Model-Based Scene 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. Separation and Inference
2. Model-based separation
3. Speech Fragment Decoder

1. Separation and Inference

- **Full separation requires "separable dimension"**
  - e.g. spatial filtering
  - but for single channel: overlap is inevitable
- **Signal knowledge provides extra constraints**
  - .. for inference of missing parts
- **Separation vs. recognition**
  - separation is *sufficient* .. but too hard
  - recognition is easier .. but too coarse
  - in-between: class plus parameters
    adequate for resynthesis?

Pattern Recognition Perspective

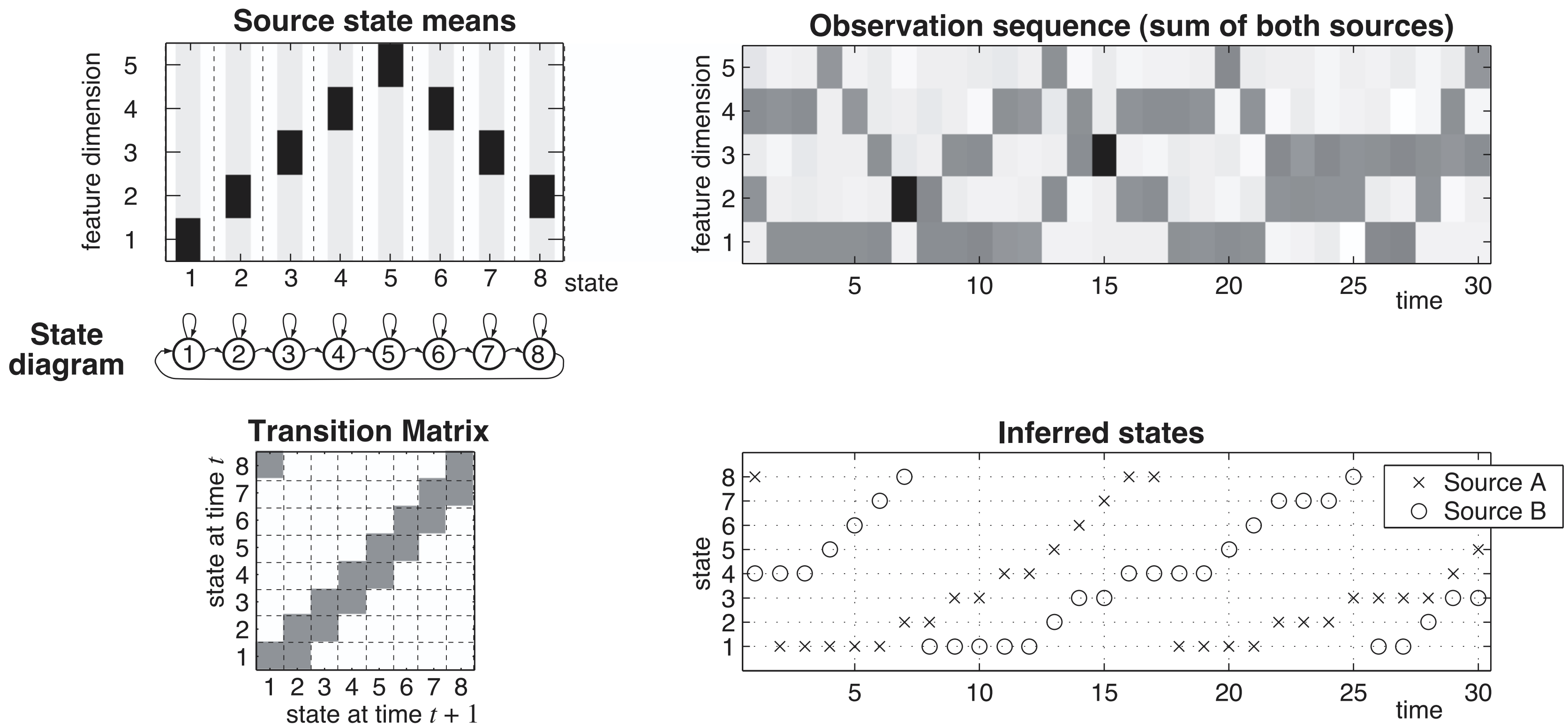
- Inferring source signal set $\{s_i\}$ from mixture signal x:

$$\underset{\{s_i\}}{\arg\max}\; p(x|\{s_i\}) \sum_i p(s_i|M_i)$$

  - $p(x|\{s_i\})$ gives physics of combination (sum)
  - $p(s_i|M_i)$ limits which s_i to consider
- How to acquire/evaluate $p(s_i|M_i)$?
  - generalize observation of solo sources
- How to search $\{s_i\}$?
  - full joint space?
  - clever pruning tricks

Factorial HMM - Toy Example

- Two sources with same underlying model

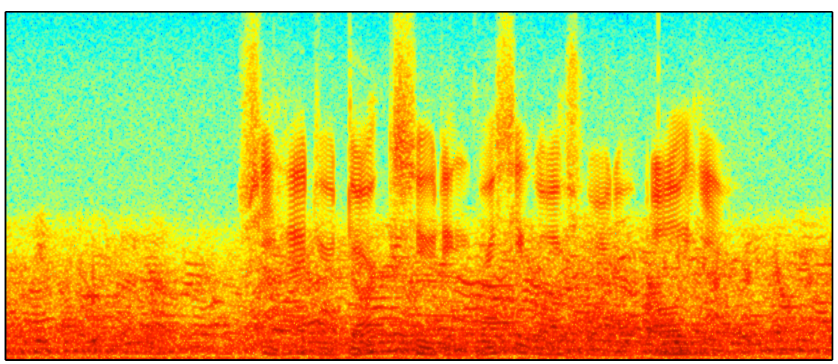


  - sequence constraints can disambiguate identical emissions

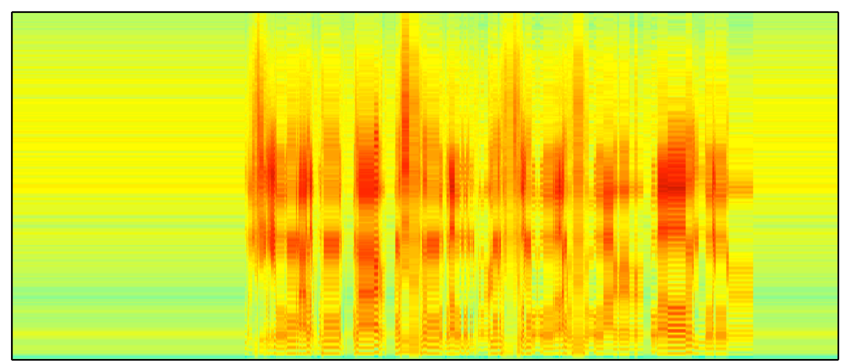
Disambiguating with Knowledge

(Roweis '03)

- Use strength of match to models as reasonableness measure for control
- e.g. MAXVQ
  - learn dictionary of spectrogram slices
  - find the ones that 'fit'
    - or max() of a combination....
  - ... then filter out excess energy



Noise-corrupt speech



Matching templates

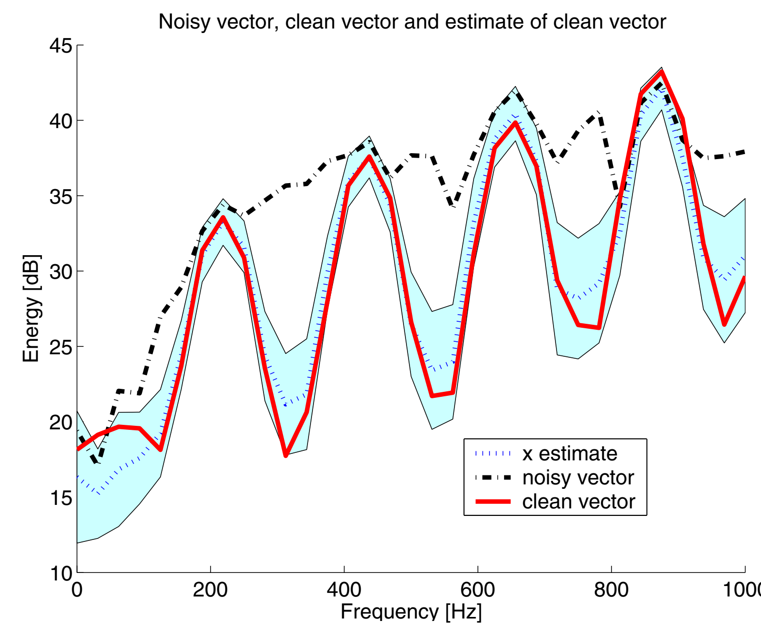
from Sam Roweis's Montreal 2003 presentation

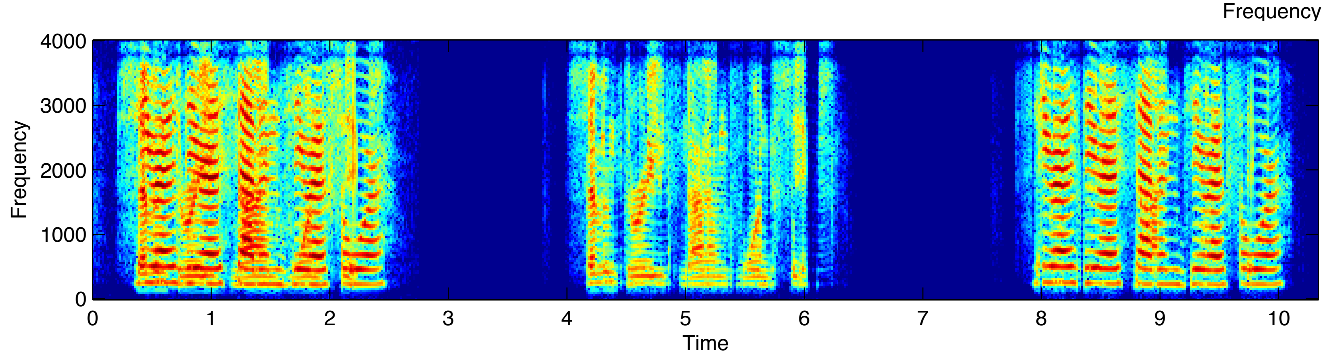
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio
Columbia University in the City of New York

Full Mixture Inference

(Kristjansson, Attias, Hershey'04)

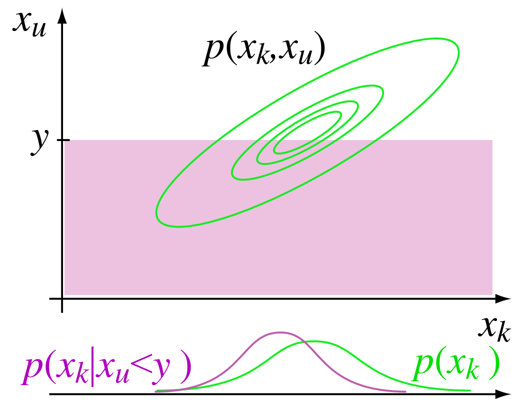
- Can model combination of magnitude spectra with stochastic model
    - phase cancellation as noise...
- Precise inference of components
    - by iterative linearization
- Works well
  (for small domains?)



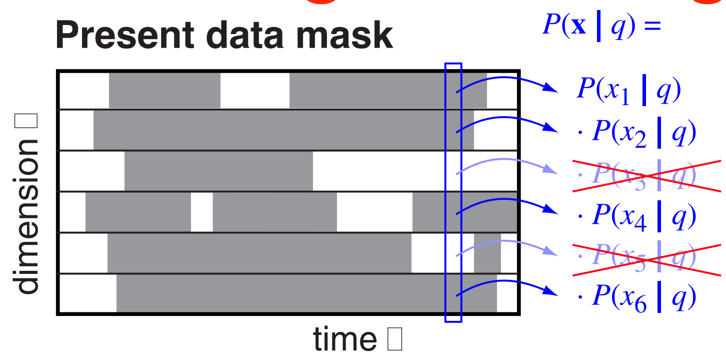


2. Missing Data Recognition

- Speech models $p(x|M)$ are multidimensional...
  - means, variances for each frequency channel
  - need values for all dimensions to get $p(\bullet)$
- But: can evaluate over a subset of dimensions x_k
  - $$p(x_k|M) = \int p(x_k, x_u|M)\,dx_u$$



- Hence, missing data recognition:

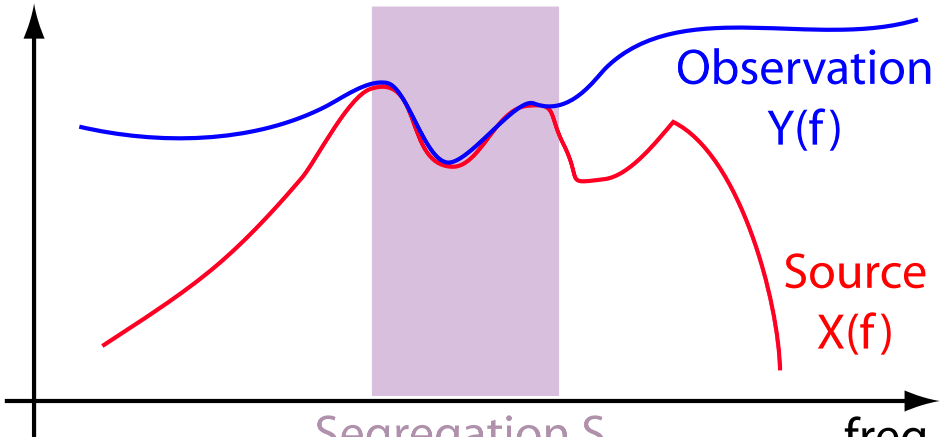


  - hard part is finding the mask (segregation)

The Speech Fragment Decoder

Barker, Cooke, Ellis '04

- Match 'uncorrupt' spectrum to ASR models using missing data



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

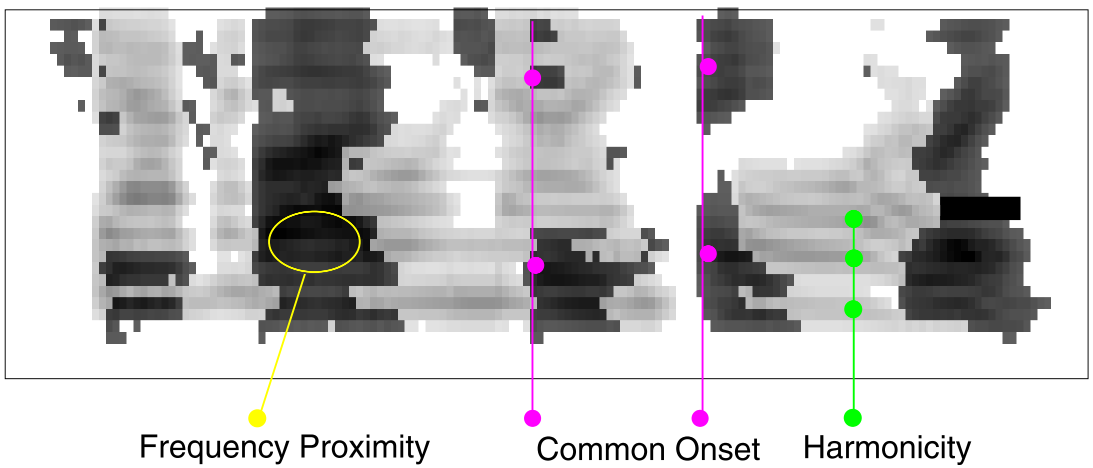
$$P(M, S|Y) = P(M)\int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)$$

Isolated Source Model *Segregation Model*

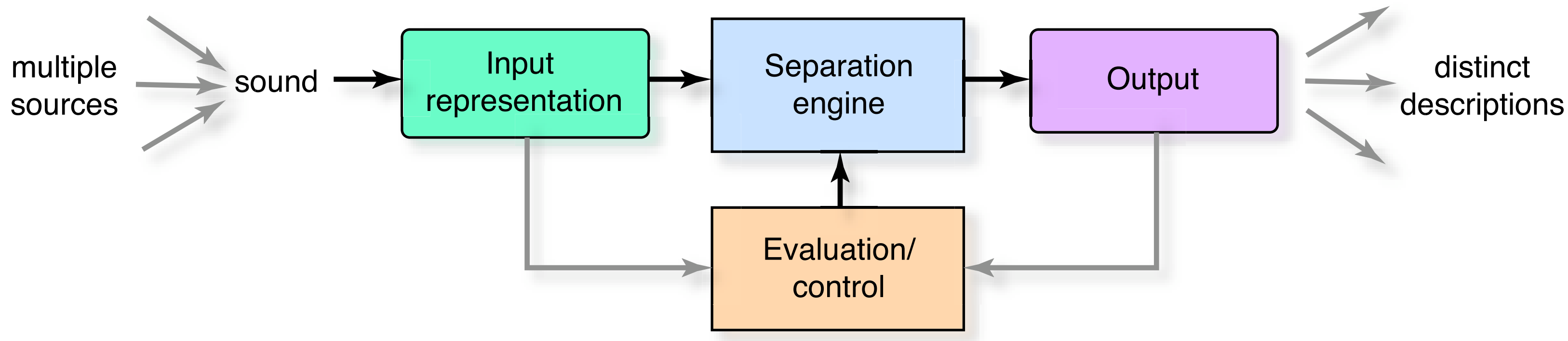
Using CASA cues

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

- CASA helps **search**
  - consider only segregations made from CASA chunks
- CASA rates **segregation**
  - construct $P(S|Y)$ to reward CASA qualities:



Learning for Separation



- Control: learn what is "reasonable"
- Input: discriminant features
  - learned subspaces
- Engine: clustering parameters
- Output: restoration...

Can Machine Learning Subsume CASA?

- ASA grouping cues describe real sounds
  - .. "anecdotally"
- Machine Learning is another way to find regularities in large datasets
  - can, e.g., Roweis templates subsume harmonicity, onset, etc.?
  - ... and handle schema at the same time?
  - "cut out the (grouping cue) middleman"
- Trick is how to represent/generalize
  - listeners can organize novel sounds

Conclusions

- Source separation needs constraints
  - e.g. prior knowledge of signal form
- Memorized signals (HMMs) can be powerful
  - but can get very large
- Speech recognition models can be co-opted
  - e.g. to identify plausible subsets of regions