Underdetermined Source Separation Using Speaker Subspace Models

Ron Weiss

Oct 2, 2009

1. Introduction
2. Speaker subspace model
3. Monaural speech separation
4. Binaural separation
5. Conclusions

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, music
- Want to infer the original source signals from the mixture
  - Robust speech recognition
  - Hearing aids
  - Un-mixing music recordings
  - Polyphonic music transcription

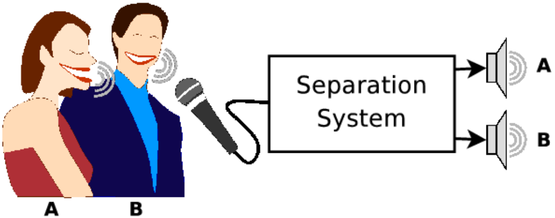
Separation approaches

Instantaneous mixing system

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

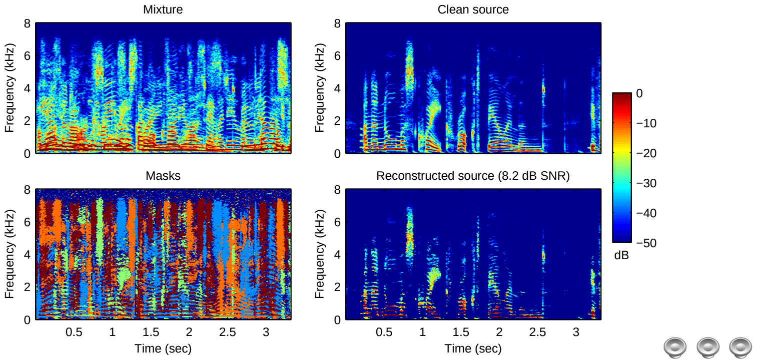
- Simplest case: more channels than sources (overdetermined)
  - Perfect separation possible
- Use constraints on source signals to guide separation
  - Statistical independence constraints (e.g. ICA)
  - Spatial constraints (e.g. beamforming)

Underdetermined source separation



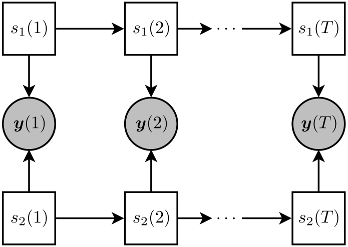
- More sources than channels, need stronger constraints
- CASA: Use perceptual cues similar to human auditory system
  - Segment STFT into short glimpses of each source
  - By harmonicity, common onset, etc.
  - Sequential grouping heuristics
  - Create time-frequency mask for each source
- Prior distribution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

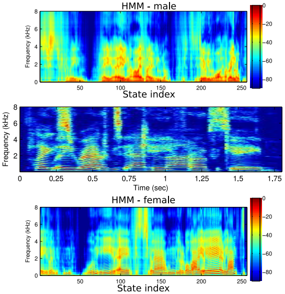
Time-frequency masking

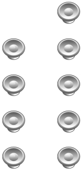


- Natural sounds tend to be sparse in time and frequency
  - 10% of spectrogram cells contain 78% of energy
- And redundant
  - Still intelligible when 22% of source energy is masked

Model-based separation







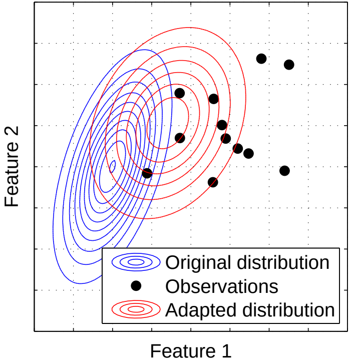
- Use constraints from prior source models to guide separation
  - Leverage differences in spectral characteristics of different sources
- Borrow machinery from speech recognition
- e.g. IBM Iroquois system [Kristjansson et al., 2006]
  - Speaker-dependent hidden Markov models
  - Acoustic dynamics *and* 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
    - "lay white in b 3 please" mixed with "bin blue in d 9 soon"
    - Separated source: "lay white in d 9 please"
- Solution: adapt speaker-independent model to compensate

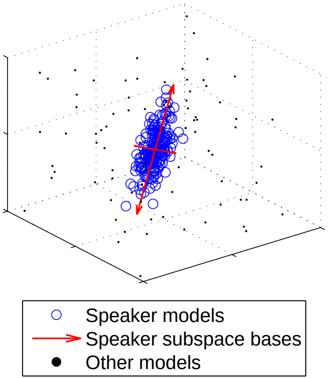
Model adaptation

- 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



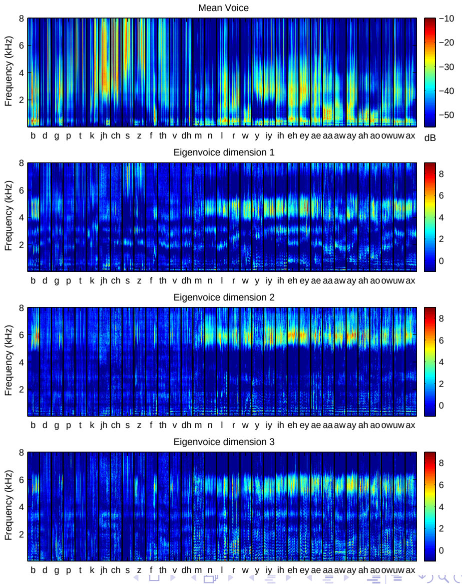
Eigenvoice adaptation

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

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

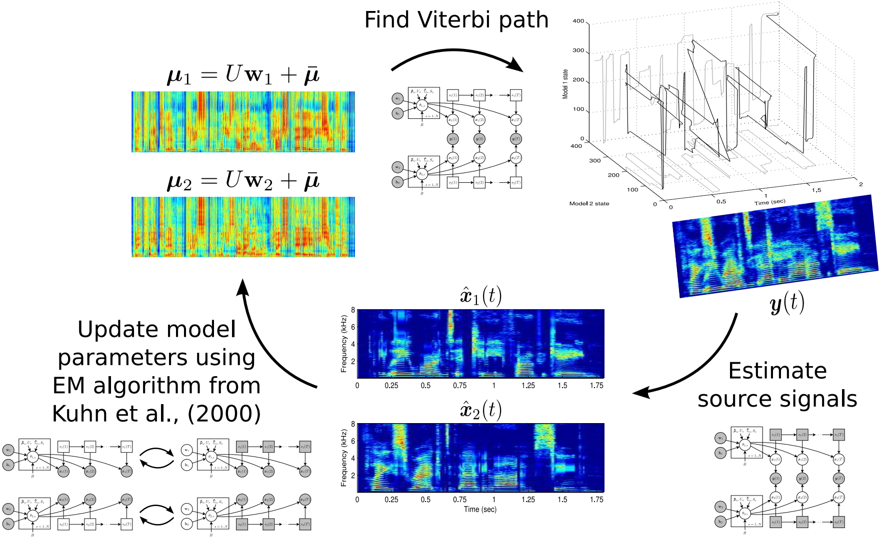
Eigenvoice bases

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

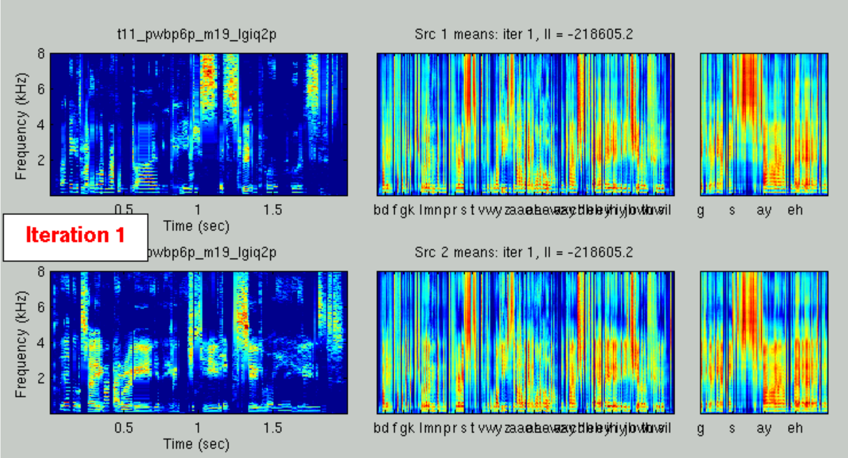


1. Introduction
2. Speaker subspace model
3. Monaural speech separation
   - Adaptation algorithm
   - Experiments
4. Binaural separation
5. Conclusions

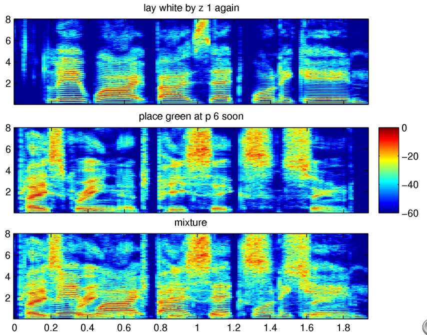
Adaptation algorithm



Adaptation example



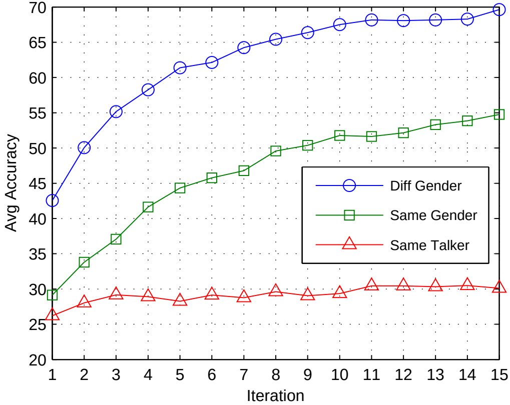
2006 Speech separation challenge [Cooke et al., 2010]



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

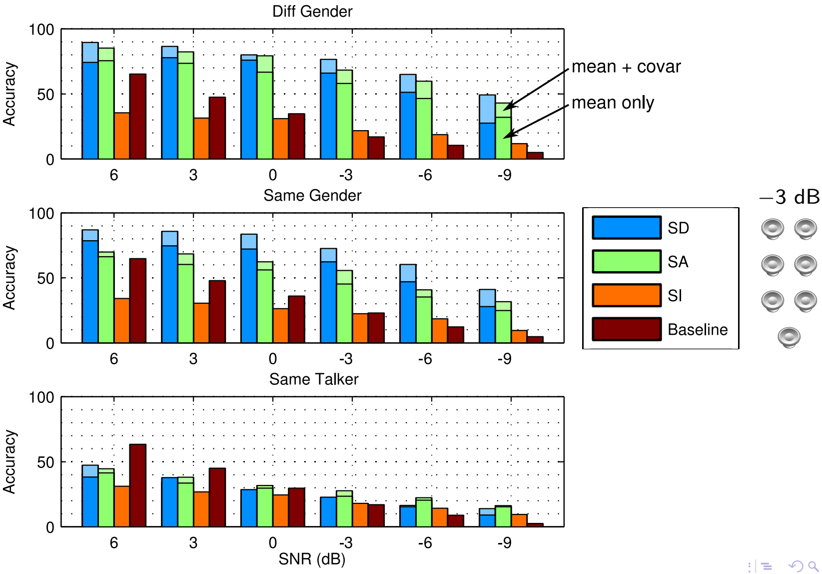
  `command(4)` `color(4)` `preposition(4)` `letter(25)` `digit(10)` `adverb(4)`
- Separation/recognition task
  - Determine letter and digit for source that said "white"

Adaptation performance



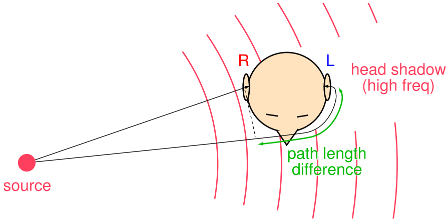
- Letter-digit accuracy averaged across all TMRs
- Adaptation clearly improves separation
- Same talker case hard – source permutations

Performance – Adapted vs. source-dependent models



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

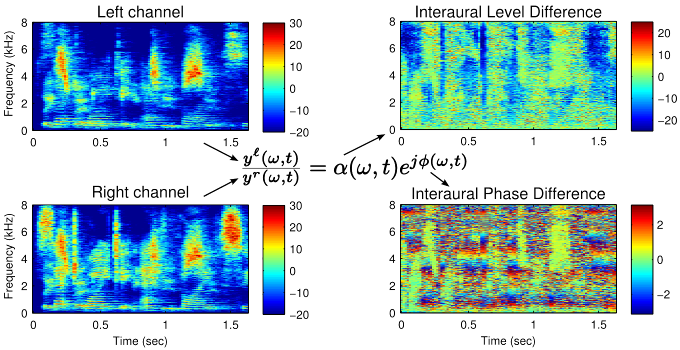


$$y^\ell(t) = \sum_i x_i(t - \tau_i^\ell) * h_i^\ell(t)$$

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

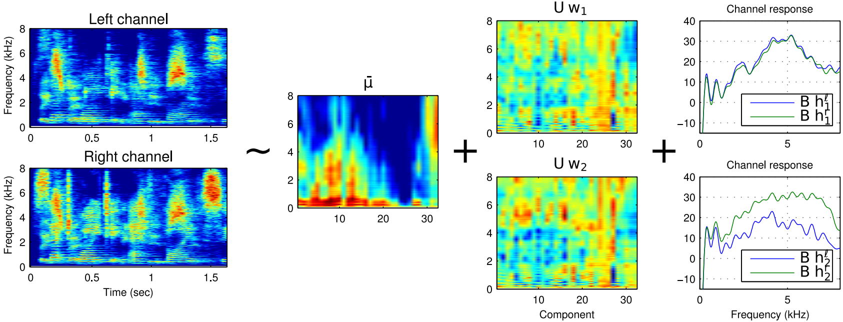
- Given stereo recording of multiple sound sources
- Utilize spatial cues to aid separation
  - Interaural time difference (ITD)
  - Interaural level difference (ILD)

MESSL: Interaural model [Mandel et al., 2009]



- 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

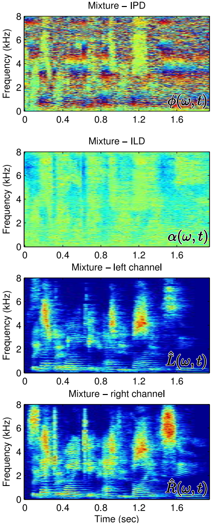
MESSL-SP: Source prior



- 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)

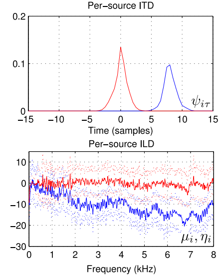
Parameter estimation and source separation

Observations

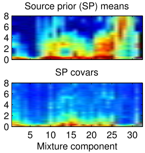


∼

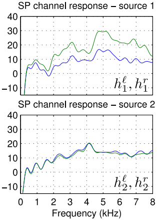
Parameters



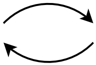
∼



+



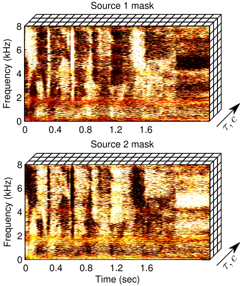
E-step
Use parameters to compute posteriors of hidden variables



M-step
Use posteriors to update parameters

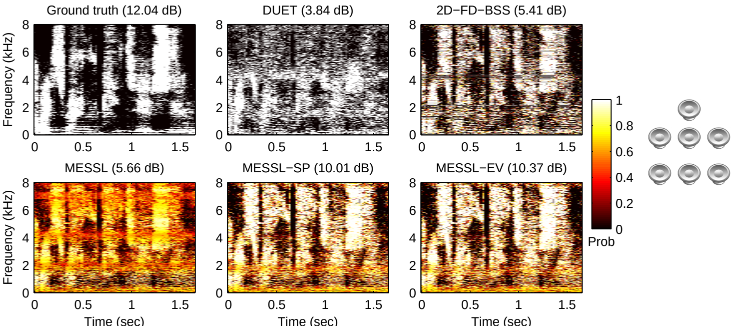
Posteriors

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



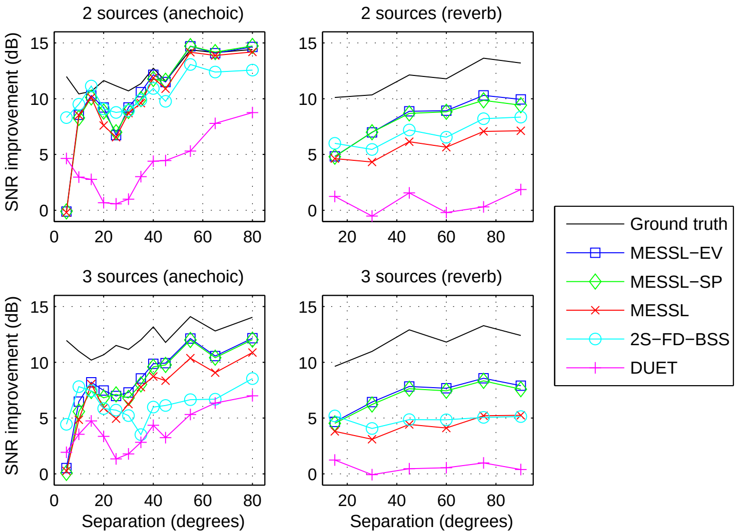
Separate sources by multiplying mixture by different masks

Experiments



- Mixtures of 2 and 3 speech sources, anechoic and reverberant
- 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]

Performance as function of distractor angle



Summary

- Prior signal models for underdetermined source separation
- Model-based source separation making minimal assumptions using subspace adaptation
- Monaural separation
    - Speaker-dependent $>$ speaker-adapted $\gg$ speaker-independent
- Binaural separation
    - Extend MESSL framework to use source models (joint with M. Mandel)
    - Substantial improvement using simple speaker-independent model

Applications to music

- Challenges
  - Very dense mixtures: 4+ sources mixed down to 2 channels
  - By design, sources are synchronized in time and frequency
- But sources contain a lot of structure
  - Very limited "vocabulary"
  - Significant repetition
- Recent work
  - Polyphonic music transcription using non-negative matrix factorization and eigeninstruments [Grindlay and Ellis, WASPAA 2009]

References

Cooke, M., Hershey, J. R., and Rennie, S. J. (2010).
Monaural speech separation and recognition challenge.
Computer Speech and Language, 24(1):1 – 15.

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., Weiss, R. J., and Ellis, D. P. W. (2009).
Model-based expectation maximization source separation and localization.
IEEE Transactions on Audio, Speech, and Language Processing.
in press.

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)*.

Weiss, R. J. (2009).
Underdetermined Source Separation Using Speaker Subspace Models.
PhD thesis, Department of Electrical Engineering, Columbia University.

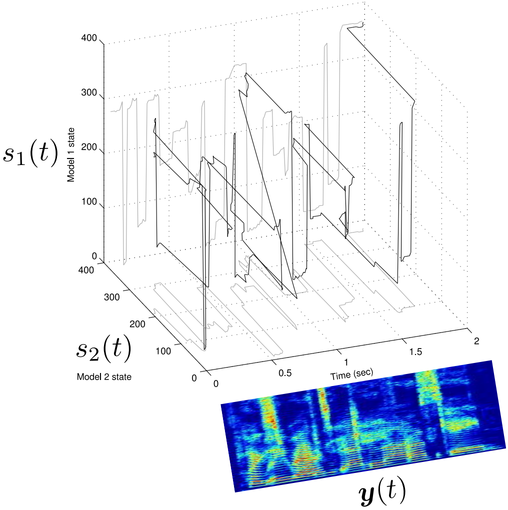
Weiss, R. J. and Ellis, D. P. W. (2010).
Speech separation using speaker-adapted eigenvoice speech models.
Computer Speech and Language, 24(1):16 – 29.

Weiss, R. J., Mandel, M. I., and Ellis, D. P. W. (2008).

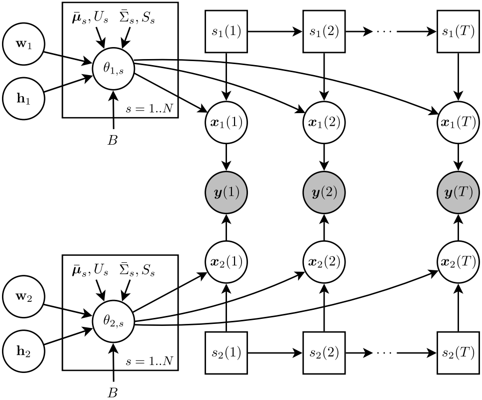
6 Extra slides

Factorial HMM separation

- Each source signal is characterized by state sequence through its HMM
- Viterbi algorithm to find maximum likelihood path through factorial HMM
- Reconstruct source signals using Viterbi path

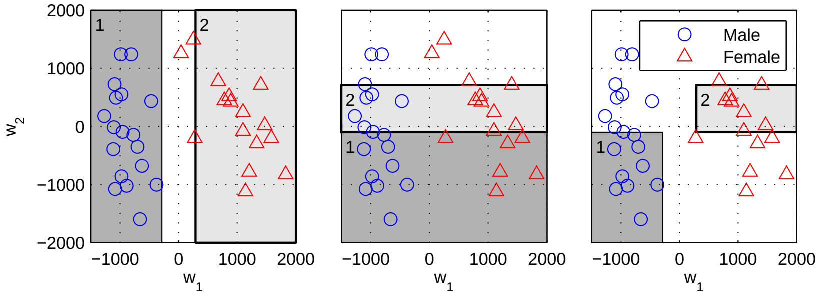


Eigenvoice factorial HMM



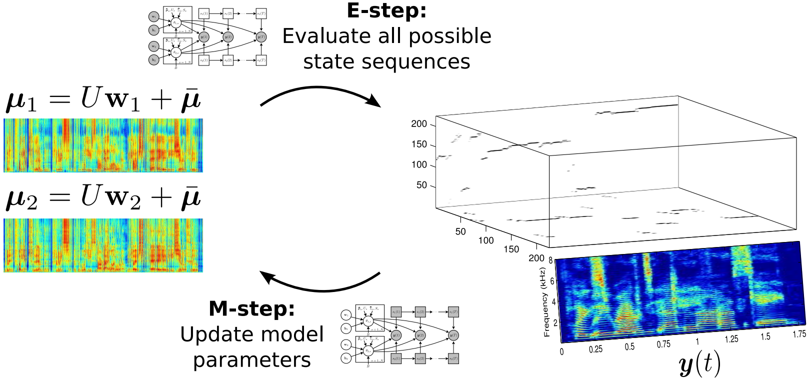
- Model mixture with combination of source HMMs
- Need adaptation parameters $\mathbf{w}_i$ to estimate source signals $x_i(t)$ and vice versa

Adaptation algorithm initialization



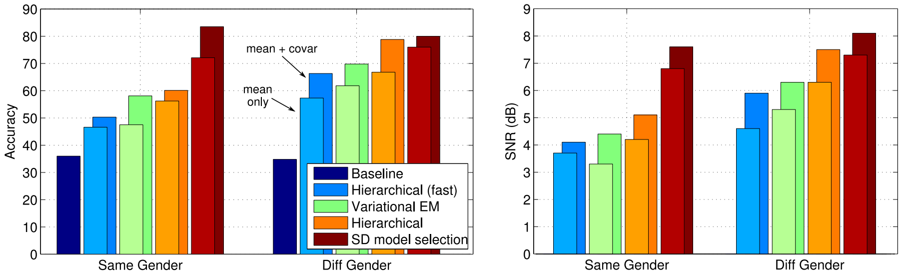
- Fast convergence needs good initialization
- Want to differentiate source models to get best initial separation
- Treat each eigenvoice dimension independently
    - Coarsely quantize weights
    - Find most likely combination in mixture

Variational learning



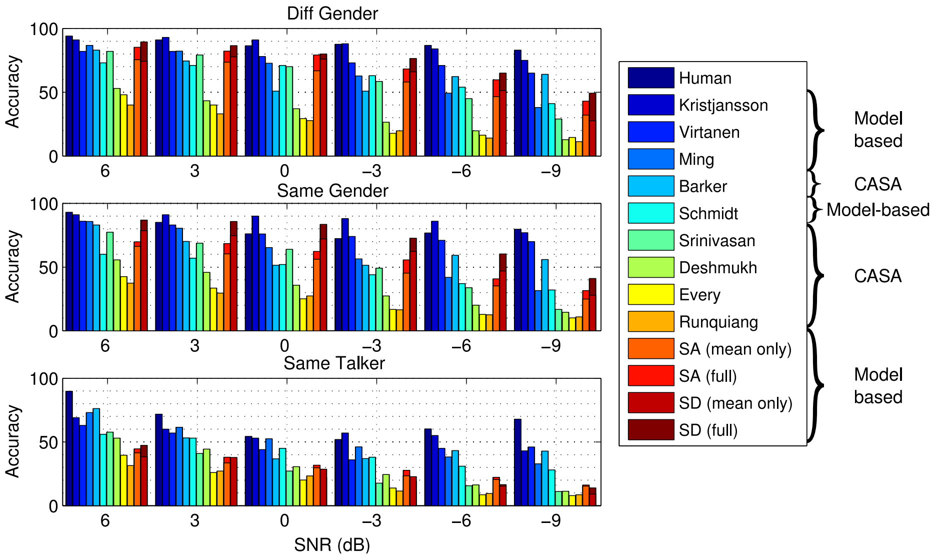
- Approximate EM algorithm to estimate adaptation parameters
- Treat each source HMM independently
- Introduce variational parameters to couple them

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

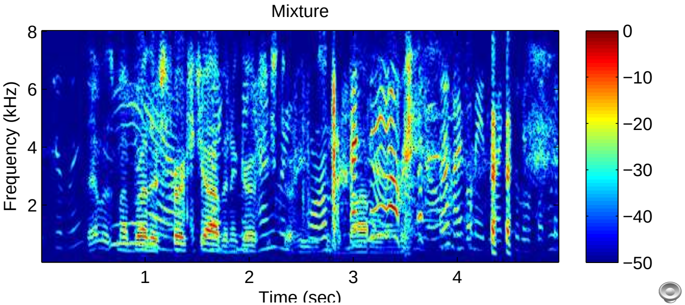
Performance – Comparison to other participants



Performance – Comparison to other participants

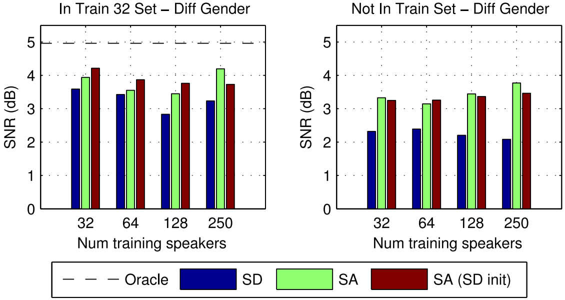
System	Description	ST	SG	DG
Human	N/A	66	81	88
Kristjansson	Source models, FHMM	56	87	87
Virtanen	Source models, FHMM	48	77	75
Ming	Source models	51	60	65
Barker	CASA, Speech fragment decoder	48	62	64
Schmidt	Source models, NMF	42	47	62
Srinivasan	CASA	28	52	61
Deshmukh	Phase Opponency	30	33	32
Every	Pitch tracking	19	23	28
Runquiang	CASA	19	22	24
SA (mean only)	Eigenvoice models, FHMM	27	48	59
SA (full)	Eigenvoice models, FHMM	30	55	70
SD (mean only)	Source models, FHMM	25	60	62
SD (full)	Source models, FHMM	26	72	74

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

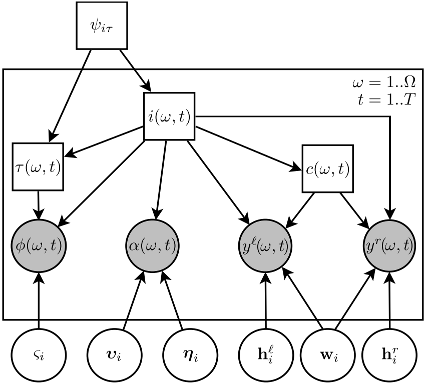
Switchboard – Results



- 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

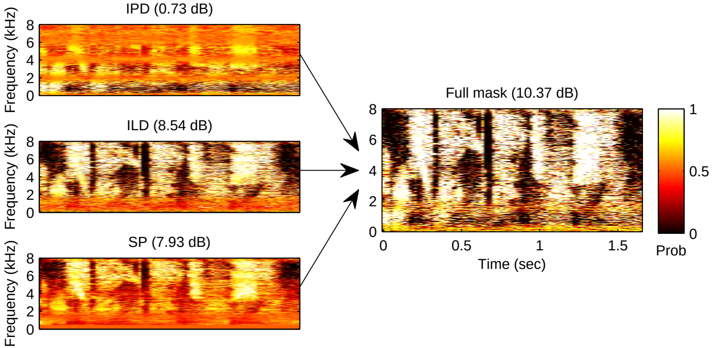
MESSL-EV: Putting it all together

- 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



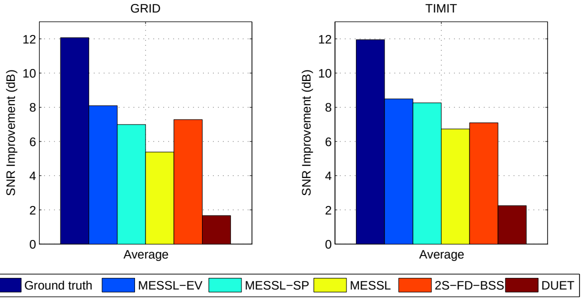
Explain each point in spectrogram by a particular source, time delay, and source model mixture component

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 due to spatial aliasing

Experiments – Matched vs. mismatched



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