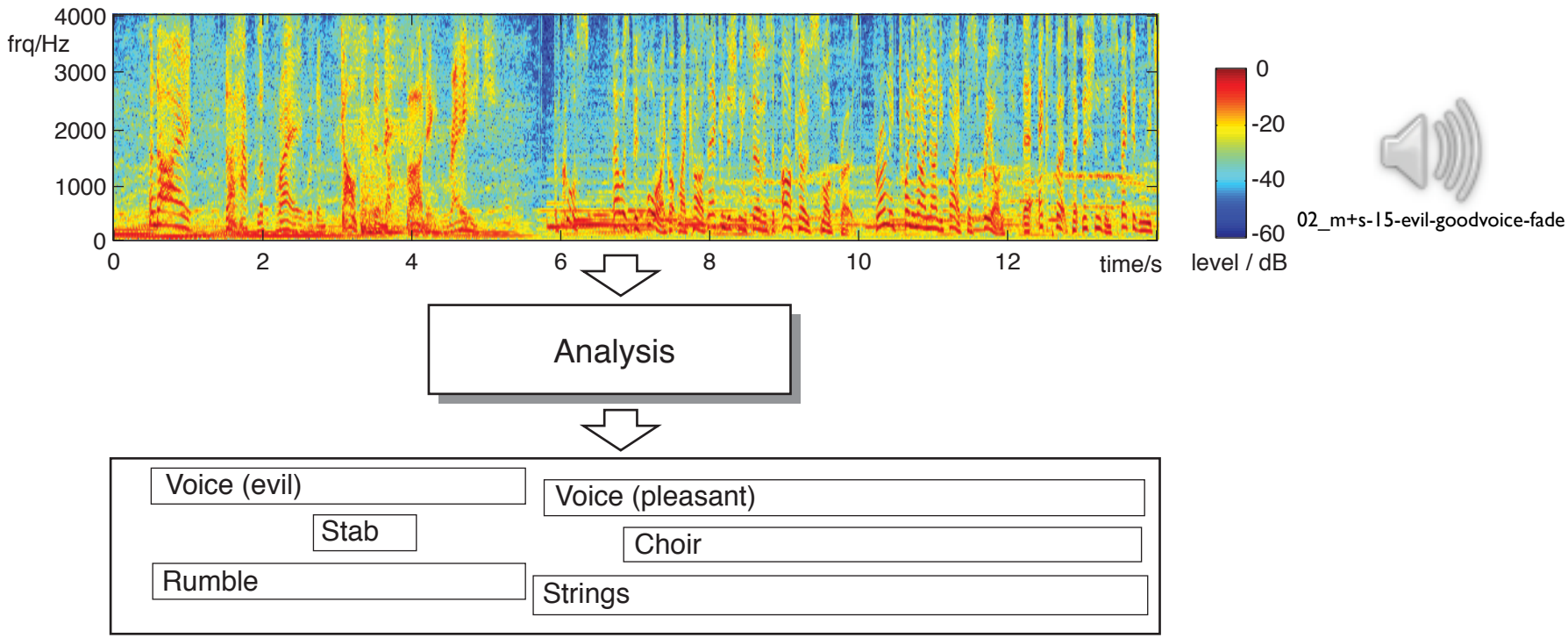
Learning, Using, and Adapting Models in Scene Analysis

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

{dpwe,ronw}@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Source Models and Scene Analysis
2. Using Source Models
3. Adapting Source Models
4. Source Model Issues

1. Source Models and Scene Analysis

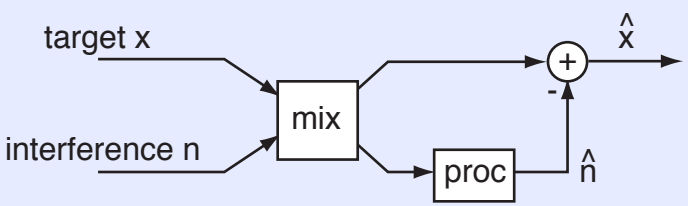


- Sounds rarely occur in isolation
  - .. so analyzing mixtures ("scenes") is a problem
  - .. for humans and machines
- How to solve this problem?

Approaches to Separation

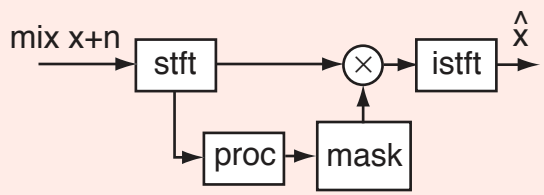
ICA

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



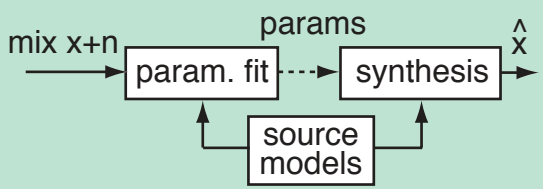
CASA

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



Model-based

- Any domain
- Param. search
- Synthetic output?



- or combinations ...

Separation vs. Inference

- Ideal separation is rarely possible
  - many situations where overlaps cannot be removed
- Overlaps → Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

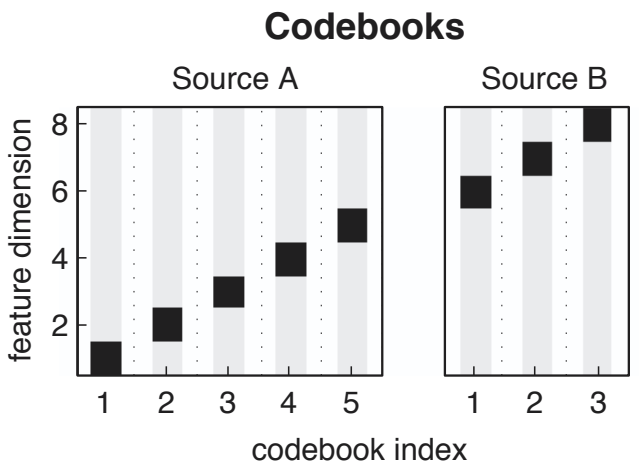
$$P(\{S_i\}|\, X) \propto P(X\,|\{S_i\})\ \Pi_i\, P(S_i|M_i)$$

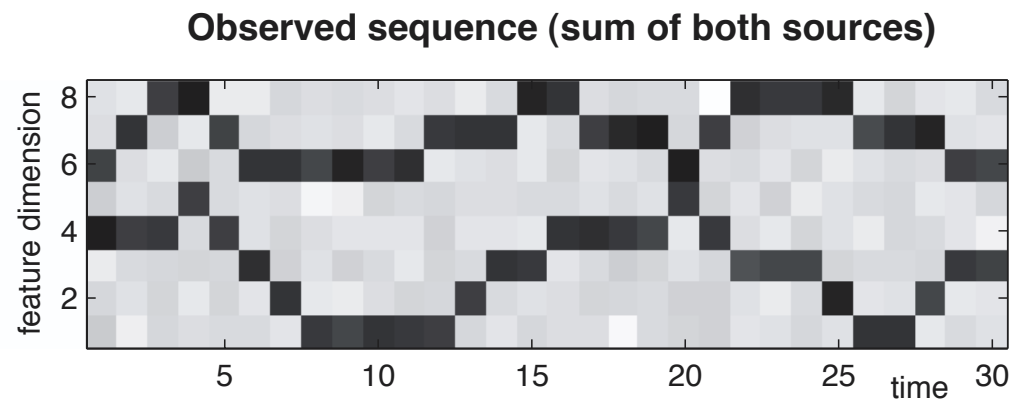
combination physics *source models*

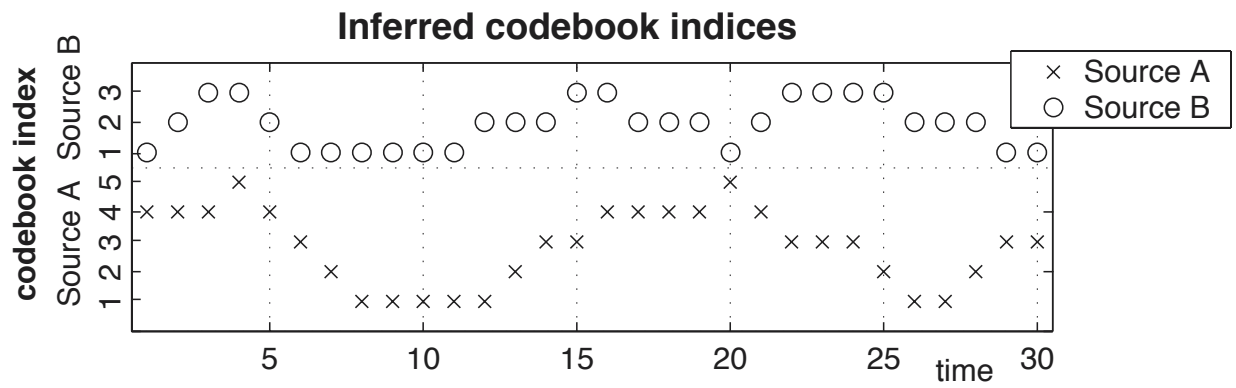
  - search over all source signal sets $\{S_i\}$??
- Better source models → better inference

A Simple Example

- Source models are codebooks from separate subspaces

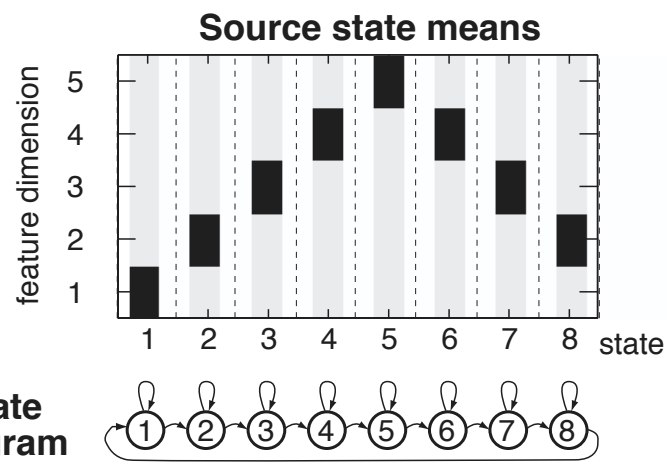


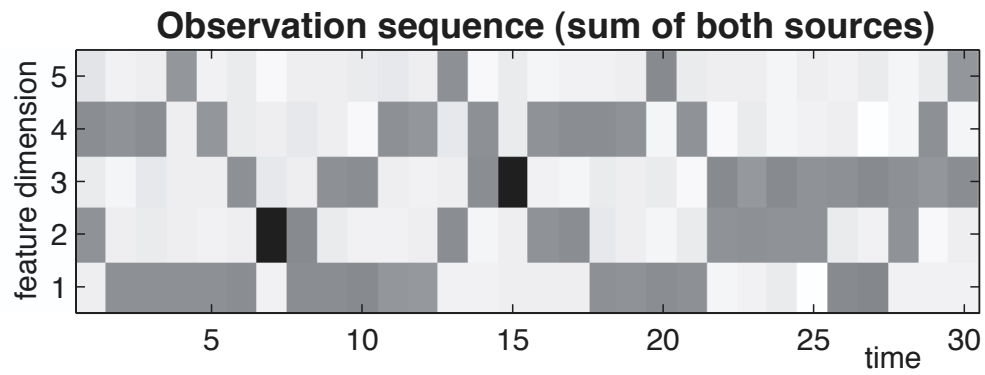


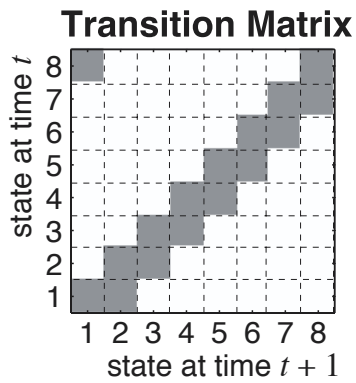


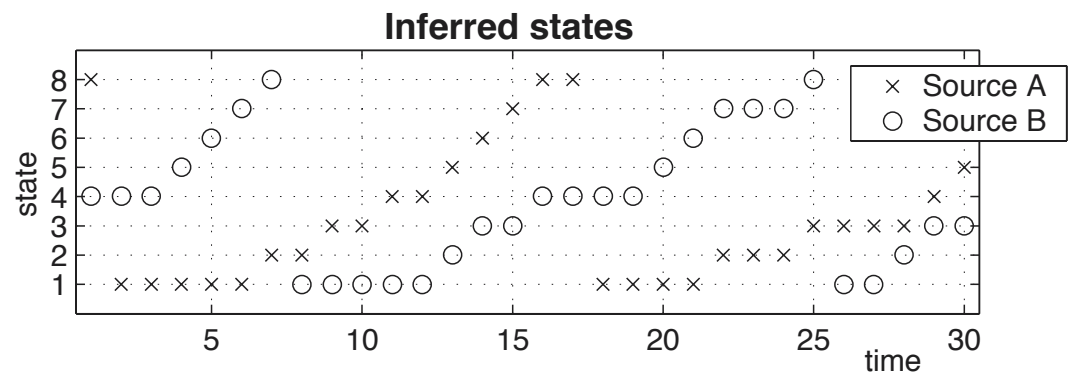
A Slightly Less Simple Example

- Sources with Markov transitions



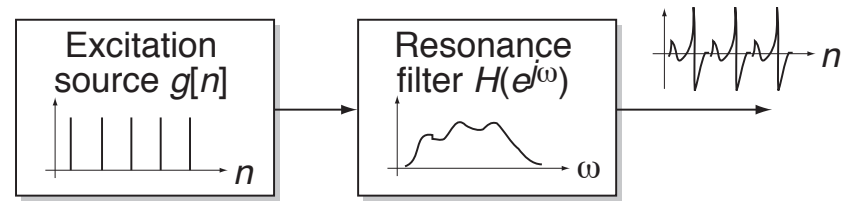






What is a Source Model?

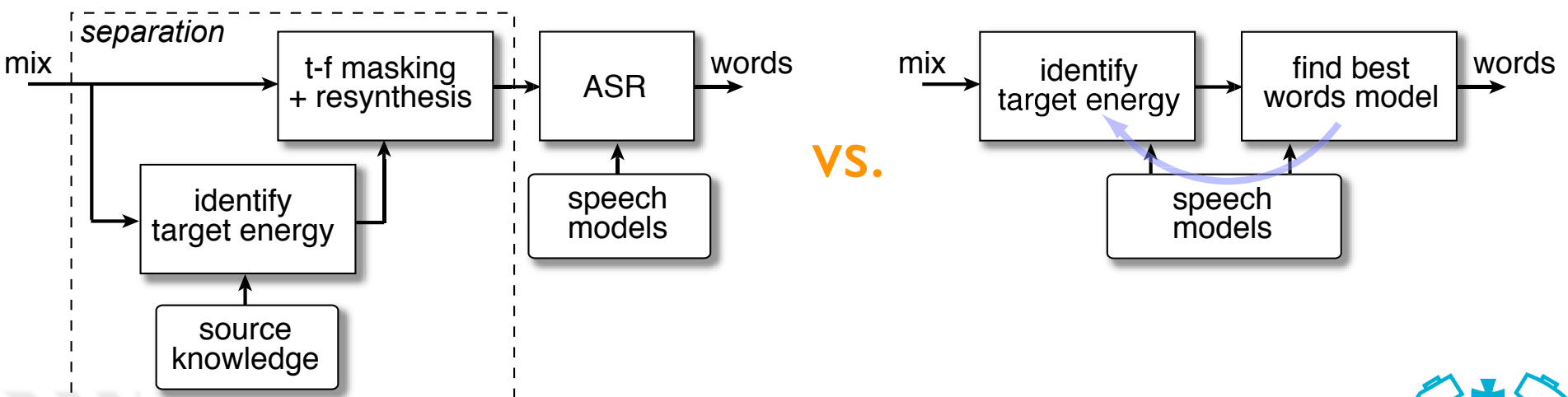
- Source Model describes signal behavior
  - encapsulates constraints on form of signal
  - (any such constraint can be seen as a model...)
- A model has parameters
  - model + parameters
    → instance



- 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

2. Using Models: Speech Separation

- Cooke & Lee's Speech Separation Challenge
  - pairs of 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"
  - (special session at Interspeech '06)
- Separation or Description?



Codebook Models

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

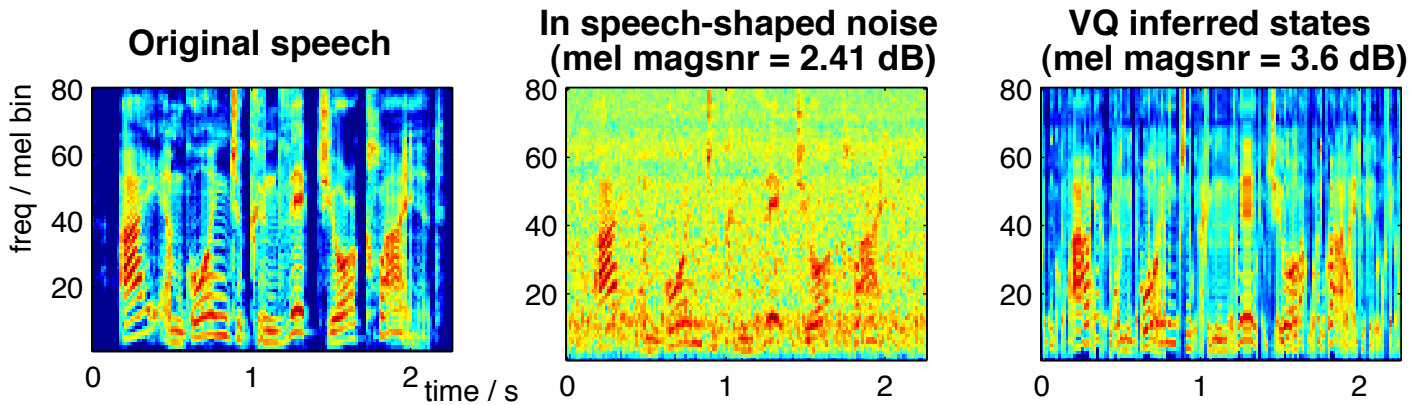
- Given models for sources, find "best" (most likely) states for spectra:

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

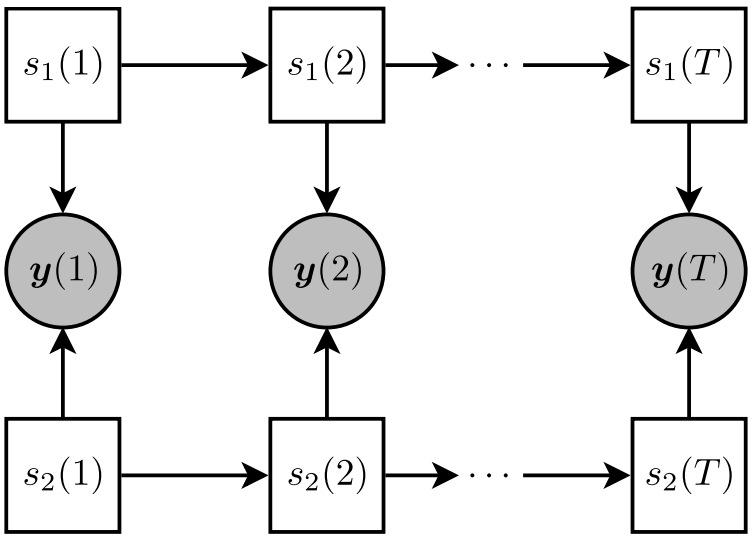
- E.g. stationary noise:

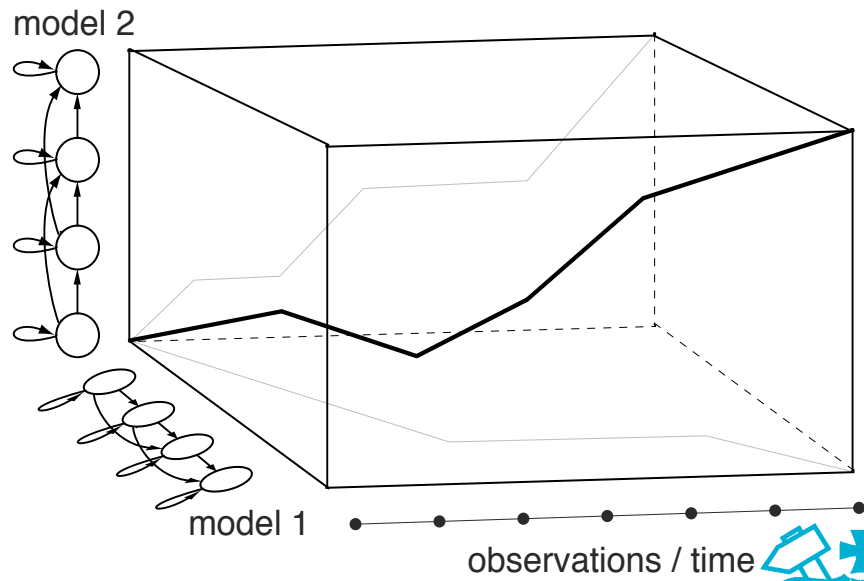


Speech Recognition Models

Varga & Moore '90

- 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

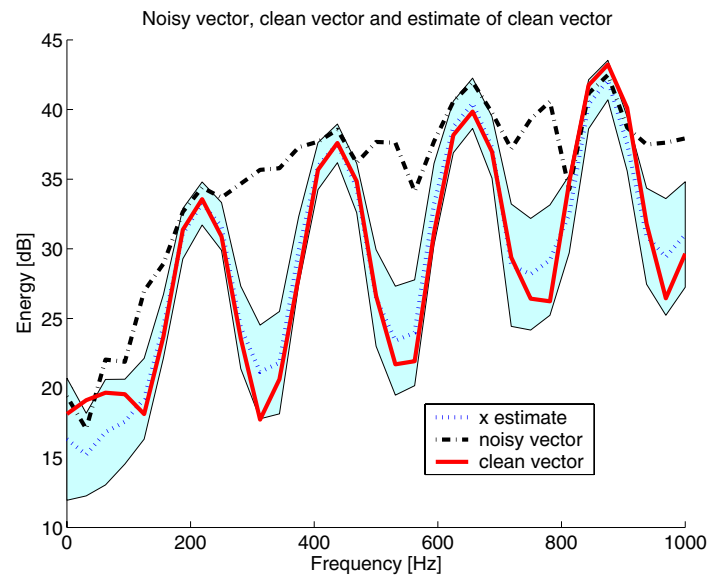


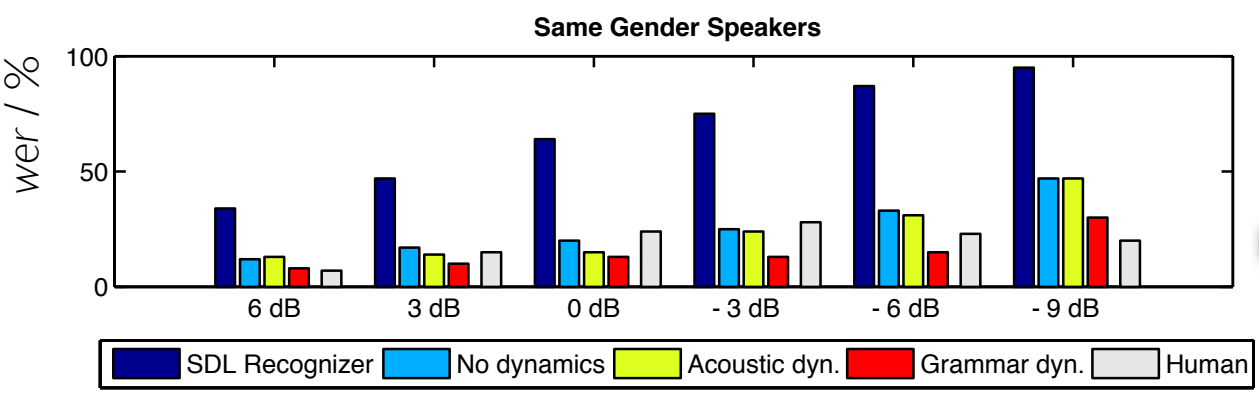


Speech Factorial Separation

Kristjansson, Hershey et al. '06

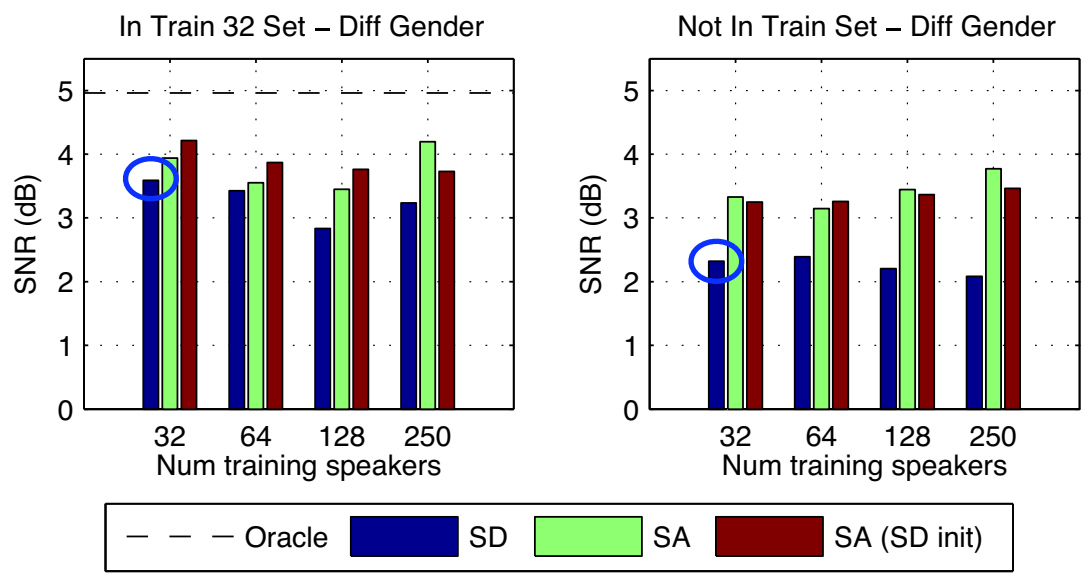
- IBM's 2006 Iroquois speech separation system
  Key features:
  - detailed state combinations
  - large speech recognizer
  - exploits grammar constraints
  - 34 per-speaker models
- "Superhuman" performance
  - ... in some conditions





3. Adapting Source Models

- Power of model-based separation depends on detail of model
- Speech separation relies on prior knowledge of every speaker?

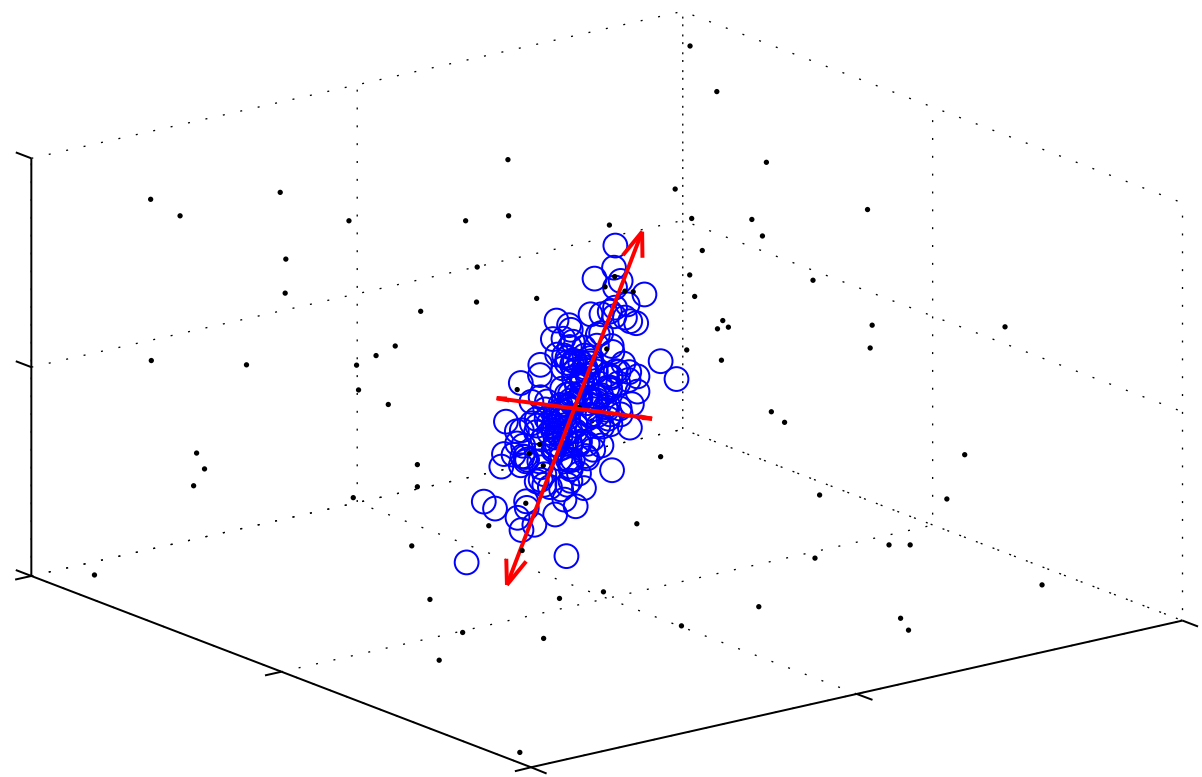


- Can this be practical?

Eigenvoices

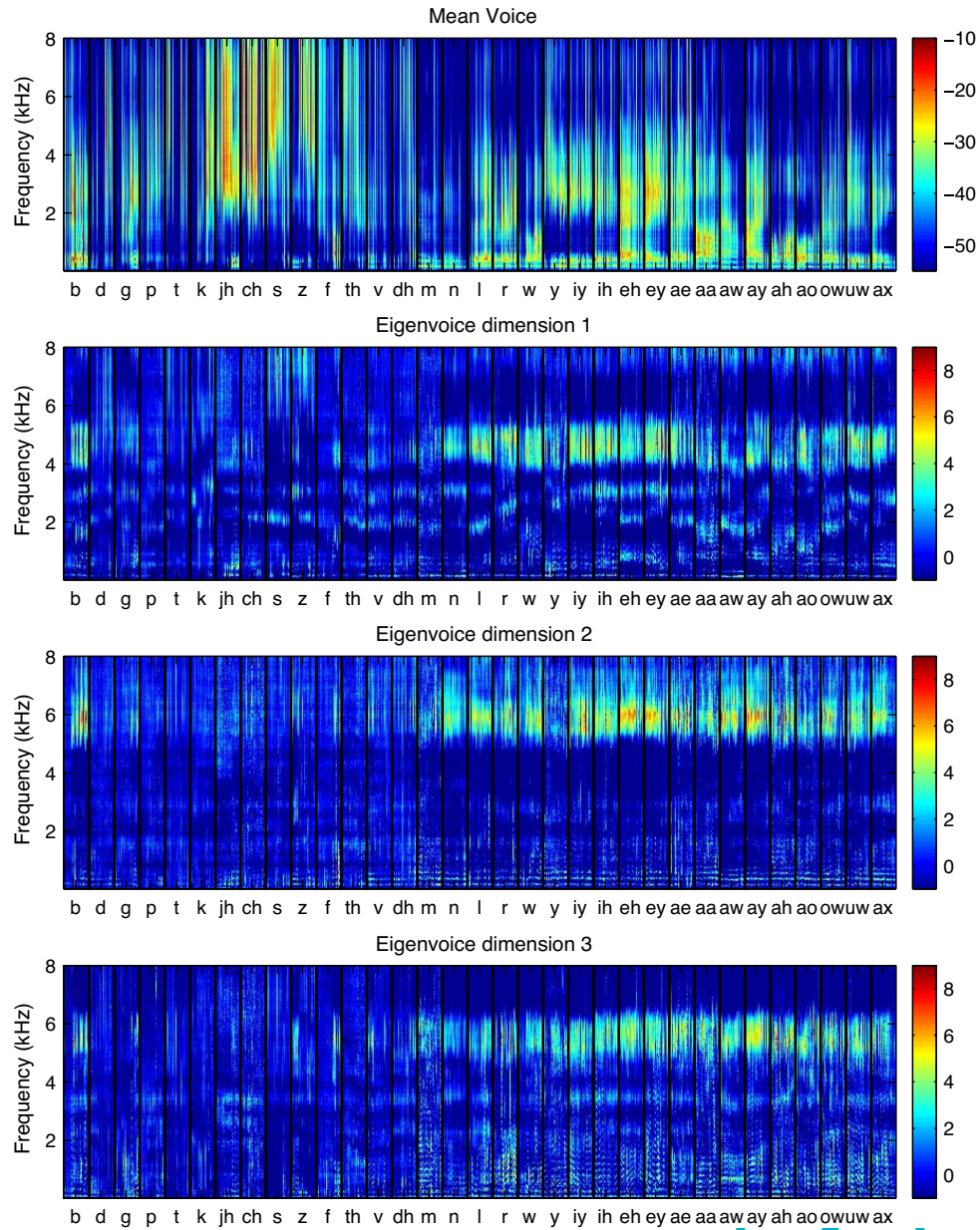
Kuhn et al. '98, '00
Weiss & Ellis '07, '08, '09

- Idea:
  Identify manifold in model parameter space
  - generalize without losing detail?



Eigenvoice Bases

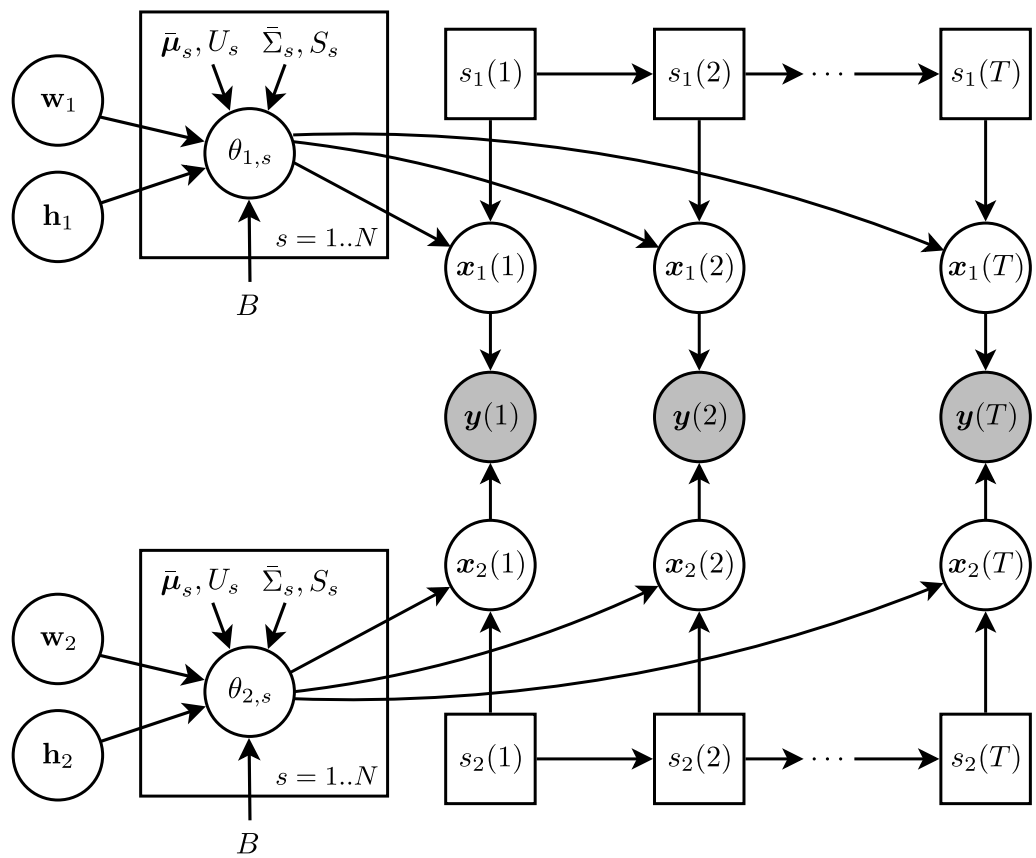
- **Mean model**
  - 280 states x 320 bins = 89,600 dimensions
- **Eigencomponents shift formants/ coloration**
  - additional components for channel



Speaker-Adapted Separation

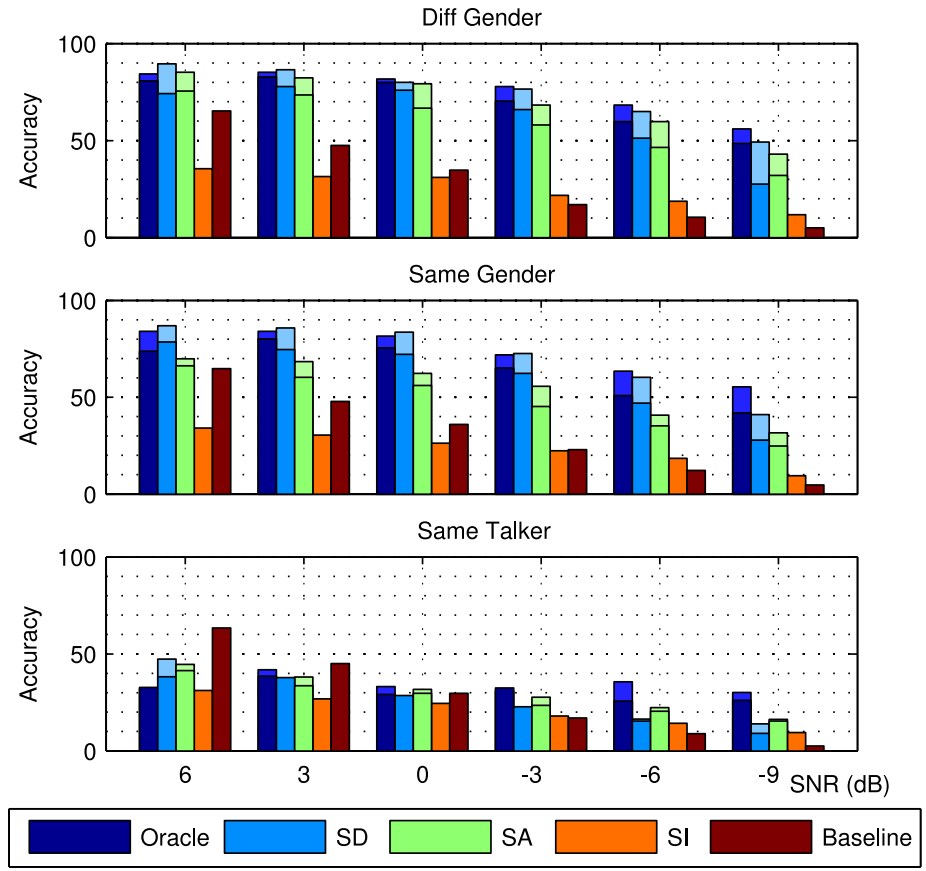
Weiss & Ellis '08

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



Speaker-Adapted Separation

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



Combining Spatial + Speech Model

Weiss, Mandel & Ellis '08

- Interaural parameters give

$$ILD_i(\omega),\ ITD_i,\ \Pr\left(X(t,\omega) = S_i(t,\omega)\right)$$

- Speech source model can give

$$\Pr\left(S_i(t,\omega) \text{ is speech signal}\right)$$

- Can combine into one big EM framework...

E-step

$$p(u|\Theta^{(n)}) = p(x,u|\Theta^{(n)})/p(x|\Theta^{(n)})$$

M-step

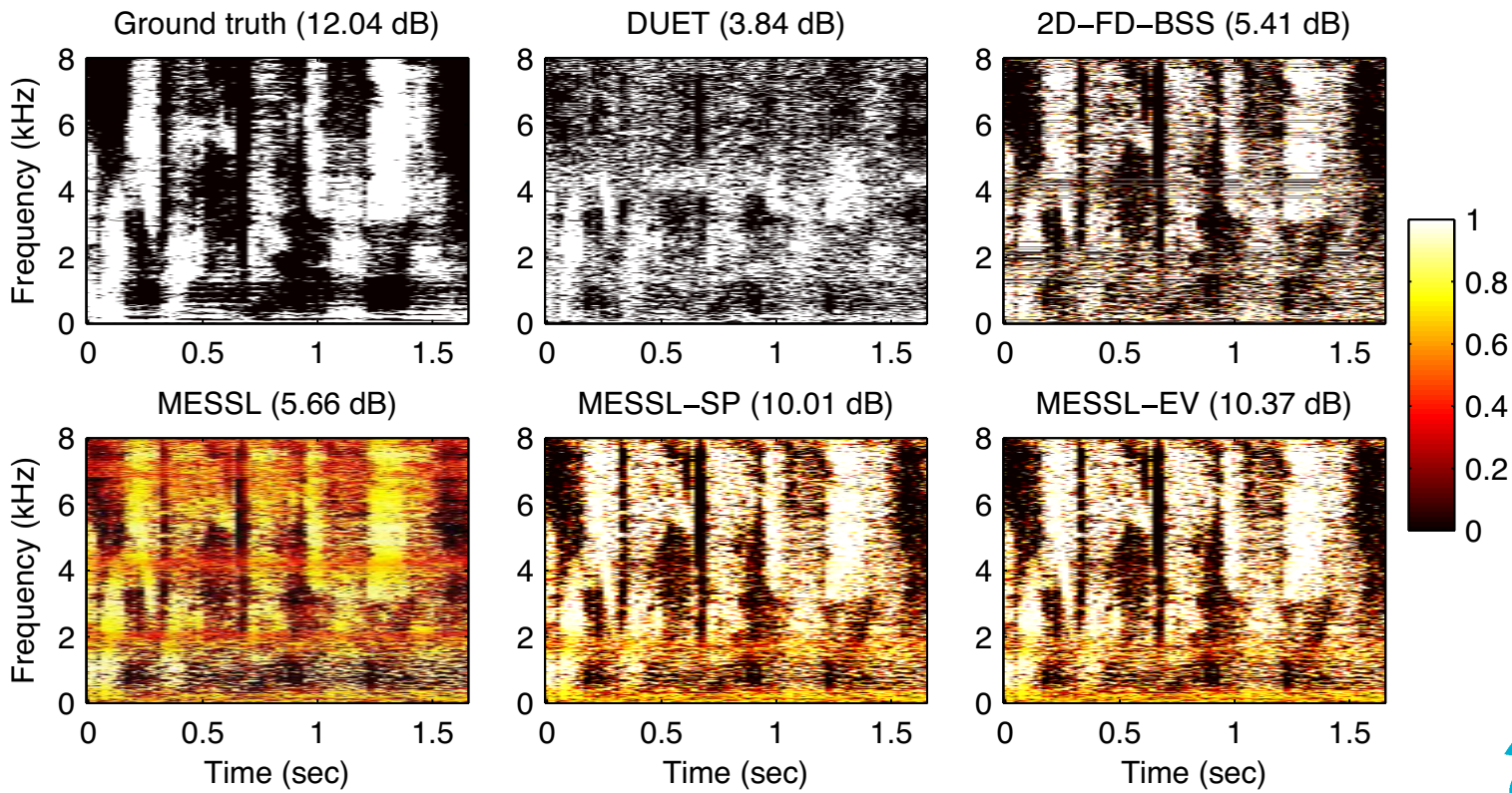
$$\Theta^{(n+1)} = \underset{\Theta}{\operatorname{argmax}}\ E_{p(u|\Theta^{(n)})}p(x,u|\Theta)$$

u is: Pr(cell from source *i*) phoneme sequence

Θ is: interaural params speaker params

Combining Spatial + Speech Model

- Source models function as priors
- Interaural parameter spatial separation
  - EM estimation of TF masks, spatial origin
  - source model prior improves spatial estimate



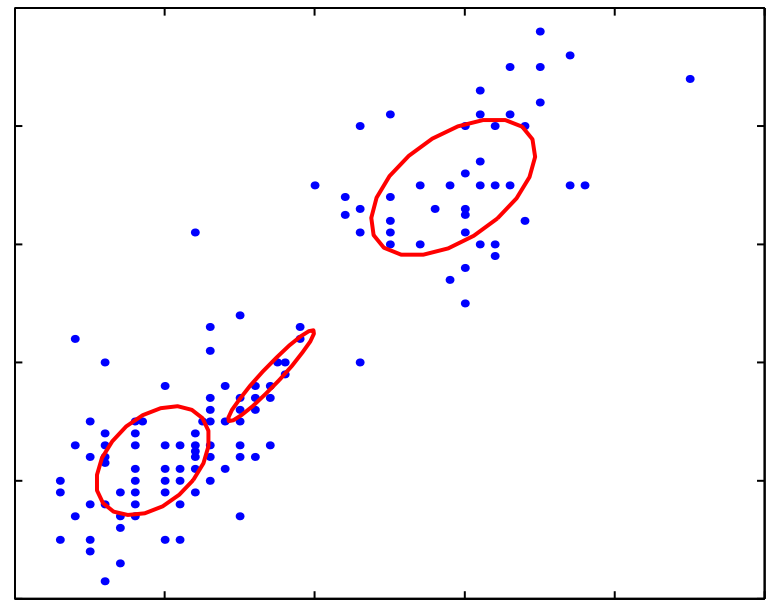
4. Source Model Issues

- Model 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 *

Learning Source Models

- **Speech models learned from labeled data**
  - single, known speaker + transcripts
  - data fully aligned to models
- **Otherwise ...**
  - wait for "clear shot"?
  - reinforce based on best-guess separation?
  - ML model updates?

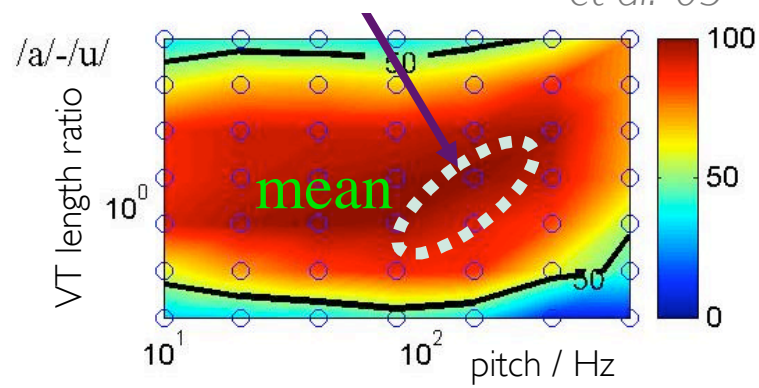
    [Ozerov et al. 2005]



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



- Time scales of model acquisition
  - innate/evolutionary (hair-cell tuning)
  - developmental (mother tongue phones)
  - dynamic - the "Bolero" effect

Summary & Conclusions

- Source models provide the constraints to make scene analysis possible
- Eigenvoices (model subspace) can be used to provide detailed models that generalize
- It is not clear how to extend this to all possible sounds, present and future
- Relevance to perception?