Environmental Sound Recognition and Classification

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. Machine Listening
2. Background Classification
3. Foreground Event Recognition
4. Speech Separation
5. Open Issues

Laboratory for the Recognition and Organization of Speech and Audio

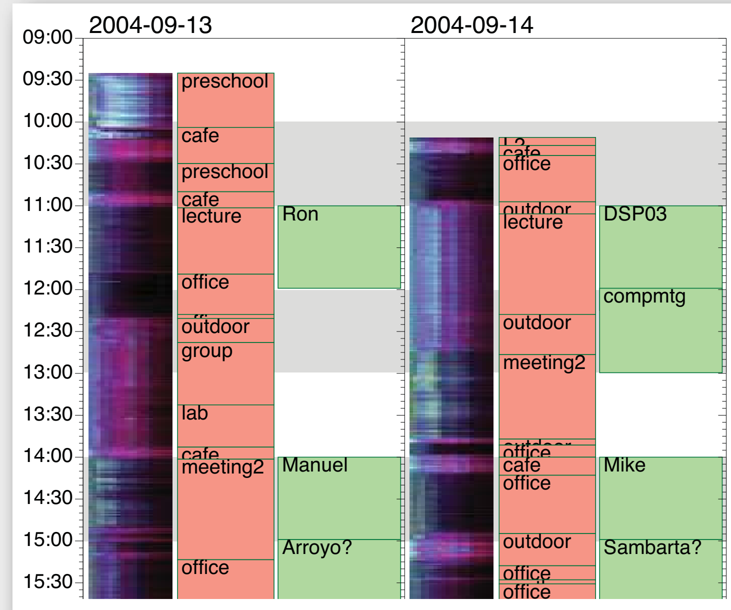
1. Machine Listening

- Extracting useful information from sound
  - ... like animals do

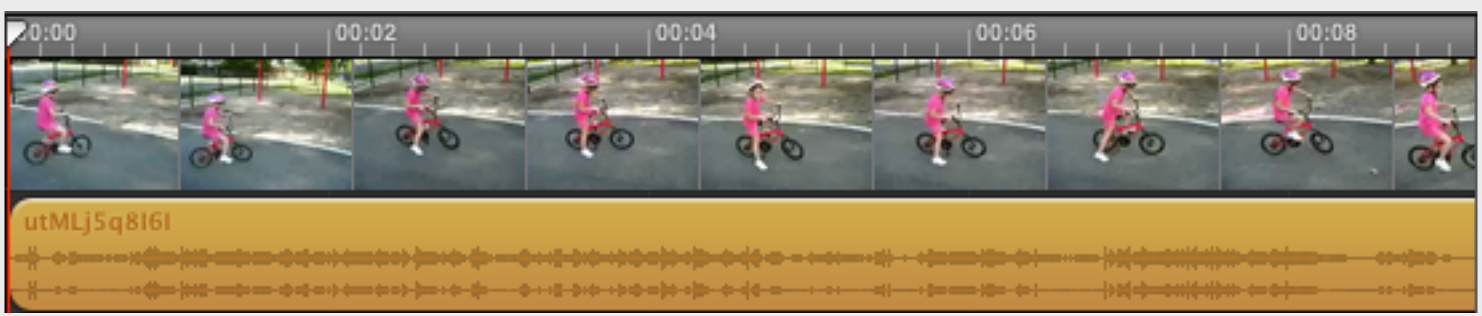
Task \ Domain	Environmental Sound	Speech	Music
Describe	Automatic Narration	Emotion	Music Recommendation
Classify	Environment Awareness	ASR	Music Transcription
Dectect	"Sound Intelligence"	VAD	Speech/Music

Environmental Sound Applications

- Audio Lifelog Diarization

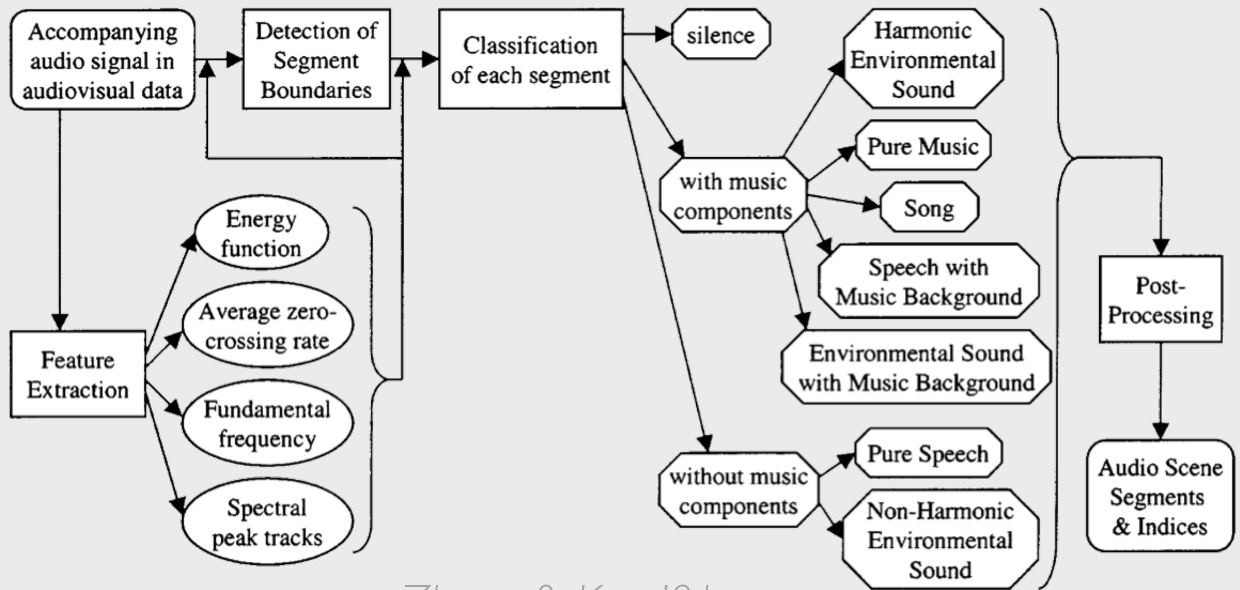


- Consumer Video Classification & Search

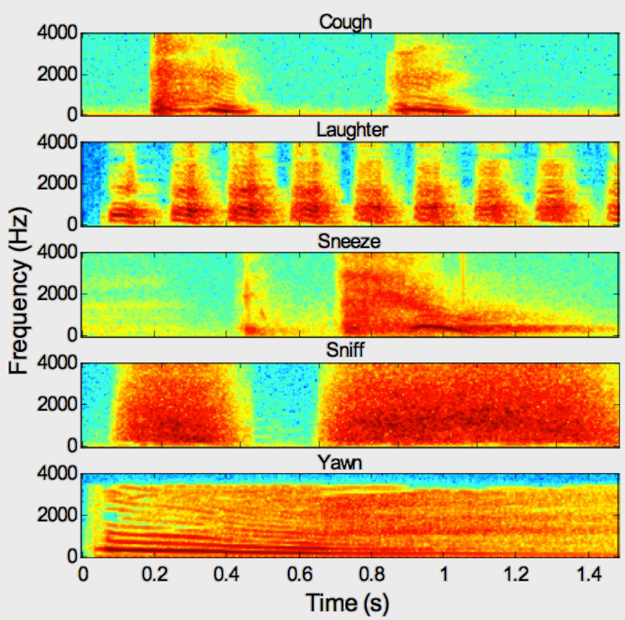


Prior Work

- Environment Classification
  - speech/music/silent/machine



Zhang & Kuo '01

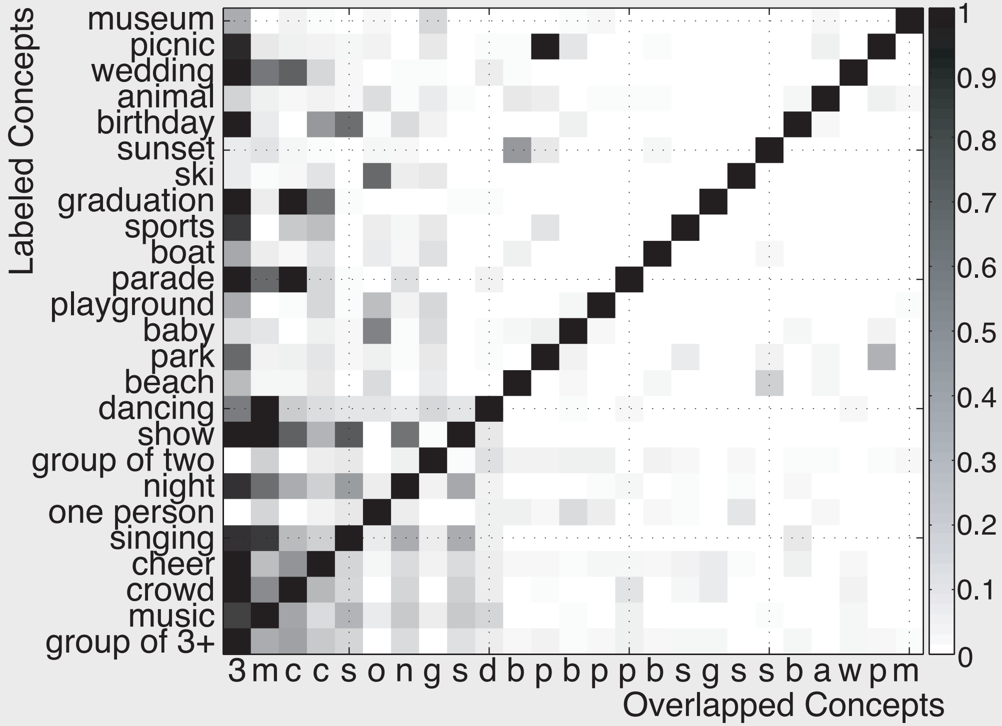


Temko & Nadeu '06

- Nonspeech Sound Recognition
  - Meeting room Audio Event Classification
  - sports events - cheers, bat/ball sounds, ...

Consumer Video Dataset

- 25 "concepts" from Kodak user study
  - boat, crowd, cheer, dance, ...



- Grab top 200 videos from YouTube search
  - then filter for quality, unedited = 1873 videos
  - manually relabel with concepts

Obtaining Labeled Data

Y-G Jiang et al. 2011

- Amazon Mechanical Turk
  - 10s clips
  - 9,641 videos in 4 weeks

Mark all the categories that appear in any part of the video.

Description:

- Watch the entire video as more categories may appear over time.
- Mark all the categories that appear in any part of the video.
- Make sure the audio is on.
- If no matching category is found, mark the box in front of "None of the categories matches".
- For categories that appears to be relevant but you're not completely sure, please still mark it.
- Please move over or click on the category name for detailed description.

Sport	Animal	Celebration	Others
☐ Basketball	☐ Cat	☐ Graduation	☐ Music Performance
☐ Baseball	☐ Dog	☐ Birthday	☐ Non-music Performance
☐ Soccer	☐ Bird	☐ Wedding Reception	☐ Parade
☐ Ice Skate		☐ Wedding Ceremony	☐ Beach
☐ Ski		☐ Wedding Dance	☐ Playground

☐ Swim ☐ None of the categories matches.

☐ Biking ☐ I don't see any video playing.

Current Time: 10 sec

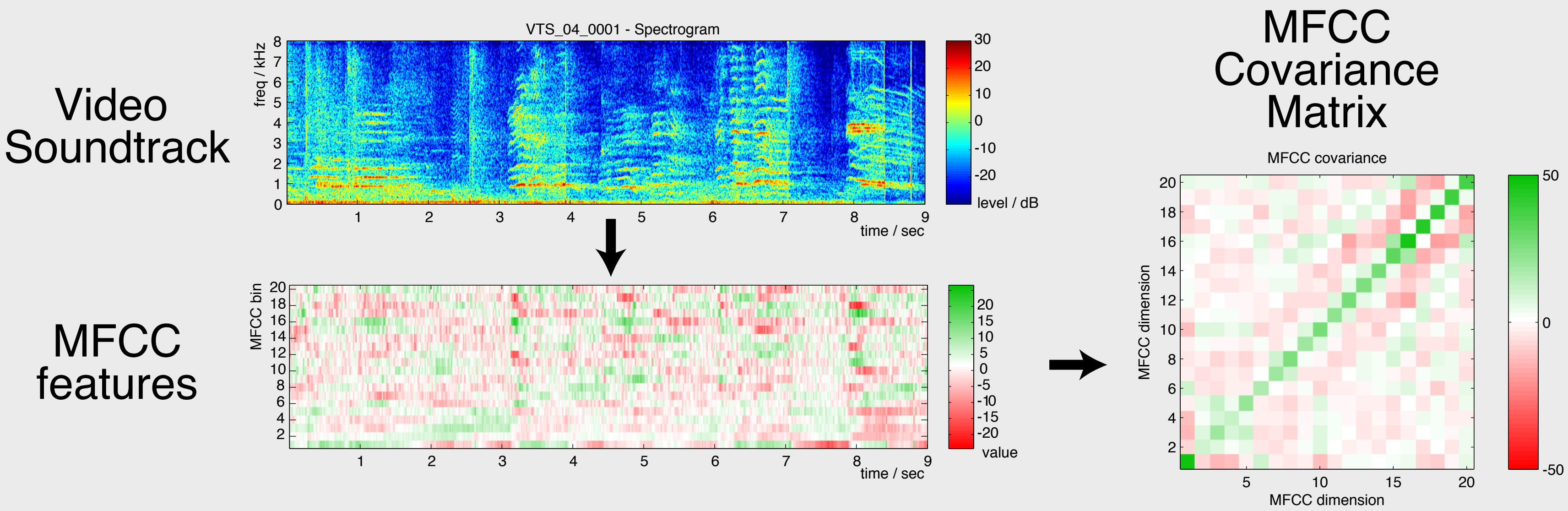
Submit

Replay Continue Playing

Original URL: http://www.youtube.com/watch?v=u_2dqWBd1L0

2. Background Classification

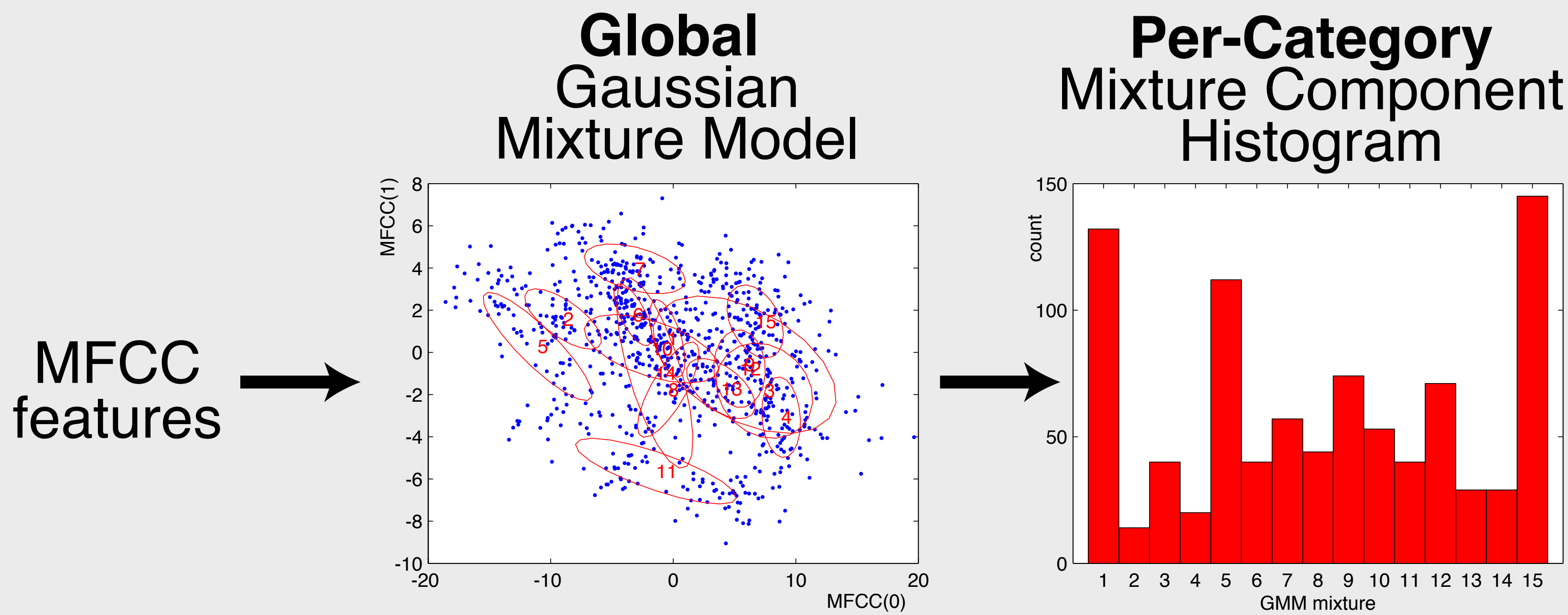
- Baseline for soundtrack classification
  - divide sound into short frames (e.g. 30 ms)
  - calculate features (e.g. MFCC) for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = "bag of features"



- Classify by e.g. KL distance + SVM

Codebook Histograms

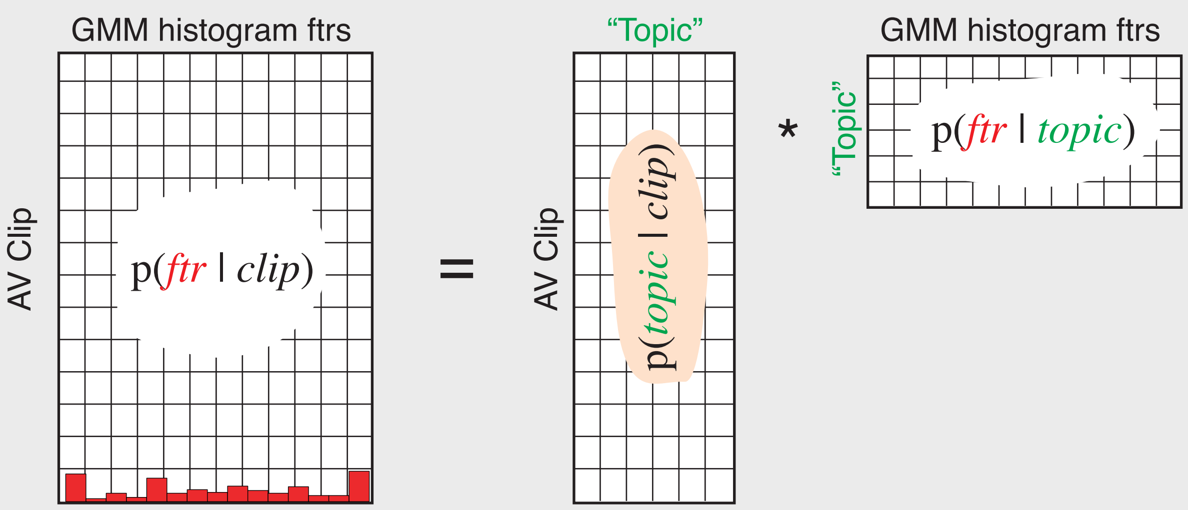
- Convert high-dim. distributions to multinomial



- Classify by distance on histograms
  - KL, Chi-squared
  - + SVM

Latent Semantic Analysis (LSA)

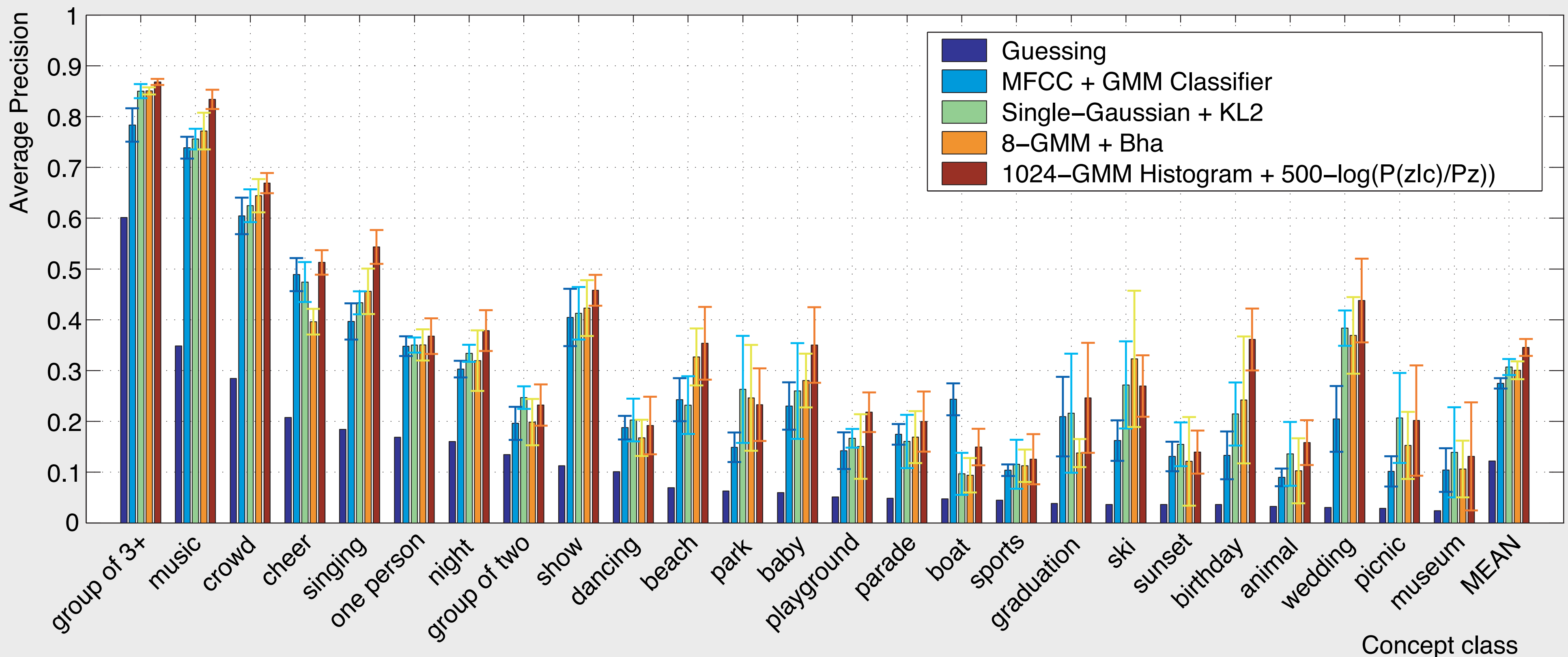
- Probabilistic LSA (pLSA) models each histogram as a mixture of several 'topics'
  - .. each clip may have several things going on
- Topic sets optimized through EM
  - p(*ftr* | *clip*) = $\sum_{topics}$ p(*ftr* | *topic*) p(*topic* | *clip*)



  - use (normalized?) p(*topic* | *clip*) as per-clip features

Background Classification Results

K Lee & Ellis '10

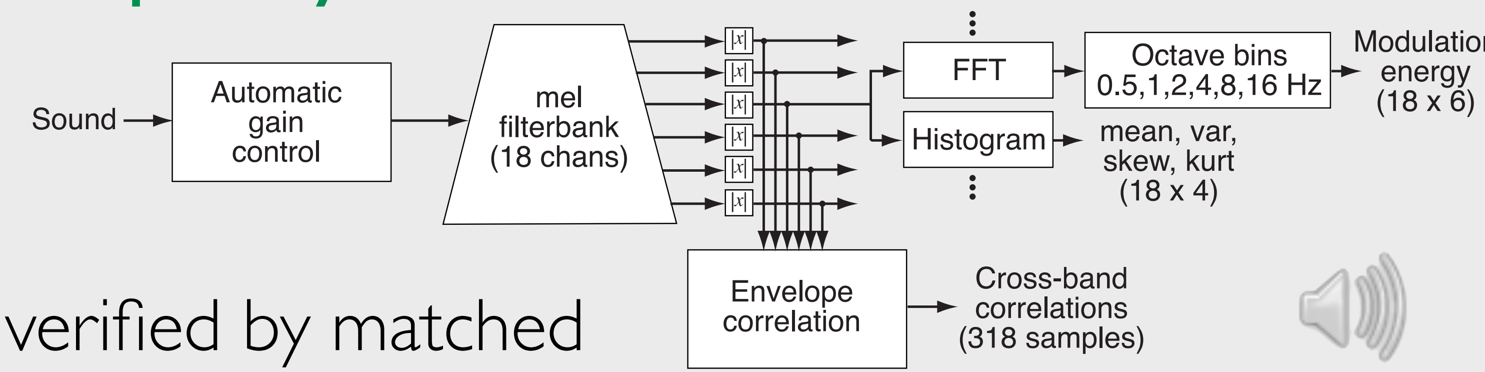


- Wide range in performance
  - audio (music, ski) vs. non-audio (group, night)
  - large AP uncertainty on infrequent classes

Sound Texture Features

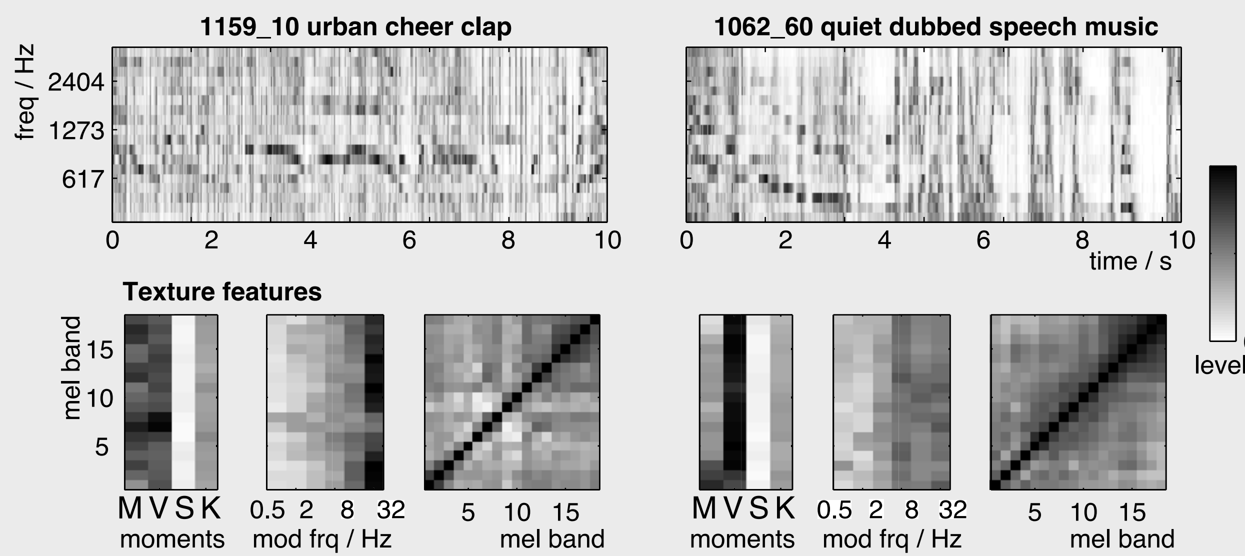
McDermott Simoncelli '09
Ellis, Zheng, McDermott '11

- Characterize sounds by perceptually-sufficient statistics



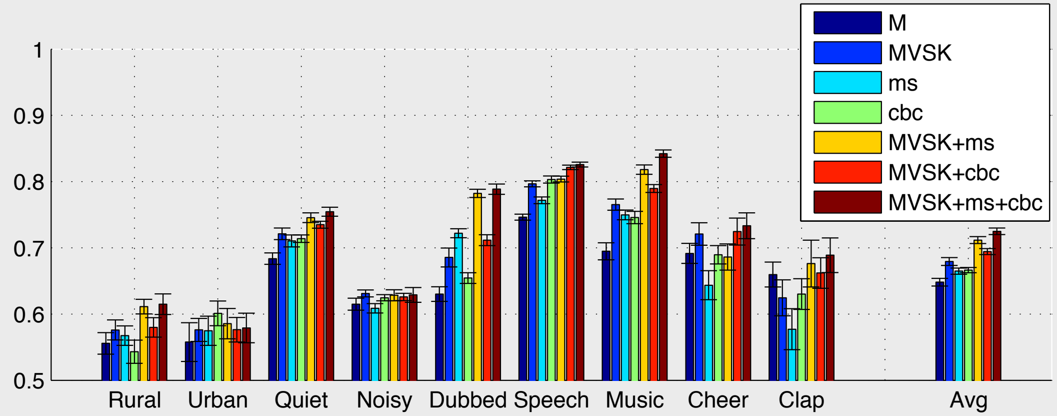
  - .. verified by matched resynthesis

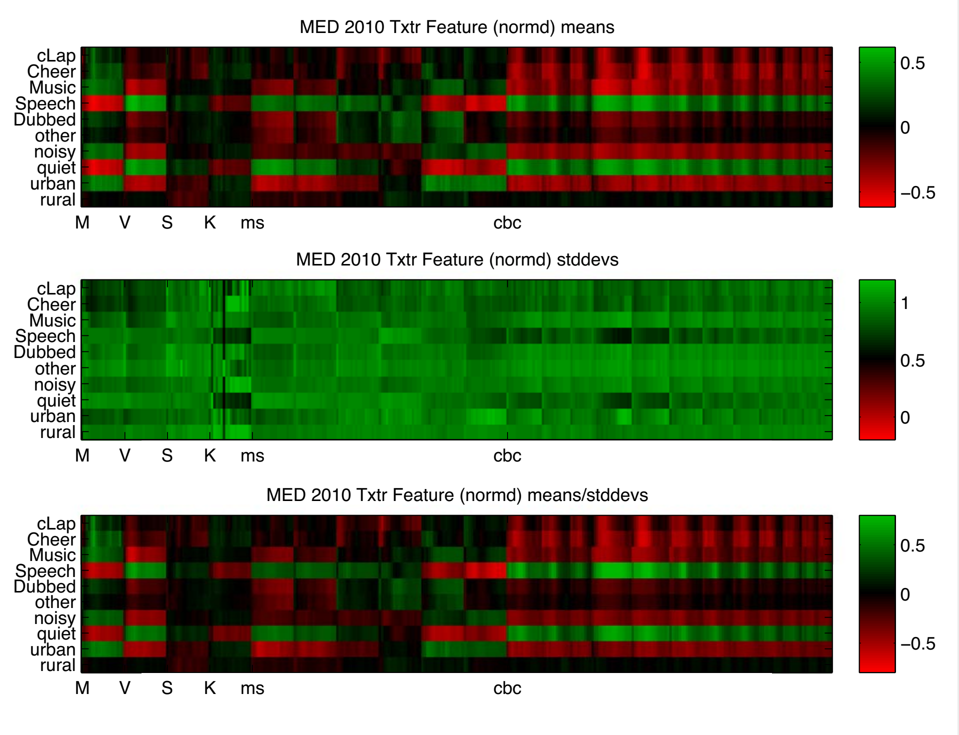
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



Sound Texture Features

- Test on MED 2010 development data
  - 10 specially-collected manual labels



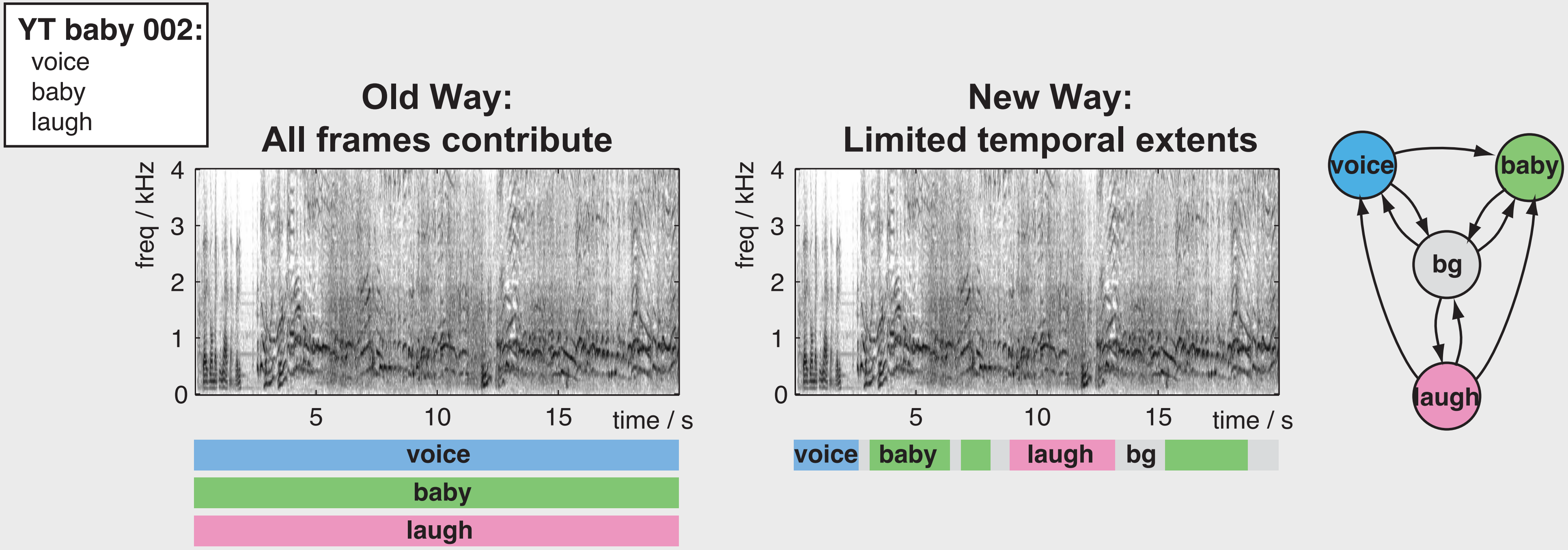


- Contrasts in feature sets
  - correlation of labels...
- Perform ~ same as MFCCs
  - combine well

3. Foreground Event Recognition

K Lee, Ellis, Loui '10

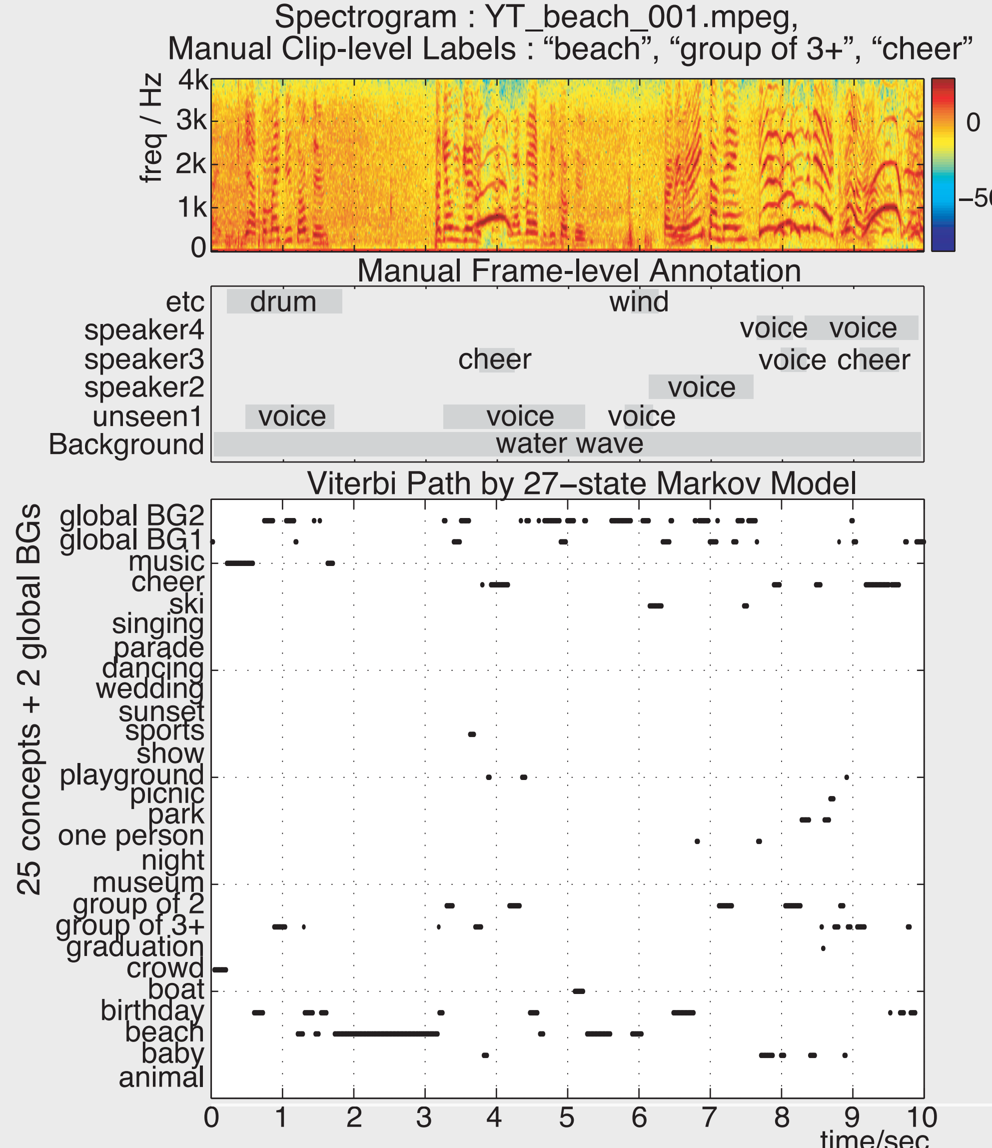
- Global vs. local class models
  - tell-tale acoustics may be 'washed out' in statistics
  - try iterative realignment of HMMs:



  - "background" model shared by all clips

Foreground Event HMMs

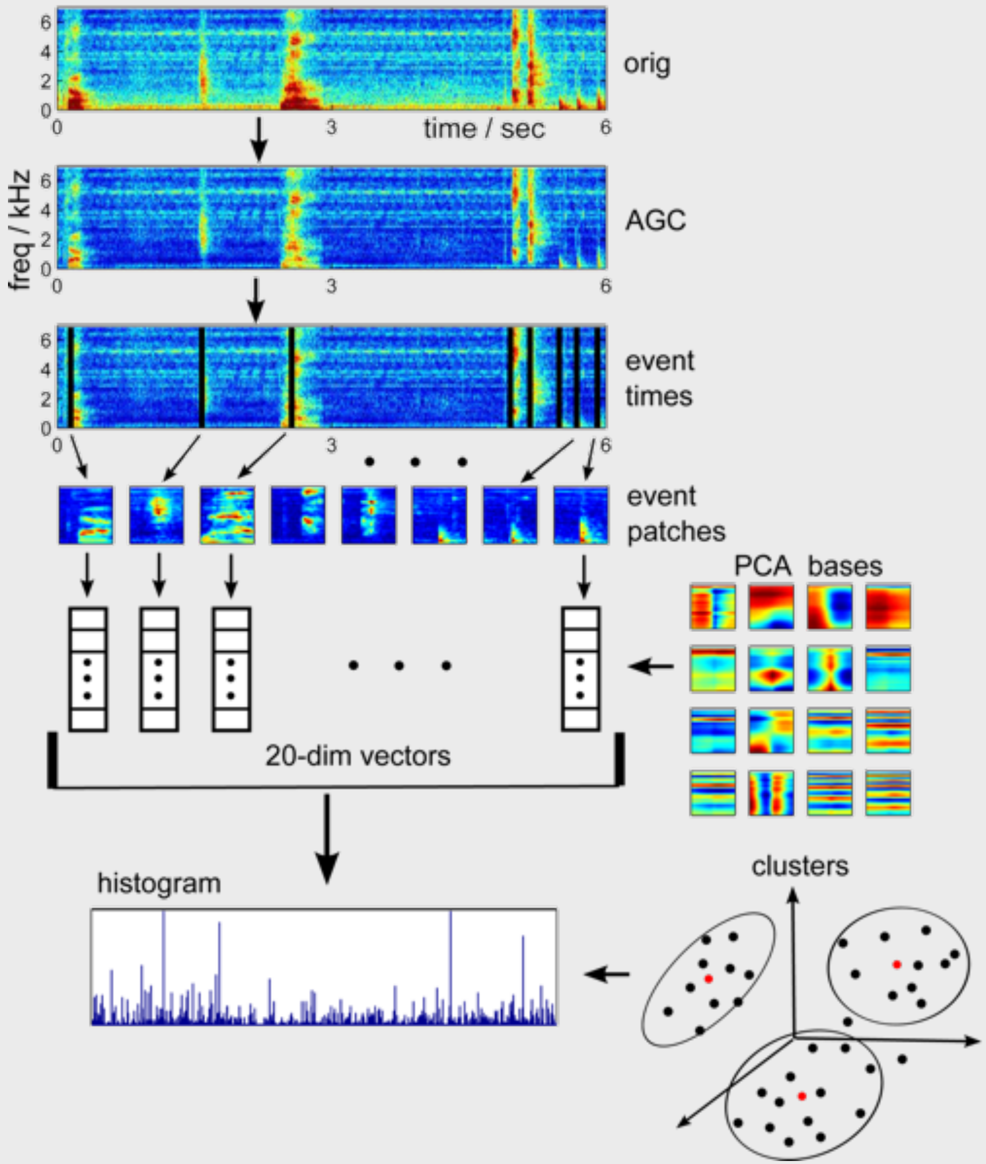
- Training labels only at clip-level
- Refine models by EM realignment
- Use for classifying entire video...
  - or seeking to relevant part



Lee & Ellis '10

Transient Features

Cotton, Ellis, Loui '11



- Transients = foreground events?
- Onset detector finds energy bursts
  - best SNR
- PCA basis to represent each
  - 300 ms x auditory freq
- "bag of transients"

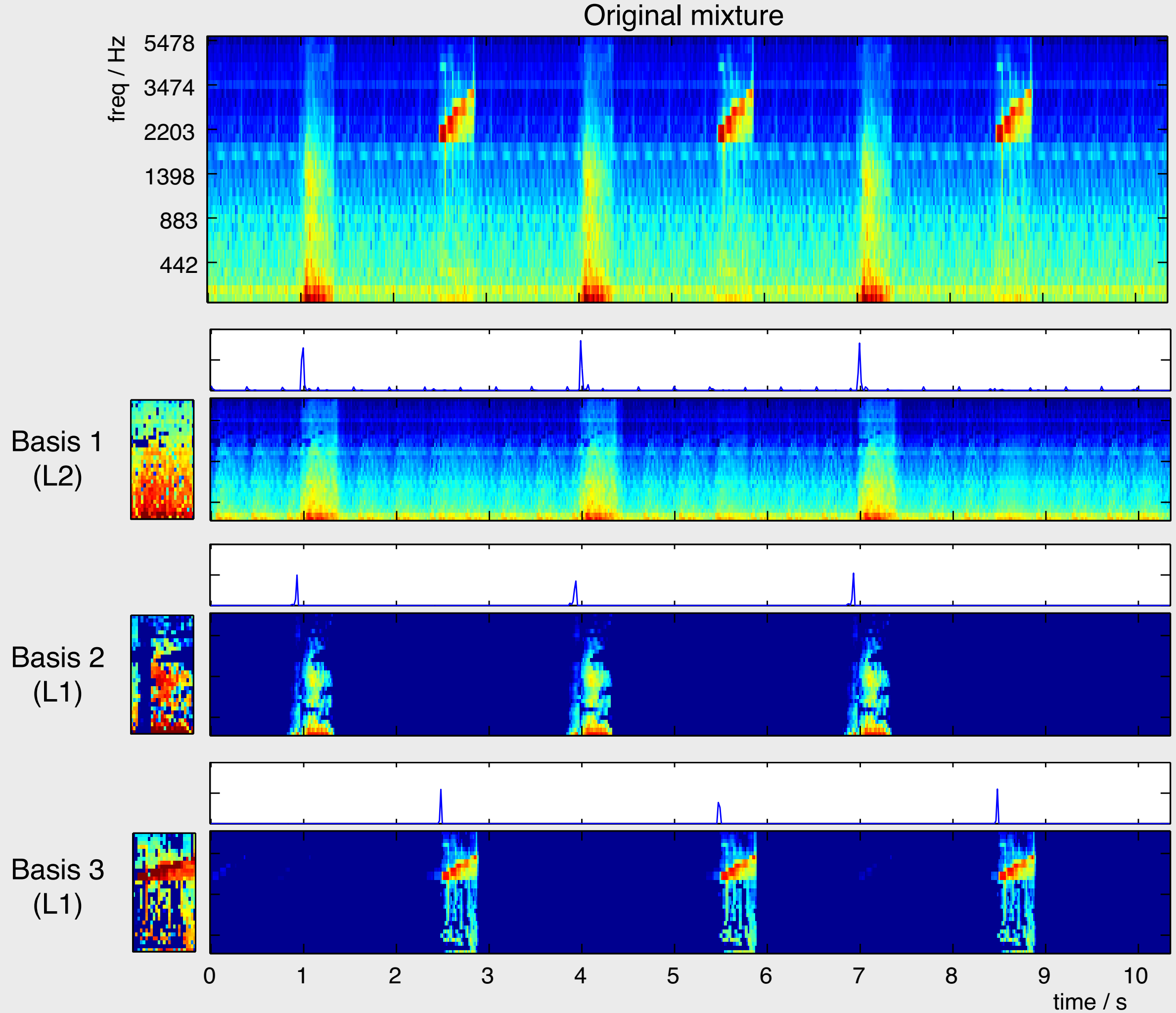
Nonnegative Matrix Factorization

Smaragdis Brown '03
Abdallah Plumbley '04
Virtanen '07

- Decompose spectrograms into
  templates
  + activation

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$$

  - fast forgiving gradient descent algorithm
  - 2D patches
  - sparsity control...

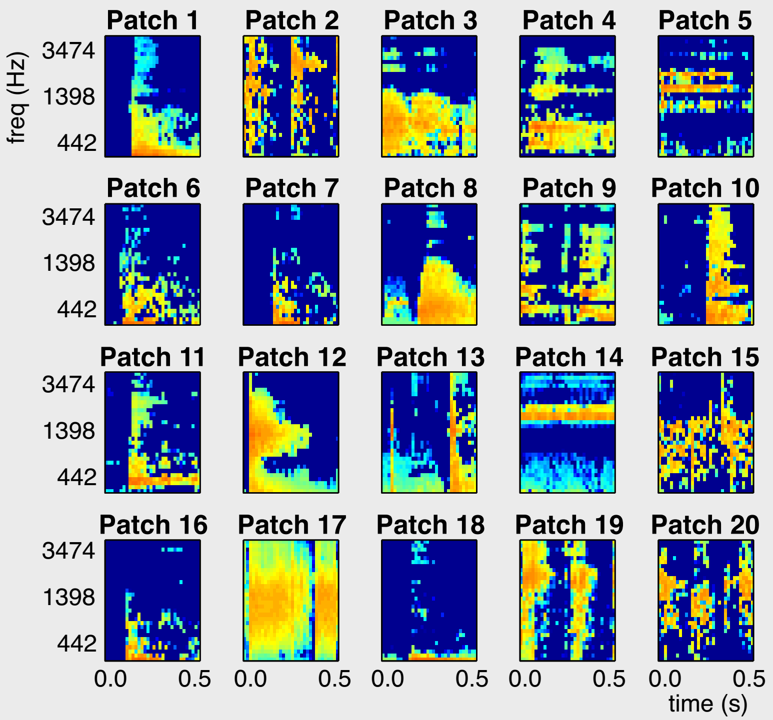


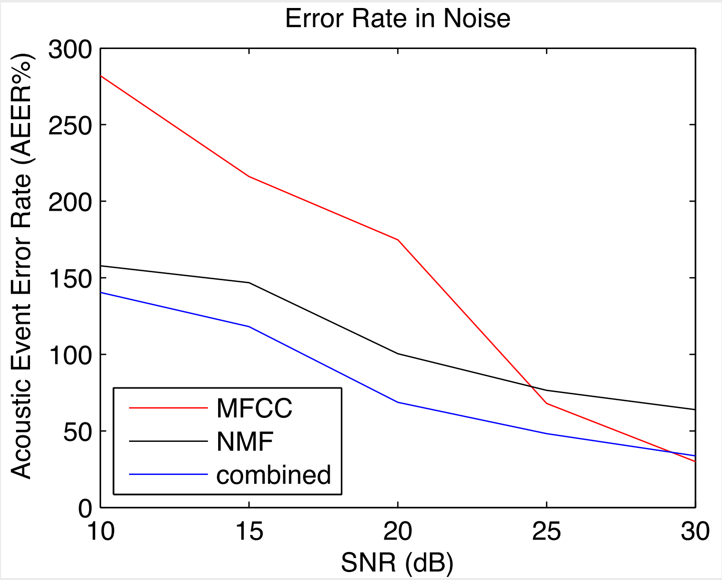
NMF Transient Features

Cotton, Ellis '11

- Learn 20 patches from Meeting Room Acoustic Event data

- Compare to MFCC-HMM detector

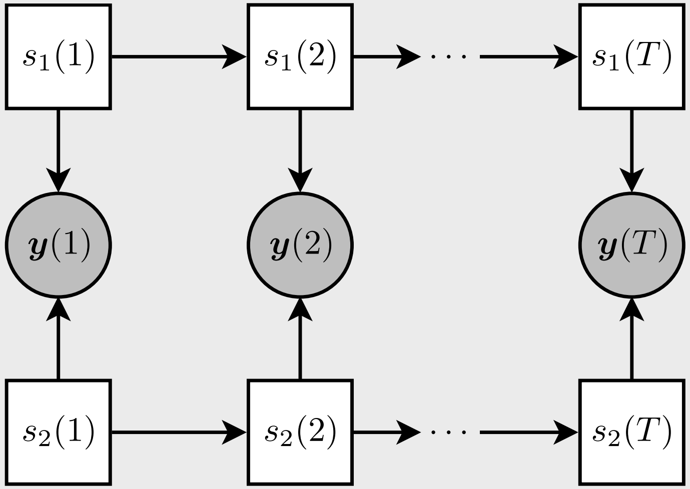


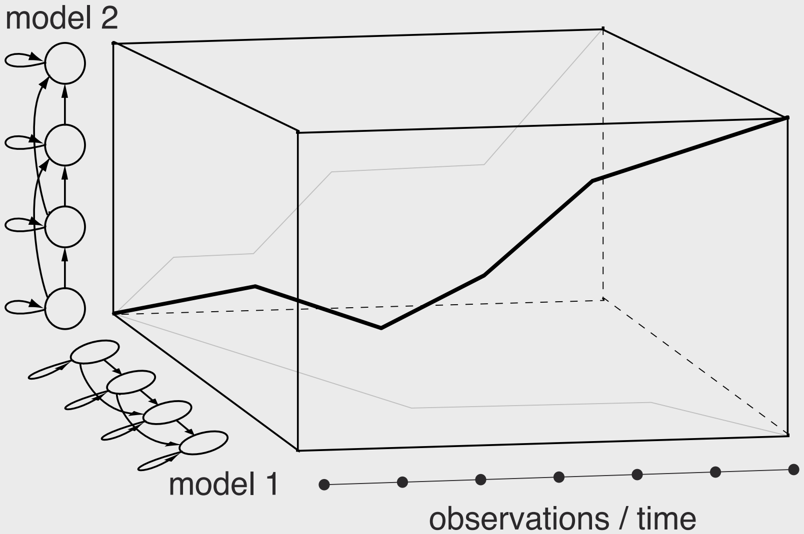


- NMF more noise-robust
  - combines well ...

4. Speech Separation

- Speech recognition is finding best-fit parameters - argmax P(W | X)
- Recognize mixtures with Factorial HMM
  - model + state sequence for each voice/source
  - exploit sequence constraints, speaker differences



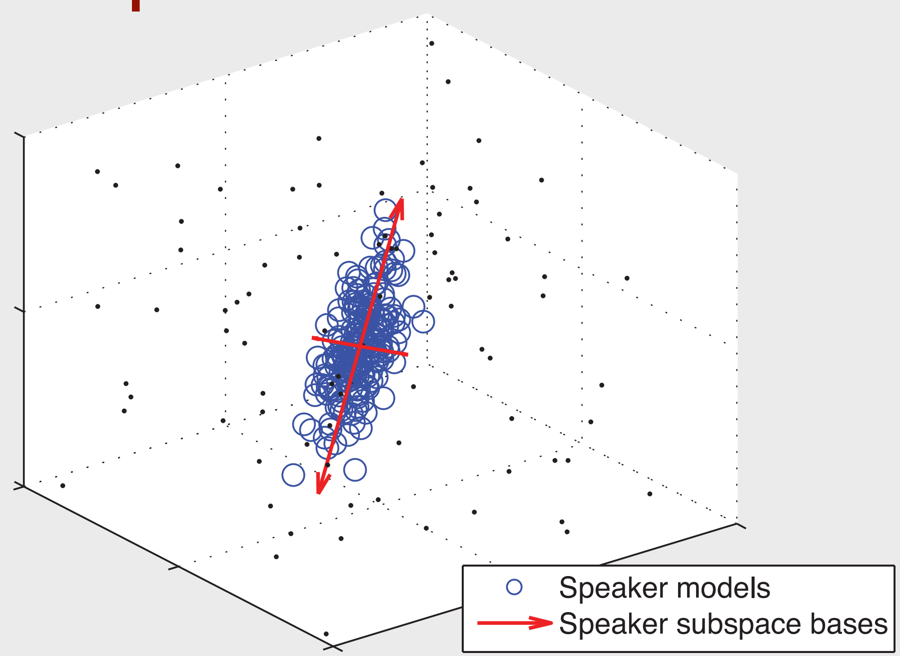


  - separation relies on detailed speaker model

Eigenvoices

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

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



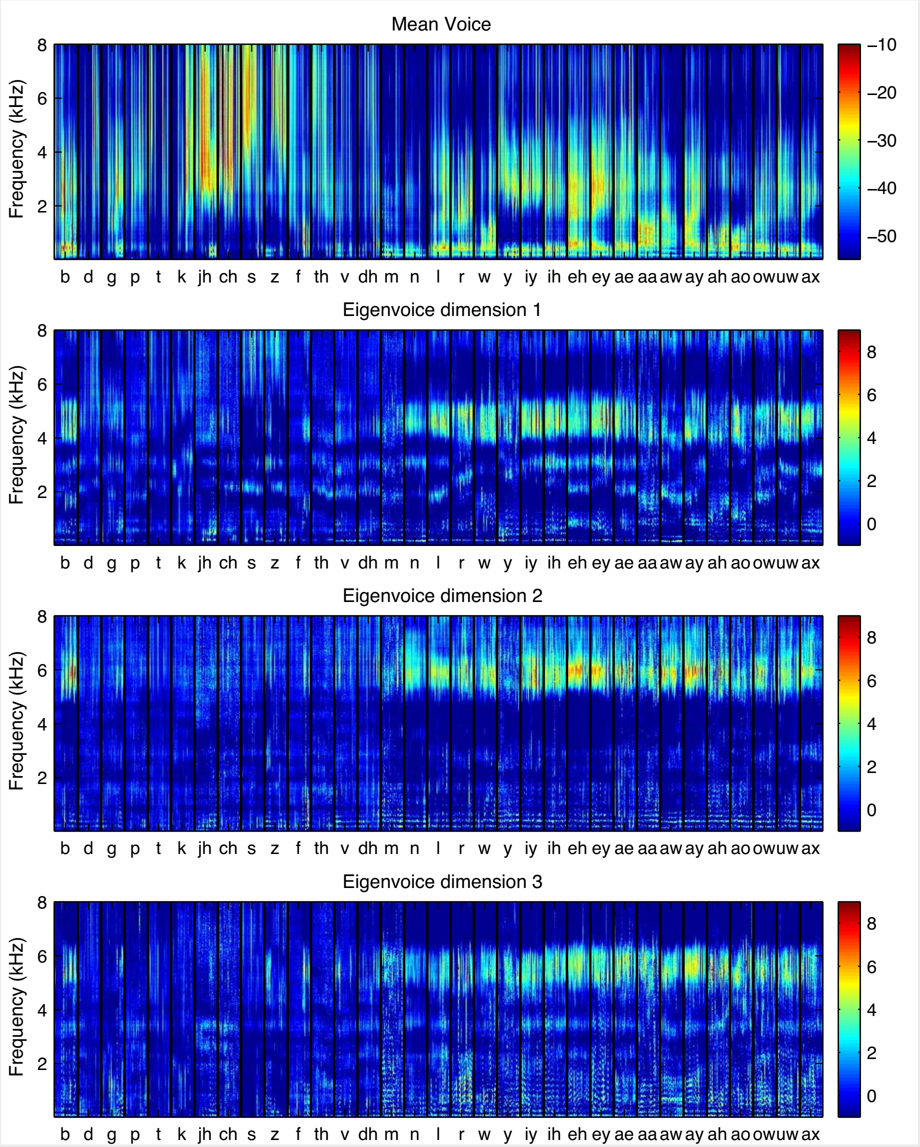
- Eigenvoice model:

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

adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

Eigenvoice Bases

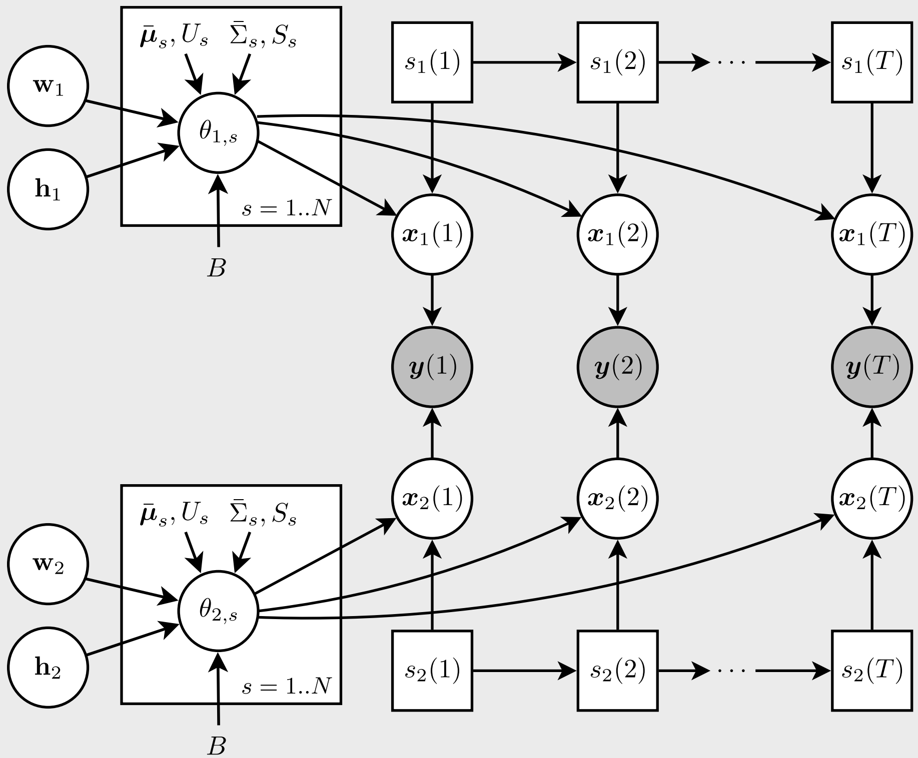
- Mean **model**
  - 280 states x 320 bins = 89,600 dimensions
- **Eigencomponents shift formants/ coloration**
  - additional components for acoustic channel



Eigenvoice Speech Separation

Weiss & Ellis '10

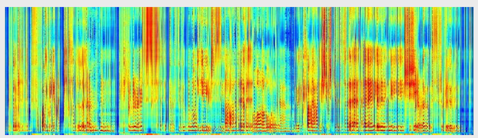
- Factorial HMM analysis
  with tuning of source model parameters
  = eigenvoice speaker adaptation



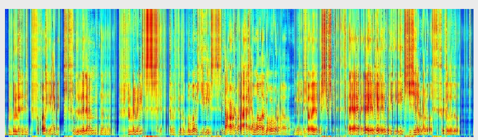
Eigenvoice Speech Separation

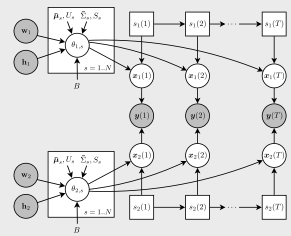
Find Viterbi path

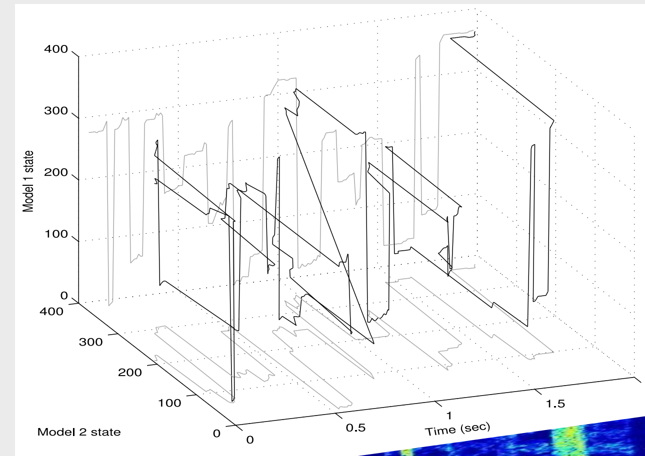
$$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$$

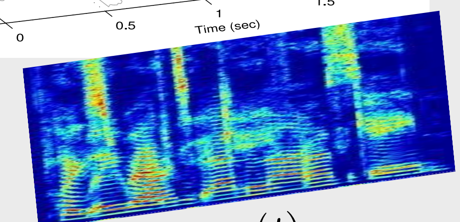


$$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$$





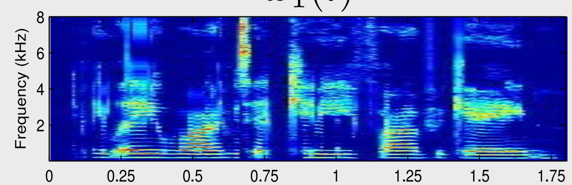




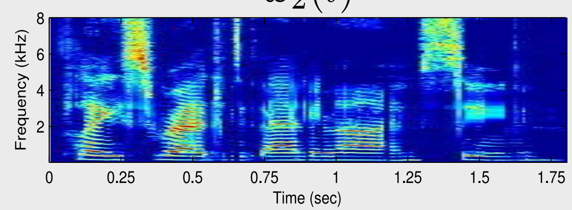
$\boldsymbol{y}(t)$

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

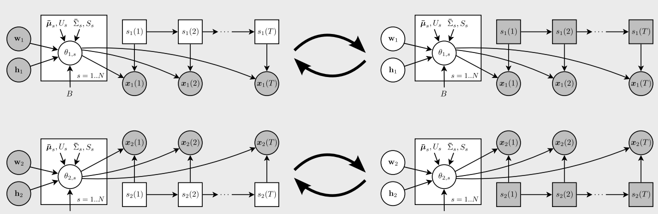
$\hat{\boldsymbol{x}}_1(t)$

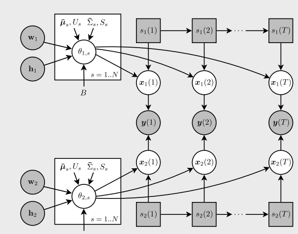


$\hat{\boldsymbol{x}}_2(t)$



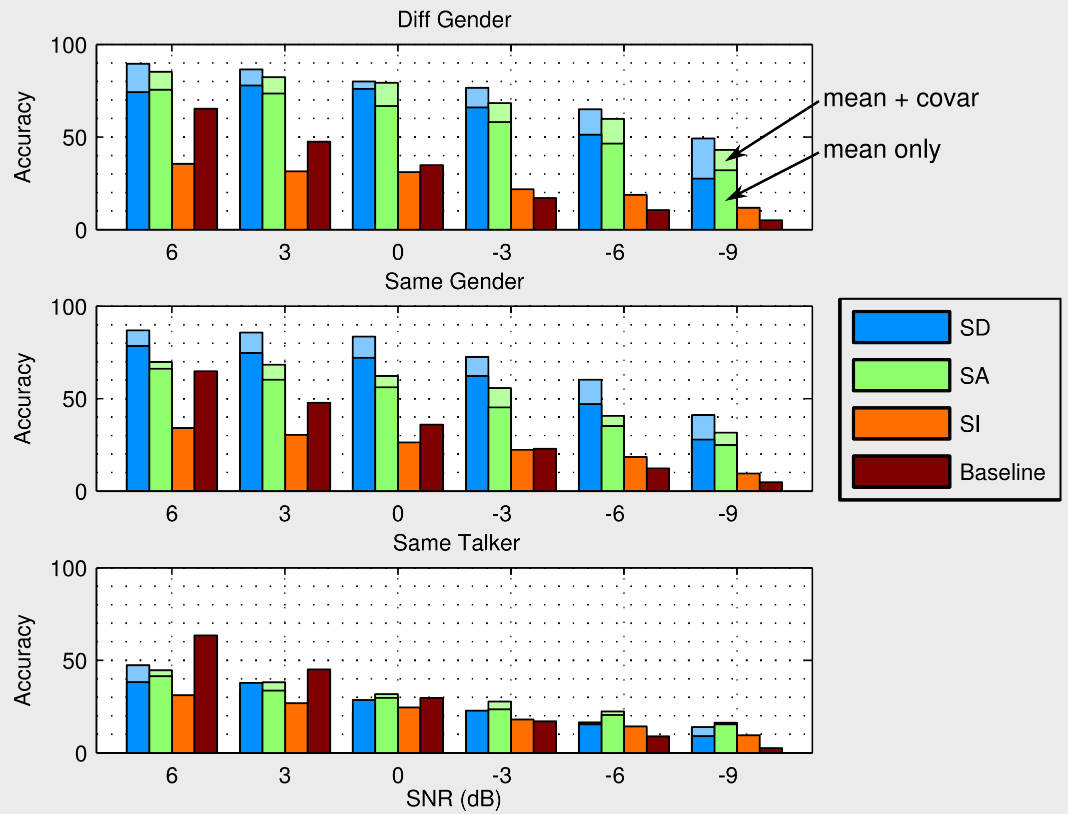
Estimate source signals





Eigenvoice Speech Separation

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



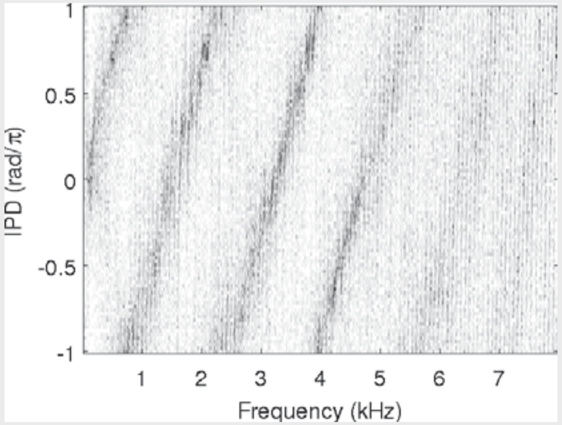
Binaural Cues

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

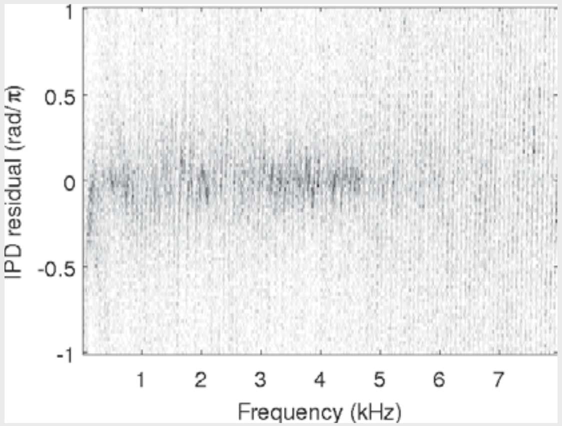
$$\frac{L(\omega,t)}{R(\omega,t)} = a(\omega)e^{j\omega\tau}N(\omega,t)$$

- e.g. at 75°, in reverb:

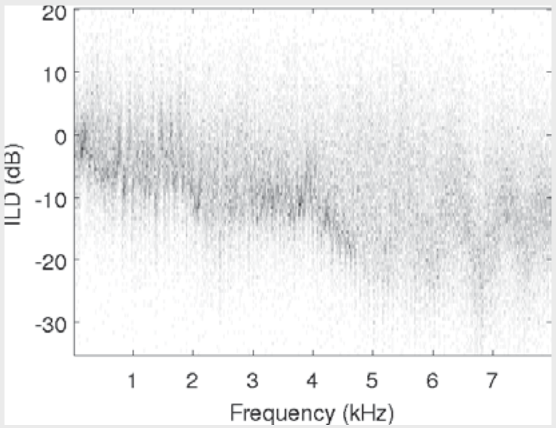
IPD



IPD residual

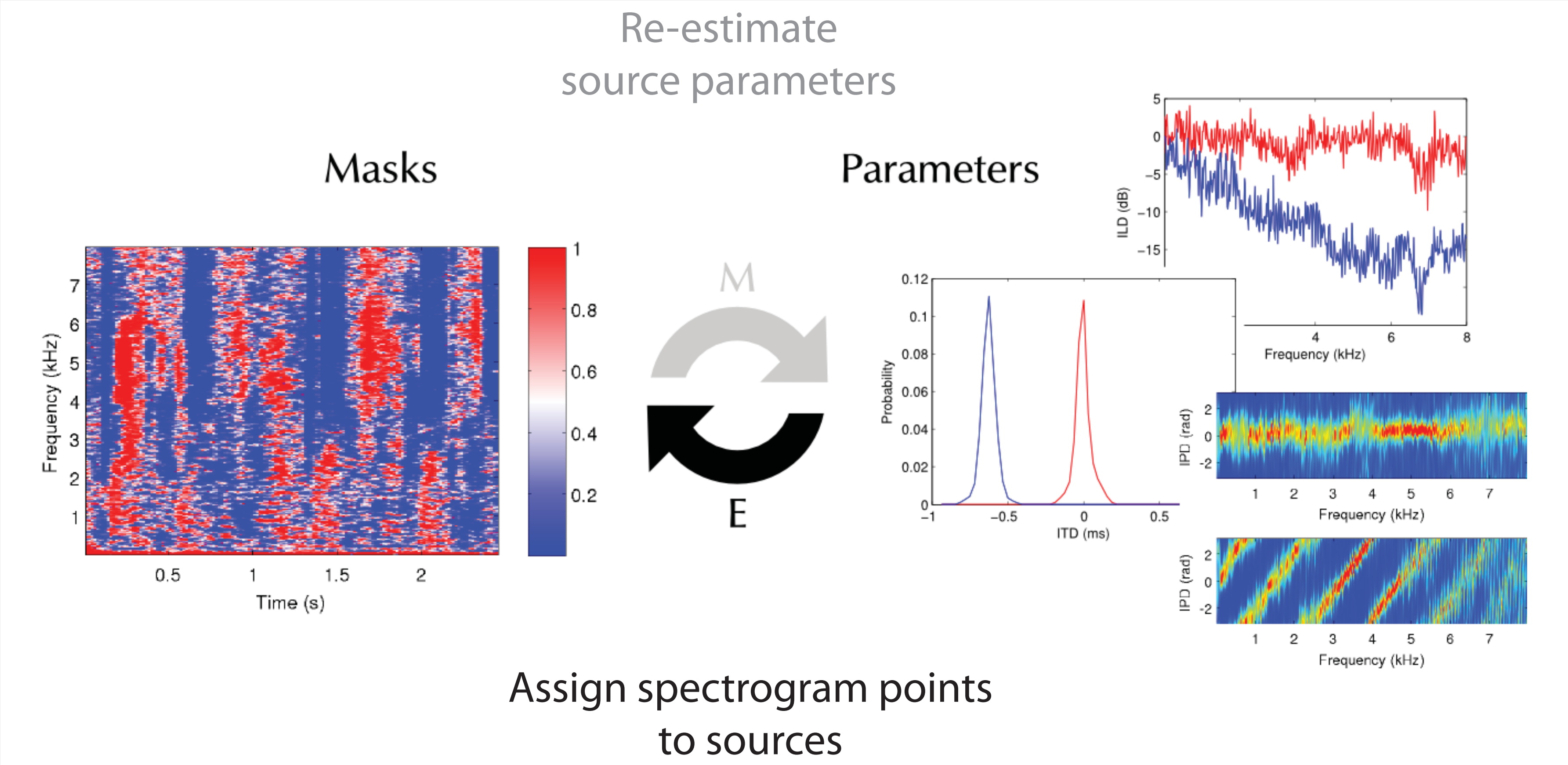


ILD



Model-Based EM Source Separation and Localization (MESSL)

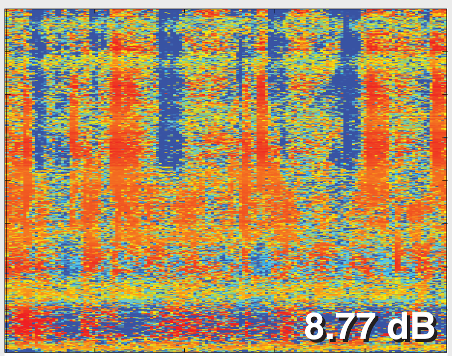
Mandel & Ellis '09



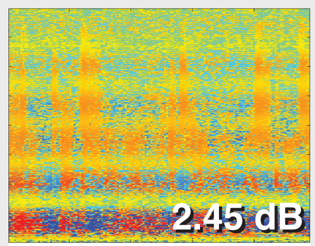
- can model more sources than sensors
- flexible initialization

MESSL Results

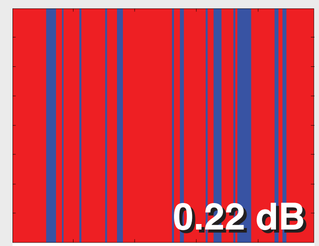
- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness



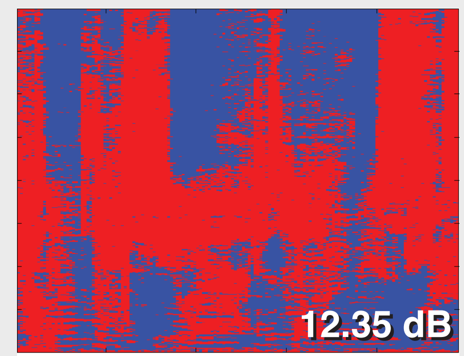
EM+ILD



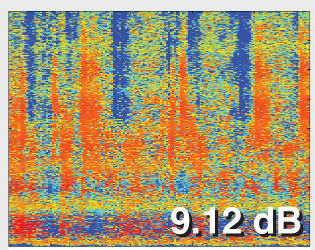
EM-ILD (only IPD)



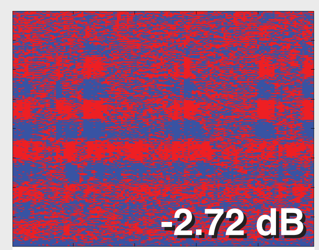
PHAT-histogram



Ground Truth



EM+1ILD (tied means)

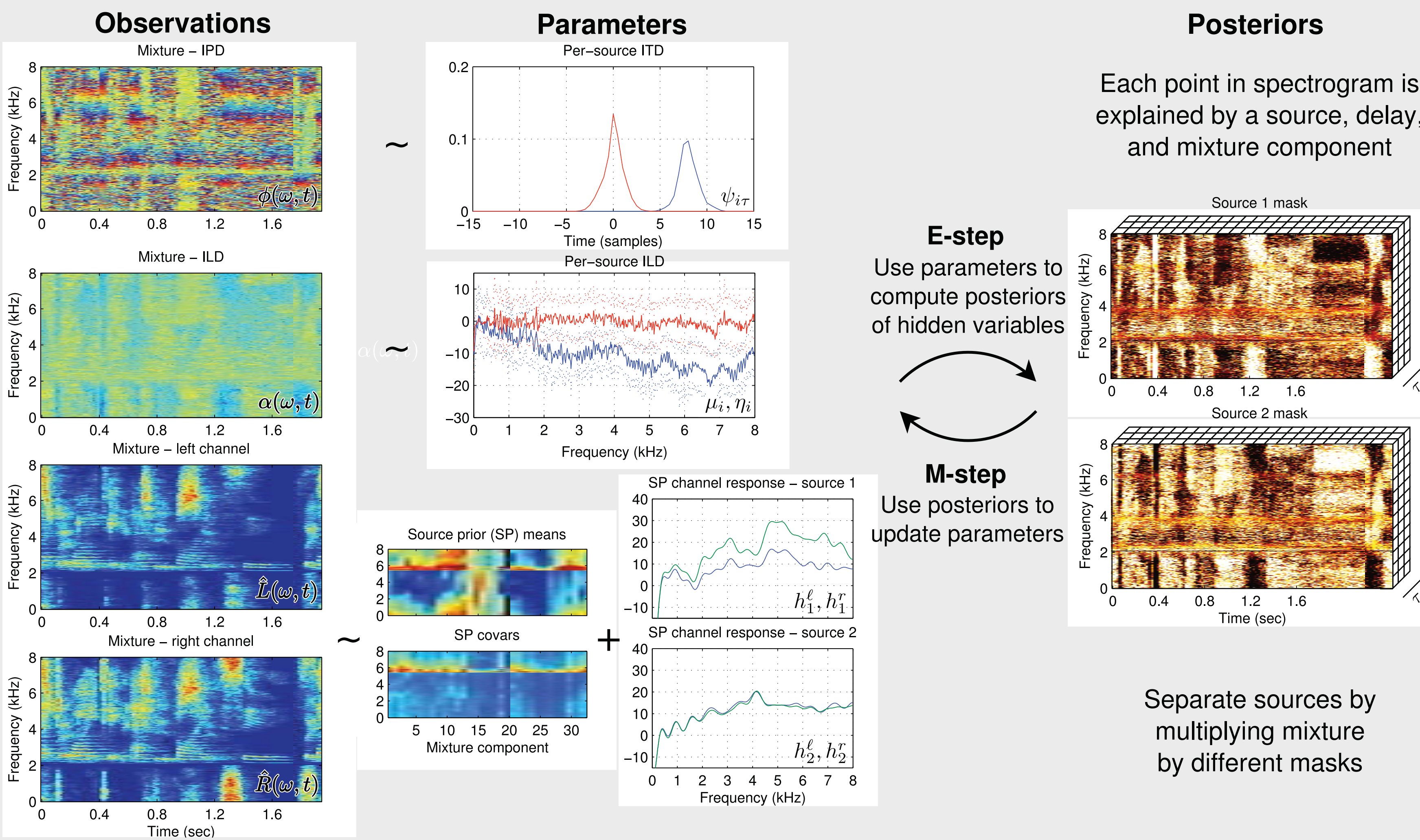


DUET

MESSL with Source Priors

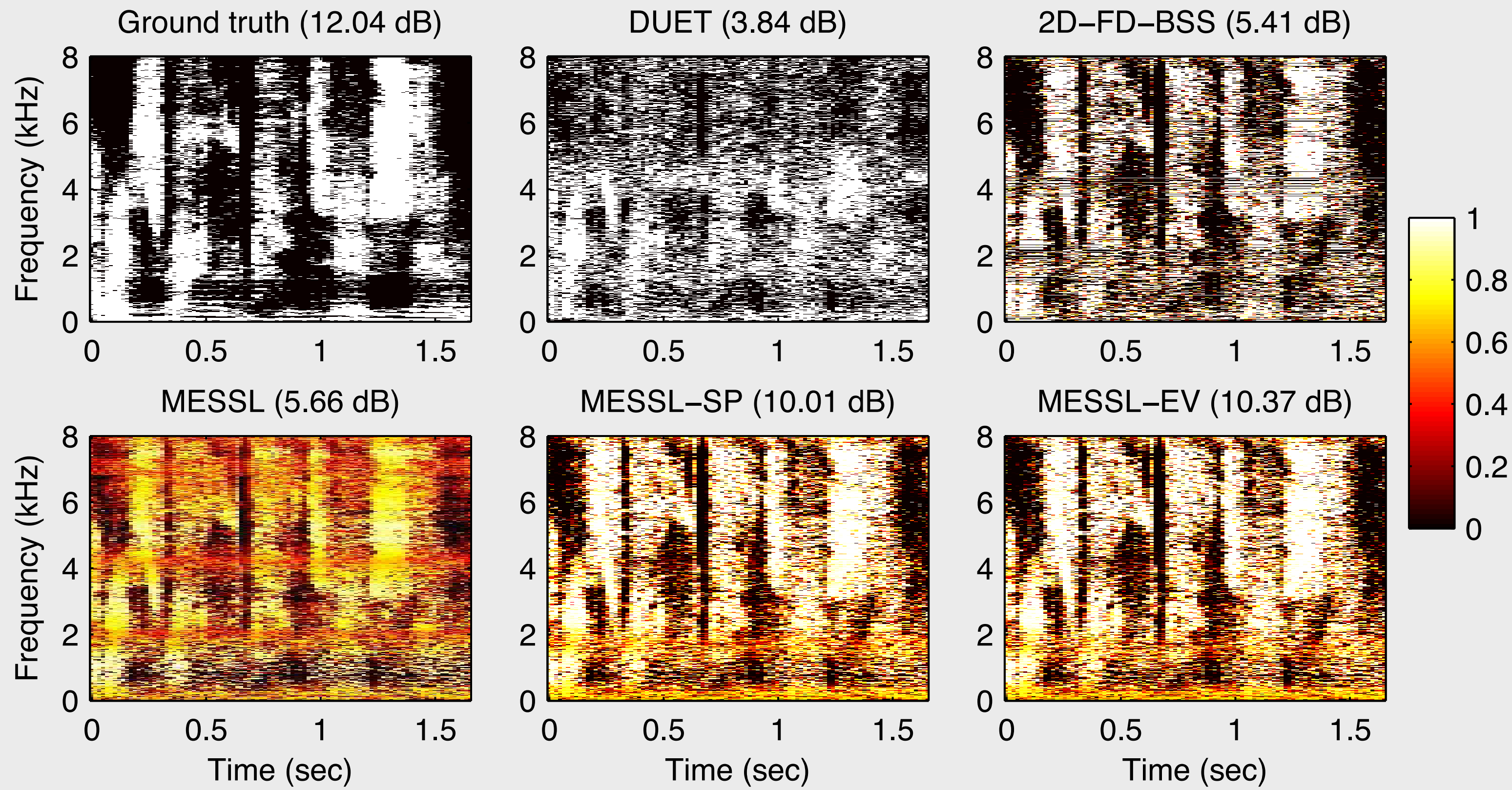
Weiss, Mandel & Ellis '11

- Fixed or adaptive speech models



MESSL-EigenVoice Results

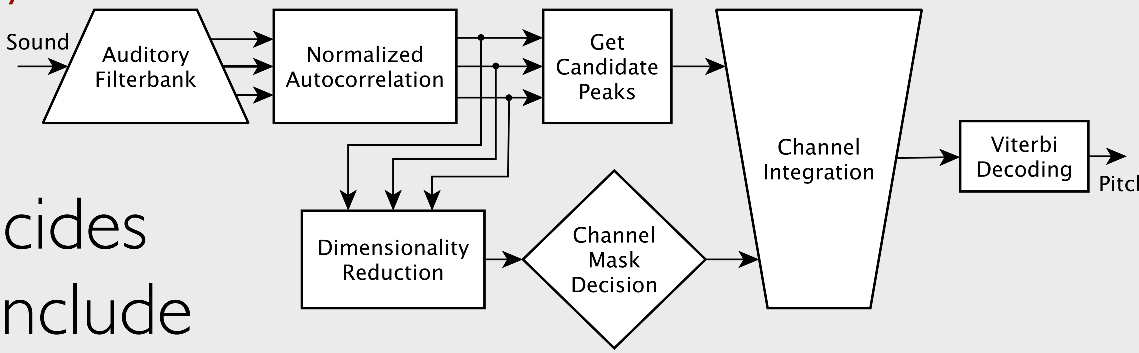
- Source models function as priors
- Interaural parameter spatial separation
  - source model prior improves spatial estimate

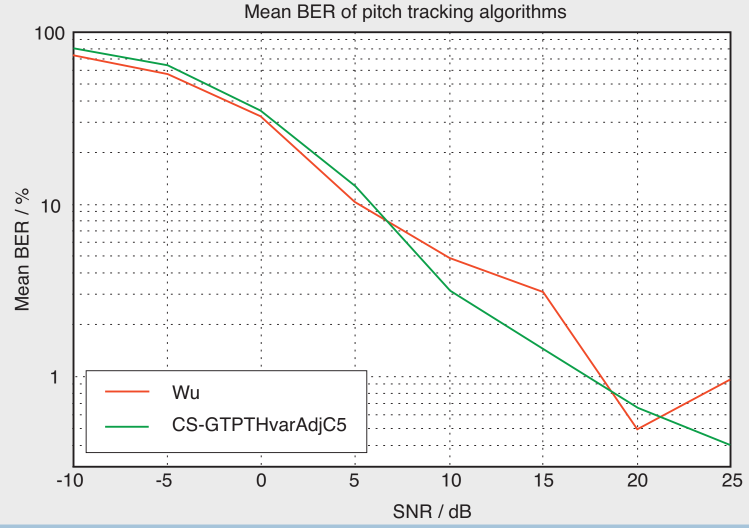


Noise-Robust Pitch Tracking

BS Lee & Ellis '11

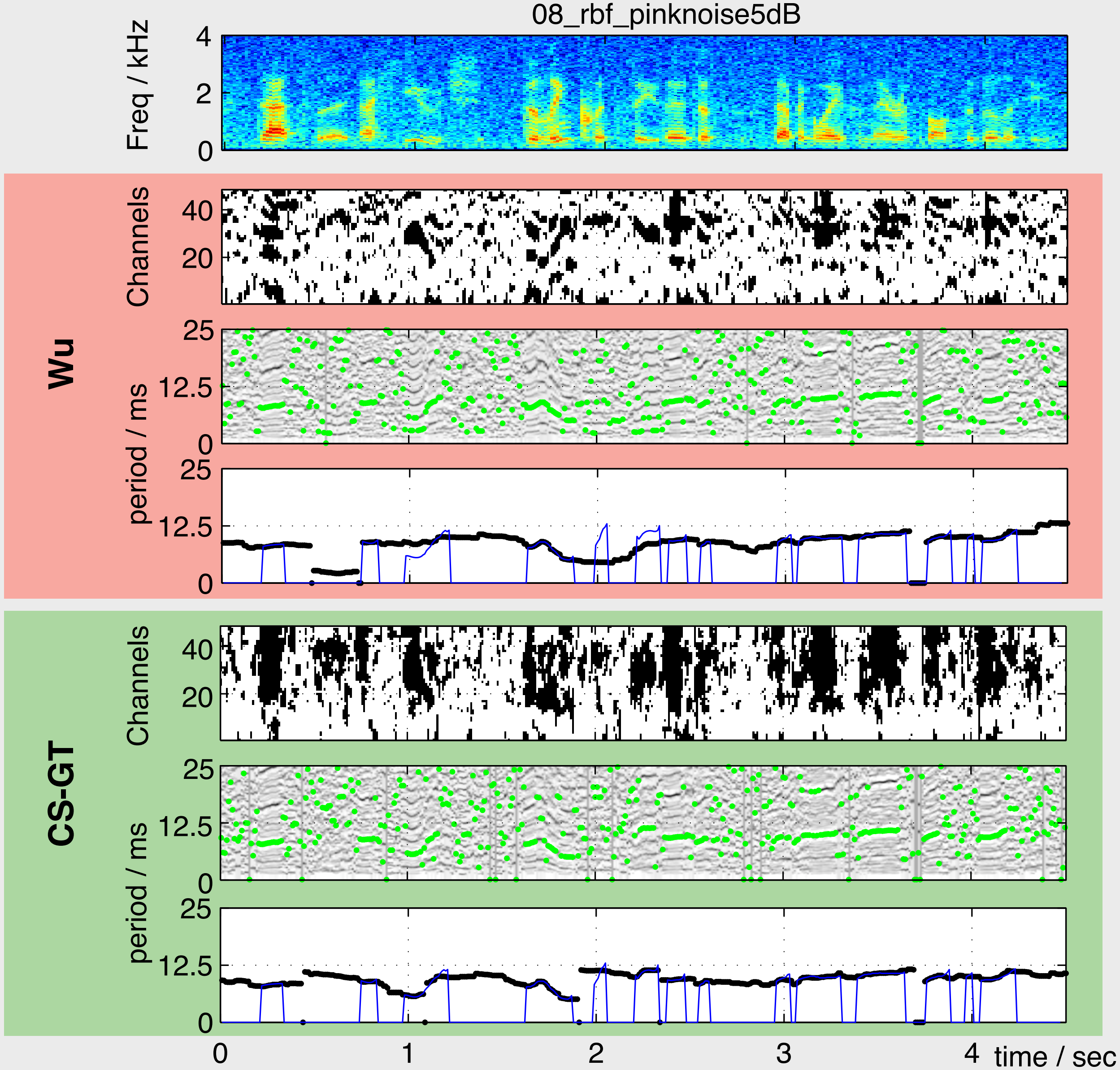
- Important for voice detection & separation
- Based on channel selection *Wu & Wang (2003)*
  - pitch from summary autocorrelation over "good" bands
  - trained classifier decides which channels to include





- Improves over simple Wu criterion
  - especially for mid SNR

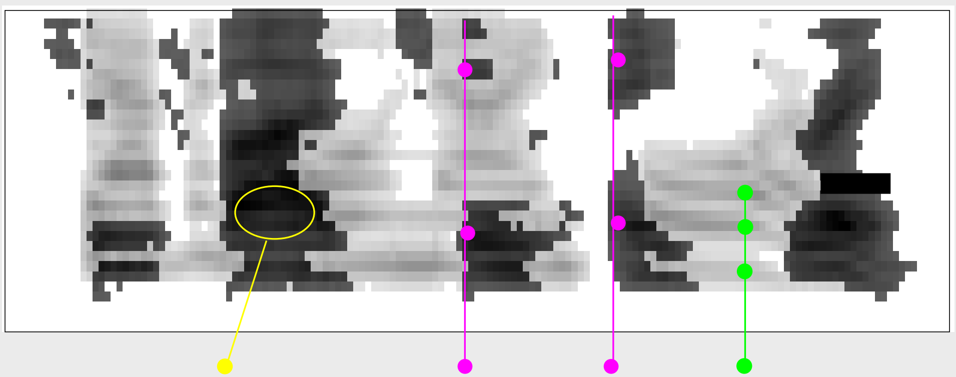
Noise-Robust Pitch Tracking



- Trained selection includes more off-harmonic channels

5. Outstanding Issues

- Better object/event separation
  - parametric models
  - spatial information?
  - computational auditory scene analysis...



Barker et al. '05

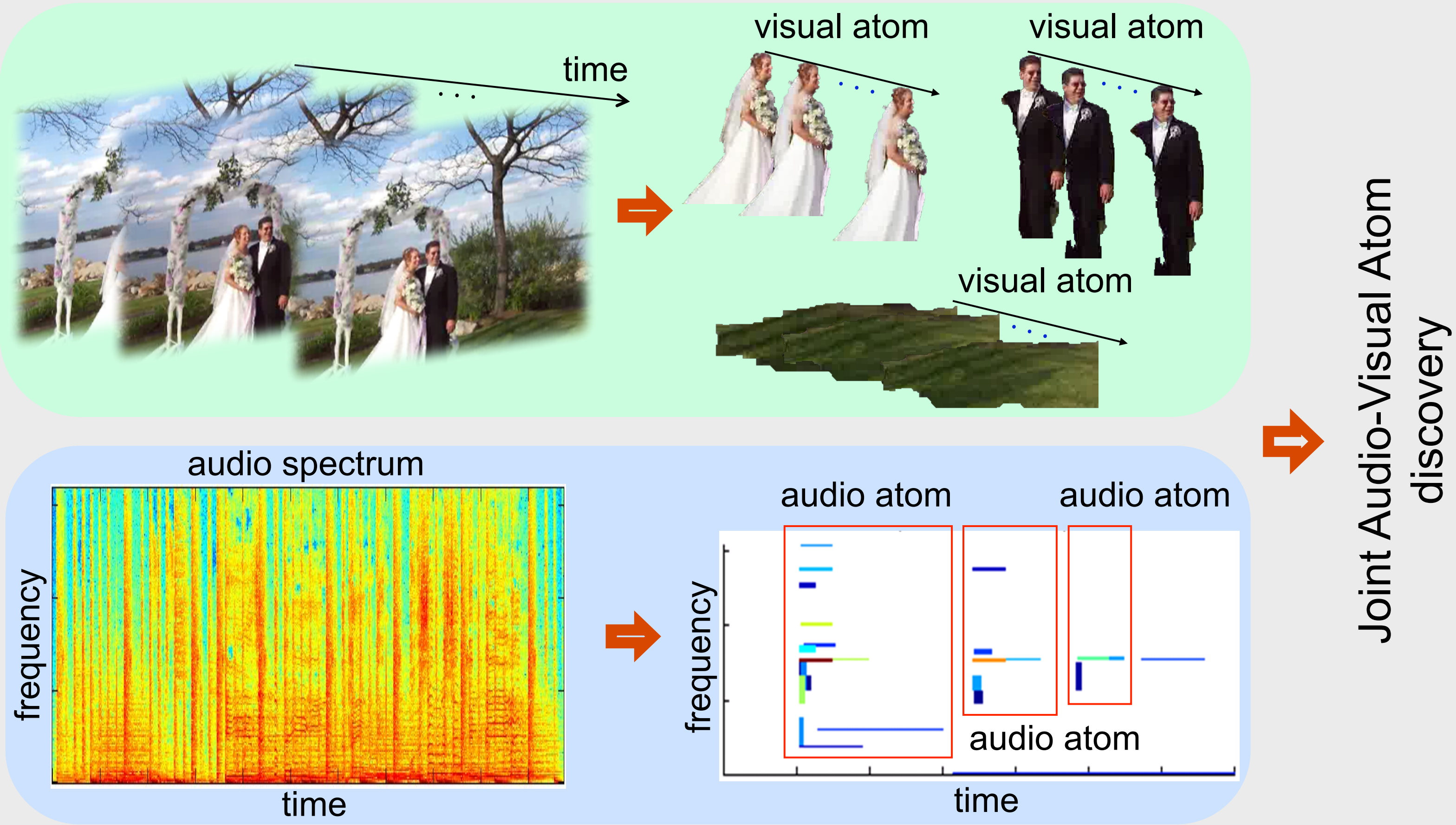
- Large-scale analysis

- Integration with video

Audio-Visual Atoms

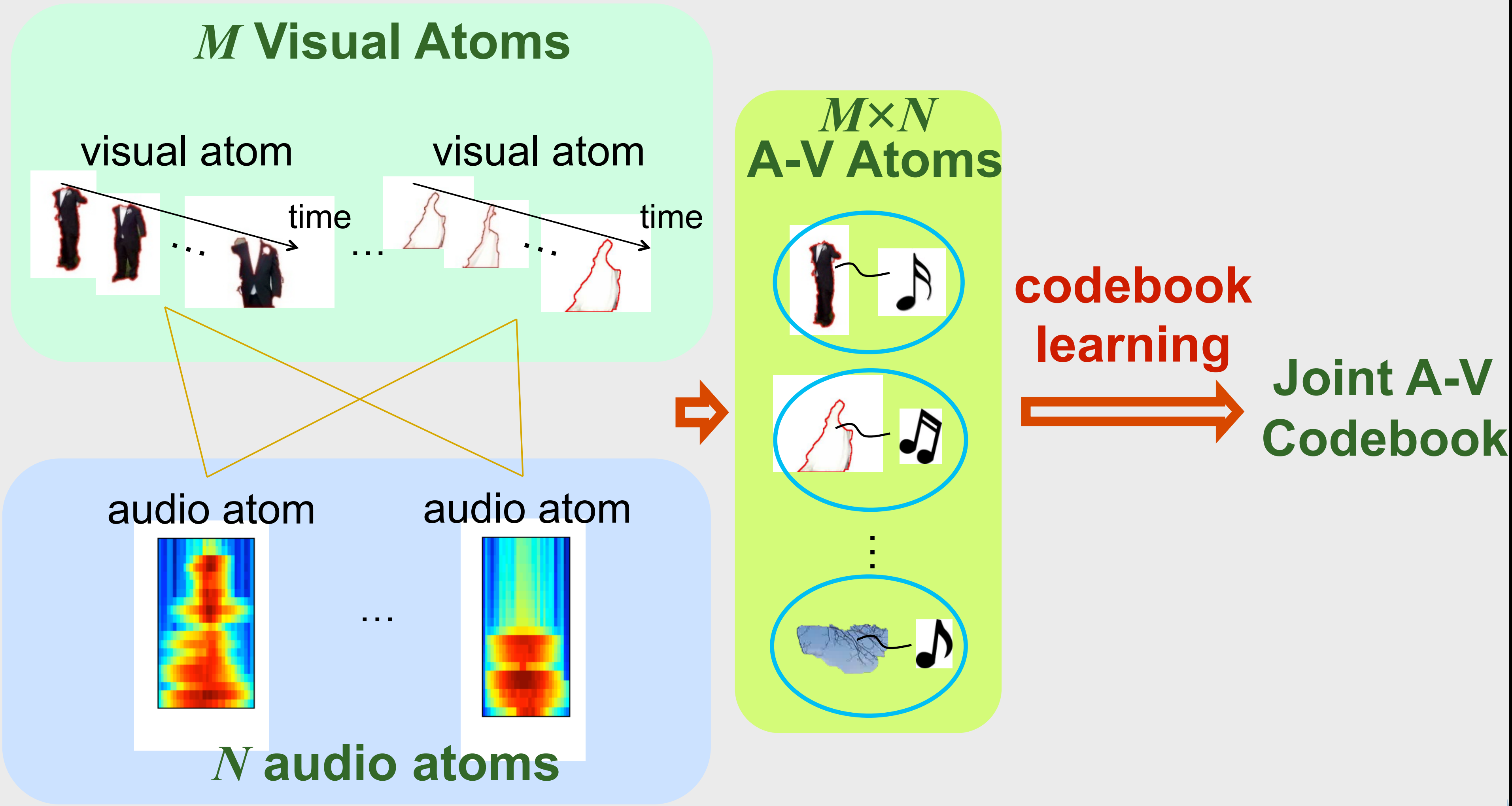
Jiang et al. '09

- Object-related features
  from both audio (transients) & video (patches)



Audio-Visual Atoms

- Multi-instance learning of A-V co-occurrences



Audio-Visual Atoms

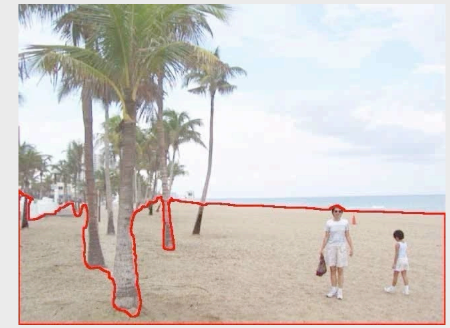
black suit
+ romantic
music

Wedding

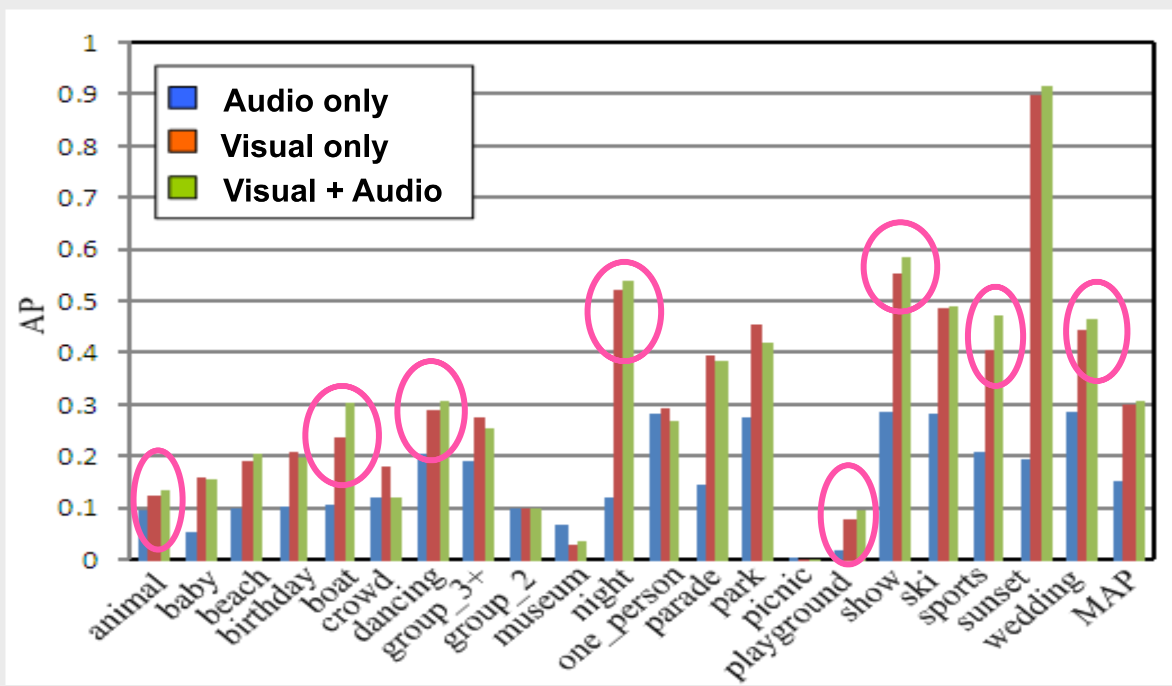
marching
people
+ parade
sound

Parade

sand
+ beach
sounds



Beach



Summary

- Machine Listening:
  Getting useful information from sound

- Background sound classification
  ... from whole-clip statistics?

- Foreground event recognition
  ... by focusing on peak energy patches

- Speech content is very important
  ... separate with pitch, models, ...

References

- Jon Barker, Martin Cooke, & Dan Ellis, "Decoding Speech in the Presence of Other Sources," *Speech Communication* 45(1): 5-25, 2005.
- Courtenay Cotton, Dan Ellis, & Alex Loui, "Soundtrack classification by transient events," *IEEE ICASSP*, Prague, May 2011.
- Courtenay Cotton & Dan Ellis, "Spectral vs. Spectro-Temporal Features for Acoustic Event Classification," submitted to *IEEE WASPAA*, 2011.
- Dan Ellis, Xiaohong Zheng, Josh McDermott, "Classifying soundtracks with audio texture features," *IEEE ICASSP*, Prague, May 2011.
- Wei Jiang, Courtenay Cotton, Shih-Fu Chang, Dan Ellis, & Alex Loui, "Short-Term Audio-Visual Atoms for Generic Video Concept Classification," *ACM MultiMedia*, 5-14, Beijing, Oct 2009.
- Keansub Lee & Dan Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Tr. Audio, Speech and Lang. Proc.* 18(6): 1406-1416, Aug. 2010.
- Keansub Lee, Dan Ellis, Alex Loui, "Detecting local semantic concepts in environmental sounds using Markov model based clustering," *IEEE ICASSP*, 2278-2281, Dallas, Apr 2010.
- Byung-Suk Lee & Dan Ellis, "Noise-robust pitch tracking by trained channel selection," submitted to *IEEE WASPAA*, 2011.
- Andriy Temko & Climent Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition* 39(4): 682-694, 2006
- Ron Weiss & Dan Ellis, "Speech separation using speaker-adapted Eigenvoice speech models," *Computer Speech & Lang*. 24(1): 16-29, 2010.
- Ron Weiss, Michael Mandel, & Dan Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Communication* 53(5): 606-621, May 2011.
- Tong Zhang & C.-C. Jay Kuo, "Audio content analysis for on-line audiovisual data segmentation," *IEEE TSAP* 9(4): 441-457, May 2001