Segmenting and Classifying Long-Duration Recordings of “Personal Audio”

Dan Ellis and Keansub Lee
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA
{dpwe,kslee}@ee.columbia.edu

1. “Personal Audio”
2. Features
3. Segmentation
4. Clustering
5. Future Work

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

Columbia University in the City of New York

1. Personal Audio

- Easy to record everything you hear
  - <2GB / week @ 64 kbps
- Very hard to find anything
  - how to scan?
  - how to visualize?
  - how to index?

- Need automatic analysis

Applications

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

Data Set

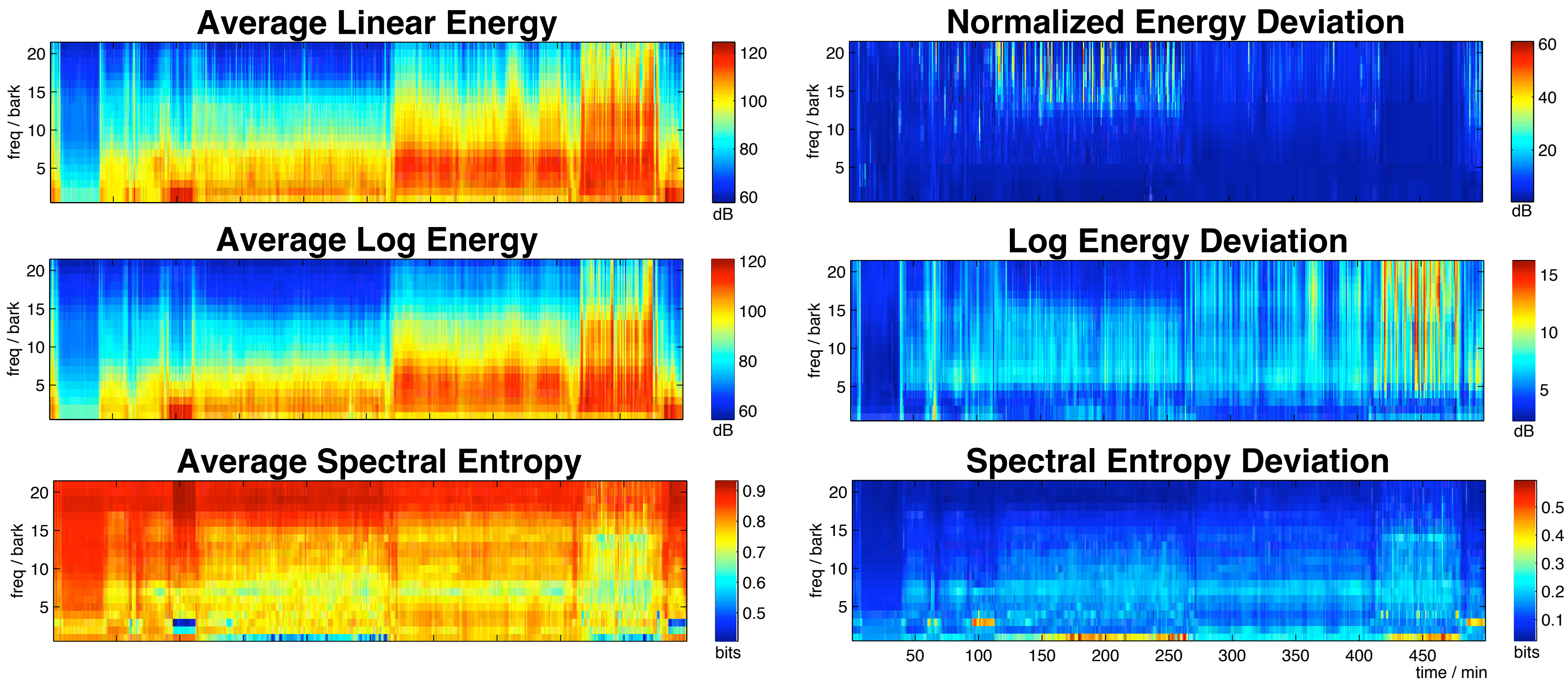
- Starting point: Collect data
  - 62 hours recorded (8 days, ~7.5 hr/day)
  - hand-mark 139 segments (26 min/seg avg.)
  - assign to 16 classes (11 have multiple instances)

Label	*total mins*	*total segs*
Library	981	27
Campus	750	56
Restaurant	560	5
Bowling	244	2
Lecture 1	234	4
Car/Taxi	165	7
Street	162	16

2. Features

- Long duration recordings
  may benefit from longer basic time-frames
  - 60s rather than 10ms?
- Perceptually-motivated features
  - broad spectrum + some detail?
- For diary application...
  - background more important than foreground?
  - smooth out uncharacteristic transients

Feature sets

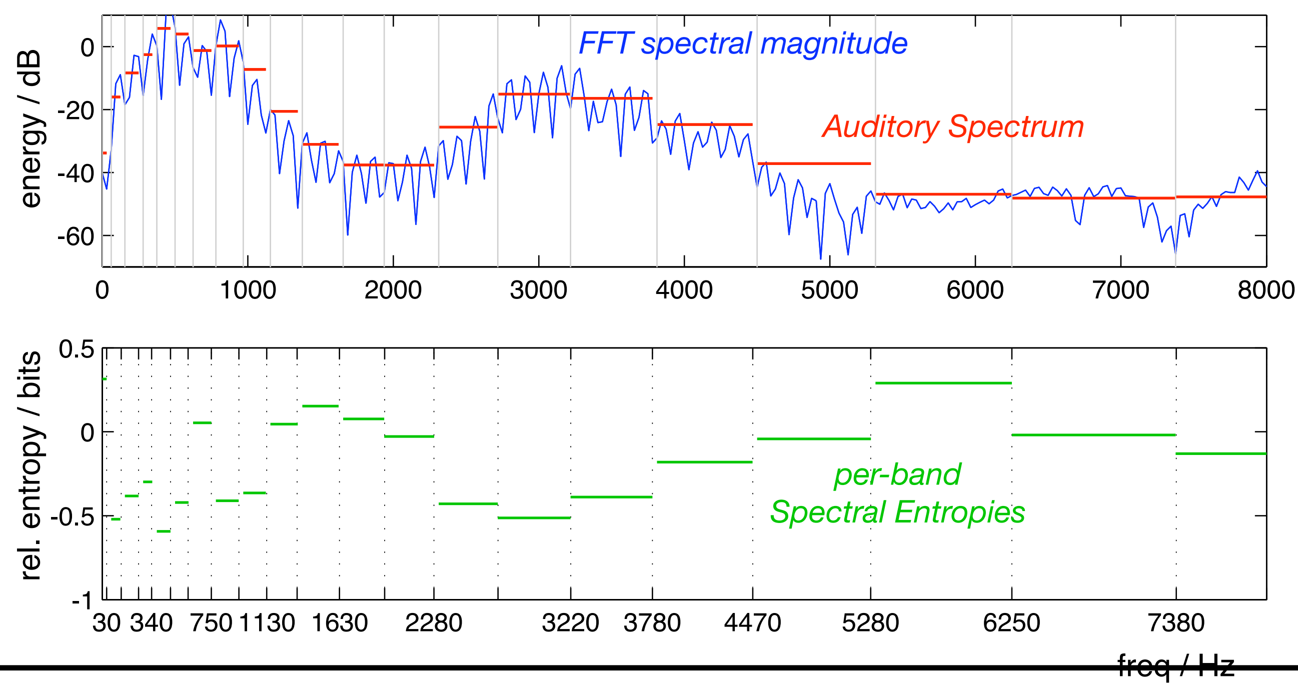


- Capture both average and variation
- Capture a little more detail in subbands...

Spectral Entropy

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

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

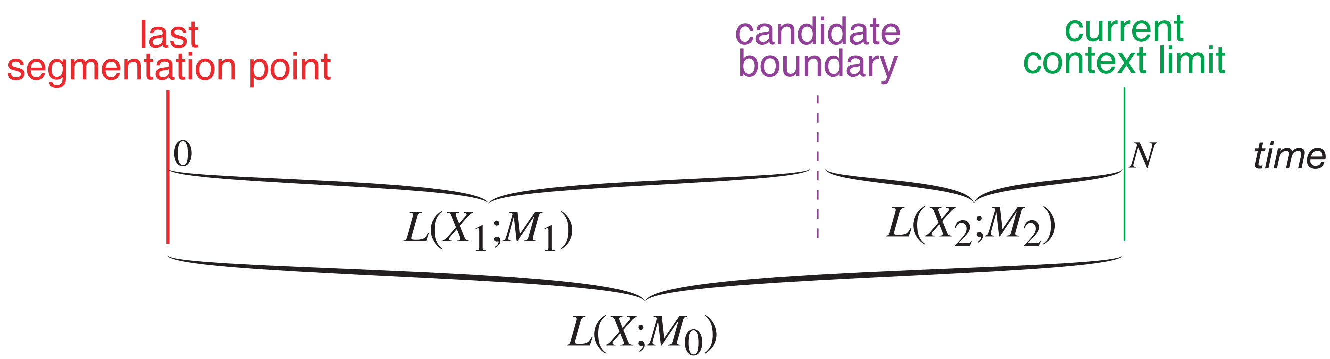


3. BIC segmentation

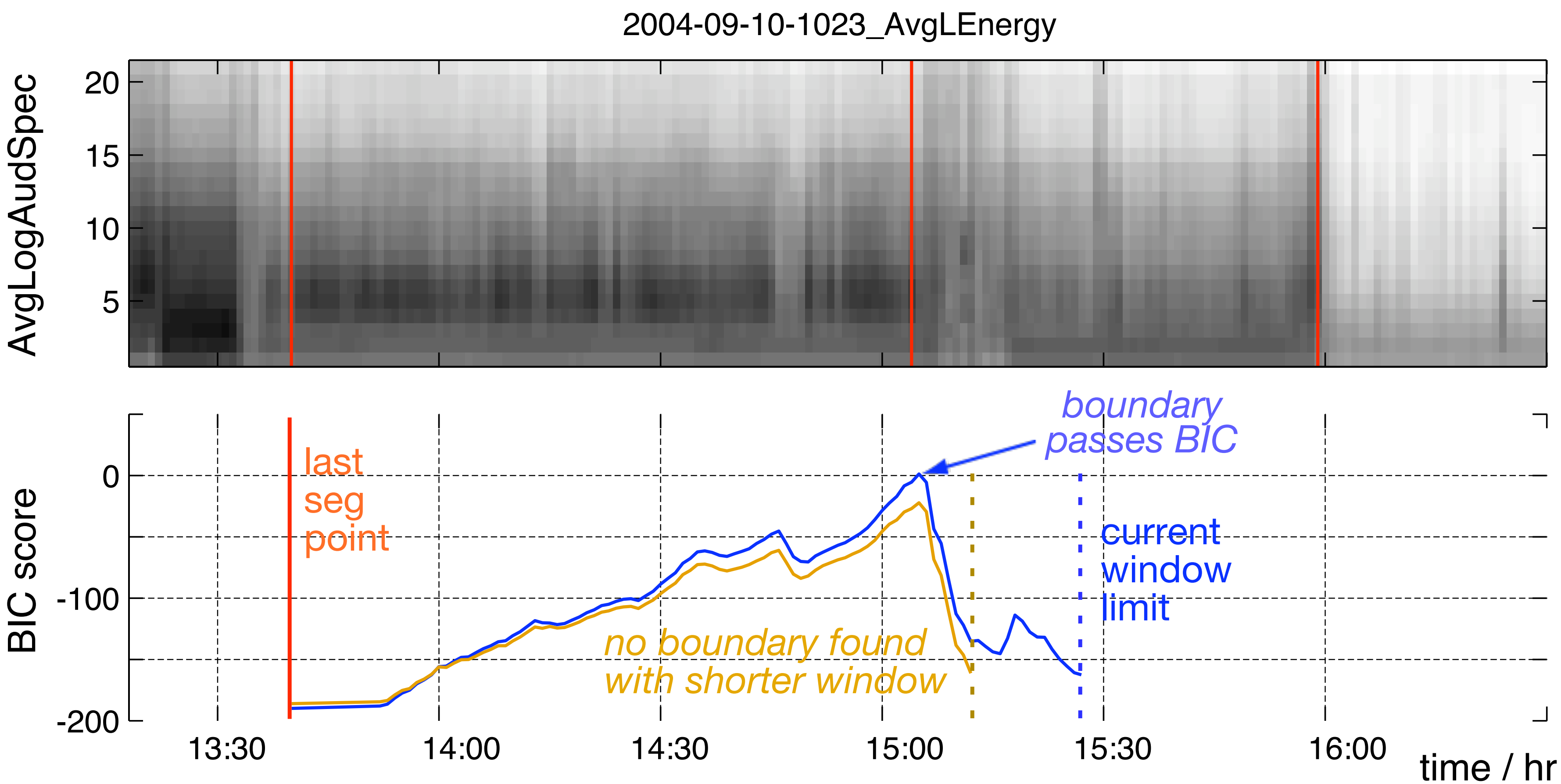
- BIC (Bayesian Information Criterion): Compare more and less complex models

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

- For segmentation:
  - Grow context window from current boundary
  - For each window, test every possible segmentation
  - When BIC is positive, mark new segment



BIC Segmentation Example

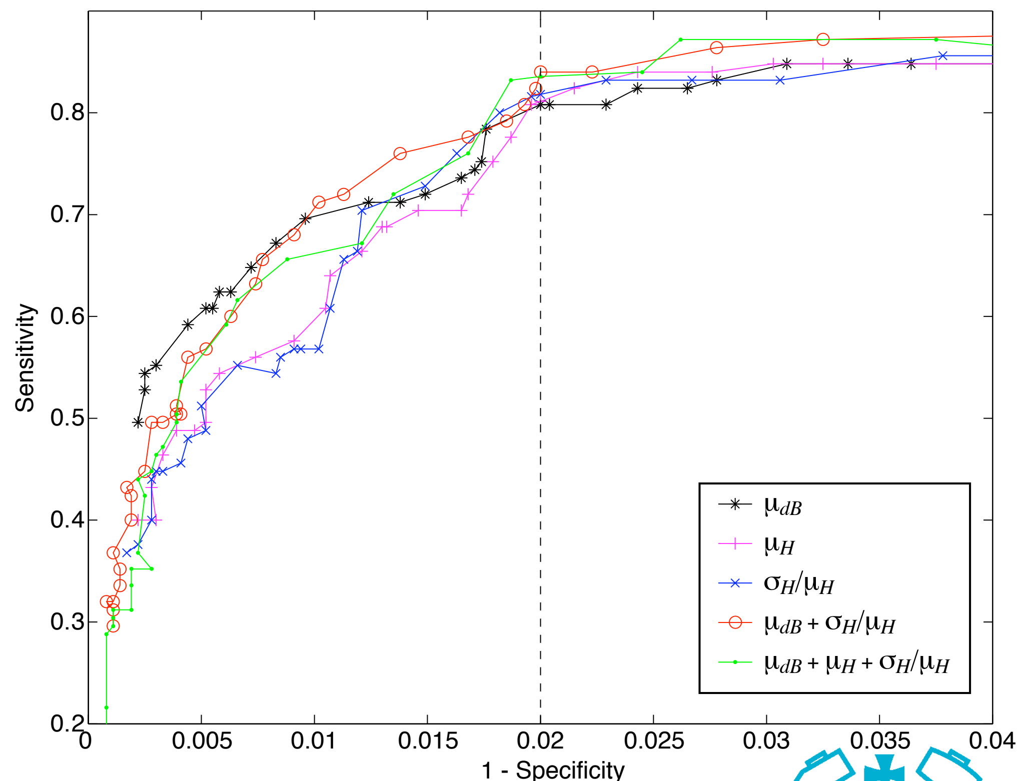


- No training or stored models

Segmentation Results

- **Evaluate: 60hr hand-marked boundaries**
  - different features & combinations
  - 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%
avg. mfcc	73.6%

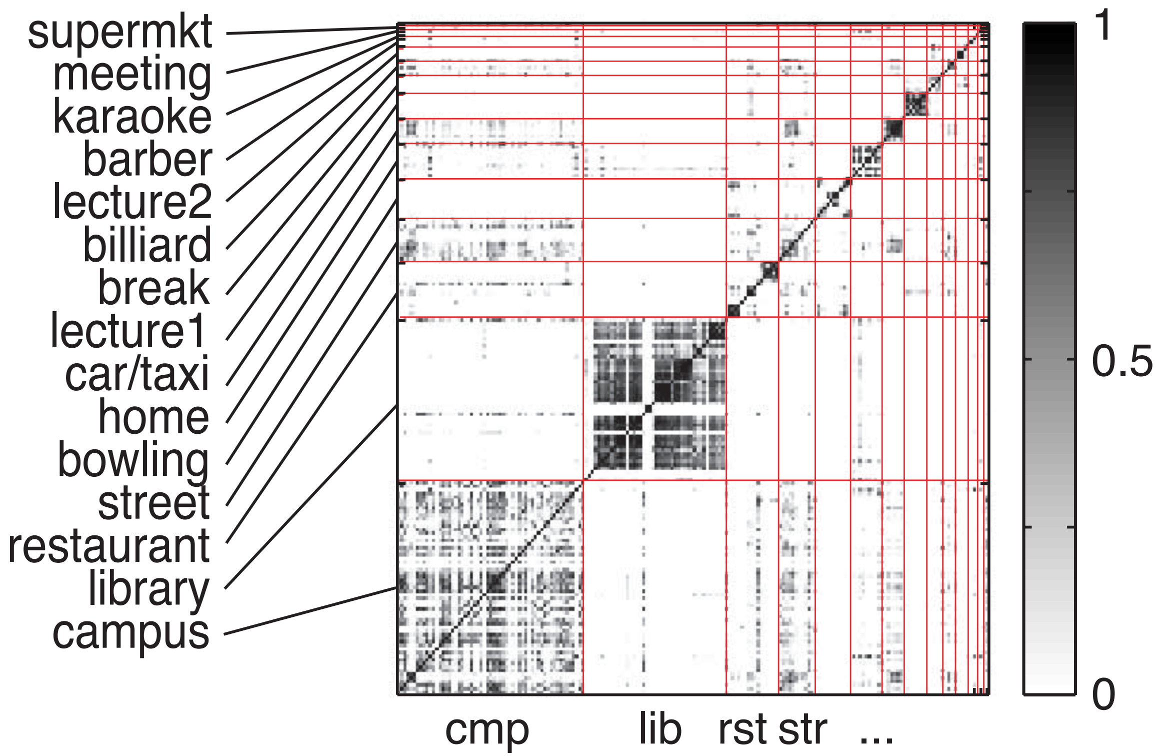


Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

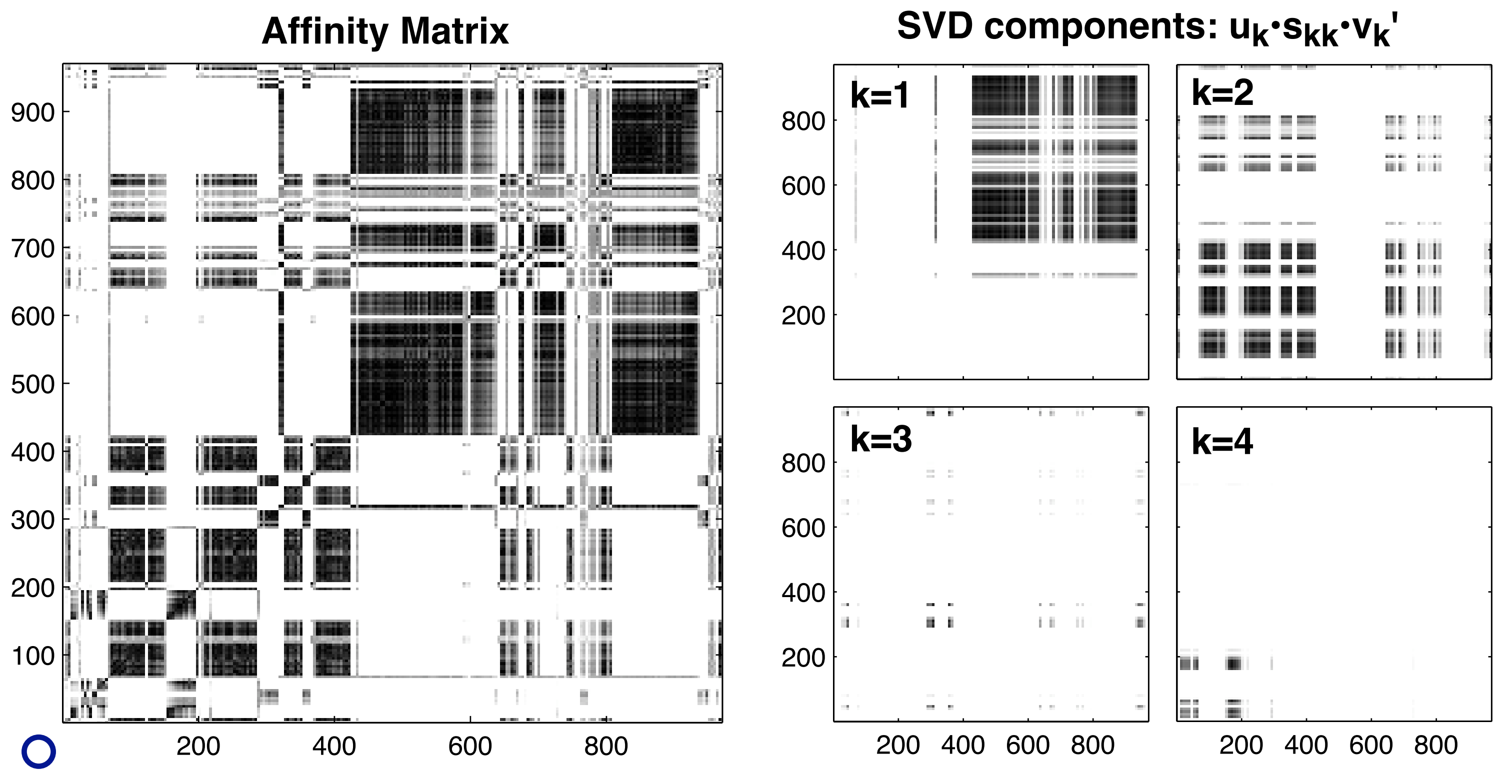
4. Segment clustering

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



Spectral Clustering

- Eigenanalysis of affinity matrix: A = U•S•V'

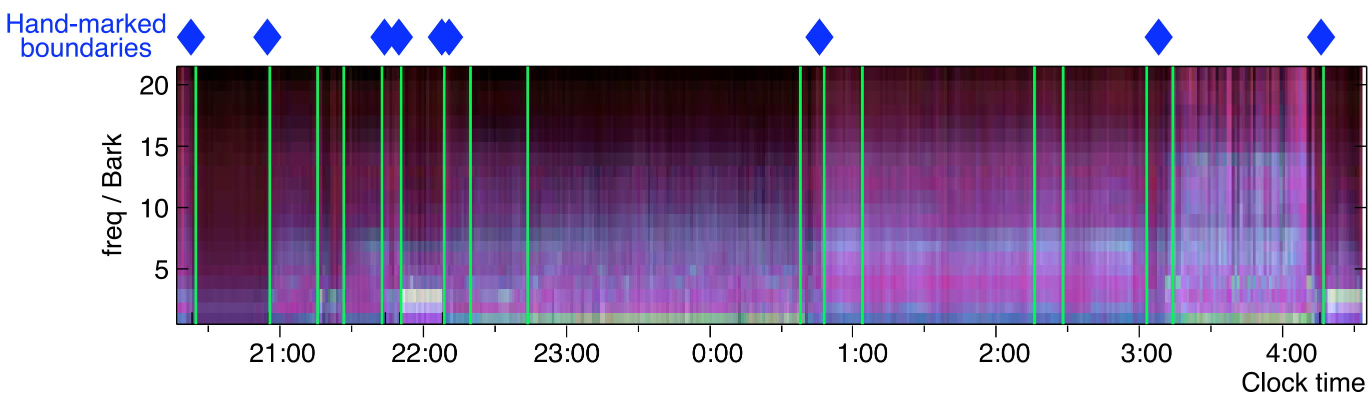


  - eigenvectors v_k give cluster memberships

- Number of clusters?

Clustering Results

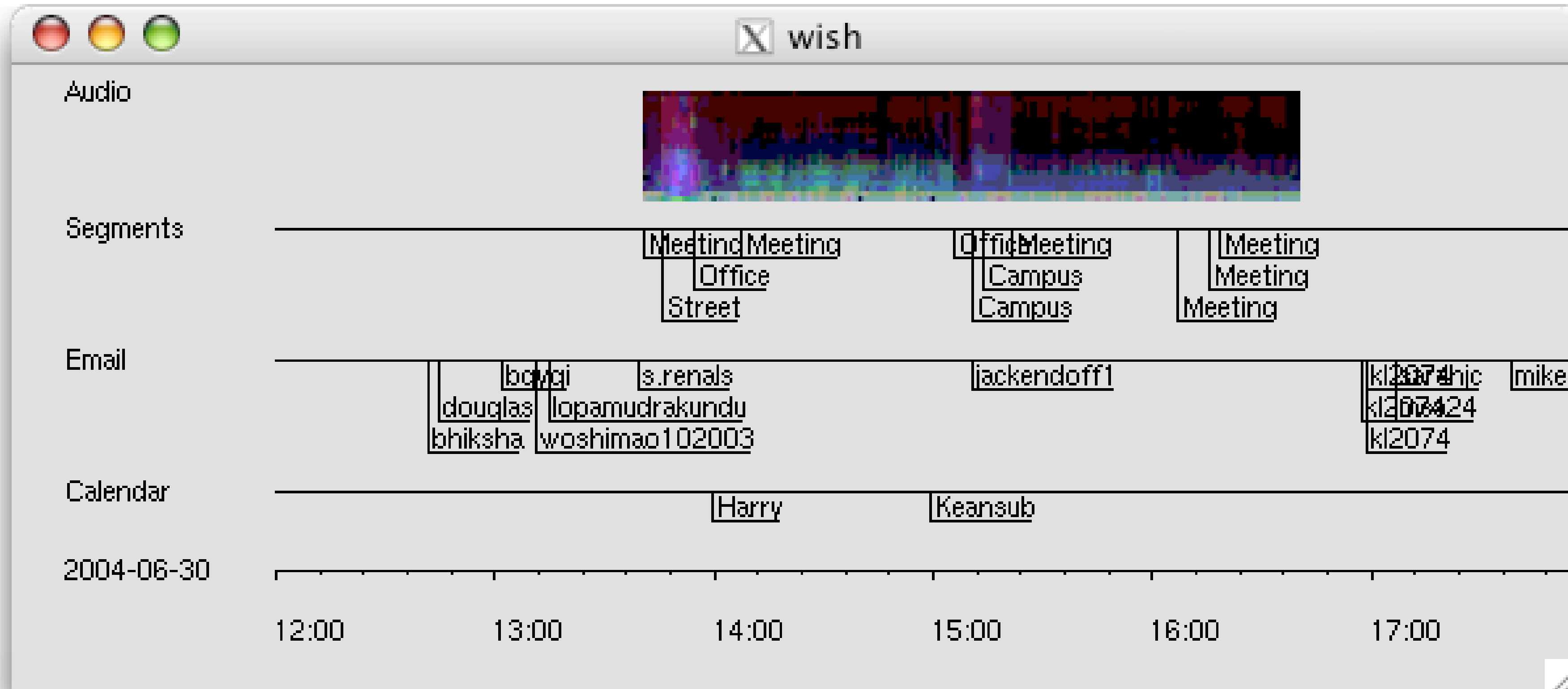
- **Clustering of automatic segments gives 'anonymous classes'**
  - BIC criterion to choose number of clusters
  - make best correspondence to 16 GT clusters



- **Frame-level scoring gives ~70% correct**
  - errors when same 'place' has multiple ambiences
  - clusters formed by strong foregrounds (voices)

5. Future Work

- Visualization / browsing / diary inference
  - link in other information sources



Privacy

- Recording conversations conflicts with expectations of privacy
  - critical barrier to progress
- Technical solutions to improve acceptance? Speaker/speech "search and destroy"
  - scramble 100ms segs of speech (preserving longer-term statistics)
  - high-confidence speaker ID to bypass

Conclusions

- "Personal Audio" is easy & cheap to collect
  - but is it any use?
- Boundaries quite easy to spot
  - moving to a new location
  - change in activity (talking <> reading)
- Repeated activities can cluster together
  - .. so user's labels can propagate
- Still gaining experience with the data
  - speech is the most interesting part
    - .. but very hard to transcribe
  - speaker ID, privacy, ...