Monaural speech separation using source-adapted models

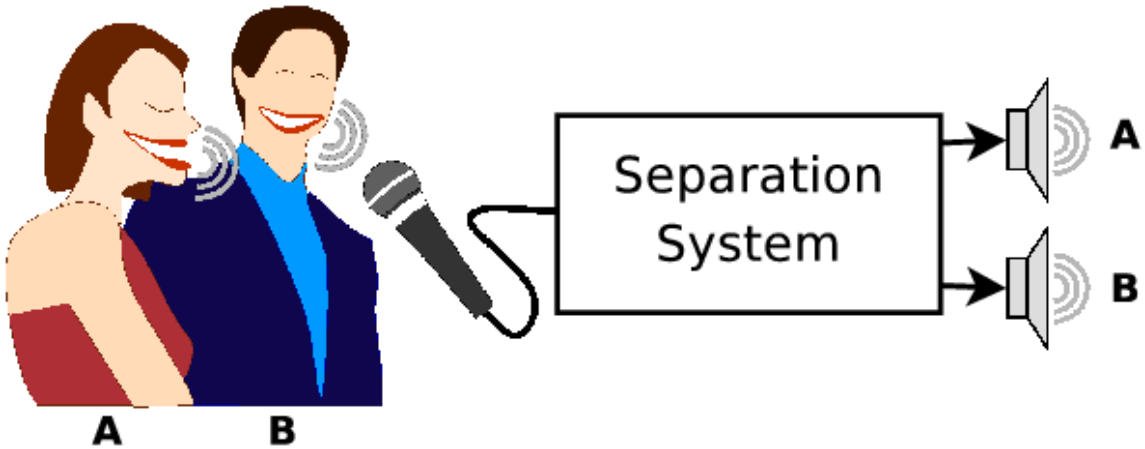
Ron Weiss, Dan Ellis
{ronw,dpwe}@ee.columbia.edu

LabROSA
Department of Electrical Enginering
Columbia University

2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

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

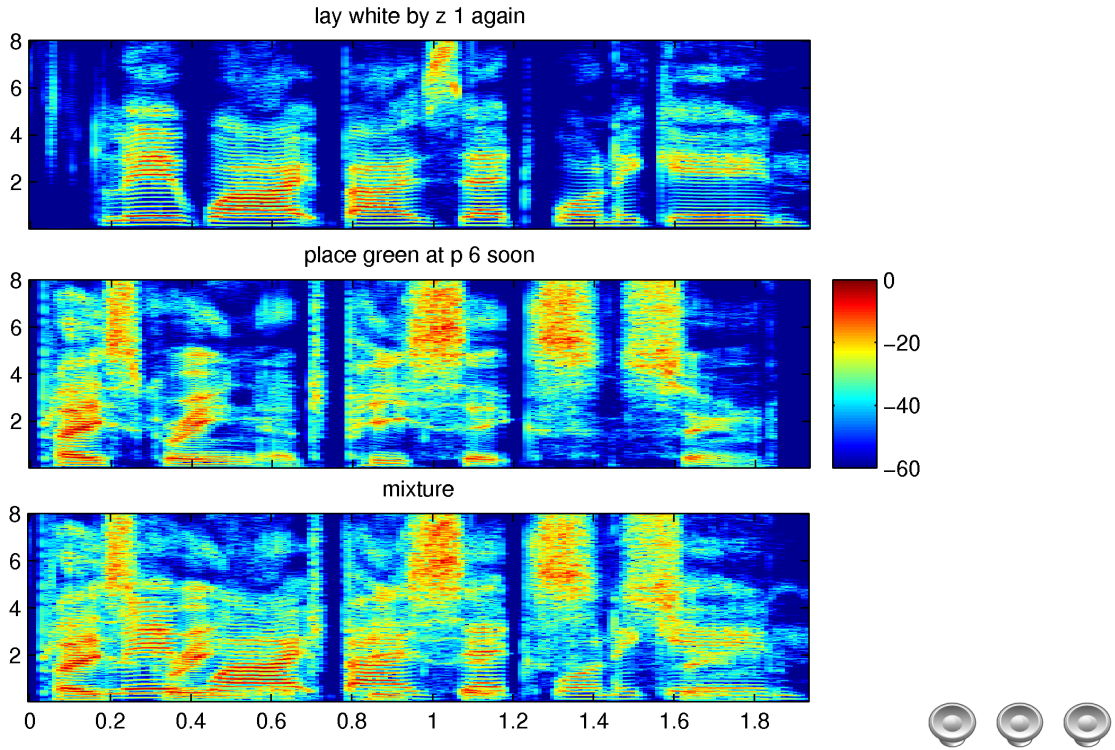
Monaural speech separation



- Given single channel recording of multiple talkers
- Infer the original source signals from mixture
- Under-determined - more unknowns (sources) than observations

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

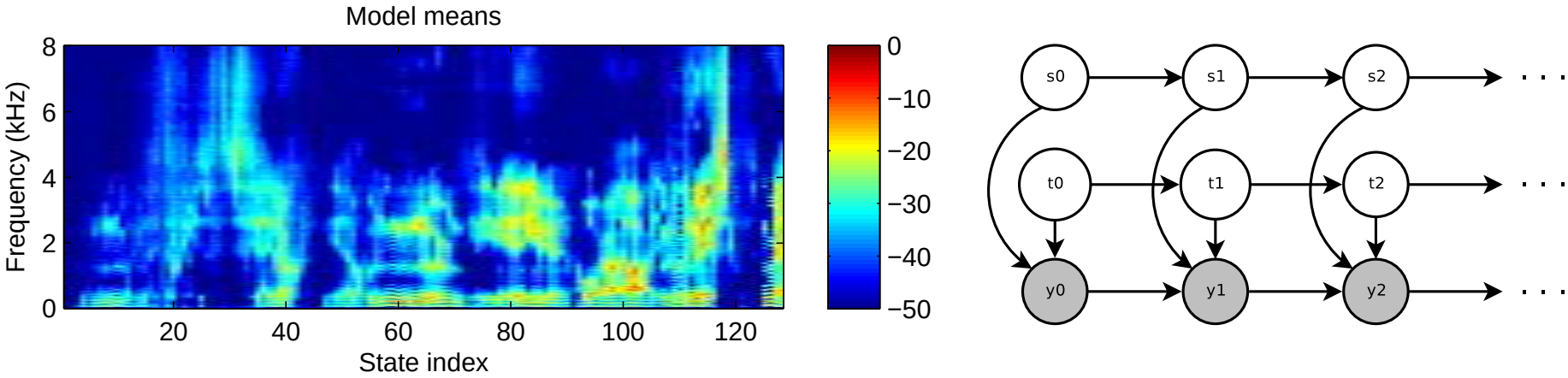
Speech separation challenge [Cooke and Lee, 2006]



- Single channel, two-talker mixtures of utterances from 34 speakers
- Constrained grammar: `command(4) color(4) preposition(4) letter(25) digit(10) adverb(4)`
- Task: determine letter and digit for source that said "white"
- -9 to 6 dB TMR

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

Model-based separation



- Use constraints from prior signal models to guide separation
- HMM, log spectral features
- Factorial model inference
  - Explain each frame of mixed signal as combination of model states
- e.g. Iroquois [Kristjansson et al., 2006]
  - Speaker-dependent models
  - Acoustic dynamics and grammar constraints
  - Superhuman performance

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

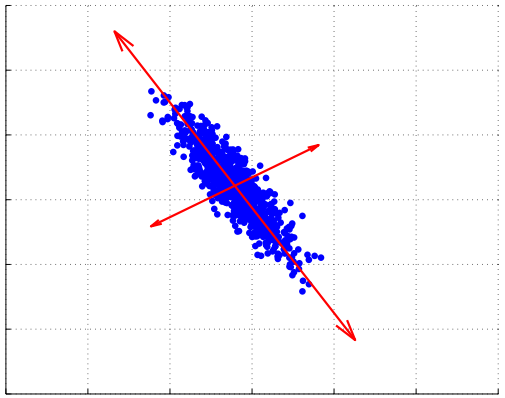
Model-based separation - Limitations

- Rely on speaker-dependent models to disambiguate sources
- What if the task isn't so well defined?
  - No a priori knowledge of speaker identities or grammar
- Adapt speaker-independent source model [Ozerov et al., 2005]
- Problems
  1. Want to adapt to a single utterance, not enough data for MLLR
     - Use PCA to reduce number of adaptation parameters - "Eigenvoices"
  2. Only observation is mixed signal
     - Iterative separation/adaptation algorithm

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

Eigenvoices [Kuhn et al., 2000]

- Train N speaker-dependent models
  - priors on space of speaker variation
- Pack model parameters (Gaussian means) into speaker supervector
- Principal component analysis to find orthonormal bases
- Speaker model is a linear combination of bases:

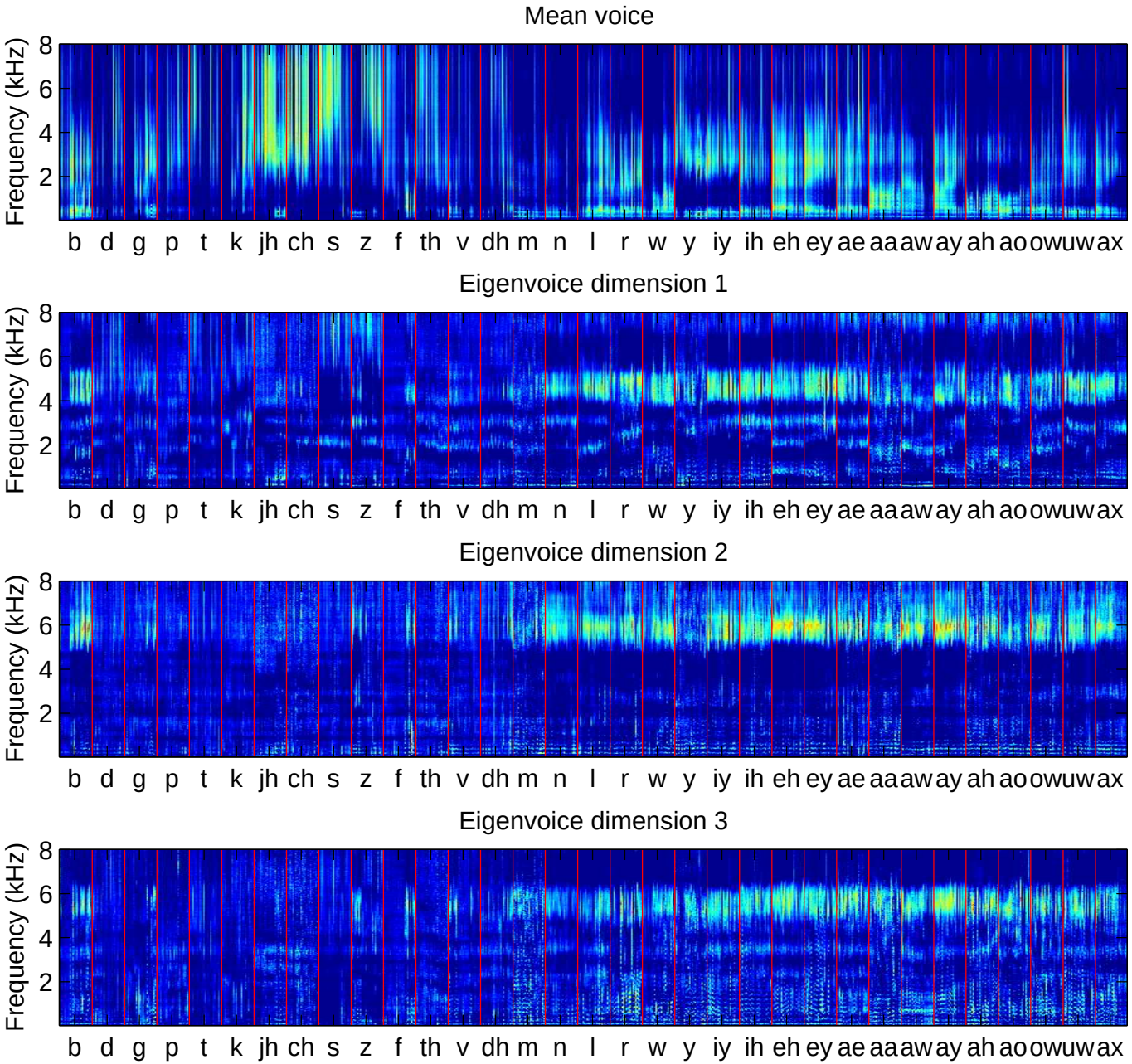


$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} + \mathbf{w} U + g$$

adapted model	=	mean voice	+	weights	eigenvoice bases	+	gain

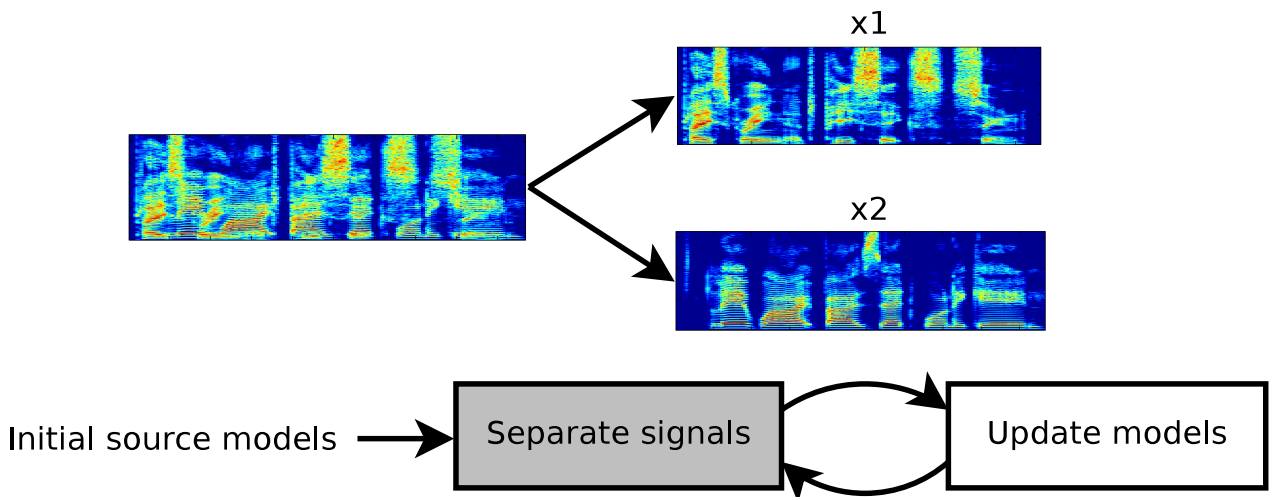
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Eigenvoice example

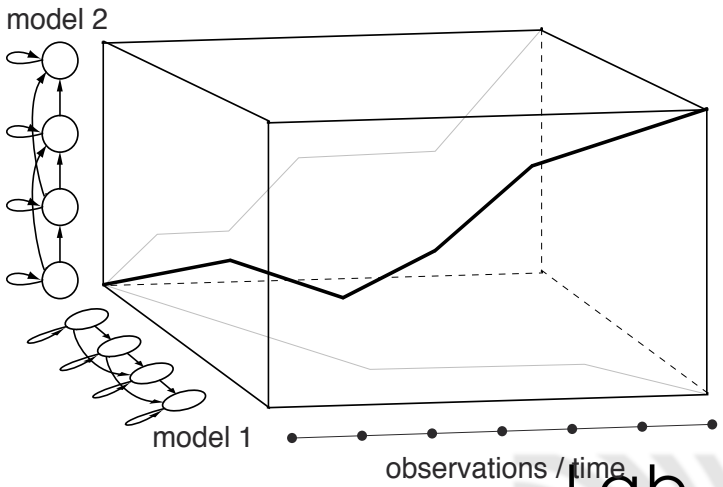


Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Separation algorithm - Signal separation

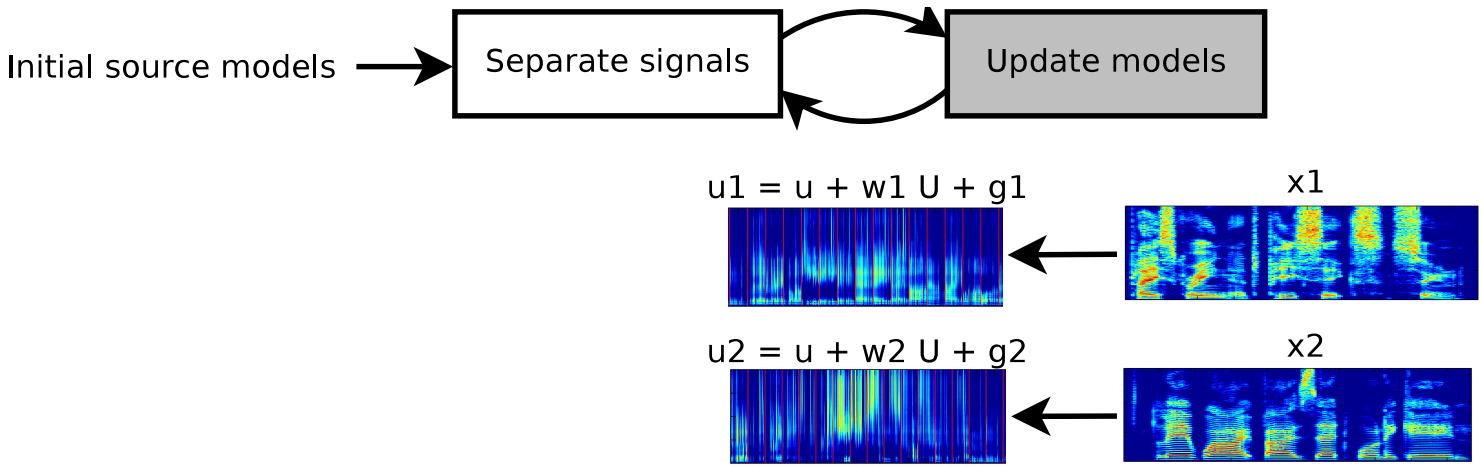


- Compose factorial HMM from adapted models
- Find maximum likelihood path using Viterbi algorithm
- Reconstruct source signals from Viterbi path



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

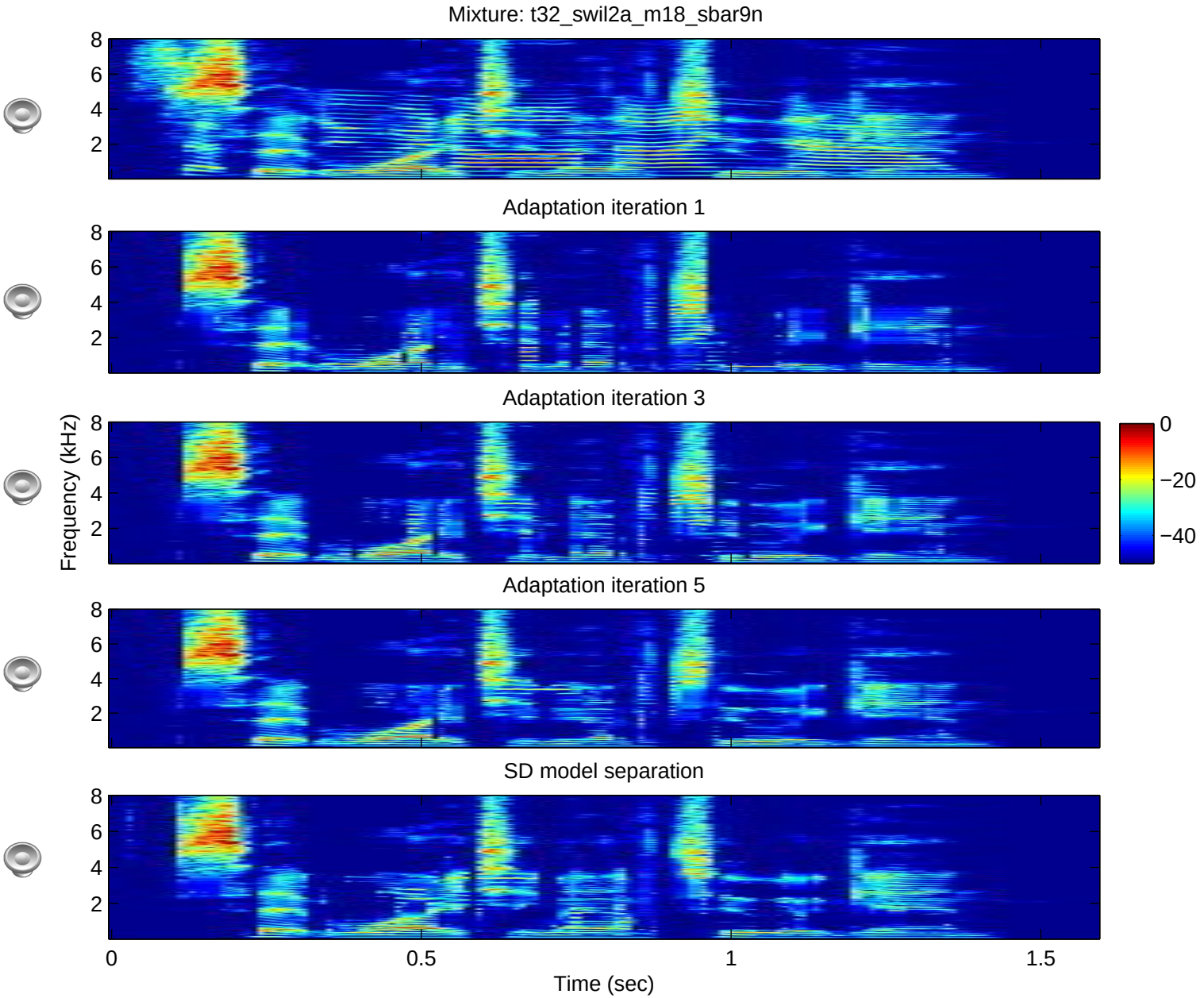
Separation algorithm - Model adaptation



- Find projection of reconstructed source signals onto eigenvoice bases
- But state sequence is hidden, need EM
  - E-step: HMM forward-backward
  - M-step: for each possible state sequence, project signal frames onto corresponding sequence of states from each eigenvoice basis vector
- Iterate...

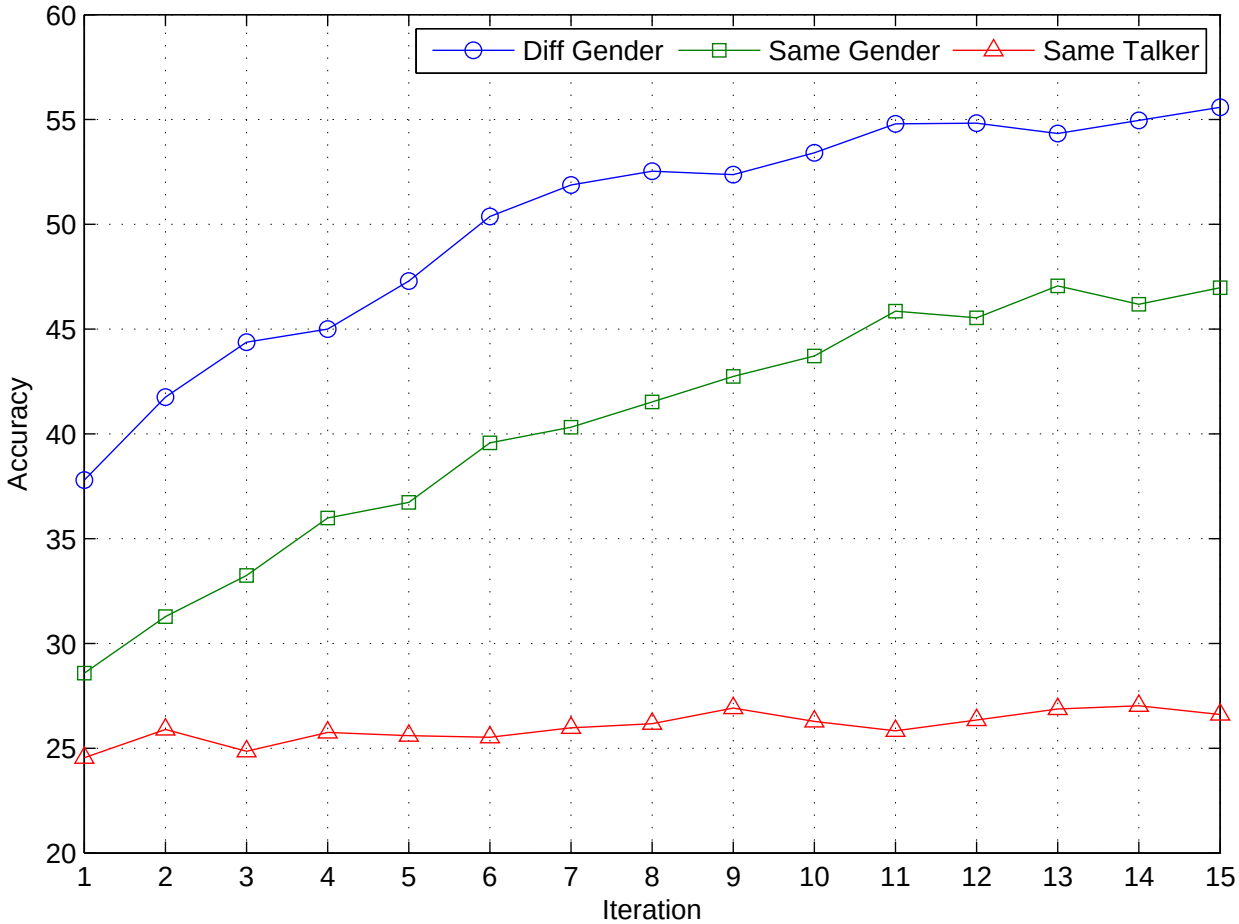
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Separation example



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

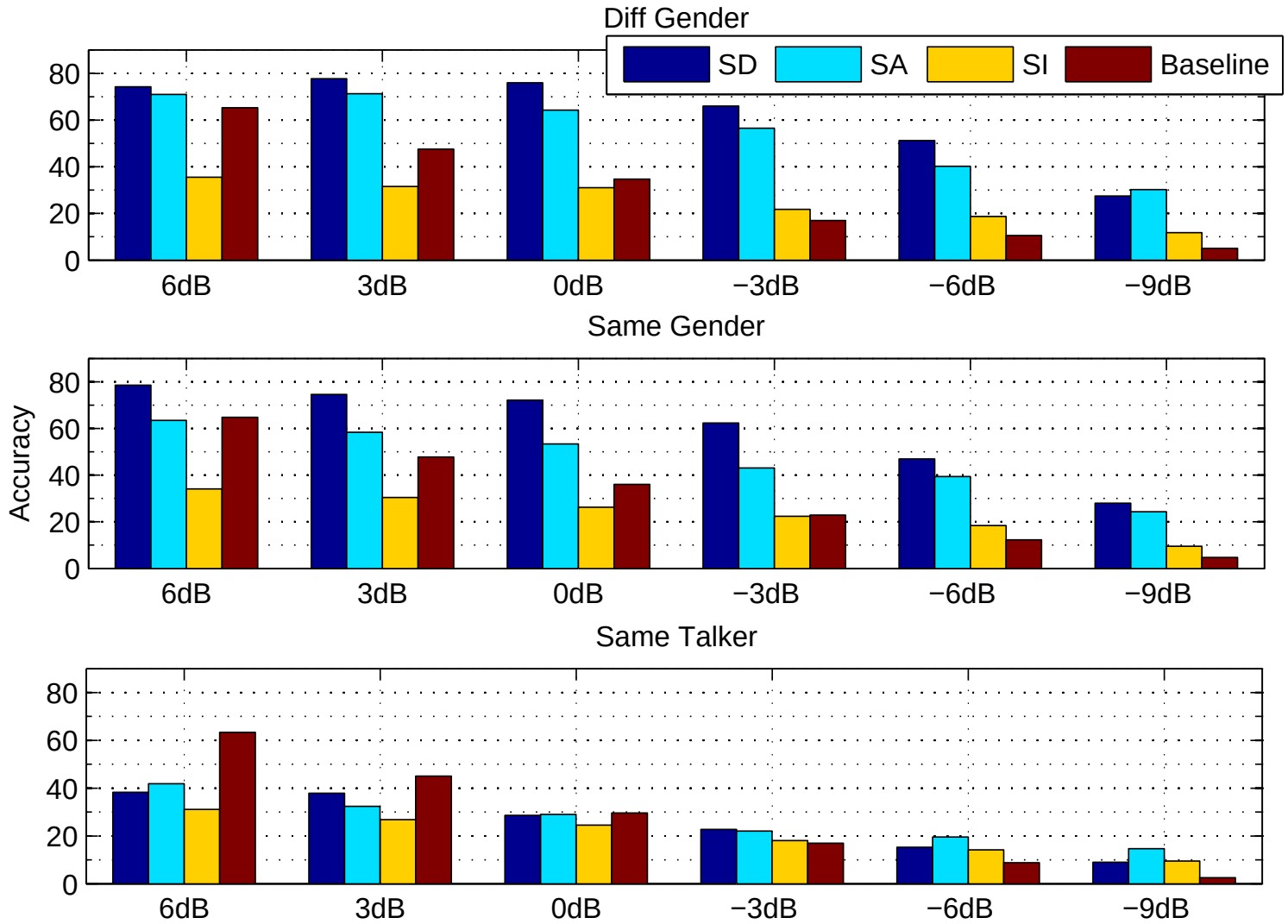
Performance



- Letter-digit accuracy averaged across all TMRs
- Adaptation improves separation
- Same talker case hard - source permutations

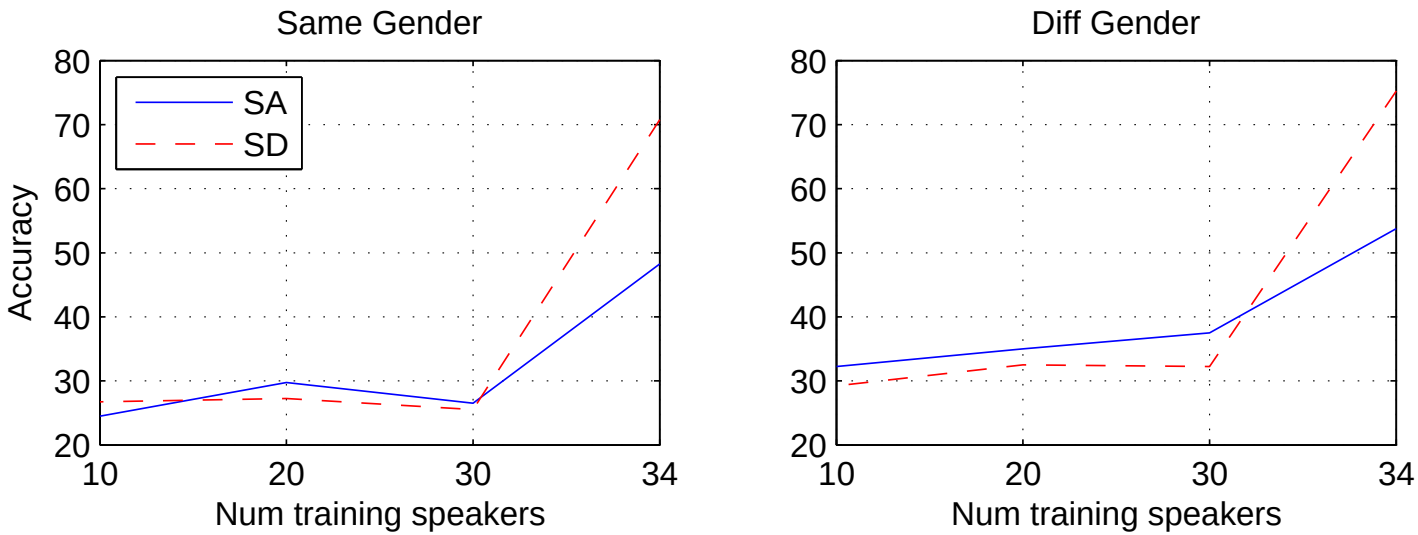
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Performance - Adapted vs. source-dependent models



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

Performance - Held out speakers



- Trained models on subset of speakers
- Tested on mixtures from held out speakers
- Performance suffers for both systems
- Relative decrease significantly bigger for SD than SA
- Open question: scale

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

Summary

- Limitations of model-based source separation
- Algorithm for model adaptation from mixed signal
- Significant improvement over speaker-independent models
- Source-dependent models better on matched training/testing data
- Adaptation helps generalize better to held out speakers

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

References

Cooke, M. and Lee, T. W. (2006).
The speech separation challenge.

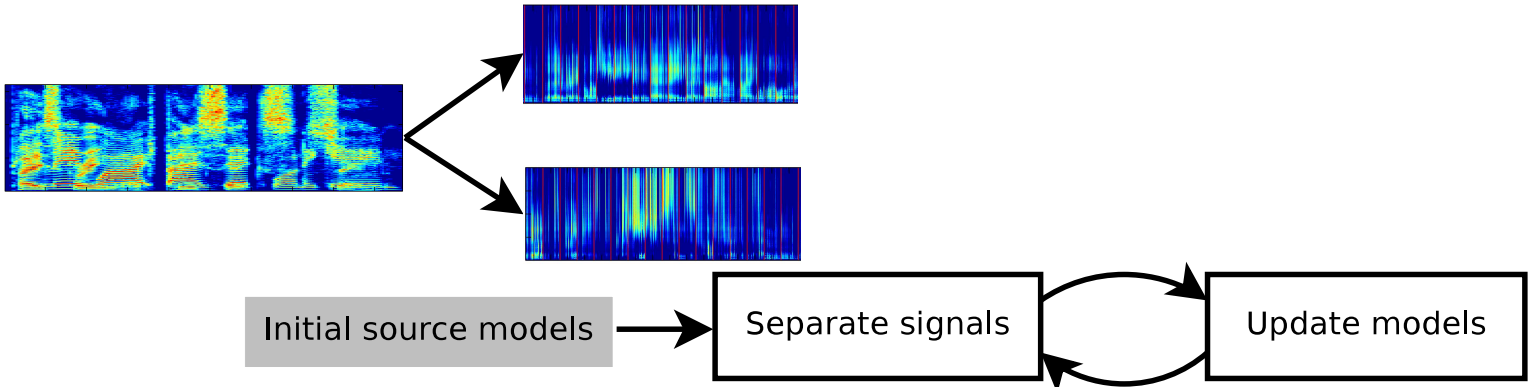
Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., and Gopinath, R. (2006).
Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system.
In *Proceedings of Interspeech*.

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

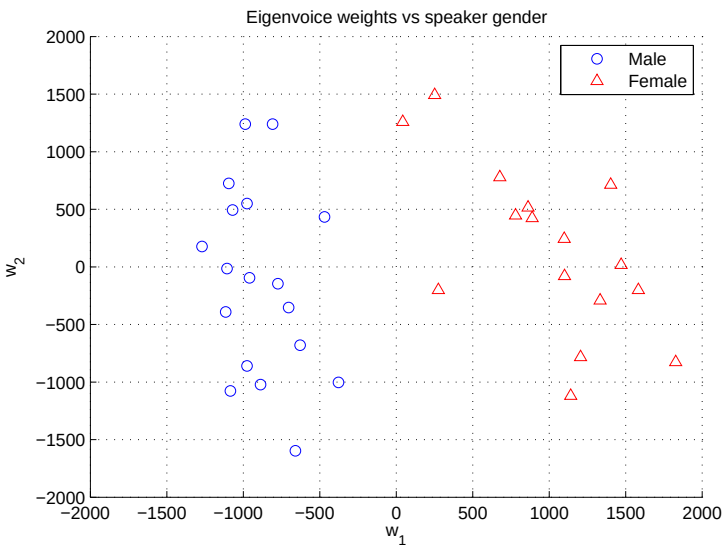
Ozerov, A., Philippe, P., Gribonval, R., and Bimbot, F. (2005).
One microphone singing voice separation using source-adapted models.
In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

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

Separation algorithm - Initialization



- Fast convergence needs good initialization
- Want to differentiate source models to get best separation
- Get initial coefficient for each eigenvoice dimension independently
  - Coarsely quantize eigenvoice weights
  - Find most likely combination in mixture



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