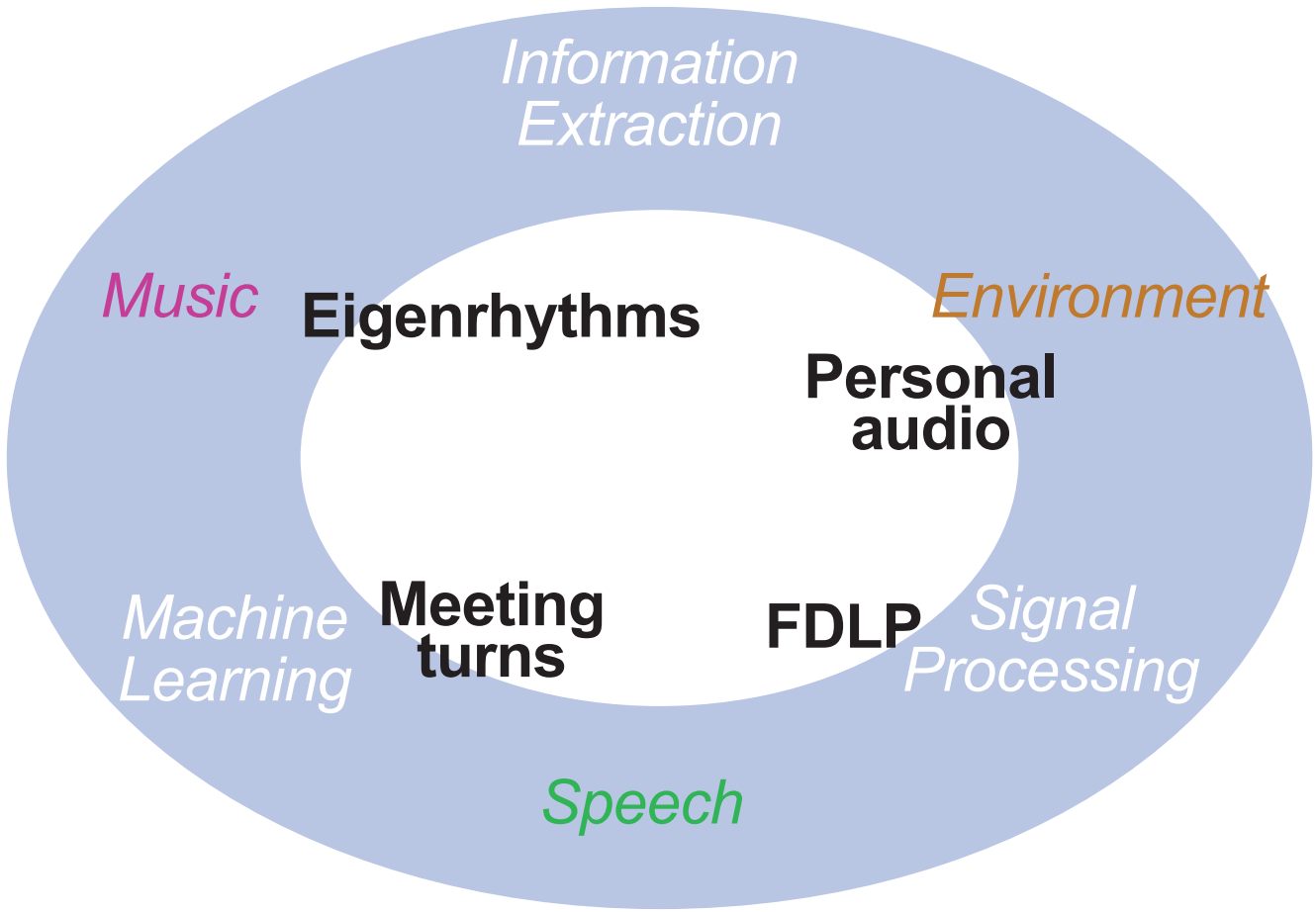
Audio & Music Research at LabROSA

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. Eigenrhythms: Representing drum tracks
2. Frequency-Domain Linear Prediction
3. Segmenting meeting turns
4. Analyzing 'personal audio' recordings

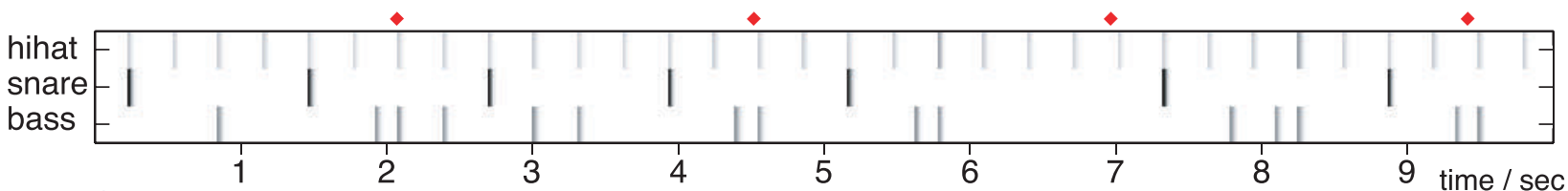
LabROSA Projects Overview



1. Eigenrhythms: Drum Pattern Space

with John Arroyo

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



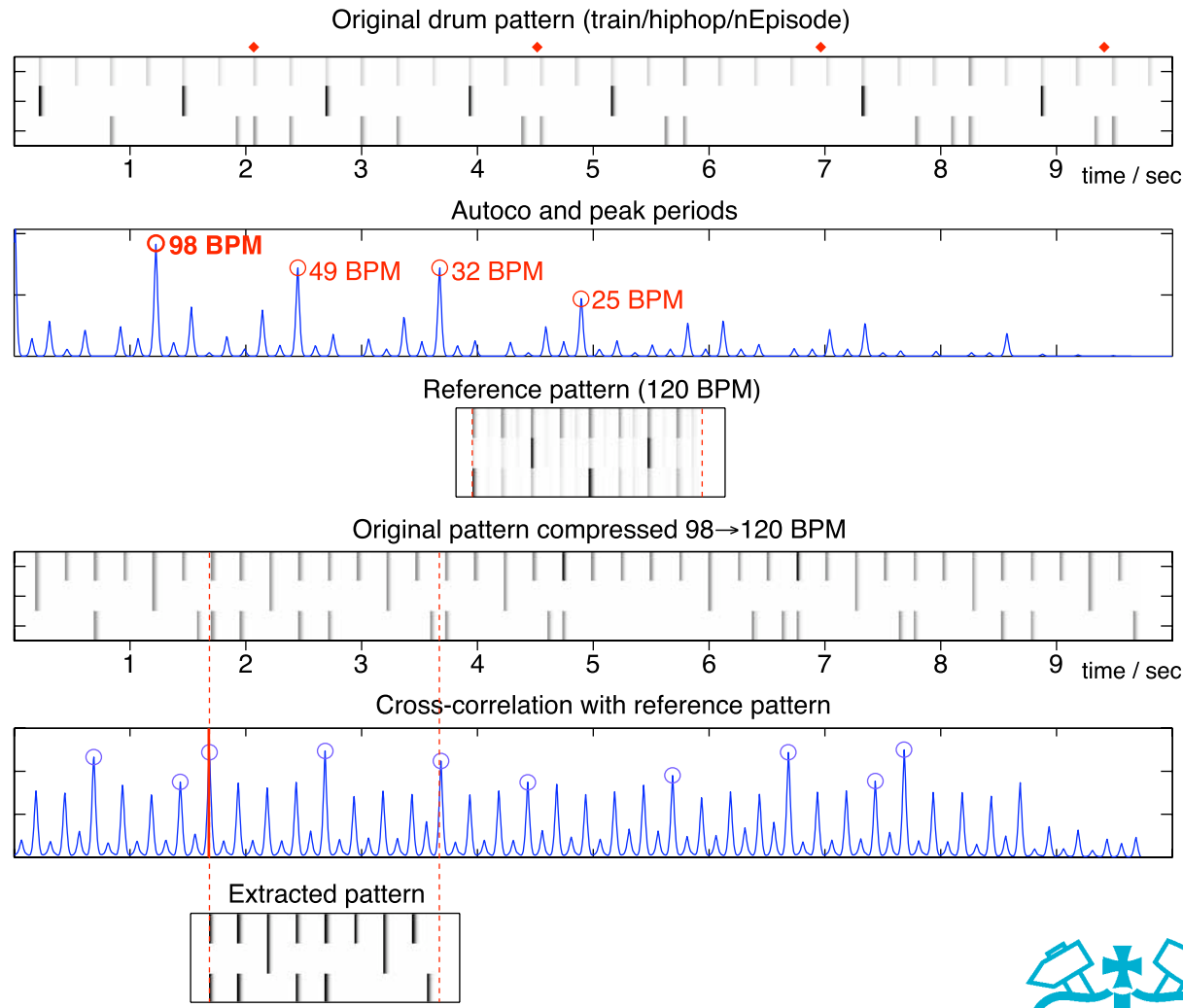
- Eigen-analysis (PCA) to capture variations?
  - by analyzing lots of (MIDI) data
- Applications
  - music categorization
  - "beat box" synthesis

Aligning the Data

- Need to align patterns prior to PCA...

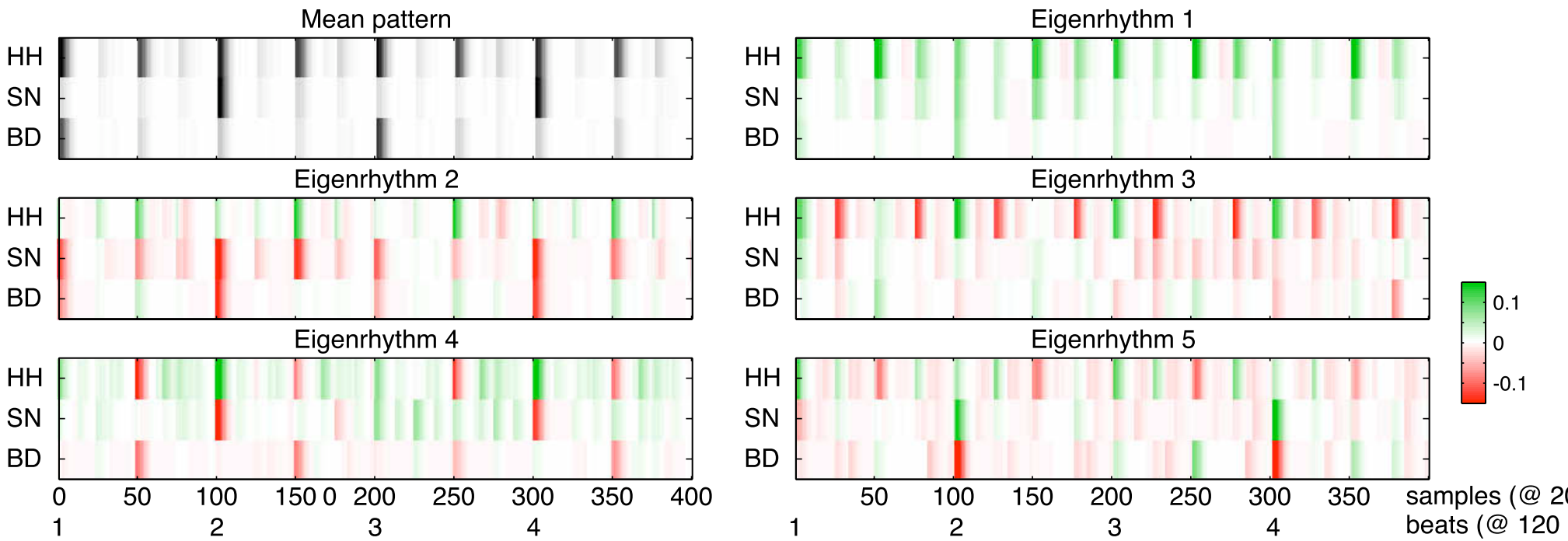
tempo (stretch): by inferring BPM & normalizing

downbeat (shift): correlate against 'mean' template



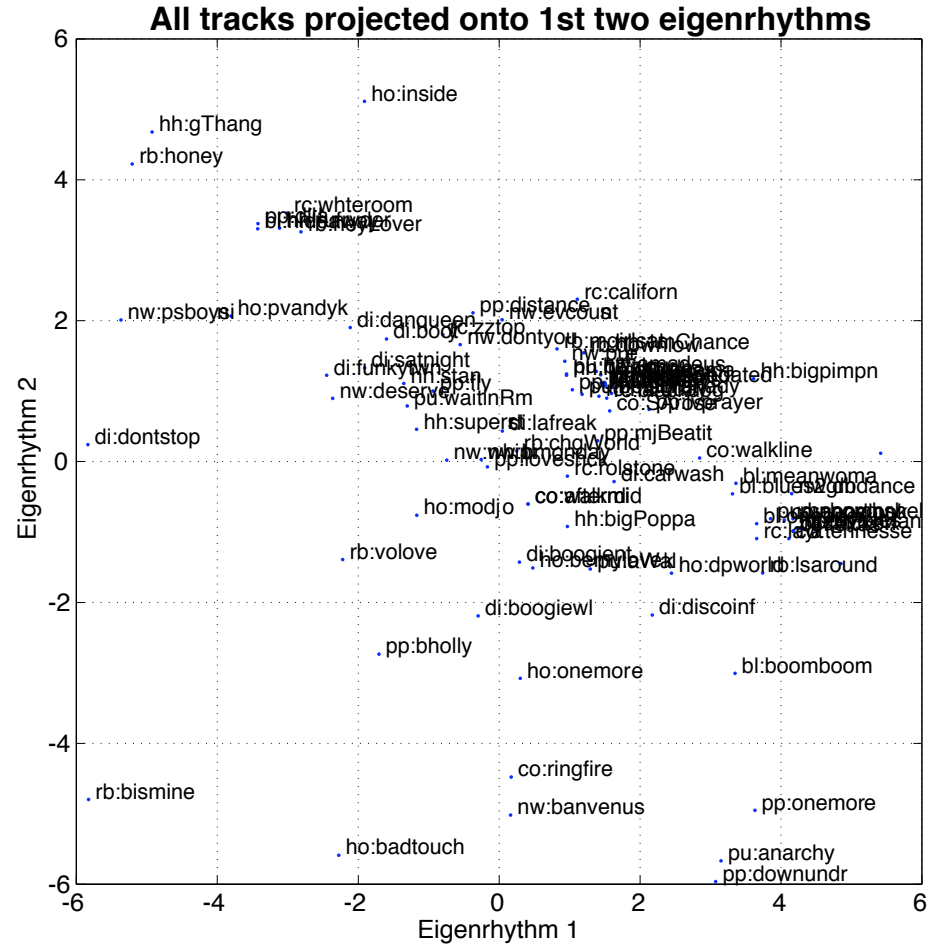
Eigenrhythms

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



Eigenrhythms for Classification

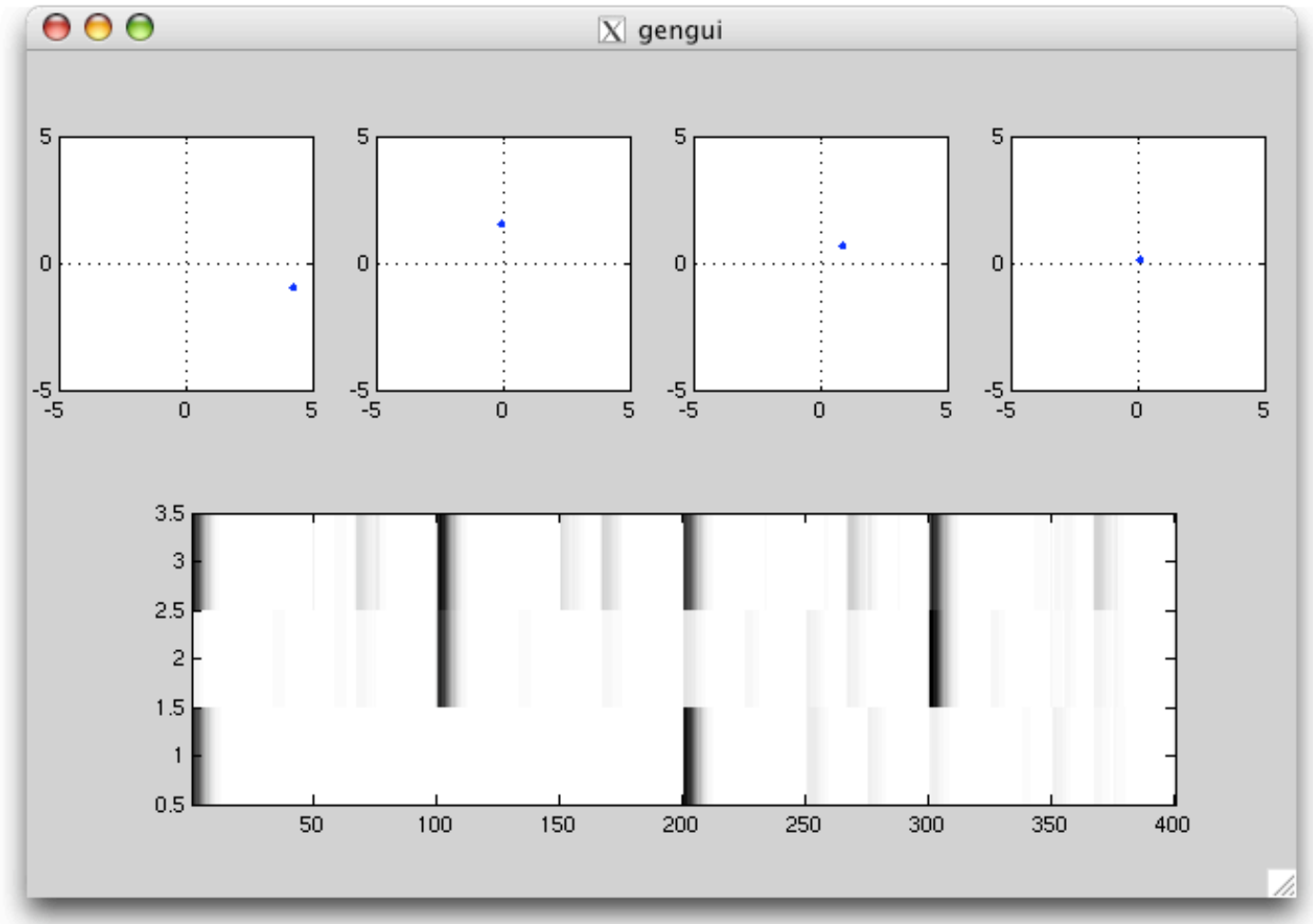
- Clusters in Eigenspace:



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

Eigenrhythm BeatBox

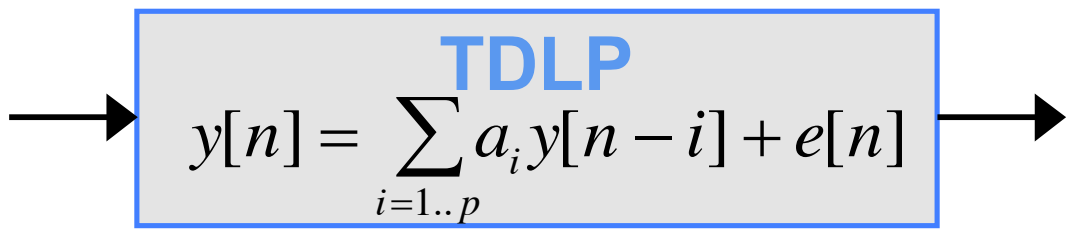
- Resynthesize rhythms from eigen-space



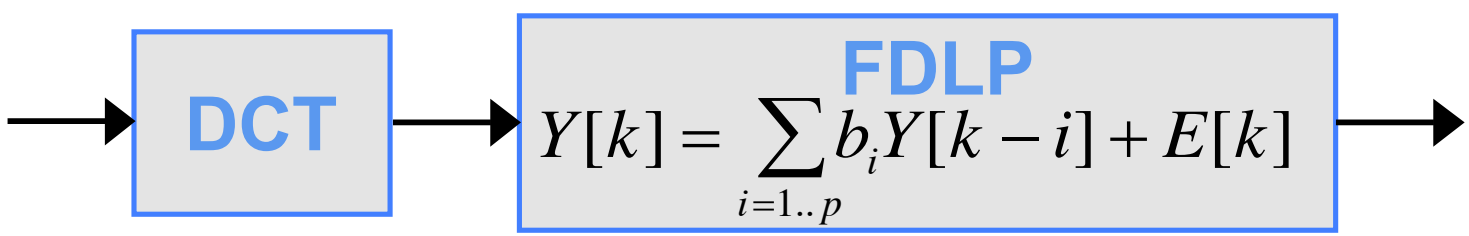
2. Frequency-Domain Lin. Pred.

with Marios Athineos

- (Time-domain) Linear Prediction
  - the well-known spectral estimator

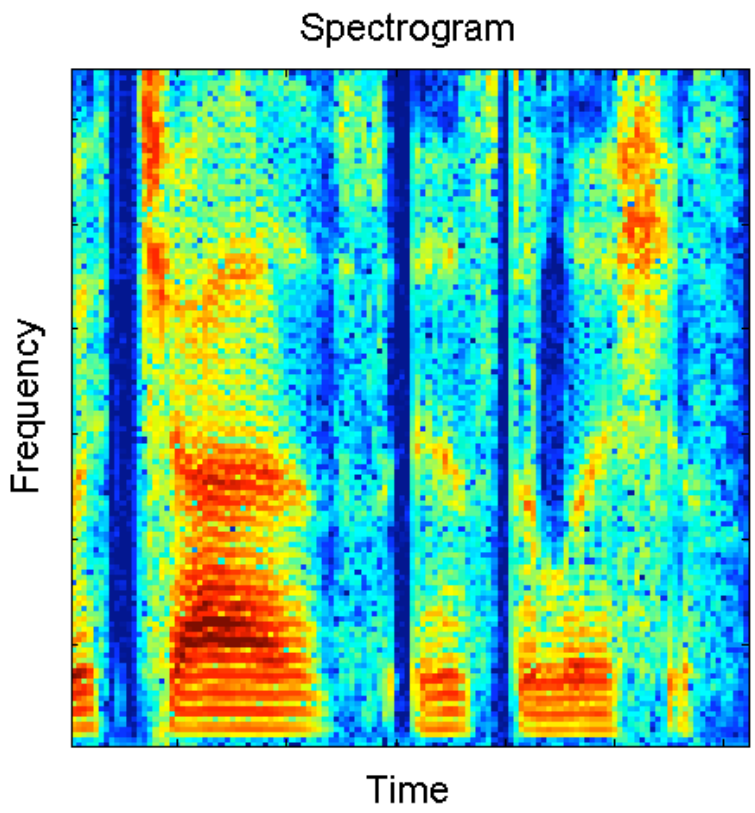


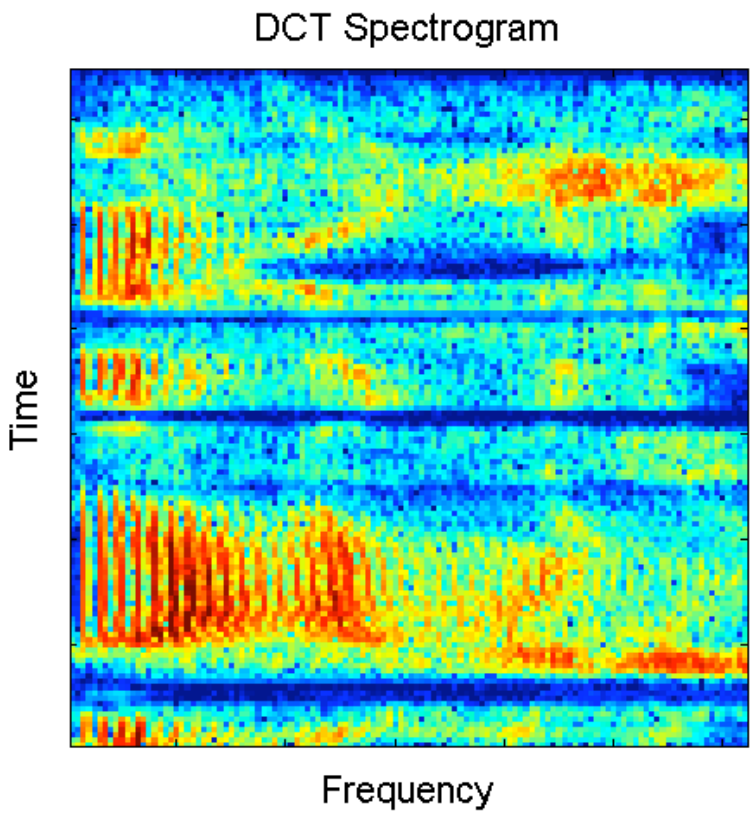
- Apply to a 'frequency domain' signal
  - dual: estimates temporal envelope



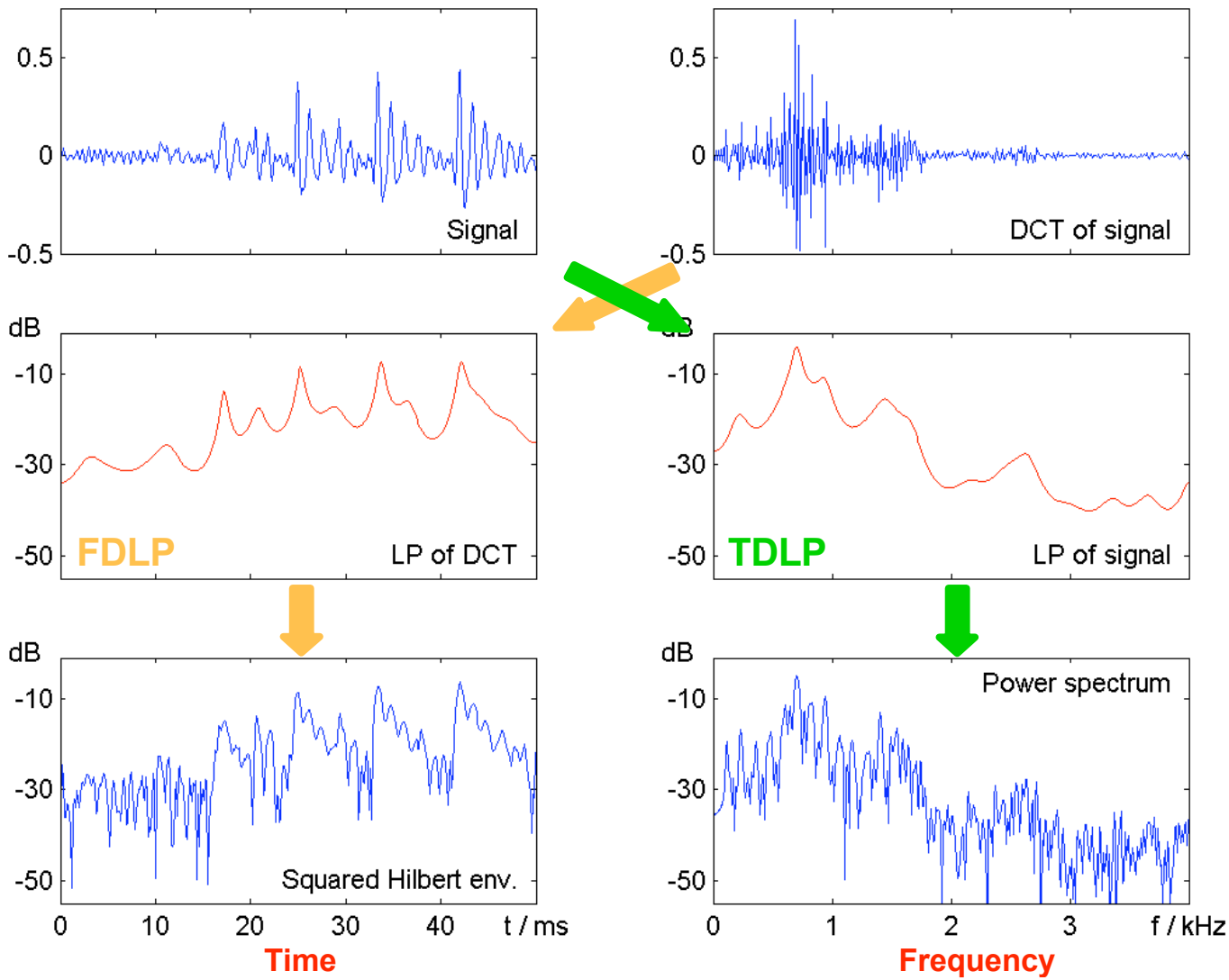
Aside: Spectrogram of the DCT

- DCT gives a pure-real signal:
  Can we treat it like a waveform?





FDLP and TDLP Duality



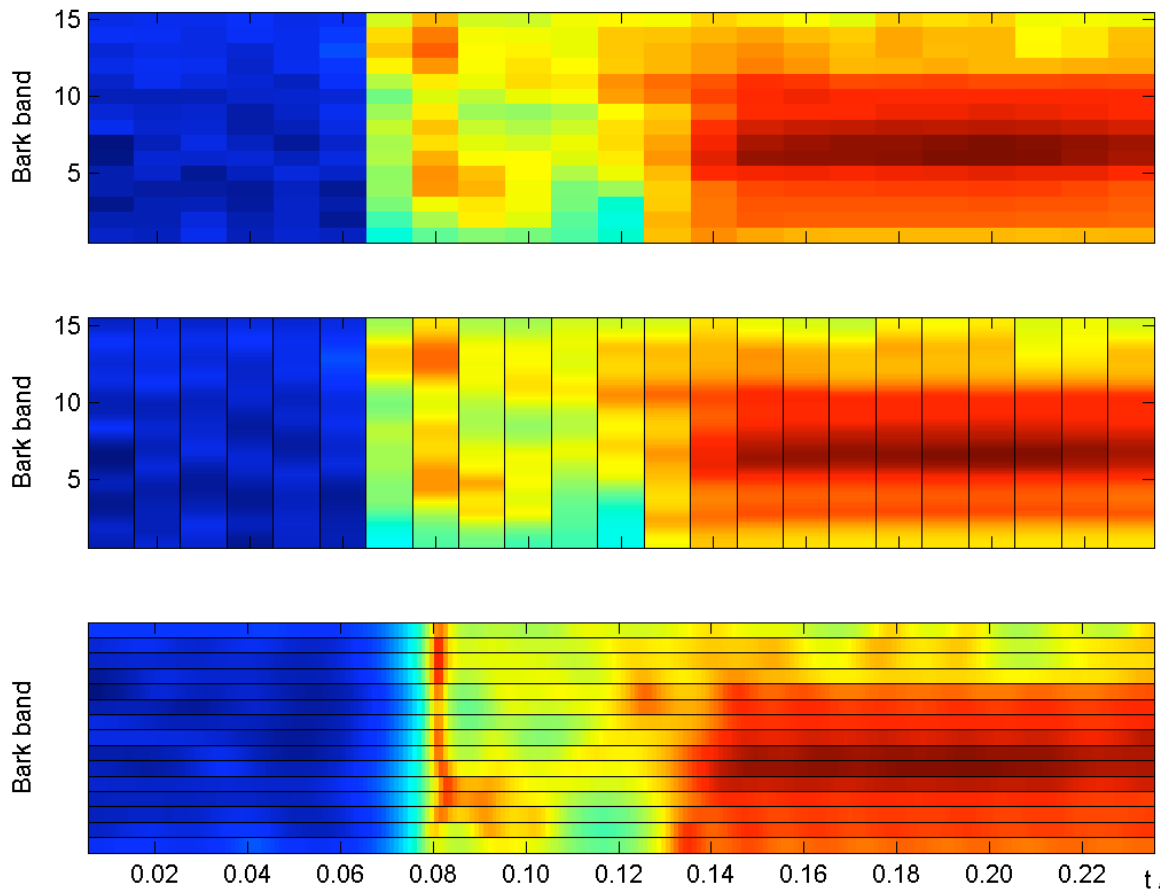
Subband FDLP

- Temporal envelopes without 25 ms windows

Auditory STFT
(10-25ms + Bark bin)

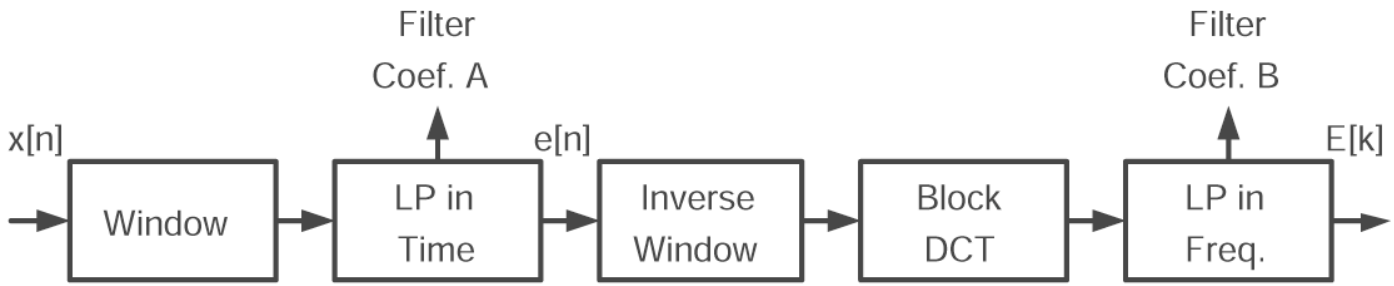
TDLP
(per time frame)

Subband FDLP
(per frequency subband)

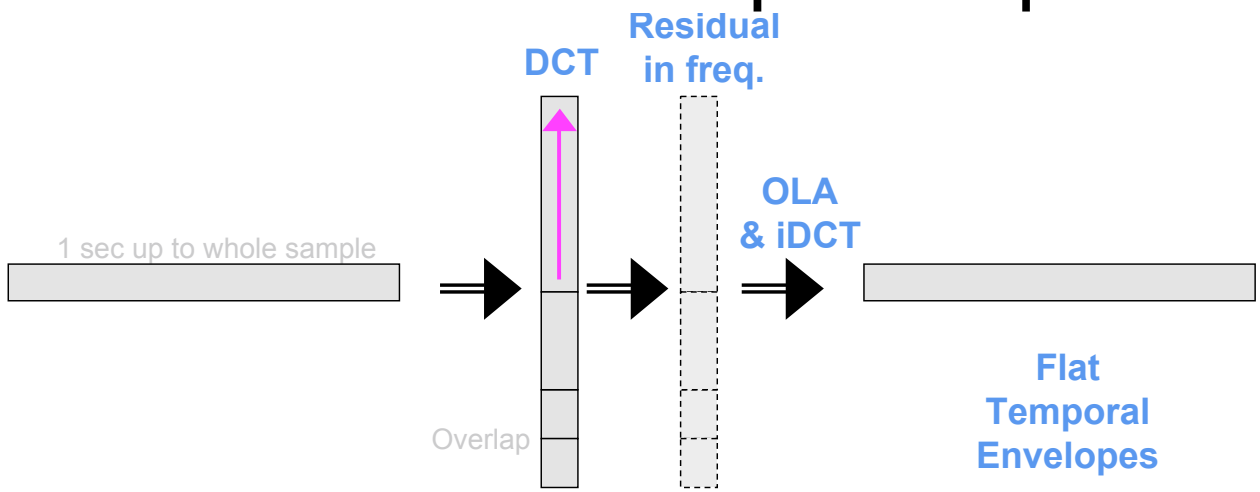


FDLP Applications

- Time-scale modification



- Modulation-domain "temporal equalization"

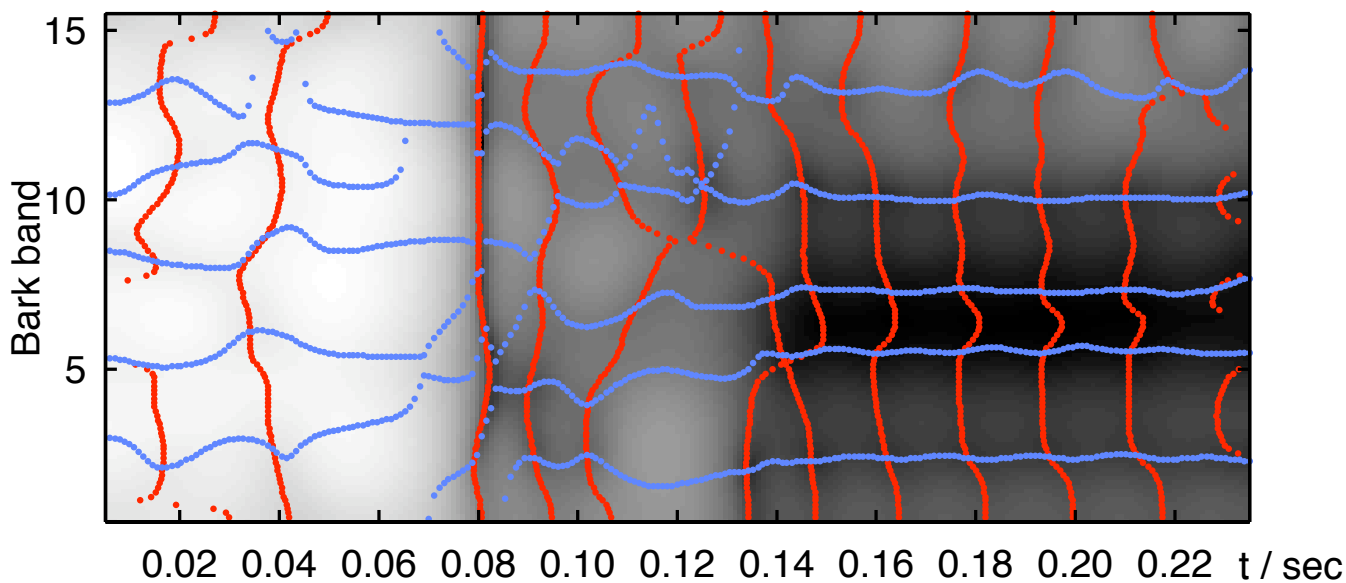


- Perceptual audio features...

PLP-squared

Marios Athineos
Hynek Hermansky

- FDLP fits temporal envelope with LP
  - Perceptual Linear Prediction (PLP) smooths across frequency
  - can we do both... iteratively?
- Speech features without ST windows

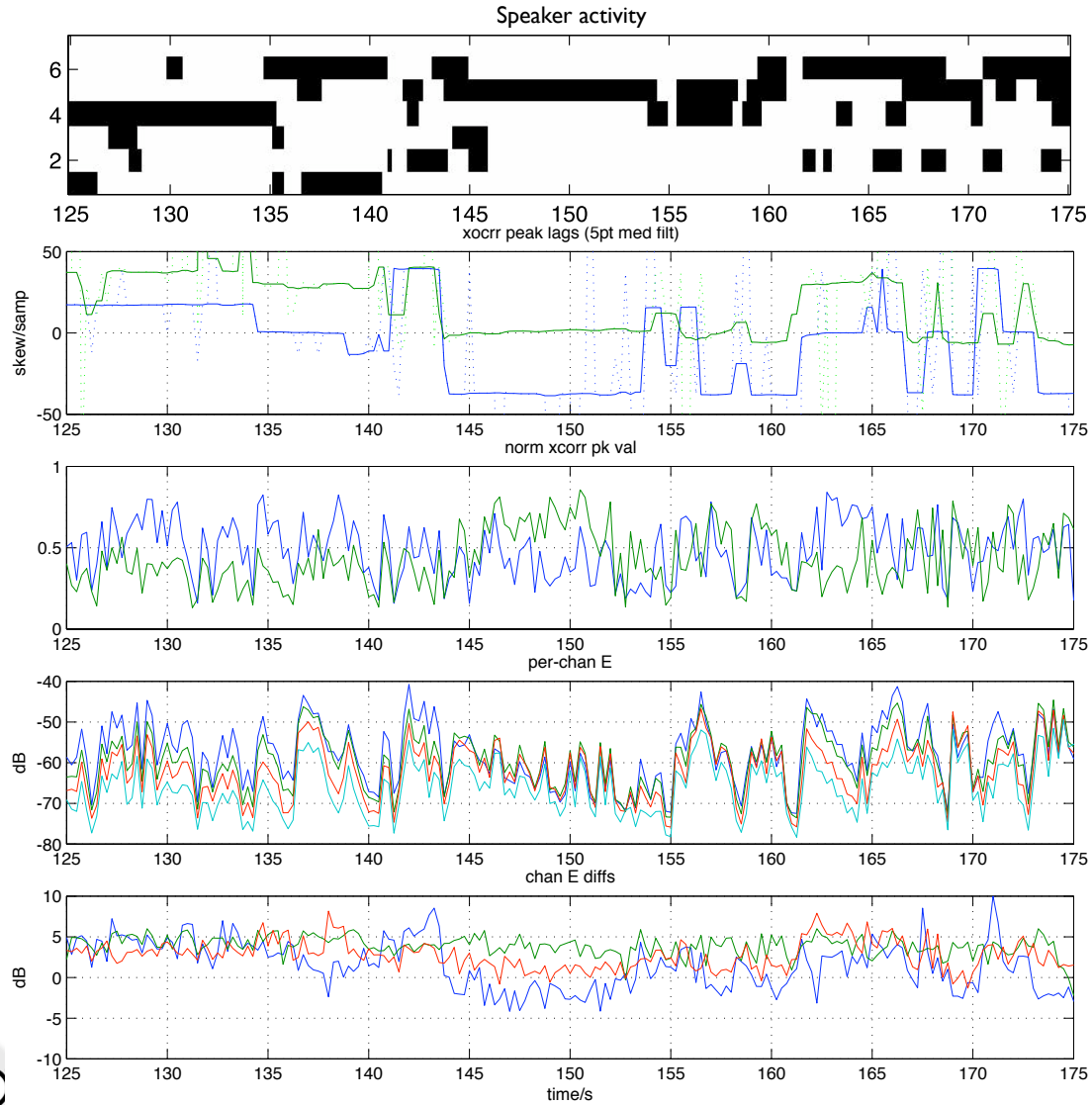


3. Meeting Turns

with Jerry Liu and ICSI

- Multi-mic recordings for 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)

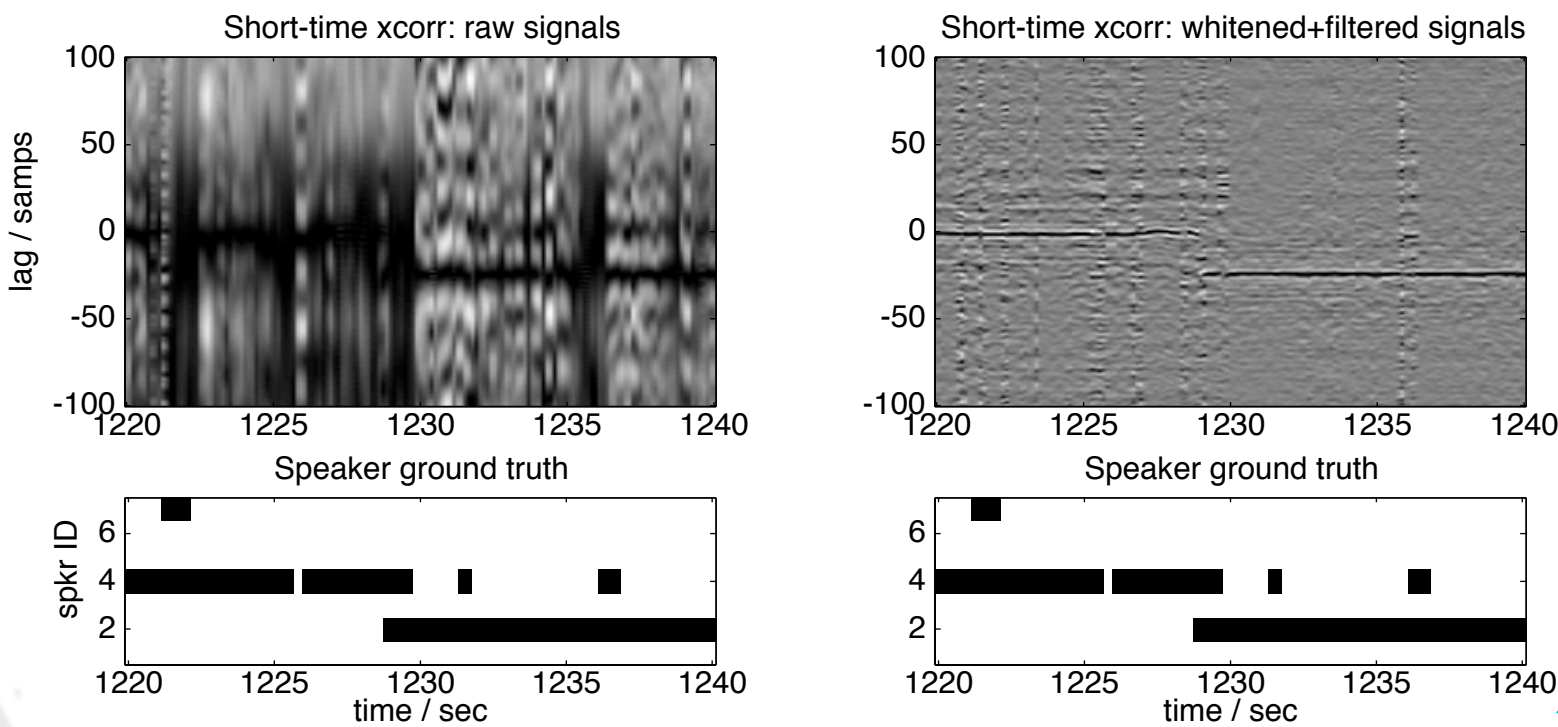
Peak correlation coefficient *r*

Per-channel energy

Between-channel energy differences

Pre-whitening for ITD

- Inverse-filter by 12-pole LPC models (32 ms windows) to remove local resonances
- Filter out noise < 500 Hz, > 6 kHz
- Then cross-correlate...

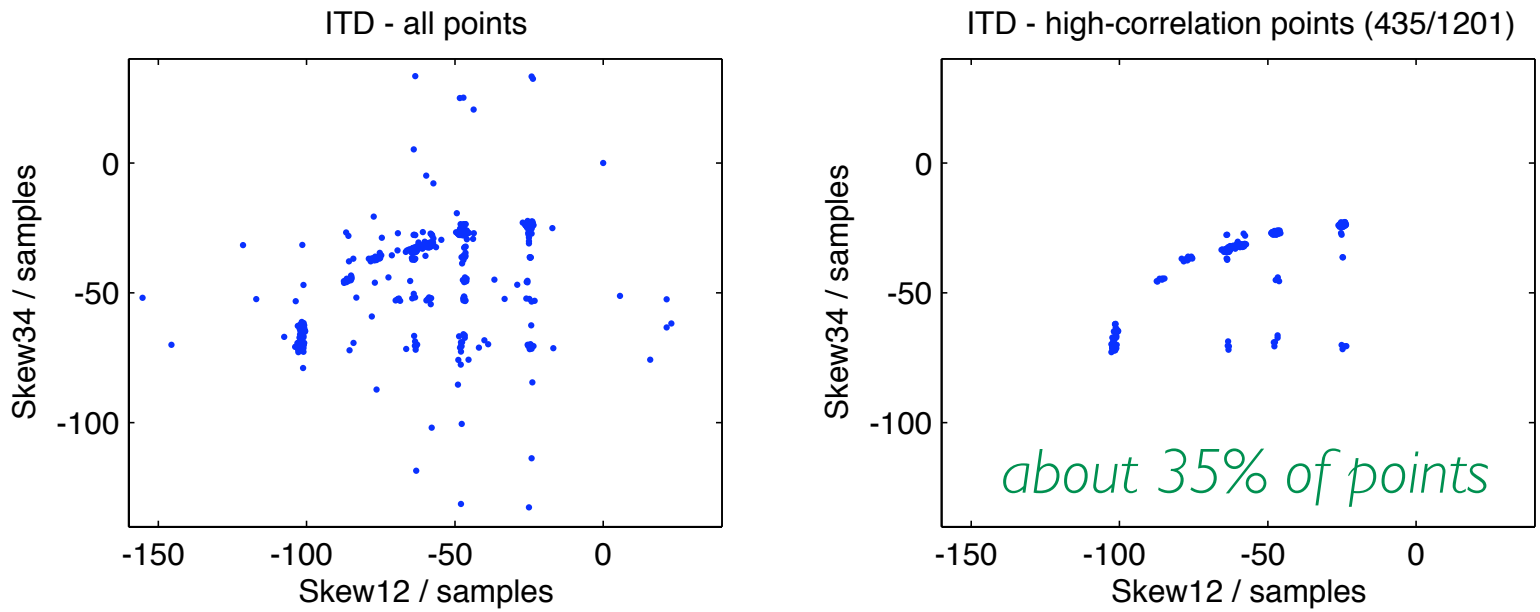


Choosing "Good" Frames

- Correlation coef. r ~ channel similarity:

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

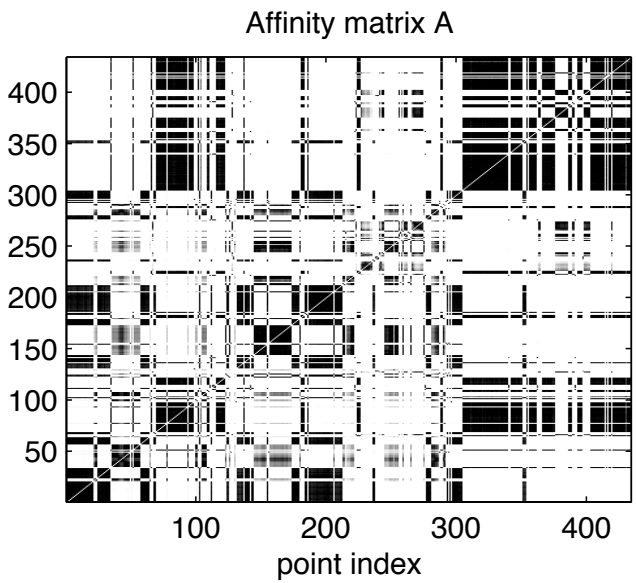
- Select frames with r in top 50% in both pairs

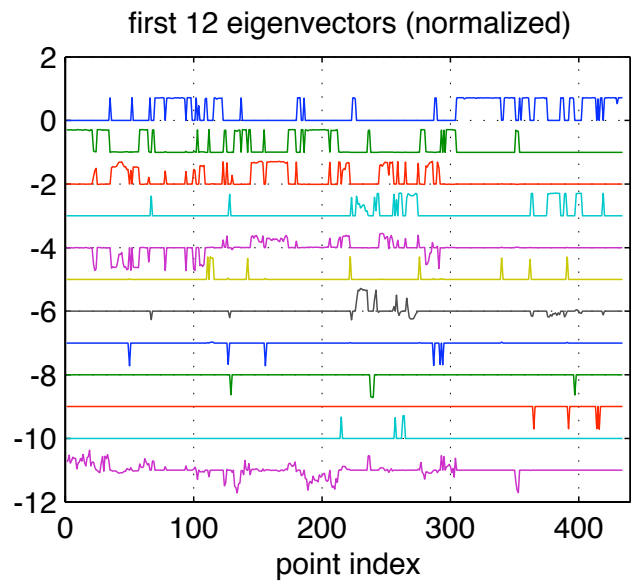


- Cleaner basis for models

Spectral clustering

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



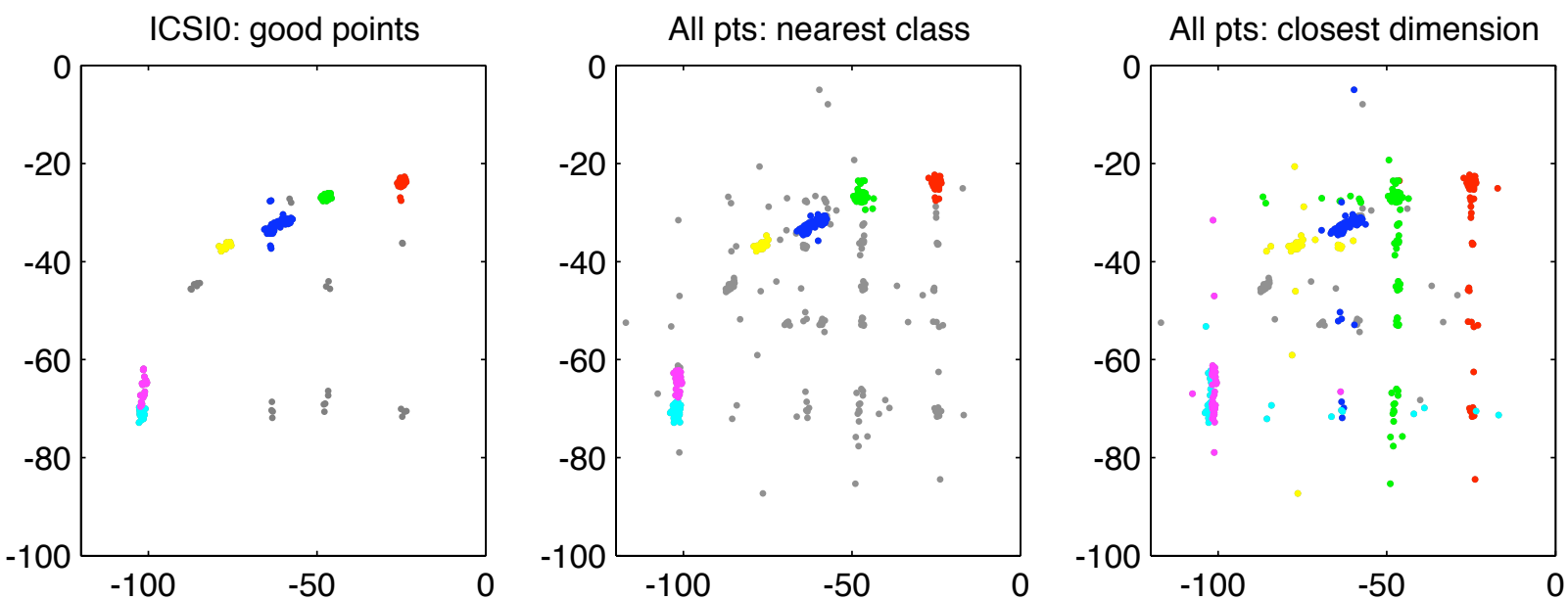


$$a_{mn} = exp\{-\|\mathbf{x}[m] - \mathbf{x}[n]\|^2 / 2\sigma^2\}$$

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

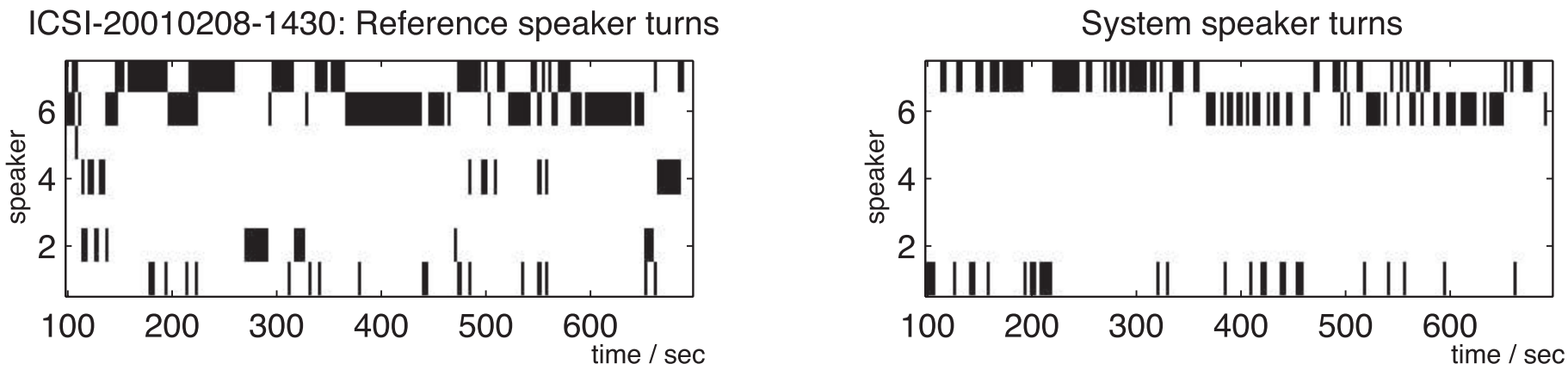
Speaker Models & Classification

- Actual clusters depend on σ and K heuristic
- Fit Gaussians to each cluster, assign that class to all frames within radius
  - or: consider dimensions 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)
- **~52% avg. error on NIST 2004 dev set**
  - speaker-characteristic-based systems ~25%

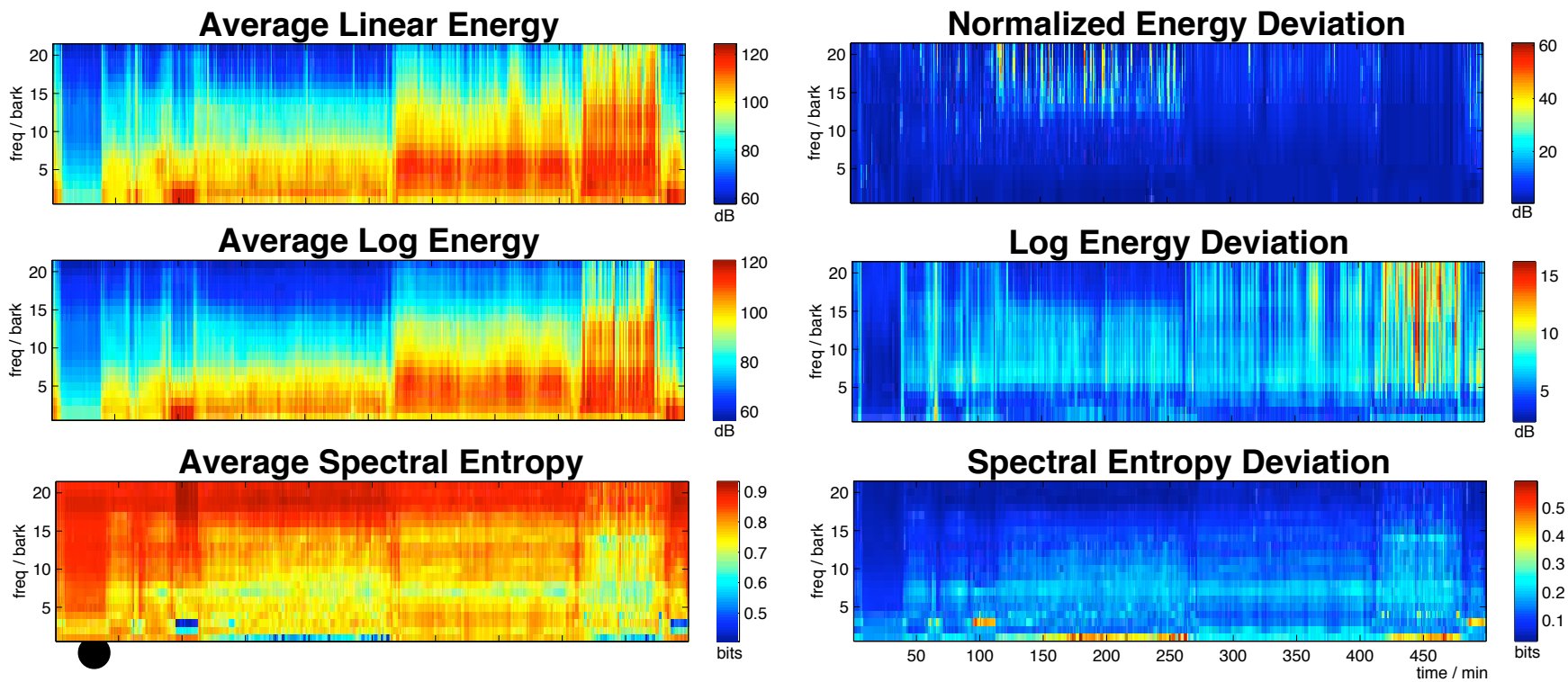
4. Segmenting Personal Audio

with Kean sub Lee

- 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 16 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

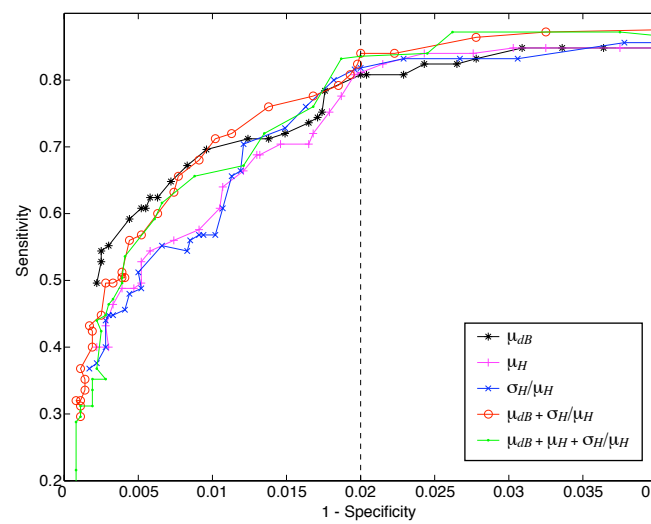
BIC Segmentation

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

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

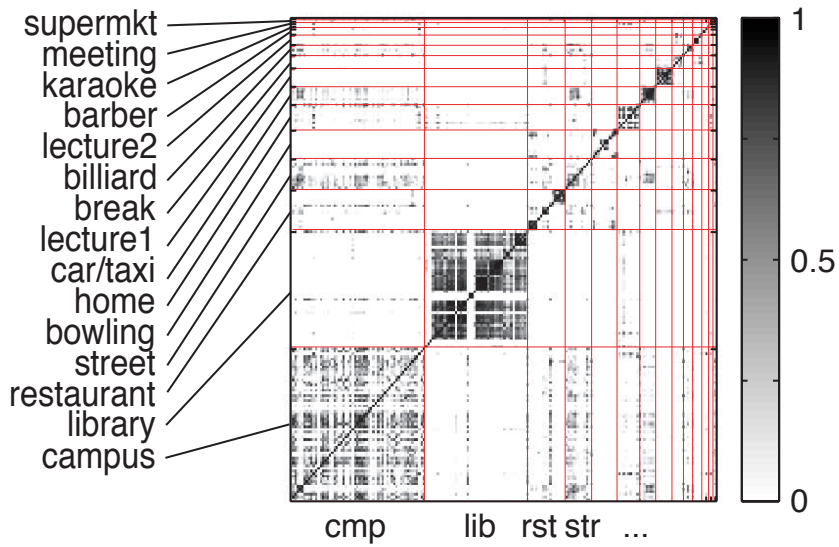
- Evaluate: 60hr hand-marked boundaries
  - different features & combinations
  - Correct Accept % @ False Accept = 2%:

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%



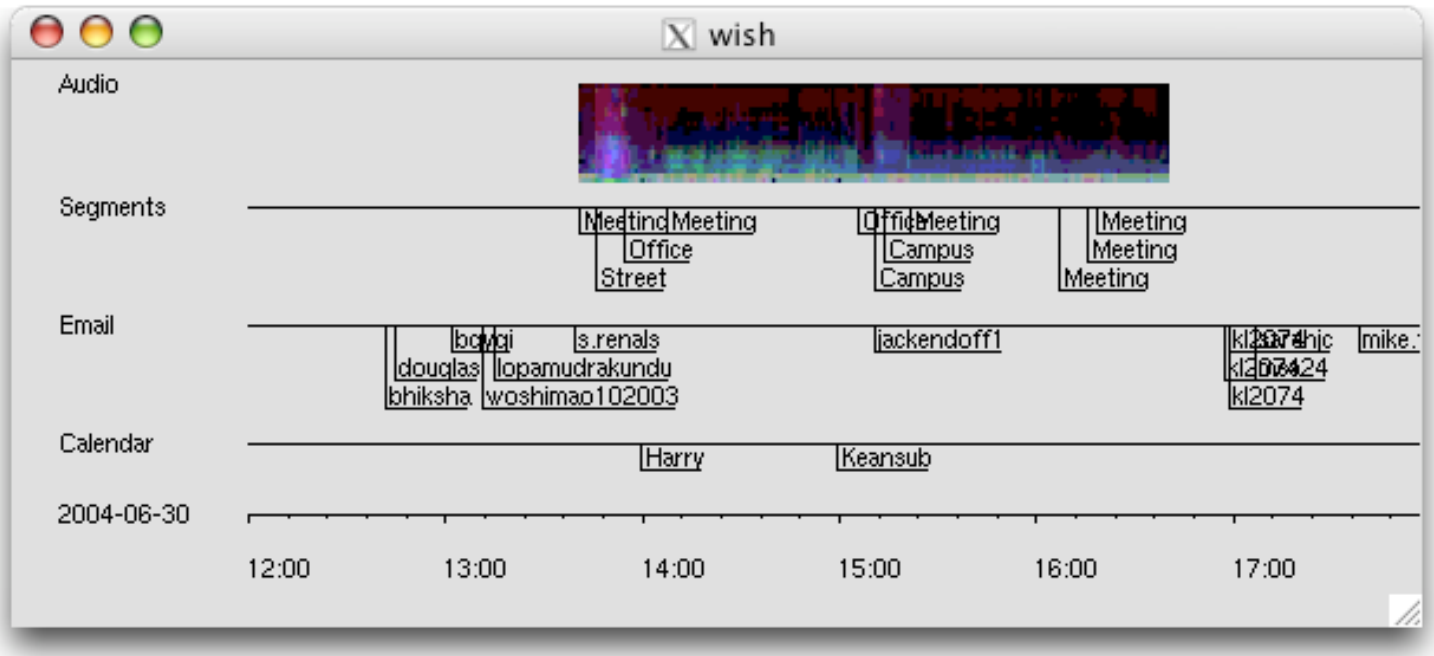
Segment clustering

- Daily activity has lots of repetition: Automatically cluster similar segments



- Spectral clustering achieves ~70% correct
  - 16-way ground truth labels
  - KL distance, smoothed covariance estimates

Future Work

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

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

LabROSA Summary

- LabROSA
  - signal processing
    + machine learning
    + information extraction
- Applications
  - Eigenrhythms: drum pattern models
  - FDLP temporal envelopes
  - Meeting recordings
  - Personal audio analysis
- Also...
  - music similarity, signal separation, ...