Sound Organization By Source Models in Humans and Machines

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

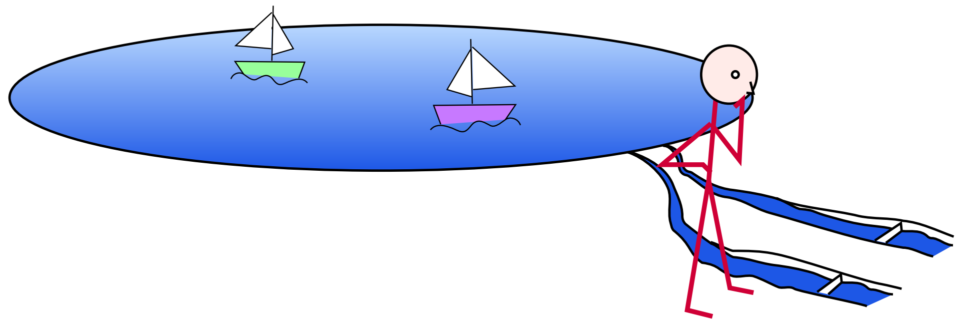
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Mixtures and Models
2. Human Sound Organization
3. Machine Sound Organization
4. Research Questions

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

Columbia University
in the City of New York

The Problem of Mixtures

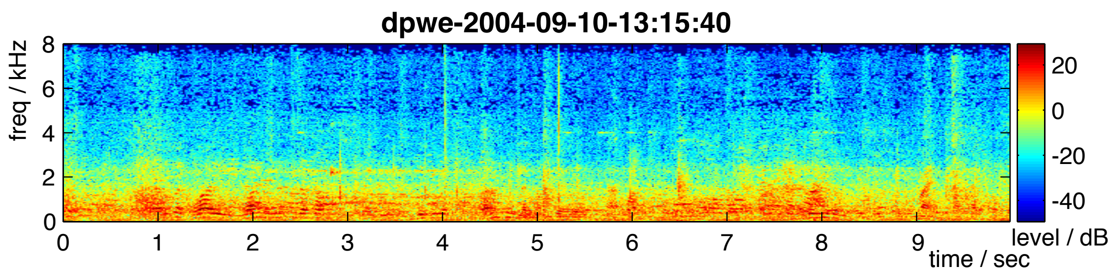


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- 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



pa-2004-09-10-131540.wav

- 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

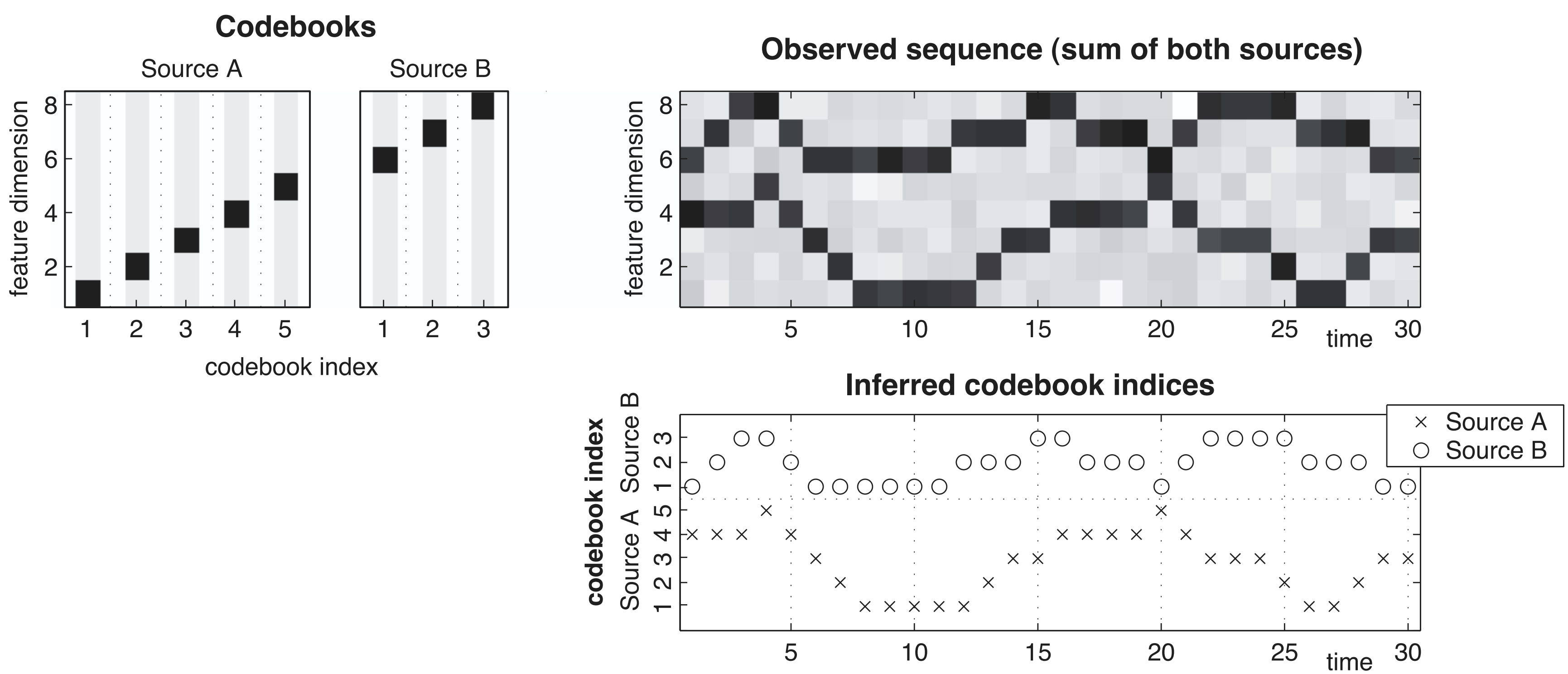
- Ideal separation is rarely possible
  - i.e. no projection can completely remove overlaps
- Overlaps → Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

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

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

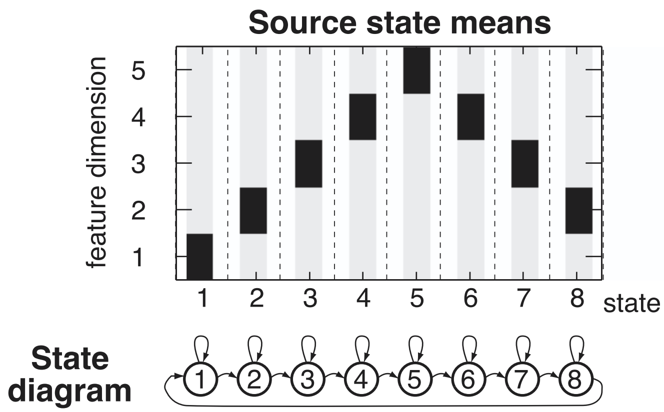
A Simple Example

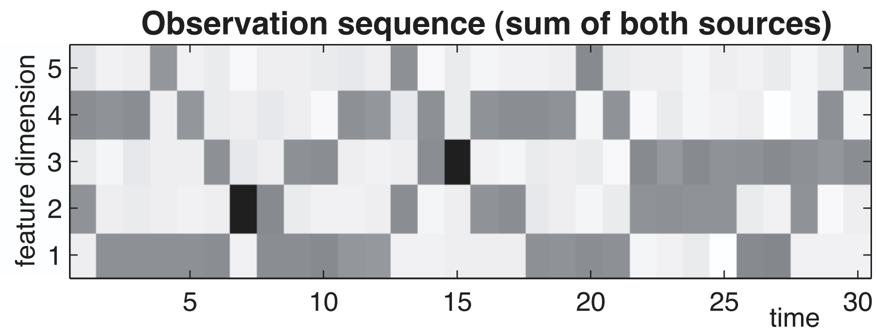
- Source models are codebooks from separate subspaces

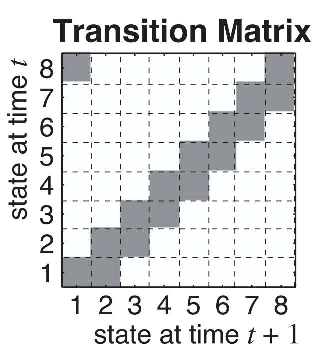


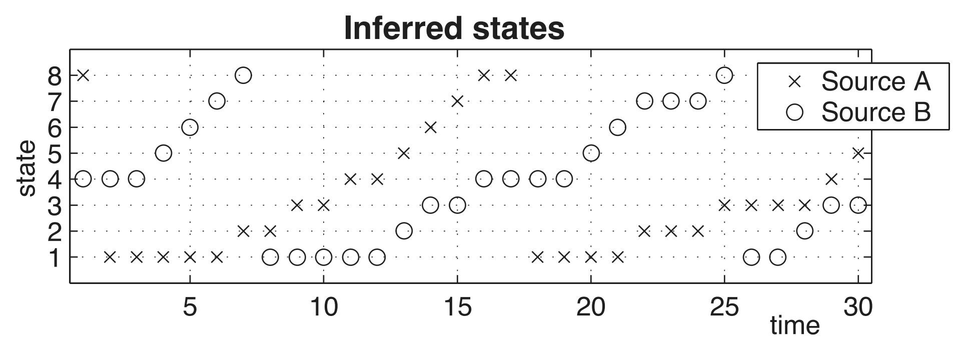
A Slightly Less Simple Example

- Sources with Markov transitions





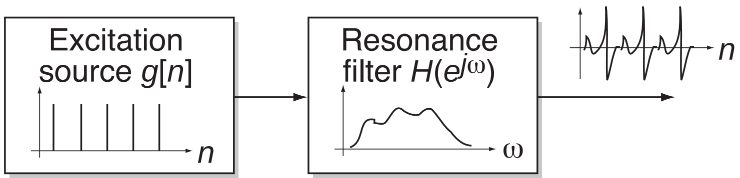




What is a Source Model?

- **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

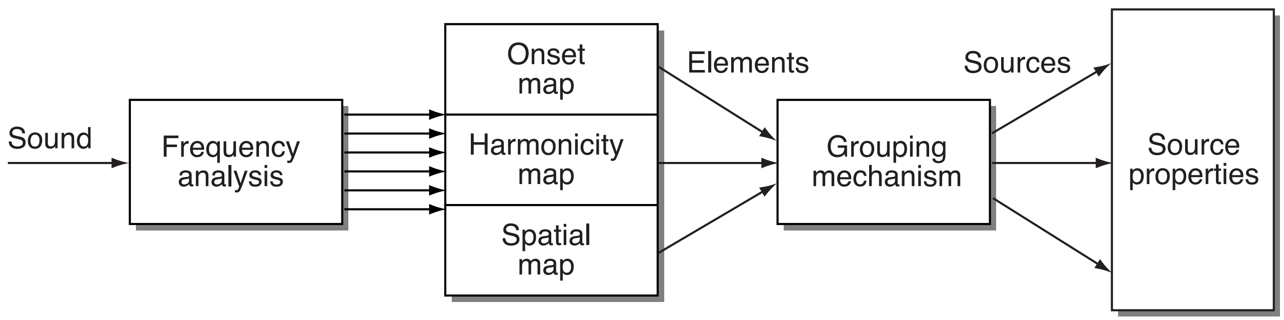
Outline

1. Mixtures and Models
2. Human Sound Organization
   - Auditory Scene Analysis
   - Using source characteristics
   - Illusions
3. Machine Sound Organization
4. Research Questions

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, ...

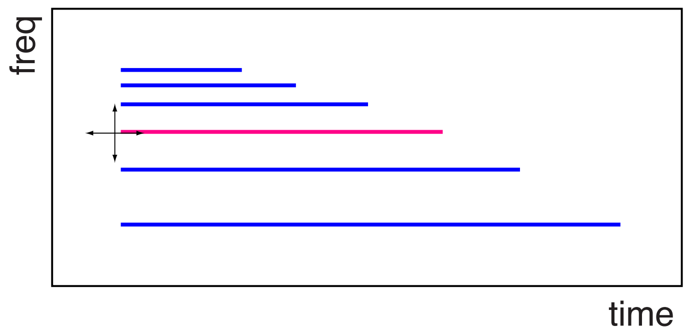


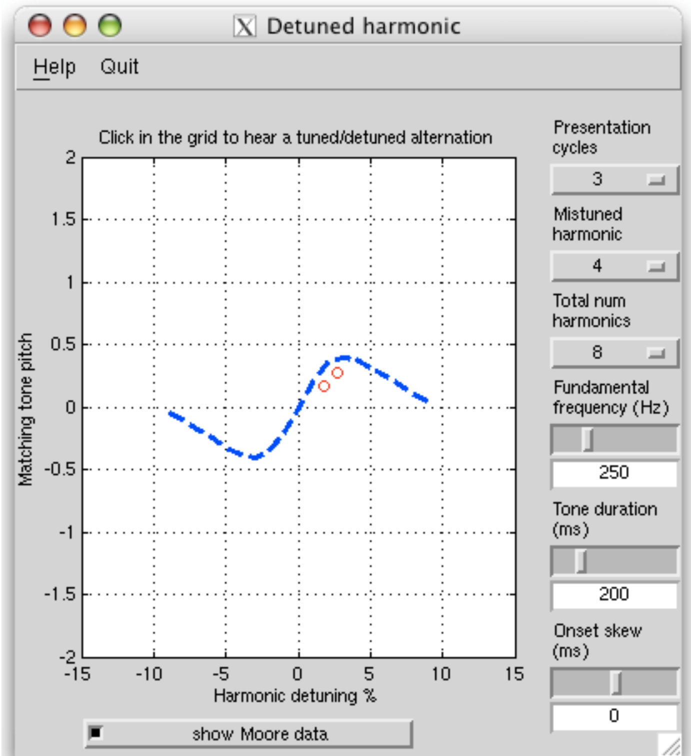
(after Darwin 1996)

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

Perceiving Sources

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

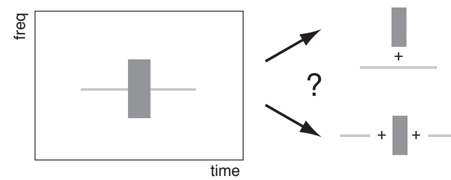


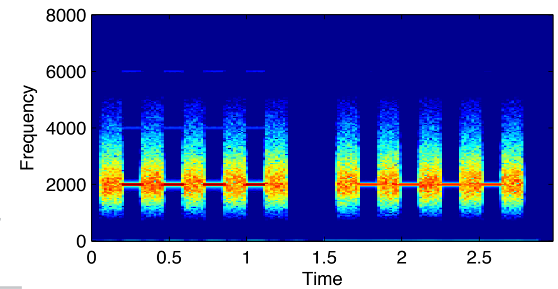


  - depends on common onset
  - depends on harmonics
- Experimental techniques
  - ask subjects "how many"
  - match attributes e.g. pitch, vowel identity
  - brain recordings (EEG "mismatch negativity")

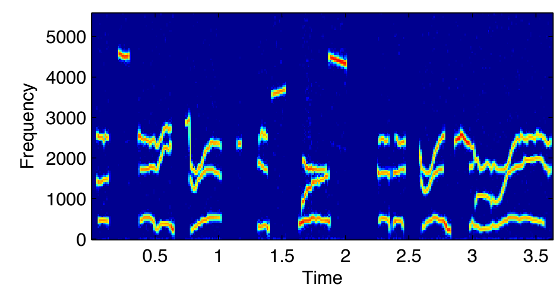
Auditory "Illusions"

- How do we explain illusions?
  - pulsation threshold

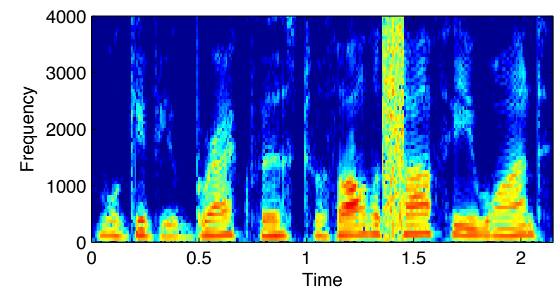




  - sinewave speech



  - phonemic restoration



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

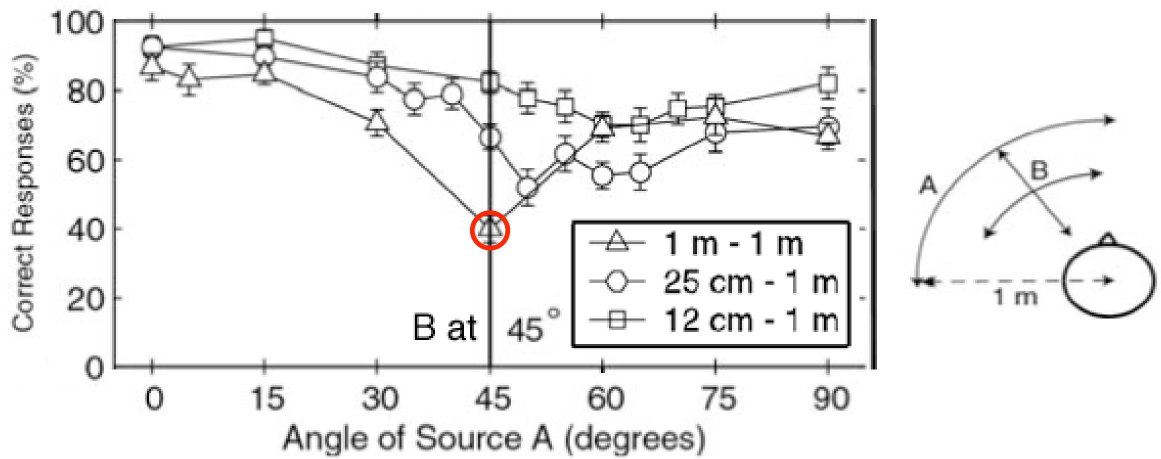
Human Speech Separation

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

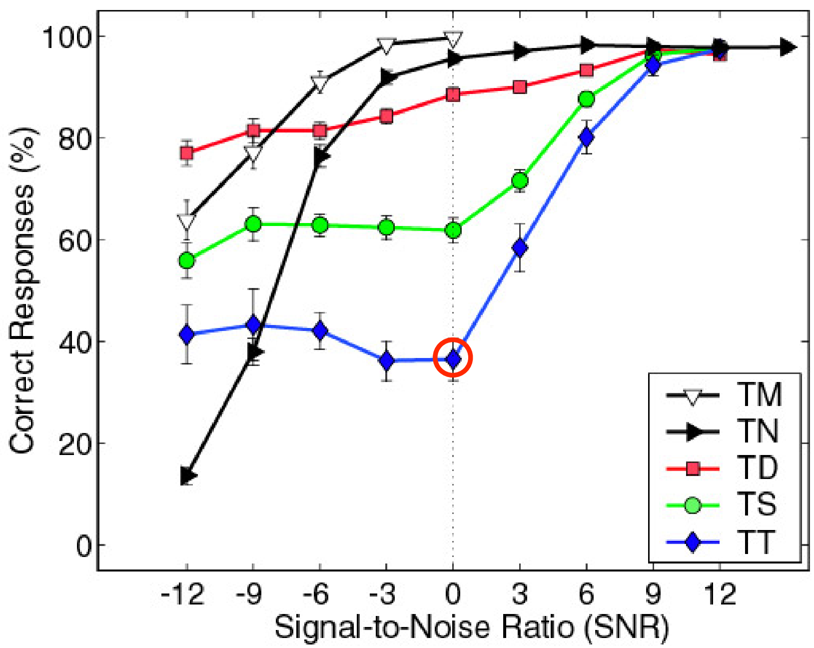
LabROSA Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

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. Research Questions

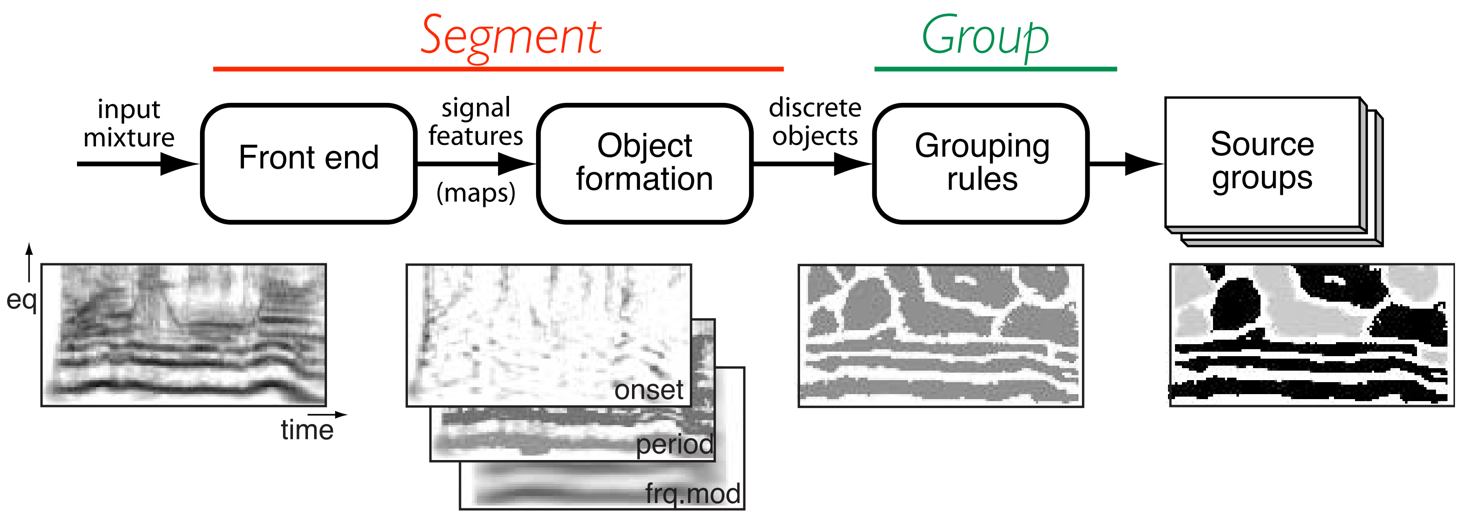
Source Model Issues

- 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 ...

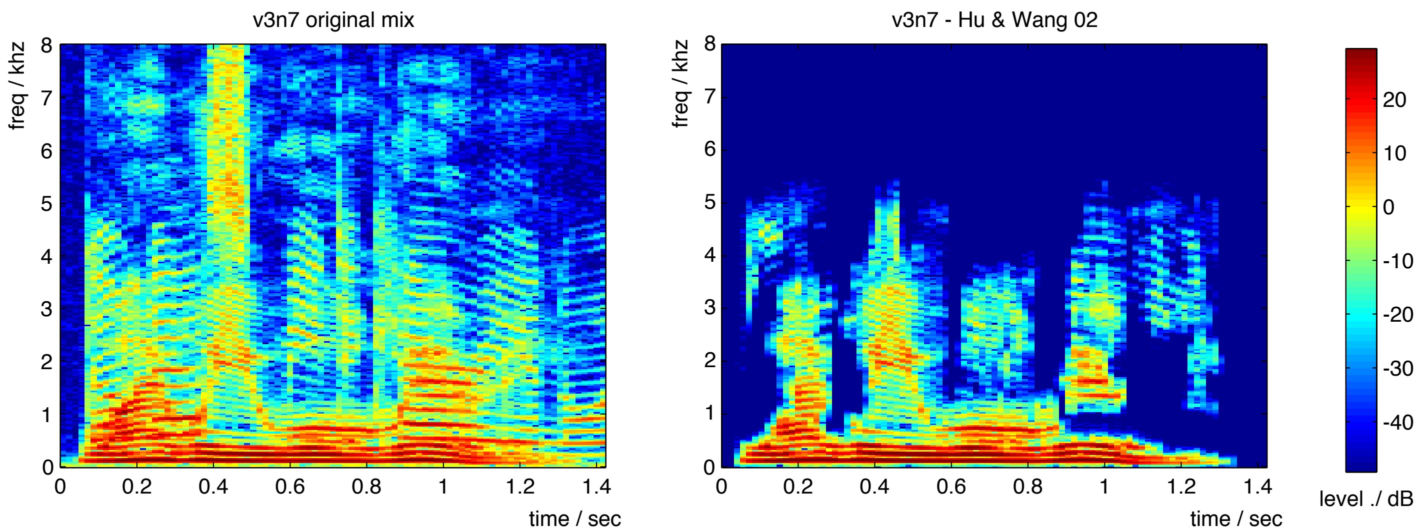
- Central idea:
  Segment time-frequency into sources based on perceptual grouping cues



  - ... principal cue is harmonicity

CASA limitations

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



from Hu & Wang '04

huwang-v3n7.wav

- Driven by local features
  - limited model scope → no inference or illusions
- Does not learn from data

Basic Dictionary Models

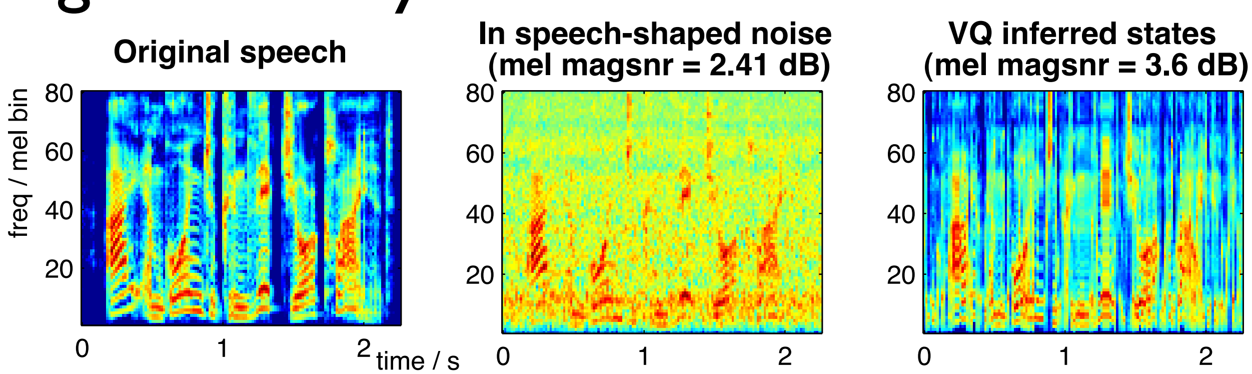
Roweis '01, '03
Kristjannson '04, '06

- 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

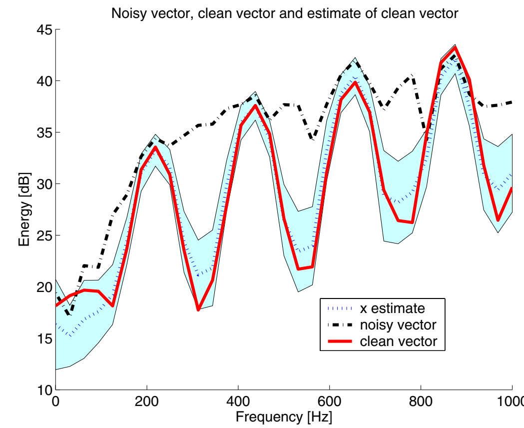
  - can include sequential constraints...
  - different domains for combining $\mathbf{c}$ and defining Σ
- E.g. stationary noise:



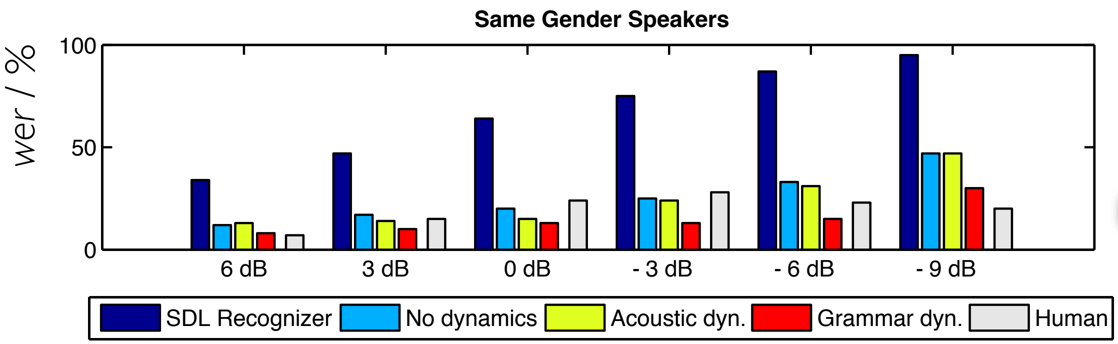
Deeper Models: Iriquois

Kristjansson, Hershey et al. '06

- Optimal inference on mixed spectra
  - speaker-specific models (e.g. 512 mix GMM)
  - Algonquin inference



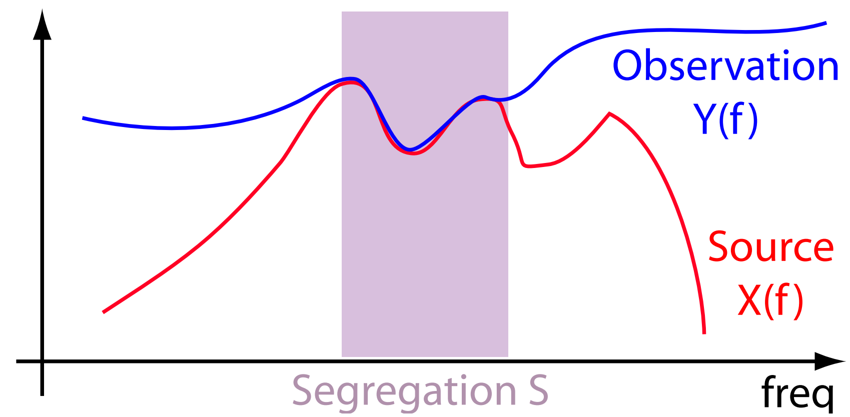
- .. for Speech Separation Challenge (Cooke/Lee'06)
  - exploit grammar constraints - higher-level dynamics



Faster Search: Fragment Decoder

Barker et al. '05

- Match 'uncorrupt' spectrum to ASR models using **missing data recognition**
  - easy if you know the segregation mask S



- Joint search for model M and segregation S to maximize:

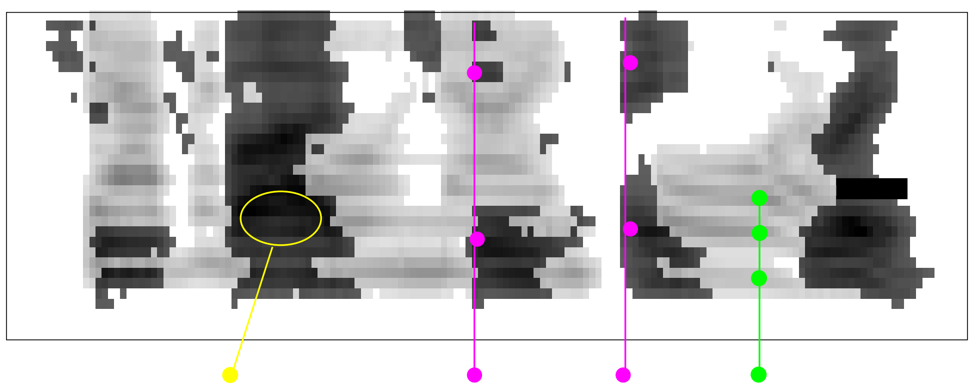
$$P(M, S|Y) = P(M)\int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)$$

Isolated Source Model *Segregation Model*

CASA in the Fragment Decoder

$$P(M,S|Y) = P(M)\int P(X|M)\cdot \frac{P(X|Y,S)}{P(X)}dX \cdot P(S|Y)$$

- CASA can help search
  - consider only segregations made from CASA chunks
- CASA can rate segregation
  - construct $P(S|Y)$ to reward CASA qualities:

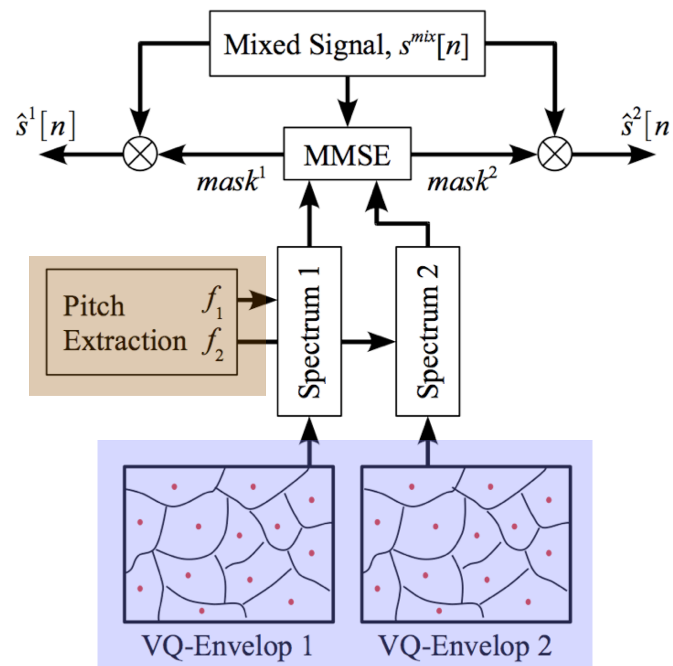


(Pitch) Factored Dictionaries

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

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

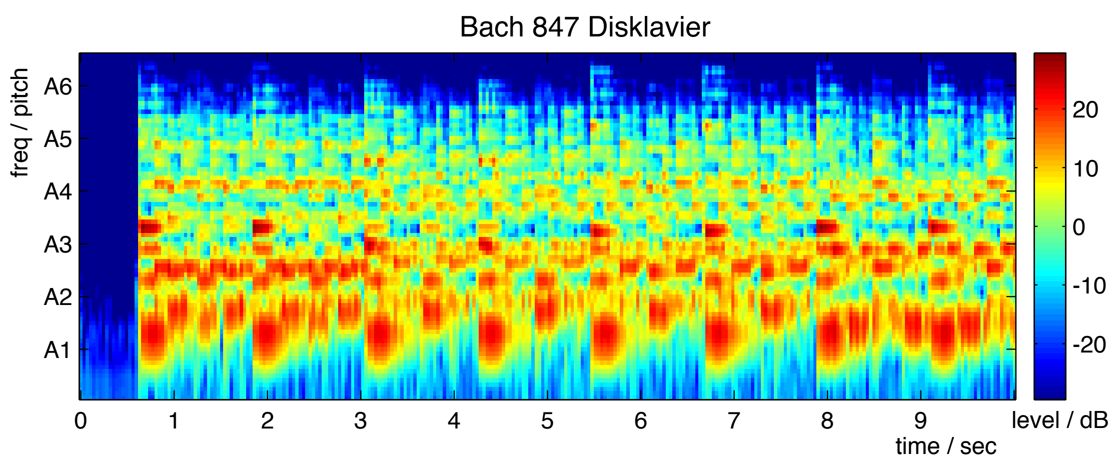
- Faster search
  - direct extraction of pitches
  - immediate separation of (most of) spectra



Discriminant Models for Music

Poliner & Ellis '06

- Transcribe piano recordings by classification
  - train SVM detectors for every piano note
  - 88 separate detectors, independent smoothing
- Trained on player piano recordings



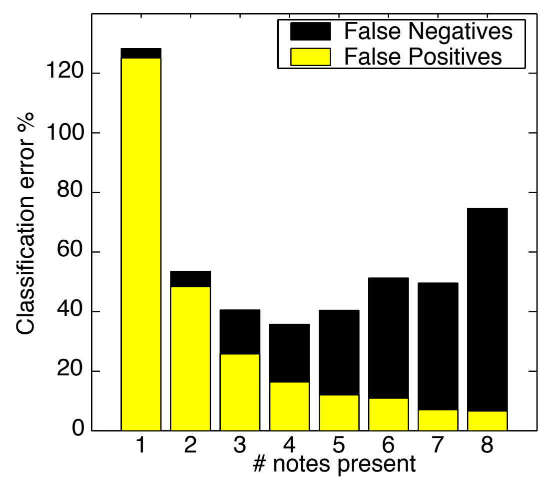
- Can resynthesize from transcript...

Piano Transcription Results

- Significant improvement from classifier:
  - frame-level accuracy results:

Algorithm	Errs	False Pos	False Neg	d'
SVM	43.3%	27.9%	15.4%	3.44
Klapuri&Ryynänen	66.6%	28.1%	38.5%	2.71
Marolt	84.6%	36.5%	48.1%	2.35

  - Breakdown by frame type:
  -

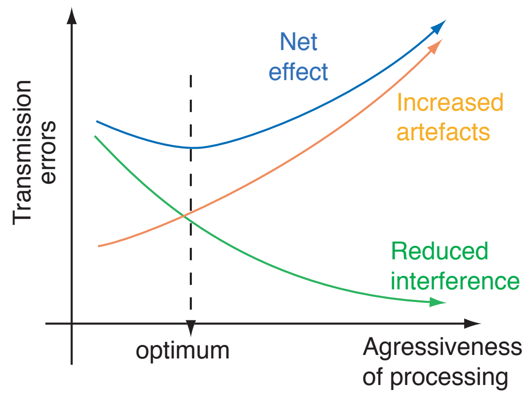


Outline

1. Mixtures & Models
2. Human Sound Organization
3. Machine Sound Organization
4. **Research Questions**
    - Task and Evaluation
    - Generic vs. Specific

Task & Evaluation

- How to measure separation performance?
  - depends what you are trying to do
- SNR?
  - energy (and distortions) are not created equal
  - different nonlinear components *[Vincent et al. '06]*
- Human Intelligibility?
  - rare for nonlinear processing to improve intelligibility
  - listening tests expensive

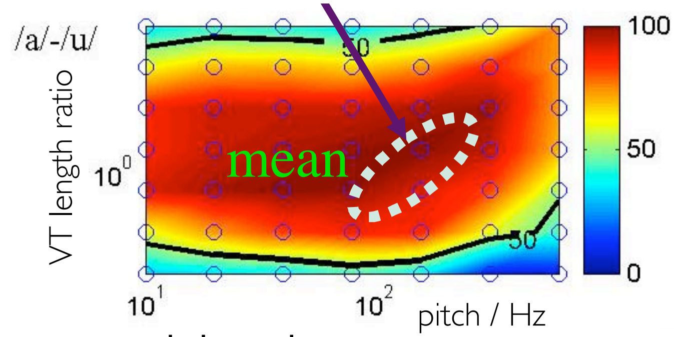


- ASR performance?
  - separate-then-recognize too simplistic; ASR needs to accommodate separation

How Many Models?

- More specific models → better separation
  - need individual dictionaries for "everything"??
- Model adaptation and hierarchy
  - speaker adapted models : base + parameters
  - extrapolation beyond normal

Smith, Patterson et al. '05



  - generic-specific: pitched → piano → this piano
- Time scales of model acquisition
  - innate/evolutionary (hair-cell tuning)
  - developmental (mother tongue phones)
  - dynamic - the "slung mugs" effect; Ozerov

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