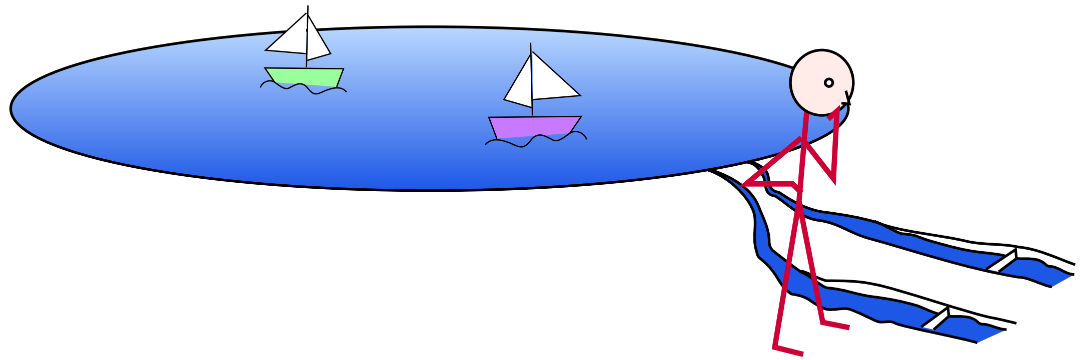
Using Sound Source Models to Organize Mixtures

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. Mixtures and Models
2. Human Sound Organization
3. Machine Sound Organization
4. Ambient Sounds

The Problem of Mixtures

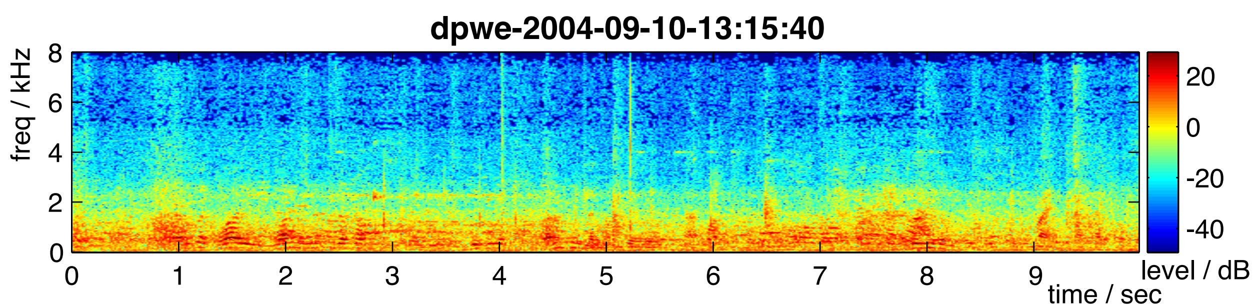


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- **Received waveform is a mixture**
  - 2 sensors, N sources - underconstrained
- **Undoing mixtures: hearing's primary goal?**
  - .. by any means available

Sound Organization Scenarios

- Interactive voice systems
  - human-level understanding is expected
- Speech prostheses
  - crowds: #1 complaint of hearing aid users
- Archive analysis
  - identifying and isolating sound events



- Unmixing/remixing/enhancement...

How Can We Separate?

- By between-sensor differences (spatial cues)
  - ‘steer a null’ onto a compact interfering source
  - the filtering/signal processing paradigm
- By finding a ‘separable representation’
  - spectral? sources are broadband but sparse
  - periodicity? maybe – for pitched sounds
  - something more signal-specific...
- By inference (based on knowledge/models)
  - acoustic sources are redundant
    - → use part to guess the remainder
    - - limited possible solutions

Separation vs. Inference

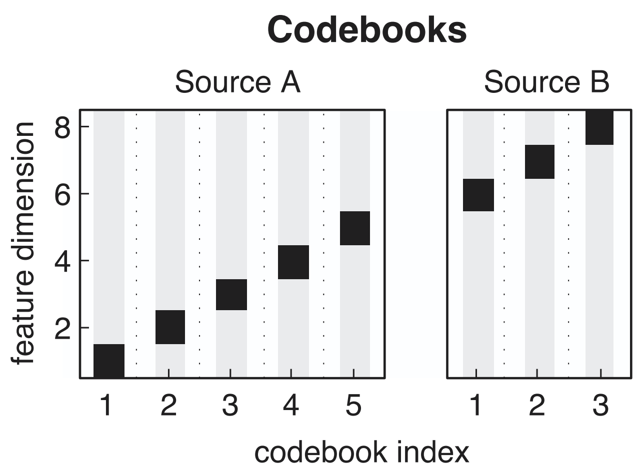
- Ideal separation is rarely possible
  - i.e. no projection can completely remove overlaps
- Overlaps → Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

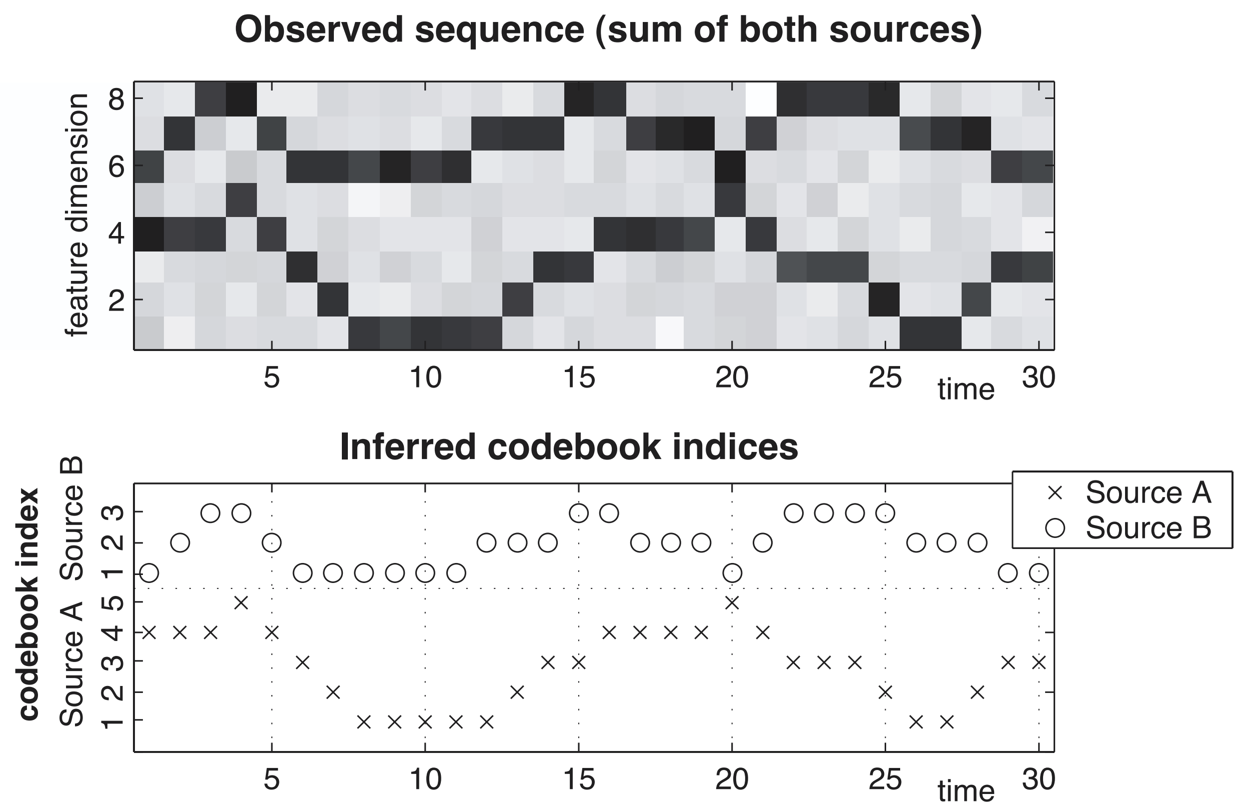
$$P(\{S_i\}|\,X) \propto \underbrace{P(X\,|\{S_i\})}_{\text{combination physics}}\;\underbrace{P(\{S_i\})}_{\text{source models}}$$

- Better source models → better inference
  - .. learn from examples?

A Simple Example

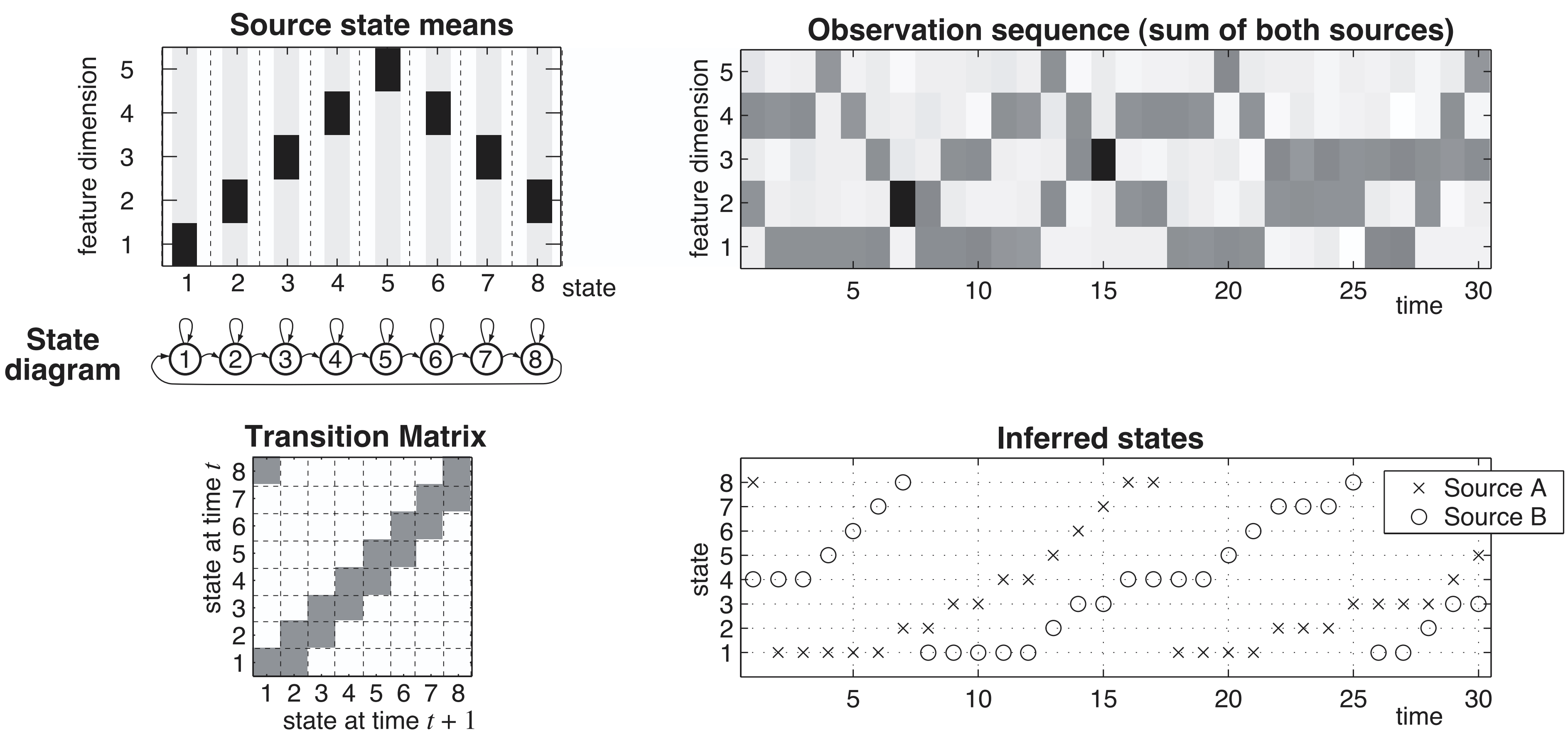
- Source models are codebooks from separate subspaces





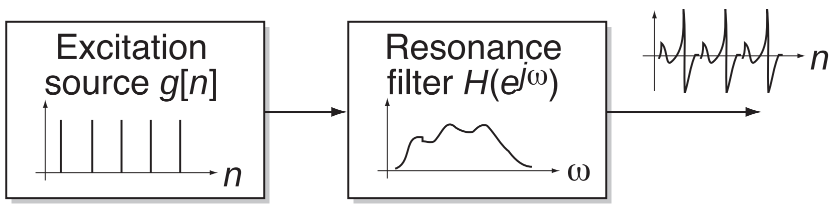
A Slightly Less Simple Example

- Sources with Markov transitions



What is a Source Model?

- Source Model describes signal behavior
  - encapsulates constraints on form of signal
  - (any such constraint can be seen as a model...)
- A model has parameters
  - model + parameters
    → instance



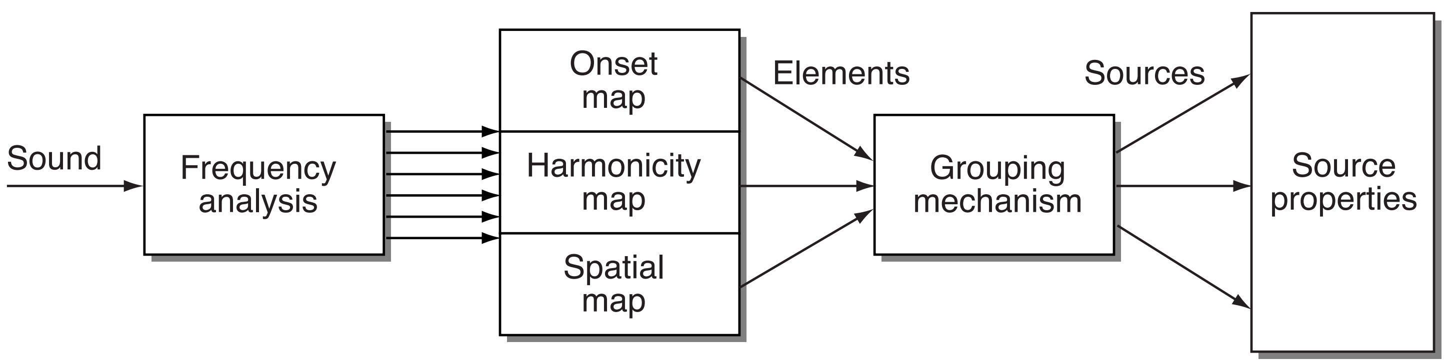
- What is *not* a source model?
  - detail not provided in instance
    e.g. using phase from original mixture
  - constraints on interaction between sources
    e.g. independence, clustering attributes

Outline

Auditory Scene Analysis

Bregman'90
Darwin & Carlyon'95

- How do people analyze sound mixtures?
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes
- Grouping rules (Darwin, Carlyon, ...):
  - cues: common onset/modulation, harmonicity, ...

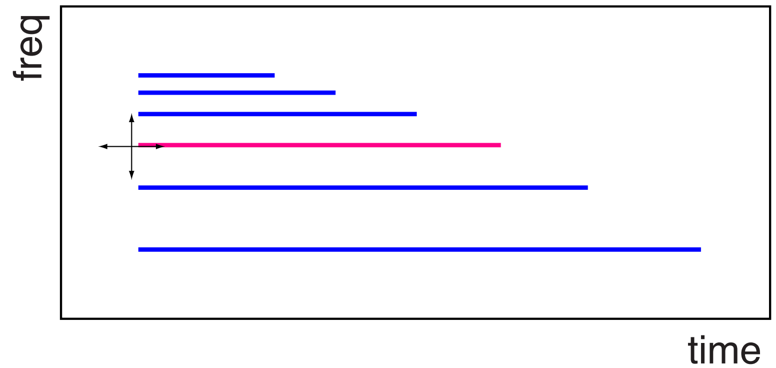


(after Darwin 1996)

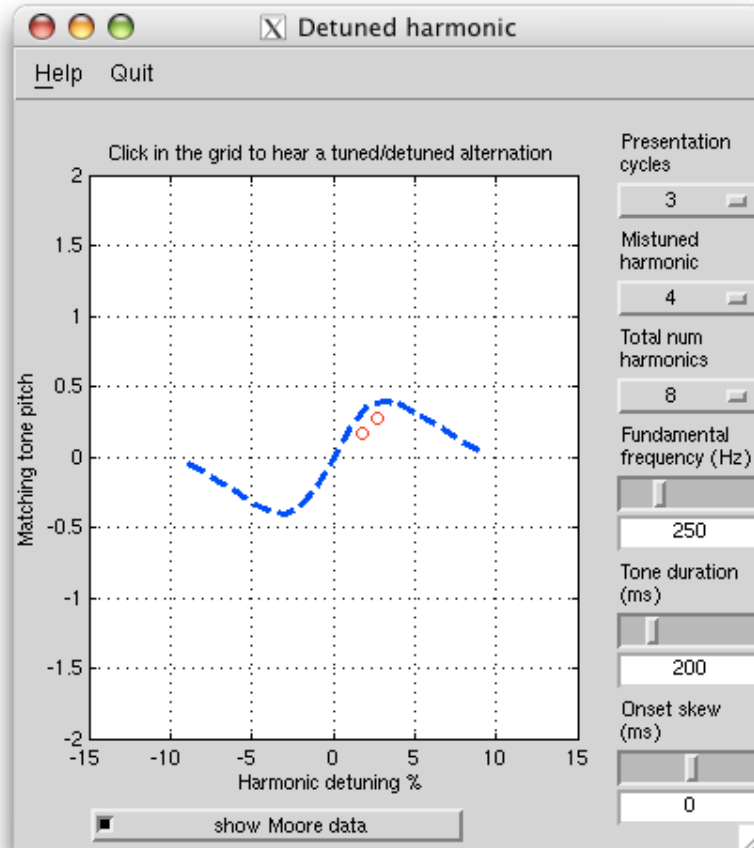
- Also learned "schema" (for speech etc.)

Perceiving Sources

- Harmonics distinct in ear, but perceived as one source ("fused"):



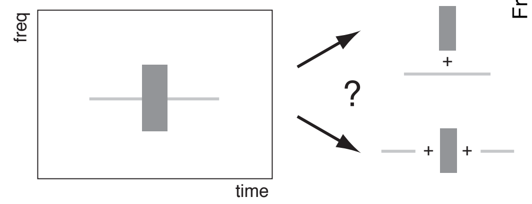
  - depends on common onset
  - depends on harmonics

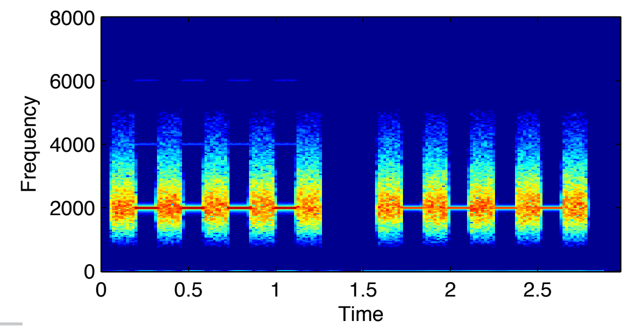


- Experimental techniques
  - ask subjects "how many"
  - match attributes e.g. pitch, vowel identity
  - brain recordings (EEG "mismatch negativity")

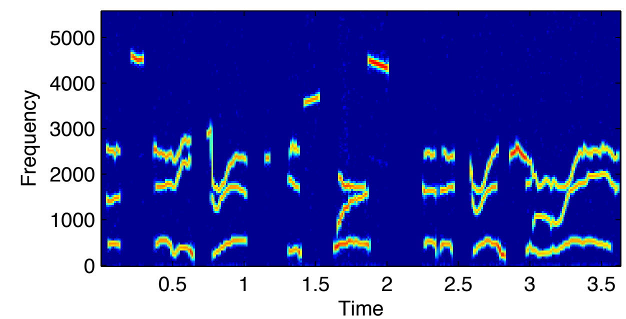
Auditory "Illusions"

- How do we explain illusions?
  - pulsation threshold

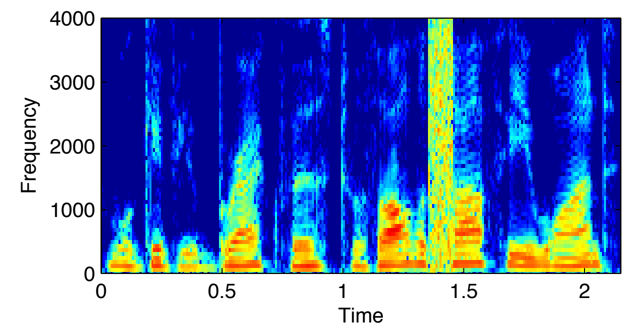




  - sinewave speech



  - phonemic restoration



- Something is providing the missing (illusory) pieces ... source models

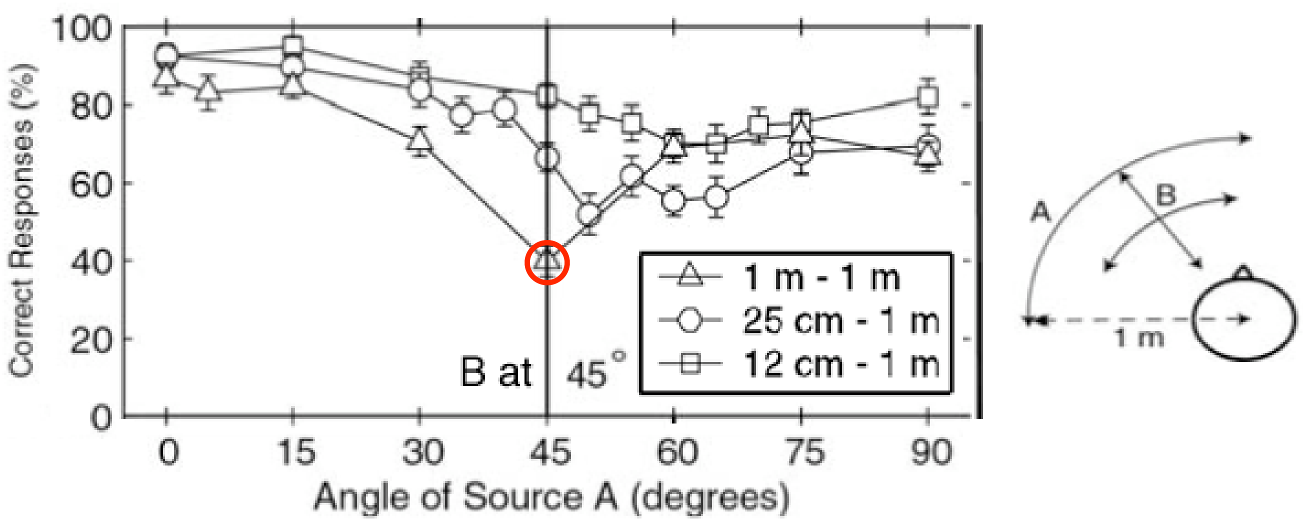
Human Speech Separation

Brungart et al.'02

- Task: Coordinate Response Measure
  - "Ready Baron go to green eight now"
  - 256 variants, 16 speakers
  - correct = color and number for "Baron"

crm-11737+16515.wav

- Accuracy as a function of spatial separation:

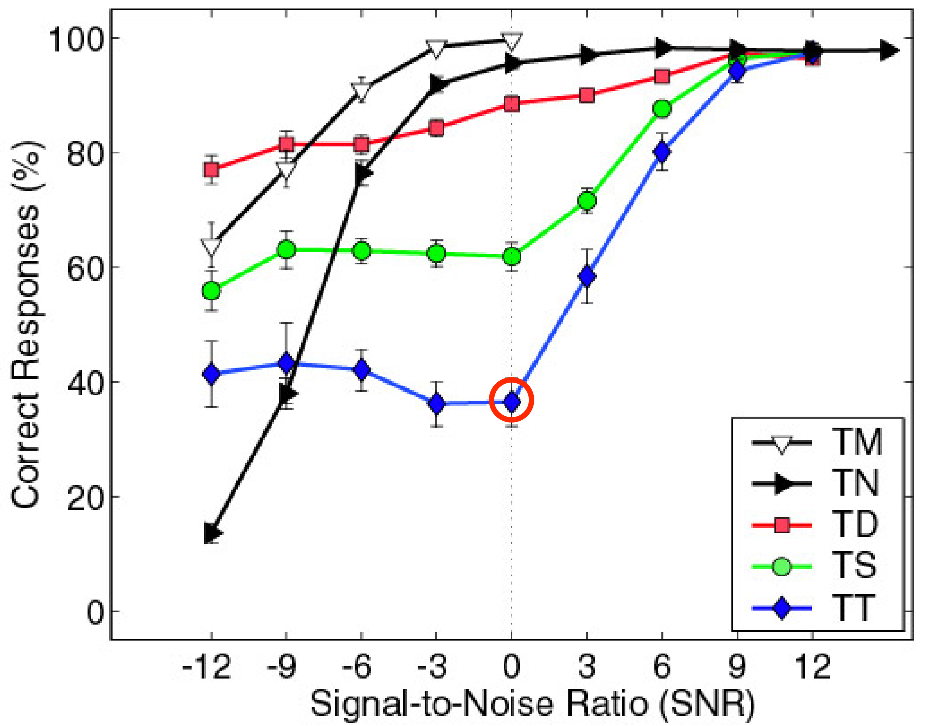


- A, B same speaker
- Range effect

Separation by Vocal Differences

Brungart et al.'01

- CRM varying the level and voice character



(same spatial location)

- energetic vs. informational masking
- more than pitch .. source models

Outline

1. Mixtures and Models
2. Human Sound Organization
3. Machine Sound Organization
   - Computational Auditory Scene Analysis
   - Dictionary Source Models
4. Ambient Sounds

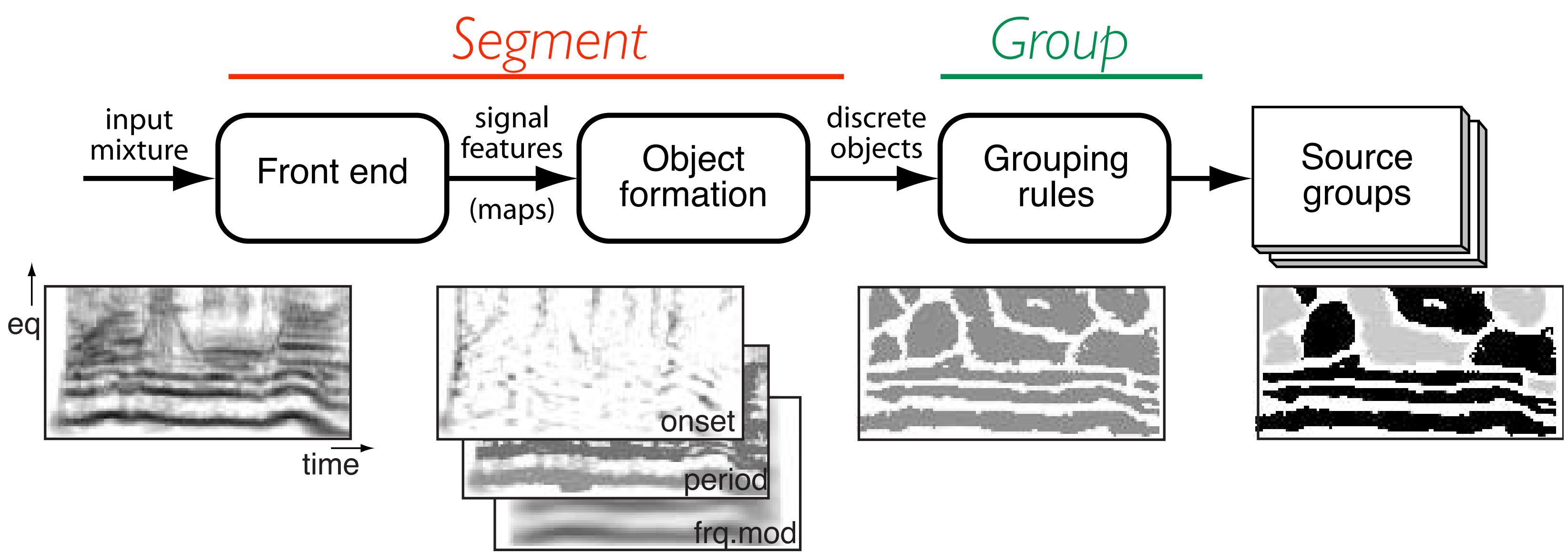
Source Model Issues

- Domain
  - parsimonious expression of constraints
  - nice combination physics
- Tractability
  - size of search space
  - tricks to speed search/inference
- Acquisition
  - hand-designed vs. learned
  - static vs. short-term
- Factorization
  - independent aspects
  - hierarchy & specificity

Computational Auditory Scene Analysis

Brown & Cooke'94
Okuno et al.'99
Hu & Wang'04 ...

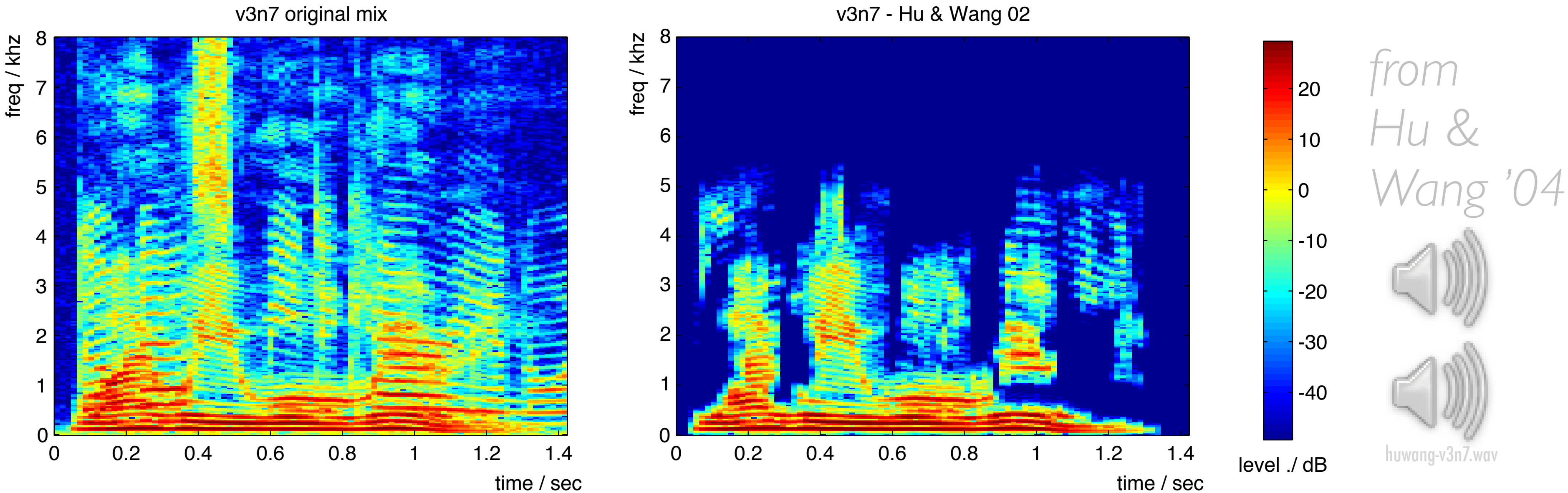
- Central idea:
  Segment time-frequency into sources based on perceptual grouping cues



  - ... principal cue is harmonicity

CASA limitations

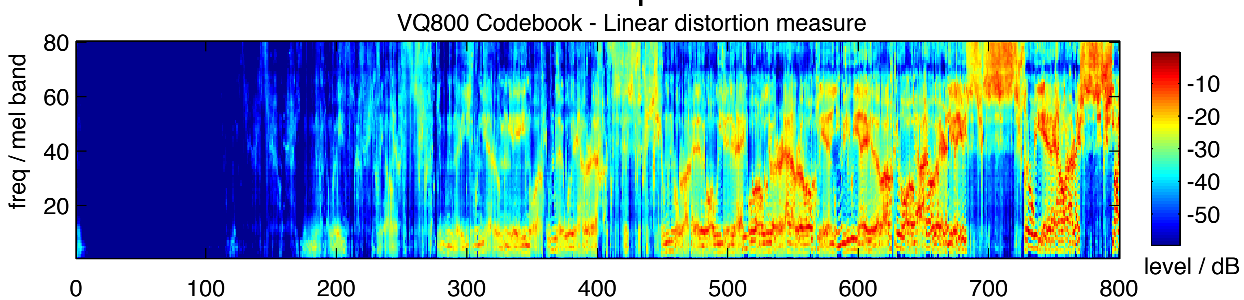
- **Limitations of T-F masking**
  - cannot undo overlaps – leaves gaps



- **Typically driven by local features**
  - limited model scope → no inference or illusions
- **Processing hand-defined, not learned**

Can Models Do CASA?

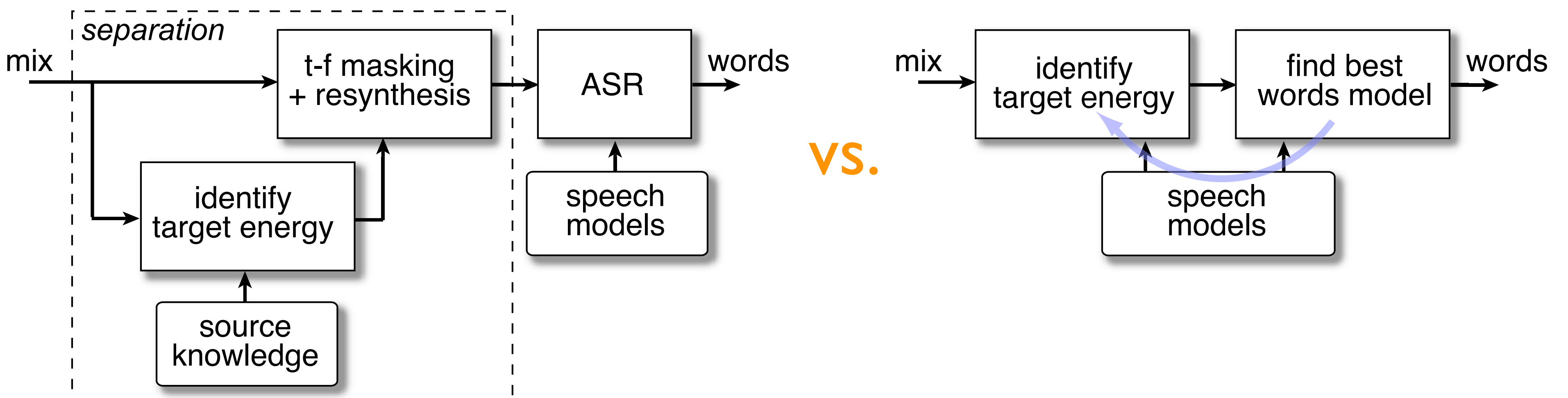
- Source models can learn harmonicity, onset
  - ... to subsume rules/representations of CASA



  - can capture spatial info too *[Pearlmutter & Zador'04]*
- Can also capture sequential structure
  - e.g. consonants follow vowels
  - ... like people do?
- But: need source-specific models
  ... for every possible source
  - use model adaptation? *[Ozerov et al. 2005]*

Separation or Description?

- Are isolated waveforms required?
  - clearly sufficient, but may not be necessary
  - not part of perceptual source separation!
- Integrate separation with application?
  - e.g. speech recognition



  - words output = abstract description of signal

Dictionary Models

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

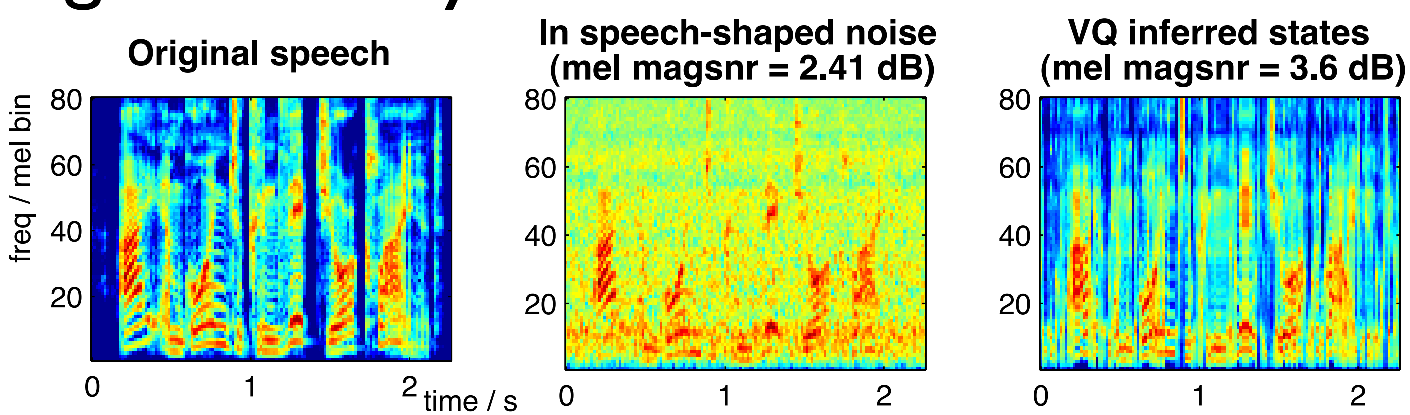
- Given models for sources, find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1, i_2) = \mathcal{N}(\mathbf{x}; \mathbf{c}_{i1} + \mathbf{c}_{i2}, \Sigma)$$ 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...
  - different domains for combining $\mathbf{c}$ and defining Σ

- E.g. stationary noise:



Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

Speech Recognition Models

- Cooke & Lee Speech Separation Challenge
  - short, grammatically-constrained utterances:

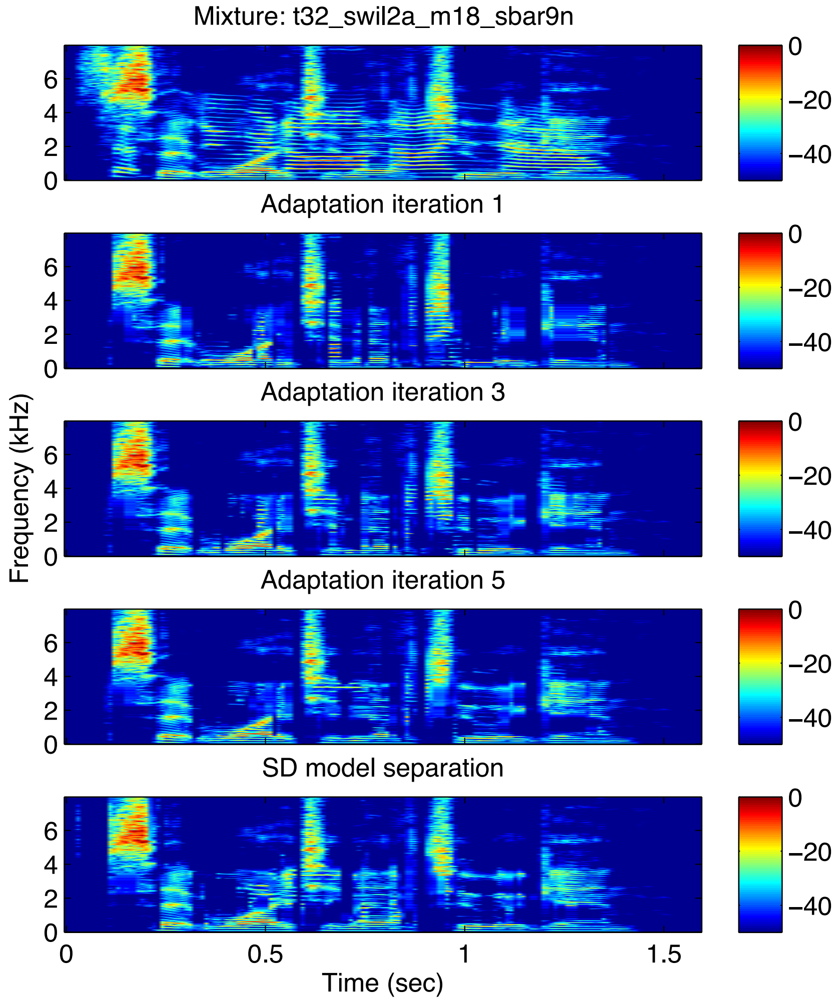
    <command:4><color:4><preposition:4><letter:25><number:10><adverb:4>

    e.g. "bin white by R 8 again"
  - task: report letter+number for "white"
- Decode with Factorial HMM
  - i.e. two state sequences, one model for each voice
  - exploit sequence constraints
  - exploit speaker differences
- IBM "superhuman" system *Kristjansson, Hershey et al. '06*
  - fewer errors than people for same speaker, level
  - exploits known speakers, limited grammar

Speaker-Adapted (SA) Models

Ron Weiss

- Factorial HMM needs distinct speakers

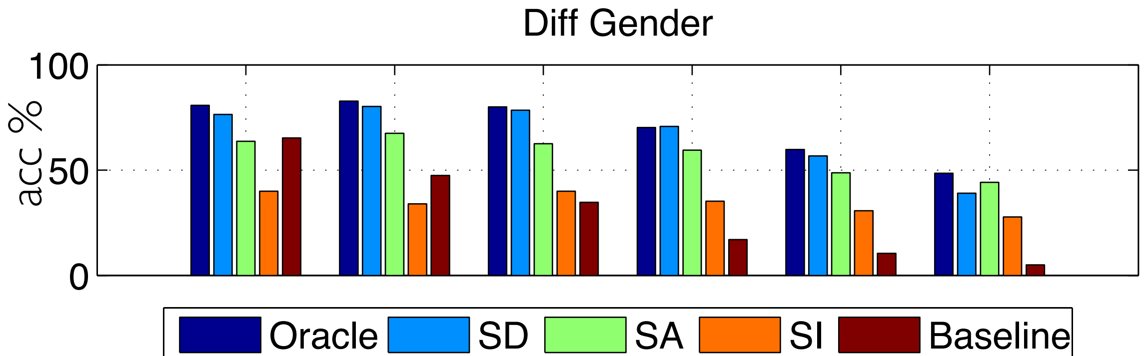


- use "eigenvoice" speaker space
- iterate estimating voice & separating speech
- performs midway between speaker-independent (SI) and speaker-dependent (SD)

SI

SA

SD

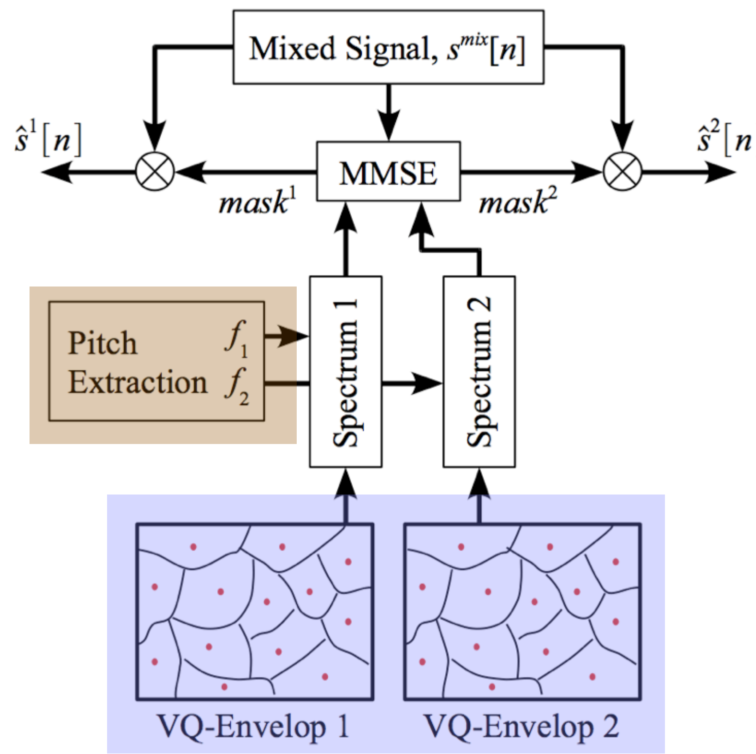


(Pitch) Factored Dictionaries

Ghandi & Has-John. '04
Radfar et al. '06

- Separate representations for "source" (pitch) and "filter"
  - *NM* codewords from *N+M* entries
  - but: overgeneration...

- Faster search
  - direct extraction of pitches
  - immediate separation of (most of) spectra



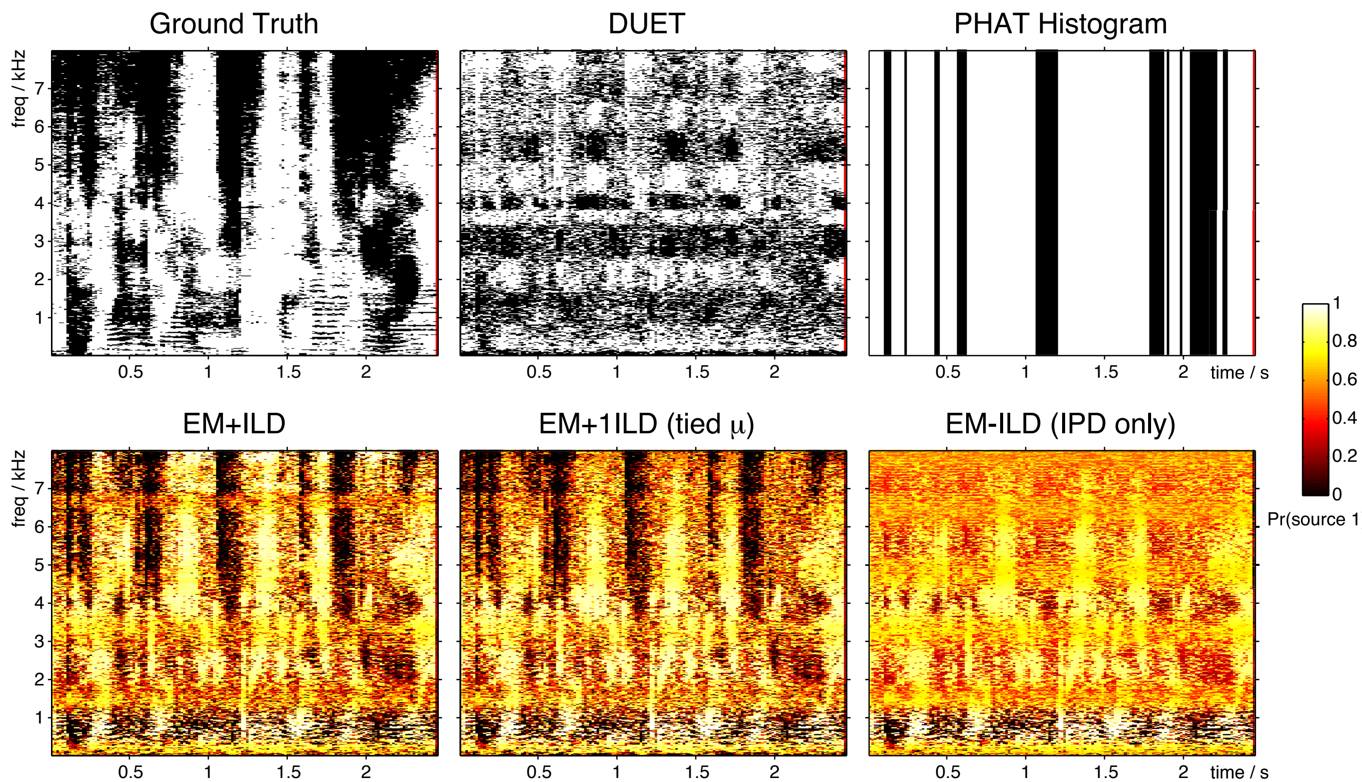
Outline

1. Mixtures & Models
2. Human Sound Organization
3. Machine Sound Organization
4. **Ambient Sounds**
    - binaural separation
    - “personal audio” analysis

Binaural Localization by EM

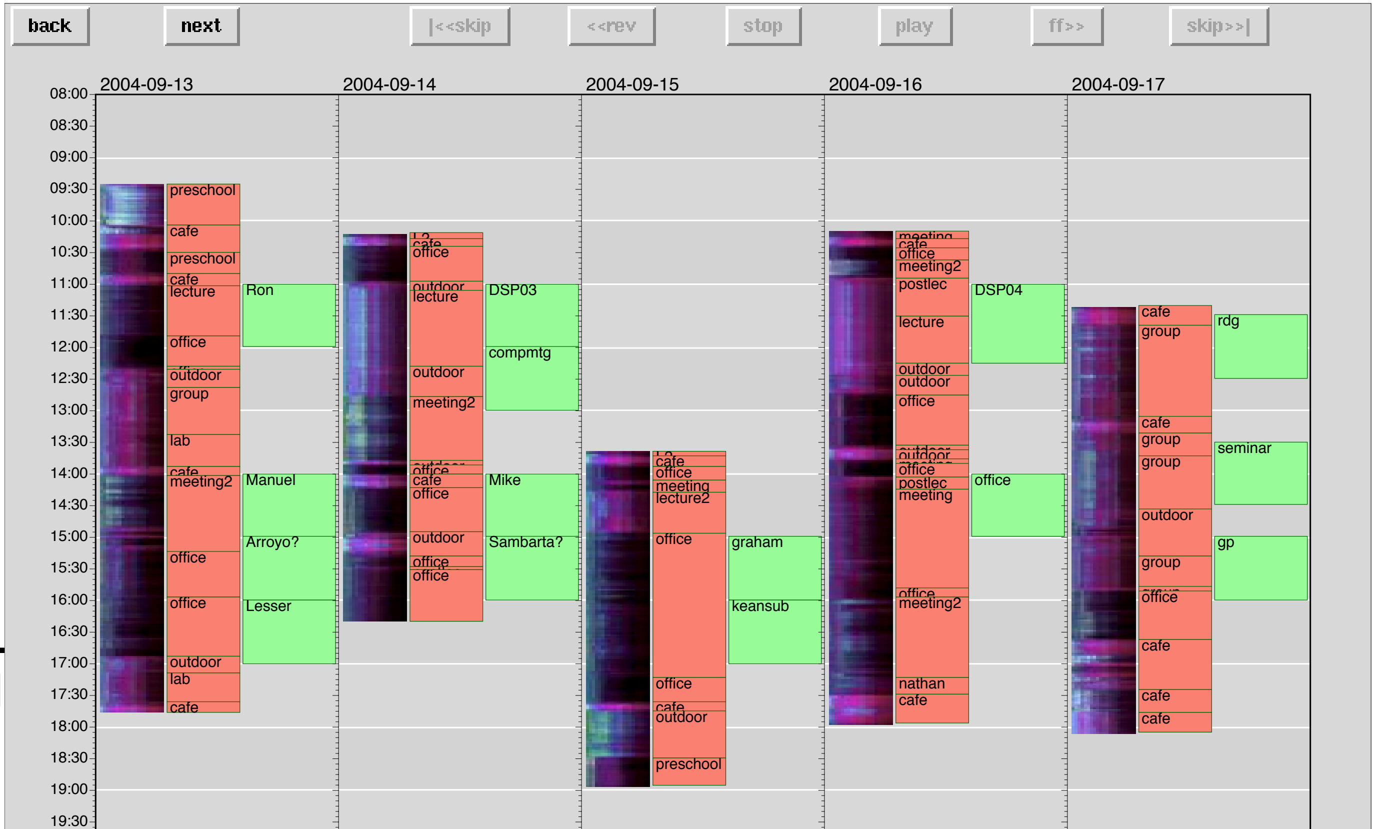
Mike Mandel, NIPS'06

- 2 or 3 sources in reverberation
- Iteratively estimate ILD, IPD
  - initialize from PHAT ITD histogram
  - output is soft TF mask



“Personal Audio” Archives

- Continuous recordings with MP3 player
- Segment / cluster “episodes”
  - .. by statistics of ~10 s segments
  - .. for browsing interface



Lab ROSA

Laboratory for the Recognition and Organization of Speech and Audio

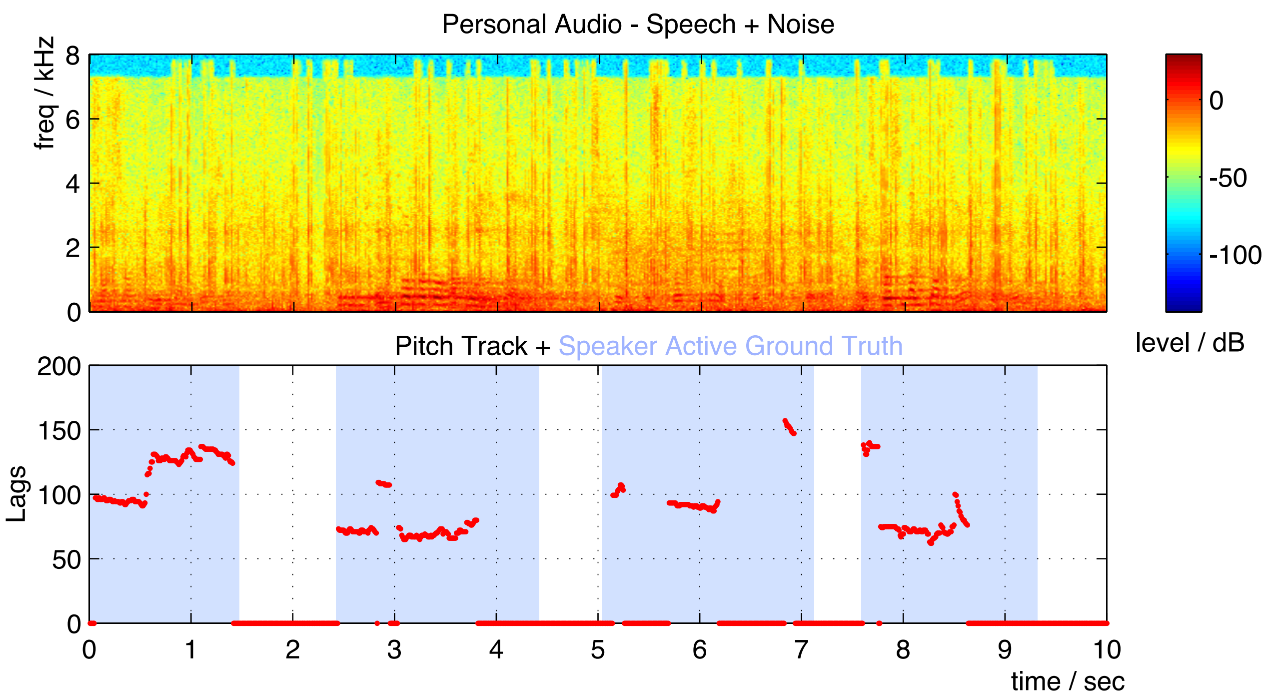
Sound

Columbia University in the city of New York

Personal Audio Speech Detection

Keansub Lee, Interspeech'06

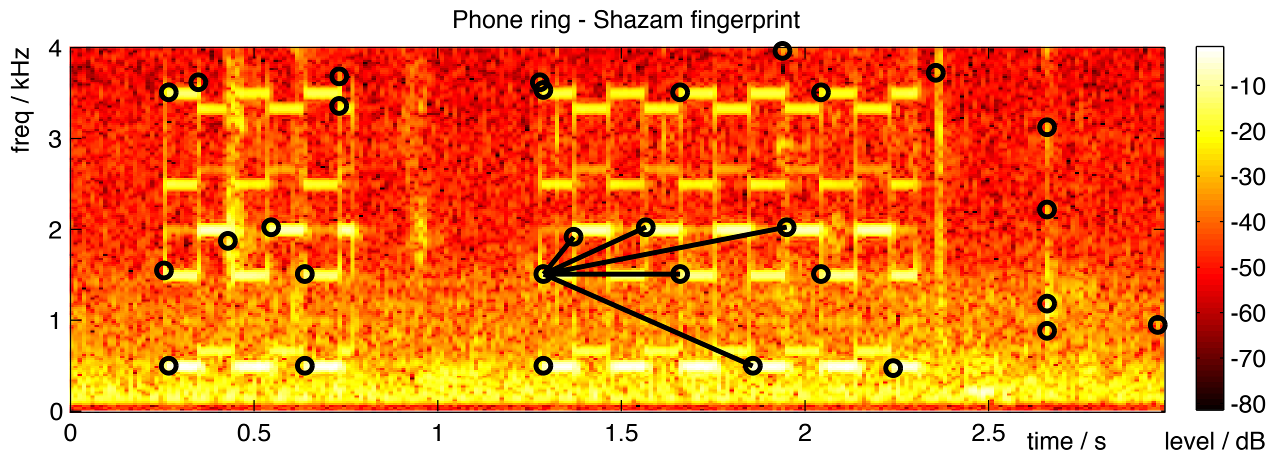
- **Pitch is last speech cue to disappear**
  - noise robust pitch tracker for voice detection
  - biggest problem was periodic noise (air conditioning)



Repeating Events in Personal Audio

Jim Ogle, ICASSP'07

- "Unsupervised" feature to help browsing
- Full NxN search is very expensive
  - use Shazam fingerprint hashes to find repeats



  - only works for exact repeats (alarms, jingles)
- O(N) scan for repeats
  - fixed-size hash table
  - multiple common hashes → confident match

Summary & Conclusions

- Listeners do well separating sound mixtures
  - using signal cues (location, periodicity)
  - using source-property variations
- Machines do less well
  - difficult to apply enough constraints
  - need to exploit signal detail
- Models capture constraints
  - learn from the real world
  - adapt to sources
- Separation feasible only sometimes
  - describing source properties is easier