Speech Separation for Recognition and Enhancement

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia University, NY
and
International Computer Science Institute, Berkeley CA

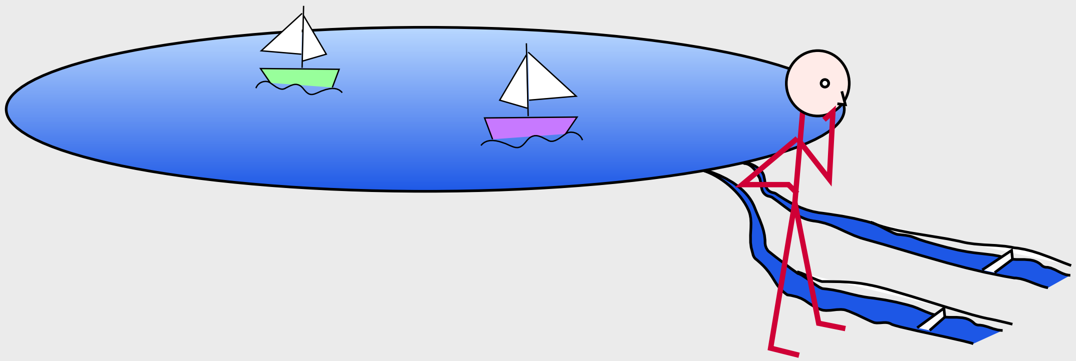
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Speech in the Wild
2. Separation by Space
3. Separation by Pitch
4. Separation by Model

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

ICSI

1. Speech in the Wild



- The world is cluttered
  sound is transparent
  - mixtures are inevitable

- Useful information is structured by 'sources'
  - specific definition of a 'source':
    intentional independence

Speech in the Wild: Examples

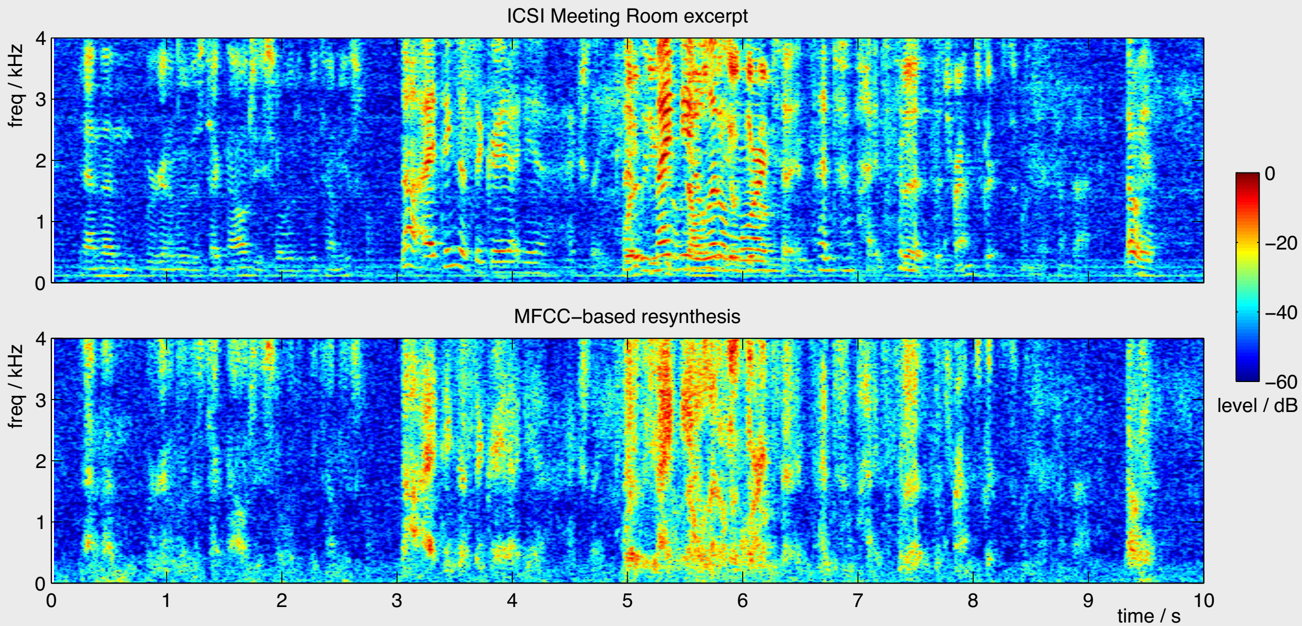
- Multi-party discussions

- Ambient recordings

- Applications:
  - communications
  - robots
  - lifelogging/archives

Recognizing Speech in the Wild

- Current ASR relies on low-D representations
  - e.g. 13 dimensional MFCC features every 10ms

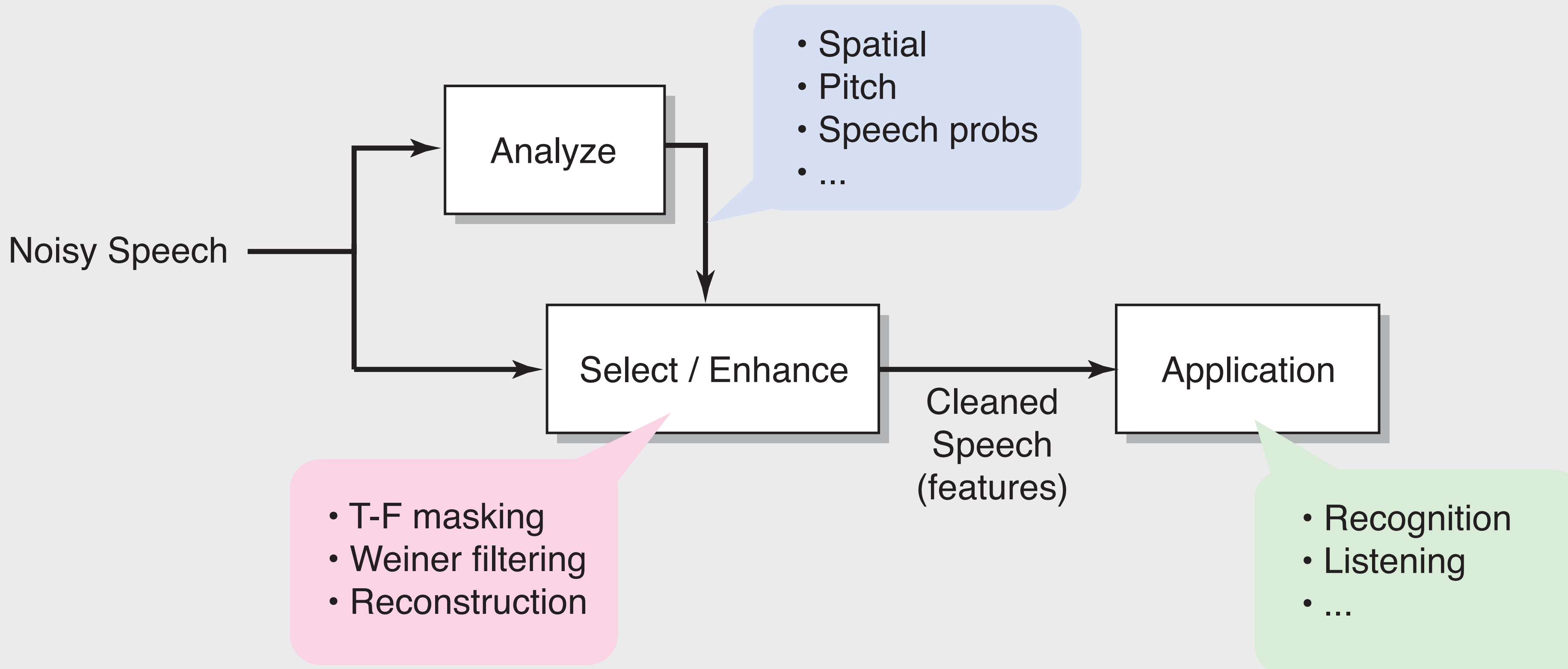


  - very successful for clean speech!
  - inadequate for mixtures

- We need separation!

2. Speech Separation

- How can we separate speech information?

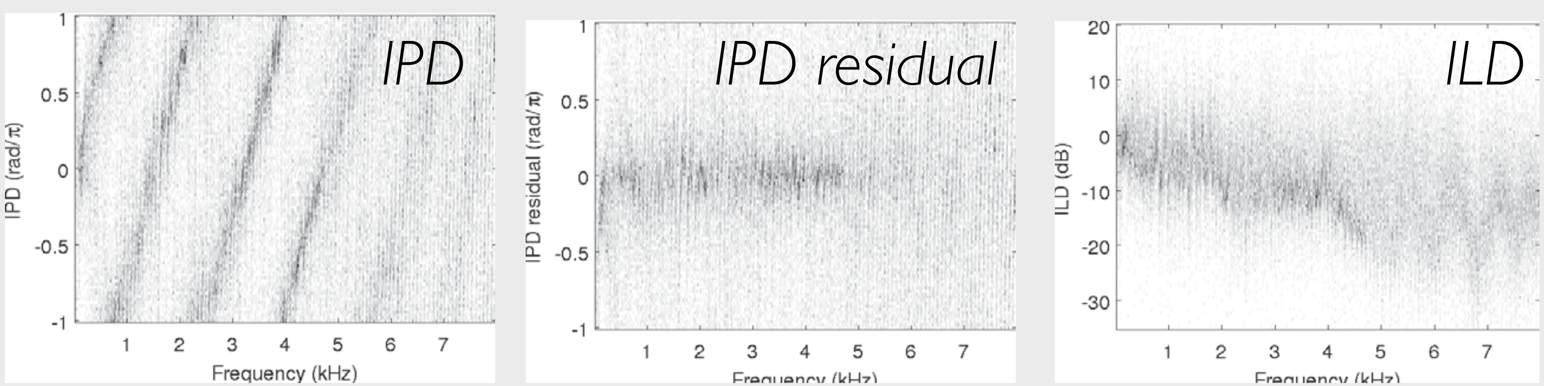


Separation by Spatial Info

- Given multiple microphones,
  sound carries spatial information about source
- E.g. model interaural spectrum of each source
  as stationary level and time differences:

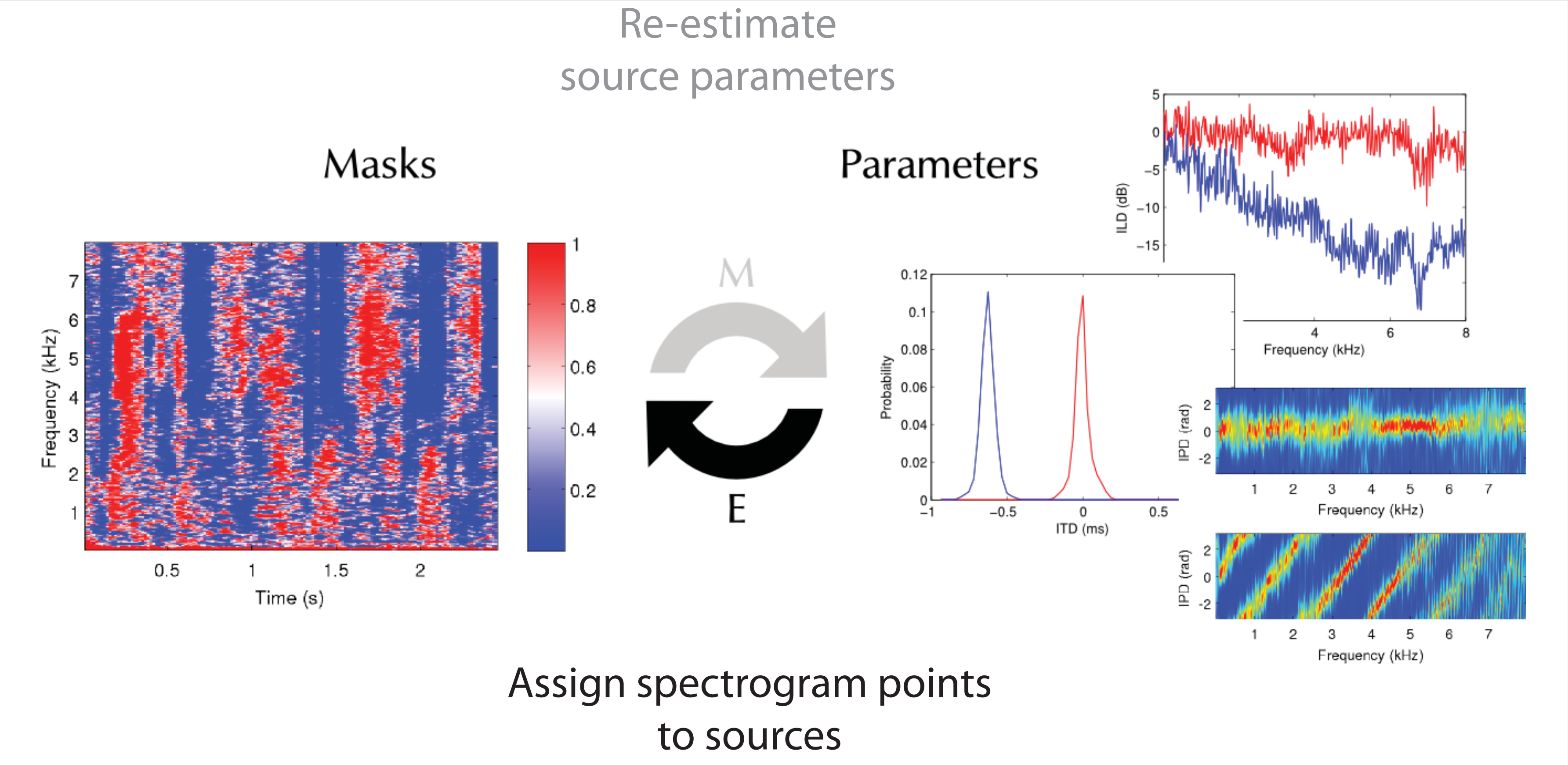
$$\frac{L(\omega,t)}{R(\omega,t)} = a(\omega)e^{j\omega\tau}N(\omega,t)$$

- e.g. at 75°, in reverb:



Model-Based EM Source Separation and Localization (MESSL)

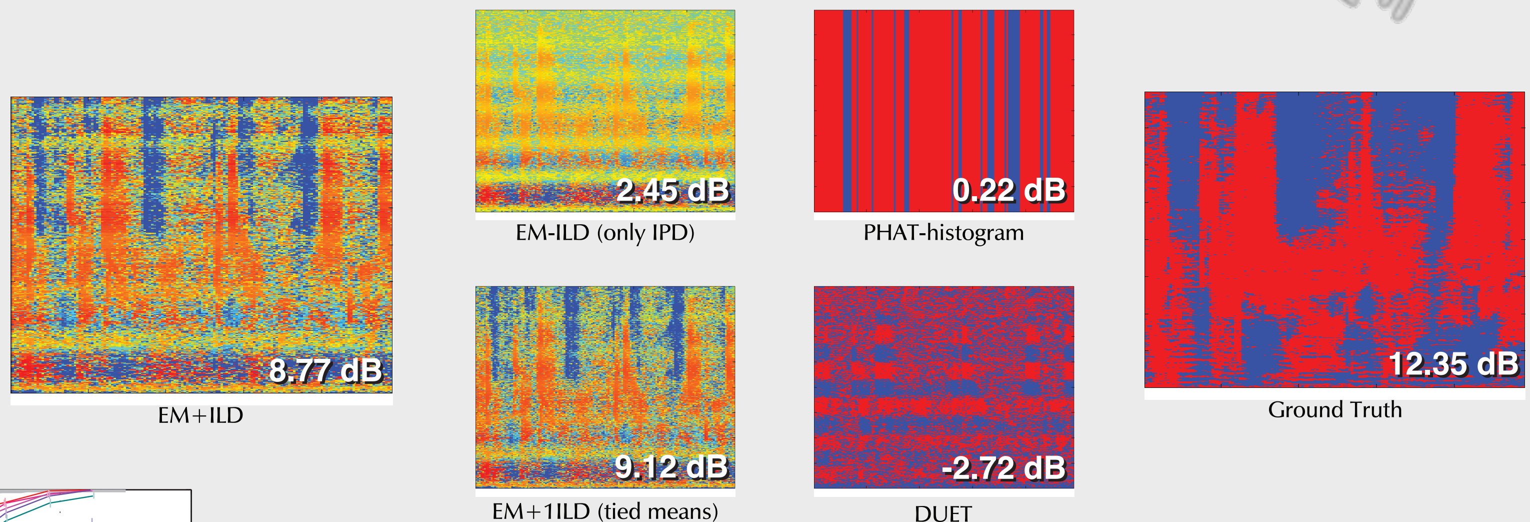
Mandel et al. '10

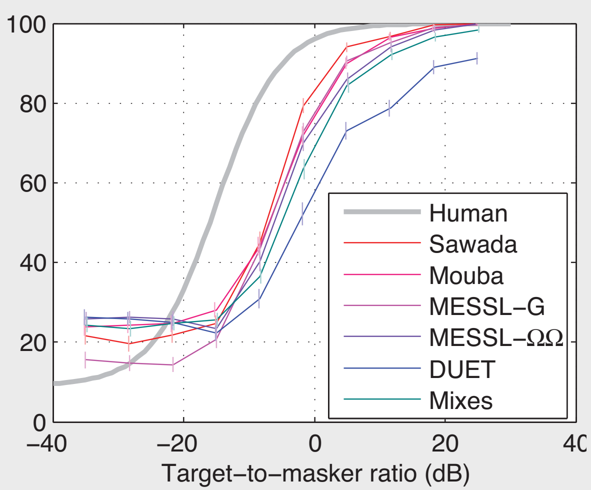


- can model more sources than sensors

MESSL Results

- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness

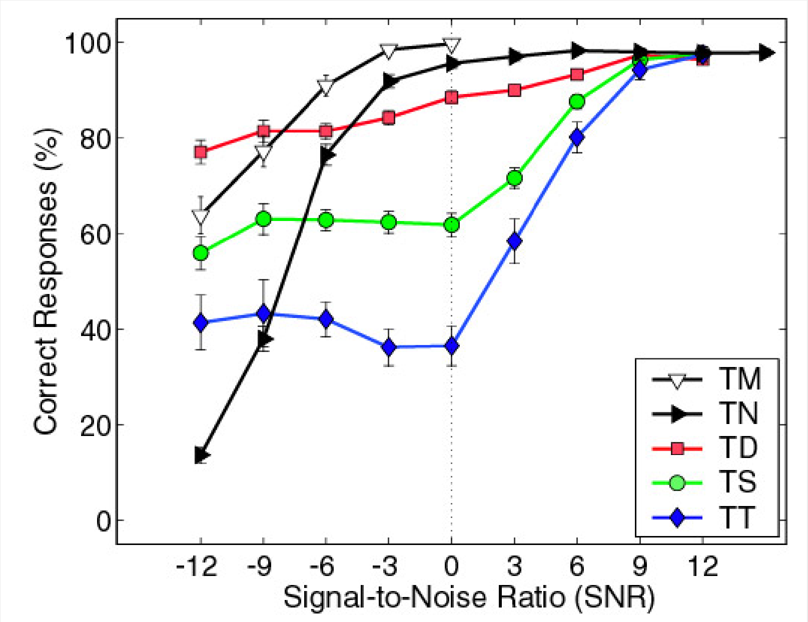




- Helps with recognition
  - digits accuracy

3. Separation by Pitch

- Voiced syllables have near-periodic "pitch"
  - perceptually salient
  - lost in MFCCs



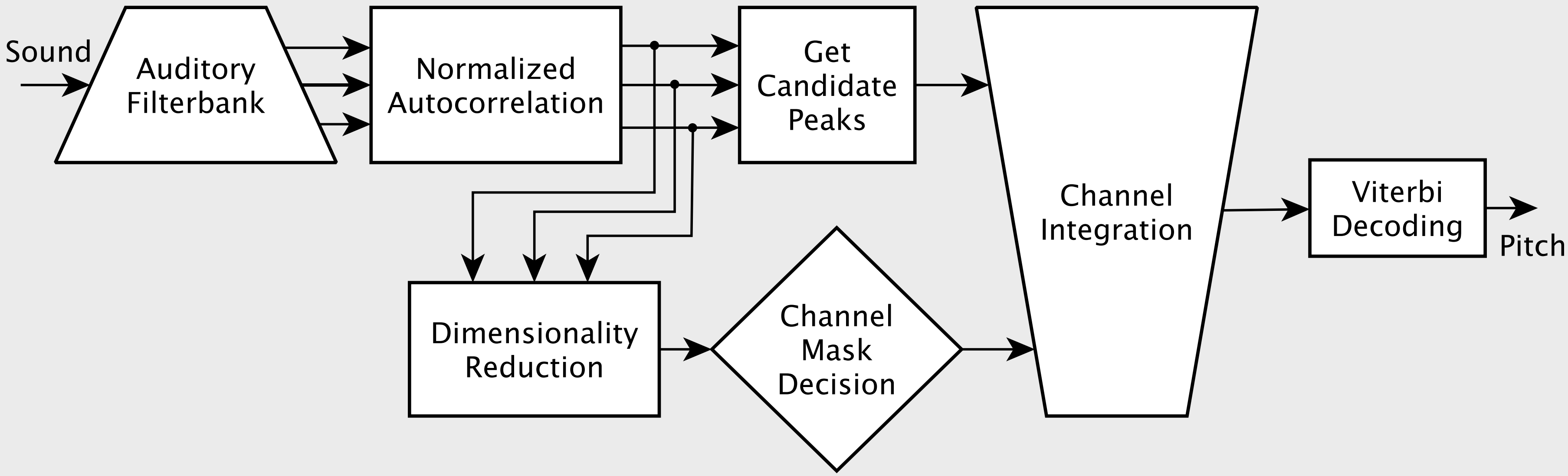
Brungart et al.'01

- Can we track pitch & use it for separation?
  - ... and other speech tasks?

Noise-Robust Pitch Tracking

BS Lee & Ellis '12

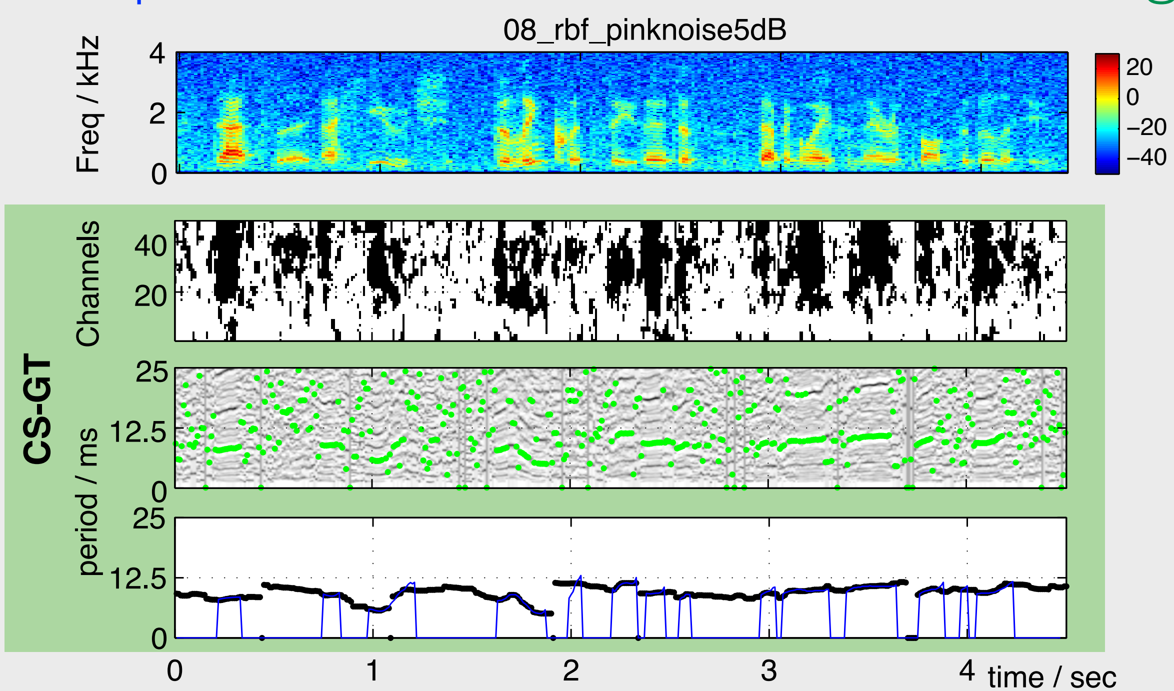
- Important for voice detection & separation
- Based on channel selection *Wu, Wang & Brown '03*
  - pitch from summary autocorrelation over "good" bands



  - trained classifier decides which channels to include

Noise-Robust Pitch Tracking

- Channel-based classifiers
  learn domain channel/noise characteristics
  - then separate, or derive features for recognition

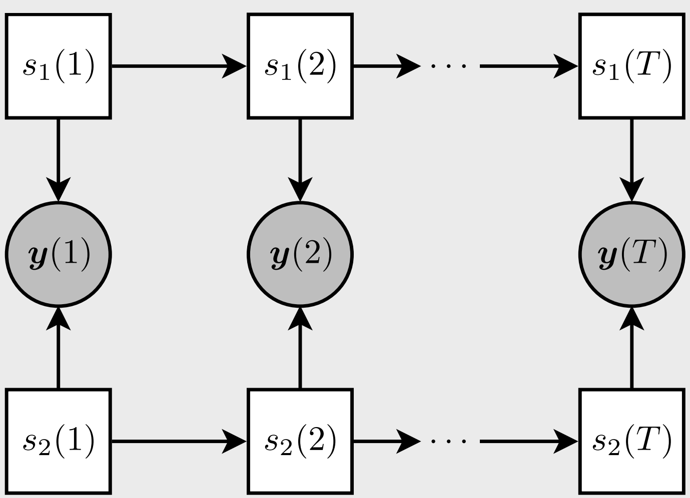


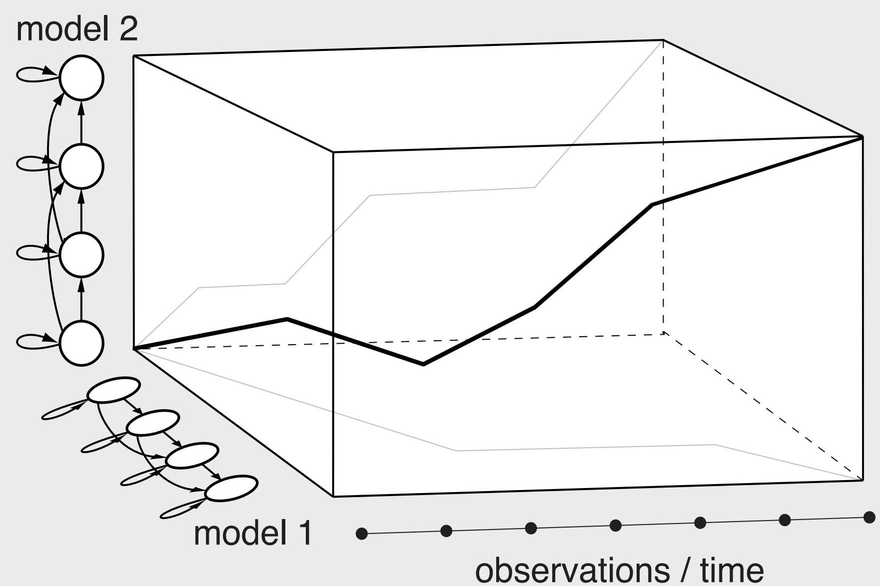
- Only works for pitched sounds
  - need a broader description of the speech source...

4. Separation by Models

Varga & Moore, '90
Hershey et al., '10

- If ASR is finding best-fit parameters
  argmax $P(W \mid X)$ ...
- Recognize mixtures with Factorial HMM
  - model + state sequence for each voice/source
  - exploit sequence constraints, speaker differences



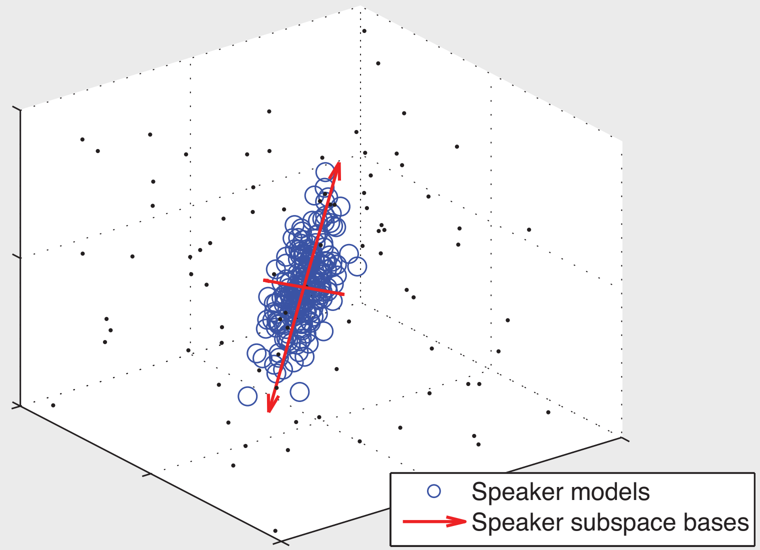


  - separation relies on detailed speaker model

Eigenvoices

Kuhn et al. '98, '00
Weiss & Ellis '10

- Idea: Find speaker model parameter space



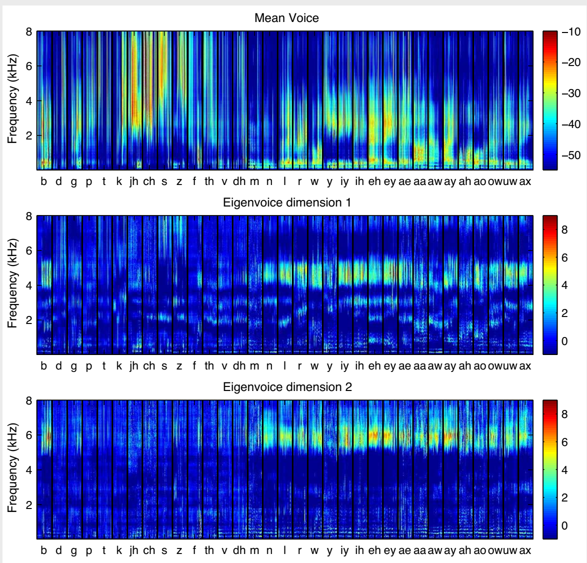
  - generalize without losing detail?

- Eigenvoice model:

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

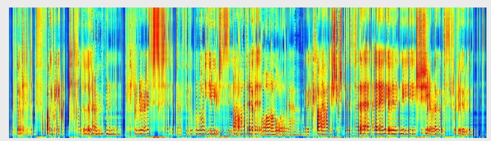
$\boldsymbol{\mu}$	$\bar{\boldsymbol{\mu}}$	U	$\mathbf{w}$	B	$\mathbf{h}$
adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

  - 89,600 dimensional space

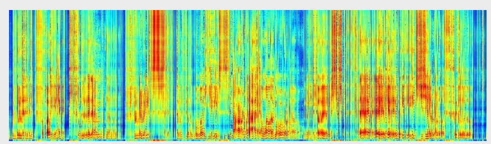


Eigenvoice Speech Separation

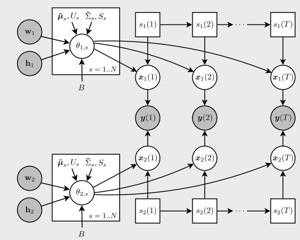
$$\boldsymbol{\mu}_1 = U\mathbf{w}_1 + \bar{\boldsymbol{\mu}}$$

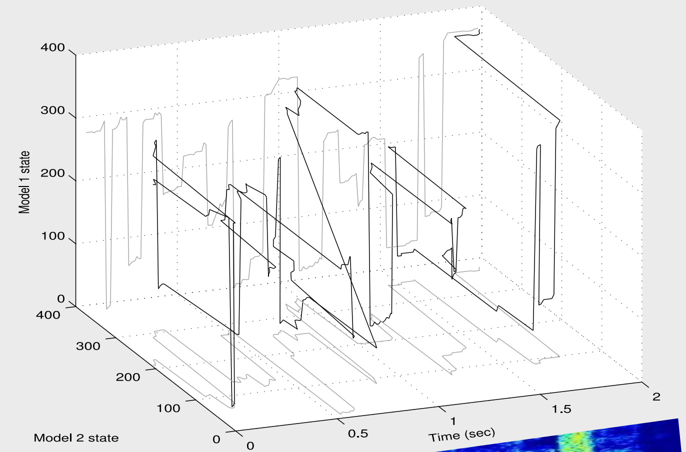


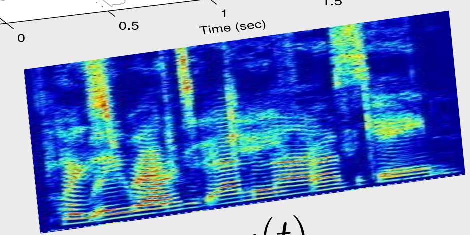
$$\boldsymbol{\mu}_2 = U\mathbf{w}_2 + \bar{\boldsymbol{\mu}}$$



Find Viterbi path

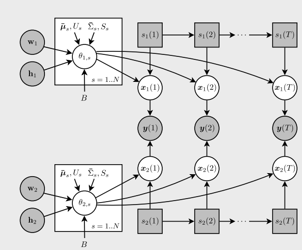






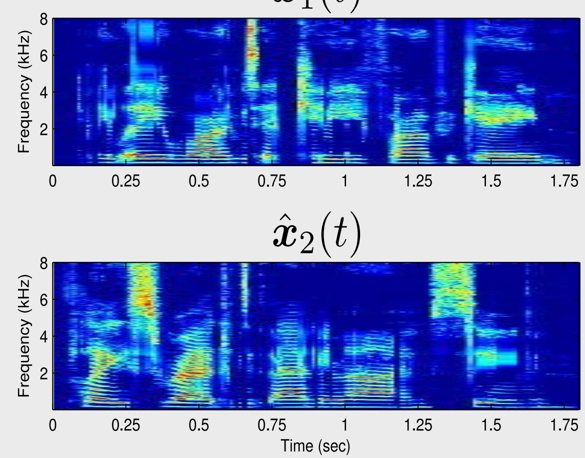
$\boldsymbol{y}(t)$

Estimate source signals

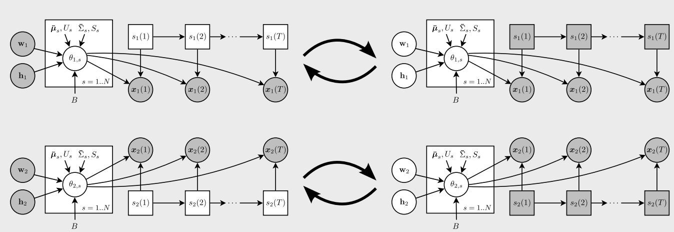


$\hat{\boldsymbol{x}}_1(t)$

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

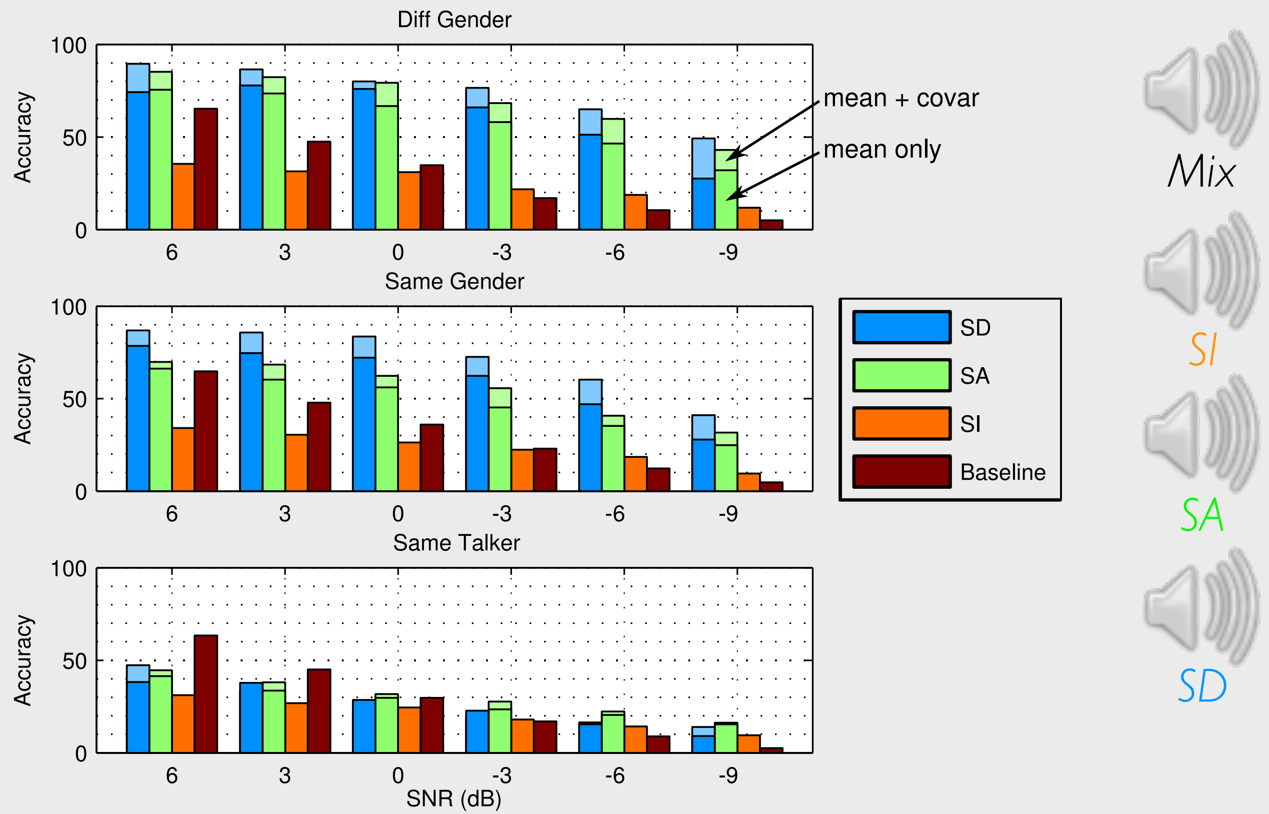


Update model parameters using EM algorithm from Kuhn et al., (2000)



Eigenvoice Speech Separation

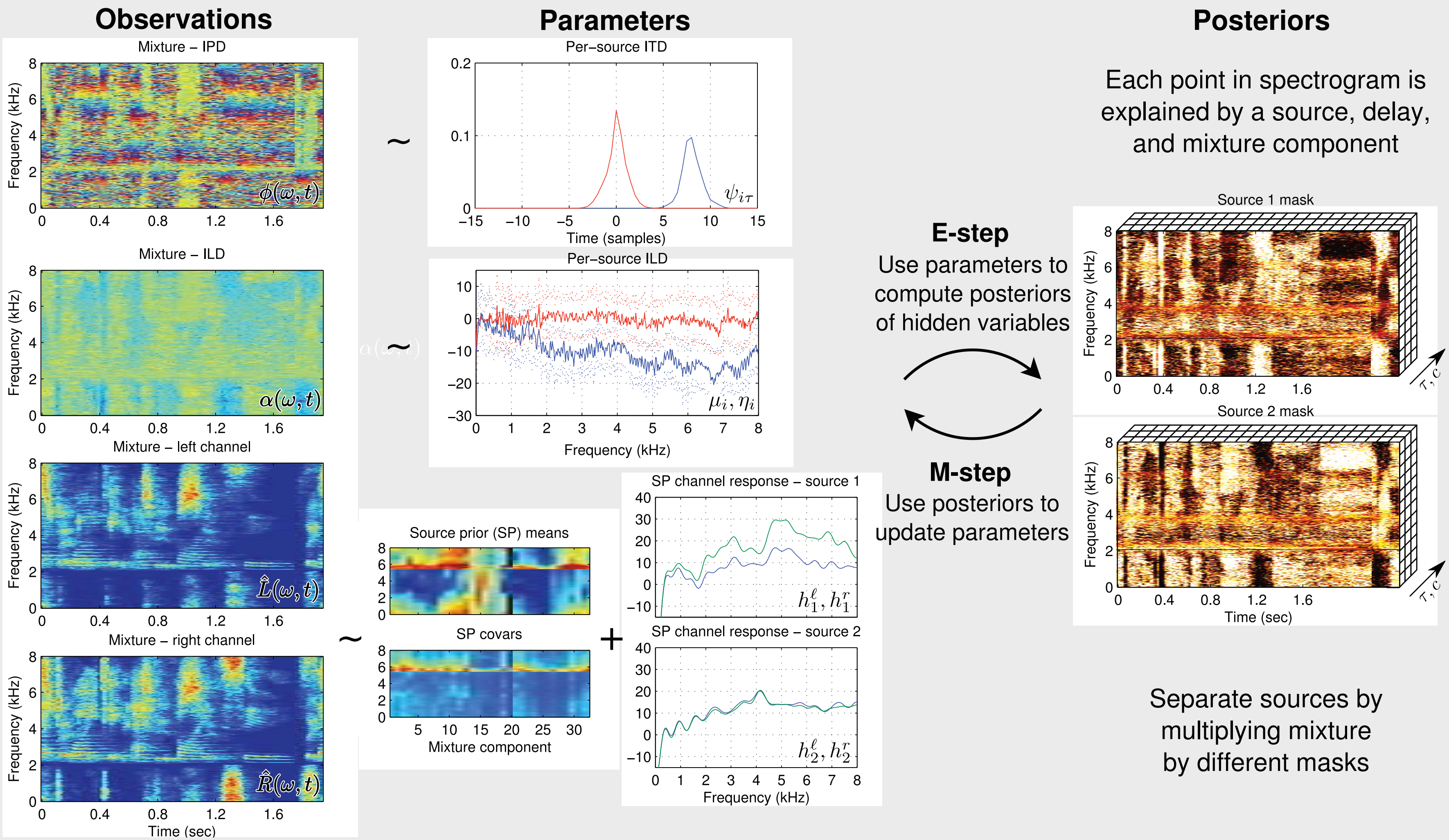
- Eigenvoices for Speech Separation task
  - speaker adapted (SA) performs midway between speaker-dependent (SD) & speaker-indep (SI)



Spatial + Model Separation

Weiss, Mandel & Ellis '11

- MESSL + Eigenvoice "priors"



Summary

- Speech in the Wild
  ... real, challenging problem
  ... applications in communications, lifelogs ...

- Speech Separation
  ... by generic properties (location, pitch)
  ... via statistical models

- Recognition and Enhancement
  ... separate-then-X, or integrated solution?

References

- John Hershey, Steve Rennie, Pedr Olsen, Trausti Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech & Lang.* 24 (1), 45-66, 2010.
- Jon Barker, Martin Cooke, Dan Ellis, "Decoding Speech in the Presence of Other Sources," *Speech Communication* 45(1): 5-25, 2005.
- R. Kuhn, J. Junqua, P. Nguyen, N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," . *IEEE Tr. Speech & Audio Proc.* 8(6): 695–707, Nov 2000.
- Byung-Suk Lee & Dan Ellis, "Noise-robust pitch tracking by trained channel selection," submitted to *ICASSP*, 2012.
- Michael Mandel, Ron Weiss, Dan Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Tr. Audio, Speech, Lang. Proc.* 18(2): 382-394, Feb 2010.
- A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," *ICASSP-90*, 845–848, 1990.
- Ron Weiss & Dan Ellis, "Speech separation using speaker-adapted Eigenvoice speech models," *Computer Speech & Lang*. 24(1): 16-29, 2010.
- Ron Weiss, Michael Mandel, Dan Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Communication* 53(5): 606-621, May 2011.
- Mingyang Wu, DeLiang Wang, Guy Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Tr. Speech & Audio Proc.* 11(3): 229–241, May 2003.