Underdetermined Source Separation Using Speaker Subspace Models

Thesis Defense

Ron Weiss

May 4, 2009

Audio source separation

Source: http://www.spring.org.uk/2009/03/the-cocktail-party-effect.php

- Many real world signals contain contributions from multiple sources
  - E.g. cocktail party
- Want to infer the original sources from the mixture
  - Robust speech recognition
  - Hearing aids

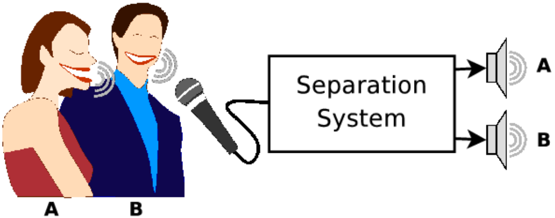
Previous work

Instantaneous mixing system

$$\begin{bmatrix} y_1(t) \\ \vdots \\ y_C(t) \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1I} \\ \vdots & \ddots & \vdots \\ a_{C1} & \dots & a_{CI} \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_I(t) \end{bmatrix}$$

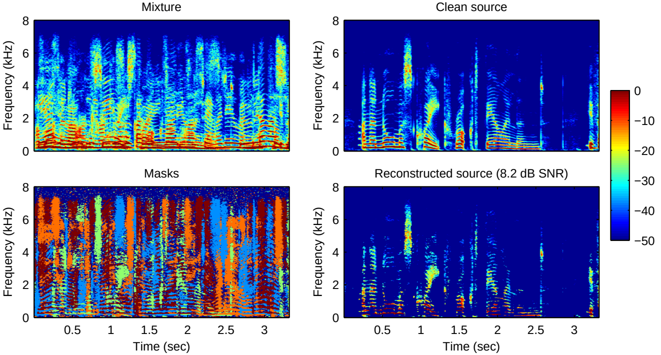
- Simplest case: more channels than sources (overdetermined)
  - Perfect separation possible
- Use constraints on source signals to guide separation
  - Independence constraints (e.g. independent component analysis)
  - Spatial constraints (e.g. beamforming)

Underdetermined source separation



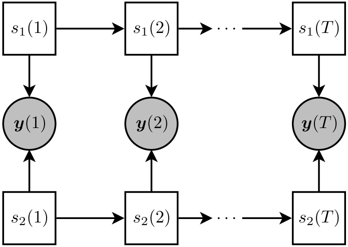
- More sources than channels, need stronger constraints
- CASA: Use perceptual cues similar to human auditory system
  - Segment STFT into short glimpses of each source
  - By harmonicity, common onset, etc.
  - Sequential grouping heuristics
  - Create time-frequency mask for each source
- Inference based on prior source models

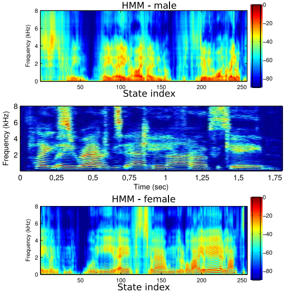
Time-frequency masking



- Natural sounds tend to be sparse in time and frequency
  - 10% of spectrogram cells contain 78% of energy
- And redundant
  - Still intelligible when 22% of source energy is masked

Model-based separation





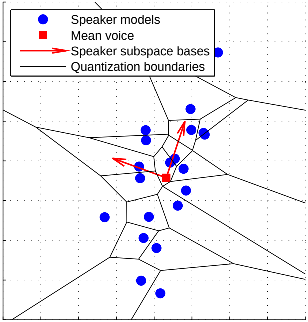
- Use constraints from prior source models to guide separation
  - Leverage differences in spectral characteristics of different sources
- Hidden Markov models, log spectral features
- Factorial model inference
- e.g. IBM Iroquois system [Kristjansson et al., 2006]
  - Speaker-dependent models
  - Acoustic dynamics *and* grammar constraints
  - Superhuman performance under some conditions

Model-based separation – Limitations

- Rely on speaker-dependent models to disambiguate sources
- What if the task isn't so well defined?
  - No prior knowledge of speaker identities or grammar
- Use speaker-independent (SI) model for all sources
  - Need strong temporal constraints or sources will permute
    - "place white by t 4 now" mixed with "lay green with p 9 again"
    - Separated source: "place white by t p 9 again"
- Solution: adapt speaker-independent model to compensate

1. Introduction
2. Speaker subspace model
   - Model adaptation
   - Eigenvoices
3. Monaural speech separation
4. Binaural separation
5. Conclusions

Model selection vs. adaptation

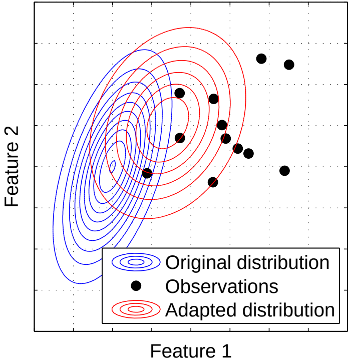


Model selection (e.g. [Kristjansson et al., 2006])

- Given set of speaker-dependent (SD) models:
  1. Identify sources in mixture
  2. Use corresponding models for separation
- How to generalize to speakers outside of training set?
  - Selection – choose closest model
  - Adaptation – interpolate

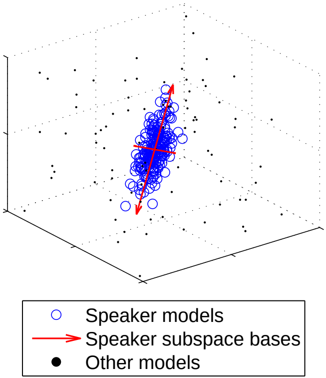
Model adaptation

- Adjust model parameters to better match observations
- Caveats
  1. Want to adapt to a single utterance, not enough data for MLLR, MAP
     - Need adaptation framework with few parameters
  2. Observations are mixture of multiple sources
     - Iterative separation/adaptation algorithm



Eigenvoice adaptation [Kuhn et al., 2000]

- Train a set of SD models
  - Pack params into speaker supervector
  - Samples from space of speaker variation
- Principal component analysis to find orthonormal bases for speaker subspace
- Model is linear combination of bases



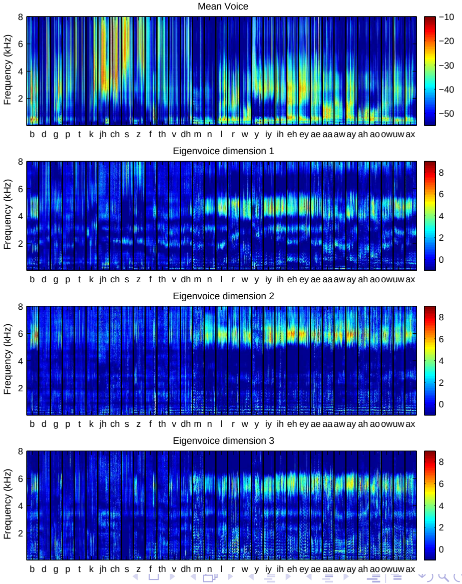
Eigenvoice adaptation

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

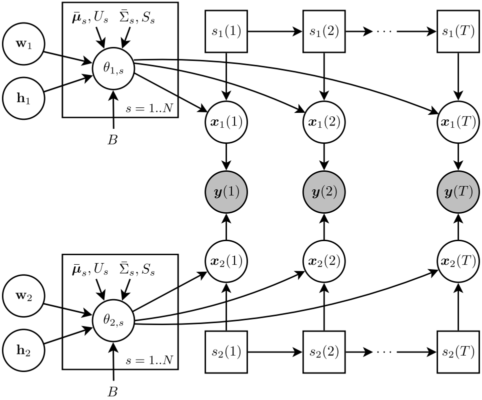
$\boldsymbol{\mu}$	$\bar{\boldsymbol{\mu}}$	U	$\mathbf{w}$	B	$\mathbf{h}$
adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

Eigenvoice bases

- Mean voice
  = speaker-independent model
- Eigenvoices shift formant frequencies, add pitch
- Independent bases to capture channel variation

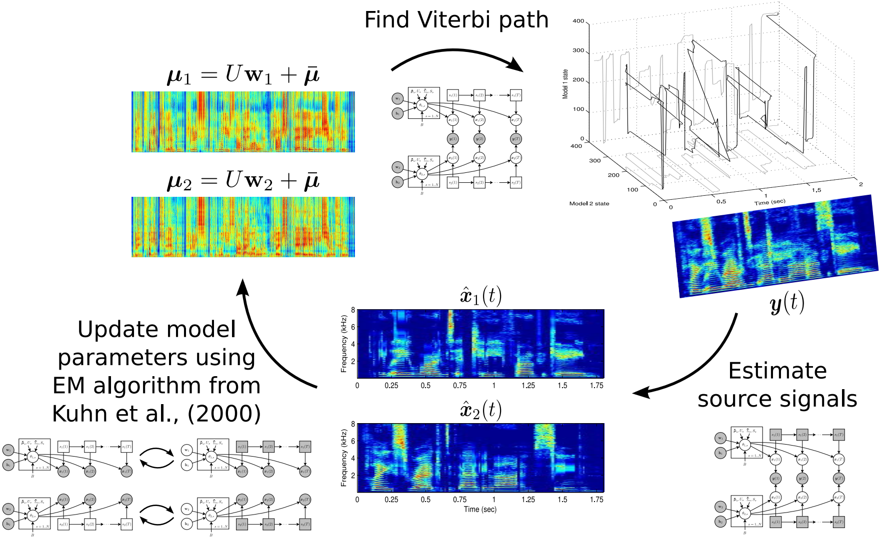


Eigenvoice factorial HMM

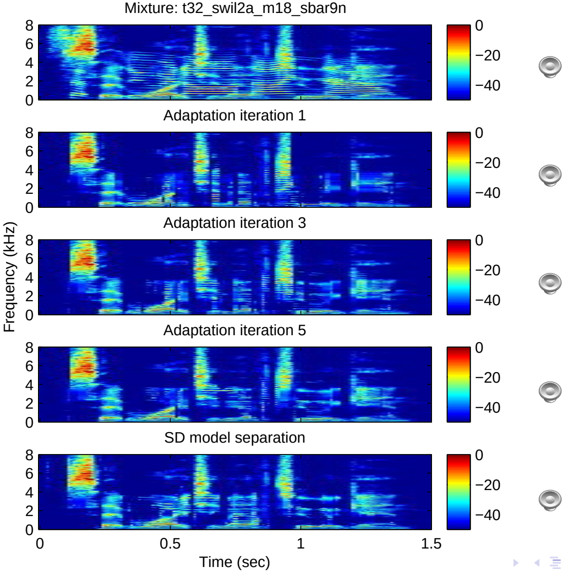


- Model mixture with combination of source HMMs
- Need adaptation parameters $\mathbf{w}_i$ to estimate source signals $x_i(t)$ and vice versa

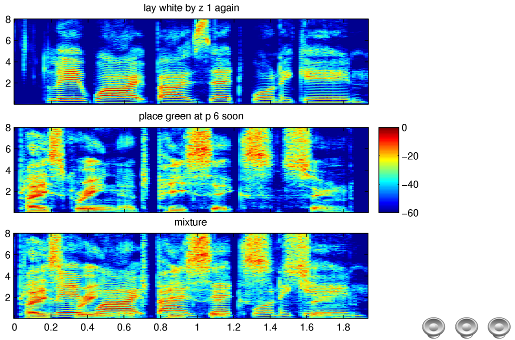
Adaptation algorithm



Adaptation example



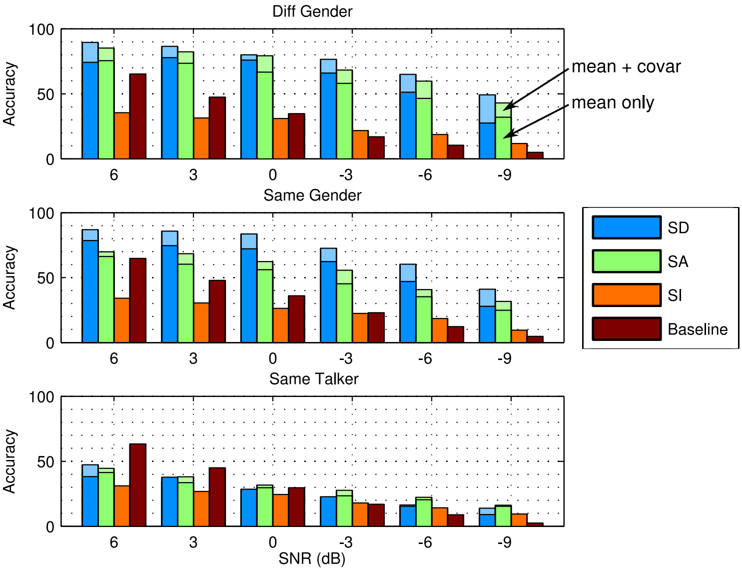
2006 Speech separation challenge [Cooke and Lee, 2006]



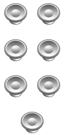
- Single channel mixtures of utterances from 34 different speakers
- Constrained grammar:

  `command(4) color(4) preposition(4) letter(25) digit(10) adverb(4)`
- Separation/recognition task
  - Determine letter and digit for source that said "white"

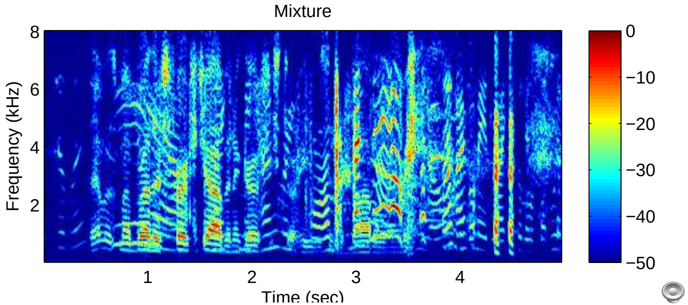
Performance – Adapted vs. source-dependent models



−3 dB

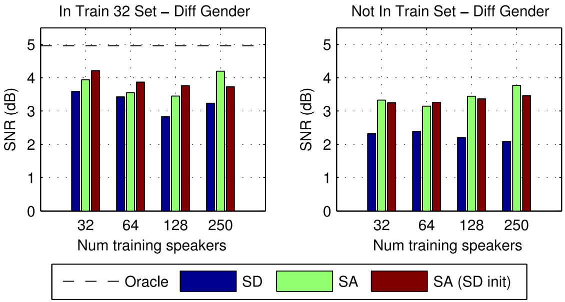


Experiments – Switchboard



- What about previously unseen speakers?
- Switchboard: corpus of conversational telephone speech
  - 200+ hours, 500+ speakers
- Task significantly more difficult than Speech Separation Challenge
  - Spontaneous speech
  - Large vocabulary
  - Significant channel variation across calls

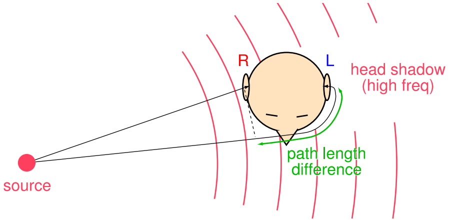
Switchboard – Results



- Adaptation outperforms SD model selection
    - Model selection errors due to channel variation
- SD performance drops off under mismatched conditions
- SA performance improves as number of training speakers increases

1. Introduction
2. Speaker subspace model
3. Monaural speech separation
4. Binaural separation
   - Mixed signal model
   - Parameter estimation and source separation
   - Experiments
5. Conclusions

Binaural audition

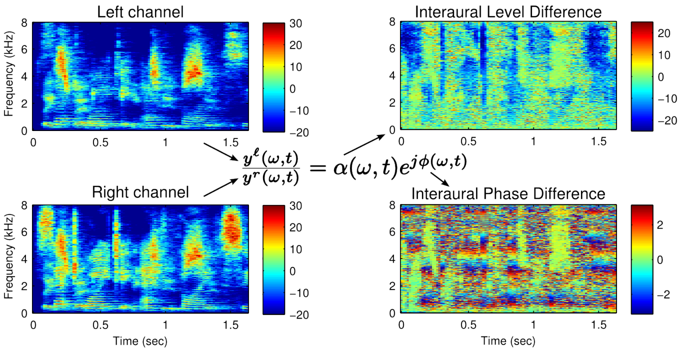


$$y^{\ell}(t) = \sum_i x_i(t - \tau_i^{\ell}) * h_i^{\ell}(t)$$

$$y^{r}(t) = \sum_i x_i(t - \tau_i^{r}) * h_i^{r}(t)$$

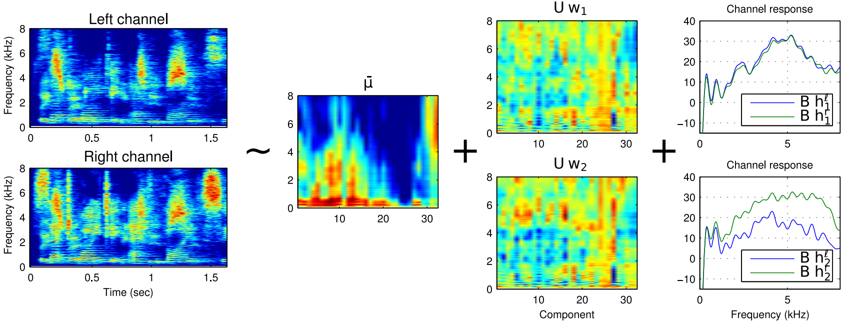
- Given stereo recording of multiple sound sources
- Utilize spatial cues to aid separation
  - Interaural time difference (ITD)
  - Interaural level difference (ILD)

MESSL: Interaural model [Mandel and Ellis, 2007]



- Model-based EM Source Separation and Localization
- Probabilistic model of interaural spectrogram
    - Independent of underlying source signals
- Assume each time-frequency cell is dominated by a single source
- EM algorithm to learn model parameters for each source
- Derive probabilistic time-frequency masks for separation

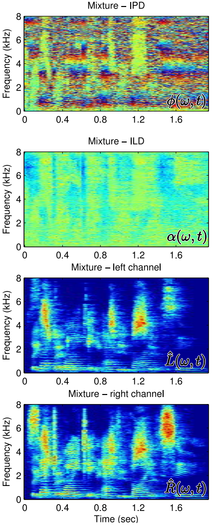
MESSL-SP: Source prior



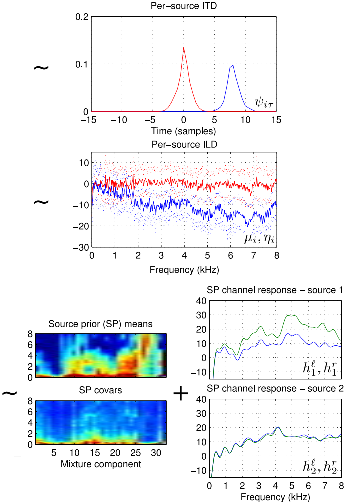
- Extend MESSL to include prior source model
- Pre-trained GMM for speech signals in mixture
- Channel model to compensate for HRTF and reverberation
- Can incorporate eigenvoice adaptation (MESSL-EV)

Parameter estimation and source separation

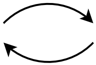
Observations



Parameters



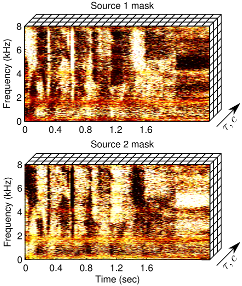
E-step
Use parameters to compute posteriors of hidden variables



M-step
Use posteriors to update parameters

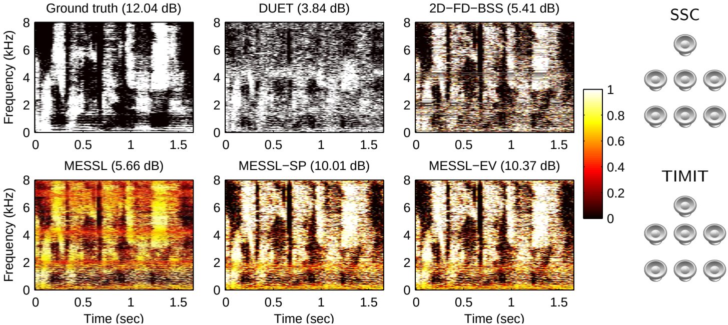
Posteriors

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



Separate sources by multiplying mixture by different masks

Experiments

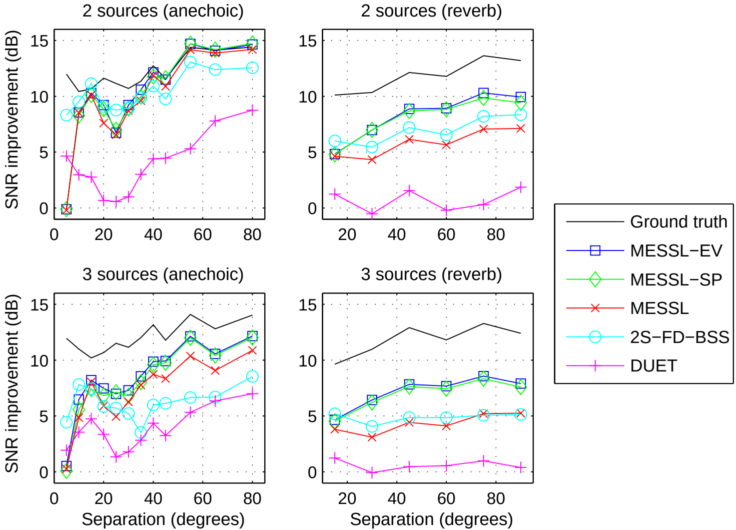


- Mixtures of 2 and 3 speech sources, anechoic and reverberant
- Evaluated on TIMIT and SSC test data
- Source models trained on SSC data (32 components)
- Compare MESSL systems to:

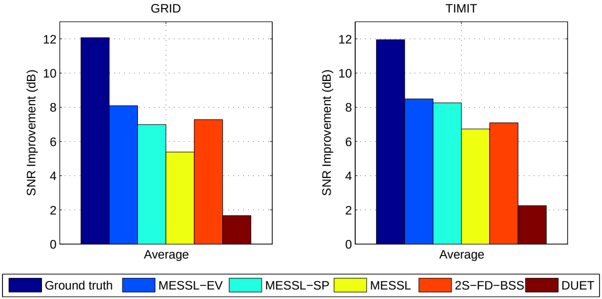
  DUET – Clustering using ILD/ITD histogram [Yilmaz and Rickard, 2004]

  2S-FD-BSS – Frequency domain ICA [Sawada et al., 2007]

Experiments – Performance as function of distractor angle



Experiments – Matched vs. mismatched



- SSC – matched train/test speakers
  - MESSL-EV, MESSL-SP beat MESSL baseline by $\sim$ 3 dB in reverb
  - MESSL-EV beats MESSL-SP by $\sim$ 1 dB on anechoic mixtures
- TIMIT – mismatched train/test speakers
  - Small difference between MESSL-EV and MESSL-SP

1 Introduction

2 Speaker subspace model

3 Monaural speech separation

4 Binaural separation

5 Conclusions

Summary

- Prior signal models for underdetermined source separation
- Subspace model for source adaptation
  - Adapt Gaussian means and covariances using a single utterance
  - Natural extension to compensate for source-independent channel effects
- Monaural separation
  - Speaker-dependent $>$ speaker-adapted $\gg$ speaker-independent
  - Adaptation helps generalize better to held out speakers
  - Improves as number of training speakers increases
- Binaural separation
  - Extend MESSL framework to use source models (joint with M. Mandel)
  - Improved performance by incorporating simple SI model
  - Smaller improvement with adaptation

Contributions

- Model-based source separation making minimal assumptions using subspace adaptation
- Extend model-based approach to binaural separation

Ellis, D. P. W. and Weiss, R. J. (2006).
Model-based monaural source separation using a vector-quantized phase-vocoder representation.
In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages V–957–960.

Weiss, R. J. and Ellis, D. P. W. (2006).
Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking.
In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, pages 31–36.

Weiss, R. J. and Ellis, D. P. W. (2007).
Monaural speech separation using source-adapted models.
In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 114–117.

Weiss, R. J. and Ellis, D. P. W. (2008).
Speech separation using speaker-adapted eigenvoice speech models.
Computer Speech and Language, In Press, Corrected Proof:–.

Weiss, R. J., Mandel, M. I., and Ellis, D. P. W. (2008).
Source separation based on binaural cues and source model constraints.
In *Proc. Interspeech*, pages 419–422.

Weiss, R. J. and Ellis, D. P. W. (2009).
A Variational EM Algorithm for Learning Eigenvoice Parameters in Mixed Signals.
In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

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 *Proc. Interspeech*, pages 97–100.

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.

Mandel, M. I. and Ellis, D. P. W. (2007).
EM localization and separation using interaural level and phase cues.
In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

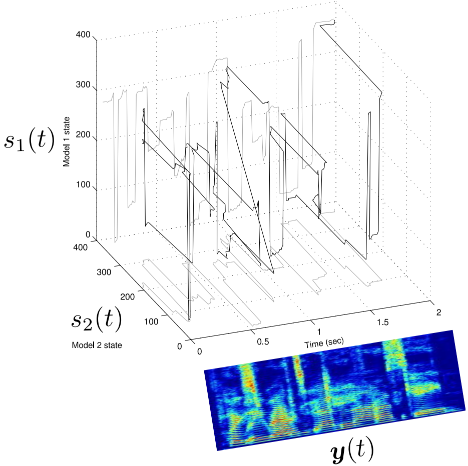
Sawada, H., Araki, S., and Makino, S. (2007).
A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures.
In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Yilmaz, O. and Rickard, S. (2004).
Blind separation of speech mixtures via time-frequency masking.
IEEE Transactions on Signal Processing, 52(7):1830–1847.

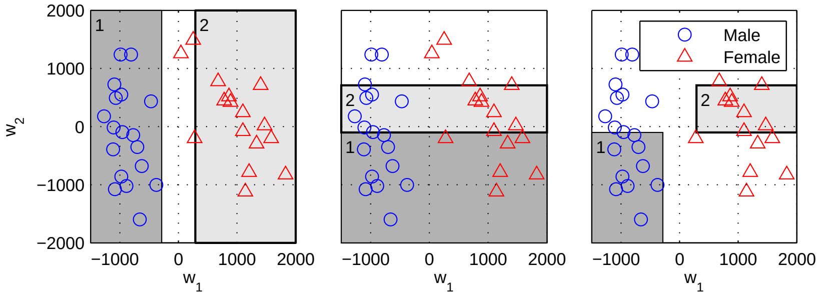
6 Extra slides

Factorial HMM separation

- Each source signal is characterized by state sequence through its HMM
- Viterbi algorithm to find maximum likelihood path through combined factorial HMM
- Reconstruct source signals using Viterbi path
- Aggressively prune unlikely paths to speed up separation

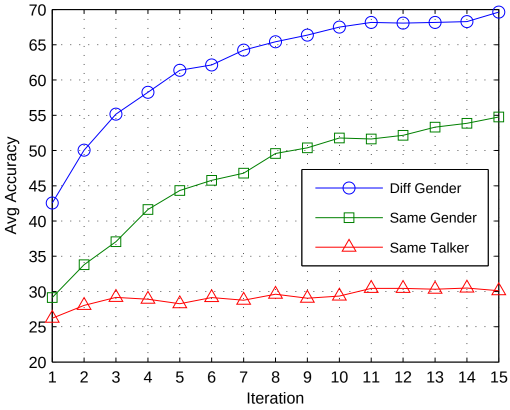


Adaptation algorithm initialization



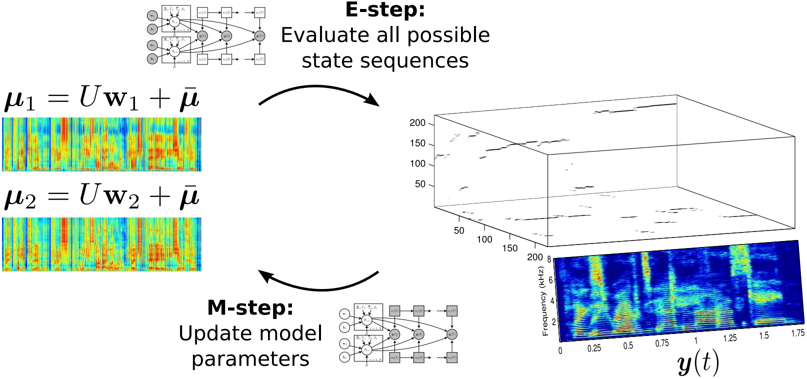
- Fast convergence needs good initialization
- Want to differentiate source models to get best initial separation
- Treat each eigenvoice dimension independently
    - Coarsely quantize weights
    - Find most likely combination in mixture

Adaptation performance



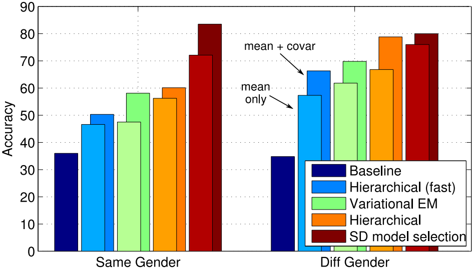
- Letter-digit accuracy averaged across all TMRs
- Adaptation clearly improves separation
- Same talker case hard – source permutations

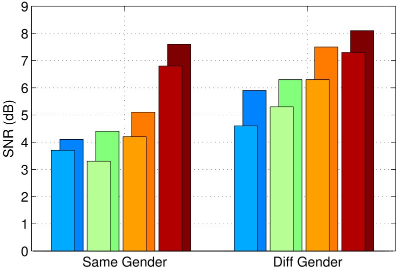
Variational learning



- Approximate EM algorithm to estimate adaptation parameters
- Treat each source HMM independently
- Introduce variational parameters to couple them

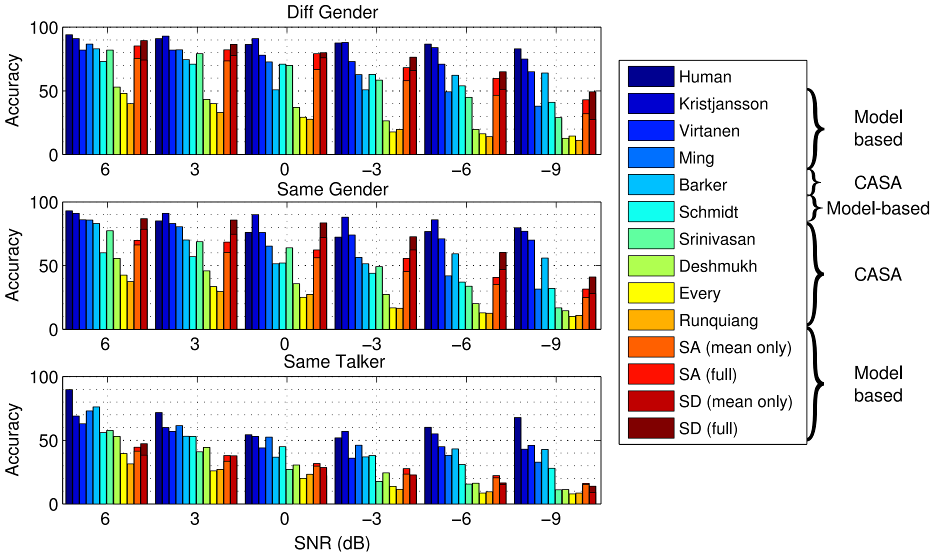
Performance – Learning algorithm comparison





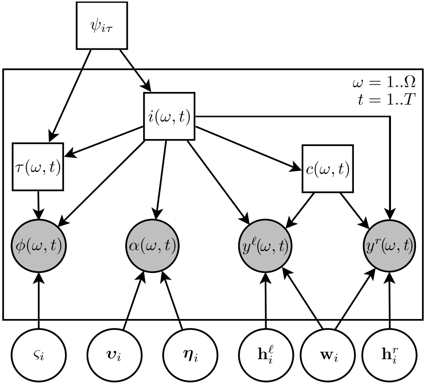
- Adapting Gaussian covariances and means significantly improves performance
- Hierarchical algorithm outperforms variational EM
- But variational algorithm is significantly ($\sim$ 4x) faster
- At same speed variational EM performs better

Performance – Comparison to other participants



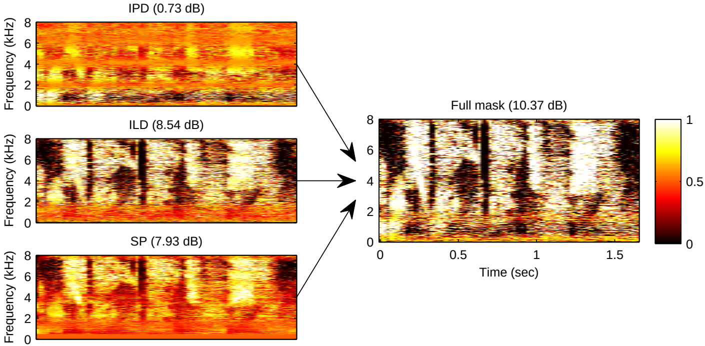
MESSL-EV: Putting it all together

- Big mixture of Gaussians
- Interaural model
  - ITD: Gaussian for each source and time delay
  - ILD: Single Gaussian for each source
- Source model
  - Separate channel responses for each source at each ear
  - Both channels share eigenvoice adaptation parameters



Explain each point in spectrogram by a particular source, time delay, and source model mixture component

MESSL-EV example



- IPD informative in low frequencies, but not in high frequencies
- ILD primarily adds information about high frequencies
- Source model introduces correlations across frequency and emphasizes reliable time-frequency regions
    - Helps resolve ambiguities in interaural parameters from reverberation and spatial aliasing

Just for fun...