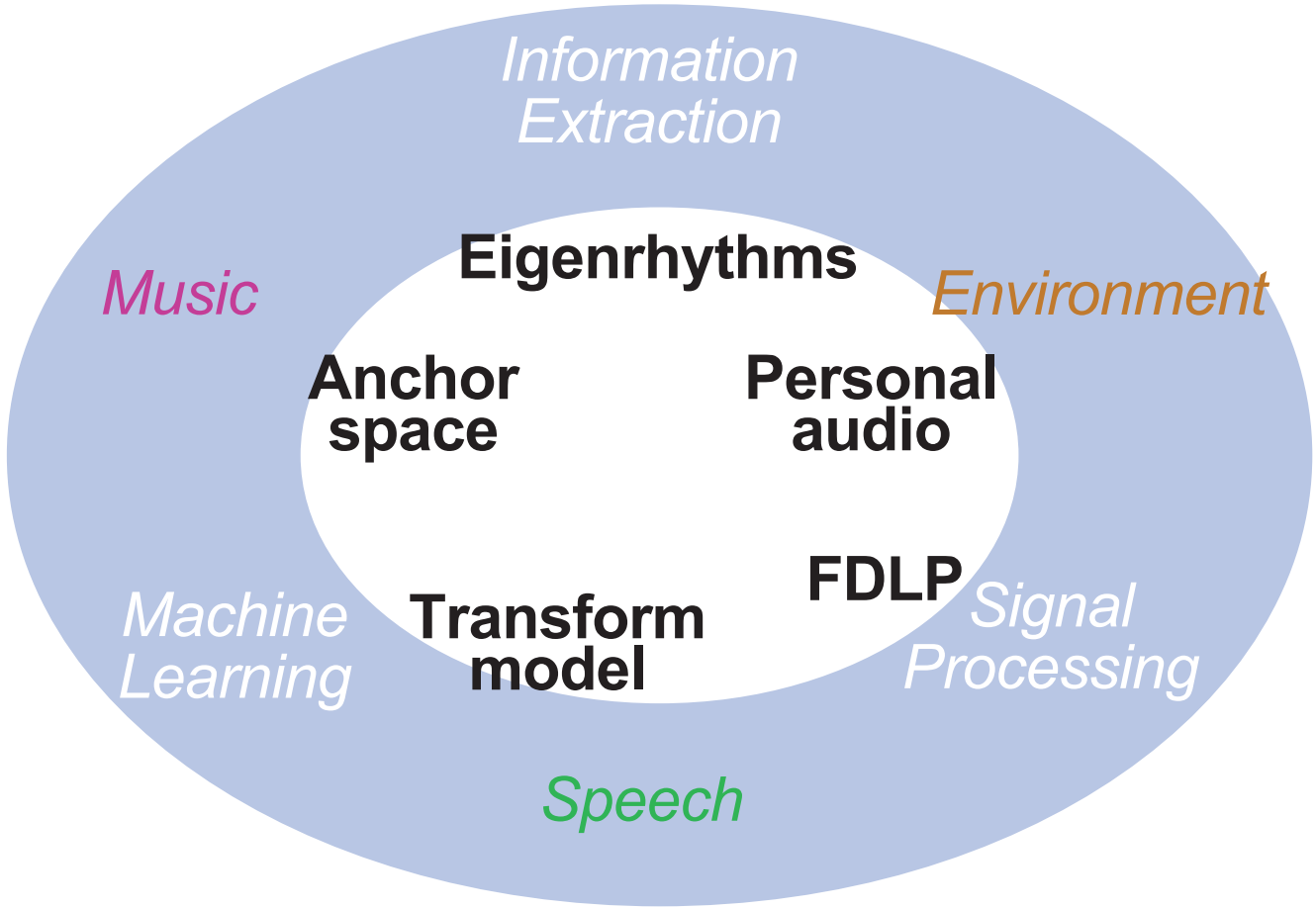
Audio & Music Research at LabROSA

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. Anchor-Space Music Similarity Browsing
4. Transformation-based generative models
5. Analyzing 'personal audio' recordings

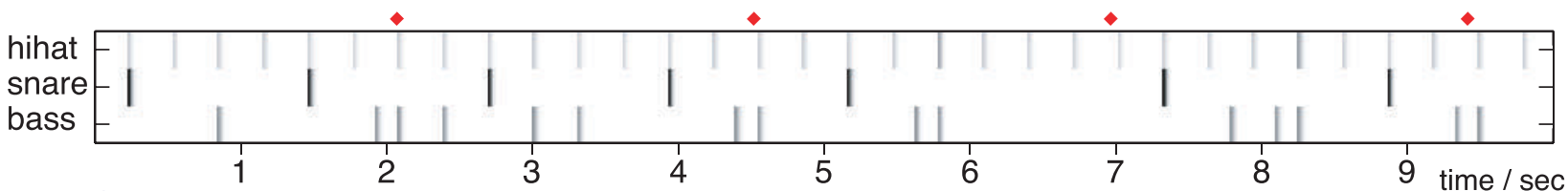
LabROSA Projects Overview



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



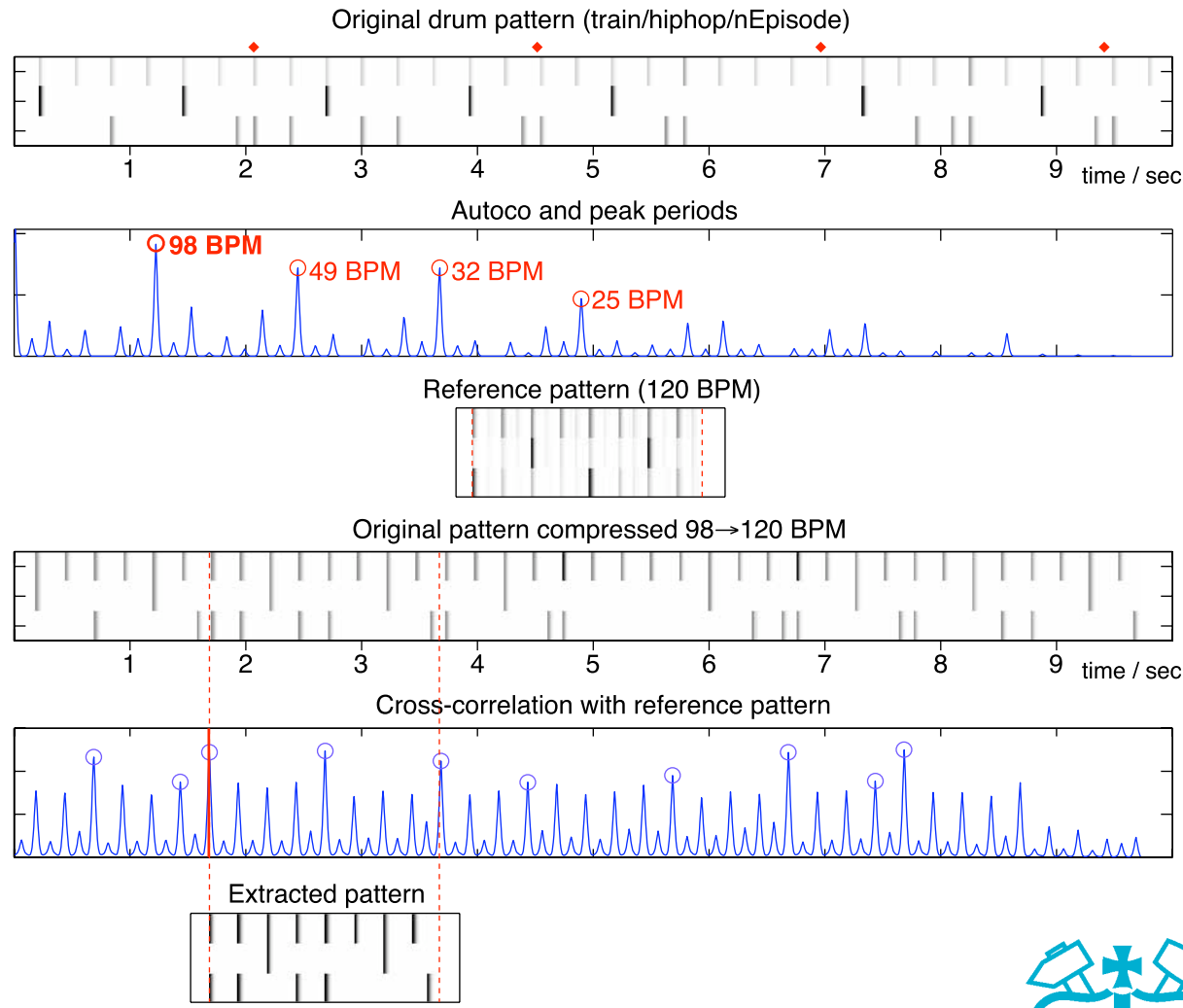
- Eigen-analysis (PCA) to capture variations?
  - by analyzing lots of (MIDI) data
- Applications
  - music categorization
  - "beat box" synthesis

Aligning the Data

- Need to align patterns prior to PCA...

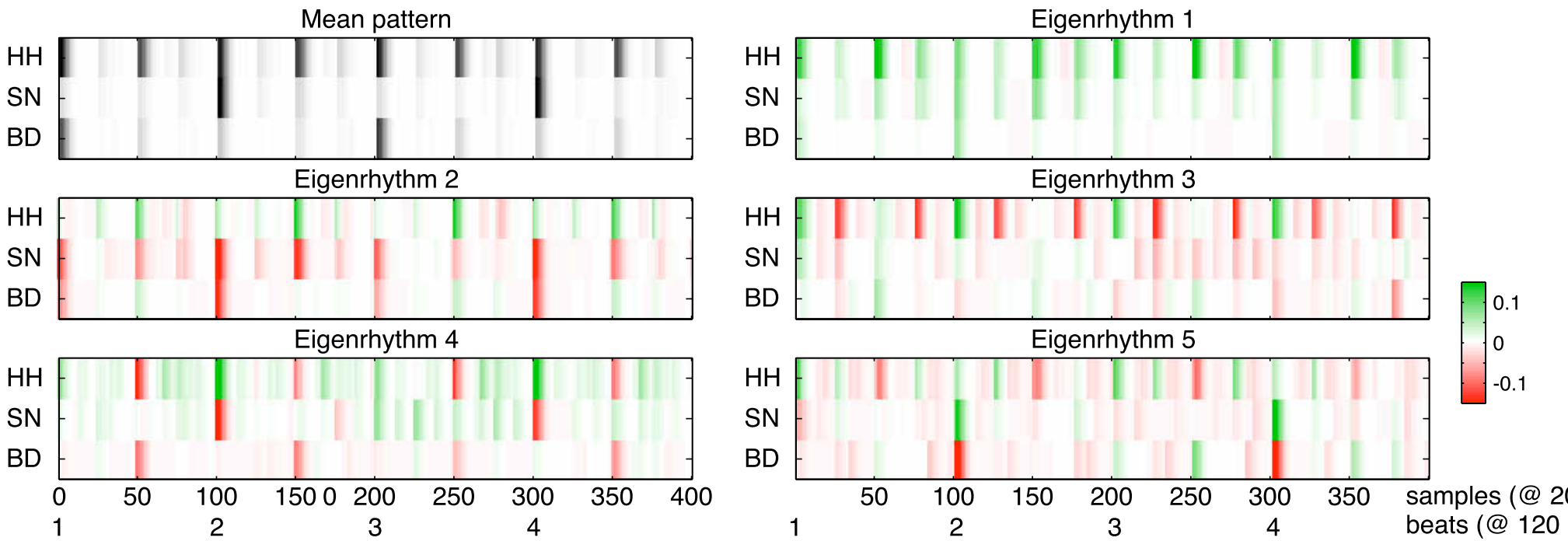
tempo (stretch): by inferring BPM & normalizing

downbeat (shift): correlate against 'mean' template



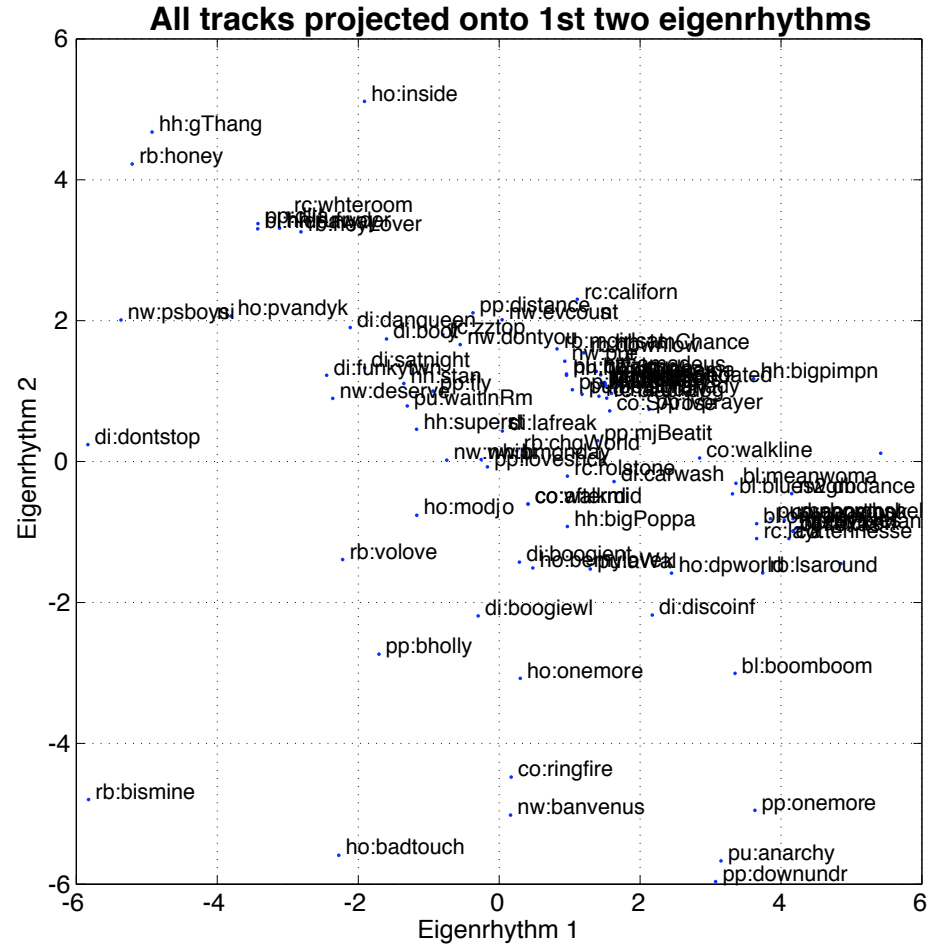
Eigenrhythms

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



Eigenrhythms for Classification

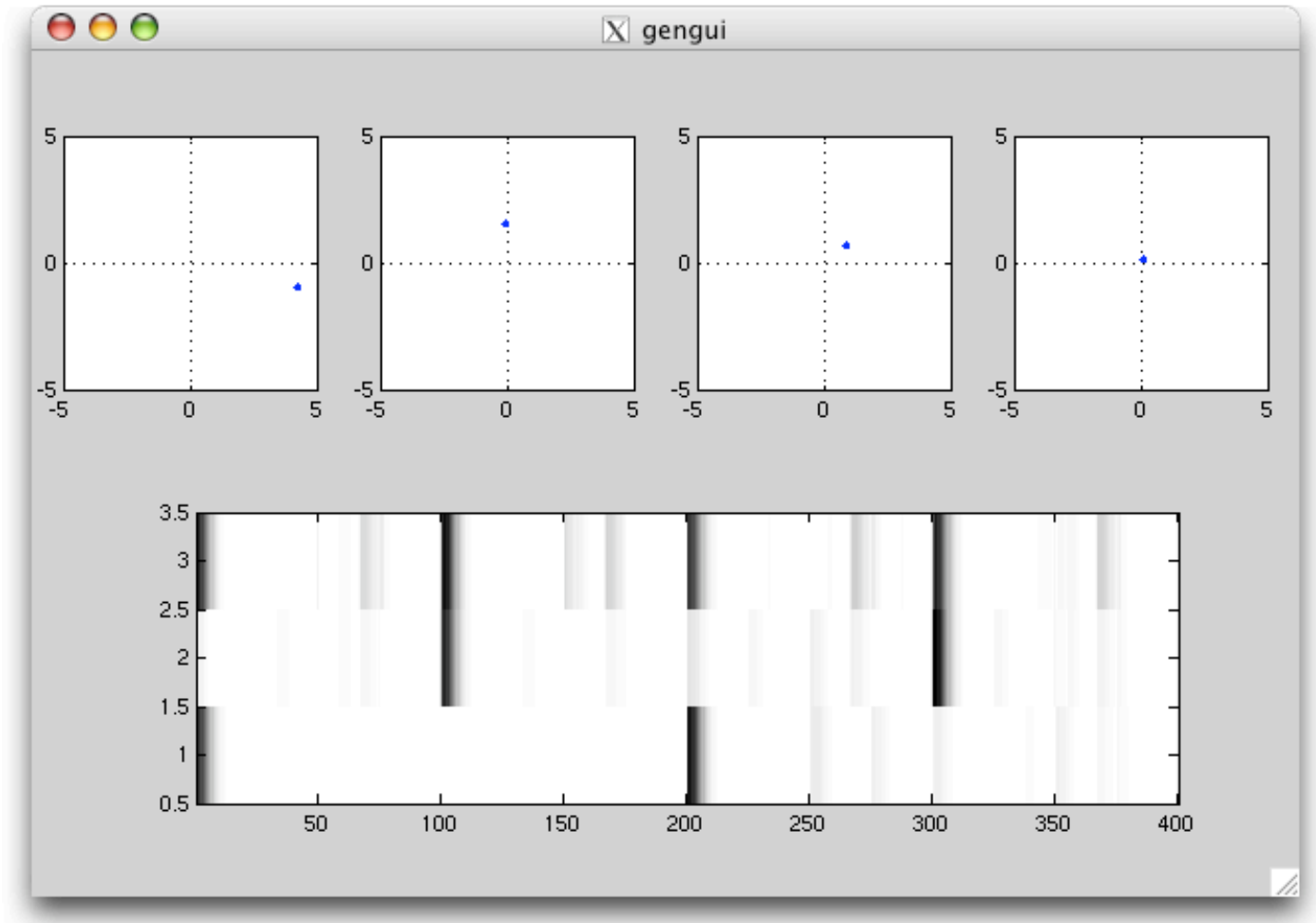
- Clusters in Eigenspace:



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

Eigenrhythm BeatBox

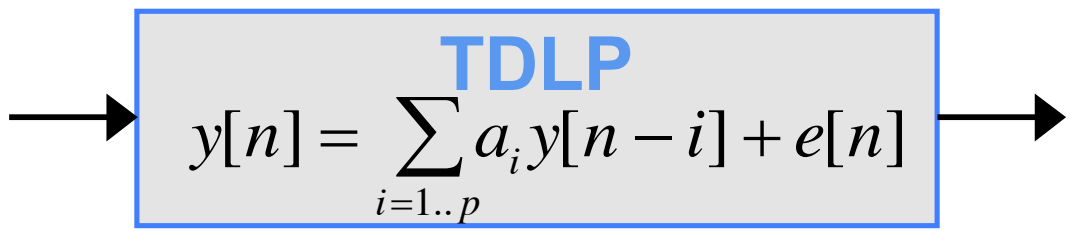
- Resynthesize rhythms from eigen space



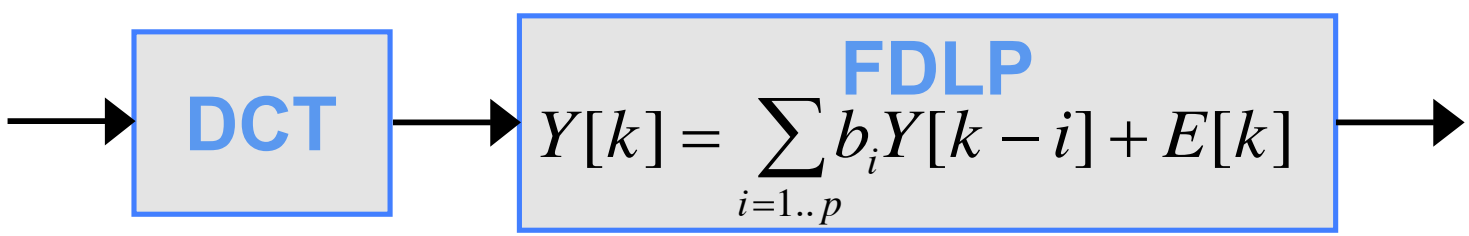
2. Frequency-Domain Lin. Pred.

with Marios Athineos

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

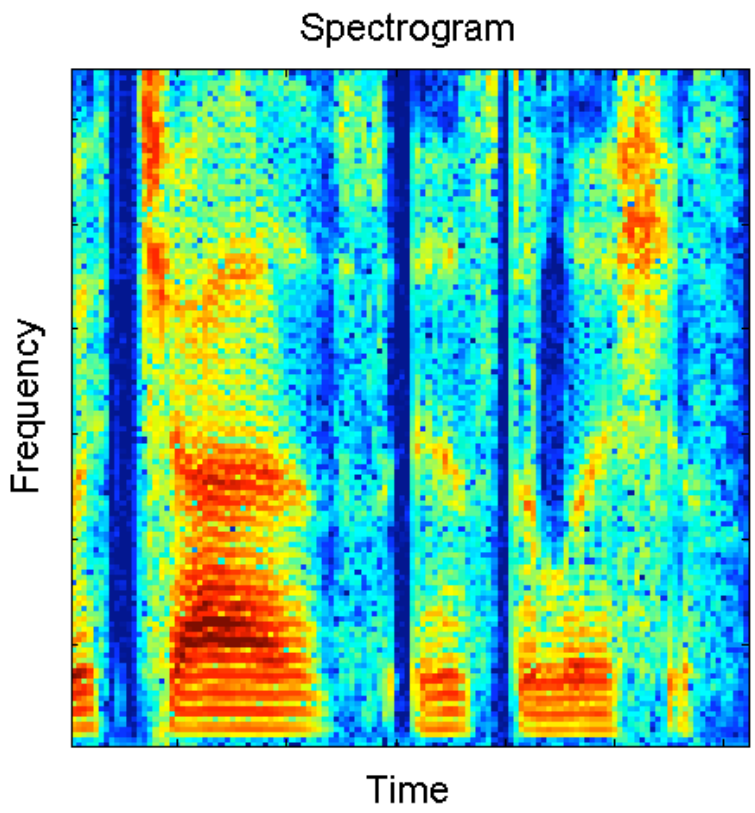


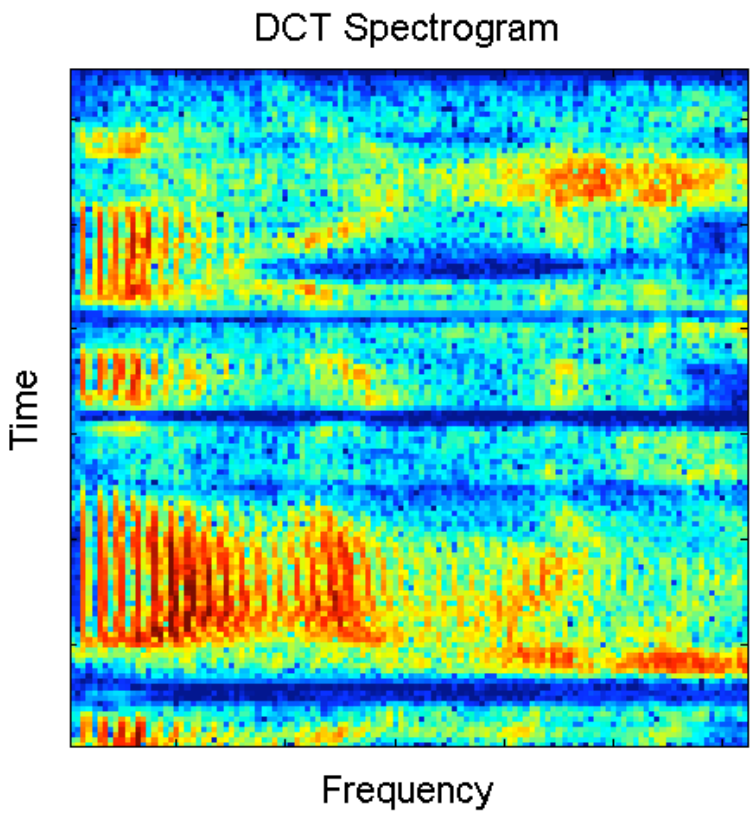
- Apply to a 'frequency domain' signal
  - dual: estimates temporal envelope



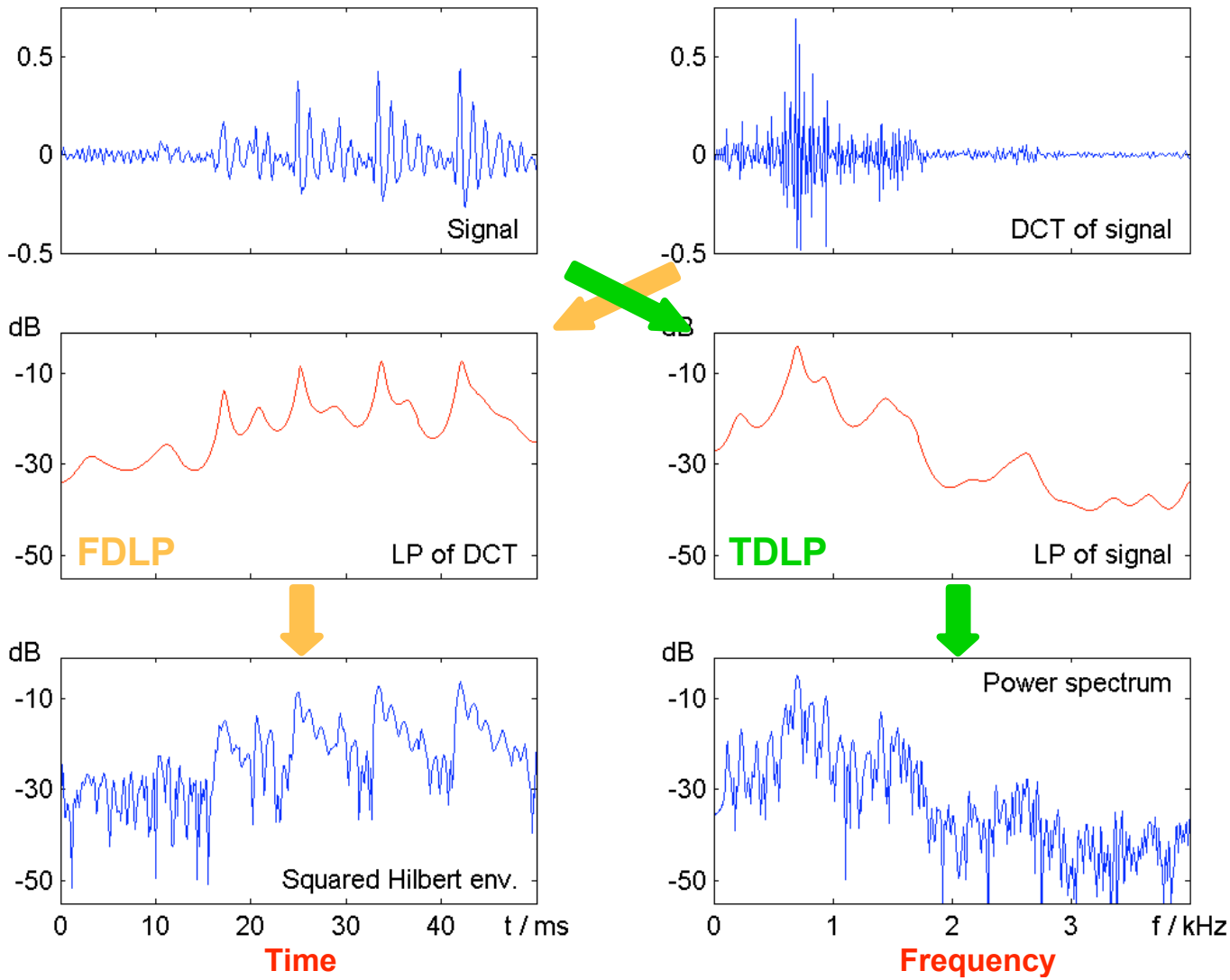
Aside: Spectrogram of the DCT

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





FDLP and TDLP Duality



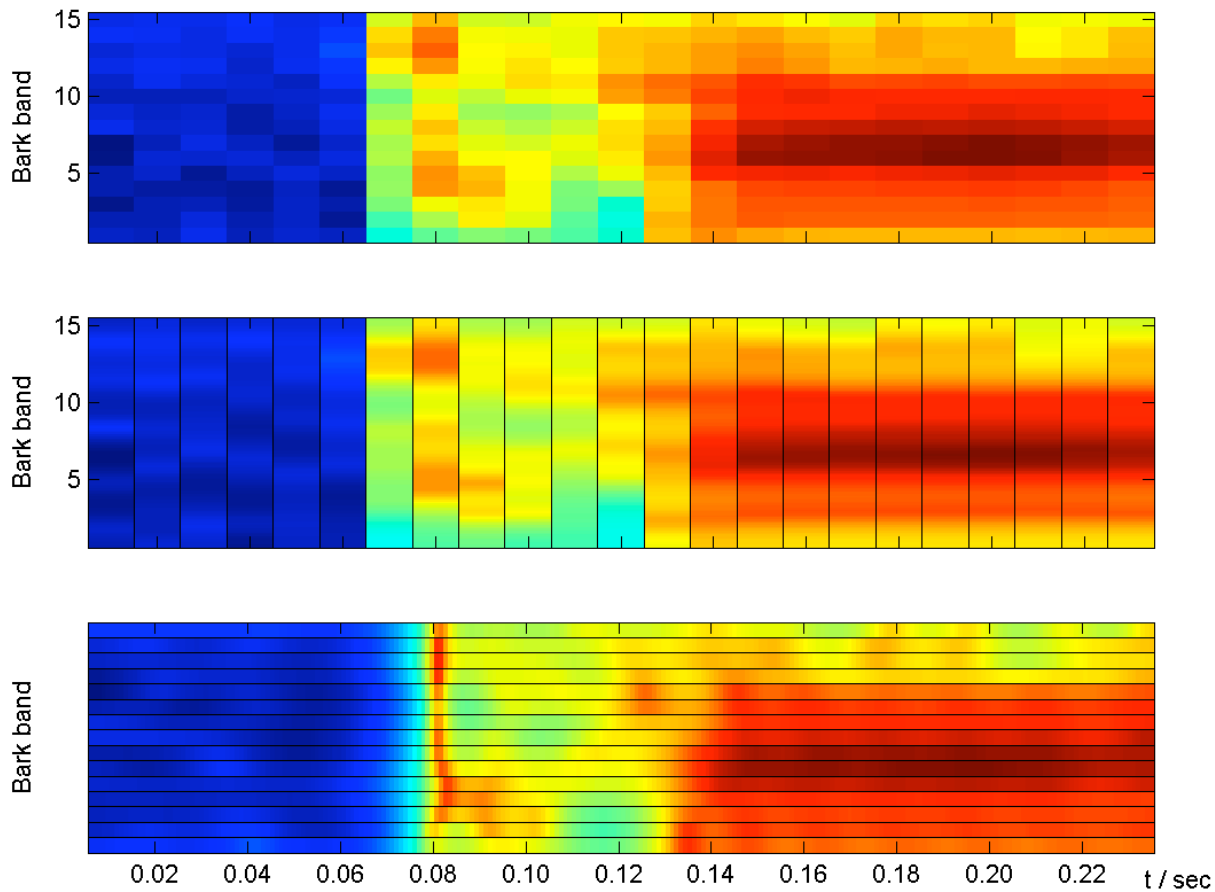
Subband FDLP

- Temporal envelopes without 25 ms windows

Auditory STFT
(10-25ms + Bark bin)

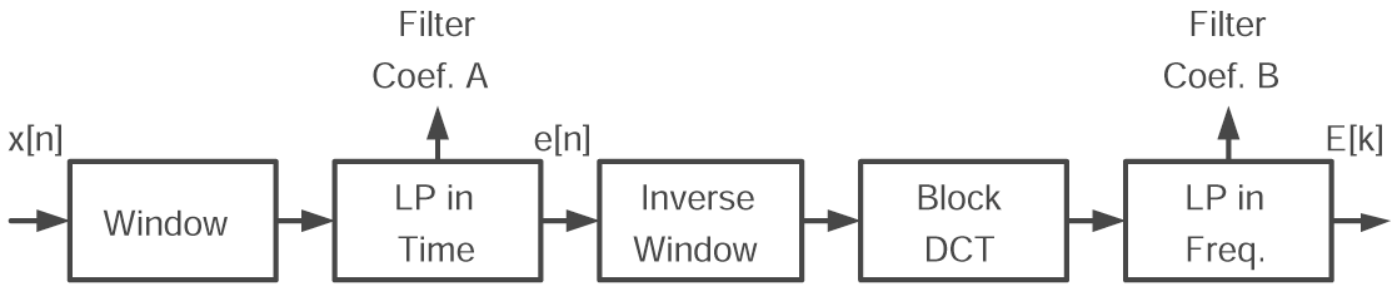
TDLP
(per time frame)

Subband FDLP
(per frequency subband)

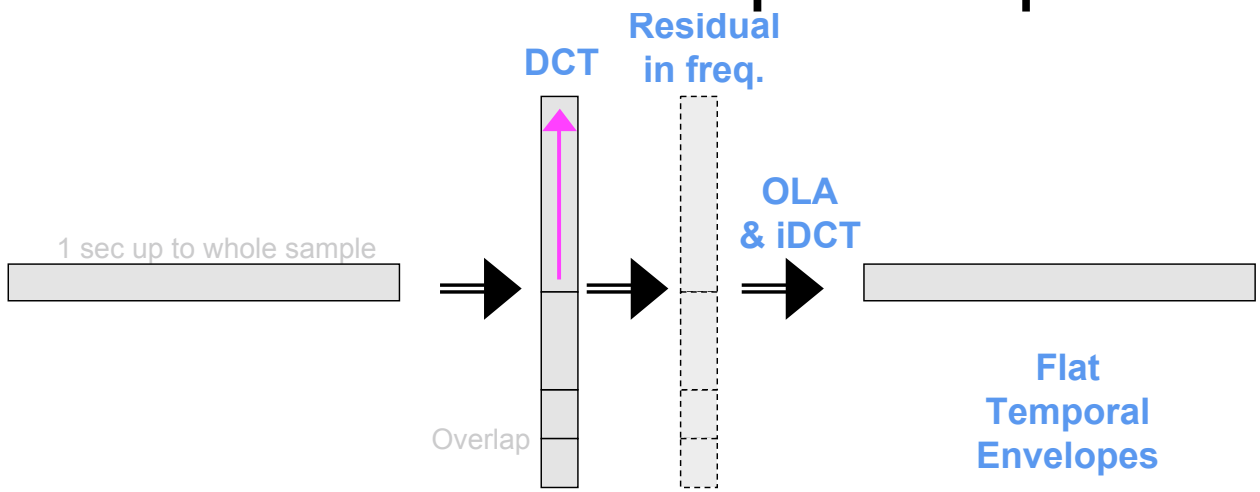


FDLP Applications

- Time-scale modification



- Modulation-domain "temporal equalization"

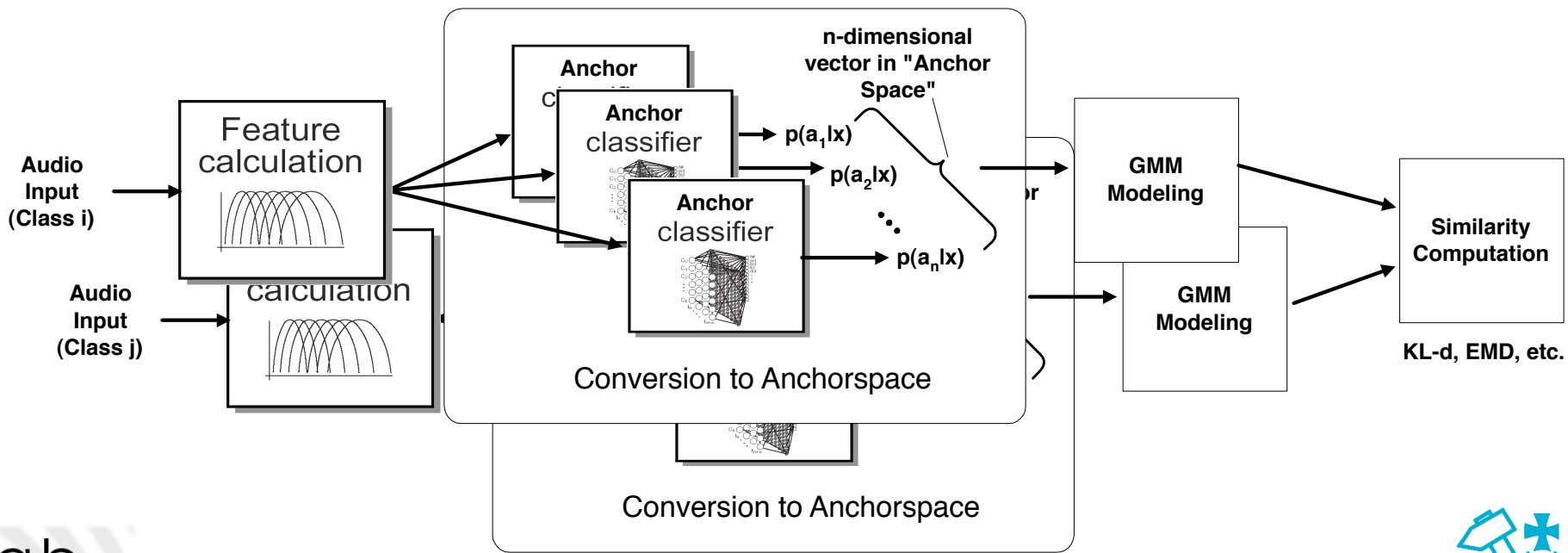


- Perceptual audio features... "PLP-squared"

3. Music Similarity Browsing

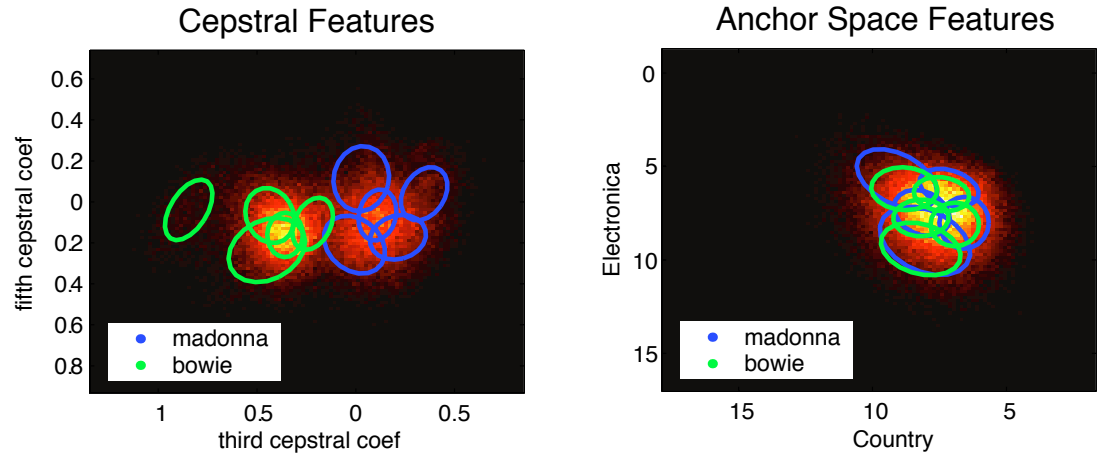
with Adam Berenzweig

- **Musical information overload**
  - record companies filter/categorize music
  - an automatic system would be less odious
- **Connecting audio and preference**
  - map to a 'semantic space'?

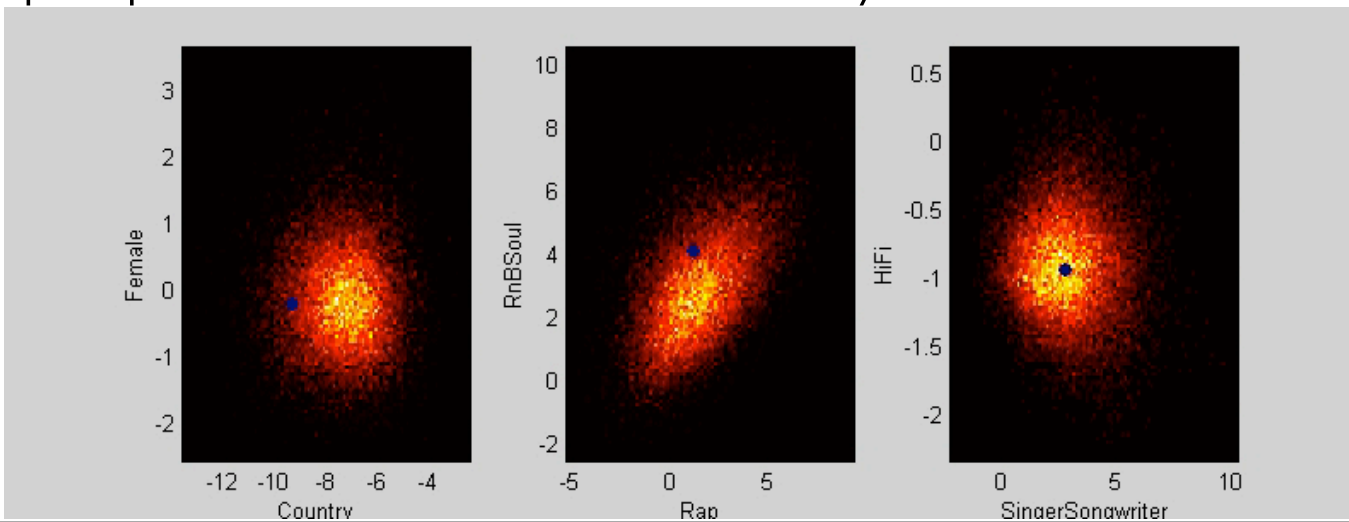


Anchor Space

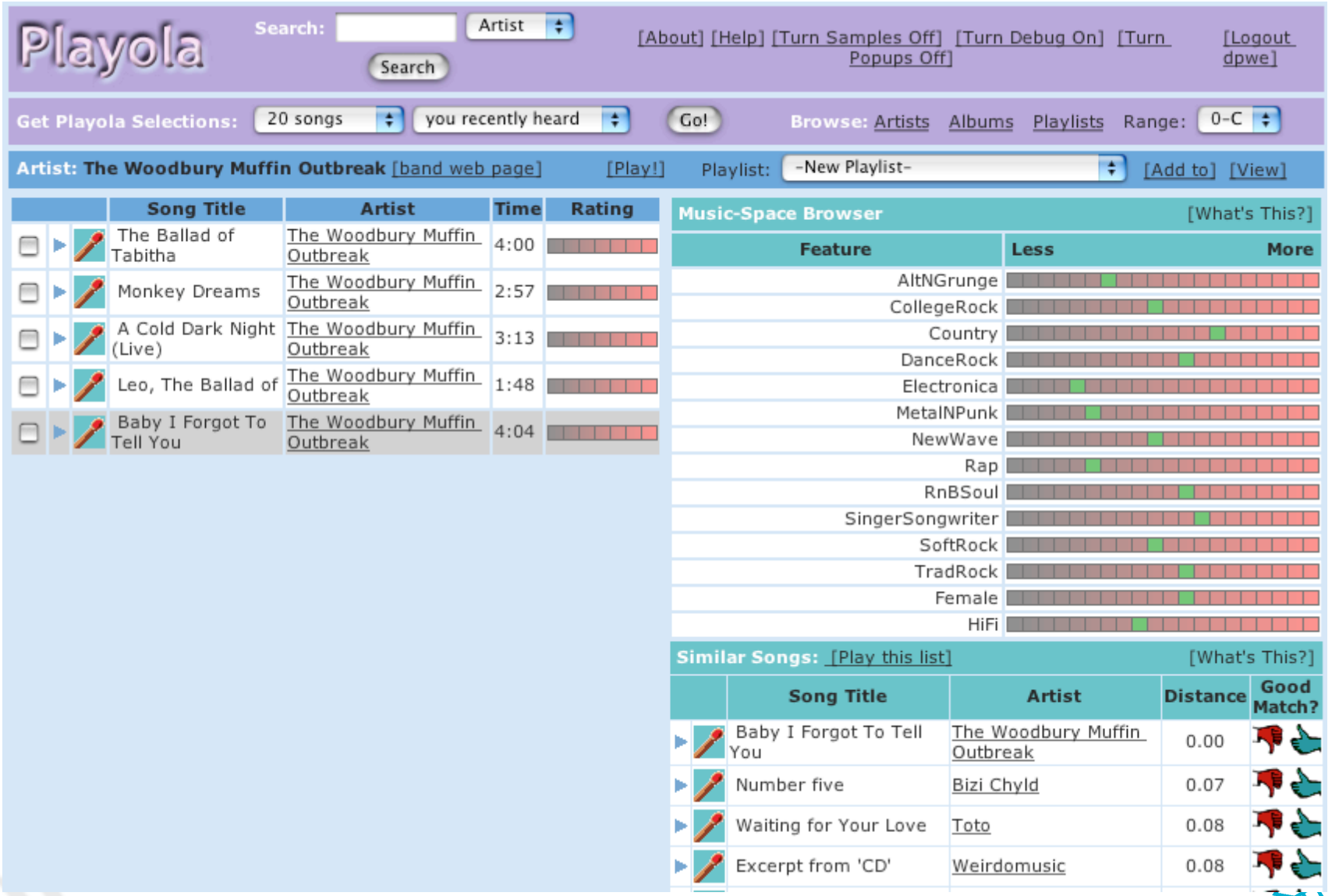
- Frame-by-frame high-level categorizations
  - compare to raw features?



  - properties in distributions? dynamics?



‘Playola’ Similarity Browser



Ground-truth data

- Hard to evaluate Playola's 'accuracy'
  - user tests...
  - ground truth?

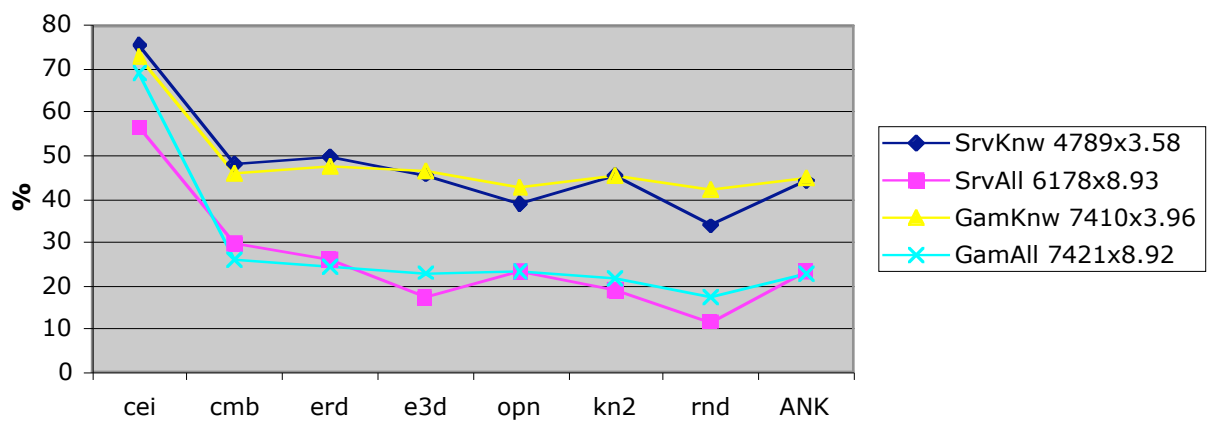
- "Musicseer" online survey:
  - ran for 9 months in 2002
  - > 1,000 users, > 20k judgments
  - http://labrosa.ee.columbia.edu/projects/musicsim/

Evaluation

- Compare Anchor Space measures against Musicseer subjective results
  - "triplet" agreement percentage
  - Top-N ranking agreement score:

$$s_i = \sum_{r=1}^{N} \alpha_r^r \alpha_c^{k_r} \qquad \alpha_r = \left(\frac{1}{2}\right)^{\frac{1}{3}} \qquad \alpha_c = \alpha_r^2$$

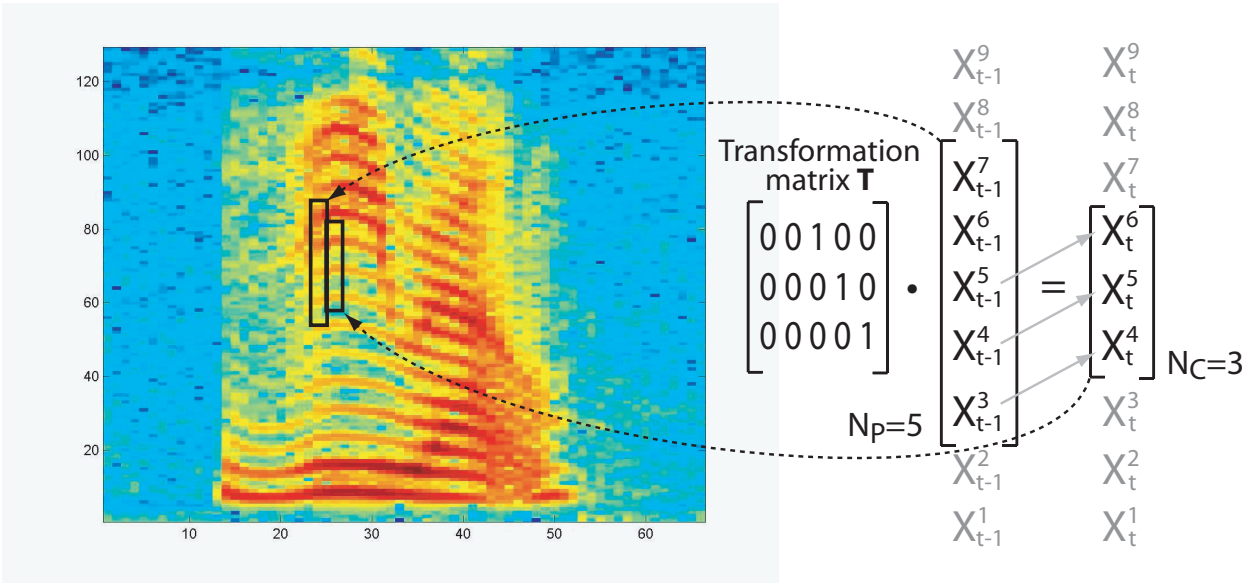
  - First-place agreement percentage
    - simple significance test



4. Transformation-based models

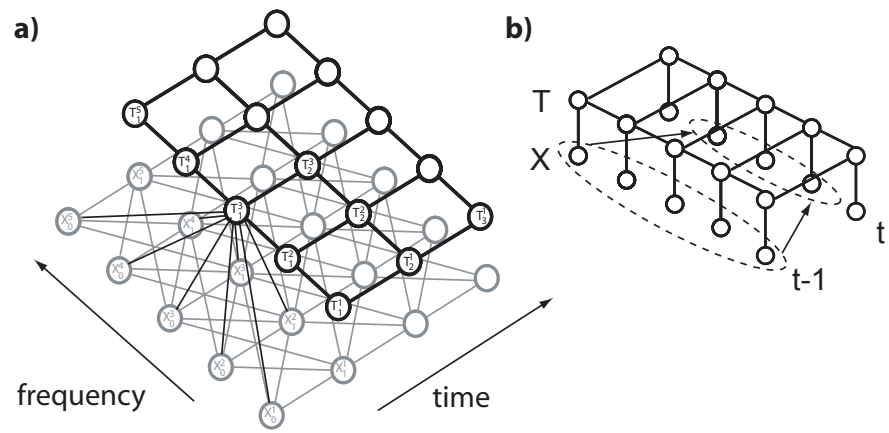
with Manuel Reyes and Nebojsa Jojic

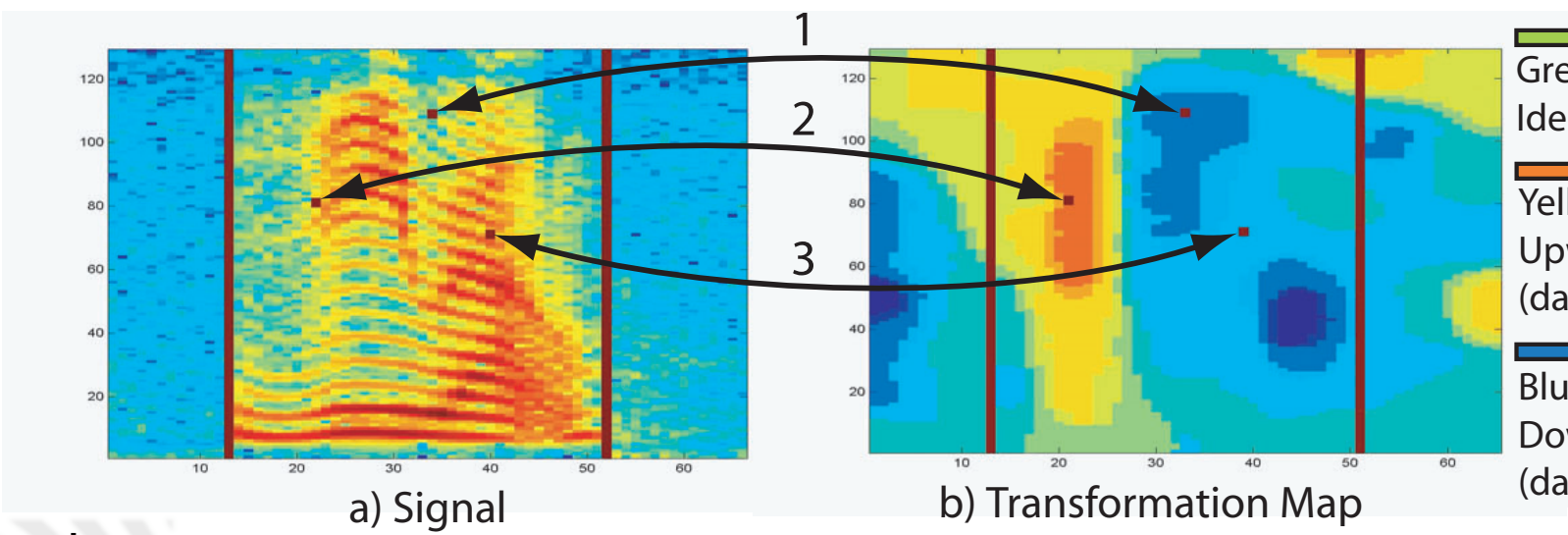
- **HMMs are poor generative models**
  - accurate modeling requires 1000s of states
- **Observation:**
  Speech spectra undergo minor deformations
  - suggests a different generative model?



States+Transformation Model

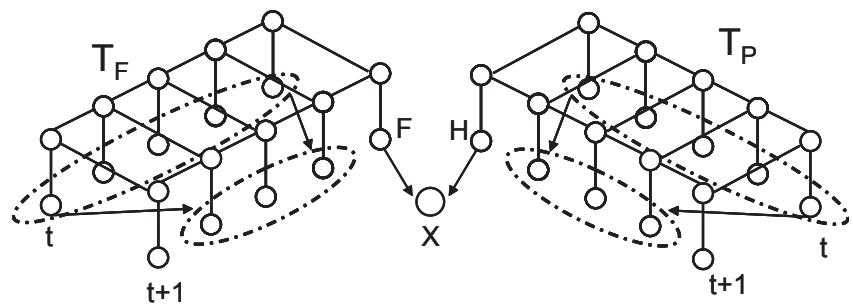
- Time-frequency state grid
- State → explicit prototype or a transformation on prior frame
- Infer underlying states

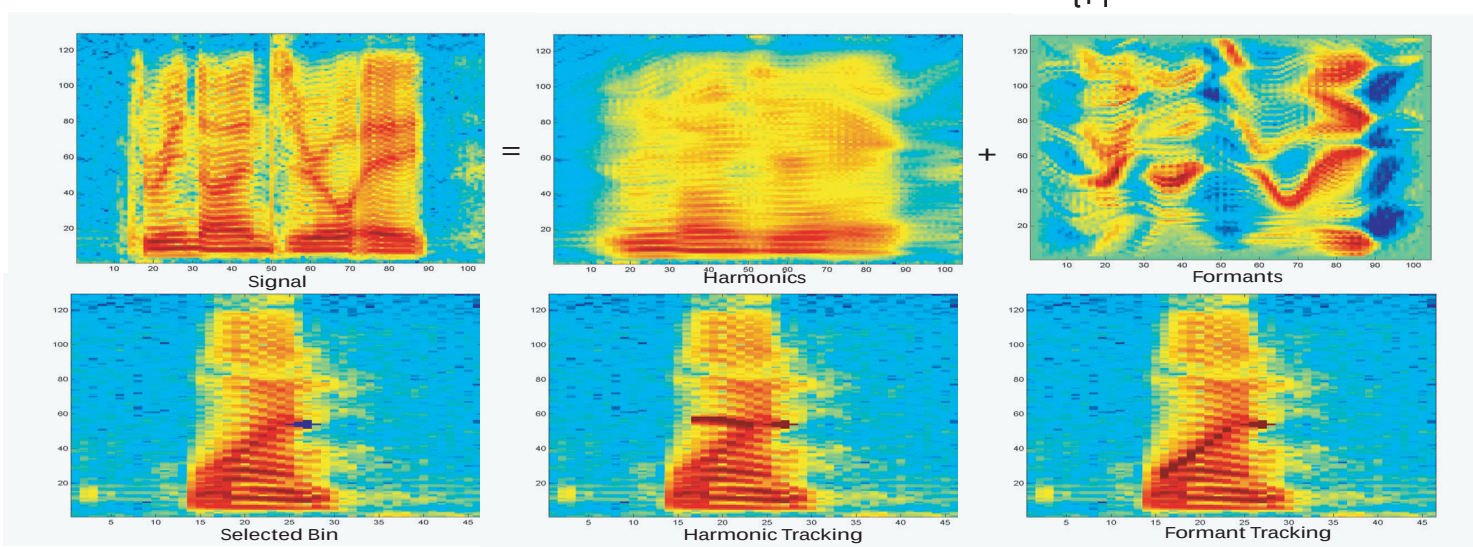




Two-layer model

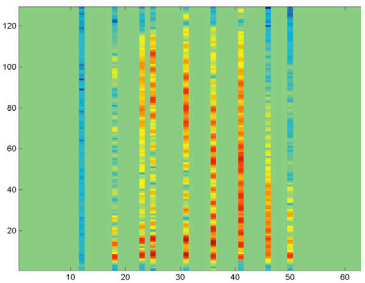
- **Source-filter decomposition**
  - pitch and formants have different dynamics
- **Apply transformation models for both**
  - log-spectra:
    sum of excitation & filter
  - inference does separation

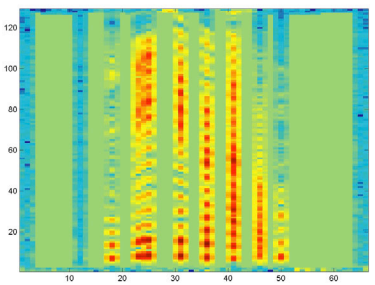


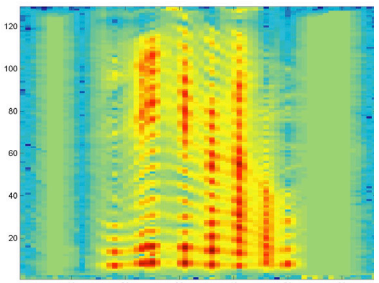


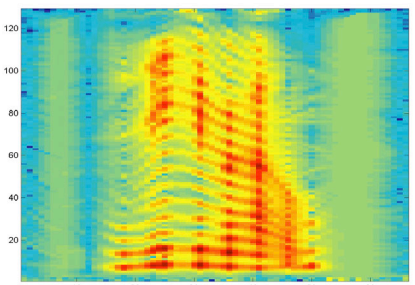
Transformation model applications

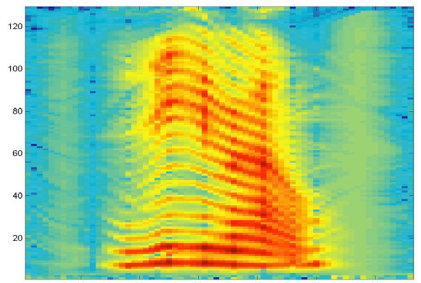
- **Compact, accurate source descriptions**
  - only a few explicit states needed


a) States

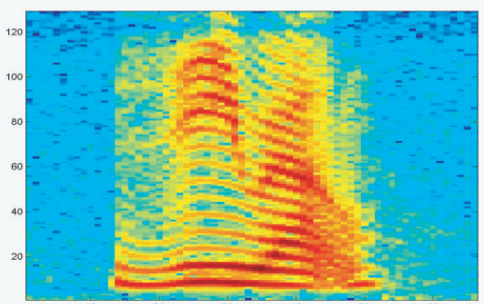

b) Reconstruction; Iter. 1

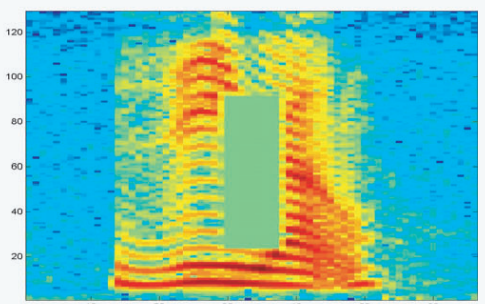

c) Reconstruction; Iter. 3

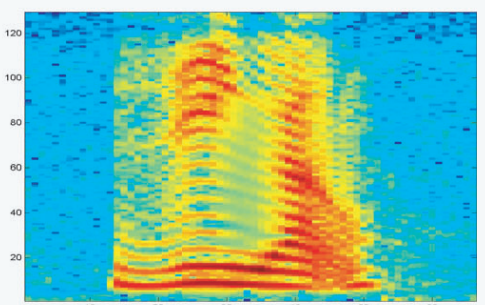

d) Reconstruction; Iter. 5

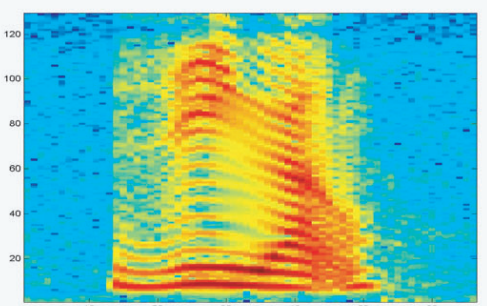

e) Reconstruction; Iter. 8

- **Belief propagation can infer missing values**
  - .. of state grid, hence magnitude spectrum


a) Original


b) Missing Data


c) After iteration 10


d) After iteration 30

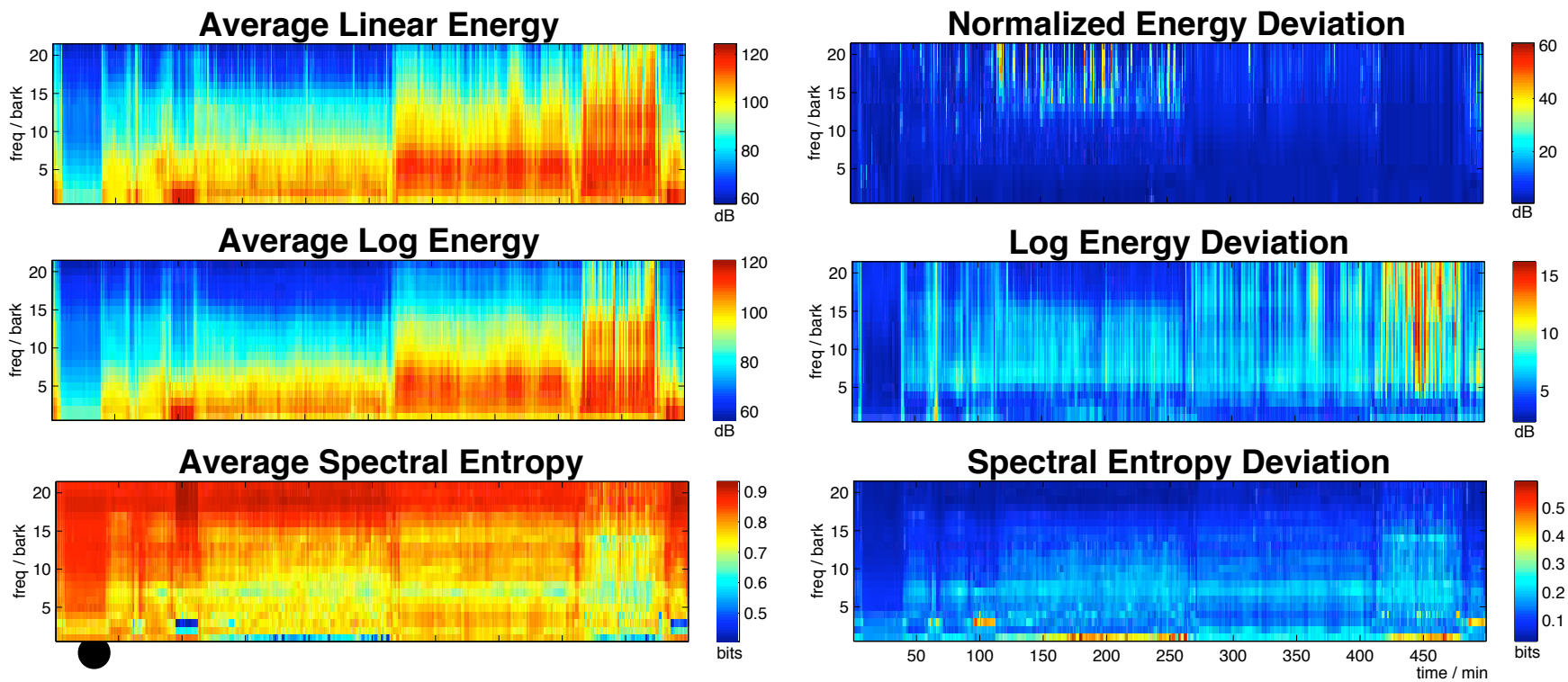
5. Segmenting Personal Audio

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 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

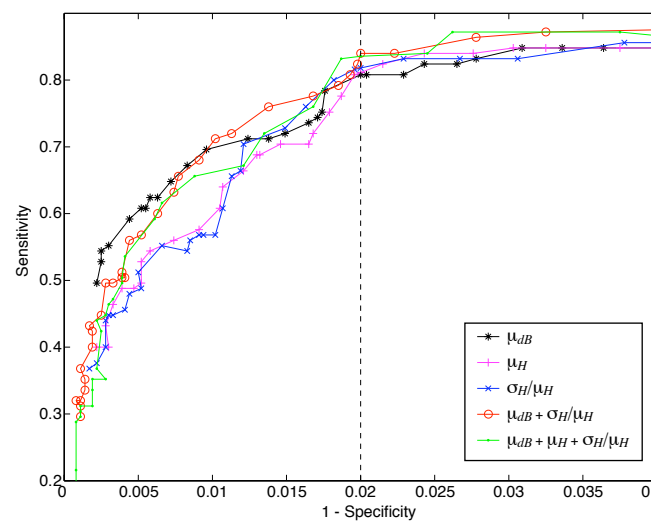
BIC Segmentation

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

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

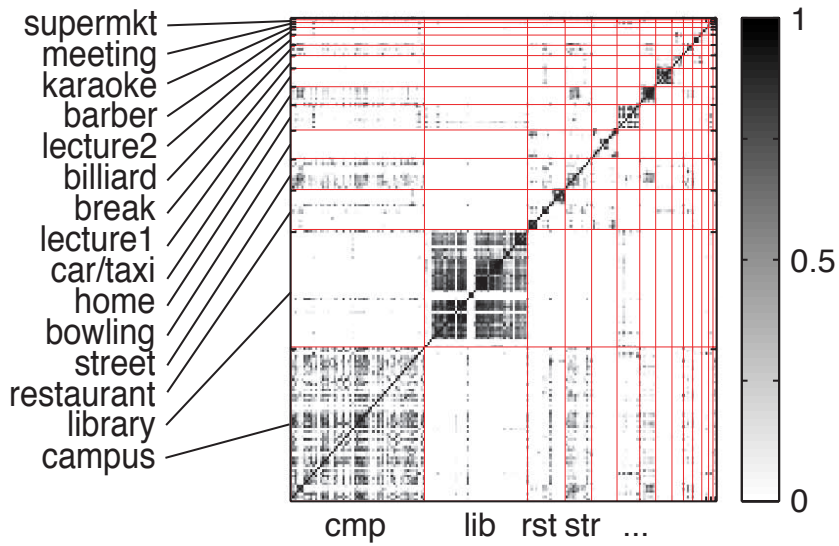
- Evaluate: 60hr hand-marked boundaries
  - different features & combinations
  - Correct Accept % @ False Accept = 2%:

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%



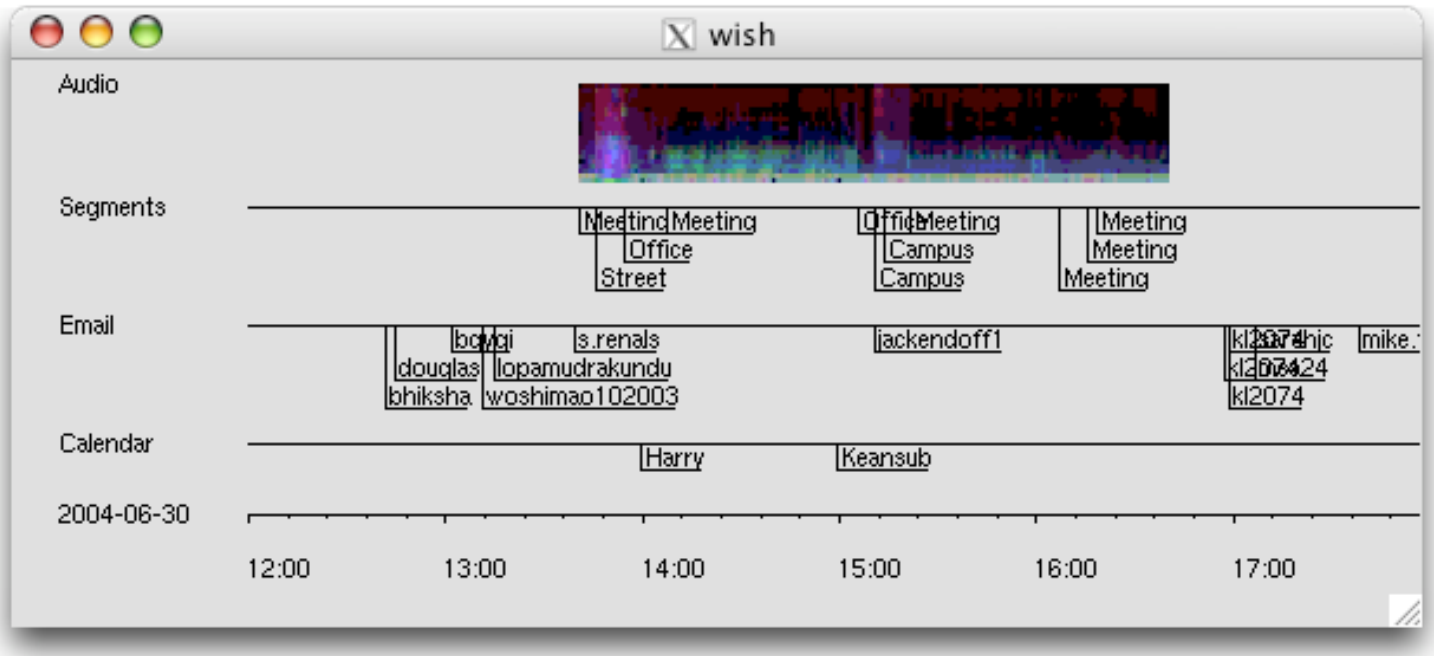
Segment clustering

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



- Spectral clustering achieves ~60% correct
  - 16-way ground truth labels

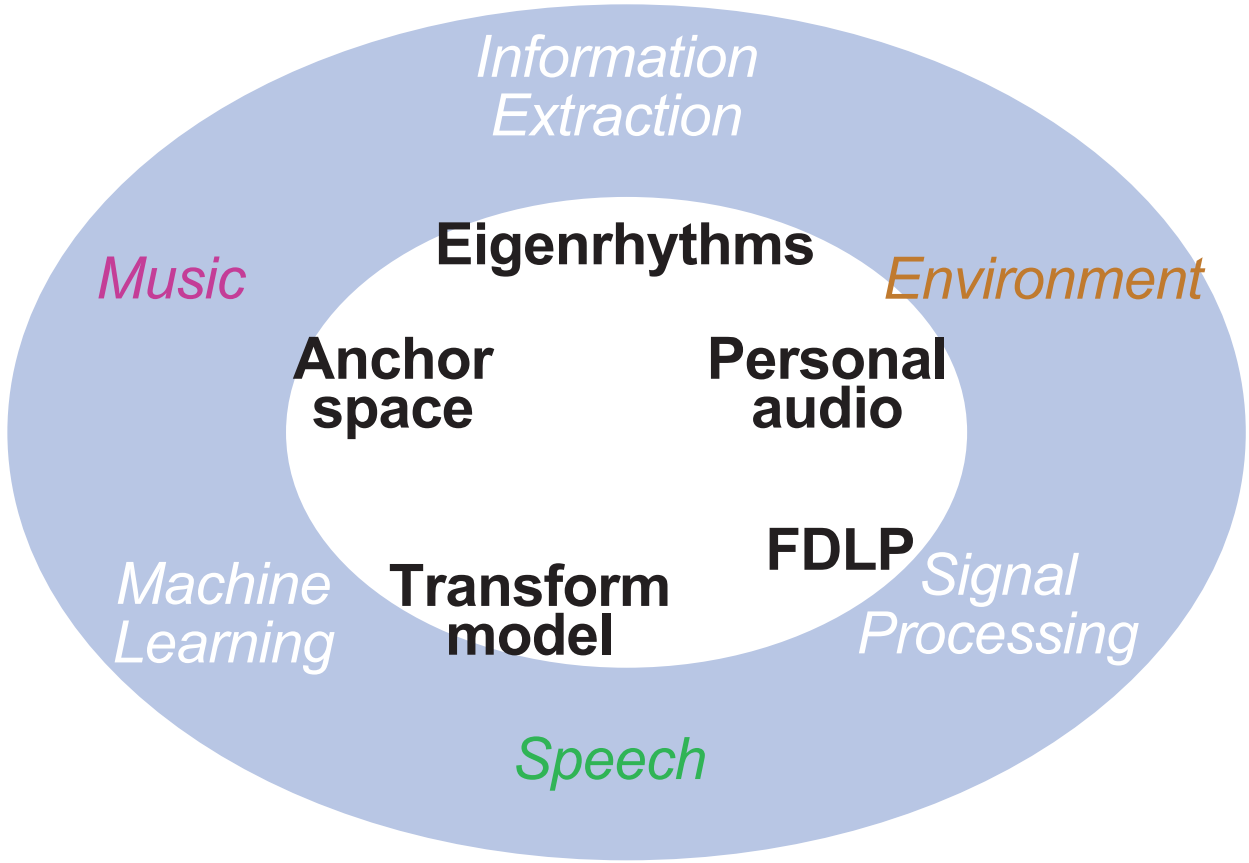
Future Work

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

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

Summary

- Today's topics:



- + Speech recognition, Meeting recordings

LabROSA Summary

- LabROSA
  - signal processing
    + machine learning
    + info extraction
- Applications
  - Eigenrhythms: drum pattern models
  - FDLP temporal envelope models
  - Music Similarity Browsing
  - Transformation-based generative models
  - Personal audio analysis
- Also...
  - speech recognition, meeting recordings, ...