Minimal-Impact Personal Audio Archives

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

dpwe@ee.columbia.edu

1. "Personal Audio" Archives
2. Segmenting & Clustering
3. Speech Detection
4. Repeated Events
5. Future

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

Information in Audio

- **Long-duration 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
  - gaze, posture, motion, ...

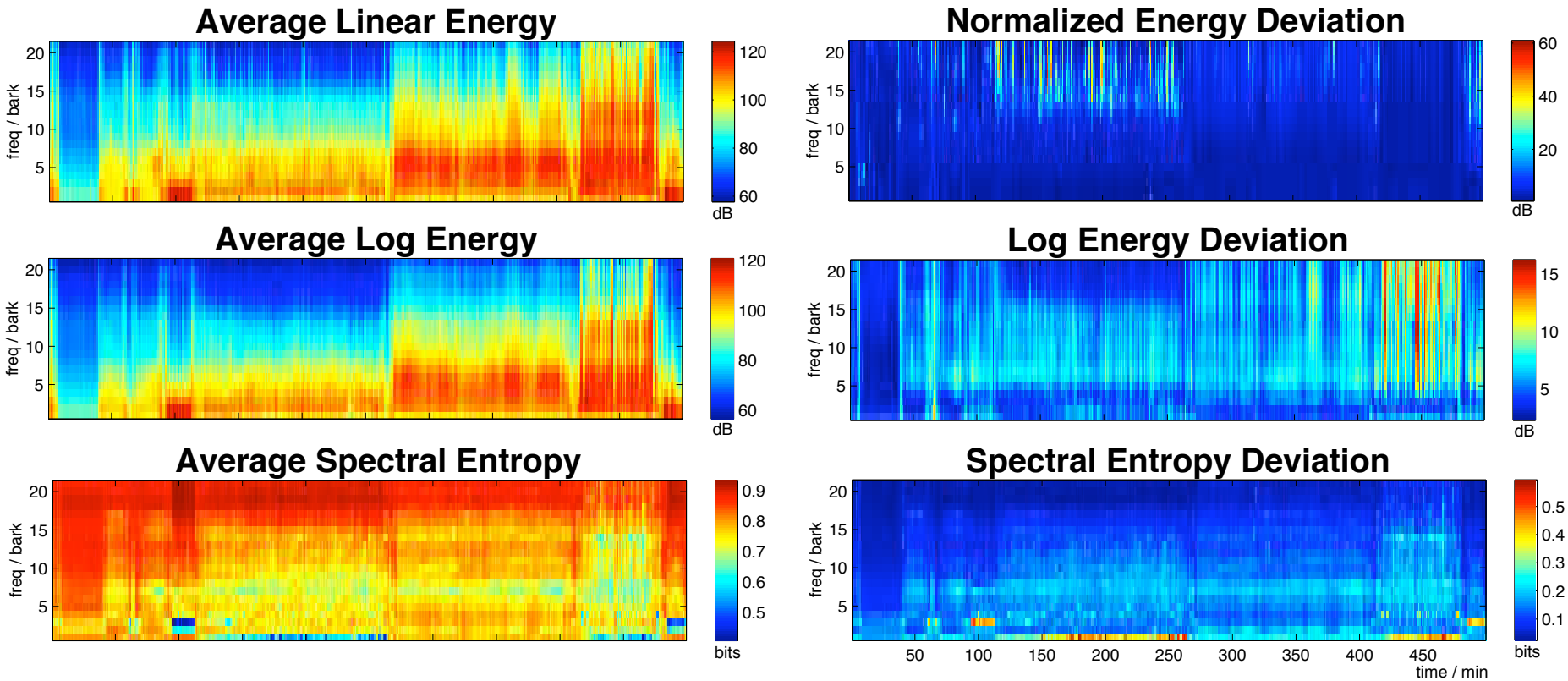
Applications

- 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?

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

Features

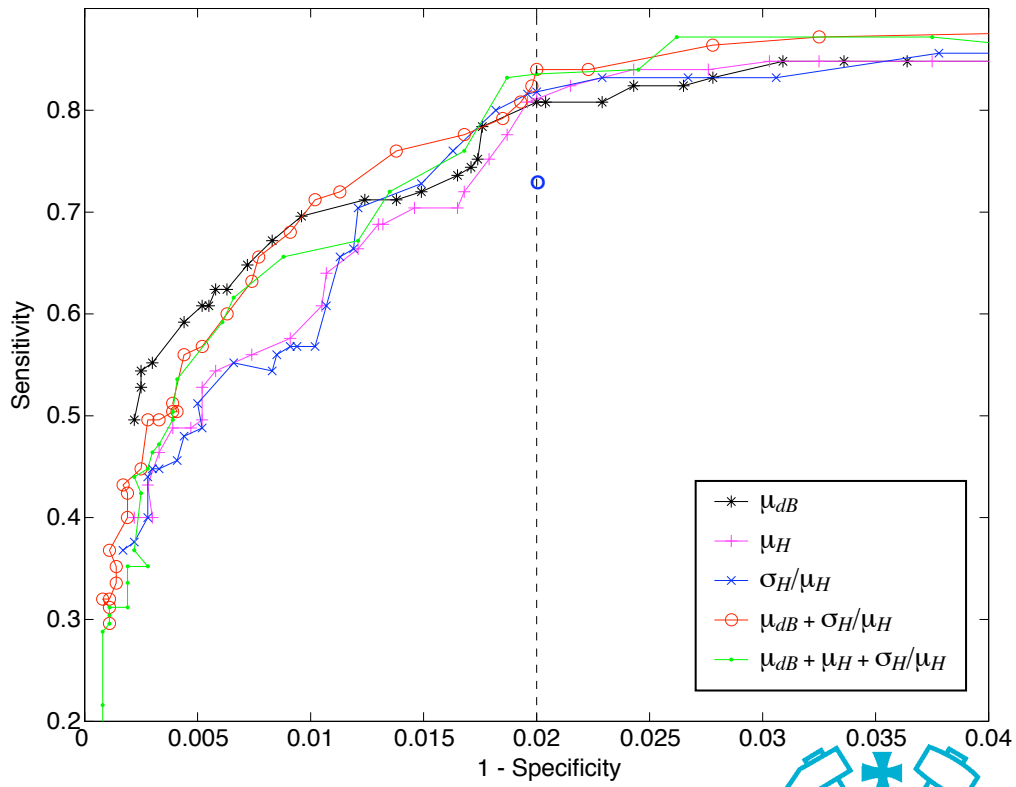


- Capture both average and variation
- Capture a little more detail in subbands...

BIC Segmentation Results

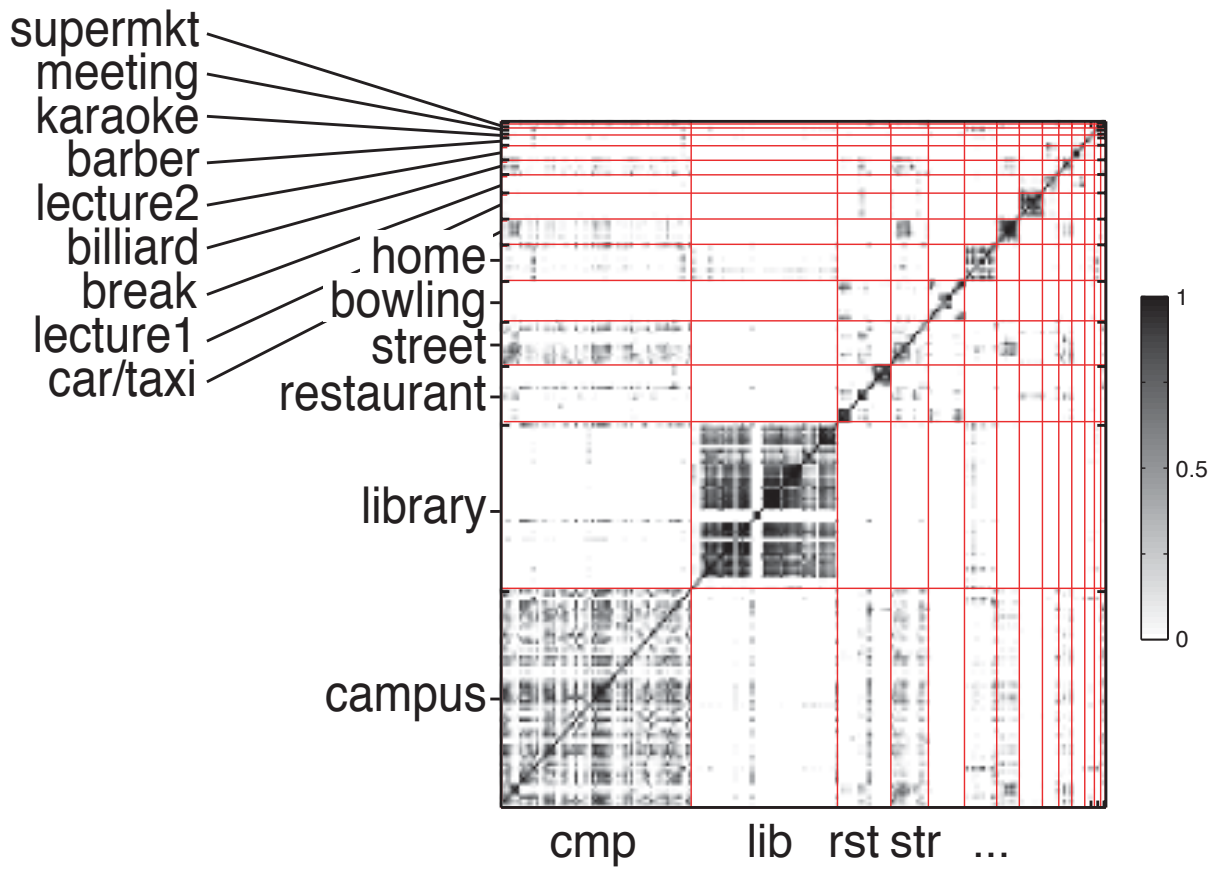
- Evaluate: 62 hr hand-marked dataset
  - 8 days, 139 segments, 16 categories
  - measure Correct Accept % @ False Accept = 2%:

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%



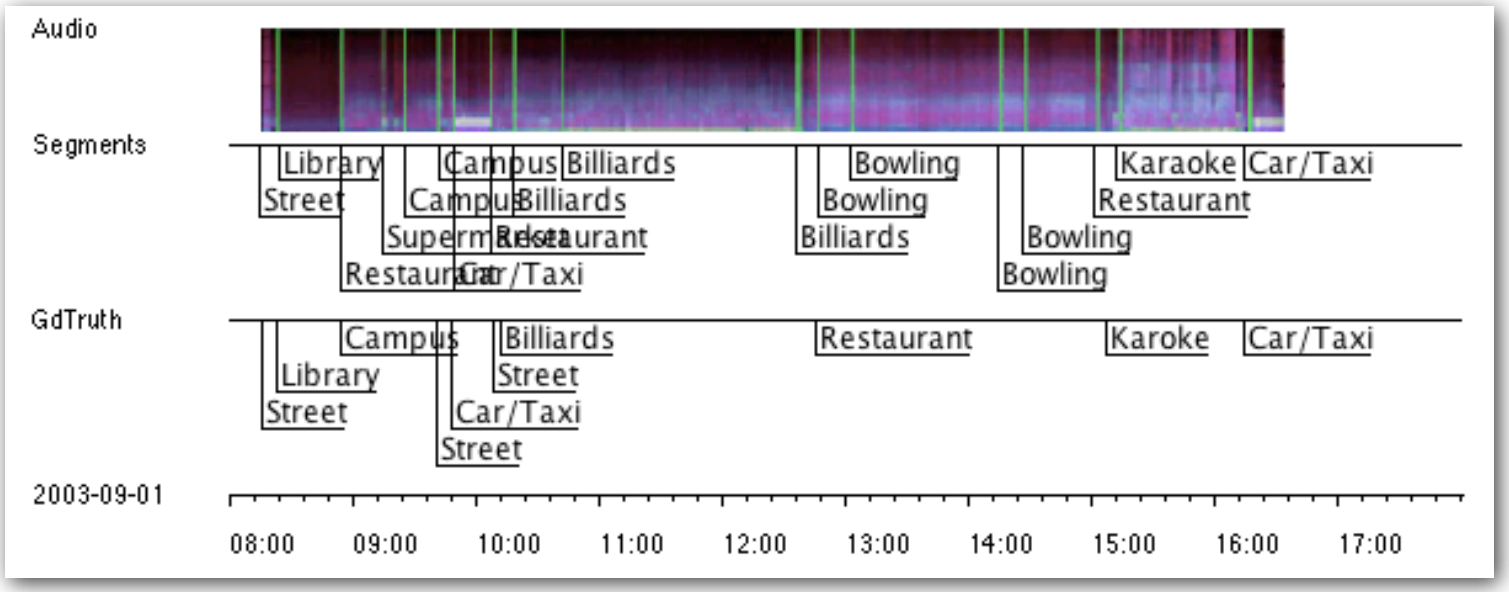
Segment Clustering

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



Clustering Results

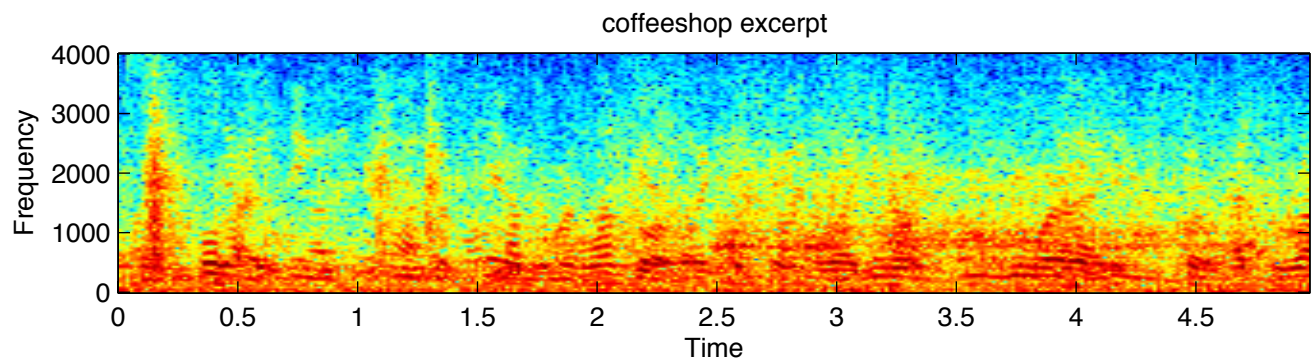
- Clustering of automatic segments gives 'anonymous classes'
  - BIC criterion to choose number of clusters
  - make best correspondence to 16 GT clusters



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

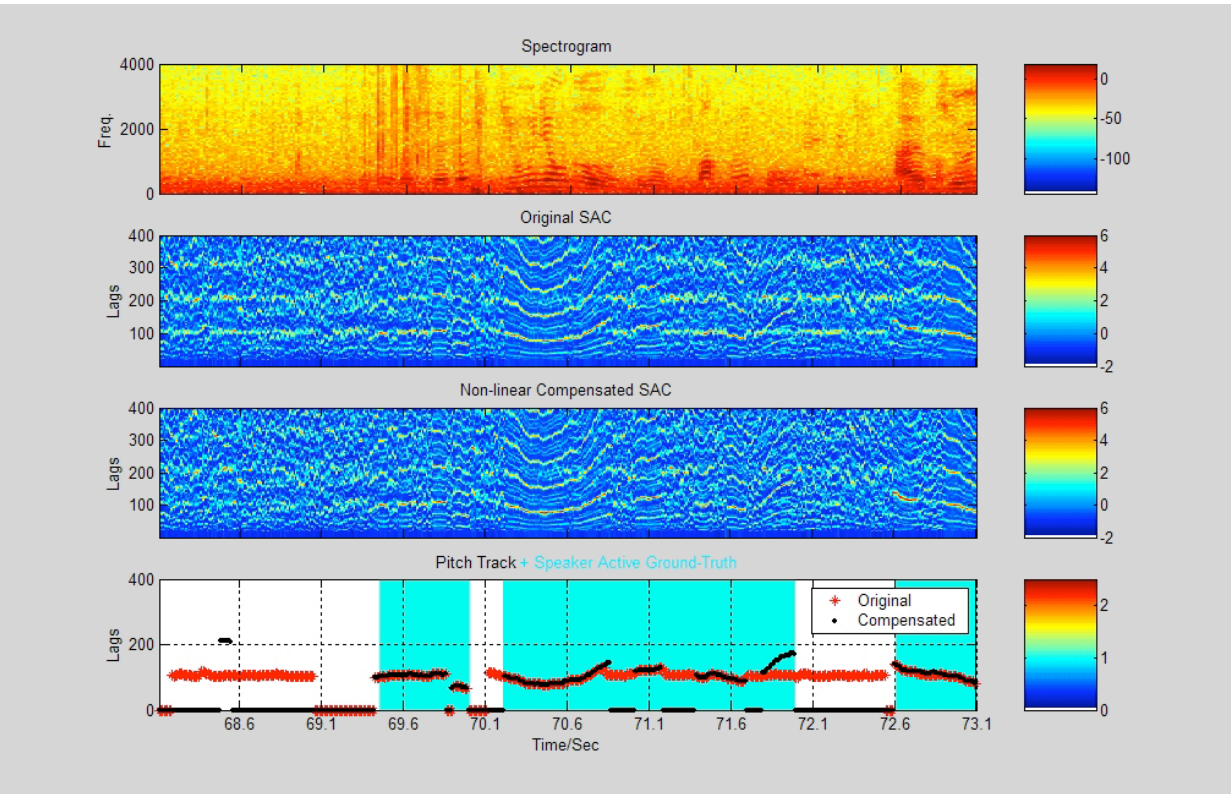
3. Speech Detection

- 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

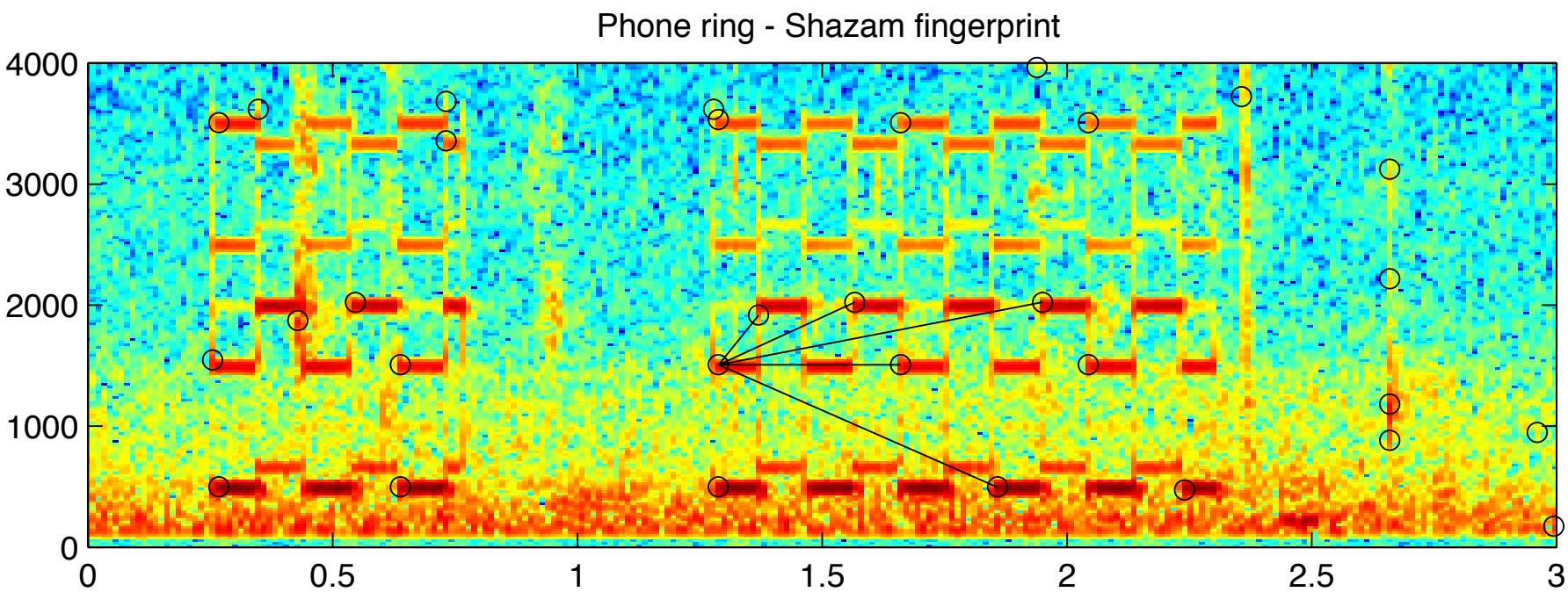
- 15 min test set; 88% acc (79% w/o enhancement)
- also for enhancing speech (harmonic filtering)

4. Repeating Events

- **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?

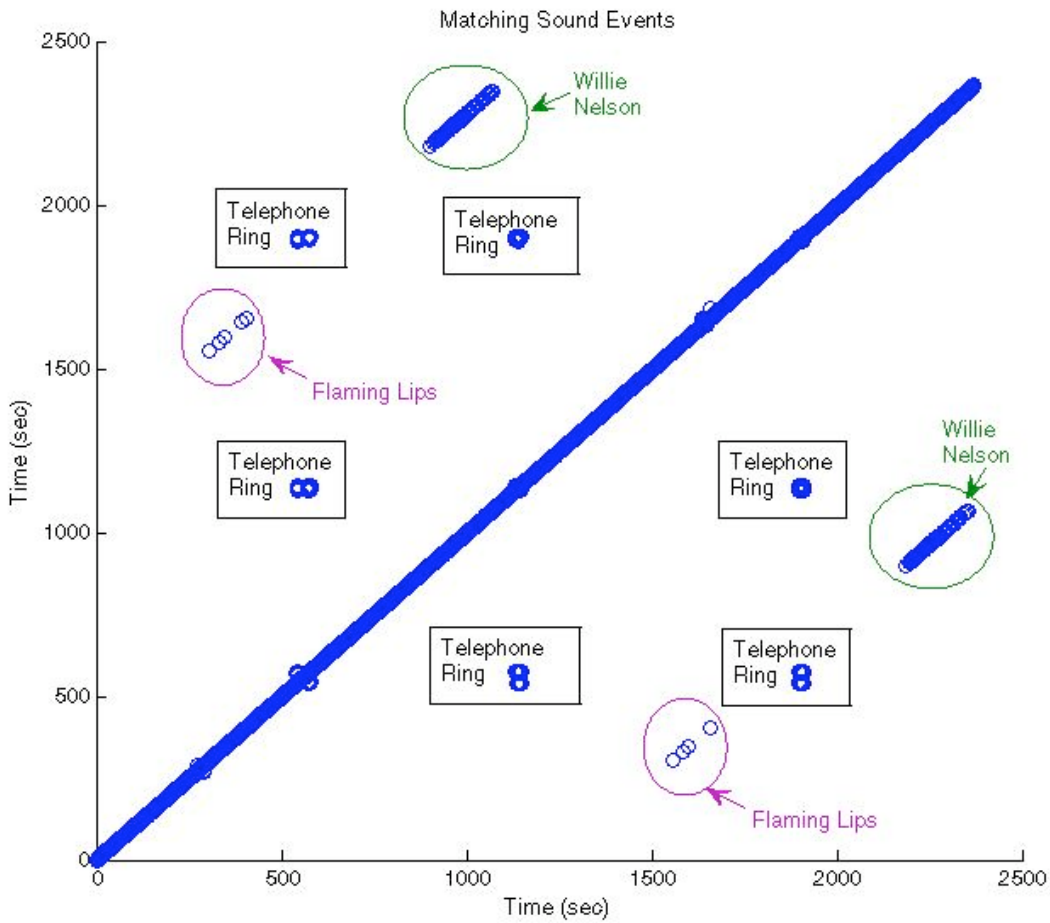
Shazam Fingerprints

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



- intrinsically robust to background noise

Exhaustive Search for Repeats

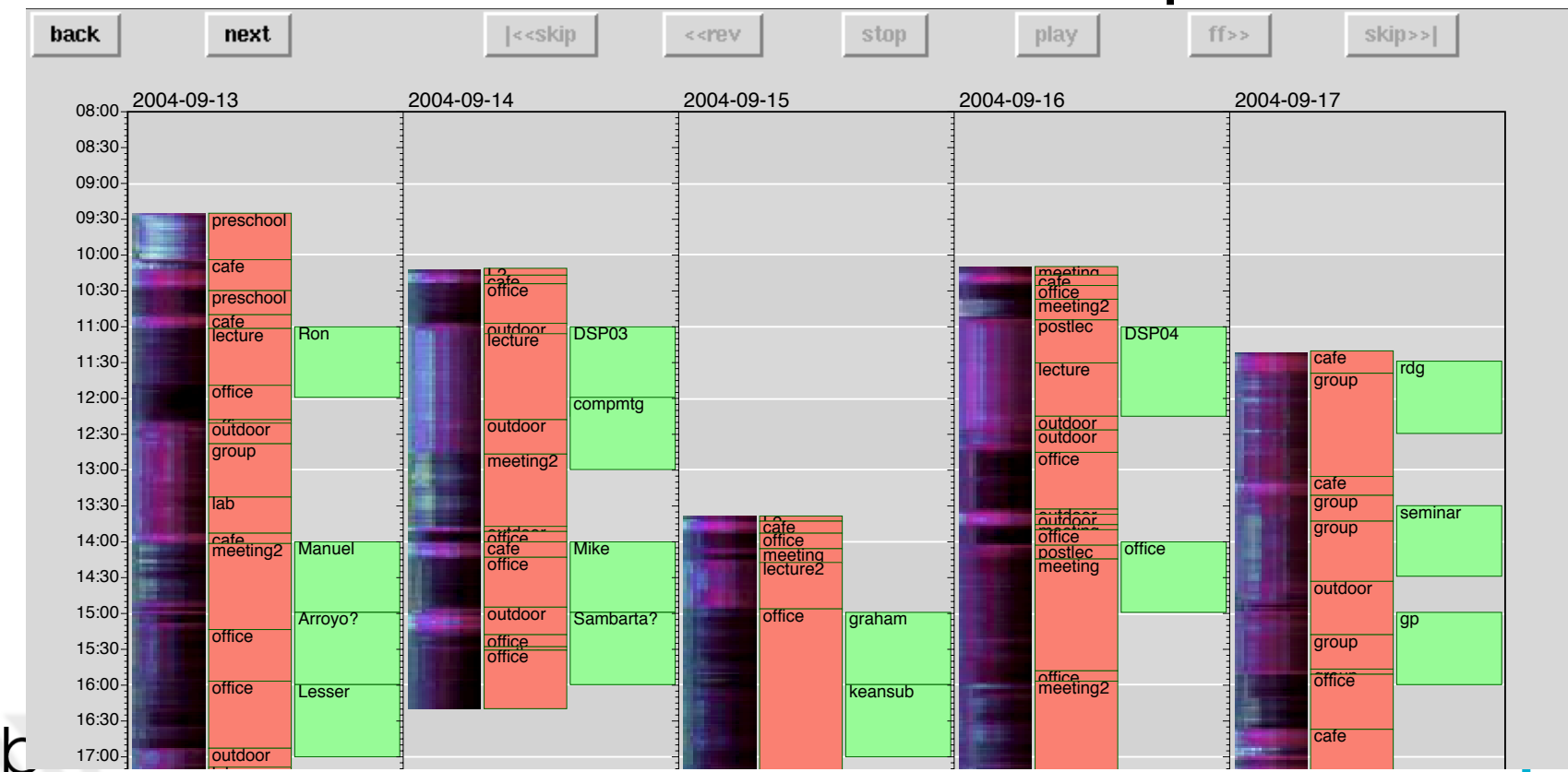


- More selective hashes →
  - few hits required to confirm match (faster; better precision)
  - but less robust to backgound (reduce recall)

- Works well when exact structure repeats
  - recorded music, electronic alerts
  - no good for "organic" sounds e.g. garage door

5. Future: Browsing Tools

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



Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

Future: Speech Recognition

- Most audio is too noisy for standard ASR
  - actually reassuring for privacy issues
- But... similar to "Meeting Recordings"
  - NIST "distant microphone" conditions

- Speech enhancement - directional filtering
  - 2 channels a big improvement over one
  - ... use a more special-purpose directional mic?

Privacy and Security

- **Recordings are controversial**
  - privacy expectations: speech should be ephemeral?
  - "Oops button", delayed review (Roy)
  - subpoenas... (Golubchik)
- **Access to recordings is very sensitive**
  - .. but preservation is important too
- **Approaches**
  - don't store intelligible audio .. but lessens utility
    - maybe store ASR output?
  - split and store on multiple machines
    - tiered, distributed trust/access protocols
- **Big issue!**

Conclusions

- "Personal Audio" is easy & cheap to collect
  - but is it any use?
- Segmentation/clustering works well
- Voice detection in noise is harder
  - prospects for speaker identification
- Hashing to find arbitrary repeating events
- Tools distribution as a goal