

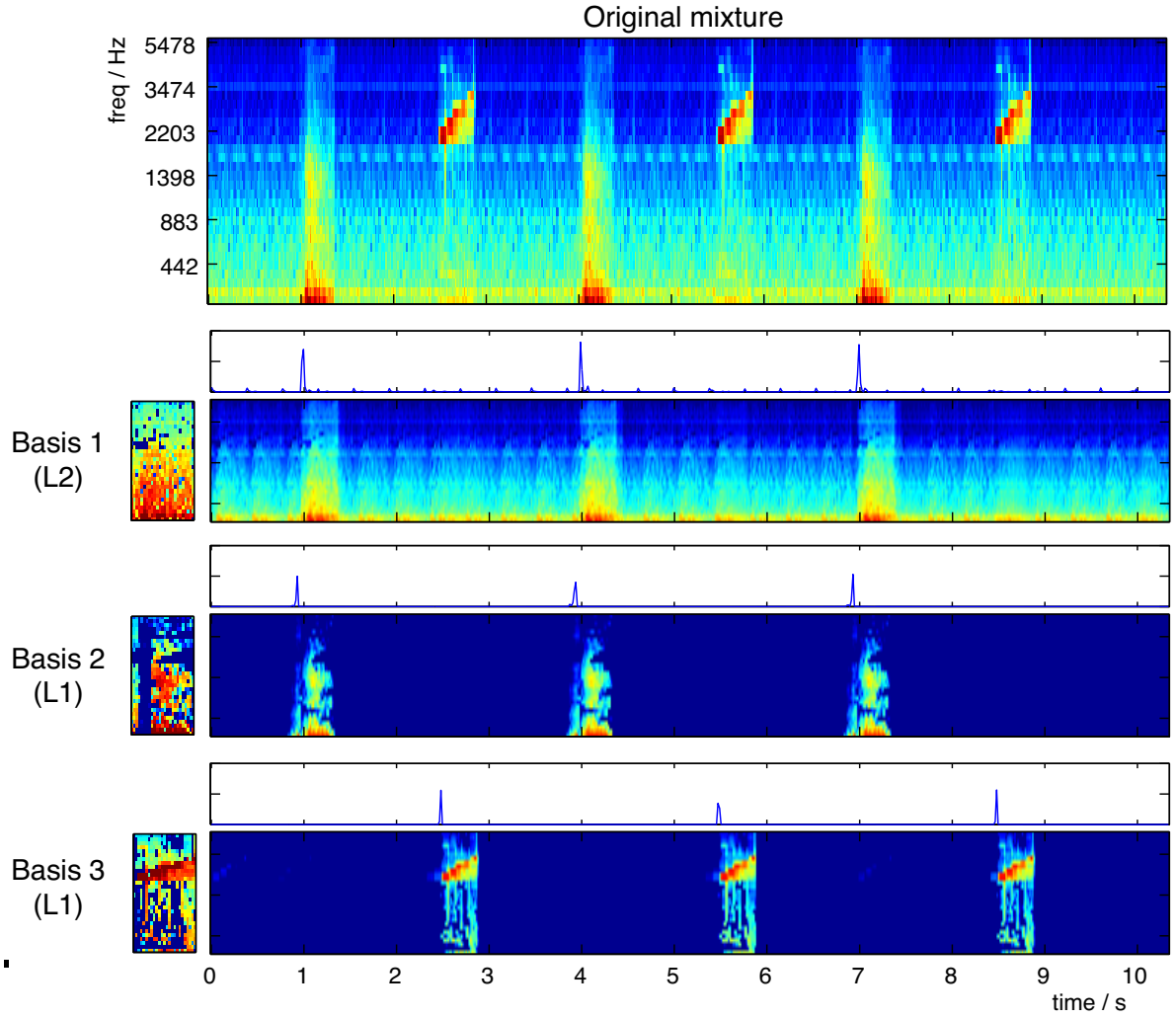
# Nonnegative Matrix Factorization

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

- Decompose spectrograms into templates + activation

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

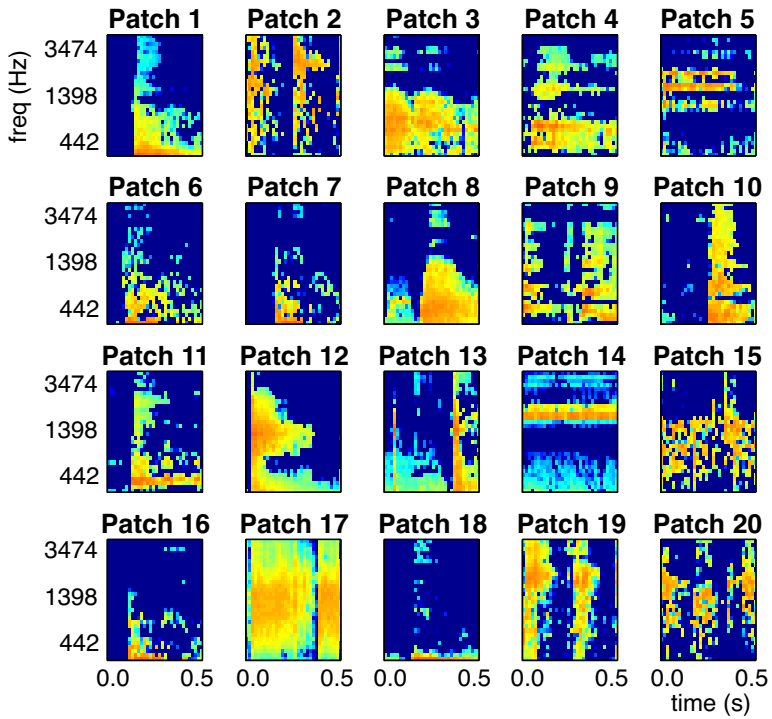
  - fast & forgiving gradient descent algorithm
  - 2D patches
  - sparsity control
  - computation time...

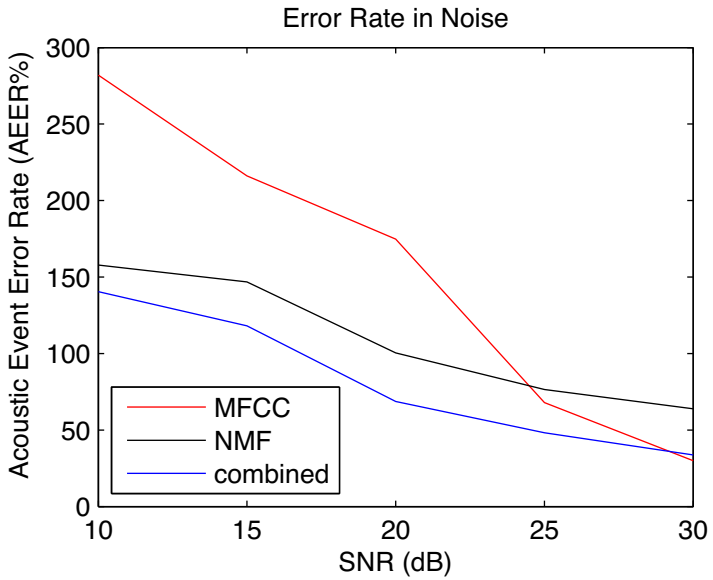


# NMF Transient Features

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

*Cotton, Ellis '11*



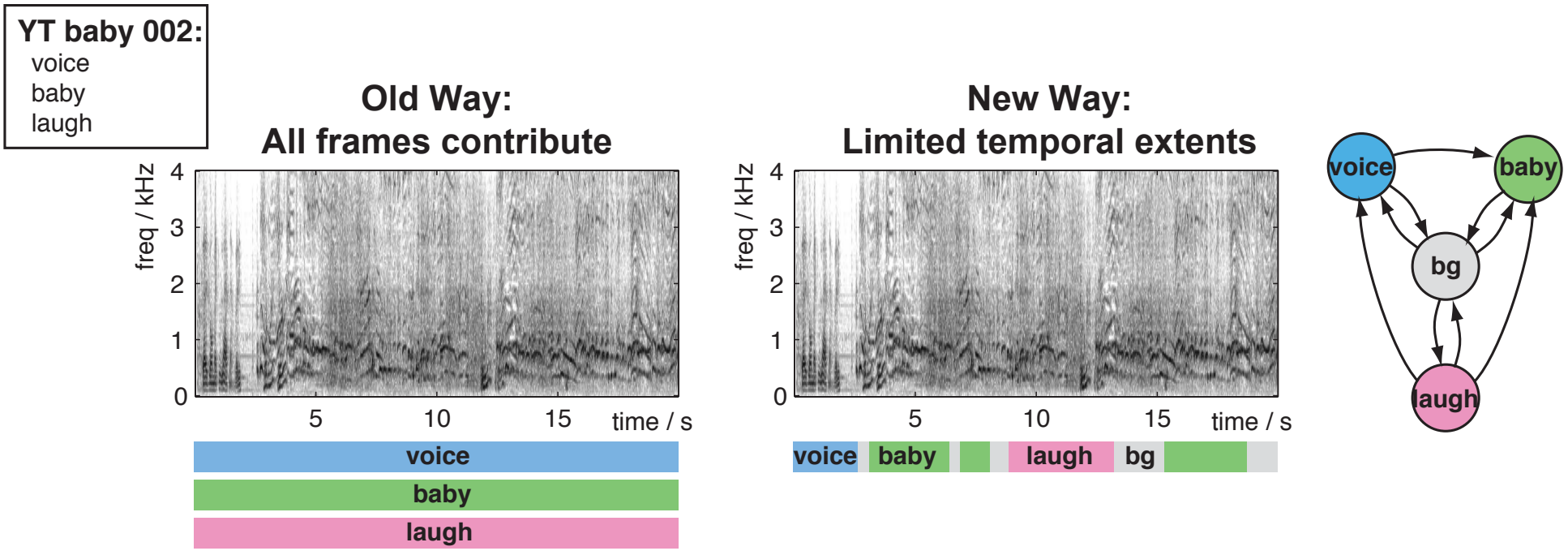

- NMF more noise-robust
  - combines well ...

# Foreground Event Localization

- Global vs. local class models

*K Lee, Ellis, Loui '10*

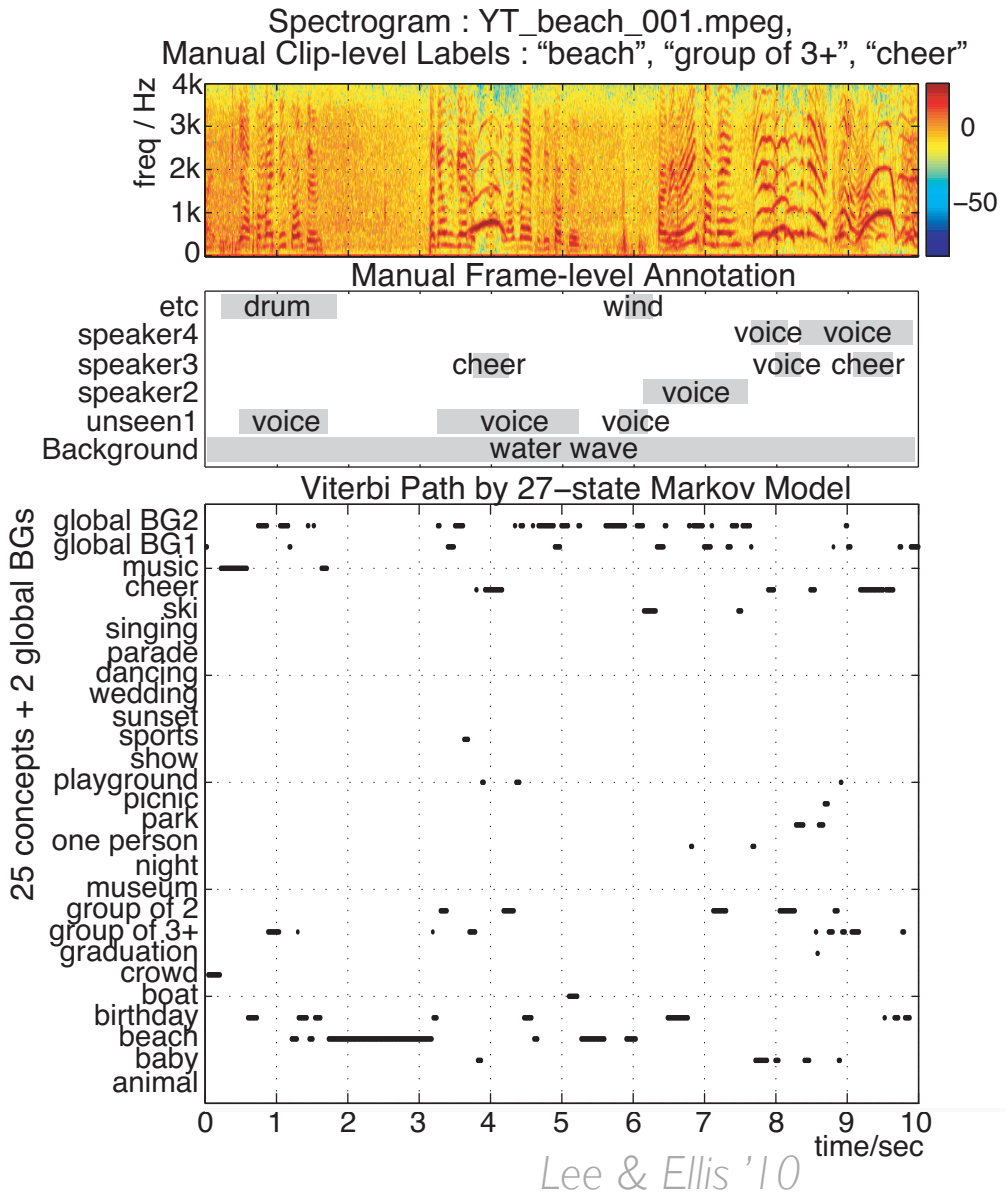
  - tell-tale acoustics may be 'washed out' in statistics
  - try iterative realignment of HMMs:




  - "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

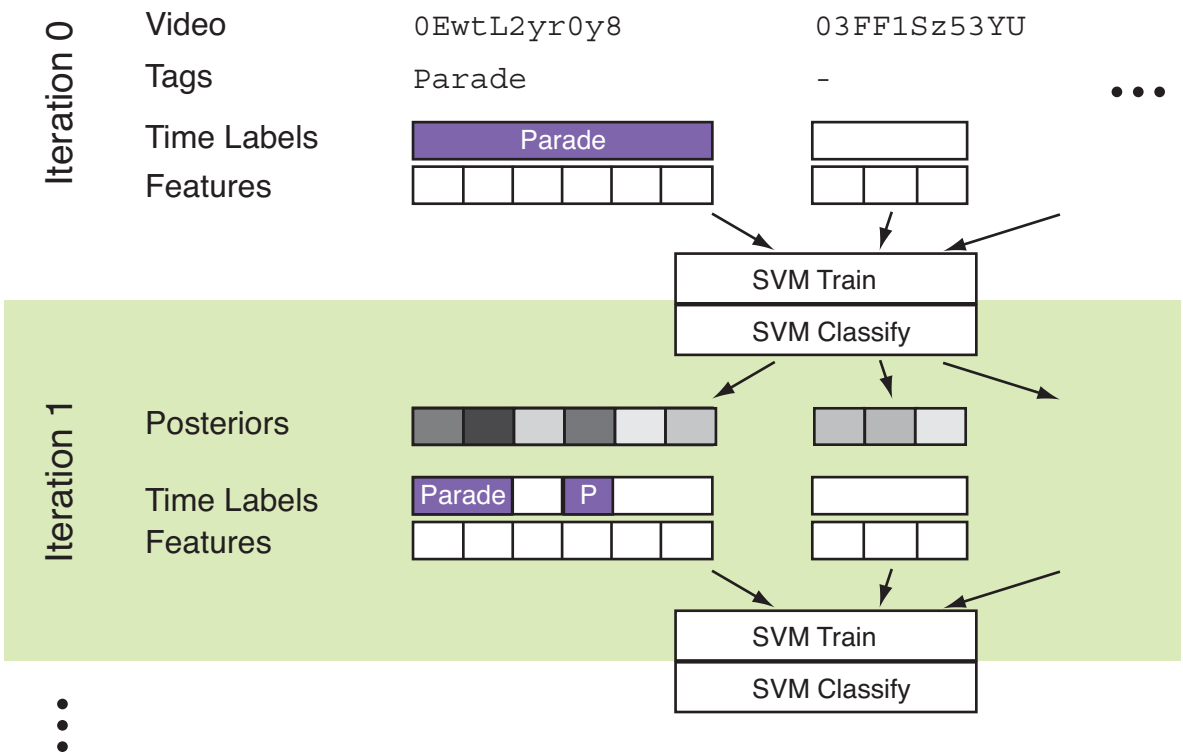



*Lee & Ellis '10*

# Refining Ground Truth

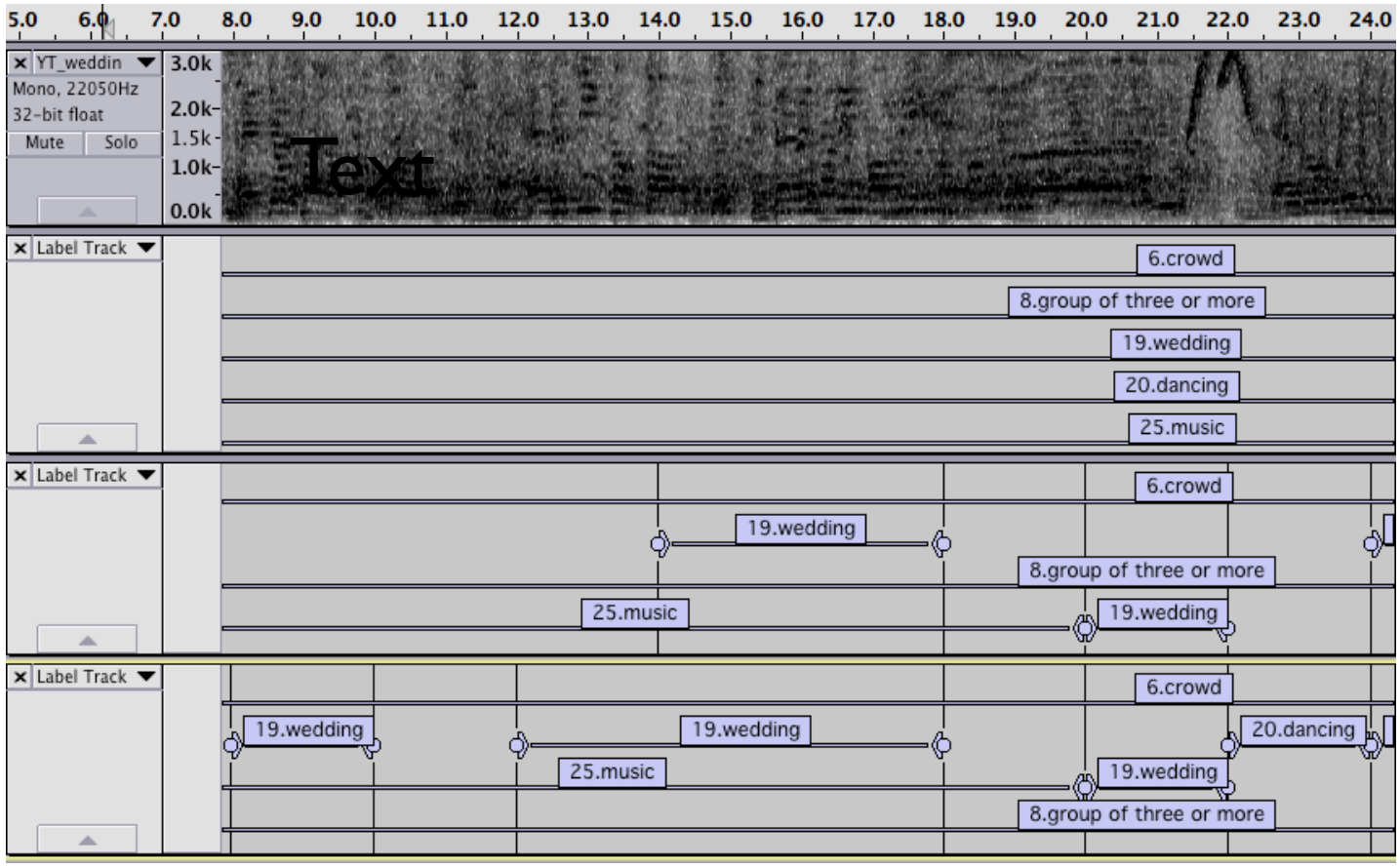
- **Audio Ground Truth at coarse time resolution**
  - better-focused labels give better classifiers?
  - but little information in very short time frames

- **Train classifiers on shorter (2 sec) segments?**
  - Initial labels apply to whole clip
  - Relabel based on most likely segments in clip
  - Retrain classifier



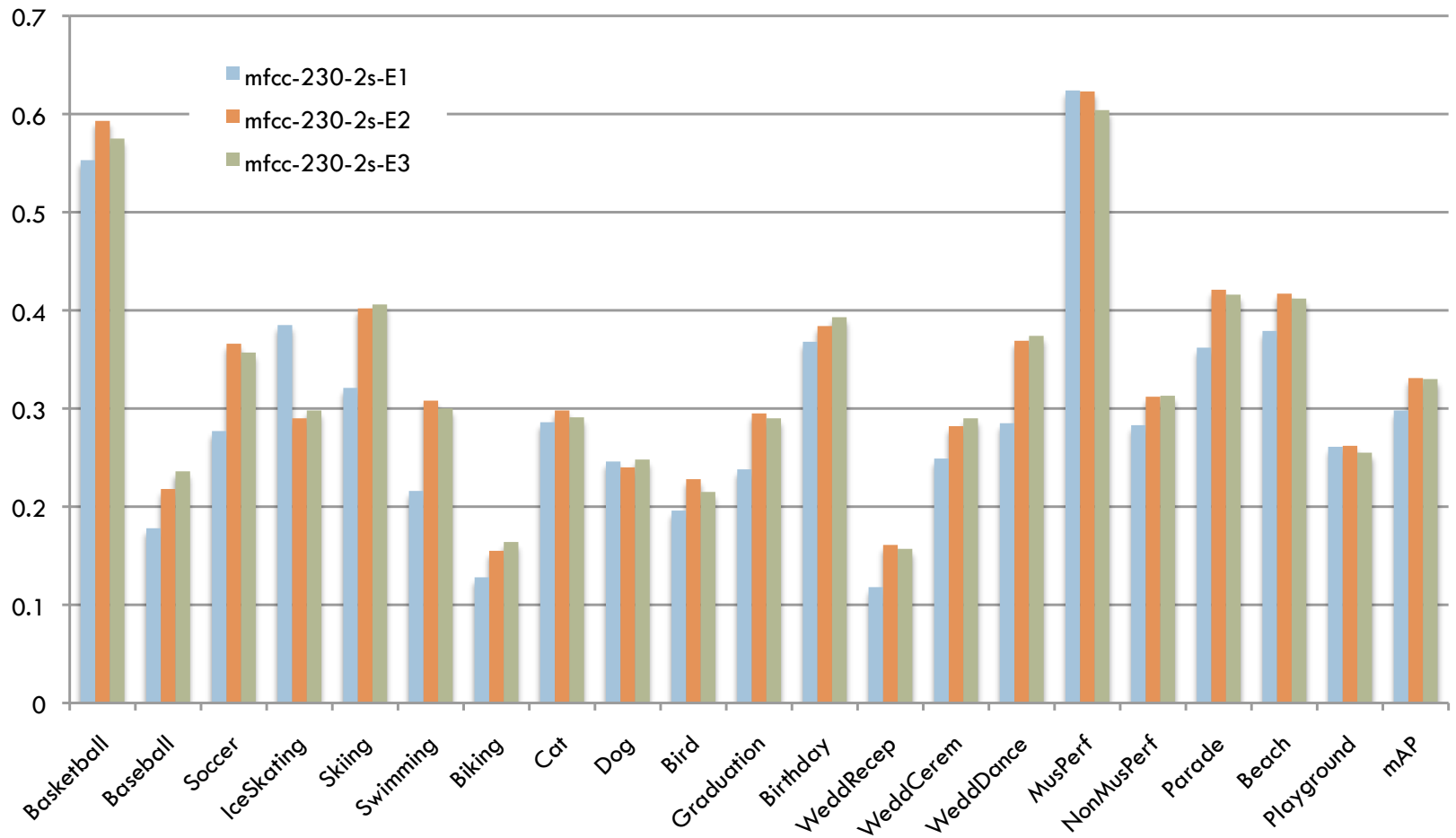
# Refining Ground Truth

- **Refining labels is "Multiple Instance Learning"**
  - "Positive" clips have at least one +ve frame
  - "Negative" clips are all –ve
- **Refine based on previous classifier's scores**
  - Threshold from CDFs of +ve and –ve frames



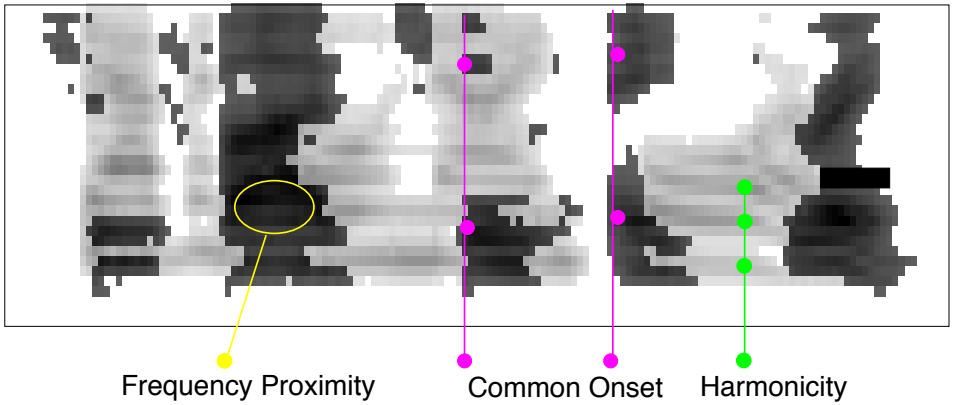
# Refining Ground Truth

- Score by aggregating frame labels to whole clip
- Systems improve for 3-5 epochs, ~10% total



# 4. Open Issues

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




*Barker et al. '05*

- Large-scale analysis
- Integration with video

# Audio Annotation

- Co-ordinating program-wide effort to share labels
  - Defining label set
  - Creating labeled data
  - CMU 43 label set as basis

| | | |
|---|---|---|
| animal | singing | clatter |
| anim_bird | music_sing | rustle |
| anim_cat | music | scratch |
| anim_ghoat | knock | hammer |
| anim_horse | thud | washboard |
| human_noise | clap | applause |
| laugh | click | whistle |
| scream | bang | squeak |
| child | beep | tone |
| mumble | engine_quiet | sirene |
| speech | engine_light | water |
| speech_ne | power_tool | micro_blow |
| radio | engine_heavy | wind |
| white_noise | cheer | |
| other_creak | crowd | |

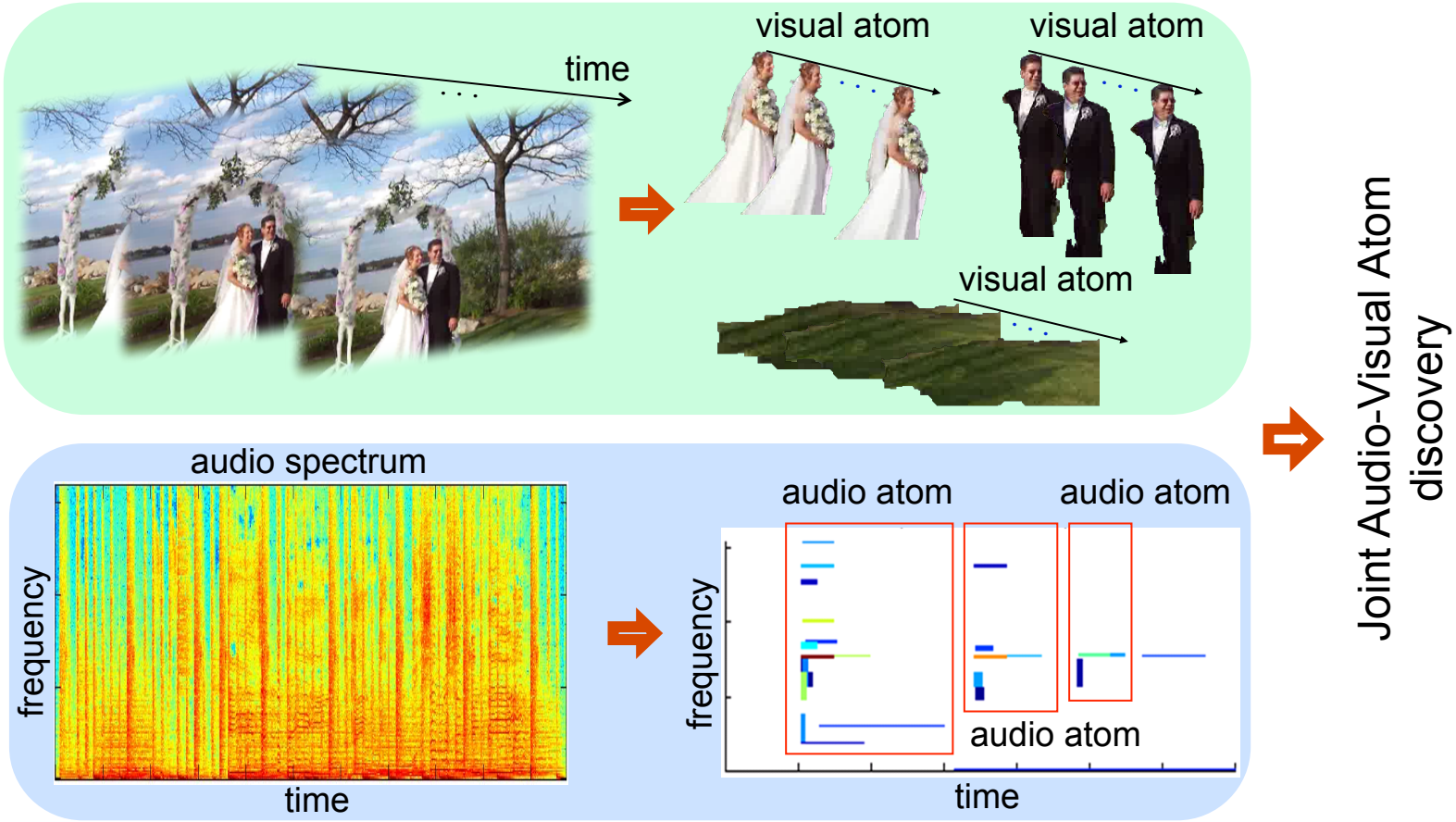
*Burger & Metze, CMU, 2012*

- Efficiency of labeling is key
  - Audio-only vs. audio+video?
  - Coarse-level human labels
    + automatic refinement?

# Audio-Visual Atoms

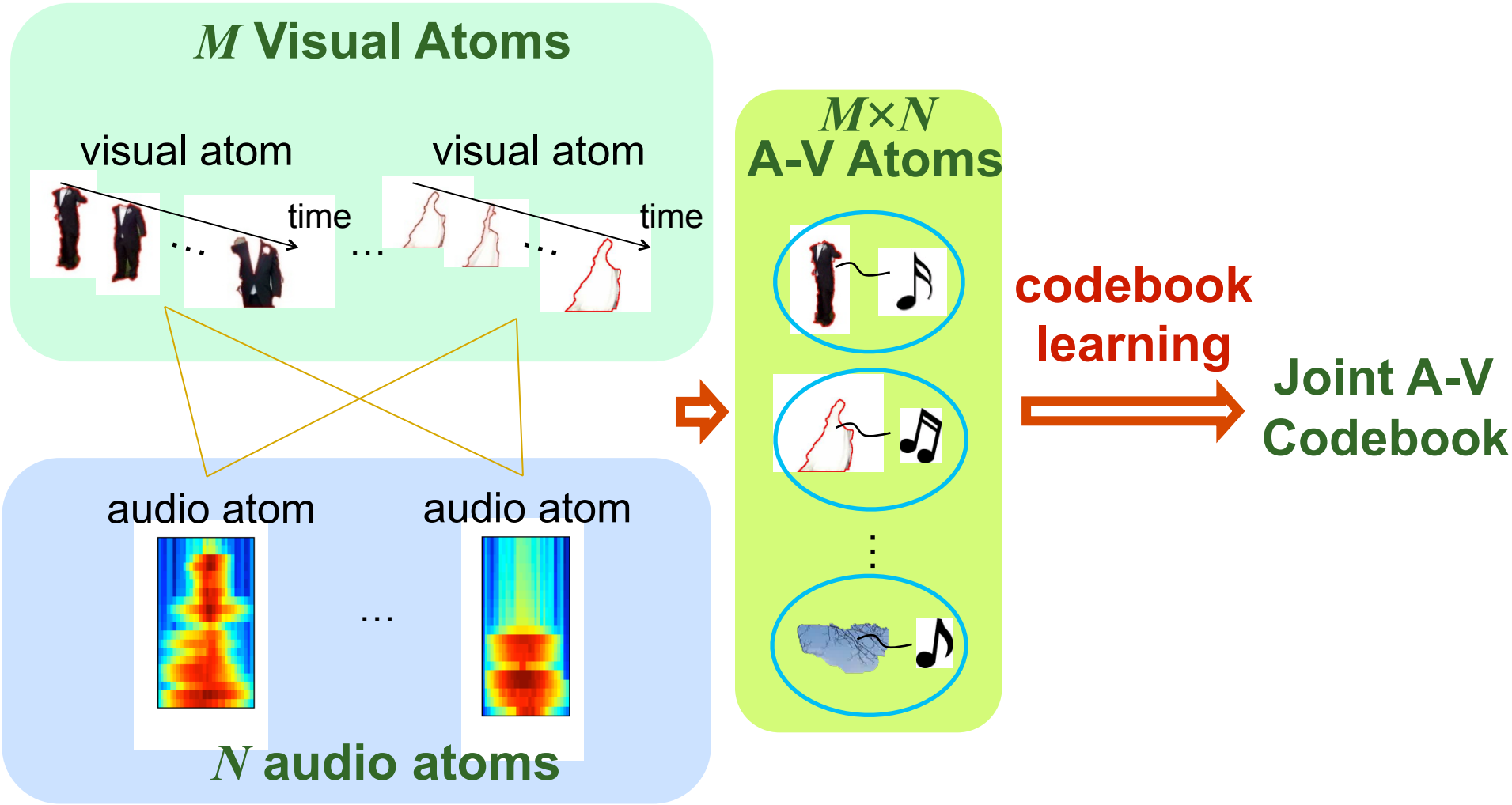
*Jiang et al. '09*

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



# Audio-Visual Atoms

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



# Audio-Visual Atoms

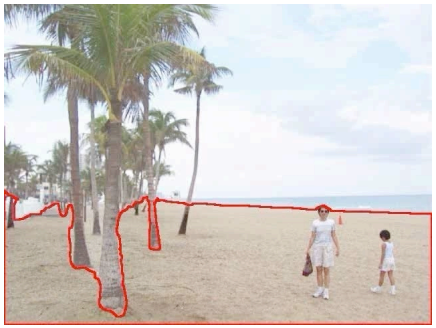
black suit
+ romantic
music

*Wedding*

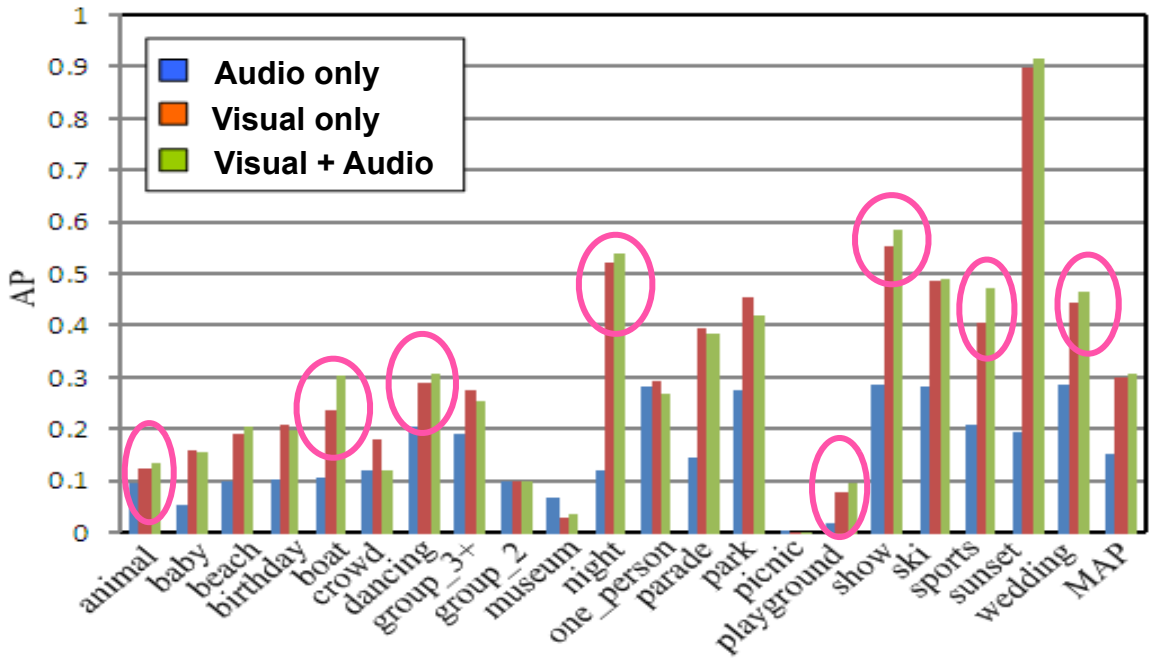
marching
people
+ parade
sound

*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, ...

# References


- Jon Barker, Martin Cooke, & Dan Ellis, "Decoding Speech in the Presence of Other Sources," *Speech Communication* 45(1): 5-25, 2005.
- Courtenay Cotton, Dan Ellis, & Alex Loui, "Soundtrack classification by transient events," *IEEE ICASSP*, Prague, May 2011.
- Courtenay Cotton & Dan Ellis, "Spectral vs. Spectro-Temporal Features for Acoustic Event Classification," submitted to *IEEE WASPAA*, 2011.
- Dan Ellis, Xiaohong Zheng, Josh McDermott, "Classifying soundtracks with audio texture features," *IEEE ICASSP*, Prague, May 2011.
- Wei Jiang, Courtenay Cotton, Shih-Fu Chang, Dan Ellis, & Alex Loui, "Short-Term Audio-Visual Atoms for Generic Video Concept Classification," *ACM MultiMedia*, 5-14, Beijing, Oct 2009.
- Keansub Lee & Dan Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Tr. Audio, Speech and Lang. Proc.* 18(6): 1406-1416, Aug. 2010.
- Keansub Lee, Dan Ellis, Alex Loui, "Detecting local semantic concepts in environmental sounds using Markov model based clustering," *IEEE ICASSP*, 2278-2281, Dallas, Apr 2010.
- Byung-Suk Lee & Dan Ellis, "Noise-robust pitch tracking by trained channel selection," submitted to *IEEE WASPAA*, 2011.
- Andriy Temko & Climent Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition* 39(4): 682-694, 2006
- Ron Weiss & Dan Ellis, "Speech separation using speaker-adapted Eigenvoice speech models," *Computer Speech & Lang*. 24(1): 16-29, 2010.
- Ron Weiss, Michael Mandel, & Dan Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Communication* 53(5): 606-621, May 2011.
- Tong Zhang & C.-C. Jay Kuo, "Audio content analysis for on-line audiovisual data segmentation," *IEEE TSAP* 9(4): 441-457, May 2001

# Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.