Analysis of Everyday Sounds

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

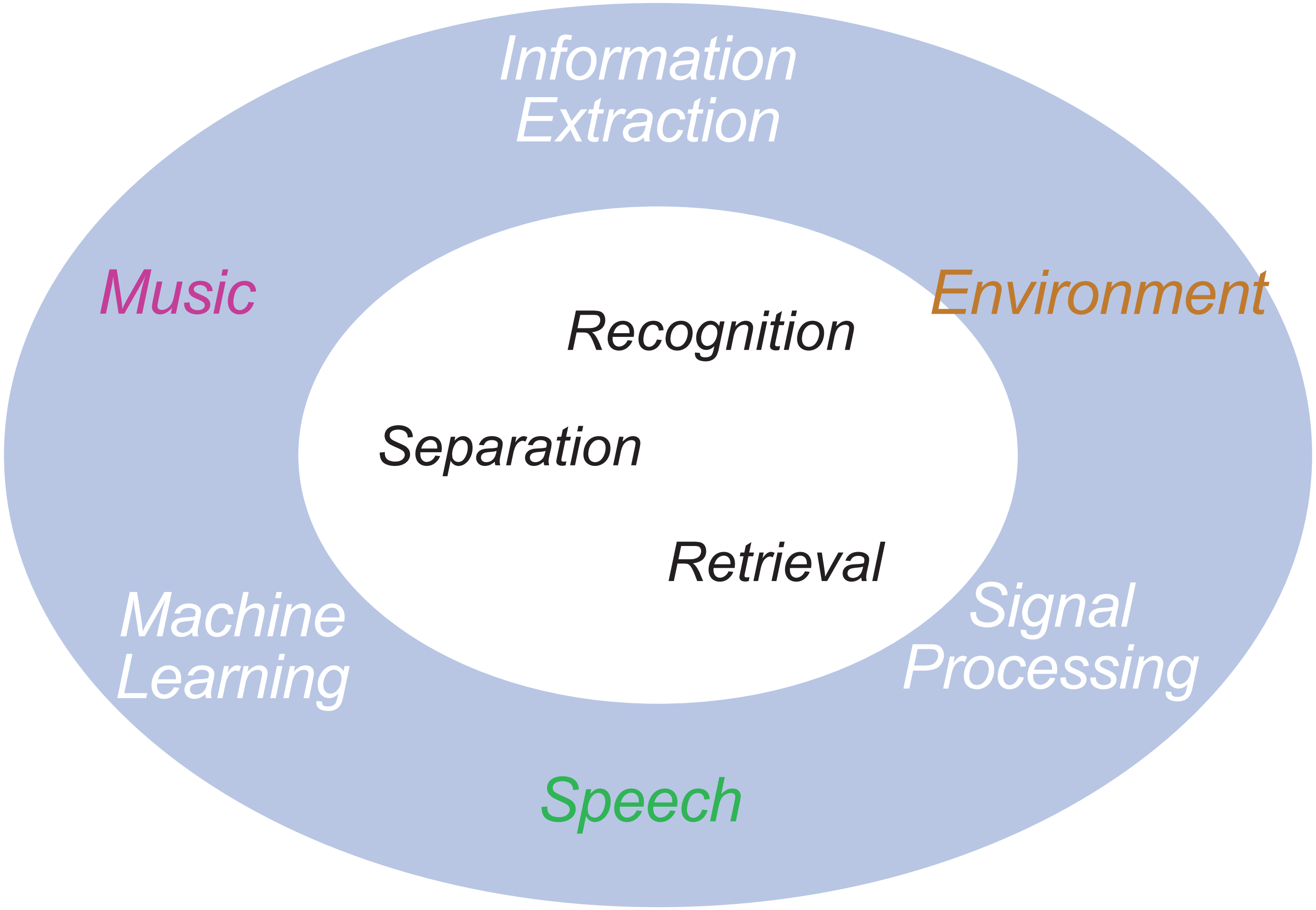
dpwe@ee.columbia.edu

1. Personal and Consumer Audio
2. Segmenting & Clustering
3. Special-Purpose Detectors
4. Generic Concept Detectors
5. Challenges & Future

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

Columbia University
in the City of New York

LabROSA Overview



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

Personal Audio Applications

- Automatic appointment-book history
  - fills in when & where of movements
- "Life statistics"
  - how long did I spend in meetings this week?
  - most frequent conversations
  - favorite phrases?
- Retrieving details
  - what exactly did I promise?
  - privacy issues...
- Nostalgia
- ... or what?

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

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

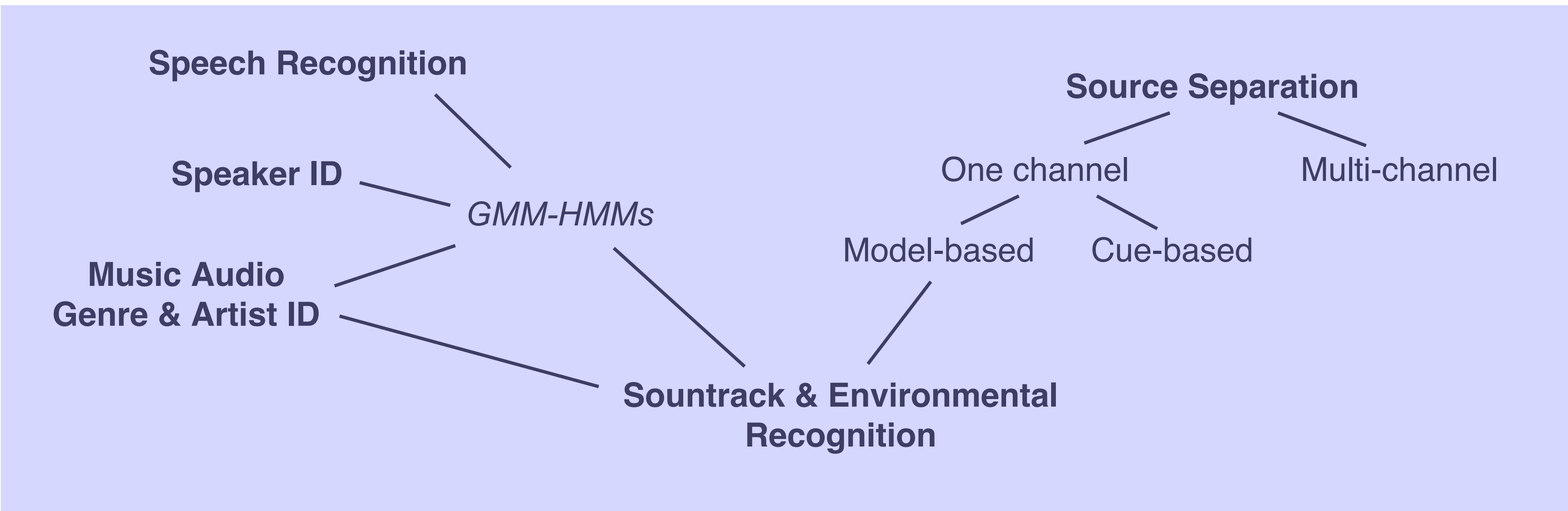
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Information in Audio

- **Environmental recordings contain info on:**
  - location – type (restaurant, street, ...) and specific
  - activity – talking, walking, typing
  - people – generic (2 males), specific (Chuck & John)
  - spoken content ... maybe
- **but not:**
  - what people and things "looked like"
  - day/night ...
  - ... except when correlated with audible features

A Brief History of Audio Processing

- Environmental sound classification draws on earlier sound classification work
  - as well as source separation...

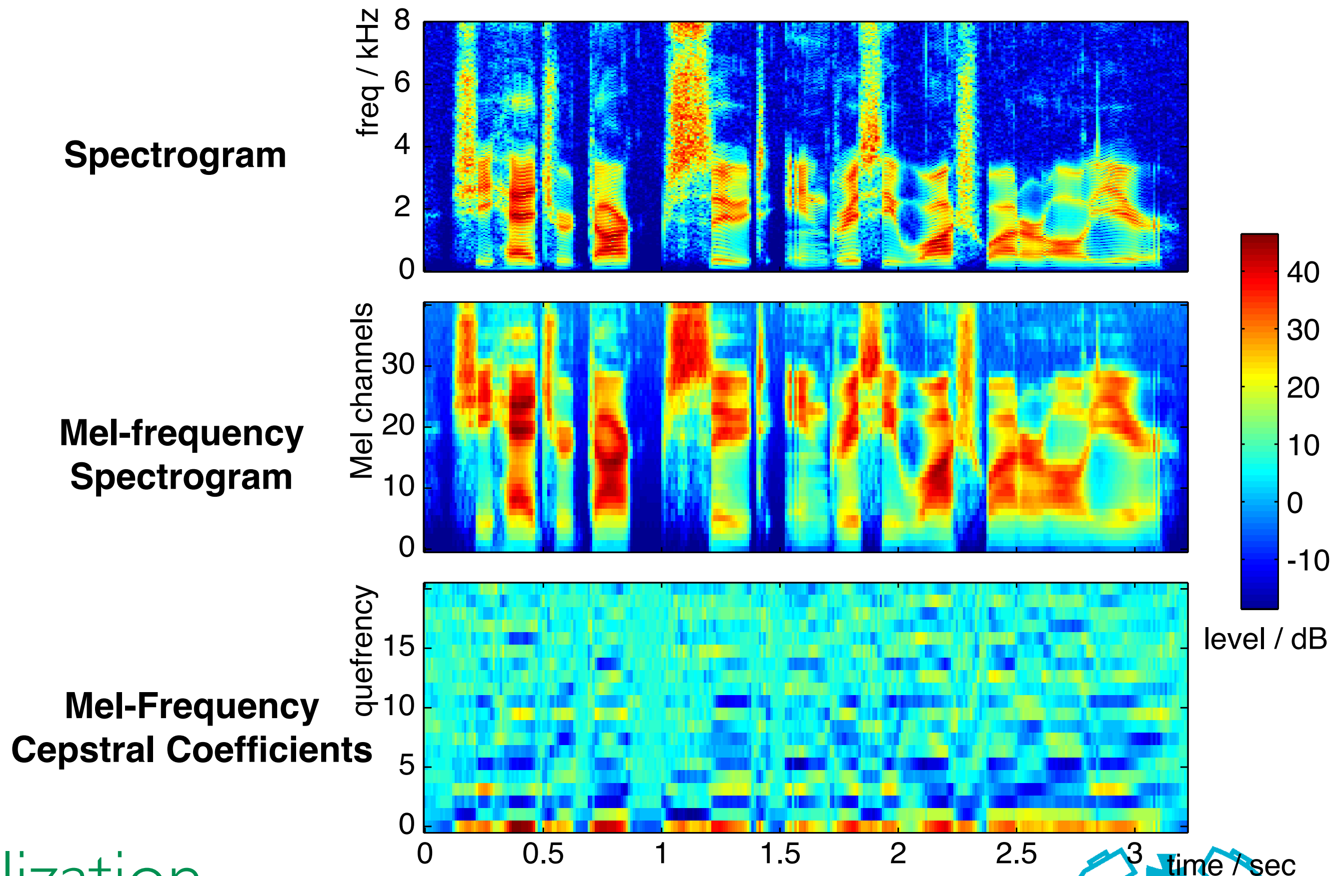


2. Segmentation & Clustering

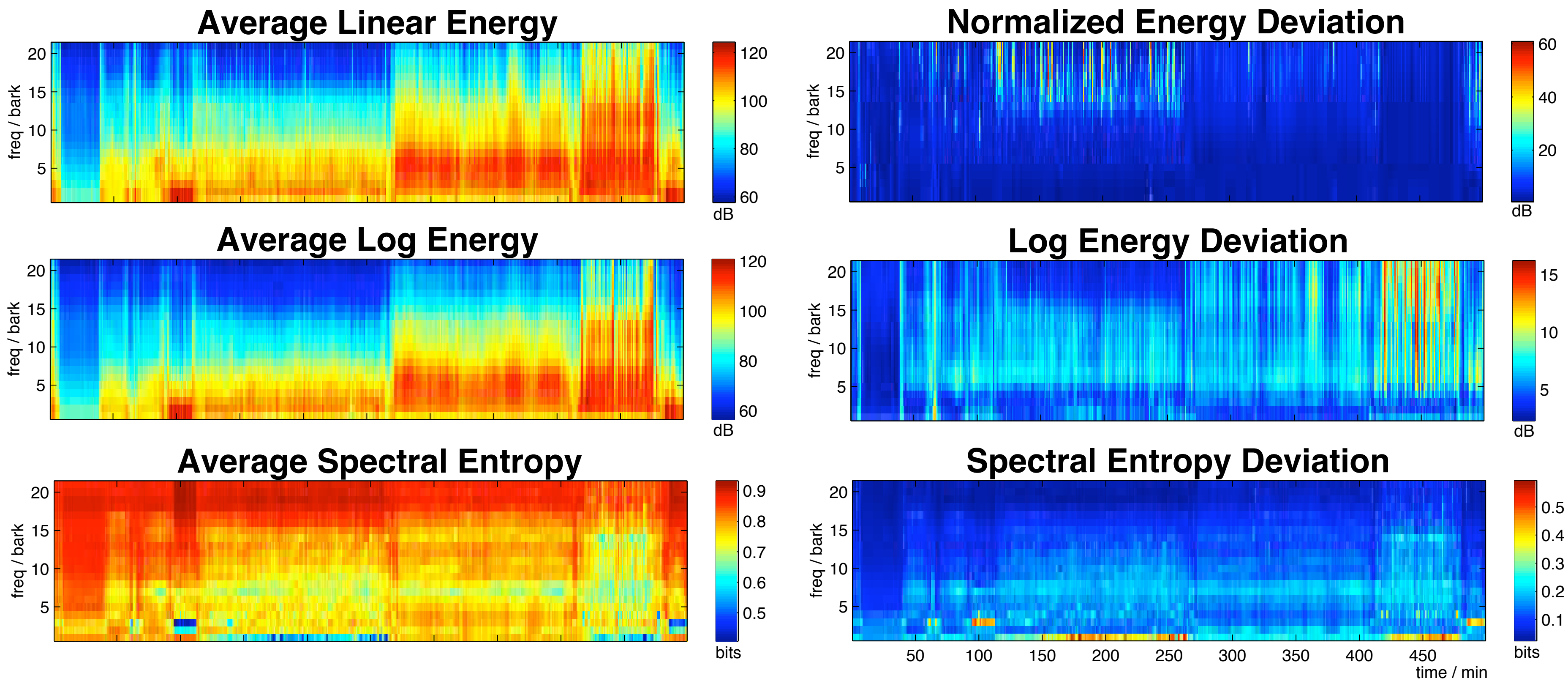
- Top-level structure for long recordings: Where are the major boundaries?
  - e.g. for diary application
  - support for manual browsing
- Length of fundamental time-frame
  - 60s rather than 10ms?
  - background more important than foreground
  - average out uncharacteristic transients
- Perceptually-motivated features
  - .. so results have perceptual relevance
  - broad spectrum + some detail

MFCC Features

- Need "timbral" features:
  Mel-Frequency Cepstral Coeffs (MFCCs)
  - auditory-like frequency warping
  - log-domain
  - discrete cosine transform = orthogonalization



Long-Duration Features

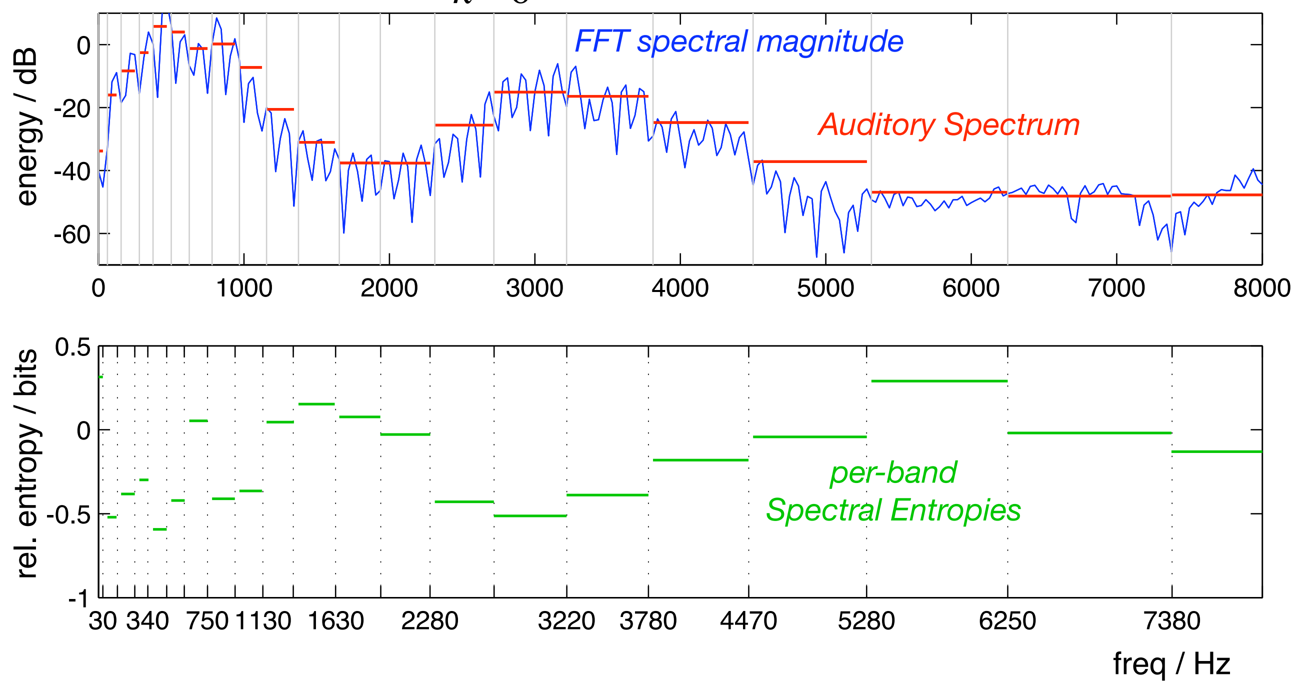


- Capture both average and variation
- Capture a little more detail in subbands...

Spectral Entropy

- Auditory spectrum: $A[n,j] = \sum_{k=0}^{N_F} w_{jk} X[n,k]$
- Spectral entropy ≈ 'peakiness' of each band:

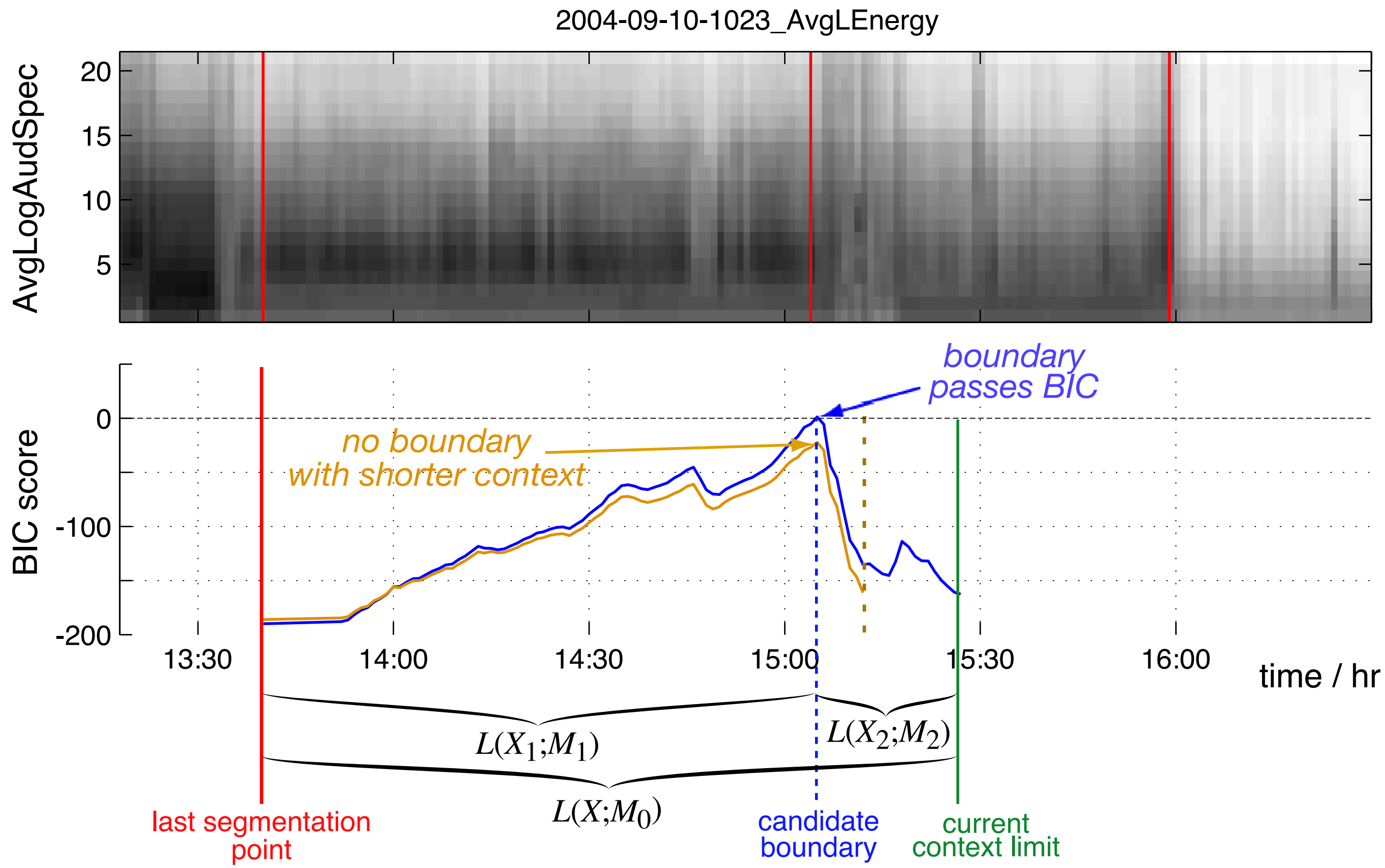
$$H[n,j] = -\sum_{k=0}^{N_F} \frac{w_{jk} X[n,k]}{A[n,j]} \cdot log\left(\frac{w_{jk} X[n,k]}{A[n,j]}\right)$$



BIC Segmentation

- BIC (Bayesian Info. Crit.) compares models:

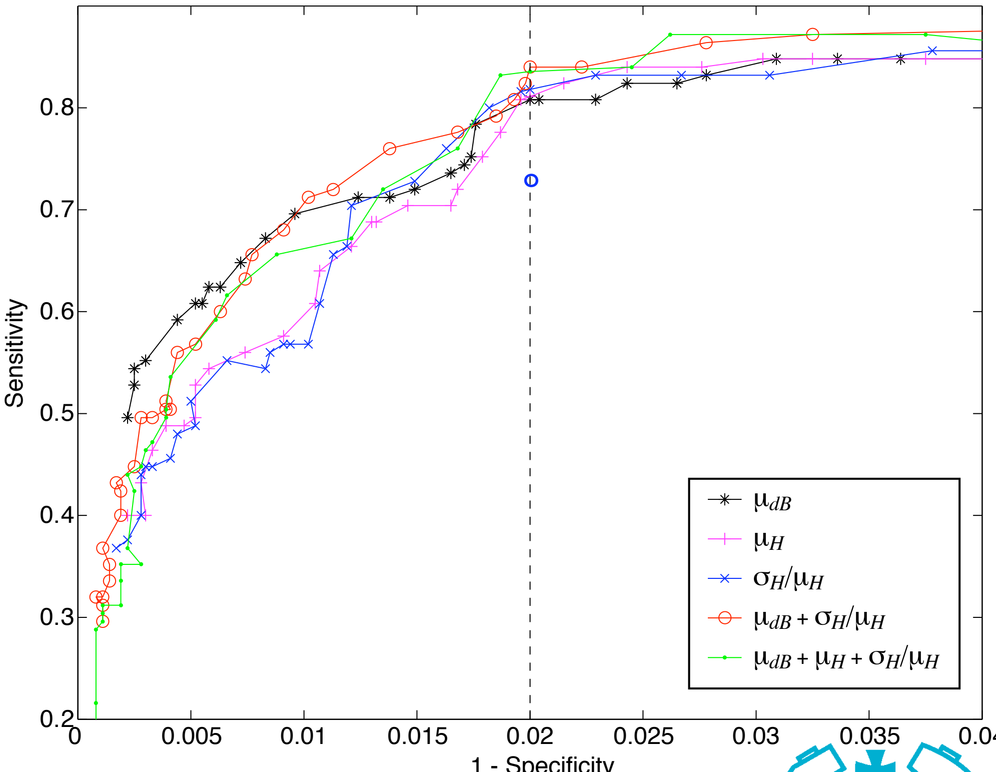
$$\log \frac{L(X_1;M_1)L(X_2;M_2)}{L(X;M_0)} \gtrless \frac{\lambda}{2} \log(N)\Delta\#(M)$$



BIC Segmentation Results

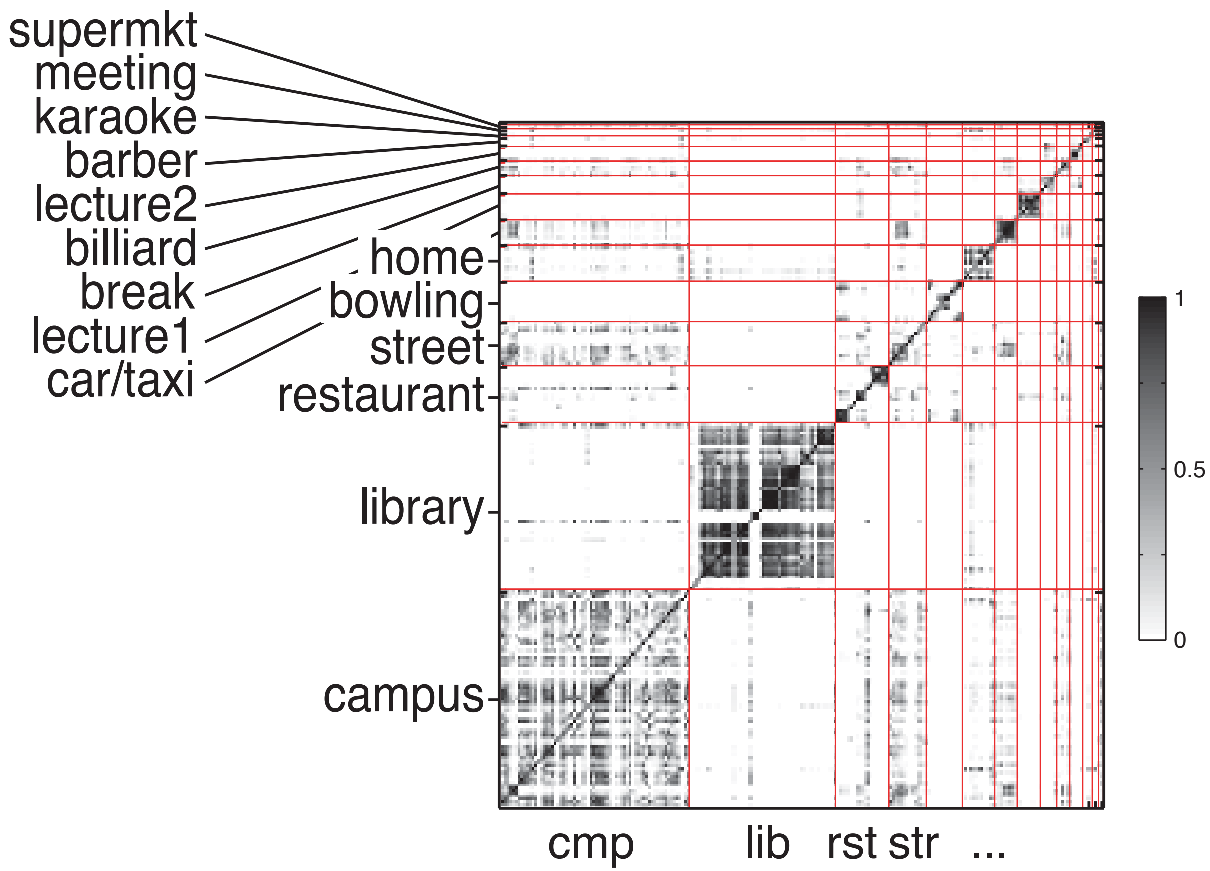
- **Evaluate: 62 hr hand-marked dataset**
  - 8 days, 139 segments, 16 categories
  - measure Correct Accept % @ False Accept = 2%:

Feature	*Correct Accept*
μ_{dB}	80.8%
μ_{H}	81.1%
σ_H/μ_H	81.6%
$\mu_{dB} + \sigma_H/\mu_H$	84.0%
$\mu_{dB} + \sigma_H/\mu_H + \mu_H$	83.6%
mfcc	73.6%



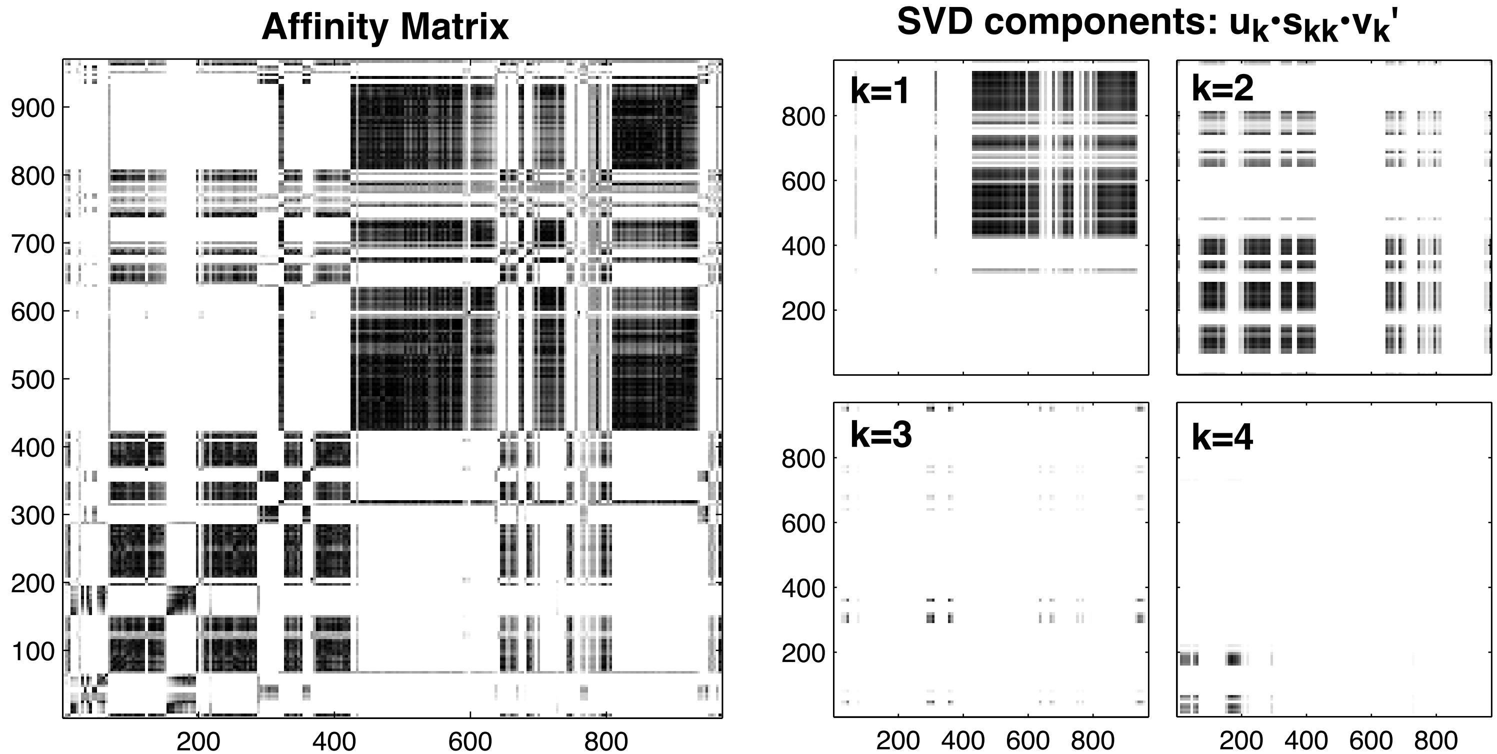
Segment Clustering

- Daily activity has lots of repetition:
  Automatically cluster similar segments
  - 'affinity' of segments as KL2 distances



Spectral Clustering

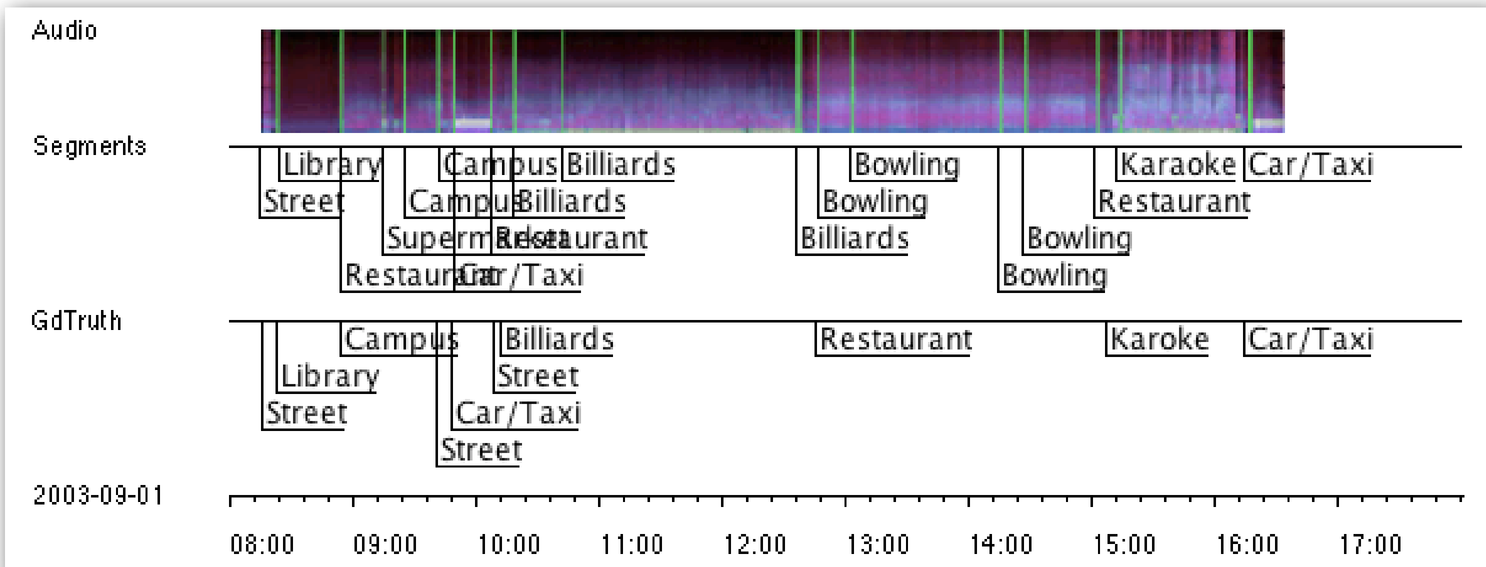
- Eigenanalysis of affinity matrix: $A = U \bullet S \bullet V'$



  - eigenvectors v_k give cluster memberships
- Number of clusters?

Clustering Results

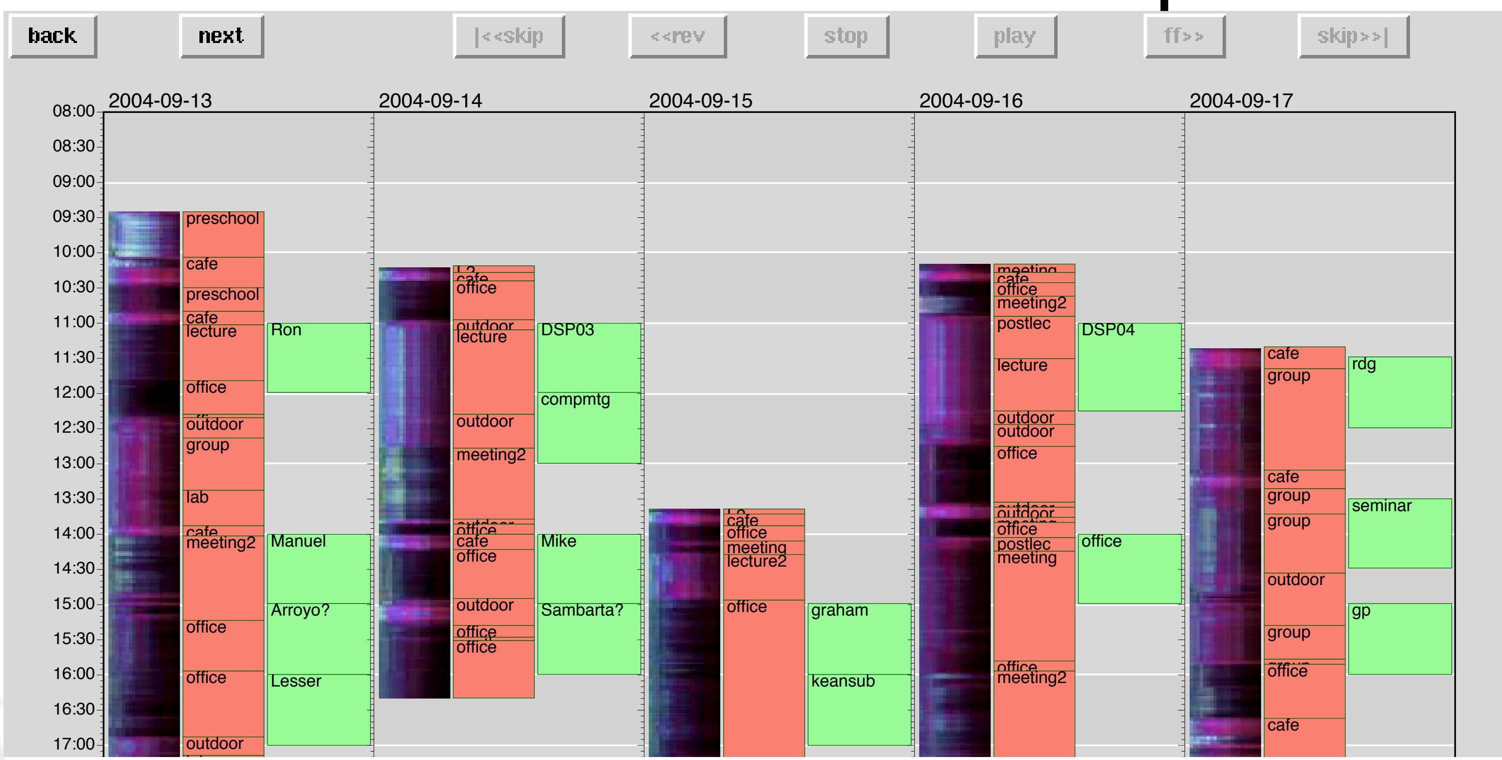
- Clustering of automatic segments gives 'anonymous classes'
  - BIC criterion to choose number of clusters
  - make best correspondence to 16 GT clusters



- Frame-level scoring gives ~70% correct
  - errors when same 'place' has multiple ambiences

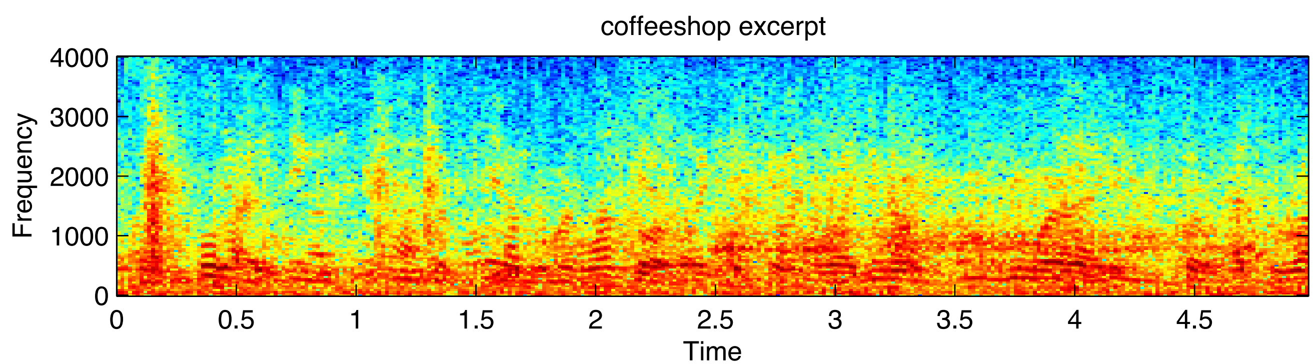
Browsing Interface

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



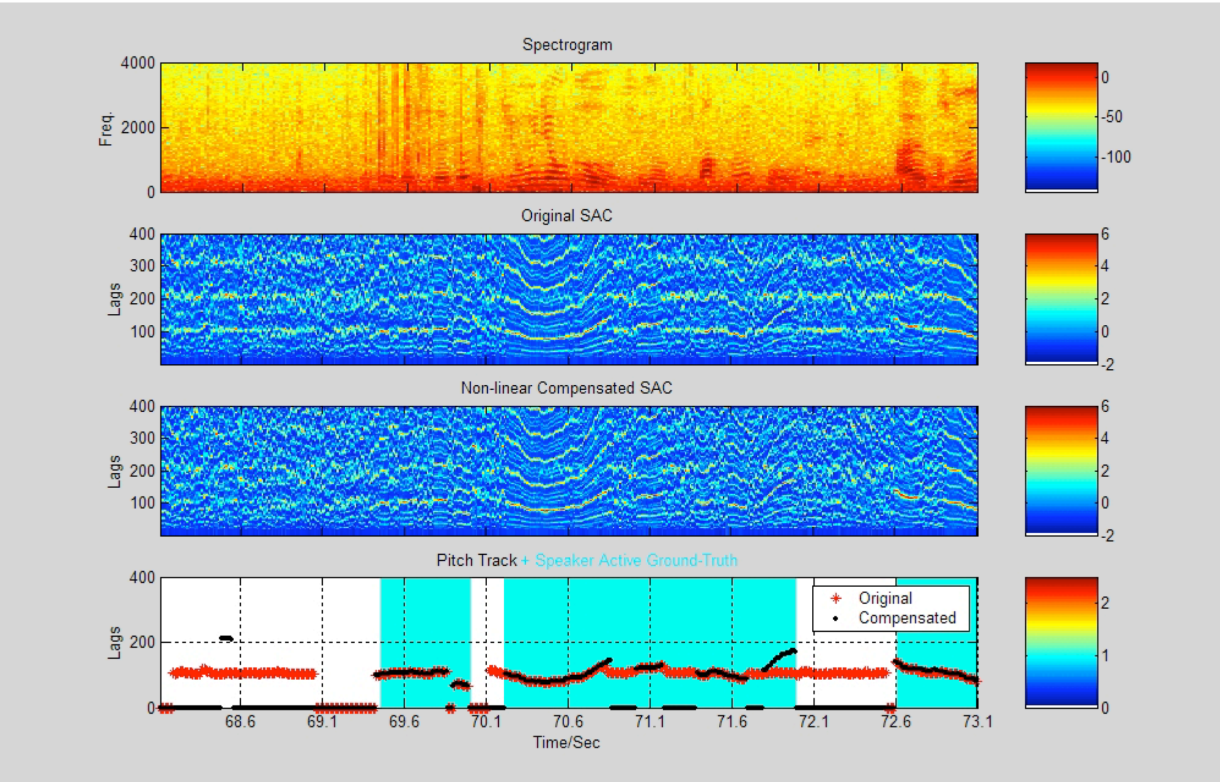
3. Special-Purpose Detectors: Speech

- Speech emerges as most interesting content
- Just identifying speech would be useful
  - goal is speaker identification / labeling
- Lots of background noise
  - conventional Voice Activity Detection inadequate
- Insight: Listeners detect pitch track (melody)
  - look for voice-like periodicity in noise



Voice Periodicity Enhancement

- Noise-robust subband autocorrelation



- Subtract local average
  - suppresses steady background e.g. machine noise
  - 15 min test set; 88% acc (no suppression: 79%)
  - also for enhancing speech by harmonic filtering

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio
Columbia University in the City of New York

Detecting Repeating Events

with Jim Ogle

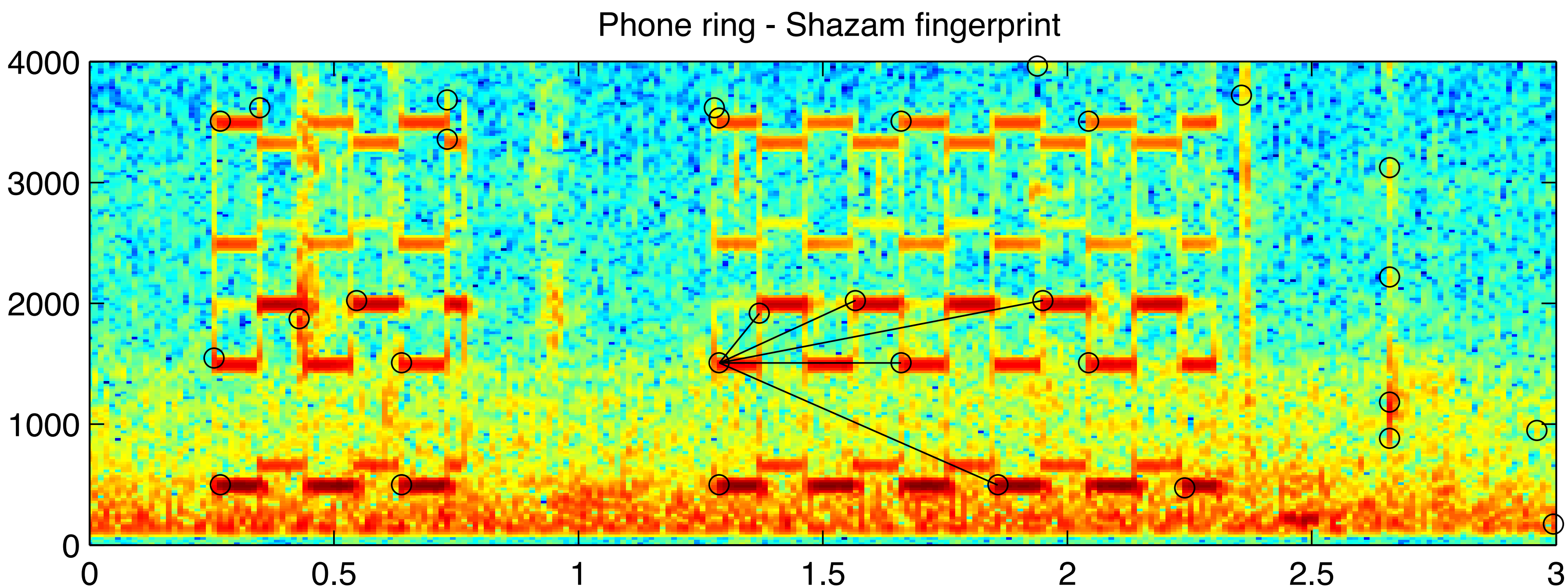
- Recurring sound events can be informative
  - indicate similar circumstance...
  - but: define "event" – sound organization
  - define "recurring event" – how similar?
  - .. and how to find them – tractable?
- Idea: Use hashing (fingerprints)
  - index points to other occurrences of each hash; intersection of hashes points to match - much quicker search
  - use a fingerprint insensitive to background?

LabROSA Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

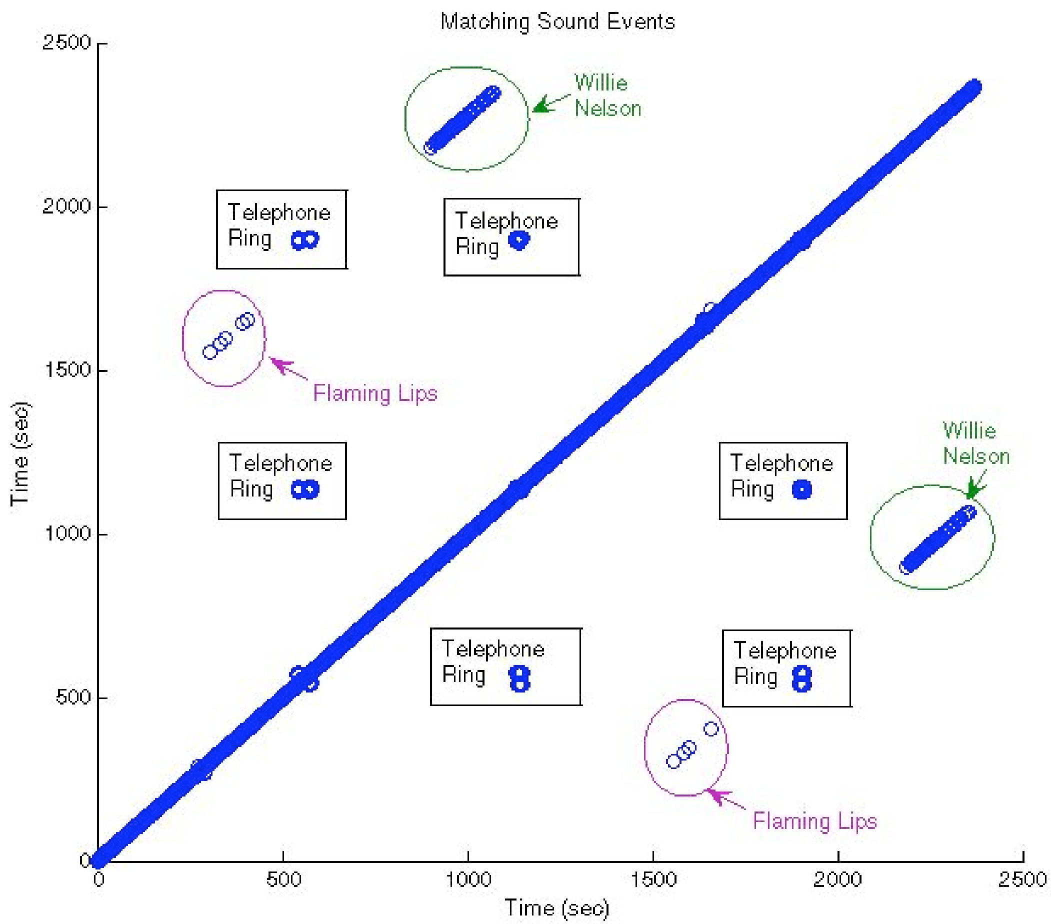
Shazam Fingerprints

- Prominent spectral onsets are landmarks;
  Use relations $\{f_1, f_2, \Delta t\}$ as hashes



  - intrinsically robust to background noise

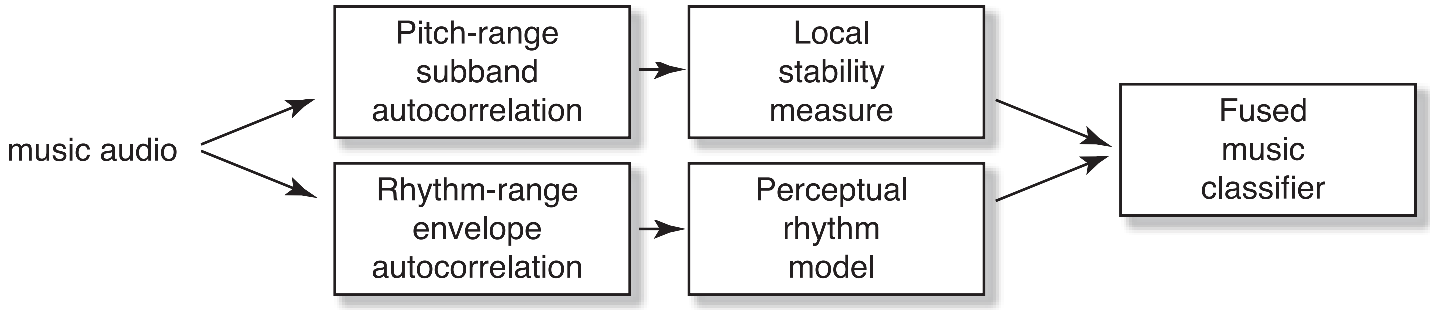
Exhaustive Search for Repeats



- More selective hashes →
  - few hits required to confirm match (faster; better precision)
  - but less robust to backgound (reduce recall)
- Works well when exact structure repeats
  - recorded music, electronic alerts
  - no good for "organic" sounds e.g. garage door

Music Detector

- Two characteristic features for music
  - strong, sustained periodicity (notes)
  - clear, rhythmic repetition (beat)
  - at least one should be present!



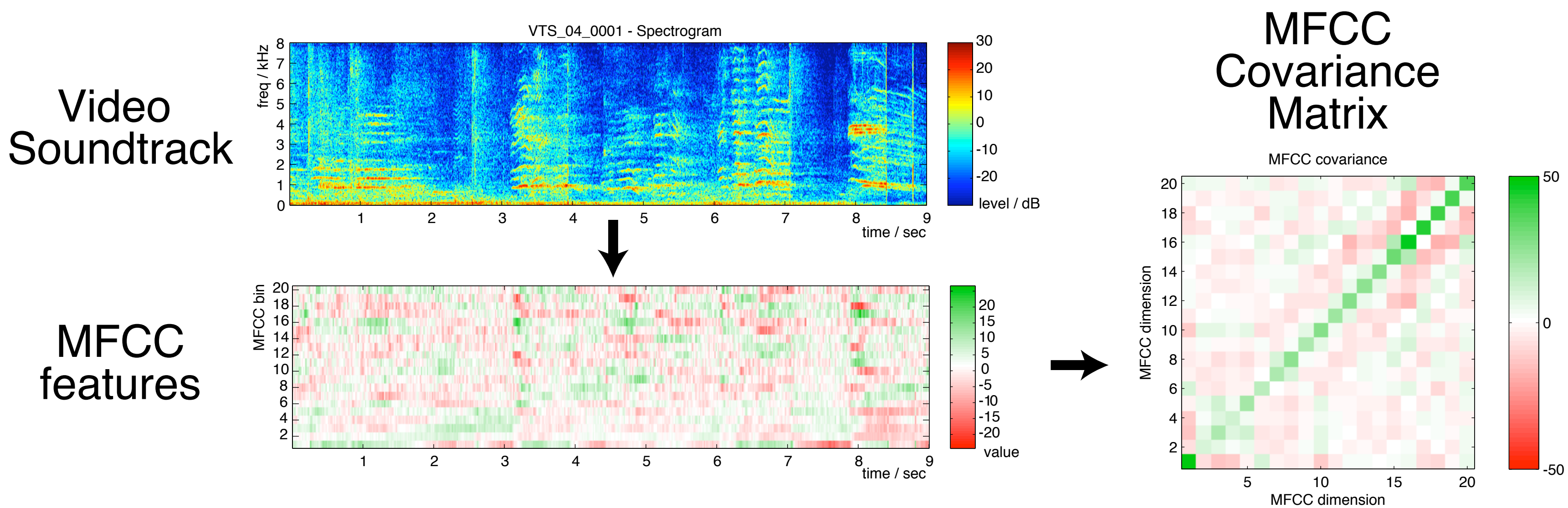
- Noise-robust pitch detector
  - looks for high-order autocorrelation
- Beat tracker
  - .. from Music IR work

4. Generic Concept Detectors

- Consumer Video application:
  How to assist browsing?
  - system automatically tags recordings
  - tags chosen by usefulness, feasibility
- Initial set of 25 tags defined:
  - "animal", "baby", "cheer", "dancing" ...
  - human annotation of 1300+ videos
  - evaluate by average precision
- Multimodal detection
  - separate audio + visual low-level detectors
  - (then fused...)

MFCC Covariance Representation

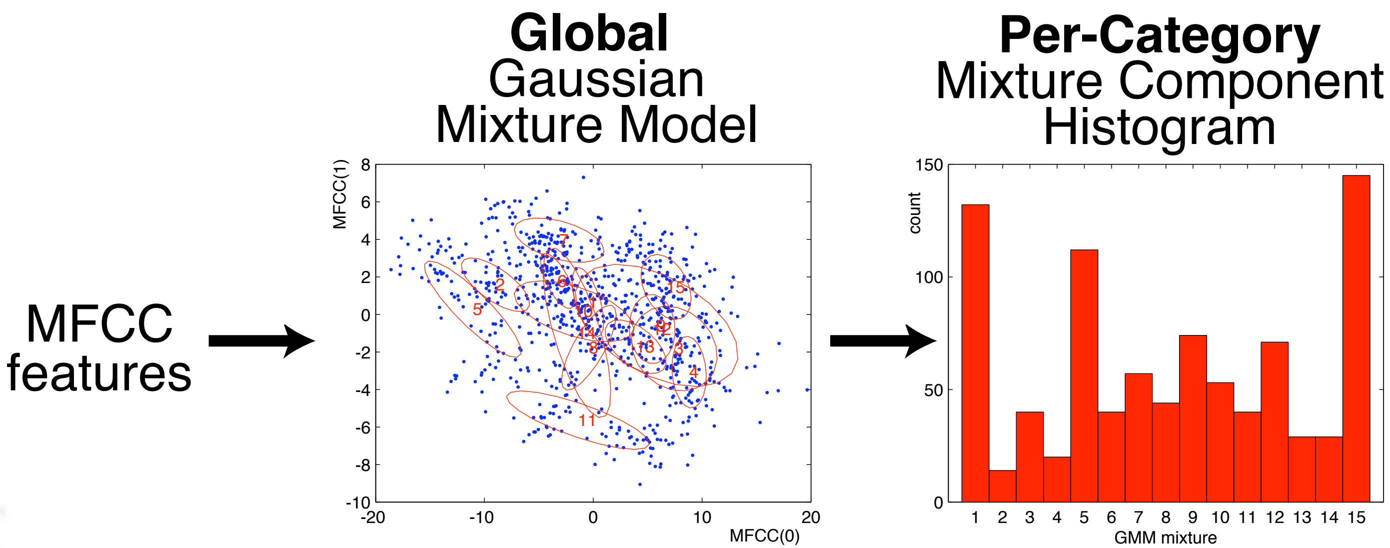
- Each clip/segment → fixed-size statistics
  - similar to speaker ID and music genre classification
- Full Covariance matrix of MFCCs
  - maps the kinds of spectral shapes present



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

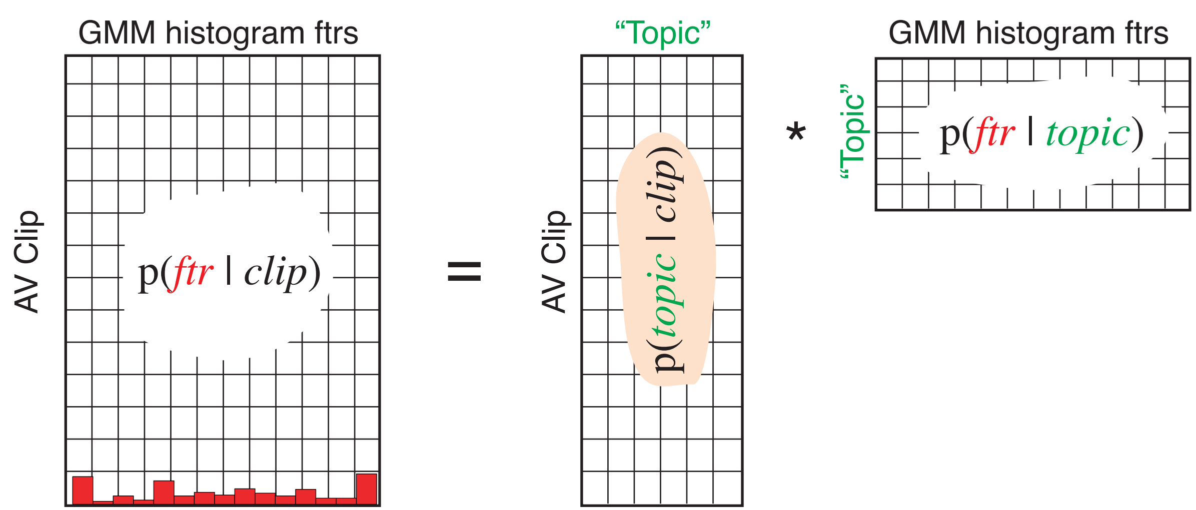
GMM Histogram Representation

- Want a more 'discrete' description
  - .. to accommodate nonuniformity in MFCC space
  - .. to enable other kinds of models...
- Divide up feature space with a single Gaussian Mixture Model
  - .. then represent each clip by the components used



Latent Semantic Analysis (LSA)

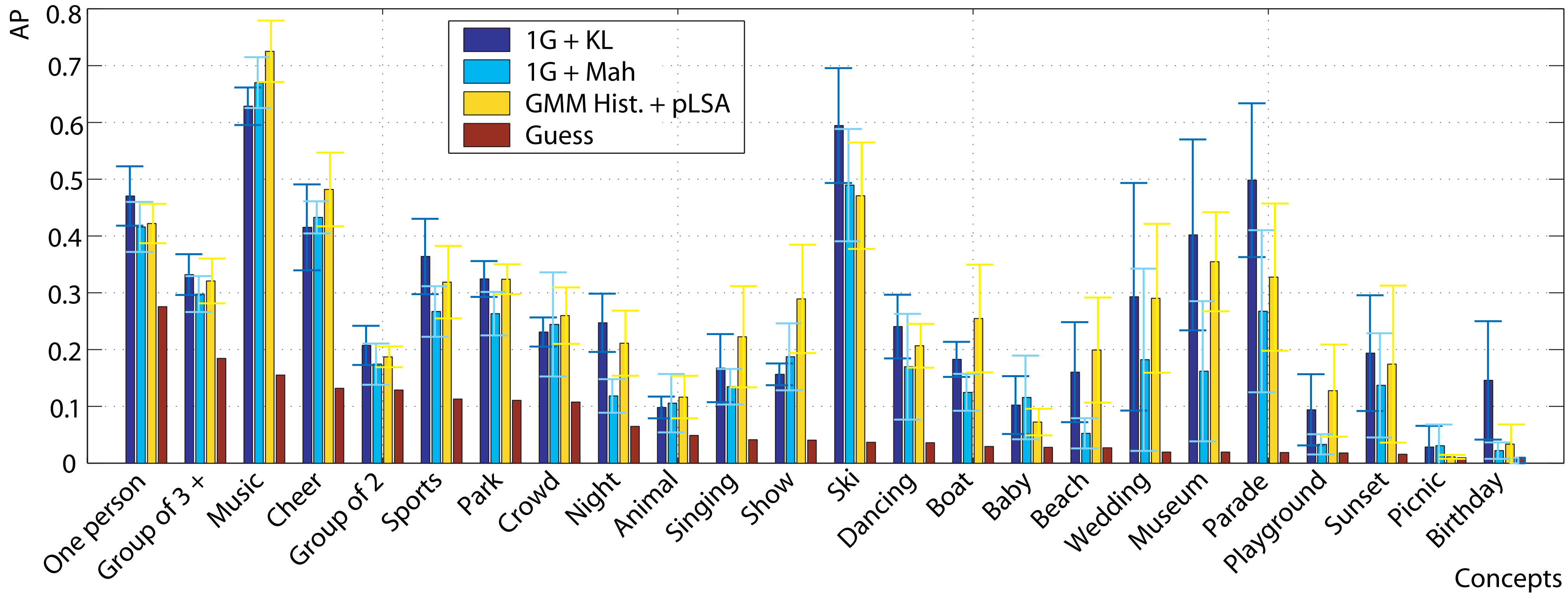
- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - $p(ftr \mid clip) = \sum_{topics} p(ftr \mid topic)\, p(topic \mid clip)$



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

Audio-Only Results

- Wide range of results:

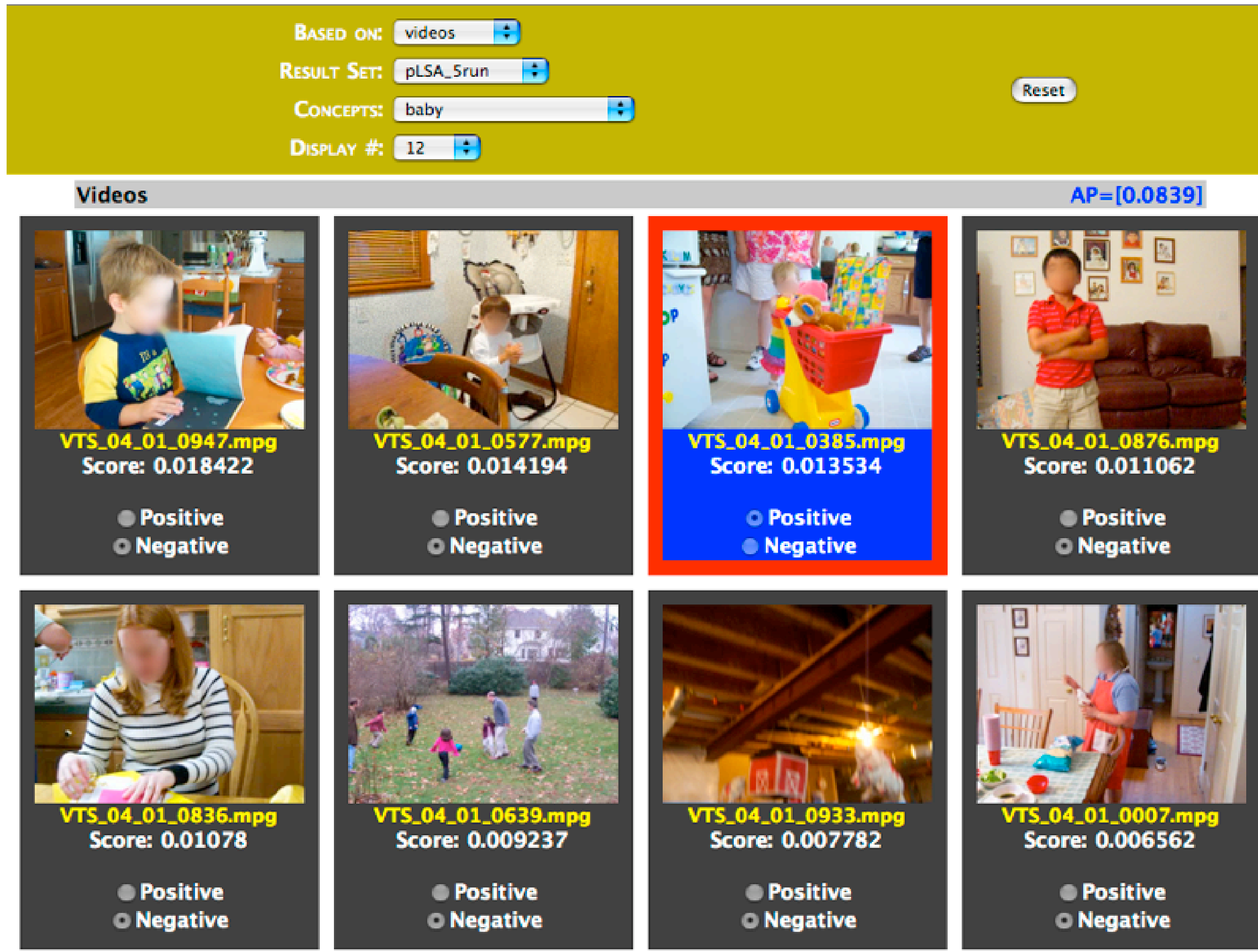


  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

How does it ‘feel’?

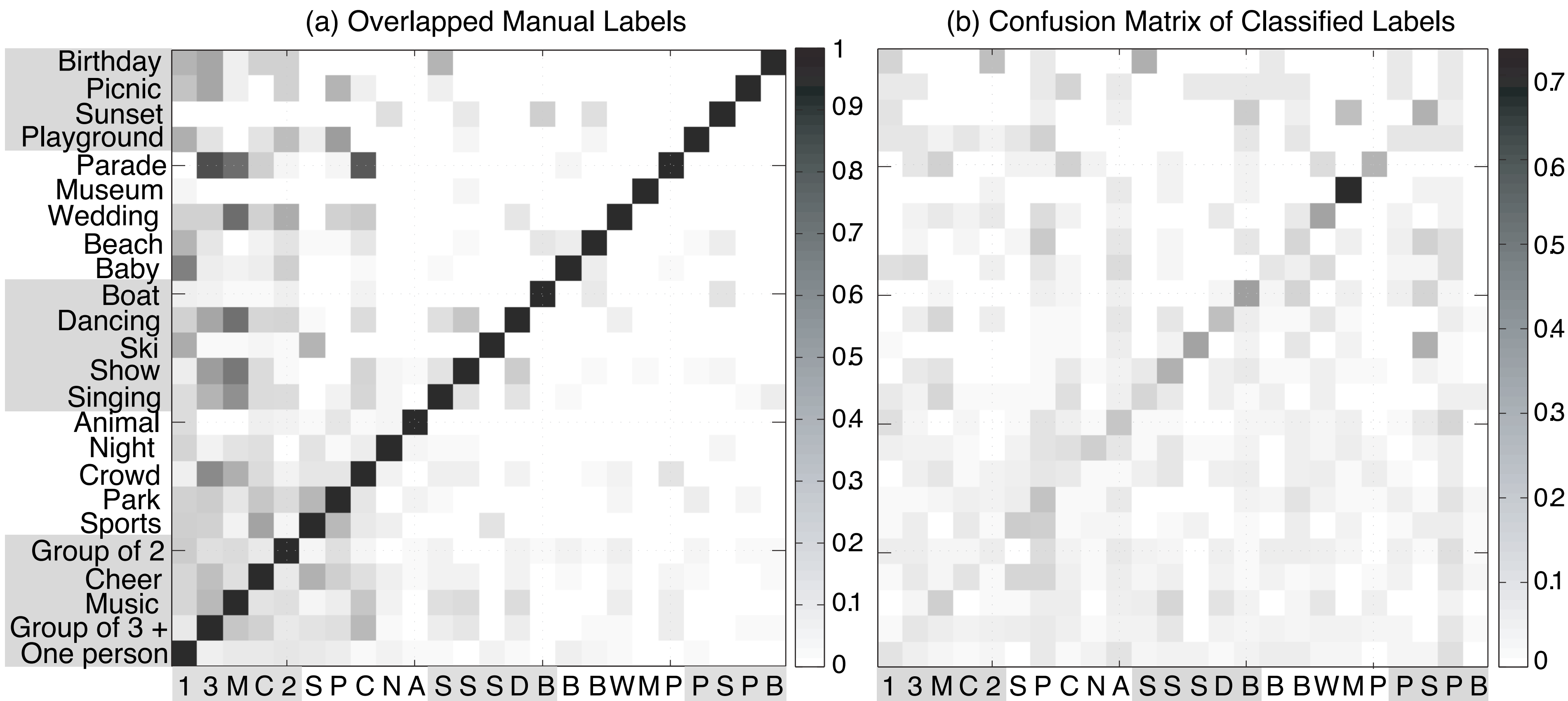
- Browser impressions: How wrong is wrong?

Top 8 hits for “Baby”



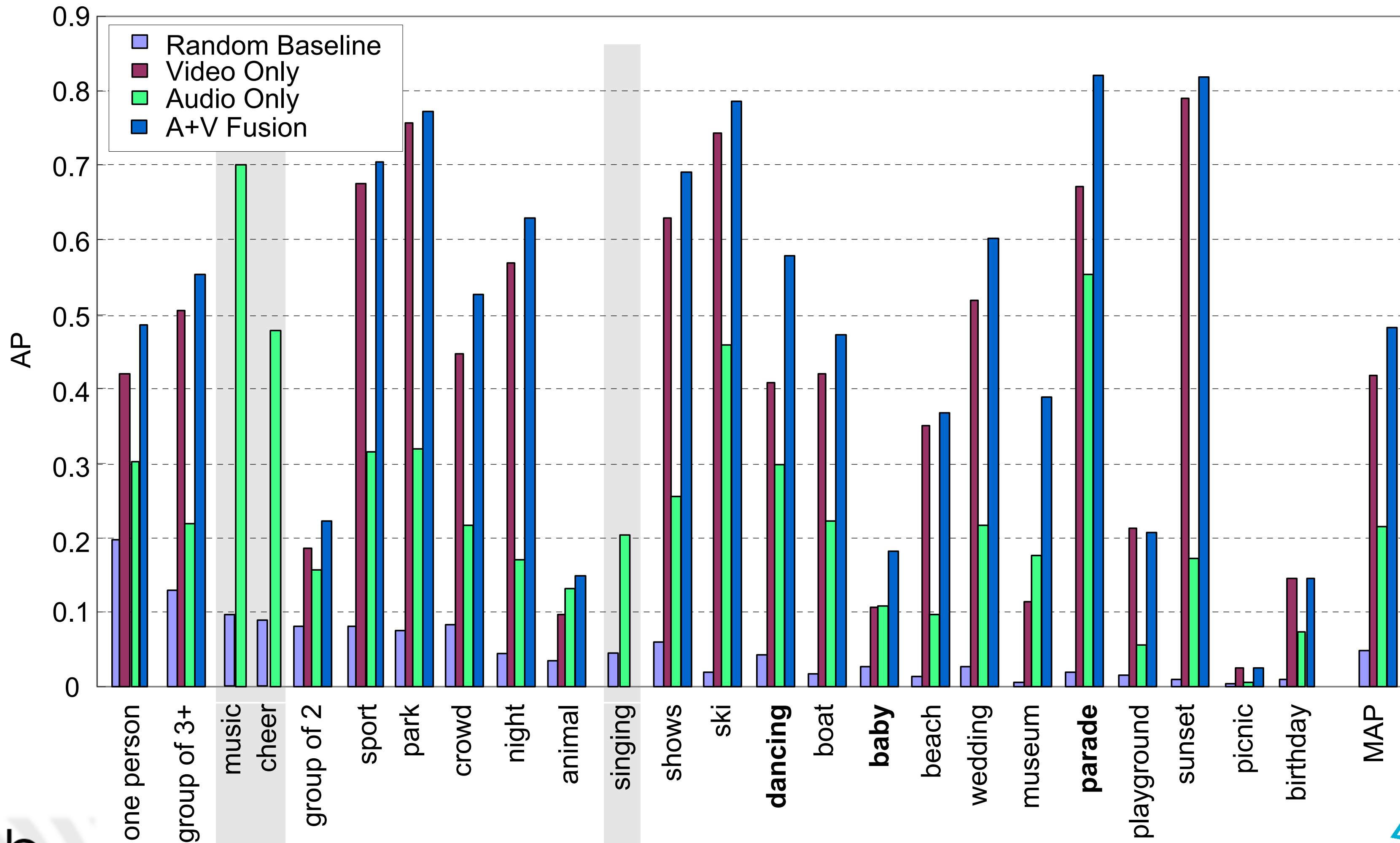
Confusion analysis

- Where are the errors coming from?



Fused Results - AV Joint Boosting

- Audio helps in many classes

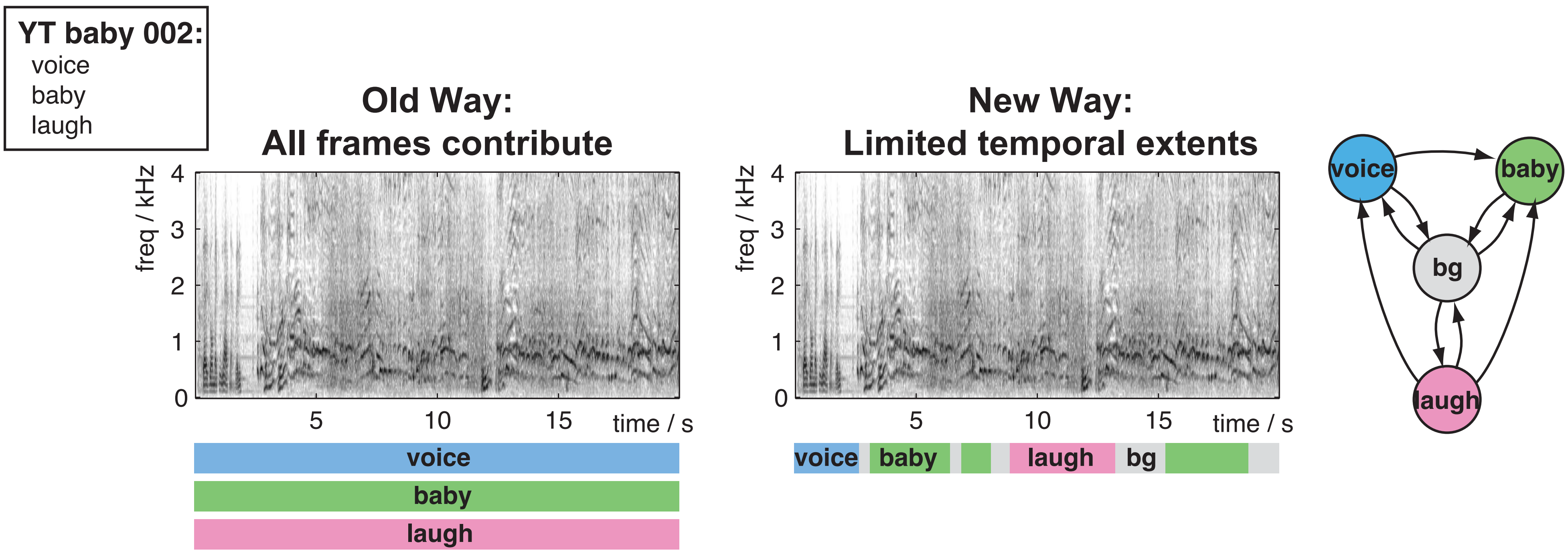


Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

5. Future: Temporal Focus

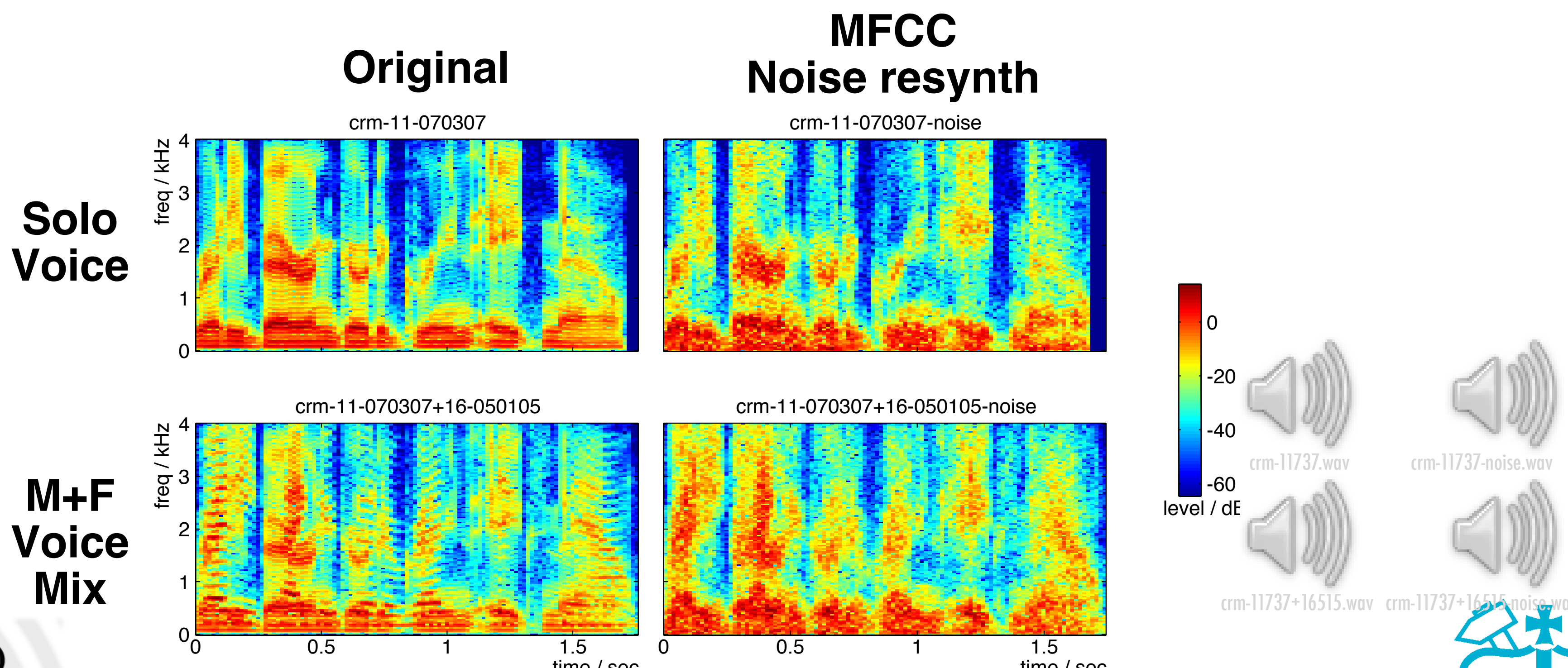
- Global vs. local class models
  - tell-tale acoustics may be 'washed out' in statistics
  - try iterative realignment of HMMs:



  - "background" (bg) model shared by all clips

Handling Sound Mixtures

- MFCCs of mixtures ≠ mix of MFCCs
  - recognition despite widely varying background?
  - factorial models / Nonnegative Matrix Factorization
  - sinusoidal / landmark techniques



Larger Datasets

- **Many detectors are visibly data-limited**
  - getting data is ~ hard
  - labeling data is expensive
- **Bootstrap from YouTube etc.**
  - lots of web video is edited/dubbed...
    - need a "consumer video" detector?
- **Preliminary YouTube results disappointing**
  - downloaded data needed extensive clean-up
  - models did not match Kodak data
- **(Freely available data!)**

Conclusions

- Environmental sound contains information
  - .. that's why we hear!
  - .. computers can hear it too
- Personal audio can be segmented, clustered
  - find specific sounds to help navigation/retrieval
- Consumer video can be 'tagged'
  - .. even in unpromising cases
  - audio is complementary to video
- Interesting directions for better models