Lab ROSA

Laboratory for the Recognition and Organization of Speech and Audio

Using Speech Models for Separation

Dan Ellis

Comprising the work of Michael Mandel and Ron Weiss

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/

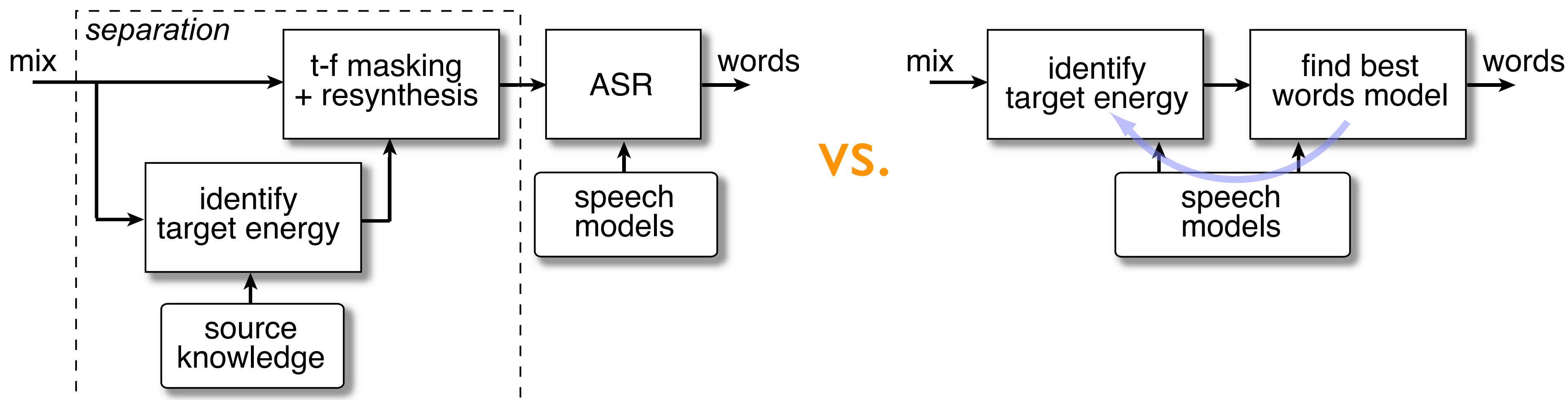
Columbia University in the City of New York

1. Eigenvoice Speaker Models
2. Spatial Parameter Models in Reverb
3. Combining Source + Spatial

1. Speech Separation Using Models

- Cooke & Lee's Monaural Speech Separation Task
  - 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"

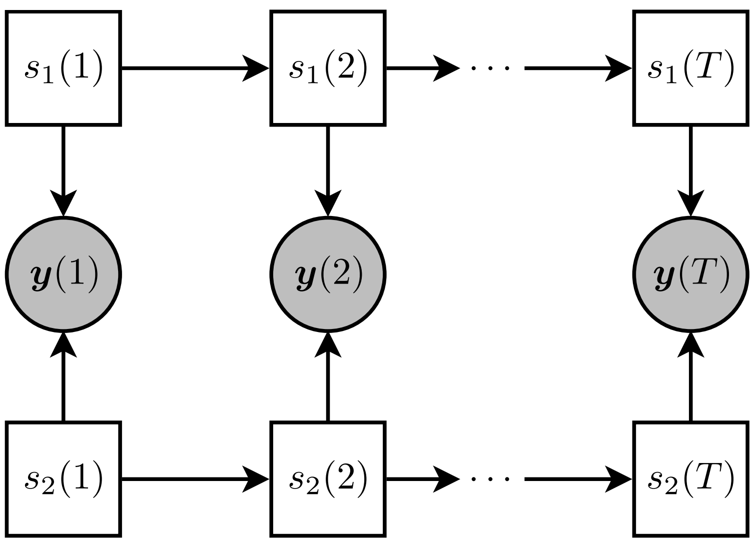
- Separation depends on source constraints
  - the more the better - ASR model

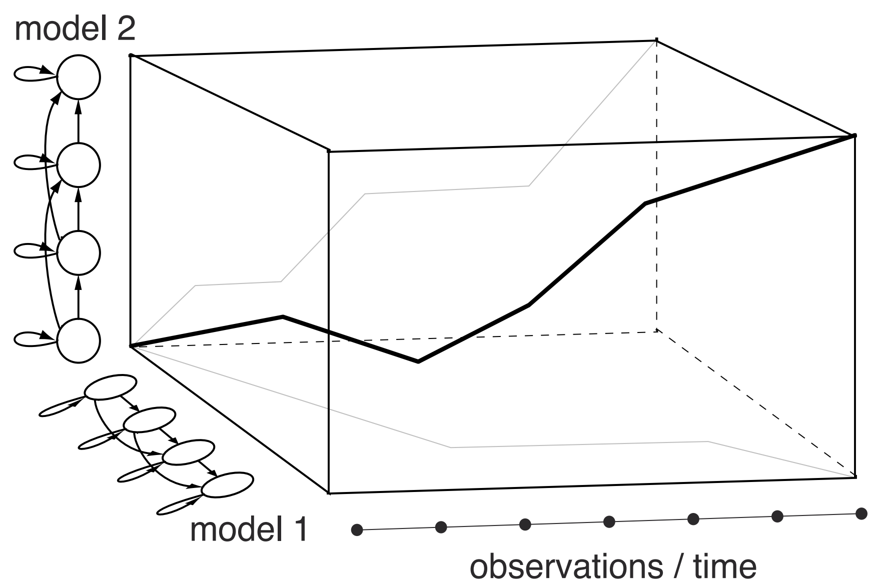


Speech Mixture Recognition

Kristjansson, Hershey et al. '06

- Speech recognizers contain speech models
  - ASR is just argmax $P(W \mid X)$
- Recognize mixtures with Factorial HMM
  - one model+state sequence for each voice
  - exploit sequence constraints, speaker differences



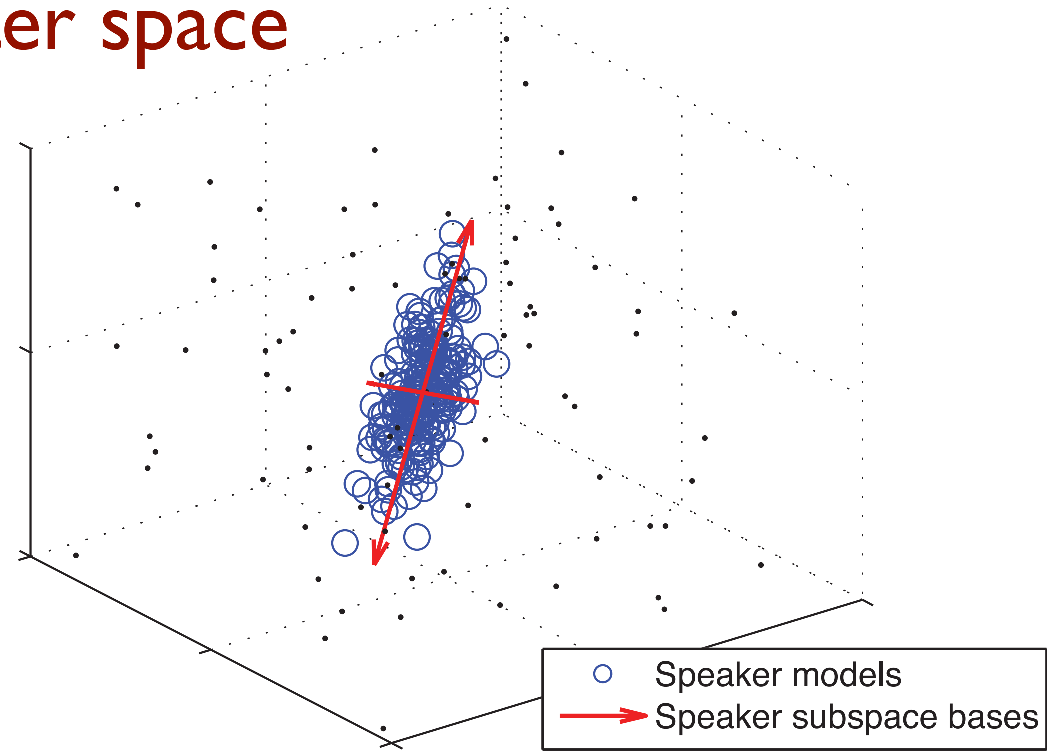


  - separation relies on detailed speaker model

Eigenvoices

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

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



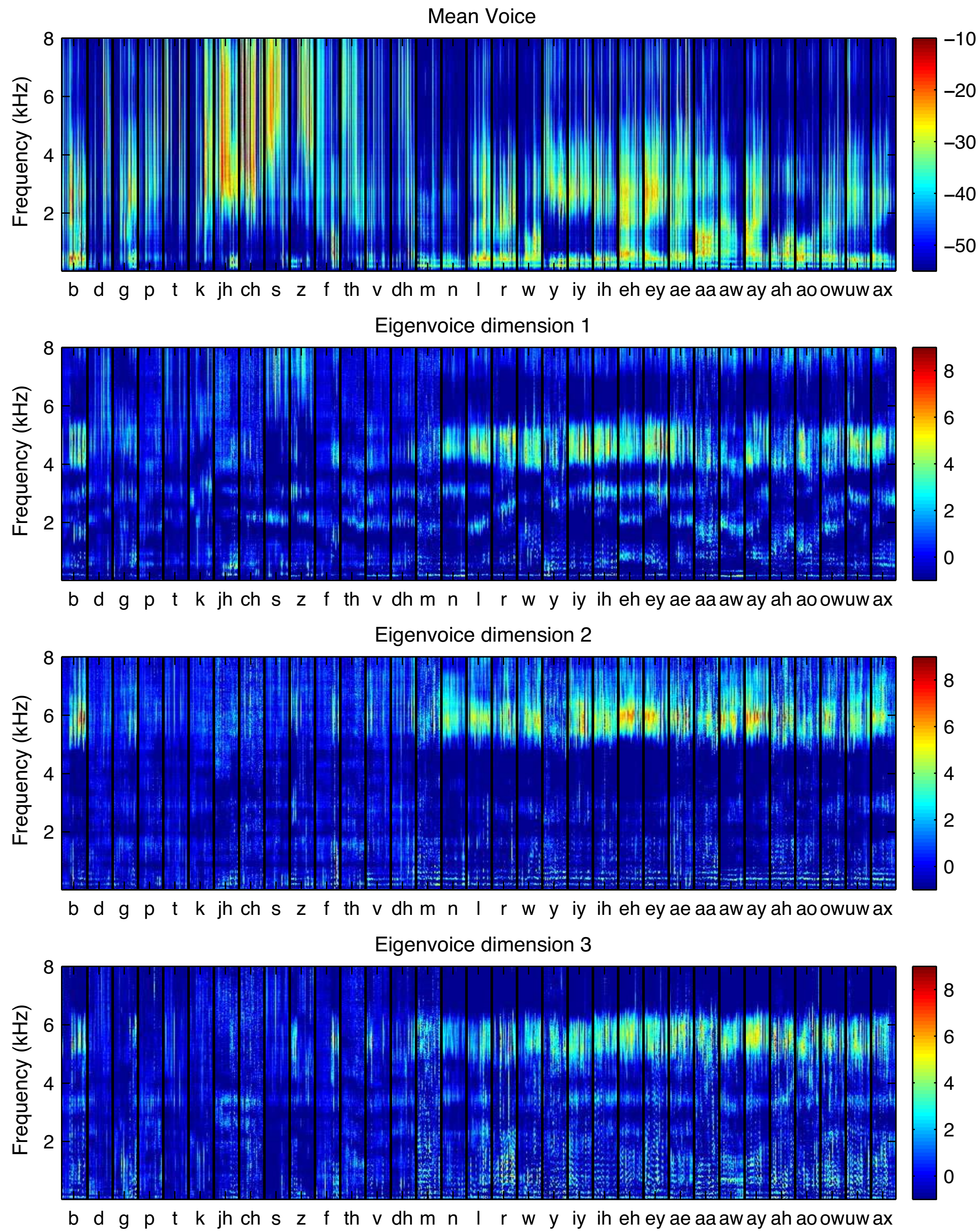
- Eigenvoice model:

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

adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

Eigenvoice Bases

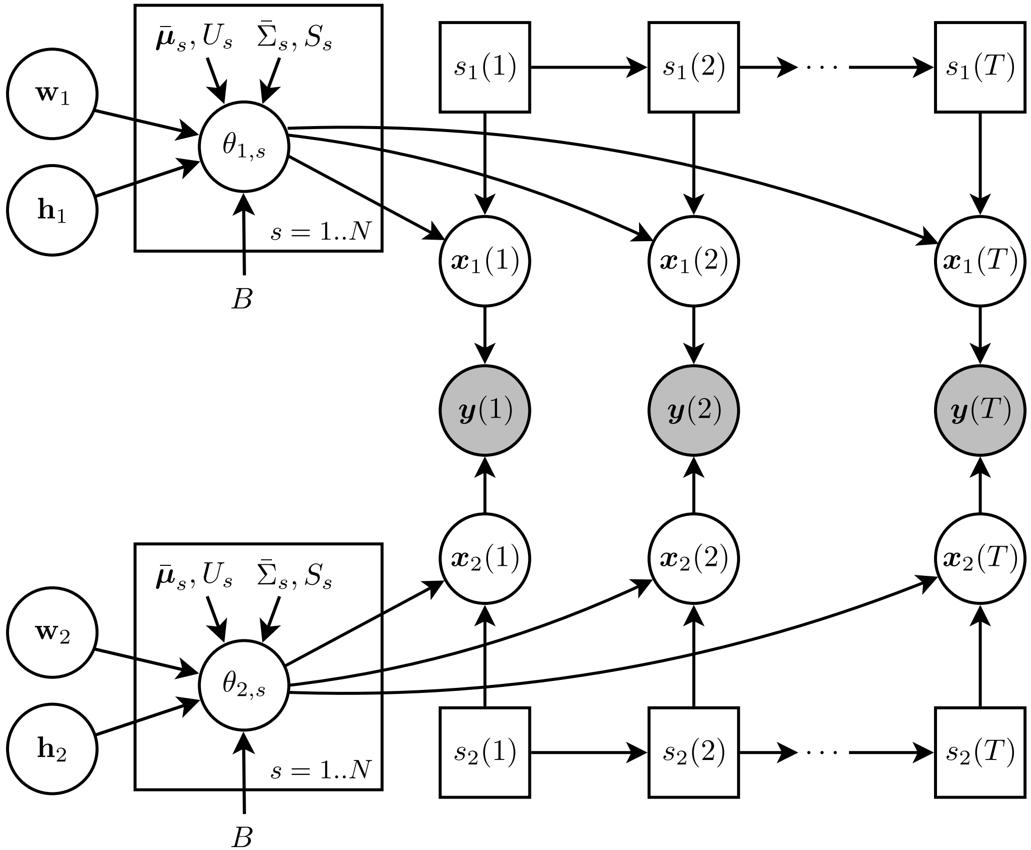
- **Mean model**
  - 280 states x 320 bins = 89,600 dimensions
- **Eigencomponents shift formants/ coloration**
  - additional components for acoustic channel



Speaker-Adapted Separation

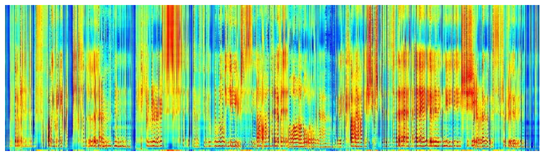
Weiss & Ellis '08

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

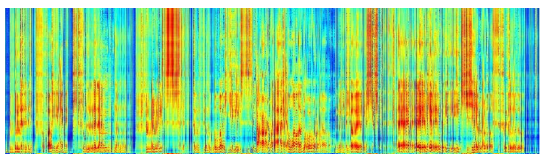


Speaker-Adapted Separation

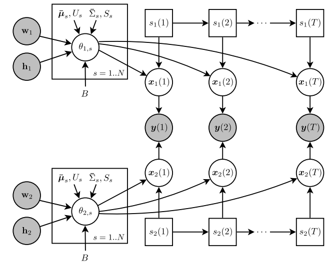
$$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$$

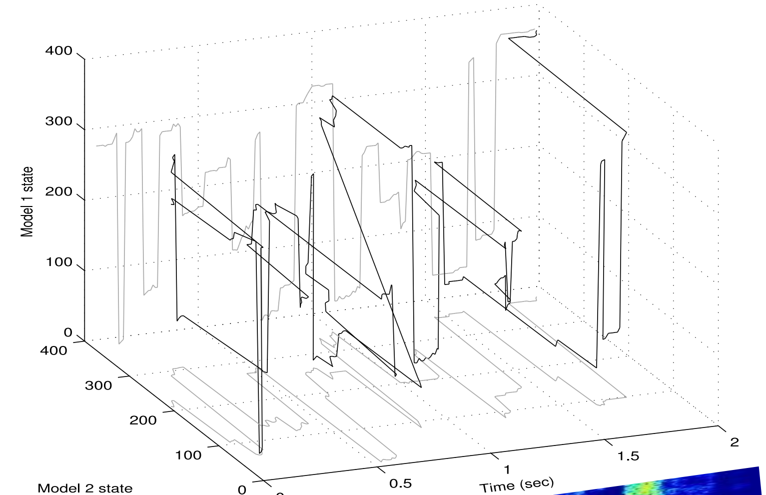


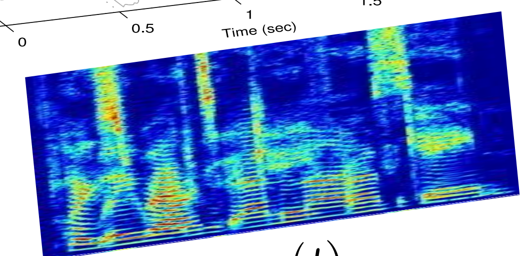
$$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$$



Find Viterbi path

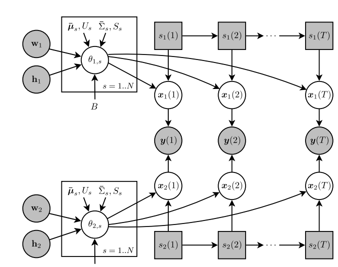




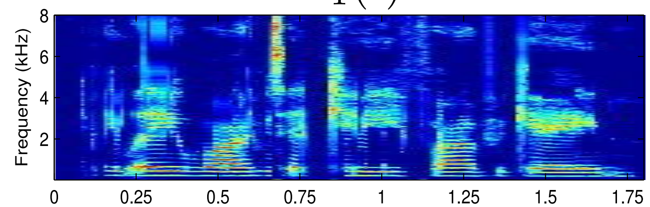


$\boldsymbol{y}(t)$

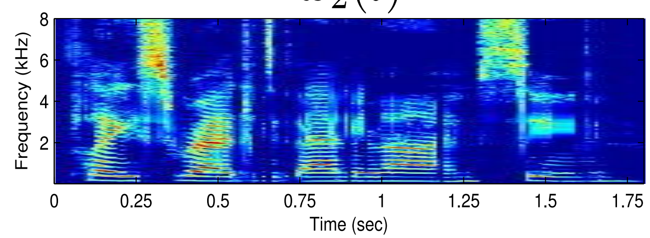
Estimate source signals



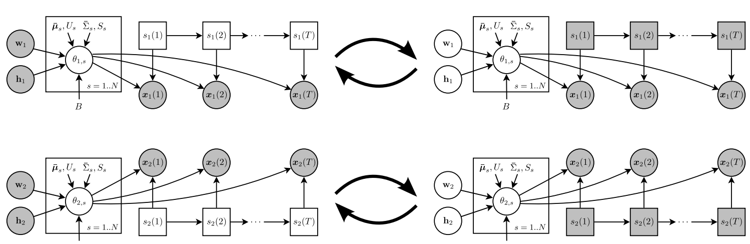
$\hat{\boldsymbol{x}}_1(t)$



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

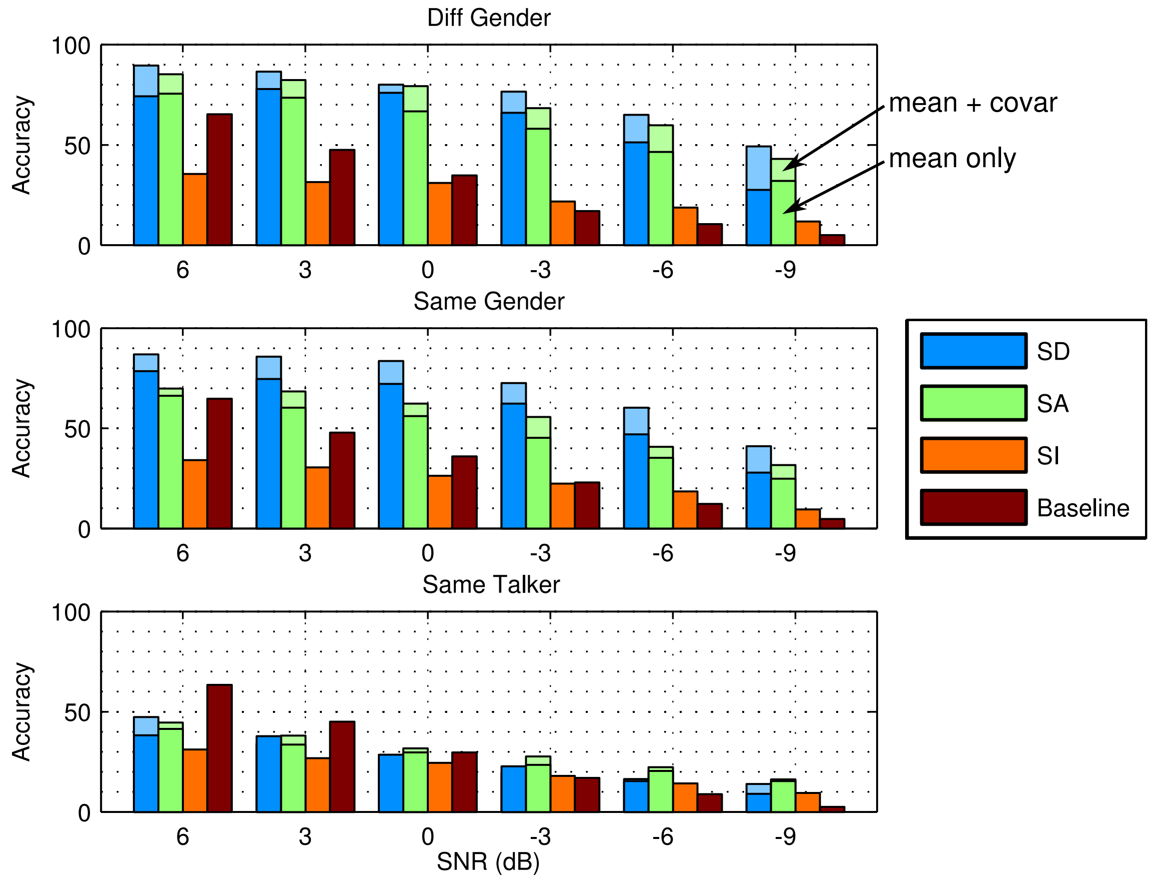


Update model parameters using EM algorithm from Kuhn et al., (2000)



Speaker-Adapted Separation

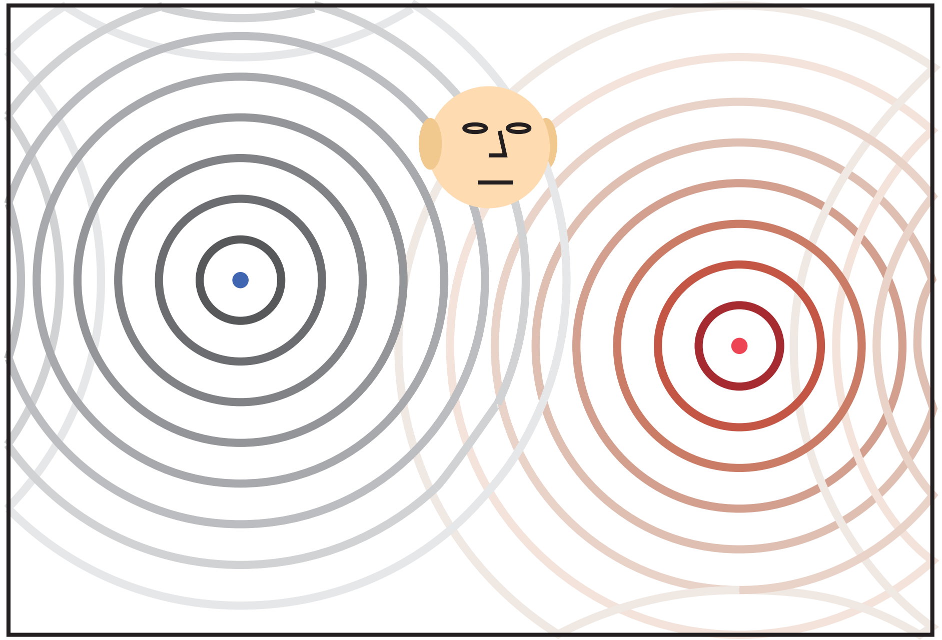
- Eigenvoices for Speech Separation task
  - speaker adapted (SA) performs midway between speaker-dependent (SD) & speaker-indep (SI)



2. Spatial Models & Reverb

Mandel & Ellis '07

- 2 or 3 sources in reverberation
  - assume just 2 'ears'

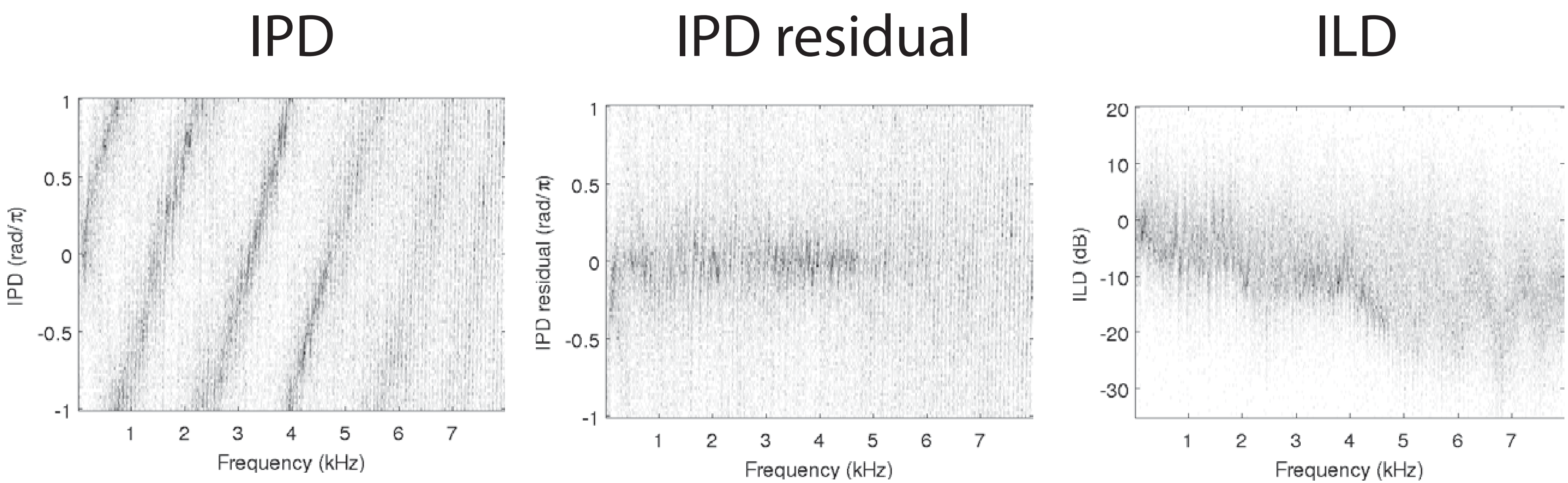


- Model interaural spectrum of each source as stationary level and time differences:

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

IPD, ILD Distributions

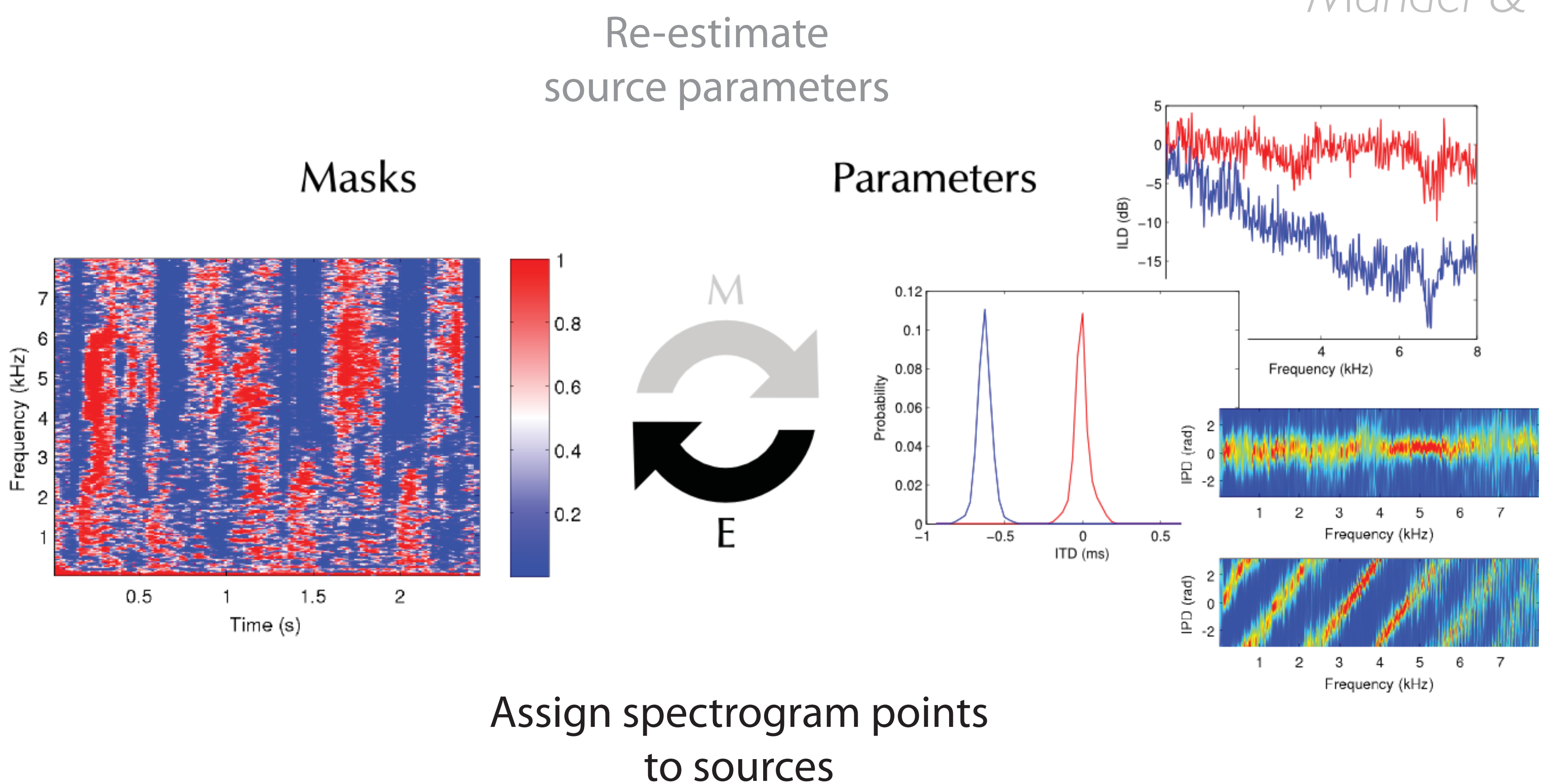
- Source at 75° in reverberation



  - IPD residual offsets phase by constant **ωτ**
  - IPD can be fit by single Gaussian
  - ILD needs frequency-dependence

Model-Based EM Source Separation and Localization (MESSL)

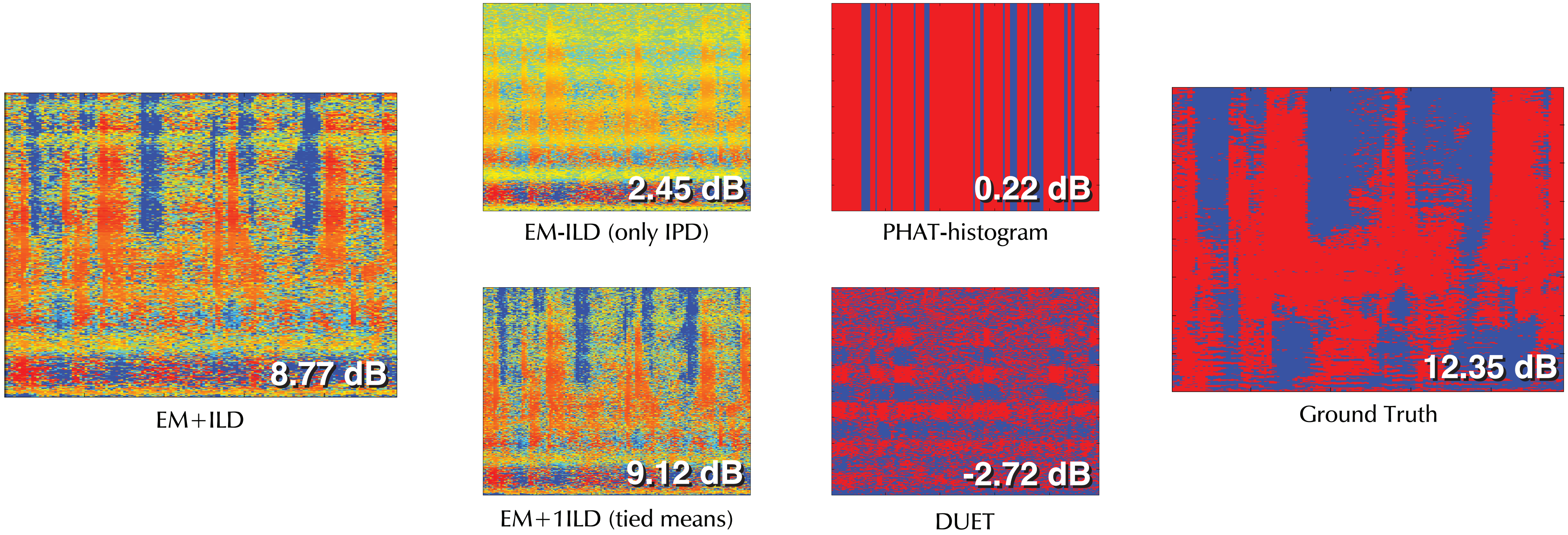
Mandel & Ellis '09



- can model more sources than sensors
- flexible initialization

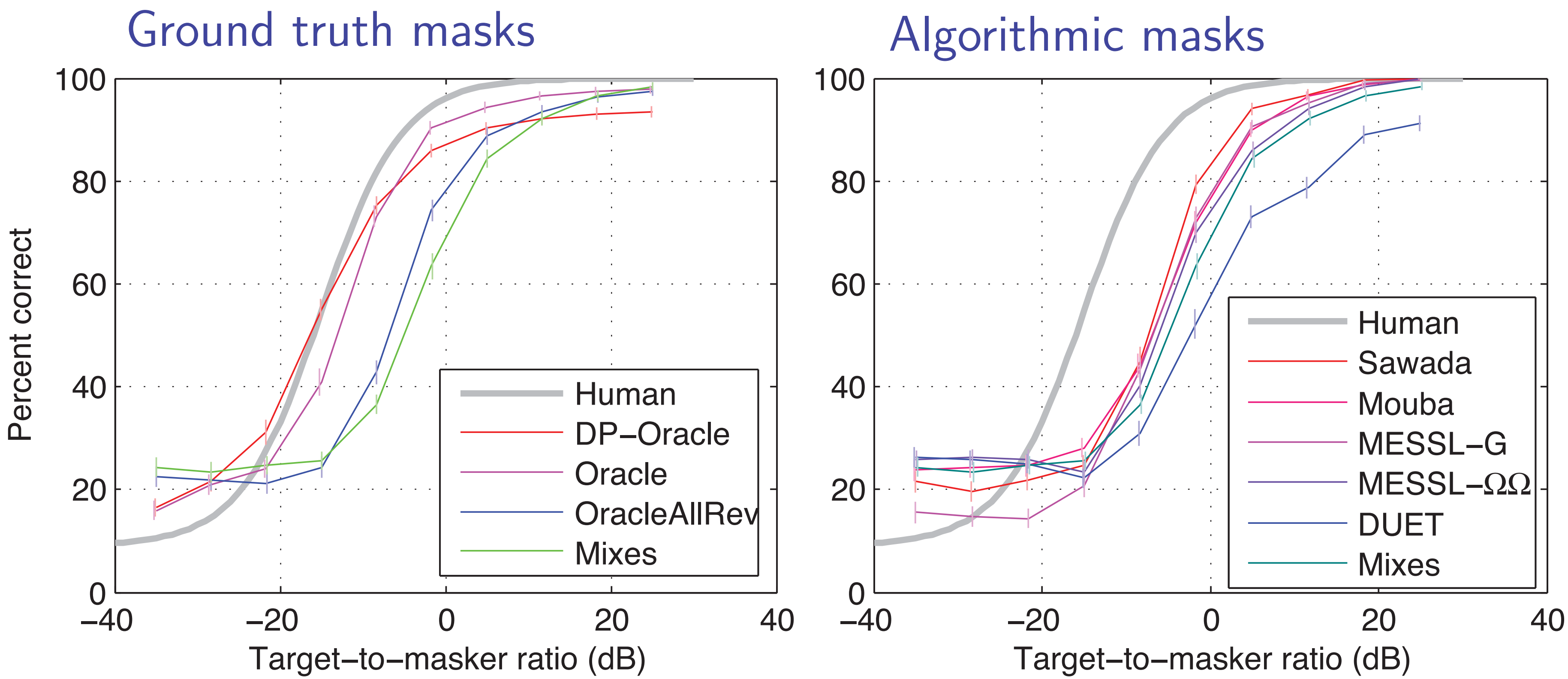
MESSL Results

- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness



MESSL Results

- Speech recognizer (Digits)



3. Combining Spatial + Speech Models

Weiss, Mandel & Ellis '08

- Interaural parameters give

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

- Speech source model can give

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

- Can combine into one big EM framework...

E-step

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

M-step

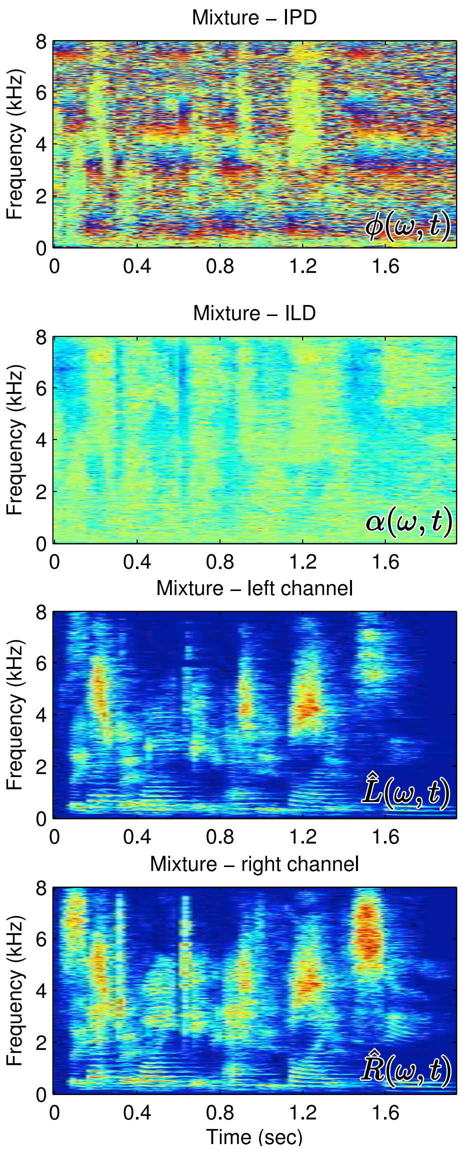
$$\Theta^{(n+1)} = \operatorname*{argmax}_{\Theta}\ E_{p(u|\Theta^{(n)})}p(x,u|\Theta)$$

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

Θ is: interaural params
speaker params

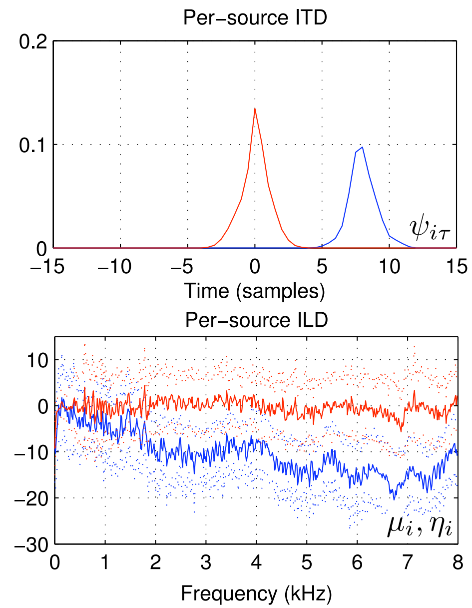
MESSL-SP (Source Prior)

Observations

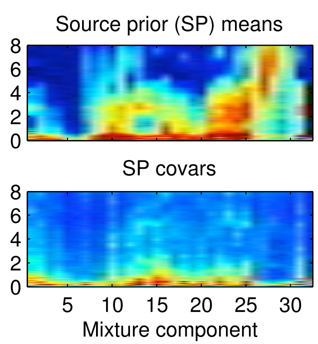


~

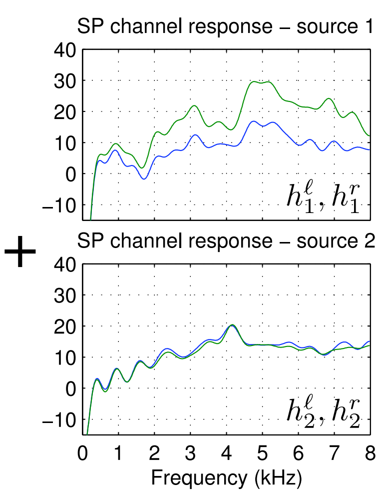
Parameters



~



+



E-step

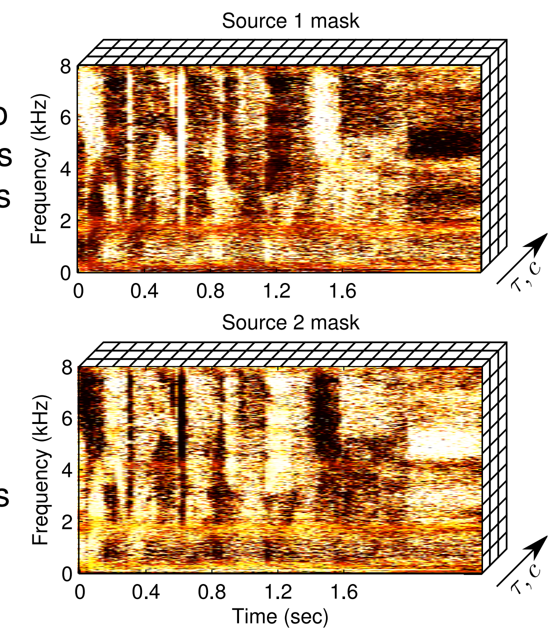
Use parameters to compute posteriors of hidden variables

M-step

Use posteriors to update parameters

Posteriors

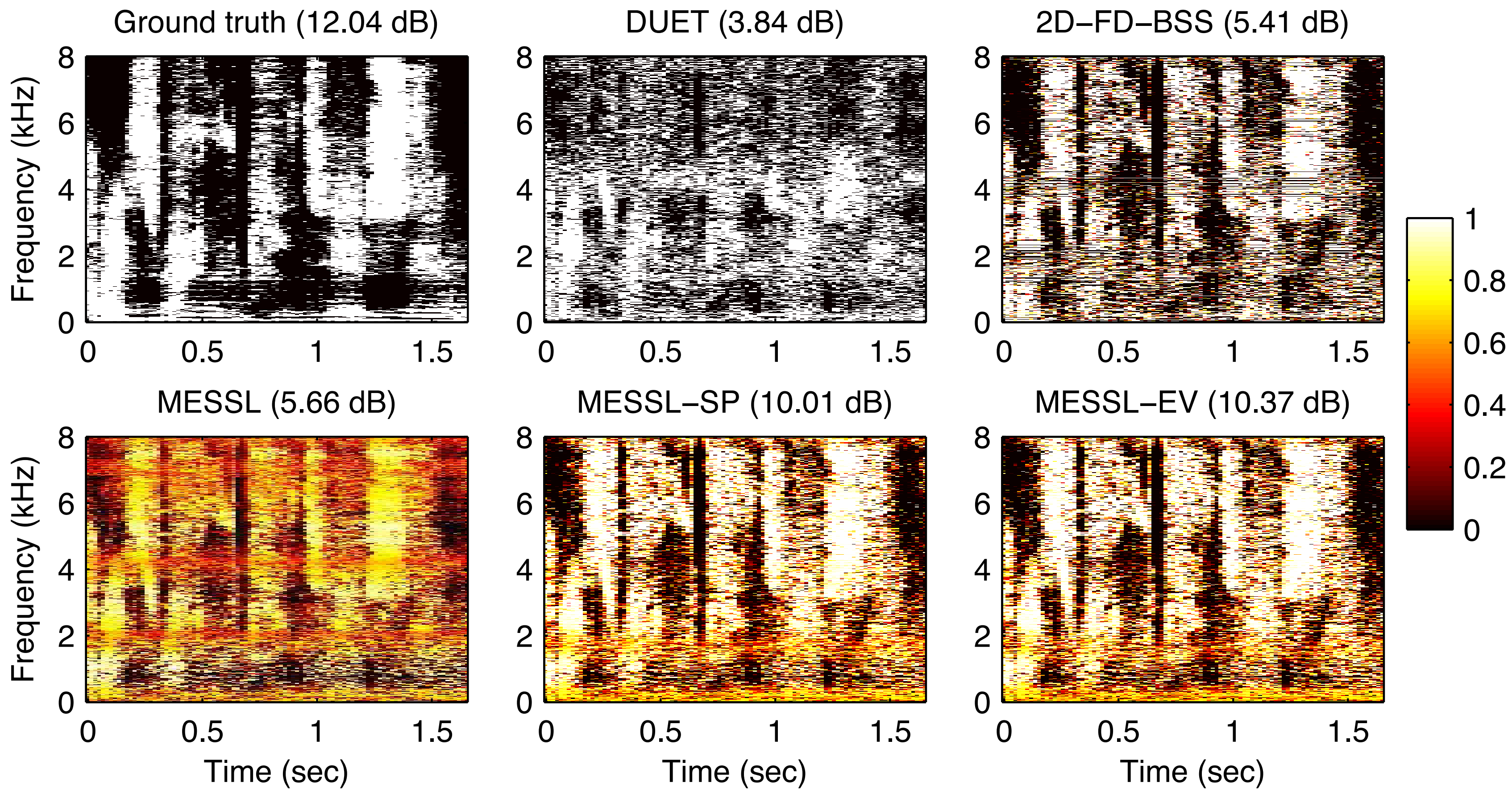
Each point in spectrogram is explained by a source, delay, and mixture component



Separate sources by multiplying mixture by different masks

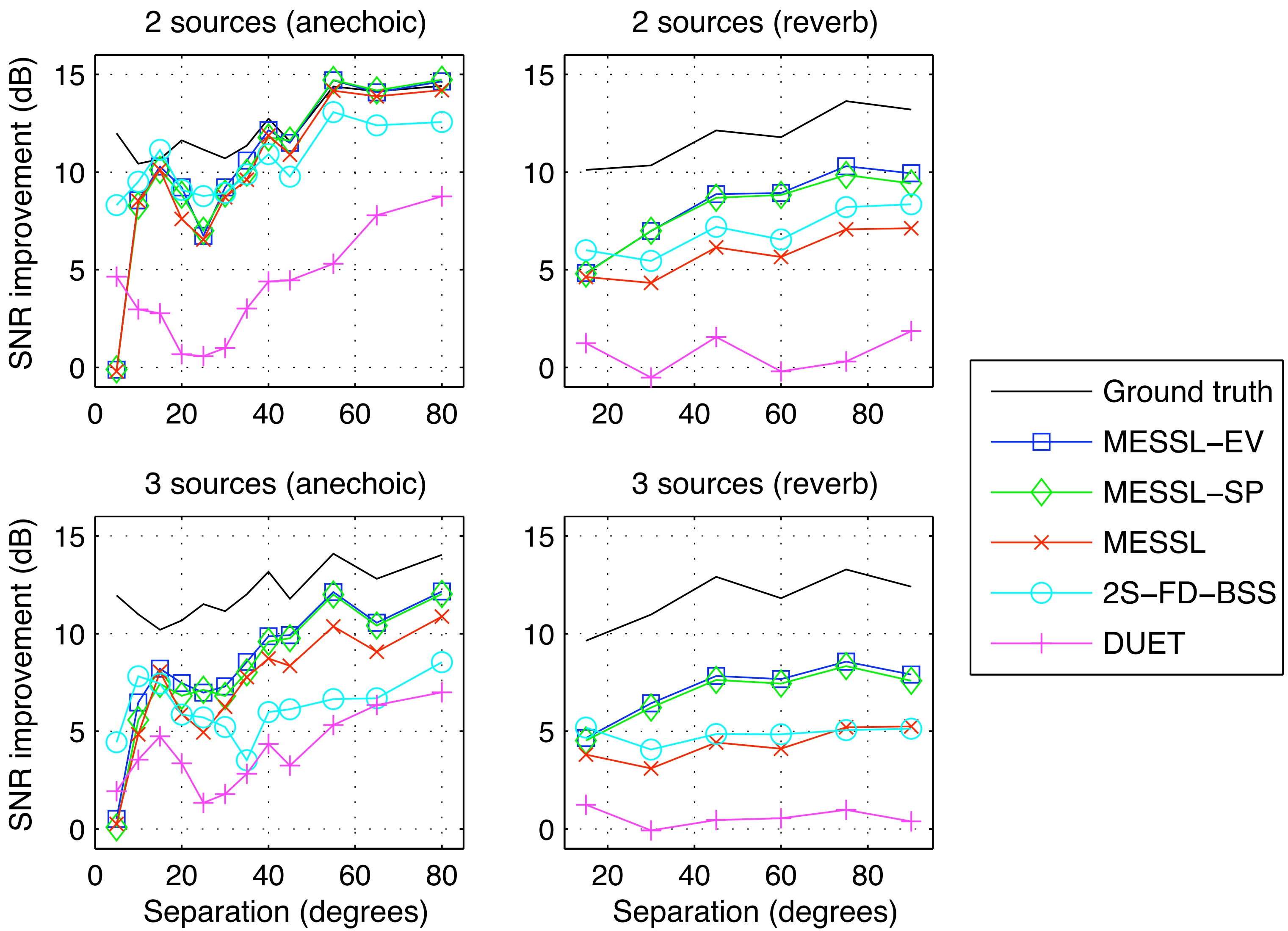
MESSL-SP Results

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



MESSL-SP Results

- SNR improvement vs. source angle separation



Future Work

- Better parametric speaker models
  - limitations of eigenvoices
  - varying style
- Understanding reverb & ASR
  - early echoes
  - what spoils ASR?
- Models of other sources
  - eigeninstruments?

Summary & Conclusions

- Source models provide the constraints to make scene analysis possible
- Eigenvoices (model subspace) can be used to provide detailed models that generalize
- Spatial parameters can identify more sources than models in reverb (MESSL)
- Can combine source + spatial models