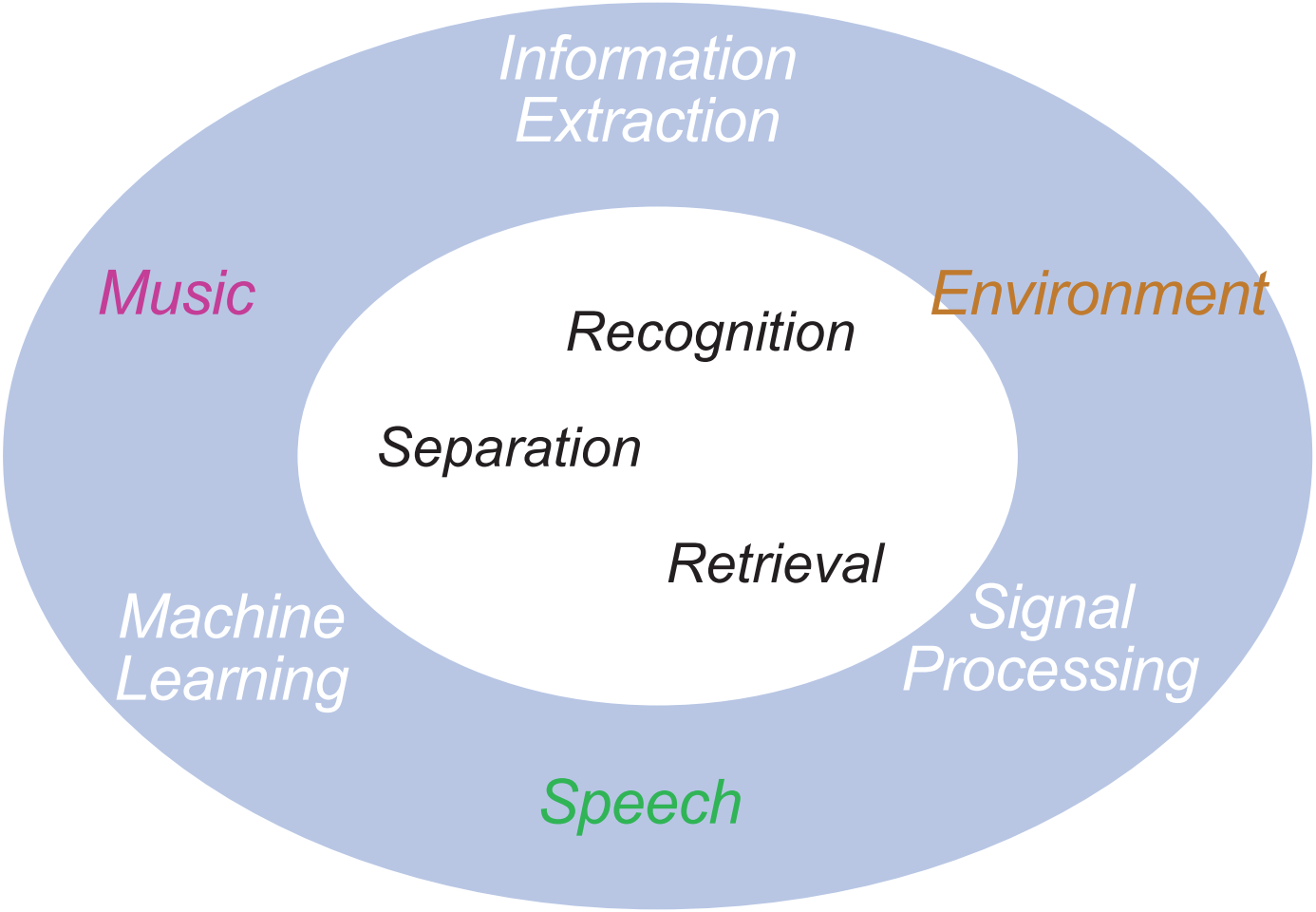
Some projects in Real-World Sound Analysis

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

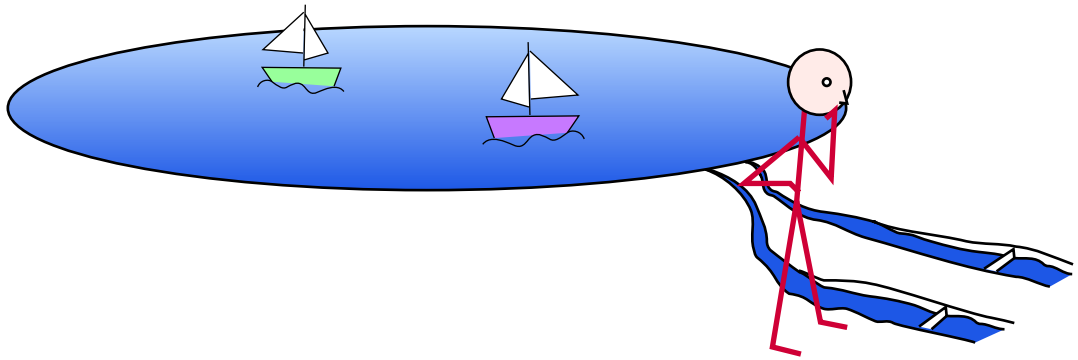
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Real-World Sound
2. Speech Separation
3. Soundtrack Classification
4. Music Audio Analysis

LabROSA Overview



1. Real-World Sound



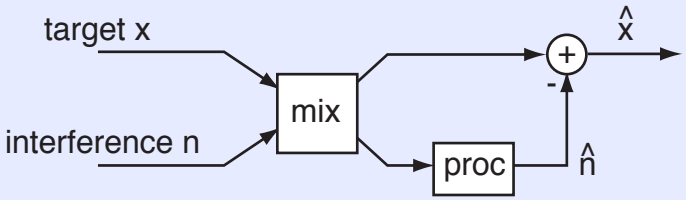
"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- Received waveform is a mixture
  - 2 sensors, N sources - underconstrained
- Use prior knowledge (models) to constrain

Approaches to Separation

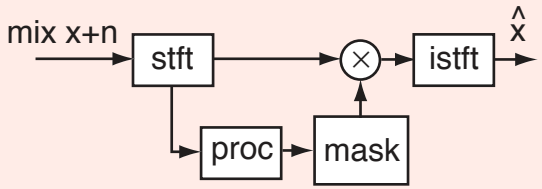
ICA

- Multi-channel
- Fixed filtering
- Perfect separation – maybe!



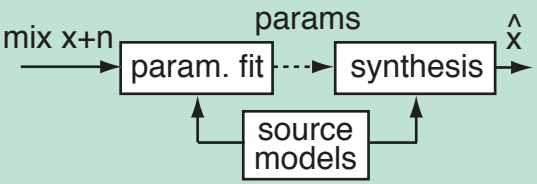
CASA

- Single-channel
- Time-var. filter
- Approximate separation



Model-based

- Any domain
- Param. search
- Synthetic output?



2. Speech Separation

Roweis '01, '03
Kristjannson '04, '06

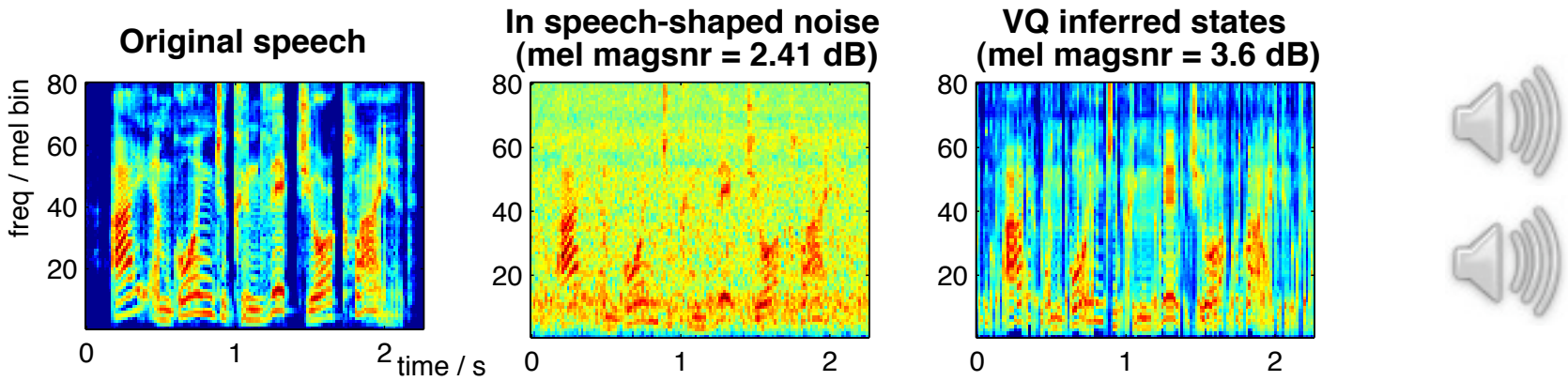
- Given models for sources, find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1,i_2) = \mathcal{N}(\mathbf{x};\mathbf{c}_{i1}+\mathbf{c}_{i2},\Sigma)$$ *combination model*

$$\{i_1(t),i_2(t)\} = argmax_{i_1,i_2} p(\mathbf{x}(t)|i_1,i_2)$$ *inference of source state*

  - can include sequential constraints...

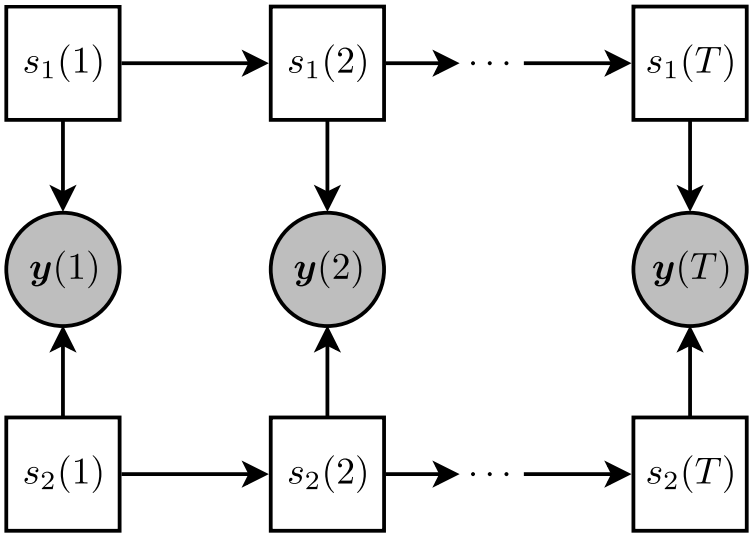
- E.g. stationary noise:

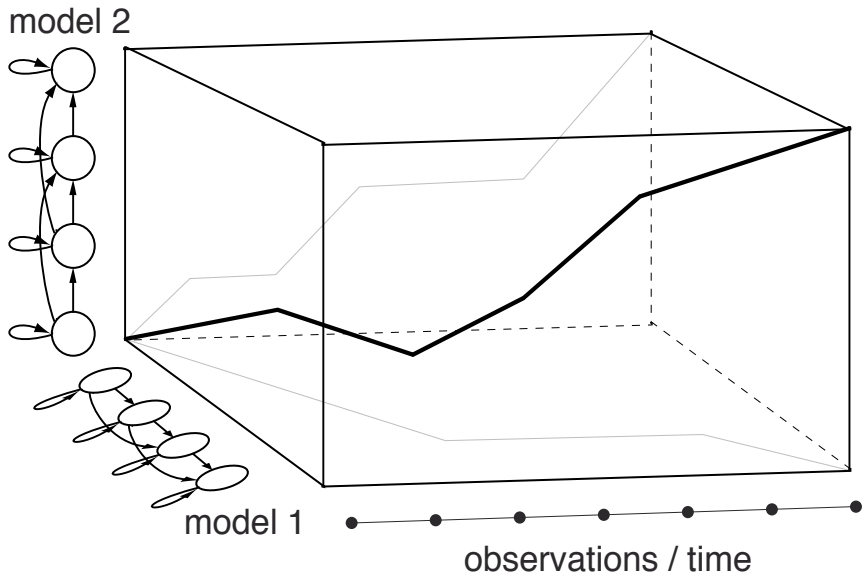


Speech Mixture Recognition

Varga & Moore '90
Kristjansson, Hershey et al. '06

- Speech recognizers contain speech models
  - ASR is just $\text{argmax}\ \text{P}(W \mid X)$
- Recognize mixtures with Factorial HMM
  - i.e. two state sequences, one model for each voice
  - exploit sequence constraints, speaker differences



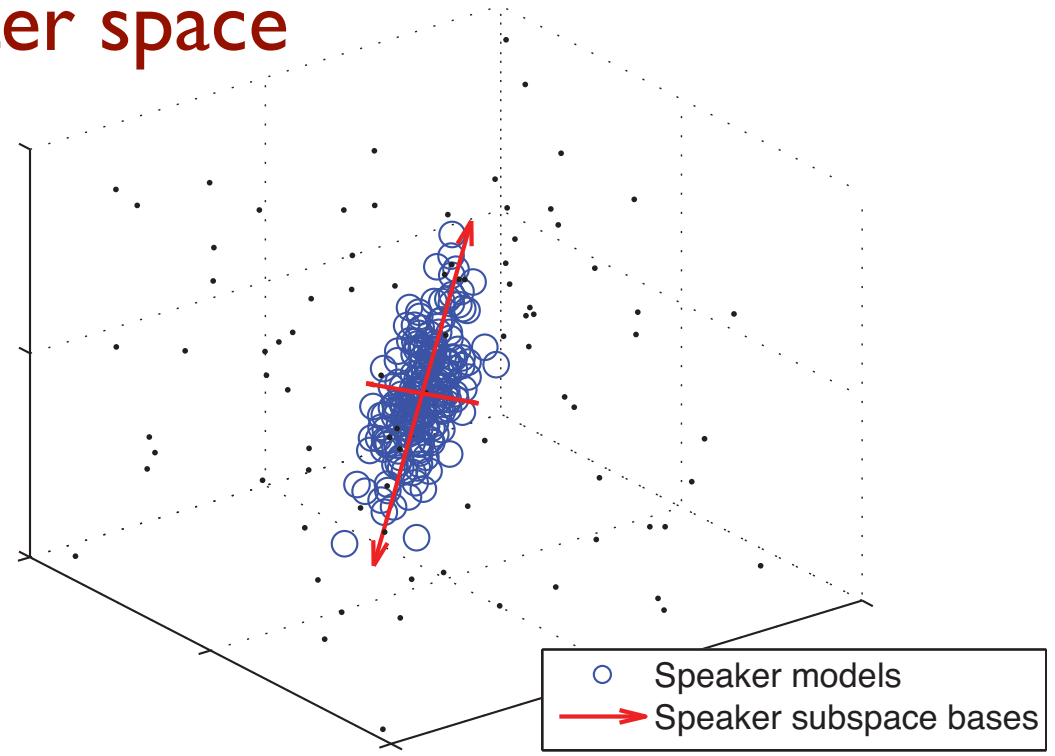


- Speaker-specific models...

Eigenvoices

Kuhn et al. '98, '00
Weiss & Ellis '07, '08, '09

- Idea: Find speaker model parameter space
  - generalize without losing detail?



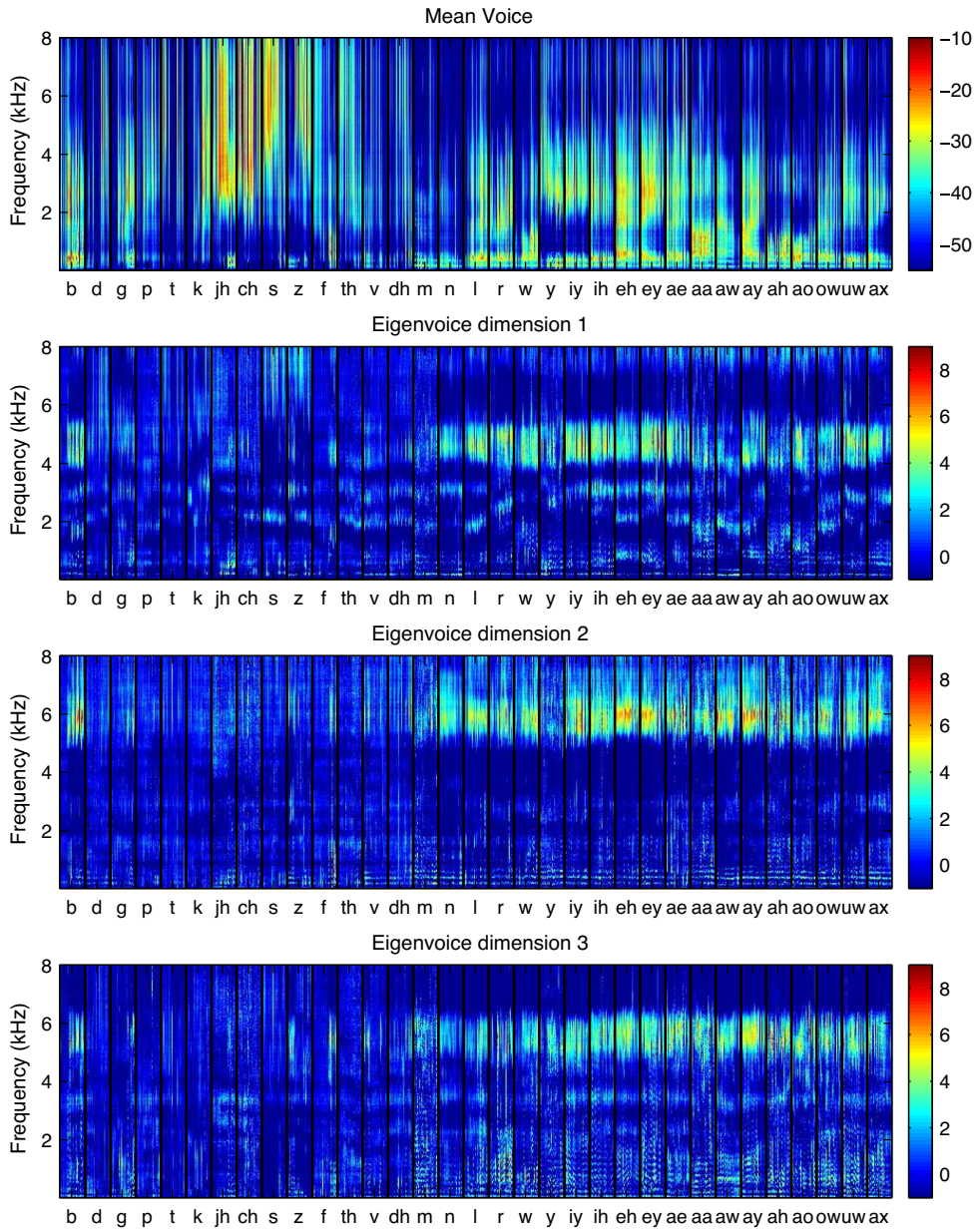
- Eigenvoice model:

$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} + U\,\mathbf{w} + B\,\mathbf{h}$$

adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

Eigenvoice Bases

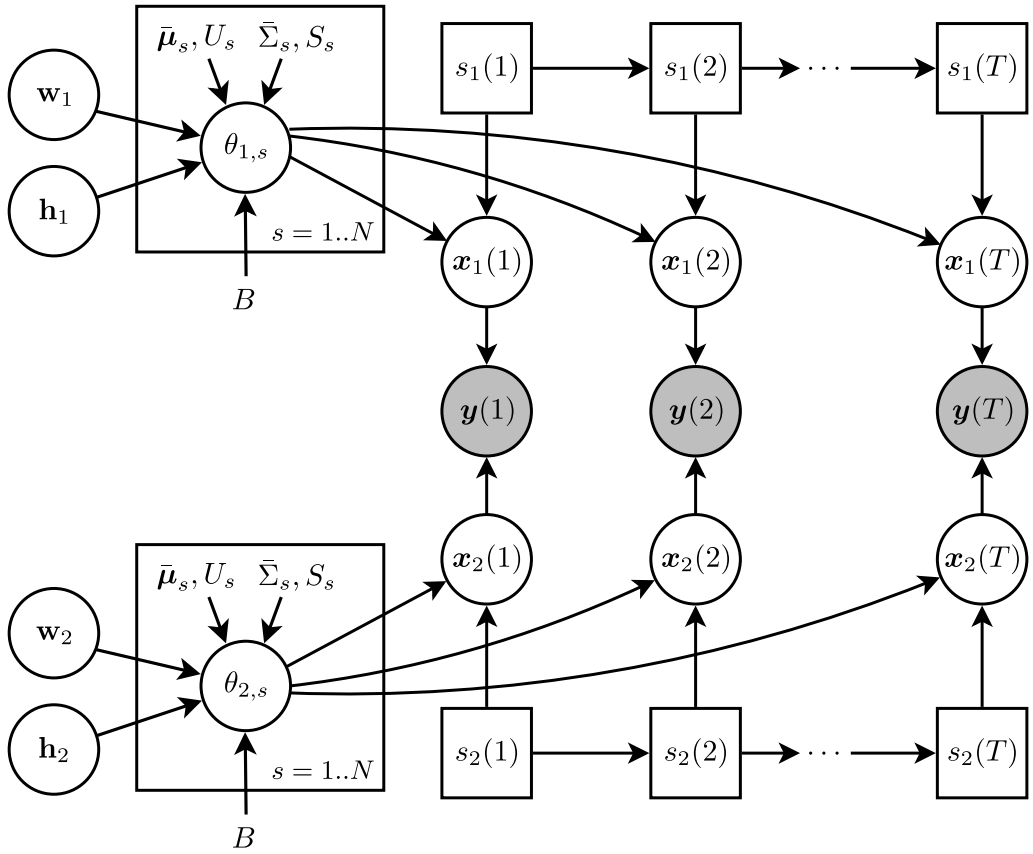
- **Mean model**
  - 280 states x 320 bins = 89,600 dimensions
- **Eigencomponents shift formants/ coloration**
  - additional components for channel



Speaker-Adapted Separation

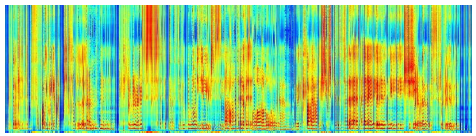
Weiss & Ellis '08

- Factorial HMM analysis
  with tuning of source model parameters
  = eigenvoice speaker adaptation

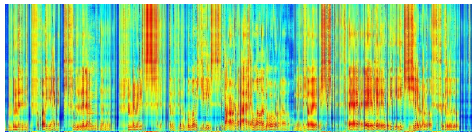


Speaker-Adapted Separation

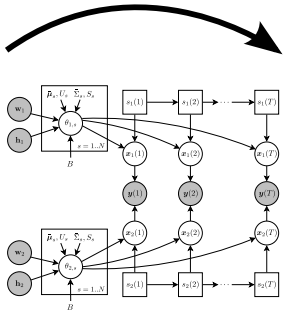
$$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$$

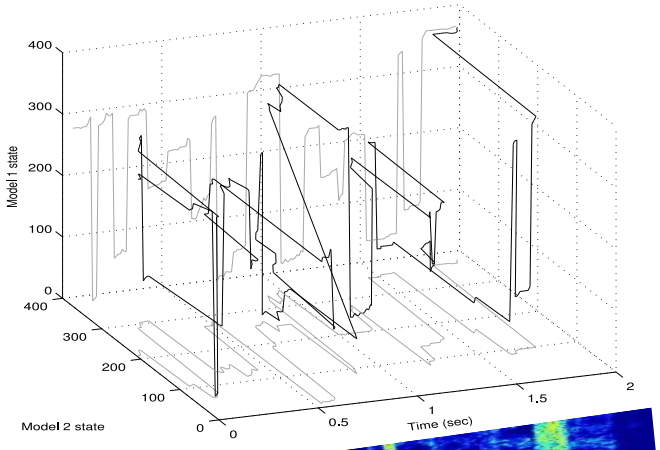


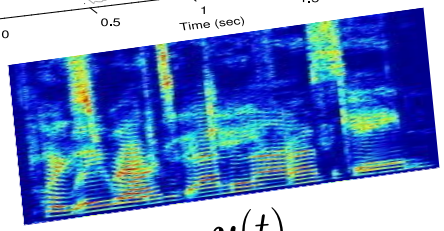
$$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$$



Find Viterbi path

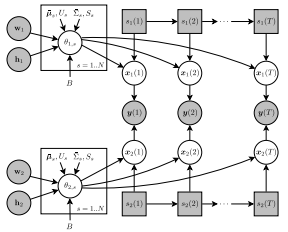




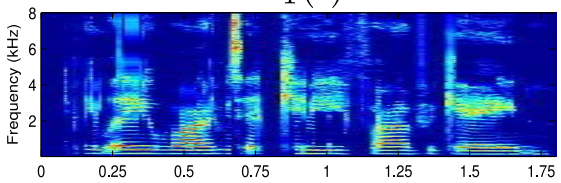


$\boldsymbol{y}(t)$

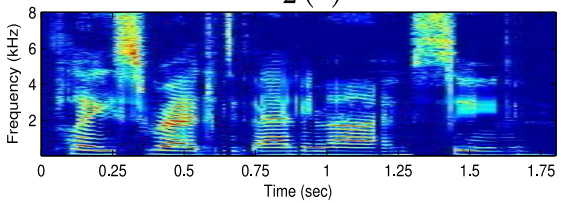
Estimate source signals



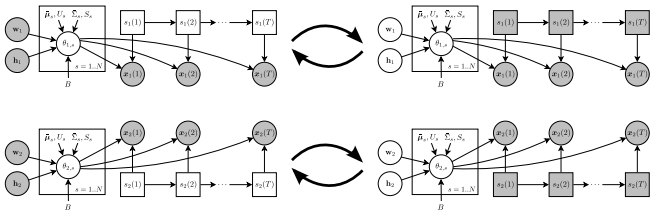
$\hat{\boldsymbol{x}}_1(t)$



$\hat{\boldsymbol{x}}_2(t)$

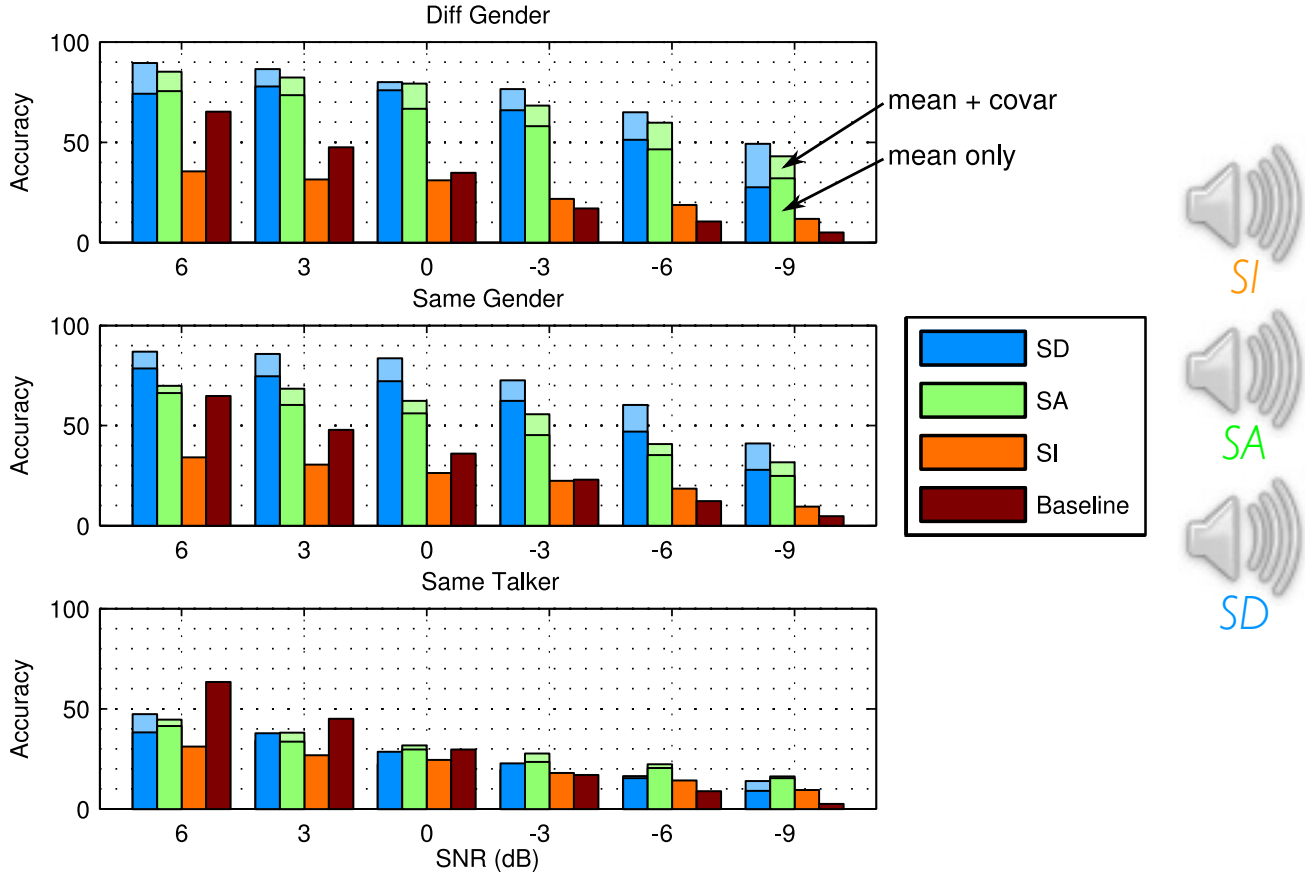


Update model parameters using EM algorithm from Kuhn et al., (2000)



Speaker-Adapted Separation

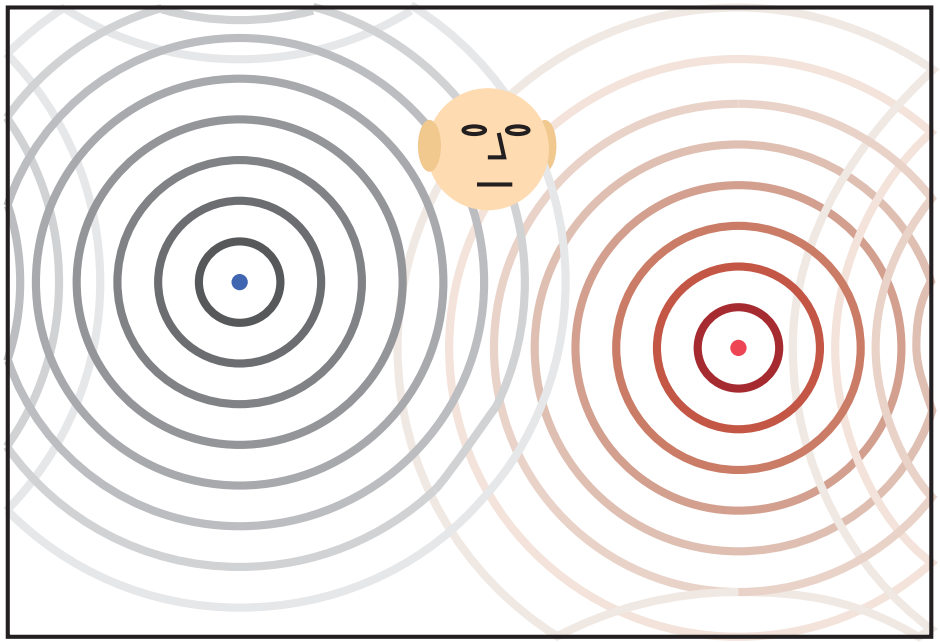
- Eigenvoices for Speech Separation task
  - speaker adapted (SA) performs midway between speaker-dependent (SD) & speaker-indep (SI)



Spatial Separation

Mandel & Ellis '07

- 2 or 3 sources in reverberation
  - assume just 2 'ears'

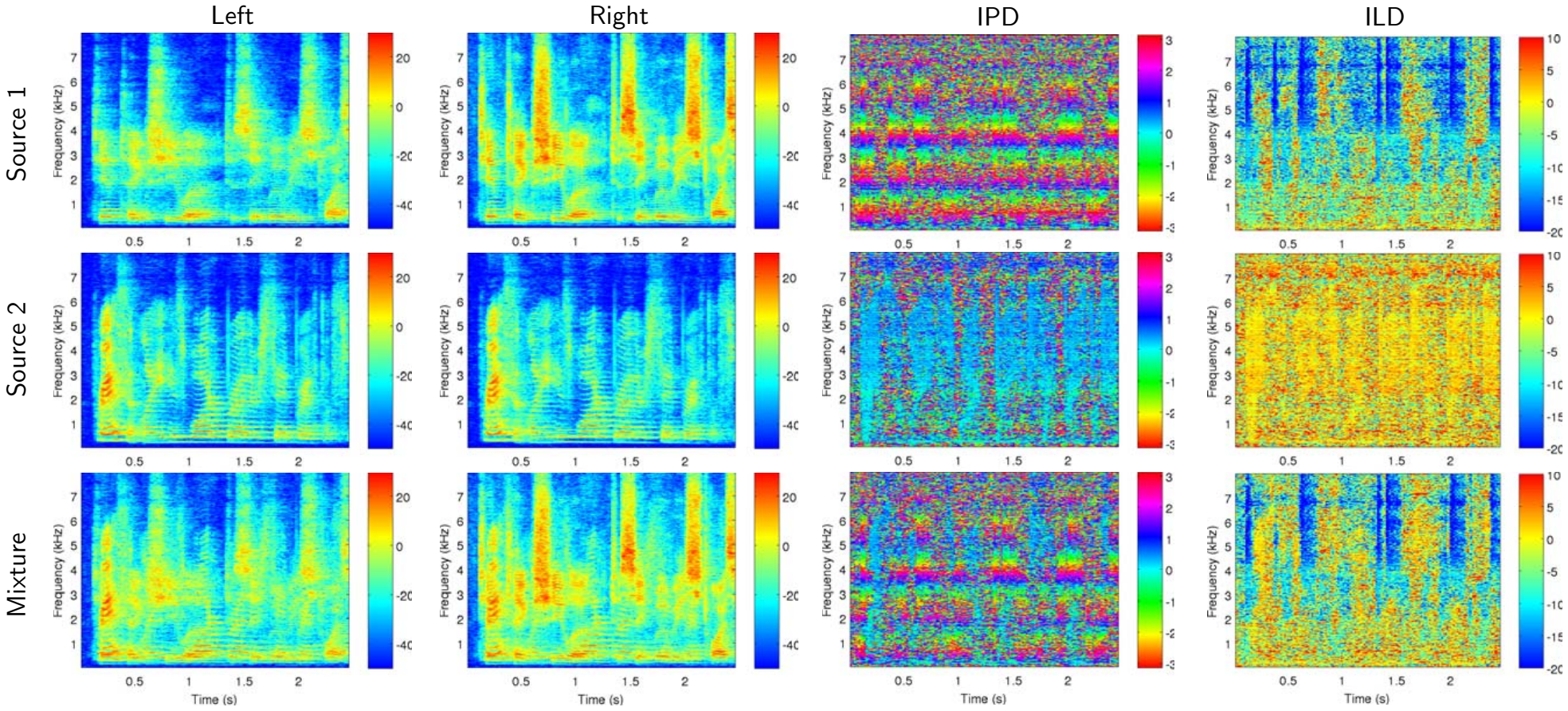


- Model interaural spectrum of each source as stationary level and time differences:

$$\frac{L(\omega,t)}{R(\omega,t)} = a(\omega)e^{j\omega\tau}N(\omega,t)$$

ILD and IPD

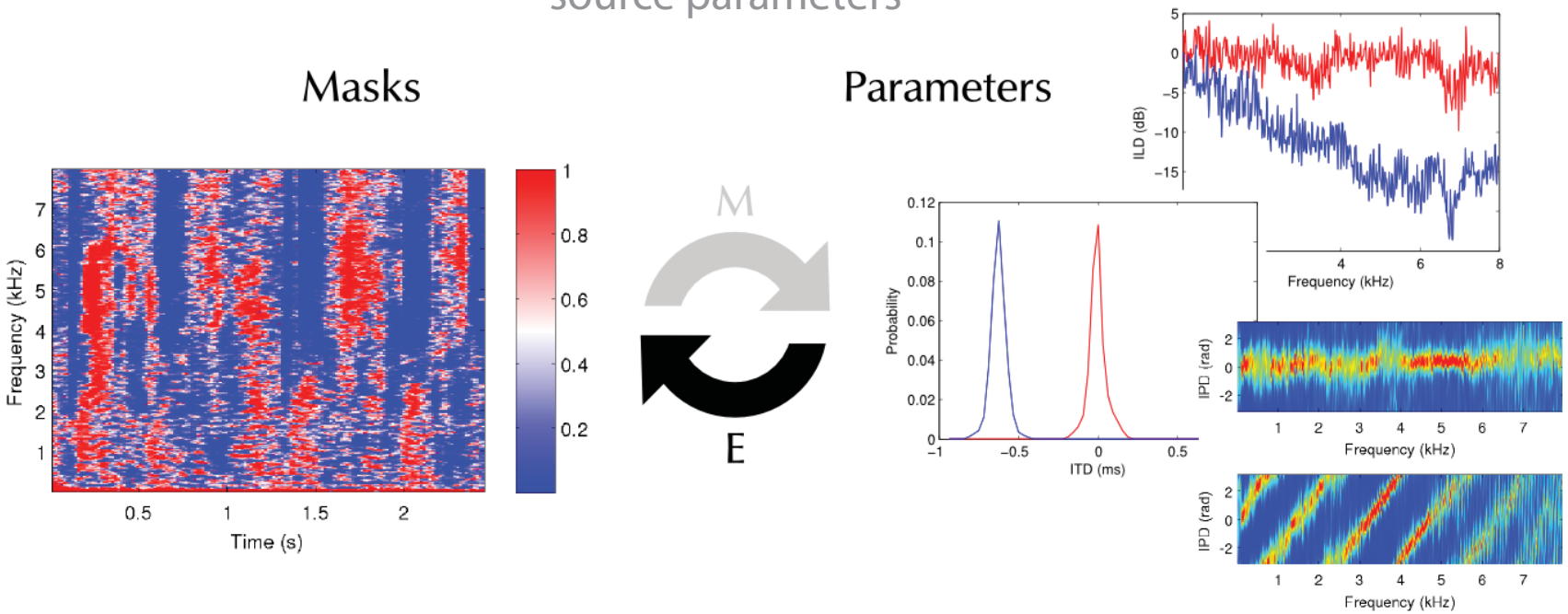
- Sources at 0° and 75° in reverb



Model-Based EM Source Separation and Localization (MESSL)

Mandel & Ellis '09

Re-estimate
source parameters

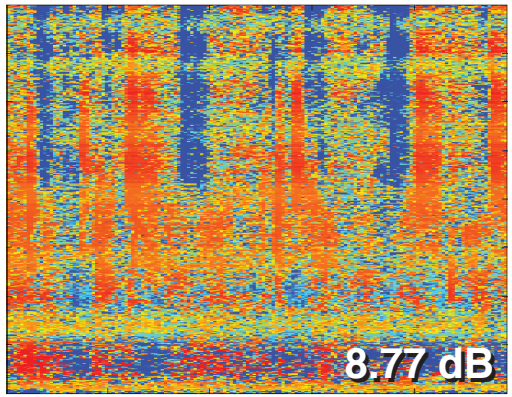


Assign spectrogram points
to sources

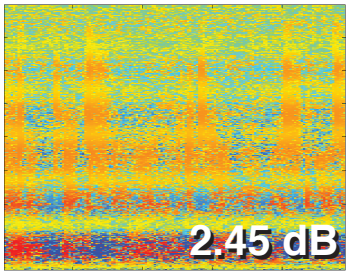
- can model more sources than sensors
- flexible initialization

MESSL Results

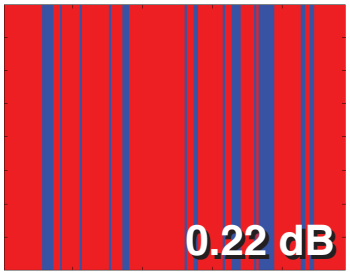
- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness



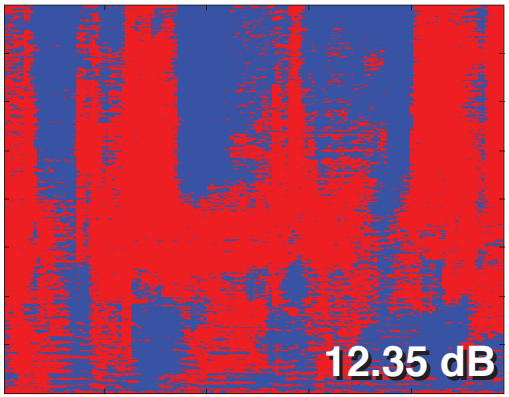
EM+ILD



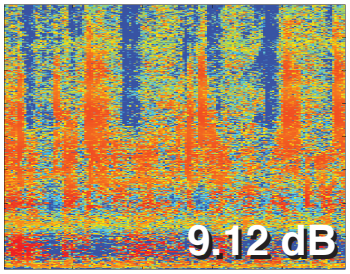
EM-ILD (only IPD)



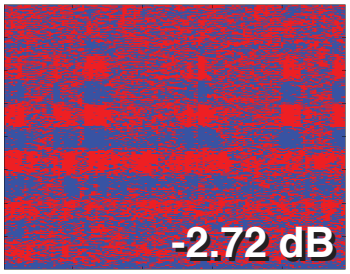
PHAT-histogram



Ground Truth



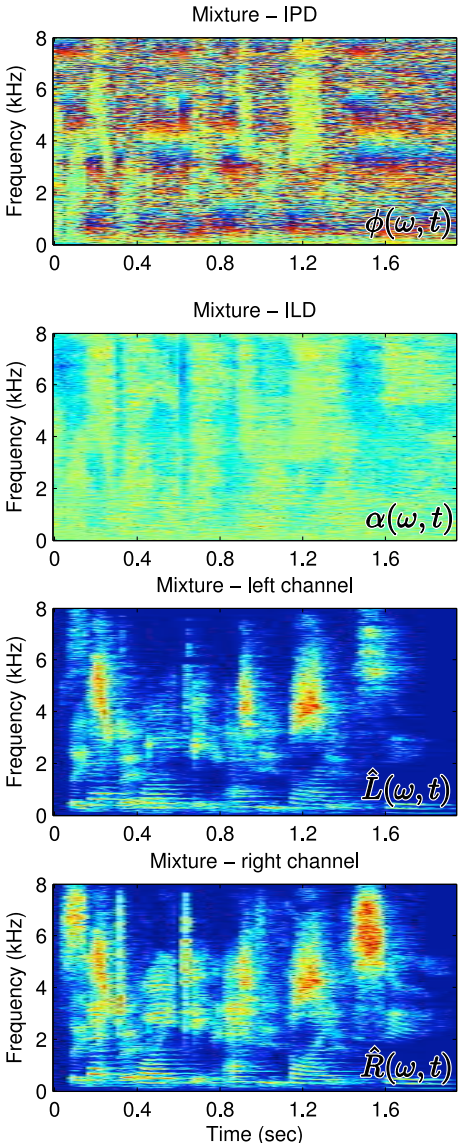
EM+1ILD (tied means)



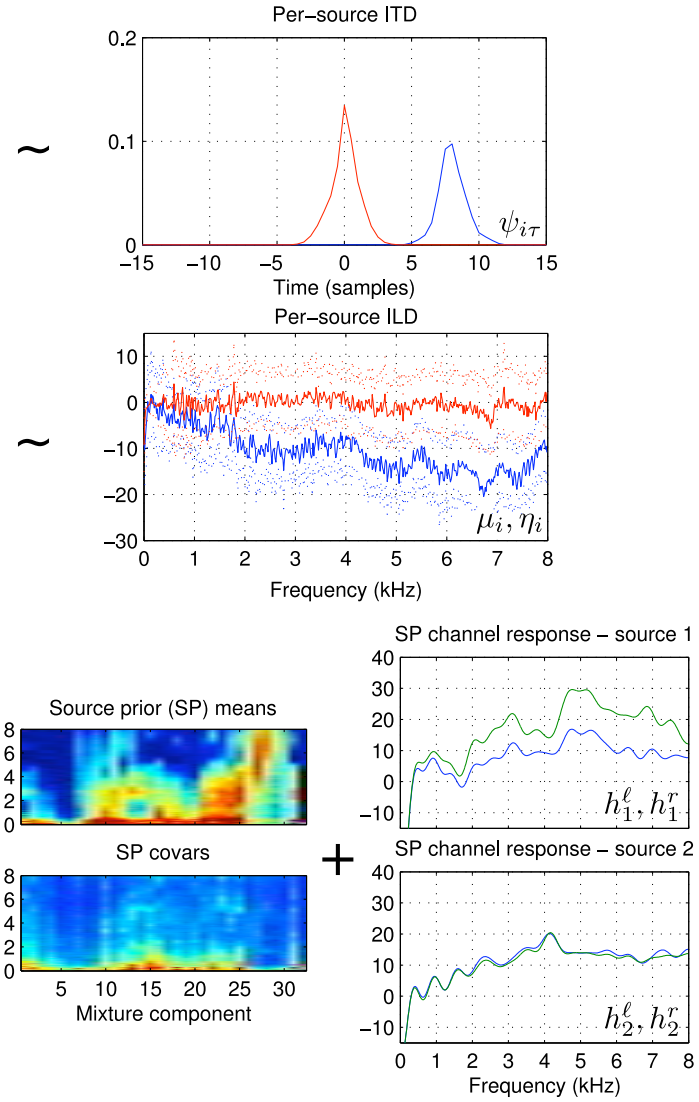
DUET

MESSL-SP (Source Prior)

Observations



Parameters

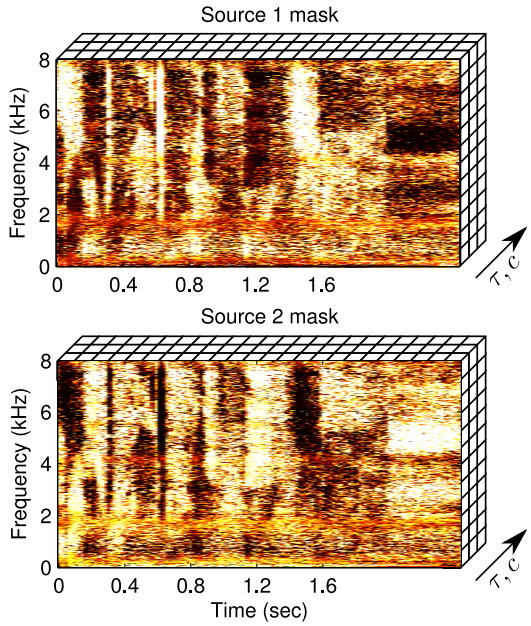


E-step
Use parameters to compute posteriors of hidden variables

M-step
Use posteriors to update parameters

Posteriors

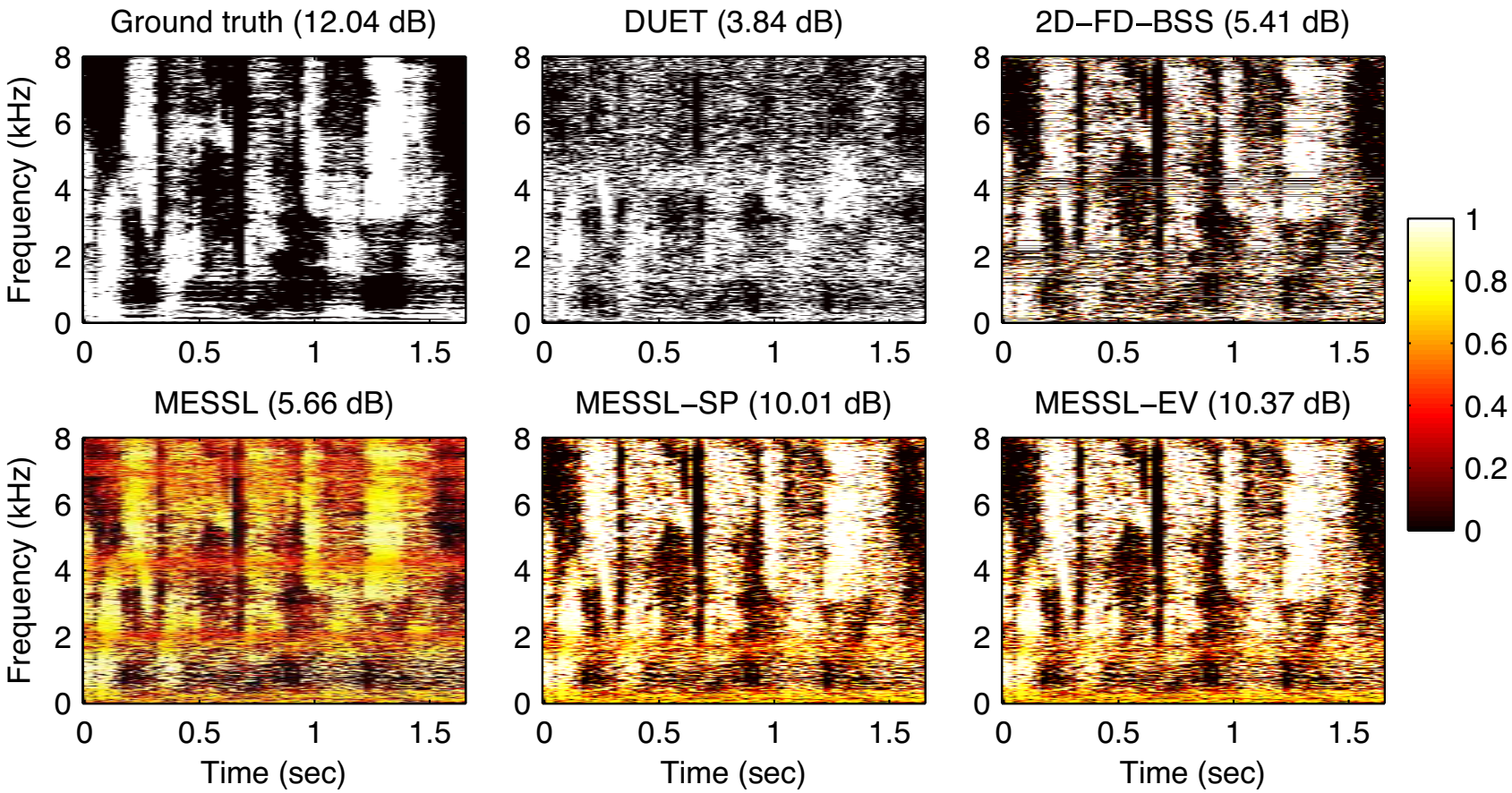
Each point in spectrogram is explained by a source, delay, and mixture component



Separate sources by multiplying mixture by different masks

MESSL-SP Results

- Source models function as priors
- Interaural parameter spatial separation
  - source model prior improves spatial estimate



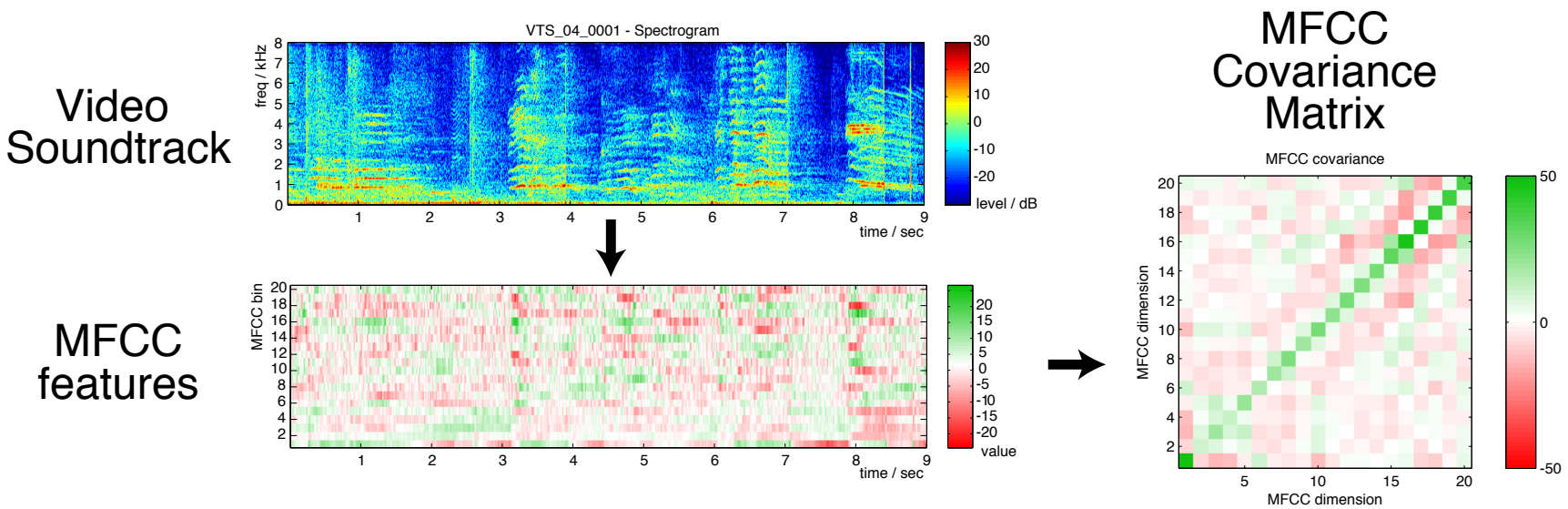
3. Soundtrack Classification

- Short video clips as the evolution of snapshots
  - 10-100 sec, one location, no editing
  - browsing?

- Need information for indexing...
  - video + audio
  - foreground + background

MFCC Covariance Representation

- Each clip/segment → fixed-size statistics
  - similar to speaker ID and music genre classification
- Full Covariance matrix of MFCCs
  - maps the kinds of spectral shapes present

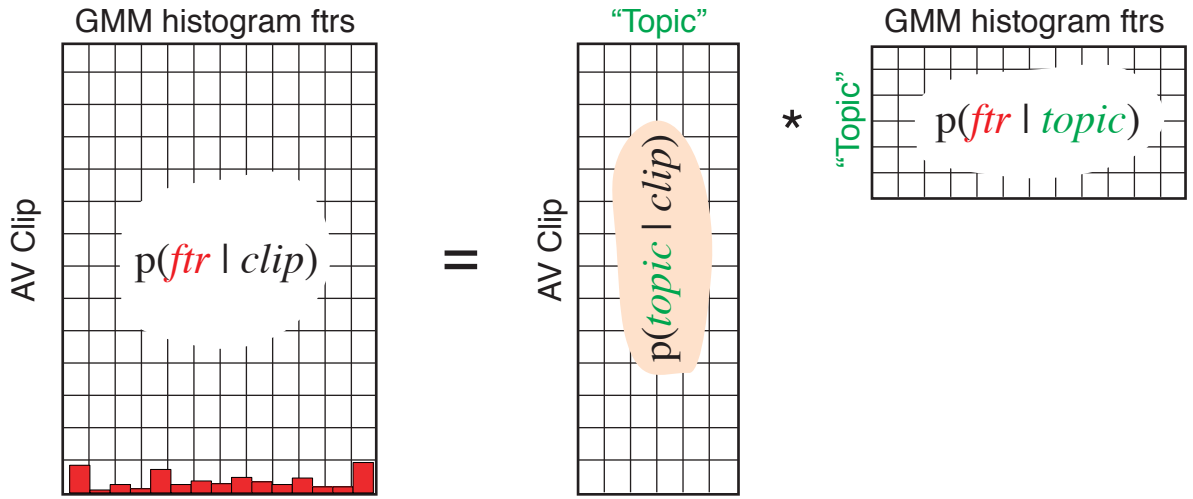


- Clip-to-clip distances for SVM classifier
  - by KL or 2nd Gaussian model

Latent Semantic Analysis (LSA)

Hofmann '99

- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - $p(ftr \mid clip) = \sum_{topics} p(ftr \mid topic)\, p(topic \mid clip)$

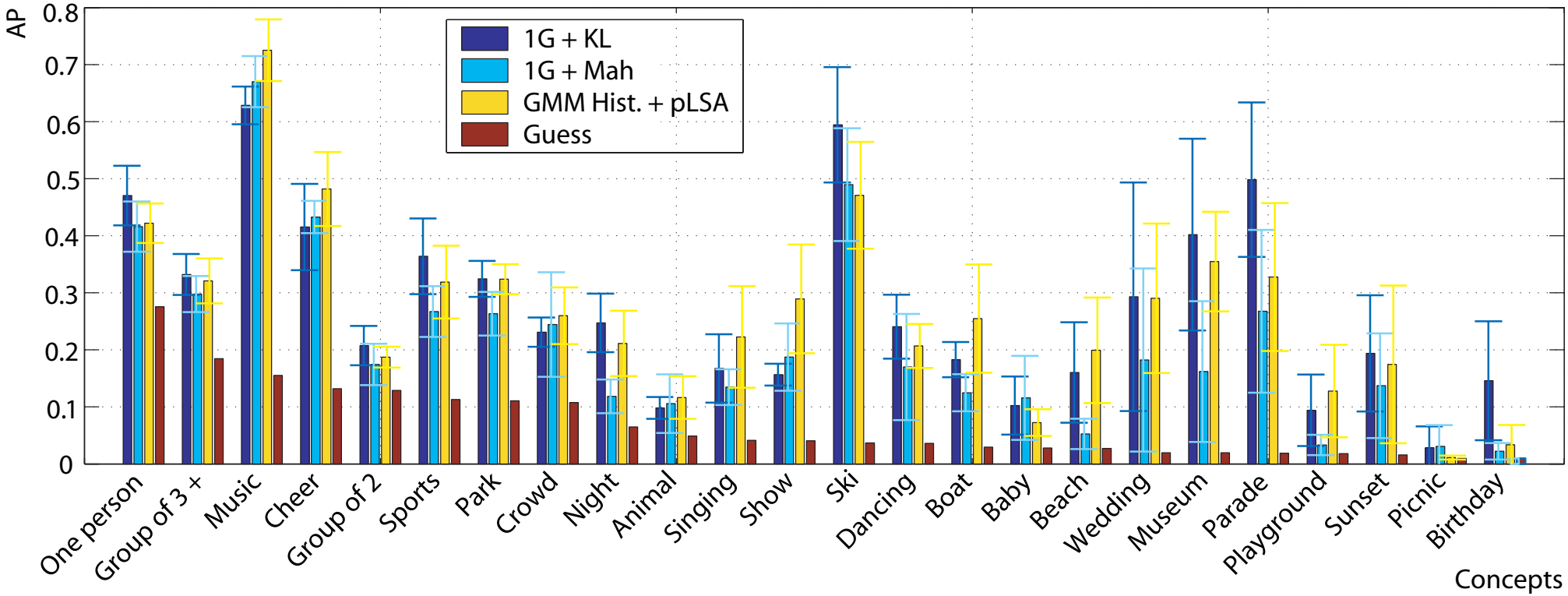


  - use $p(topic \mid clip)$ as per-clip features

Classification Results

Chang, Ellis et al. '07
Lee & Ellis '10

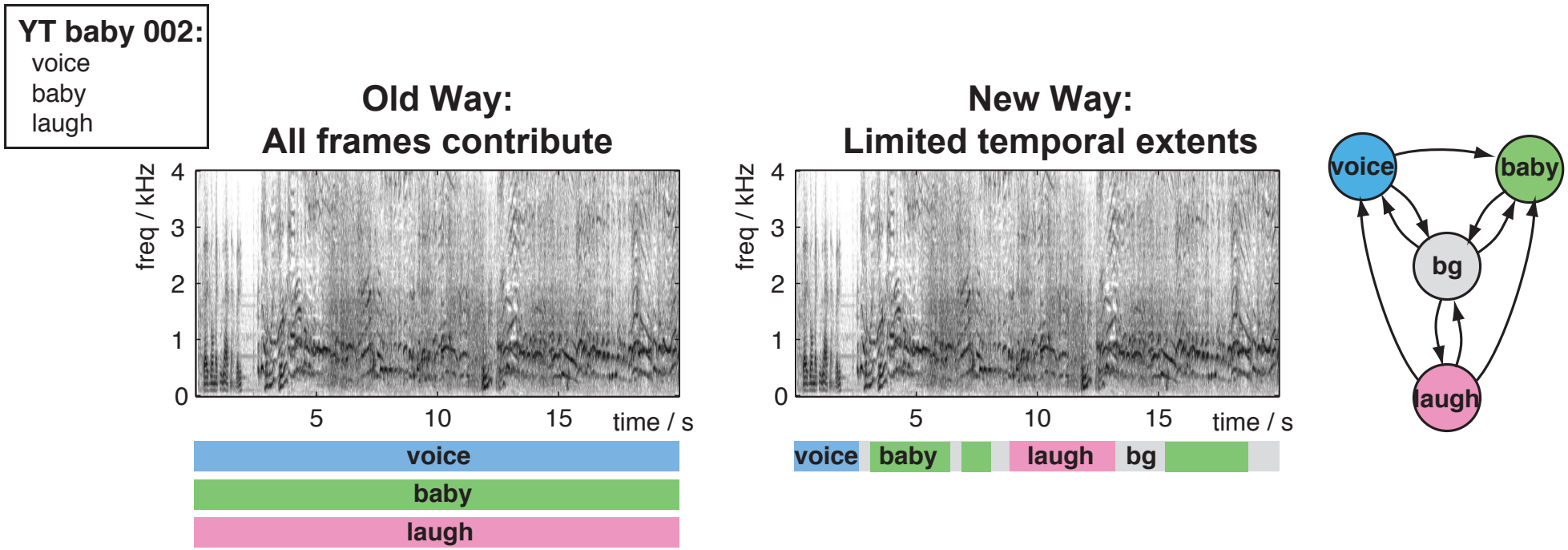
- Wide range:



- audio (music, ski) vs. non-audio (group, night)
- large AP uncertainty on infrequent classes

Temporal Refinement

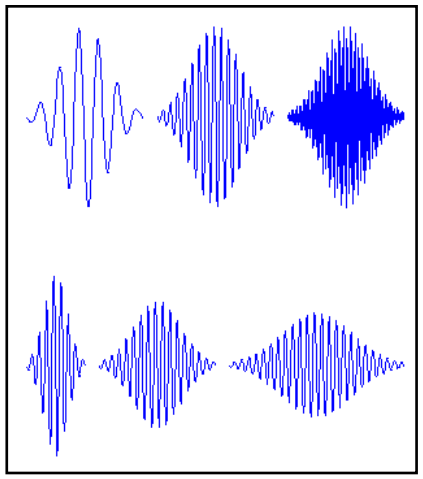
- Global vs. local class models
  - tell-tale acoustics may be 'washed out' in statistics
  - try iterative realignment of HMMs:

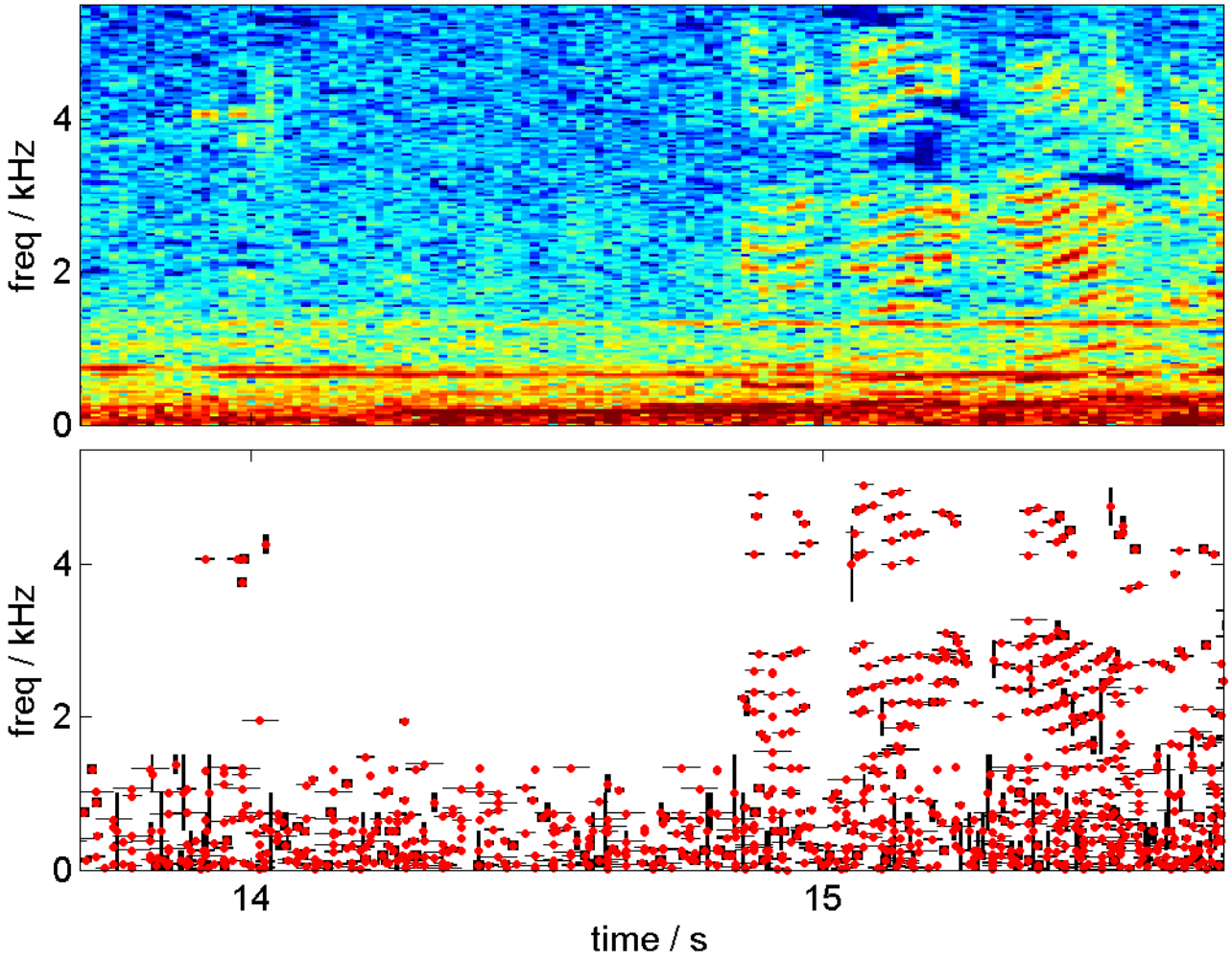


  - "background" (bg) model shared by all clips
  - Multiple-Instance Learning

Landmark-based Features

- Describe audio with biggest Gabor elements
  - extract with Matching Pursuit (MP)



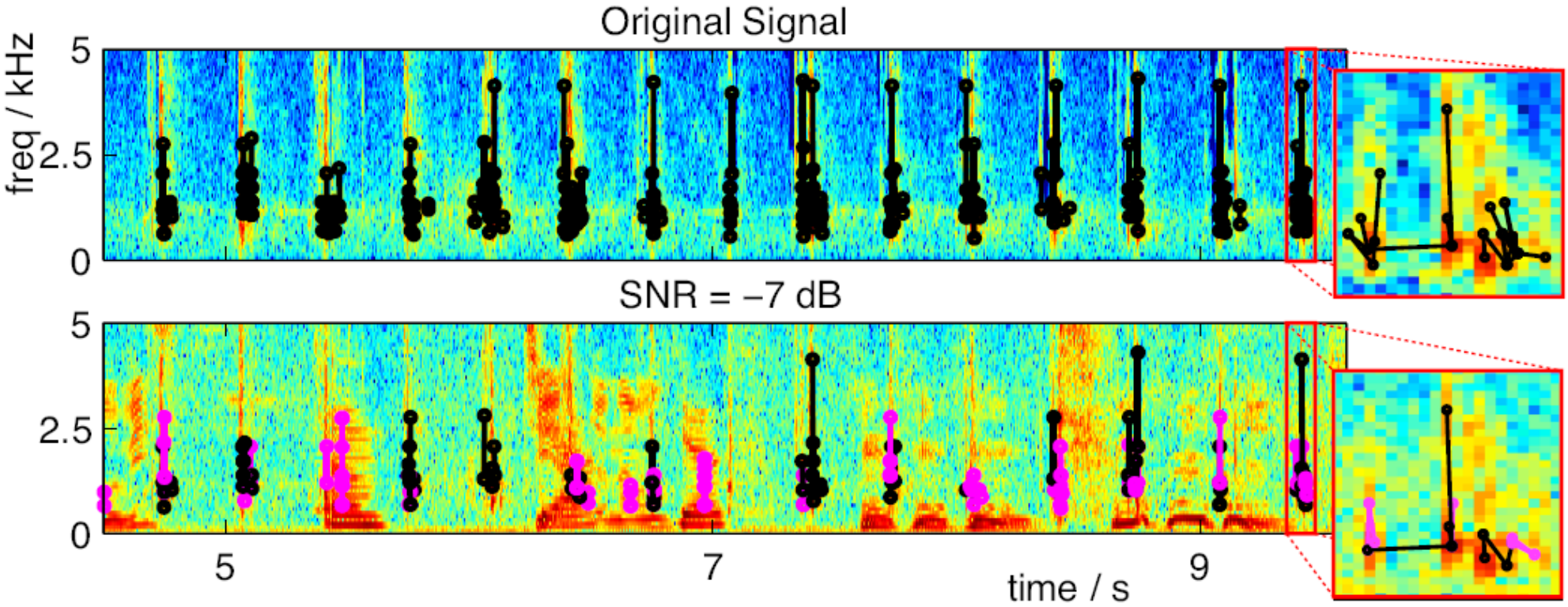


  - neighbor pairs as "features"?

Landmark models for similar events

Cotton & Ellis '09

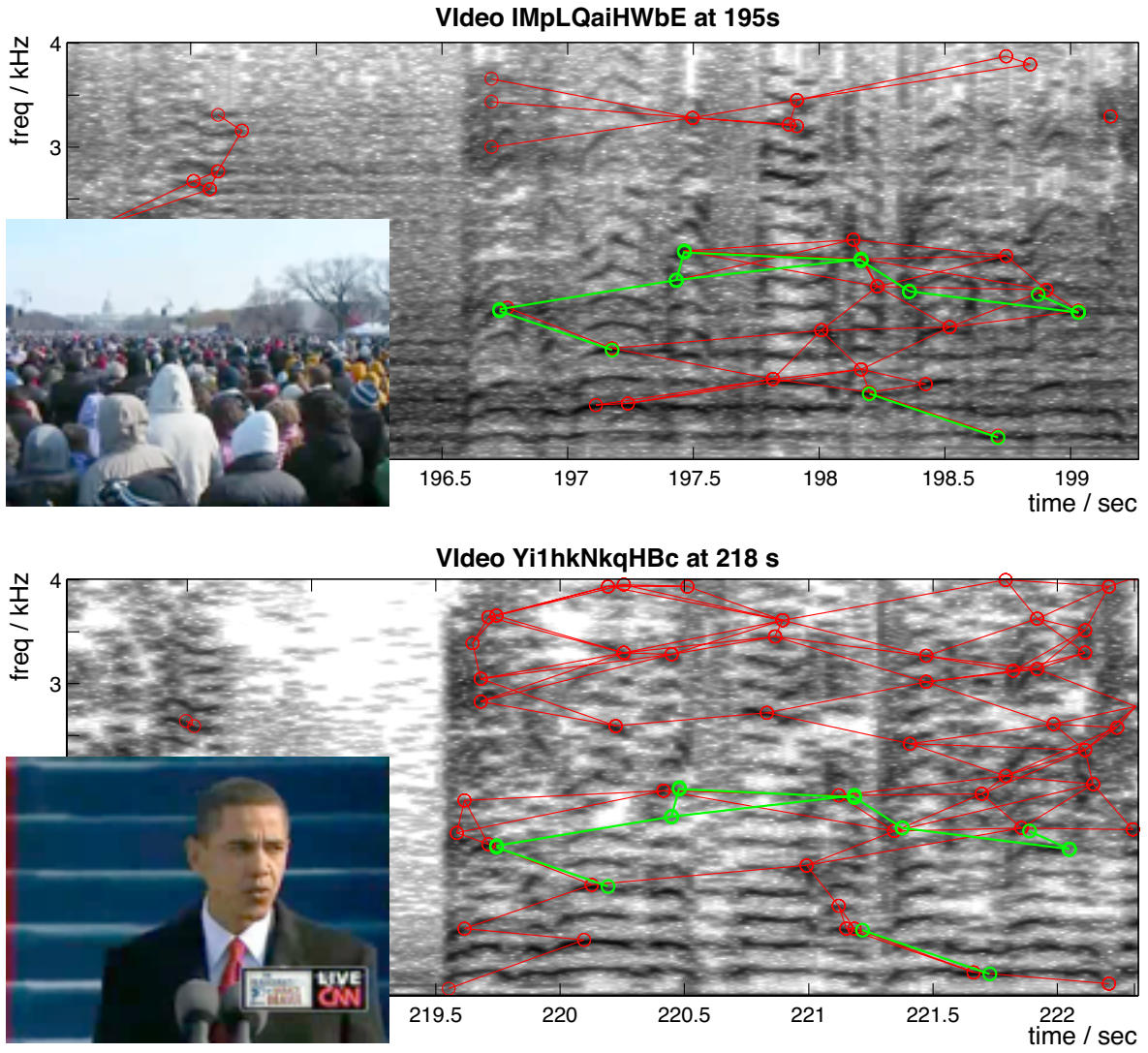
- **Build index of Gabor neighbor pairs**
  - recognize repeated events with similar pairs



Matching Videos via Fingerprints

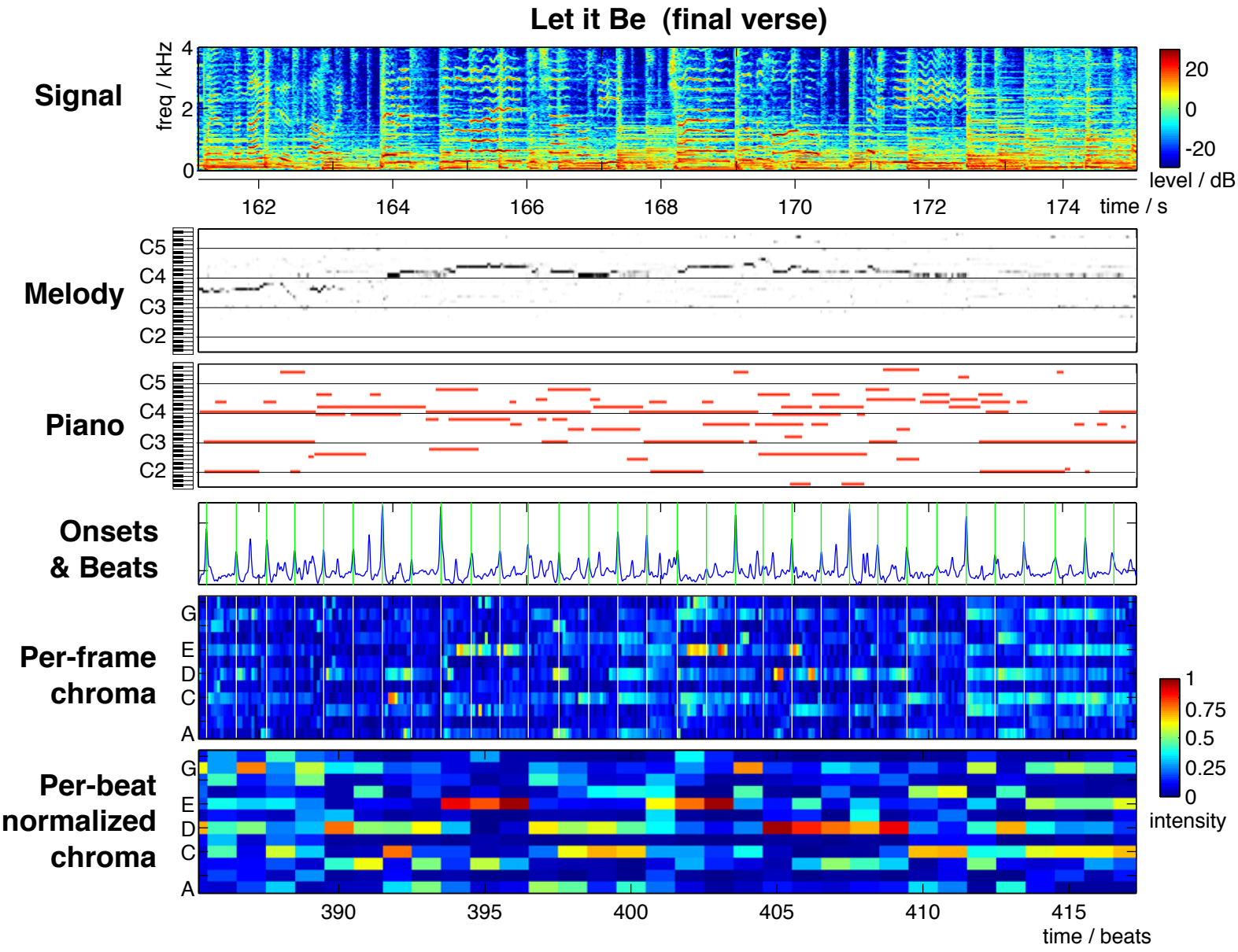
Cotton & Ellis '10?

- Landmark pairs are a noise-robust fingerprint
- Use to match distinct videos with same sound ambience



4. Music Audio Analysis

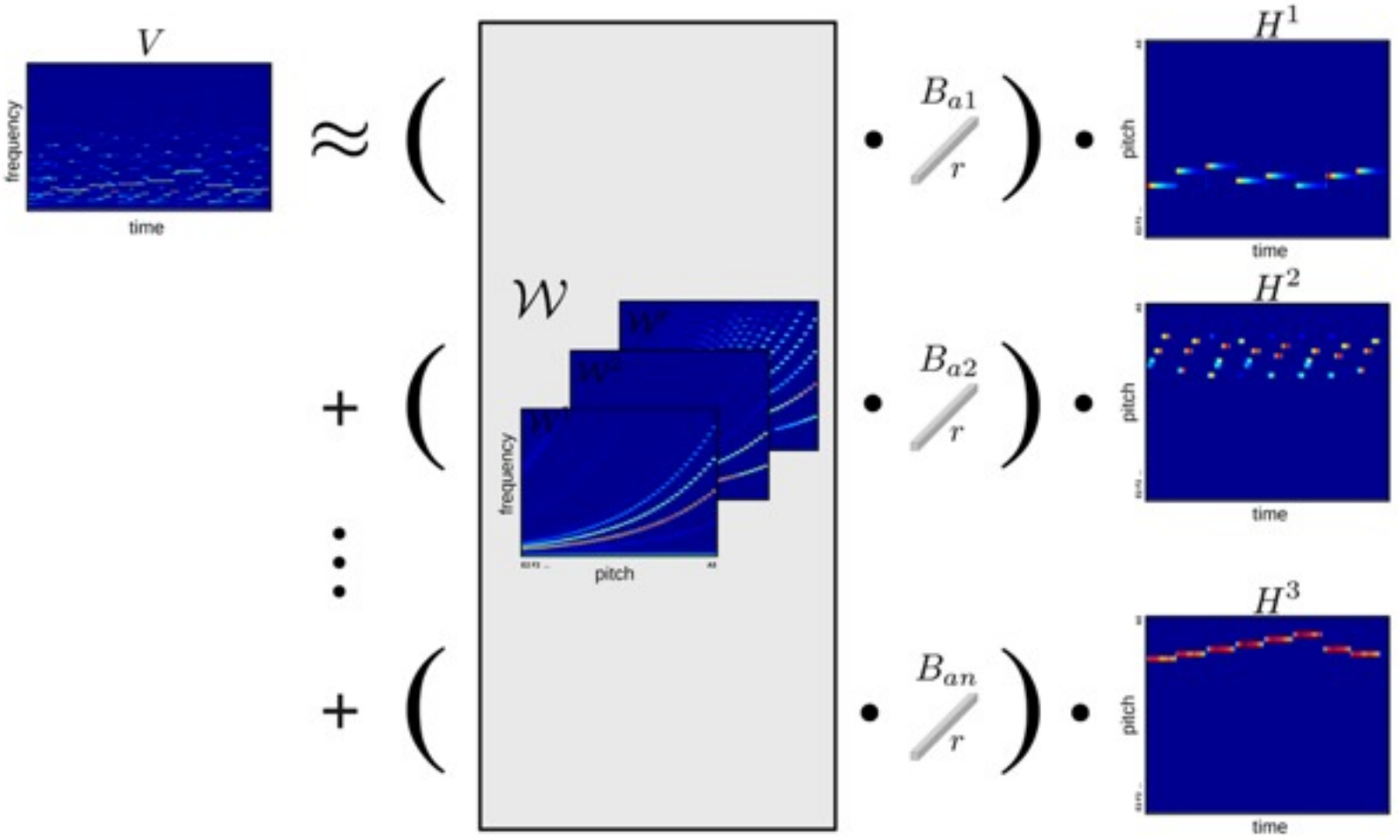
- Interested in all levels from notes to genres



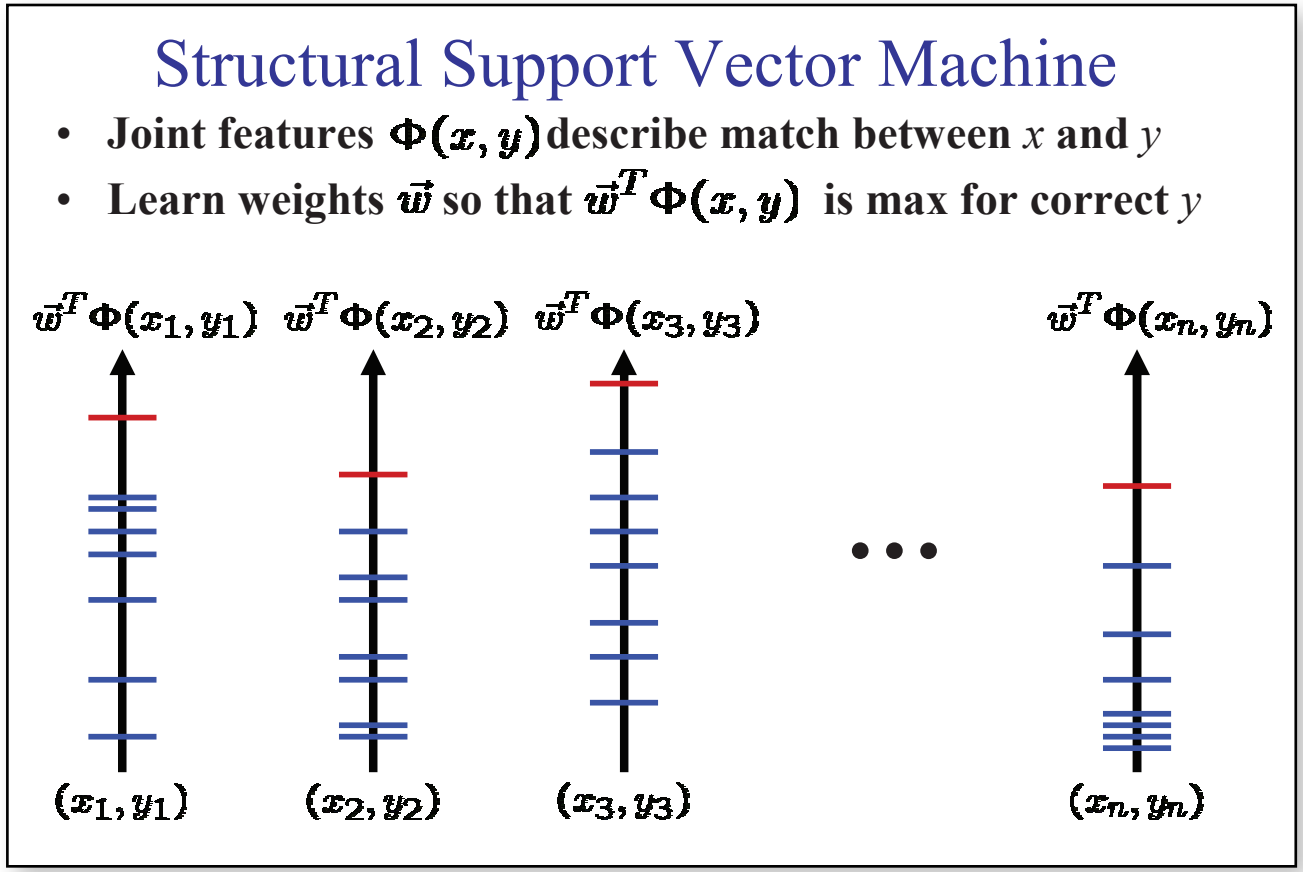
Polyphonic Transcription

Grindlay & Ellis '09

- Apply the Eigenvoice idea to music
  - eigeninstruments?
  - Subspace NMF



Chord Recognition



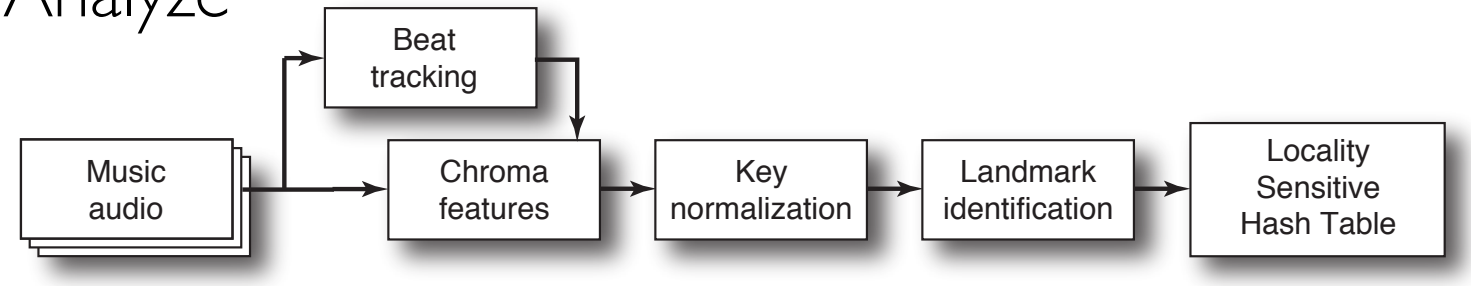
Tsochantaridis et al '05
Weller, Ellis, Jebara '09

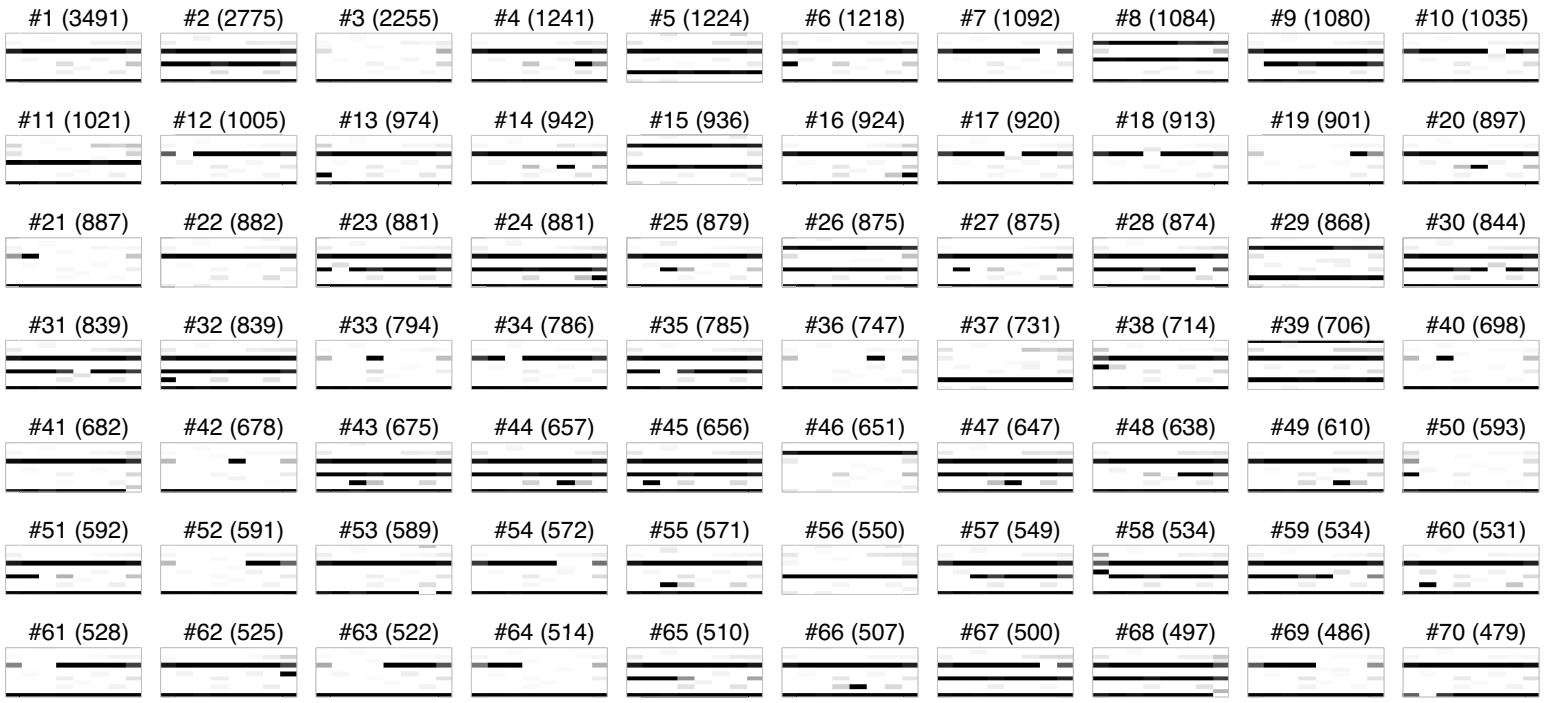
- Much better results

	RUSUSL	WEJ1	WEJ2	WEJ3	WEJ4
Wt. Ave. Overlap Score	0.701	0.704	0.723	0.723	0.742
Wt. Ave. Overlap Score (Merged maj/min)	0.760	0.743	0.762	0.760	0.777

Melodic-Harmonic Mining

- 100,000 tracks from morecowbell.dj
  - as Echo Nest Analyze
- Frequent clusters of 12 x 8 binarized event-chroma





Summary

- LabROSA : getting information from sound
- Speech
  - monaural separation using eigenvoices
  - binaural + reverb using MESSL
- Environmental
  - classification of consumer video
  - landmark-based events and matching
- Music
  - transcription of notes, chords, ...
  - large corpus mining