# Multimedia Applications of Audio Recognition

Dan Ellis
LabROSA
Columbia EE

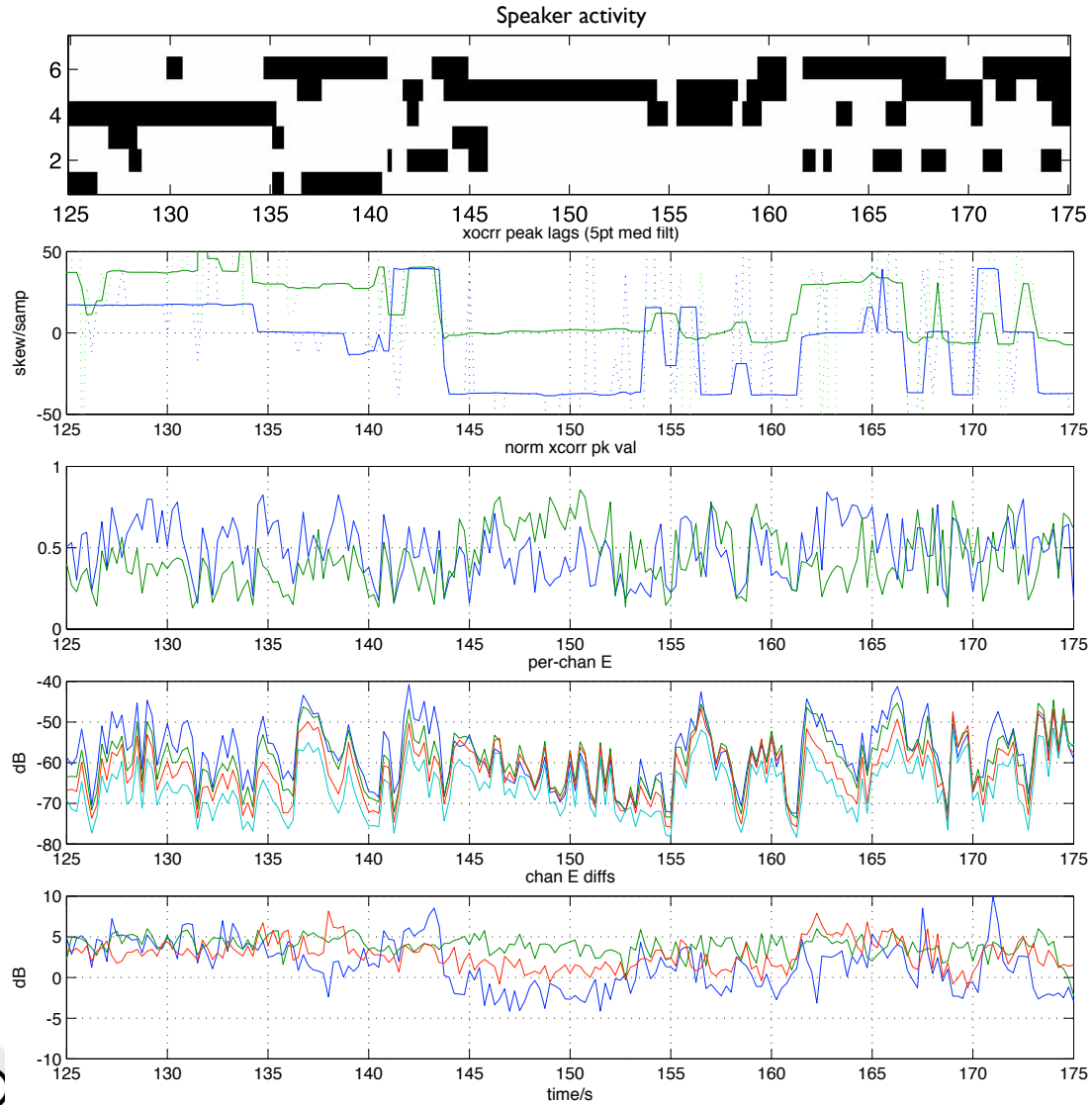
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Finding speaker turns in meetings
2. Segmenting 'personal audio' recordings
3. Eigenrhythms: representing drum tracks

# 1. Finding Speaker Turns in Meetings

- Multiple mic recordings carry information on speaker turns
  - every voice reaches every mic... (?)
  - ... but with differing coupling filters (delays, gains)
- Find turns with minimal assumptions
  - e.g. ad-hoc sensor setups (multiple PDAs)
  - differences to remove effect of source signal
    - no spectral models, < 1xRT

# Between-Channel Cues: Timing (ITD) & Level




Speaker ground-truth

Timing diffs (ITD) (2 mic pairs, 250ms win)

Peak correlation coefficient *r*

Per-channel energy

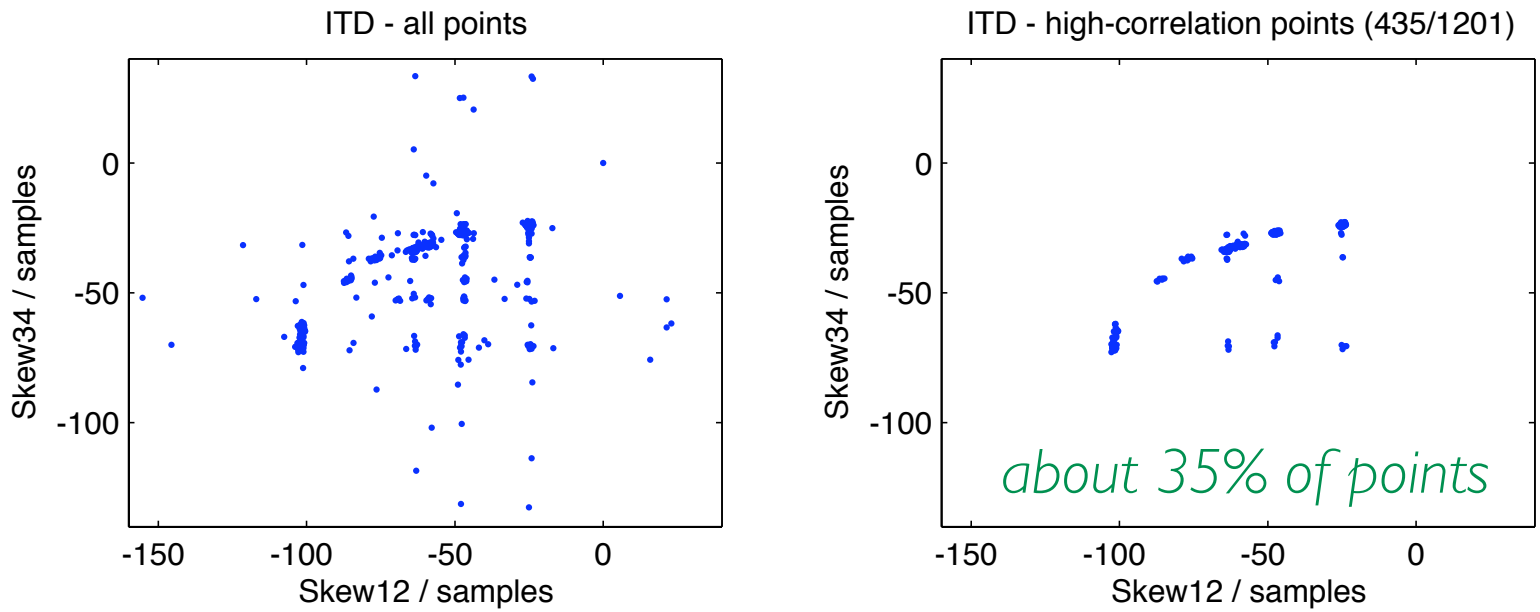
Between-channel energy differences

# Choosing "Good" Frames

- Correlation coef. $r$ ~ channel similarity:

$$r_{ij}[\ell] = \frac{\sum_n m_i[n] \cdot m_j[n+\ell]}{\sqrt{\sum m_i^2 \sum m_j^2}}$$

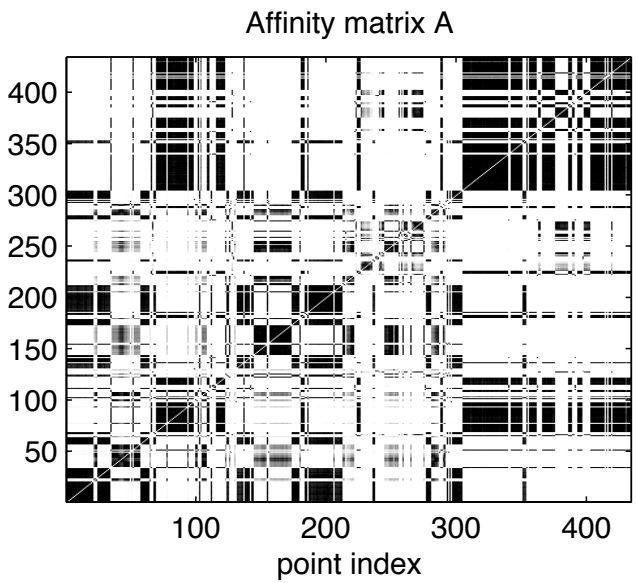
- Select frames with $r$ in top 50% in both pairs

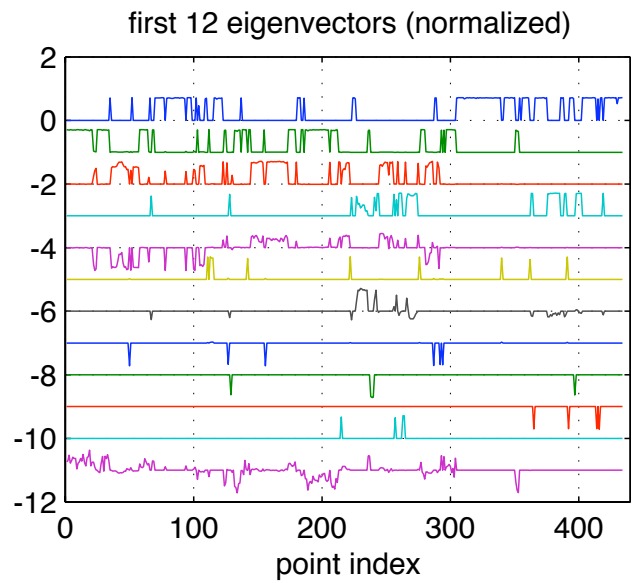



- Cleaner basis for models

# Spectral Clustering

- Eigenvectors of "affinity matrix" **A** to pick out similar points:



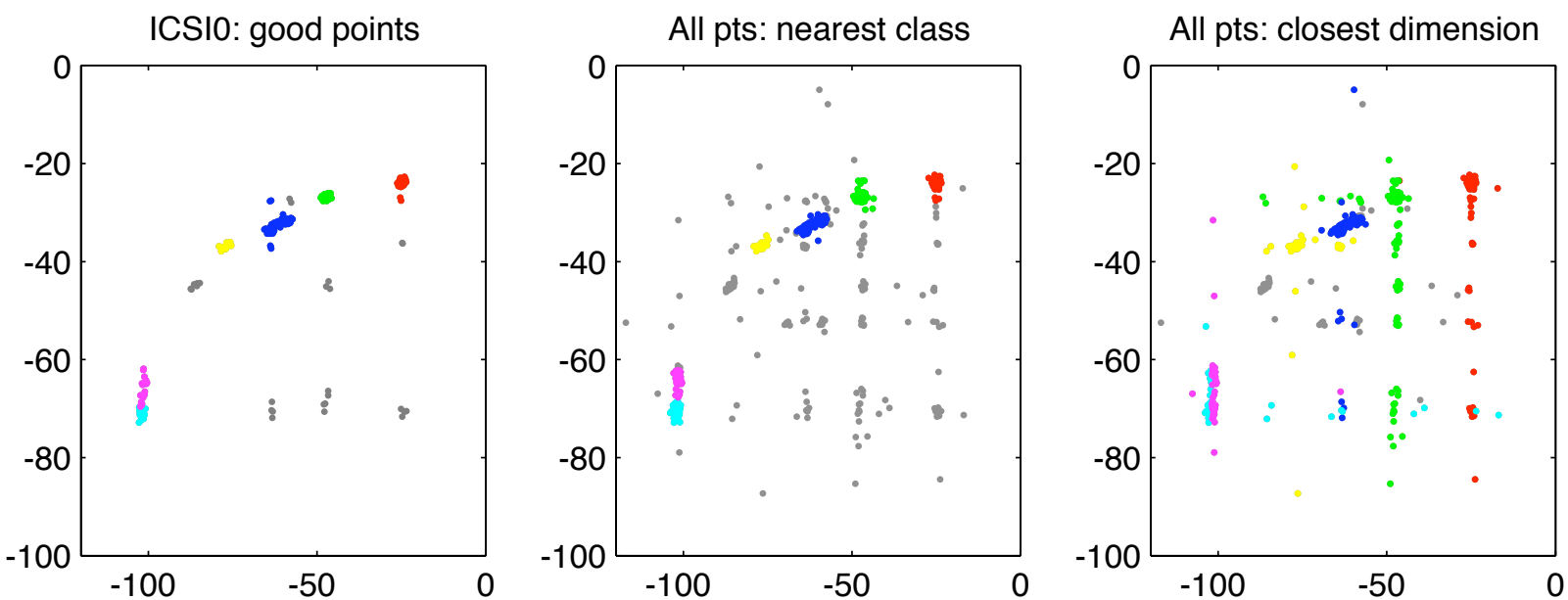




$$a_{mn} = exp\{-\|\mathbf{x}[m] - \mathbf{x}[n]\|^2 / 2\sigma^2\}$$

- Ad-hoc mapping to clusters
  - Number of clusters $K$ from eigenvalues ≈ points

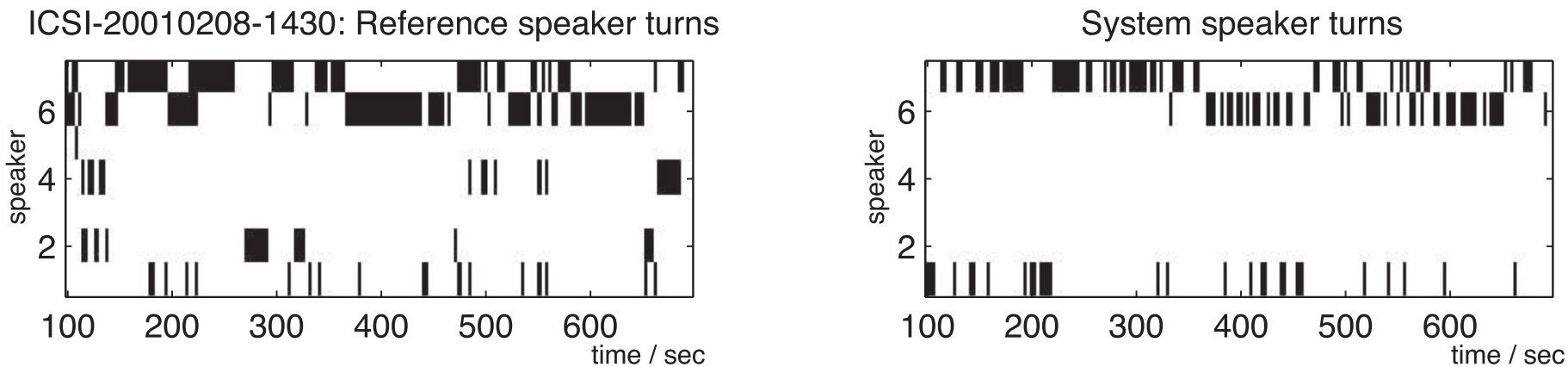
# Speaker Models & Classification

- Actual clusters depend on $\sigma$ and $K$ heuristic
- Fit Gaussians to each cluster, assign that class to all frames within radius
  - or: consider dimension independently, choose best



# Performance Analysis

- Compare reference & system activity maps:



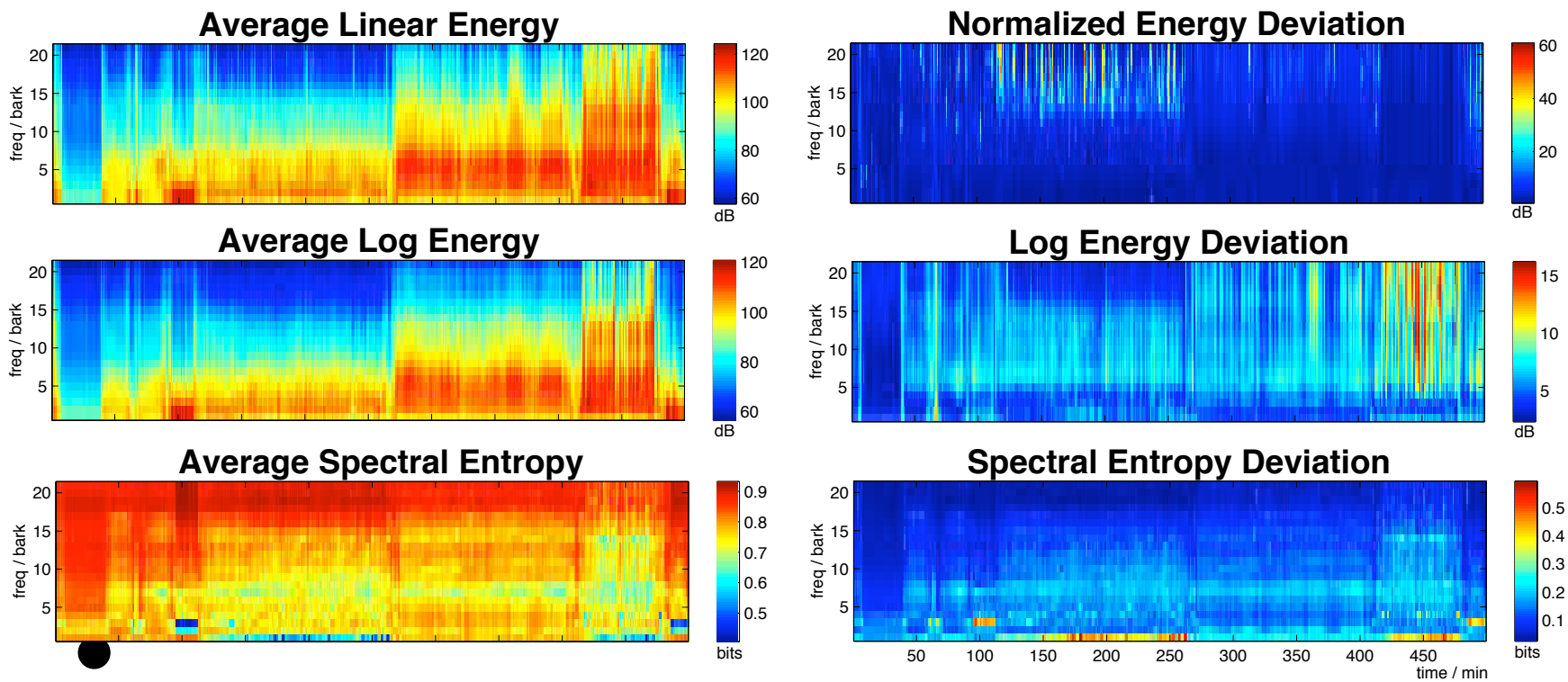

  - system misses quiet speakers 2,3,4 (deletions)
  - system splits speaker 6 (deletions+insertions)
  - many short gaps (deletions)
- **NIST RT04s dev set (80 min):**
  - correct: 34.4% ins: 16.5% del: 63.7%

# 2. Segmenting Personal Audio


- Easy to record everything you hear
  - ~100GB / year @ 64 kbps
- Very hard to find anything
  - how to scan?
  - how to visualize?
  - how to index?
- Starting point: Collect data
  - ~ 60 hours (8 days, ~7.5 hr/day)
  - hand-mark 139 segments (26 min/seg avg.)
  - assign to 41 classes (8 have multiple instances)

# Features for Long Recordings

- Feature frames = 1 min (not 25 ms!)
- Characterize variation within each frame...




- and structure within coarse auditory bands

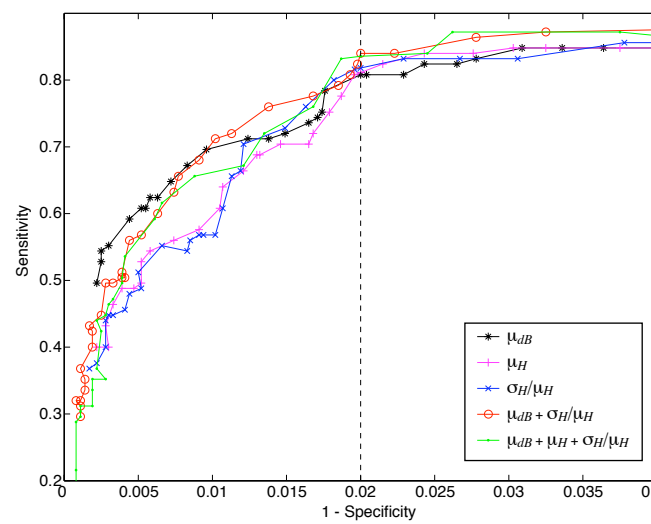
# BIC Segmentation

- Untrained segmentation technique
  - statistical test indicates good change points:

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

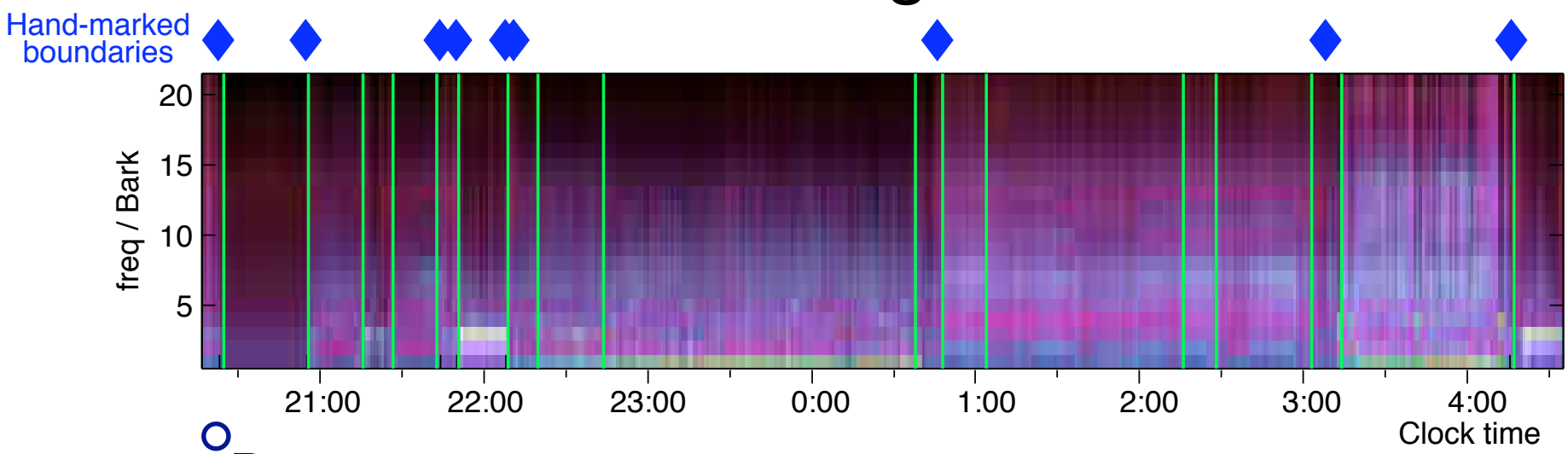
- Evaluate against hand-marked boundaries
  - different features & combinations
  - Correct Accept % @ False Accept = 2%:

| | |
|---|---|
| $\mu_{dB}$ | 80.8% |
| $\mu_H$ | 81.1% |
| $\sigma_H/\mu_H$ | 81.6% |
| $\mu_{dB} + \sigma_H/\mu_H$ | 84.0% |
| $\mu_{dB} + \sigma_H/\mu_H + \mu_H$ | 83.6% |



# Future Work

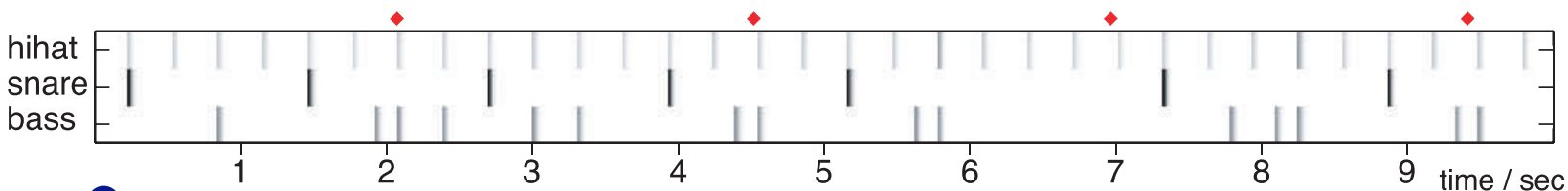
- Clustering of segments
  - .. using spectral clustering
  - compare to ground-truth, confusions
- Visualization / browsing:




- Privacy protection
  - speaker/speech "search and destroy"

# 3. Eigenrhythms: Representing Drum Tracks

- Pop songs built on repeating "drum loop"
  - bass drum, snare, hi-hat
  - small variations on a few basic patterns



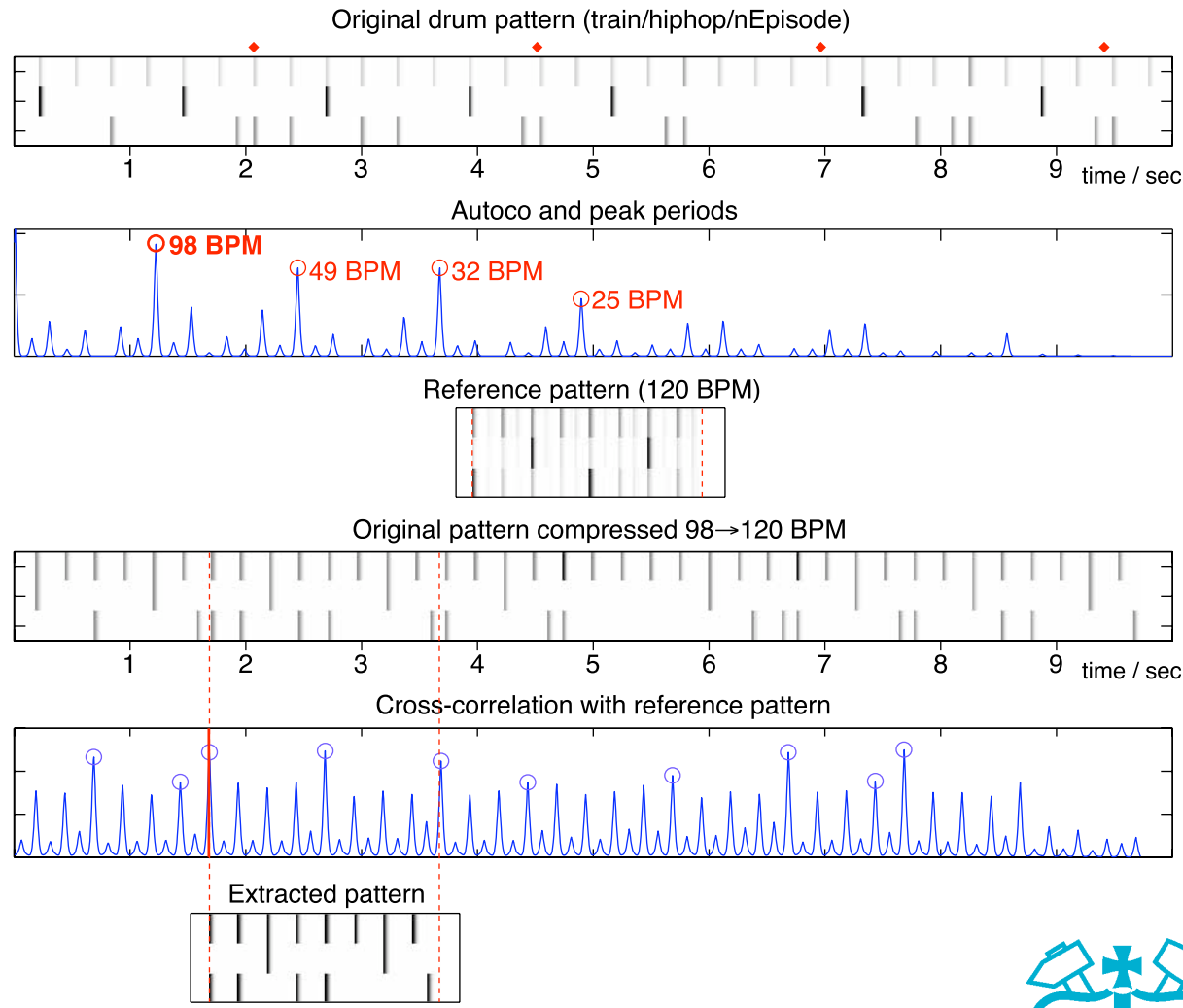

- Eigen-analysis (PCA) to capture variations?
  - by analyzing lots of (MIDI) data
- Applications
  - music categorization
  - "beat box" synthesis

# Aligning the Data

- Need to align patterns prior to PCA...

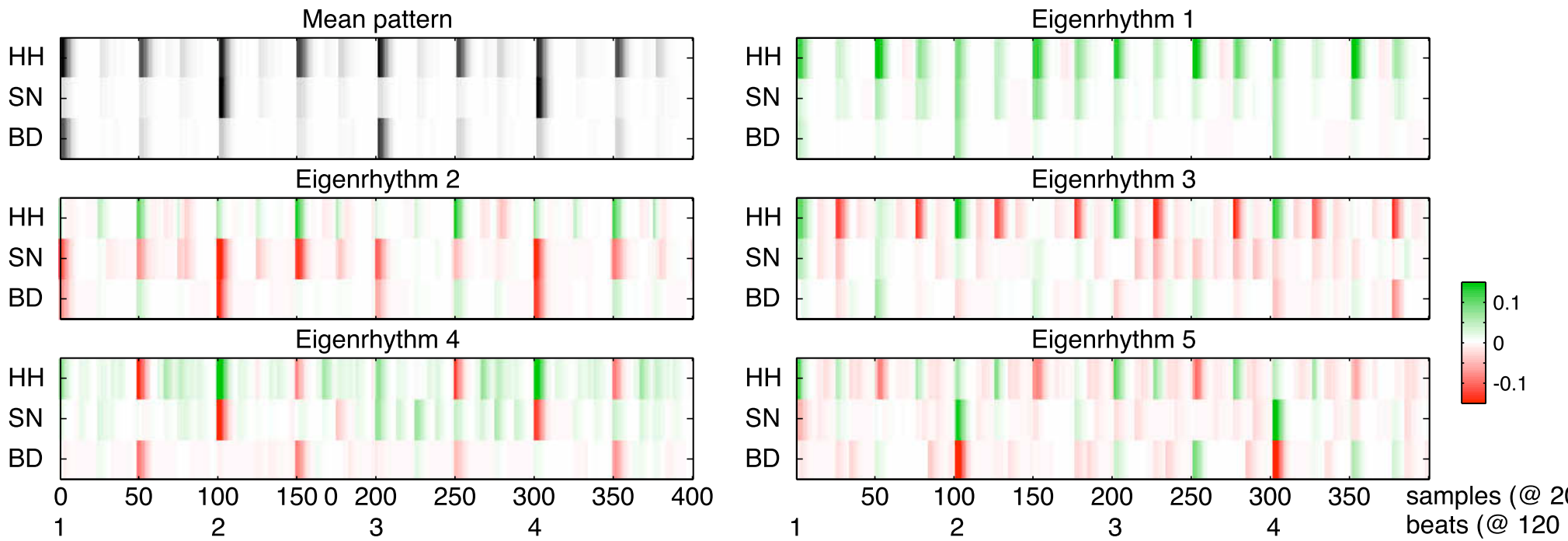
**tempo** (stretch): by inferring BPM & normalizing

**downbeat** (shift): correlate against 'mean' template



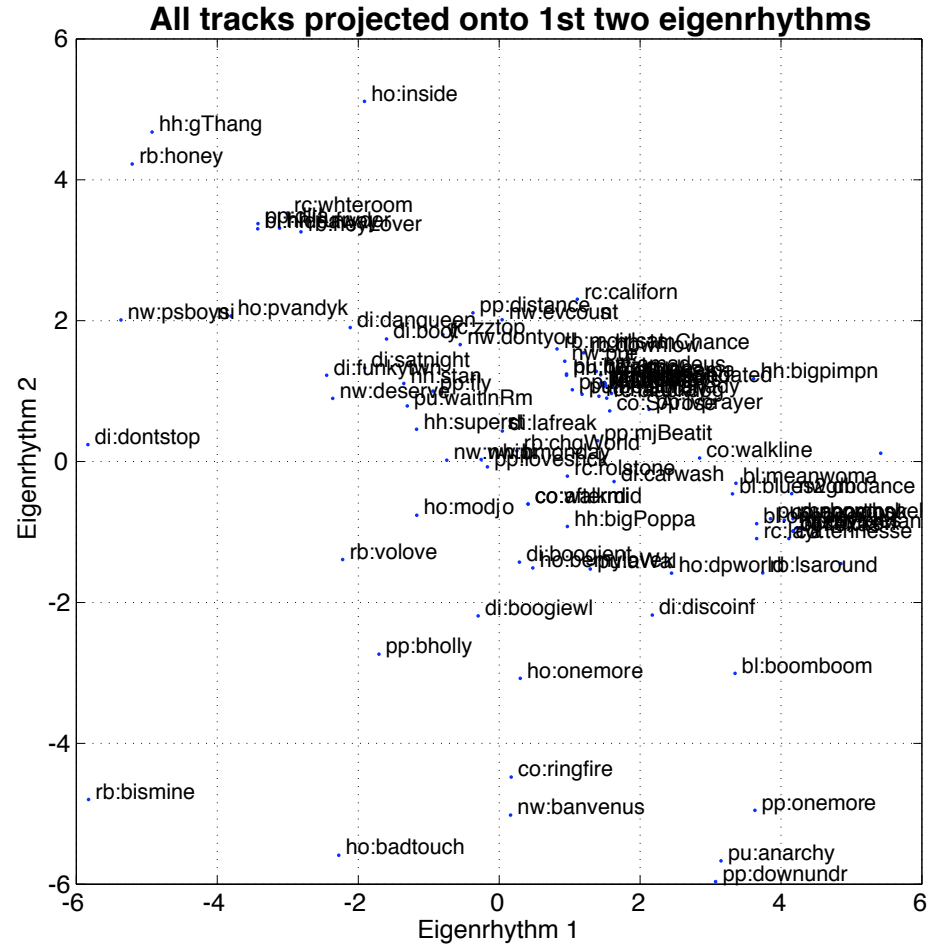
# Eigenrhythms

- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Top patterns:



# Eigenrhythms for Classification

- Clusters in Eigenspace:




- Genre classification? (10 way)
  - nearest neighbor in 4D eigenspace: 21% correct

# Summary: Audio Analysis for Multimedia

- Spatial cues for meeting recordings
  - recover meeting 'structure' from ad-hoc recordings
- Segmenting 'personal audio' recordings
  - what are good features for 1 minute frames?
- Eigenrhythm analysis of drum patterns
  - define the subspace of 'musical' rhythms