A VARIATIONAL EM ALGORITHM FOR LEARNING EIGENVOICE PARAMETERS IN MIXED SIGNALS

Ron J. Weiss and Daniel P. W. Ellis

LabROSA · Dept of Electrical Engineering · Columbia University, New York, USA

{ronw,dpwe}@ee.columbia.edu

1. Summary

- Model-based monaural speech separation where the precise source characteristics are not known a priori
- Extend original adaptation algorithm from Weiss and Ellis (2008) to adapt Gaussian covariances as well as means
- Derive a variational EM algorithm to speed up adaptation

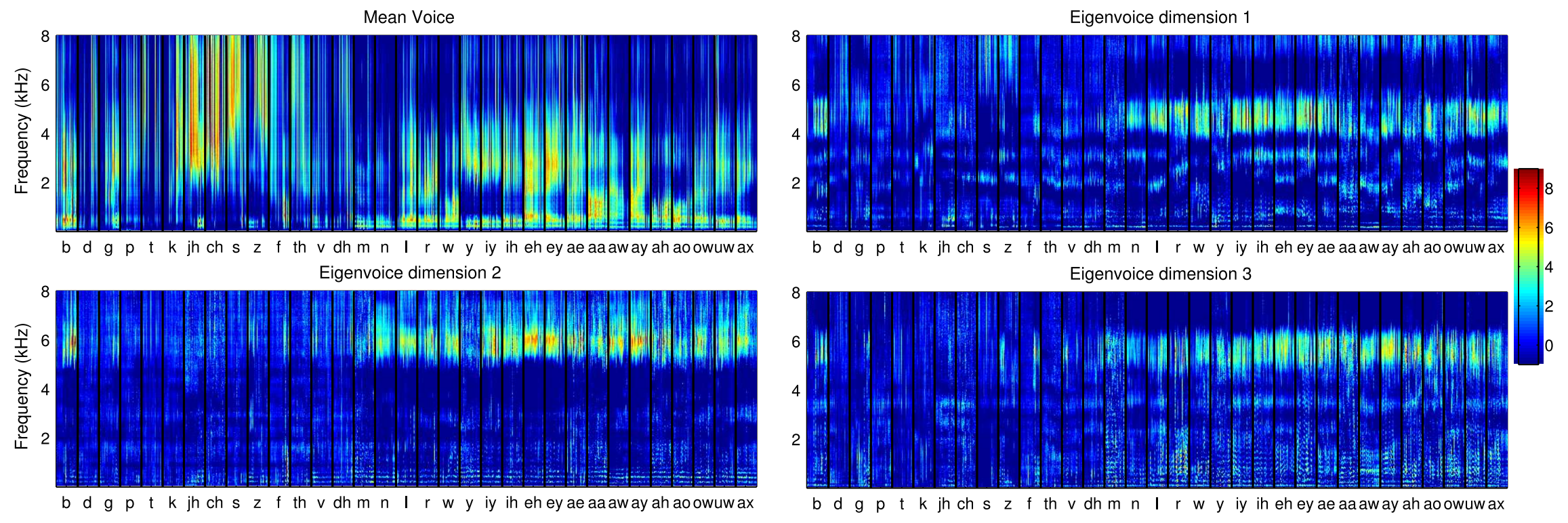
2. Mixed signal model

- Model log power spectra of source signals using hidden Markov model (HMM):

$$P\big(\boldsymbol{x}_i(1..T),\, s_i(1..T)\big) = \prod_t P\big(s_i(t) \,|\, s_i(t-1)\big)\, P\big(\boldsymbol{x}_i(t) \,|\, s_i(t)\big)$$

- Represent speaker-dependent model as linear combination of eigenvoice bases (Kuhn et al., 2000):

$$P\big(\boldsymbol{x}_i(t) \,|\, s\big) = \mathcal{N}\big(\boldsymbol{x}_i(t);\, \bar{\boldsymbol{\mu}}_s + U_s \mathbf{w}_i,\, \bar{\Sigma}_s\big)$$

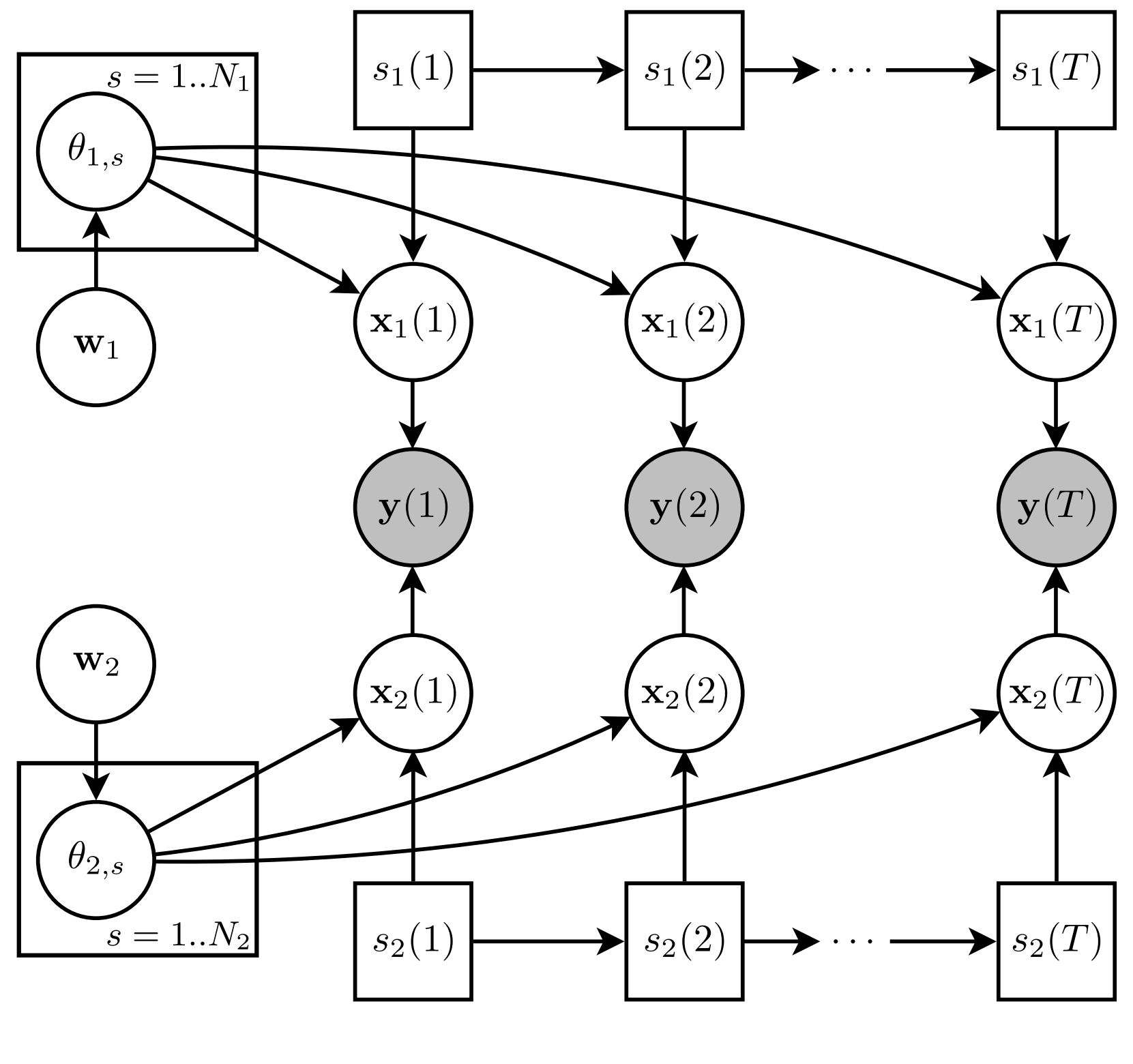


- Can incorporate covariance parameters into eigenvoice bases to adapt them as well:

$$\log \Sigma_s(w_i) = \log(S_s)\,\mathbf{w}_i + \log \bar{\Sigma}_s$$

- Combine adapted source models into factorial HMM to model mixture:

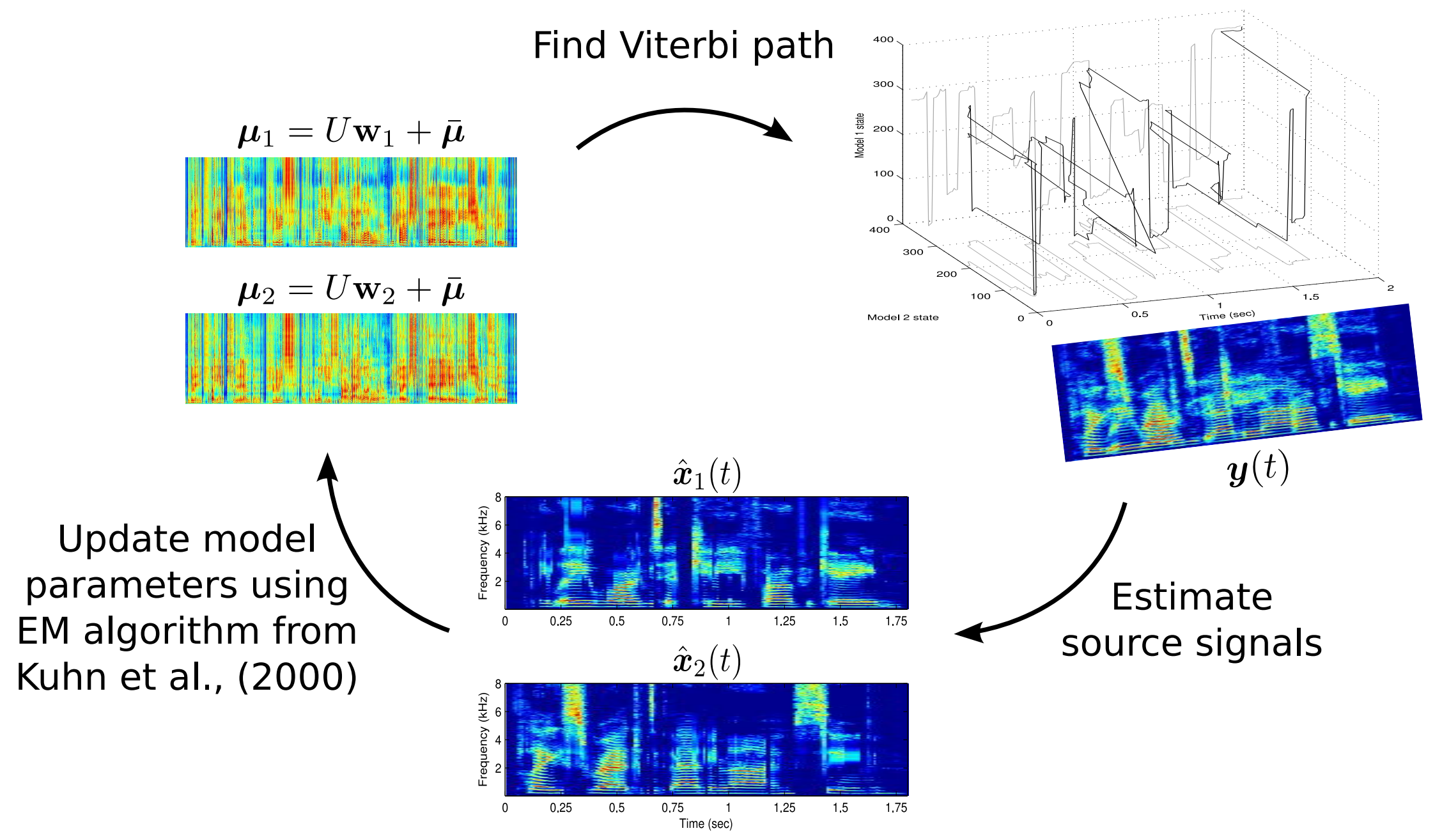
$$P\big(\boldsymbol{y}(1..T),\, s_1(1..T),\, s_2(1..T)\big) = \prod_t P\big(s_1(t) \,|\, s_1(t-1)\big)\, P\big(s_2(t) \,|\, s_2(t-1)\big)\, P\big(\boldsymbol{y}(t) \,|\, s_1(t),\, s_2(t)\big)$$



3. Adaptation algorithms

- Need to learn eigenvoice adaptation parameters $\mathbf{w}_i$ from mixture
- Exact inference in factorial HMM is intractable – $O(TN^3)$
- Propose two approximate adaptation algorithms:

1. Hierarchical algorithm (Weiss and Ellis, 2008)
   - Iteratively separate sources and learn adaptation parameters from each reconstructed source signal
   - Use aggressive pruning in factorial HMM Viterbi algorithm to make separation feasible


Find Viterbi path

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

Estimate source signals

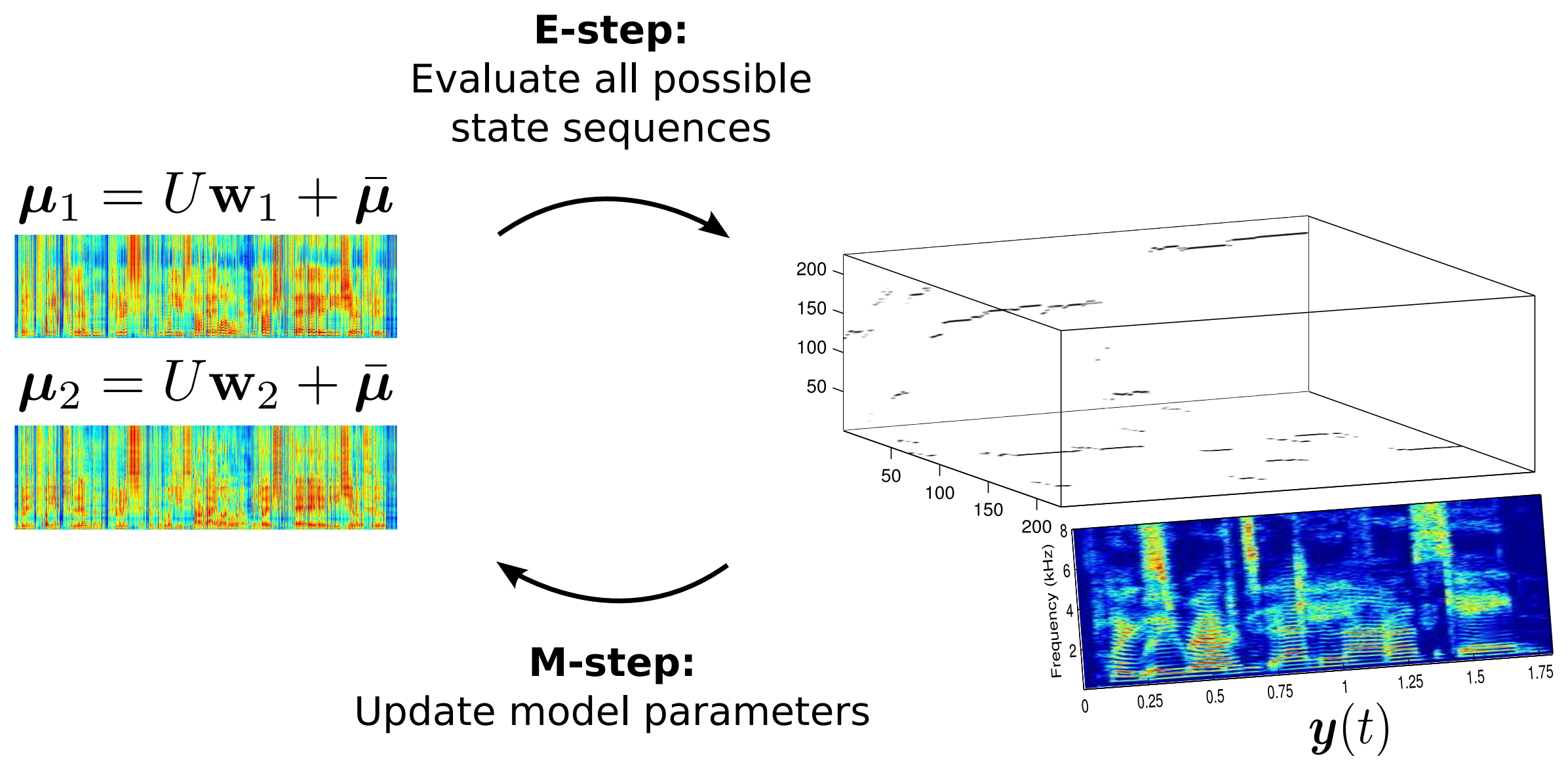
$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$
$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$
$\boldsymbol{y}(t)$
$\hat{\boldsymbol{x}}_1(t)$
$\hat{\boldsymbol{x}}_2(t)$
Frequency (kHz)
Time (sec)
0
0.25
0.5
0.75
1
1.25
1.5
1.75

2. Variational EM algorithm
   - EM algorithm based on structured variational approximation to mixed signal model (Ghahramani and Jordan, 1997)
   - Treat each source HMM independently:

$$P\big(\boldsymbol{y}(1..T),\, s_1(1..T),\, s_2(1..T)\big) \approx \prod_i Q_i\big(\boldsymbol{y}(1..T),\, s_i(1..T)\big)$$

   - Introduce variational parameters to couple them:

$$Q_i\big(\boldsymbol{y}(1..T),\, s_i(1..T)\big) = \prod_t P\big(s_i(t) \,|\, s_i(t-1)\big)\, h_{i,s_i}(t)$$


E-step:
Evaluate all possible
state sequences

M-step:
Update model parameters

$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$
$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$
$\boldsymbol{y}(t)$
50
100
150
200

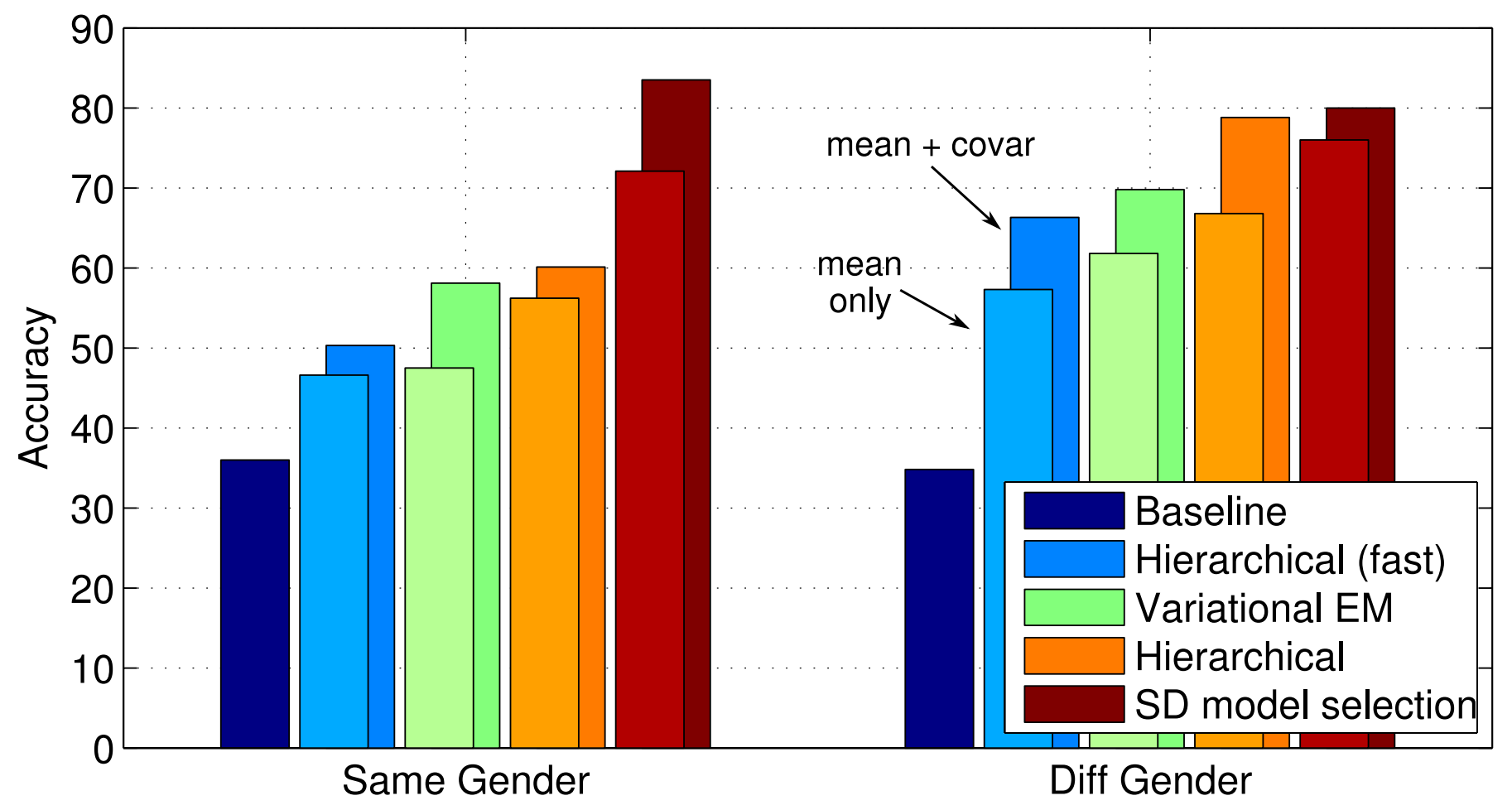
4. Experiments

- Compare two adaptation algorithms with separations based on speaker-dependent (SD) models using speaker identification algorithm from Rennie et al. (2006)
- 0 dB SNR subset of 2006 Speech Separation Challenge data set (Cooke and Lee, 2006)
- Mixtures of utterances derived from simple grammar:

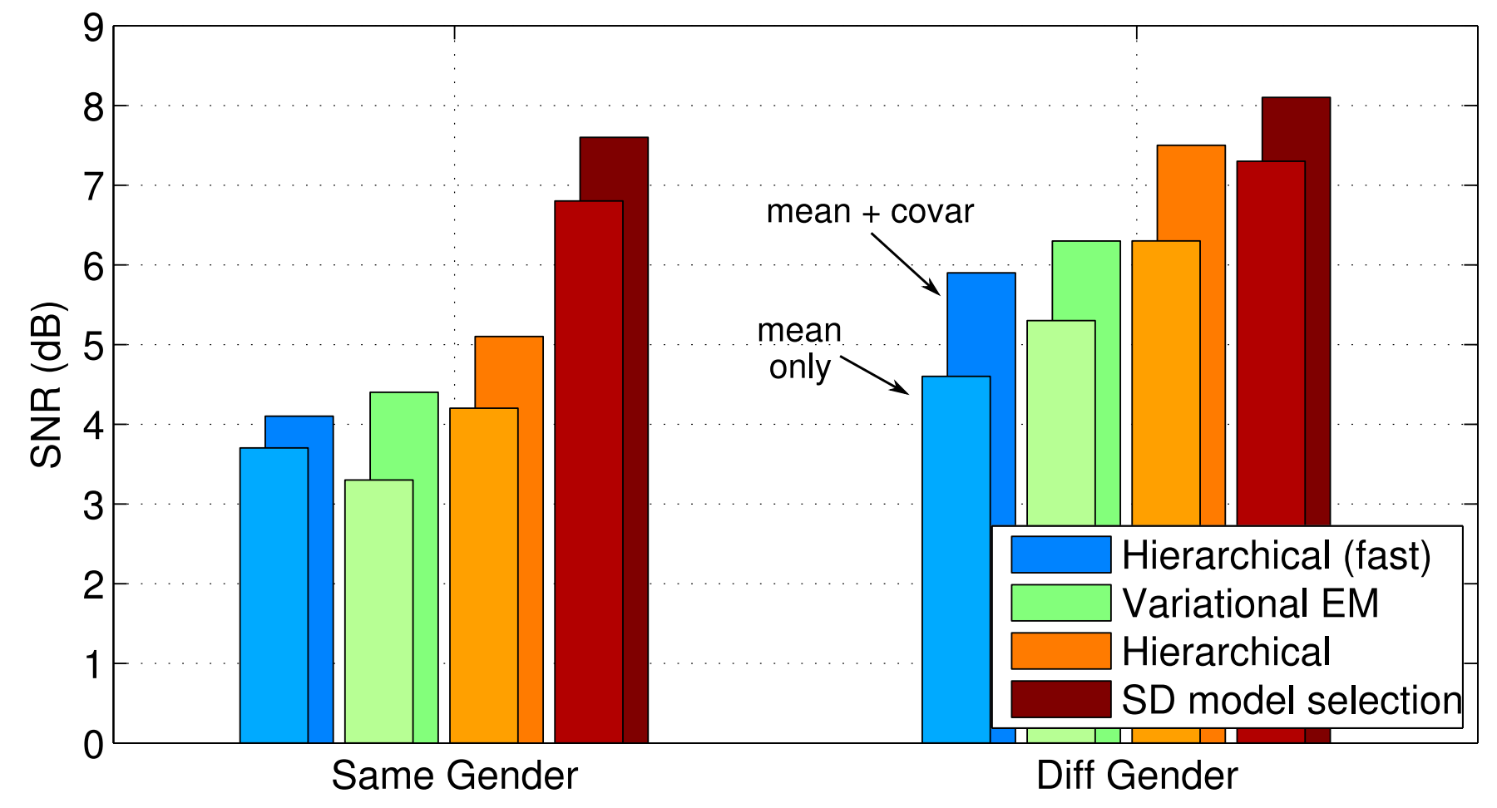
command	color	preposition	letter	digit	adverb
bin	blue	at		0-9	again
lay	green	by	a-v		now
place	red	in	x-z		please
set	white	with			soon

- Task: determine letter and digit spoken by source whose color is "white"

Digit-letter recognition accuracy:



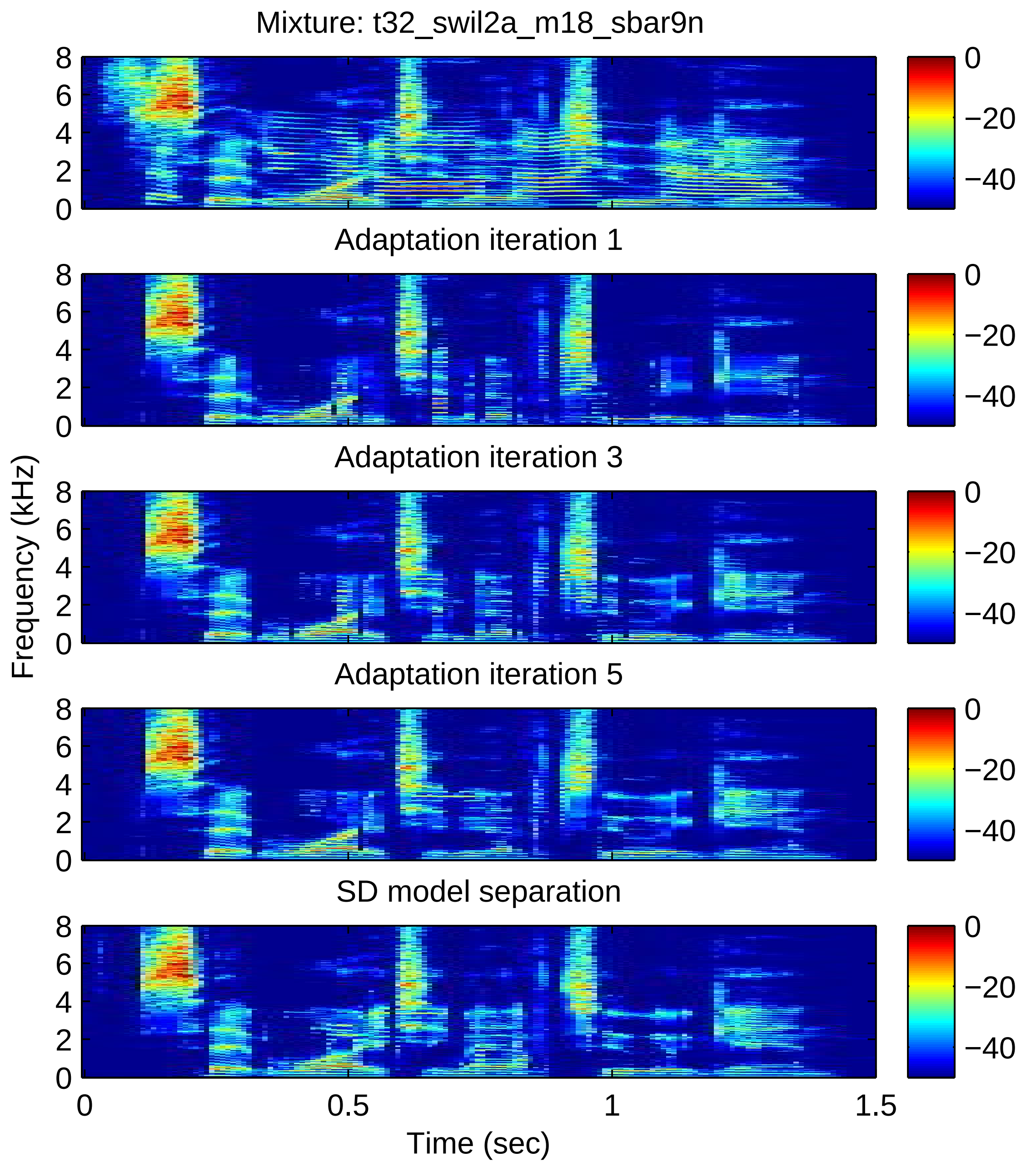
SNR of target source reconstruction:



5. Discussion

- Adapting Gaussian covariances as well as means significantly improves performance of all systems
- Adaptation comes to within 23% to 1.2% of best-case SD model performance
- Hierarchical algorithm outperforms variational EM
- But variational algorithm is significantly (∼ 4x) faster
- Performance of the hierarchical algorithm suffers when it is sped up to be as fast as the variational algorithm by pruning even more aggressively ("Hierarchical (fast)" in figures above)

6. Example



7. References

M. Cooke and T.-W. Lee. The speech separation challenge, 2006. URL http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

Z. Ghahramani and M.I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, 1997.

R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, November 2000.

S. Rennie, P. Olsen, J. Hershey, and T. Kristjansson. The Iroquois model: Using temporal dynamics to separate speakers. In *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Pittsburgh, PA, September 2006.

R. J. Weiss and D. P. W. Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech and Language*, 2008. In press.

ICASSP 2009, 19-24 April 2008, Taipei, Taiwan