Using Source Models in Speech Separation

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

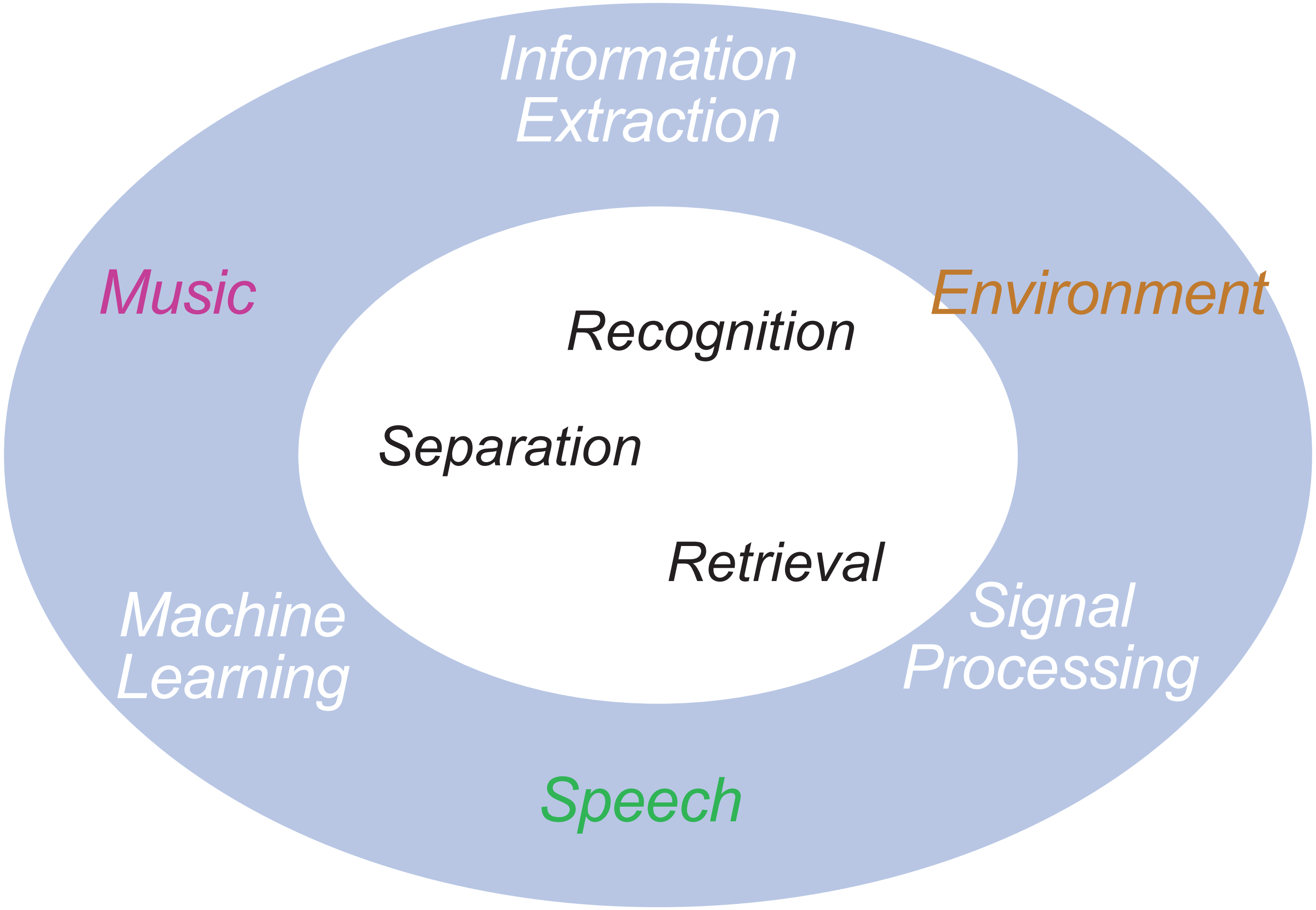
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Mixtures, Separation, and Models
2. Monaural Speech Separation
3. Binaural Speech Separation
4. Conclusions

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

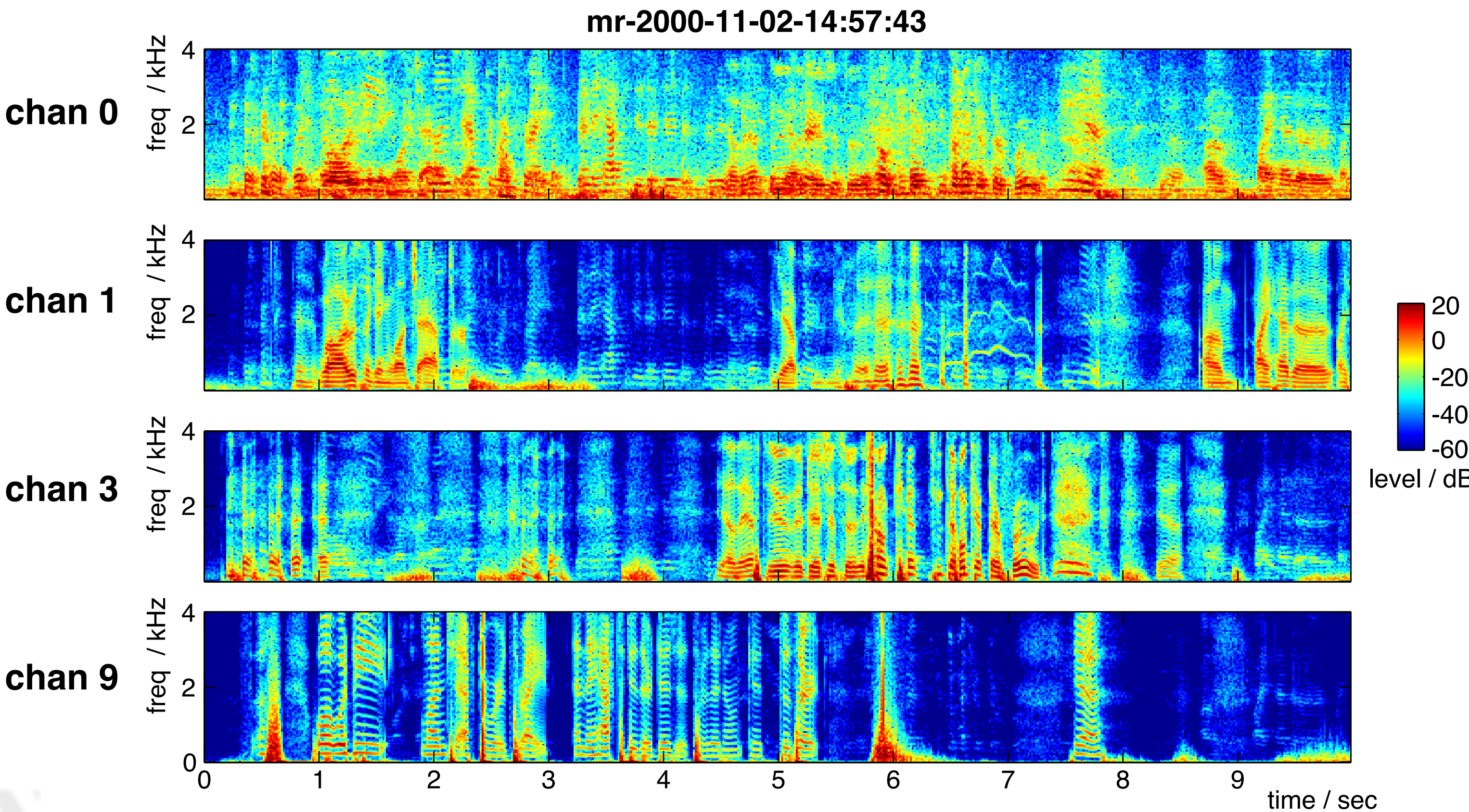
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

LabROSA Overview



1. Mixtures, Separation, and Models

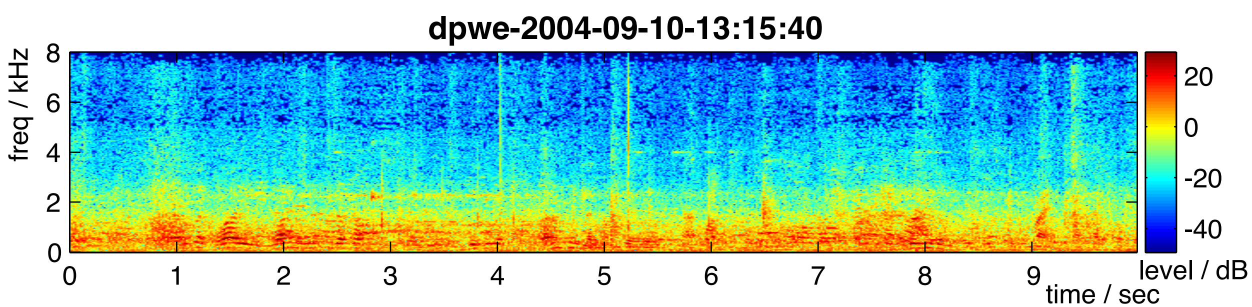
- Sounds rarely occur in isolation
  - .. so analyzing mixtures is a problem
  - .. for humans and machines



mr-20001102-1440-cE+1743.wav

Mixture Organization Scenarios

- Interactive voice systems
  - human-level understanding is expected
- Speech prostheses
  - crowds: #1 complaint of hearing aid users
- Archive analysis
  - identifying and isolating sound events



- Unmixing/remixing/enhancement...

Separation vs. Inference

- Ideal separation is rarely possible
  - many situations where overlaps cannot be removed
- Overlaps → Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

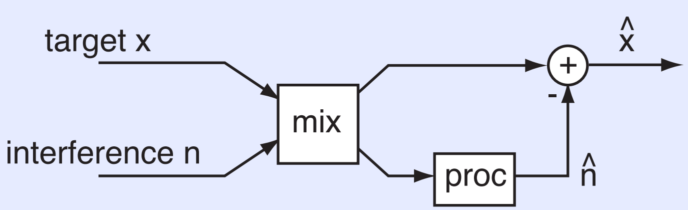
$$P(\{S_i\}|\,X) \propto \underbrace{P(X\,|\{S_i\})}_{\text{combination physics}}\;\underbrace{P(\{S_i\})}_{\text{source models}}$$

  - search over $\{S_i\}$??
- Better source models → better inference
  - .. learn from examples?

Approaches to Separation

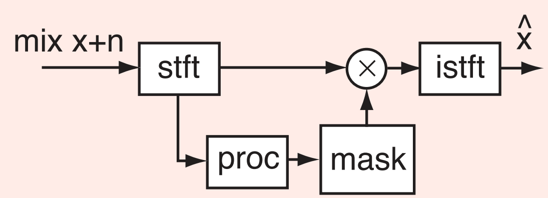
ICA

- Multi-channel
- Fixed filtering
- Perfect separation
  – maybe!



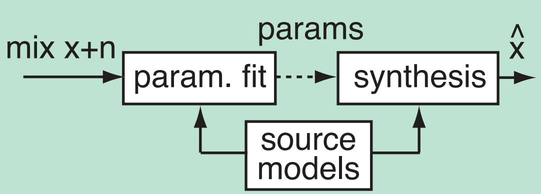
CASA

- Single-channel
- Time-var. filter
- Approximate separation



Model-based

- Any domain
- Param. search
- Synthetic output



- or combinations ...

EM for Model-based Separation

- Expectation-Maximization algorithm – for solving partially-unknown problems
  - (only local optimality guaranteed)
- EM for model-based separation
  - E-step: find distribution of unknowns $p(u)$ given current model parameters Θ and observations x
  - M-step: optimize Θ to maximize fit to x given current $p(u)$

E-step

$$p(u|\Theta^{(n)}) = p(x,u|\Theta^{(n)})/p(x|\Theta^{(n)})$$

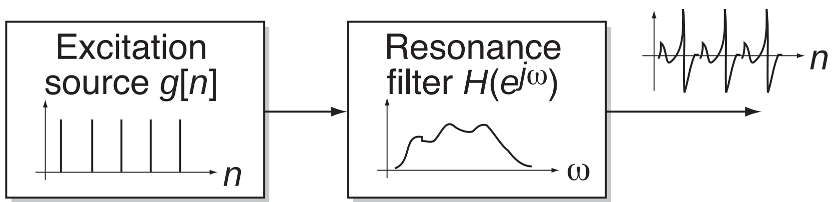
M-step

$$\Theta^{(n+1)} = \operatorname*{argmax}_{\Theta} \; E_{p(u|\Theta^{(n)})} p(x,u|\Theta)$$

u is... GMM mixture assignment
...T-F cell dominance
... current phone of voice i
...

What is a Source Model?

- Source Model describes signal behavior
  - encapsulates constraints on form of signal
  - (any such constraint can be seen as a model...)
- A model has parameters
  - model + parameters
    → instance



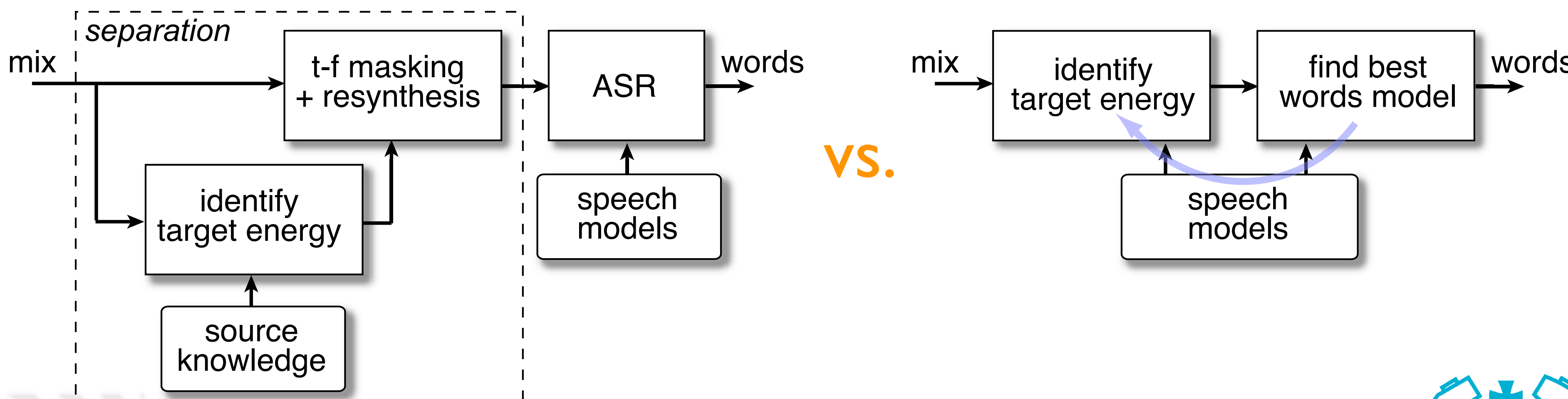
- What is *not* a source model?
  - detail not provided in instance
    e.g. using phase from original mixture
  - constraints on interaction between sources
    e.g. independence, clustering attributes

2. Monaural Speech Separation

- Cooke & Lee's Speech Separation Challenge
  - short, grammatically-constrained utterances:

    <command:4><color:4><preposition:4><letter:25><number:10><adverb:4>

    e.g. "bin white by R 8 again"
  - task: report letter + number for "white"
  - special session at Interspeech '06
- Separation or Description?



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

Columbia University in the City of New York

Codebook Models

Roweis '01, '03
Kristjannson '04, '06

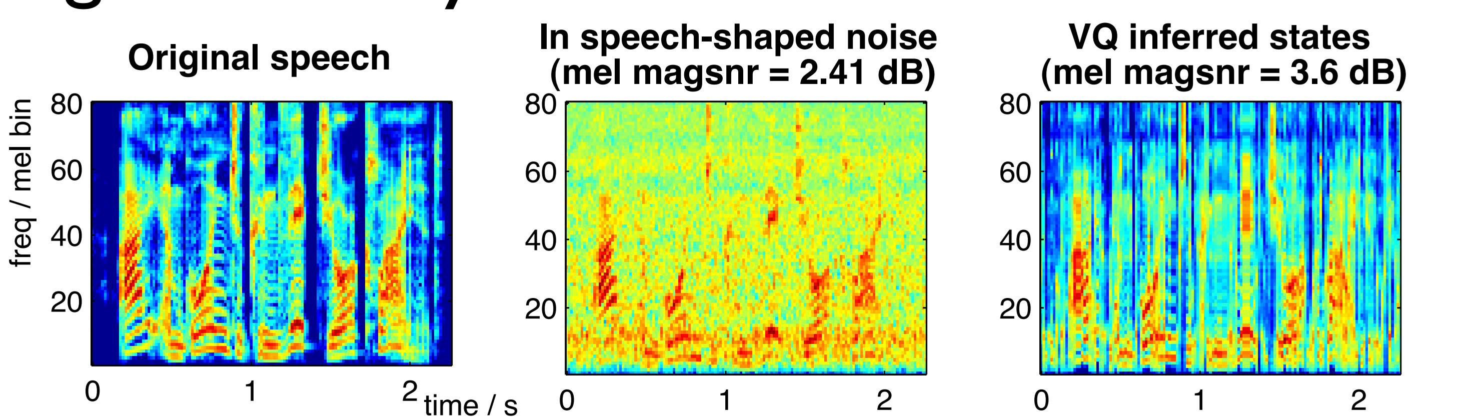
- Given models for sources, find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1, i_2) = \mathcal{N}(\mathbf{x}; \mathbf{c}_{i1} + \mathbf{c}_{i2}, \Sigma)$$ *combination model*

$$\{i_1(t), i_2(t)\} = argmax_{i_1,i_2} p(\mathbf{x}(t)|i_1, i_2)$$ *inference of source state*

  - can include sequential constraints...
  - different domains for combining $\mathbf{c}$ and defining Σ

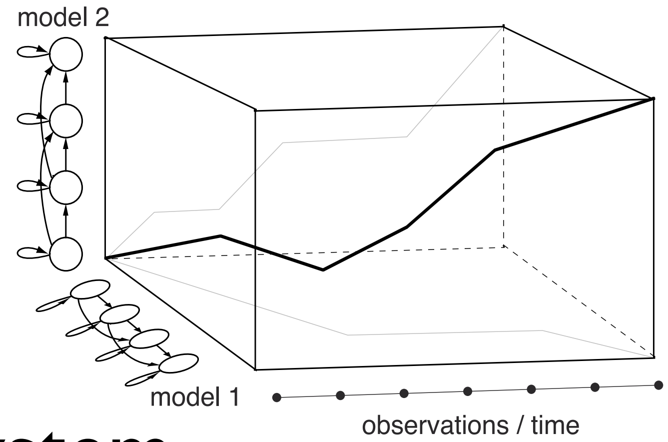
- E.g. stationary noise:



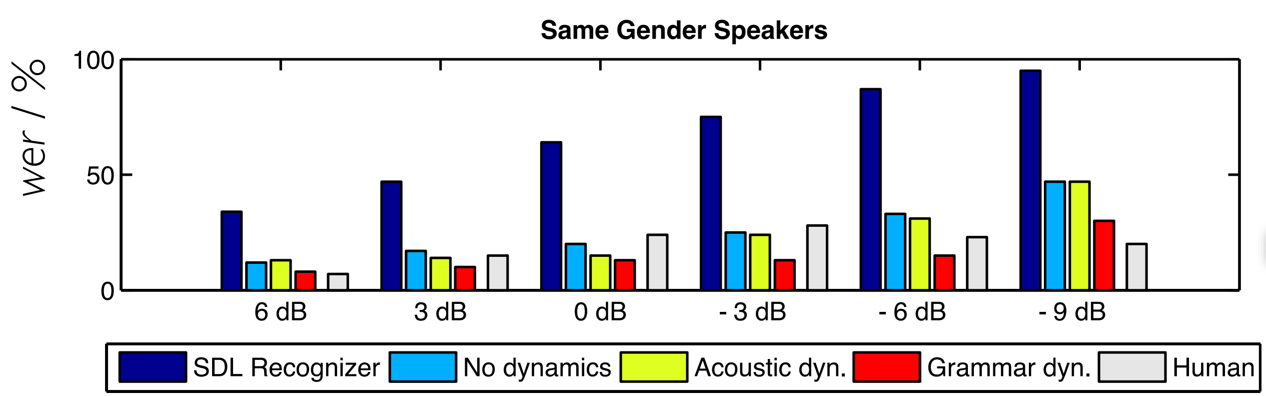
Speech Recognition Models

Kristjansson, Hershey et al. '06

- **Decode with Factorial HMM**
  - i.e. two state sequences, one model for each voice
  - exploit sequence constraints, speaker differences?



- **IBM "superhuman" Iroquois system**
  - fewer errors than people for same speaker, level
  - exploit grammar constraints - higher-level dynamics



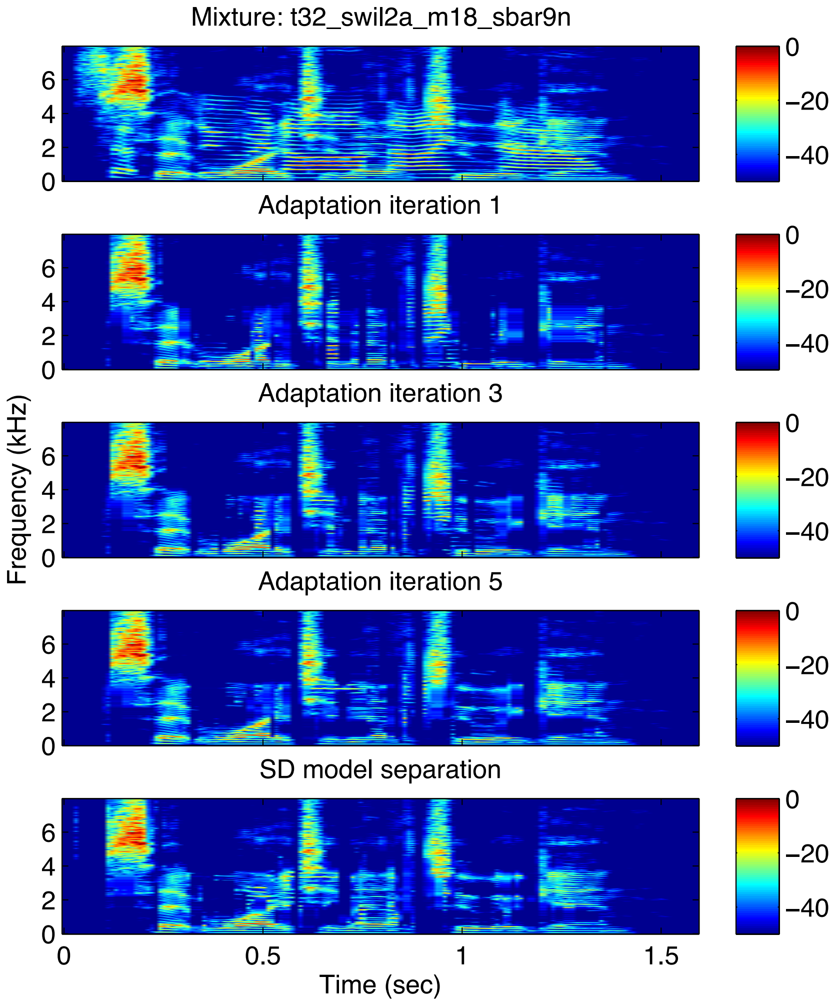
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

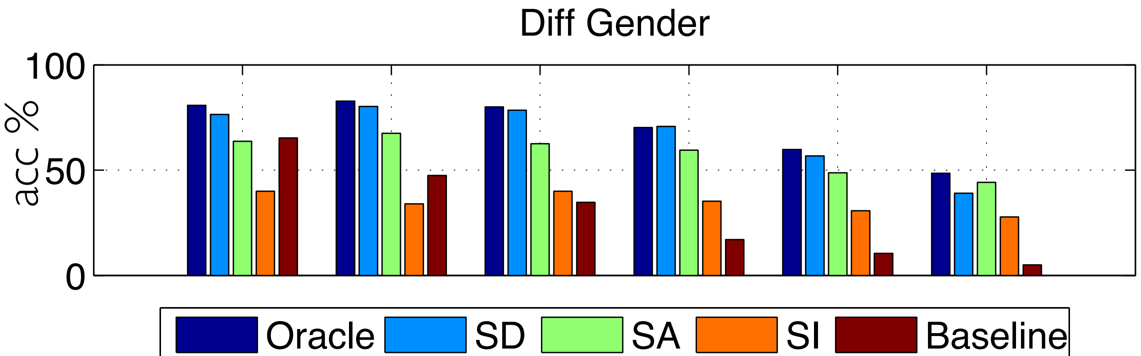
Speaker-Adapted (SA) Models

Weiss & Ellis '07

- Factorial HMM needs distinct speakers

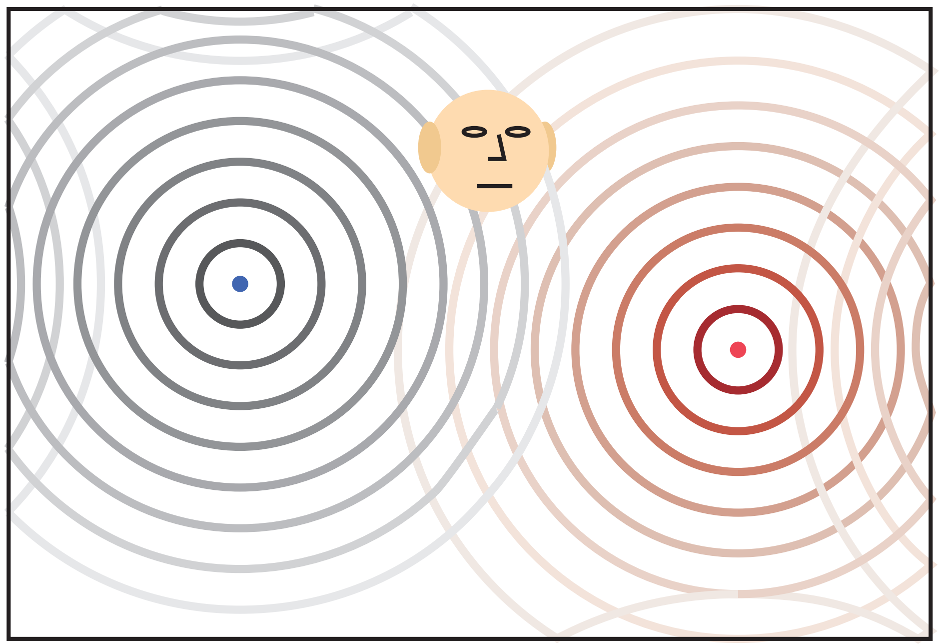


- use "eigenvoice" speaker space
- iterate estimating voice & separating speech
- performs midway between speaker-independent (SI) and speaker-dependent (SD)



3. Binaural Speech Separation

- 2 or 3 sources in reverberation
  - assume just 2 'ears'



- Tasks:
  - identify positions of sources (and number?)
  - recover source signals

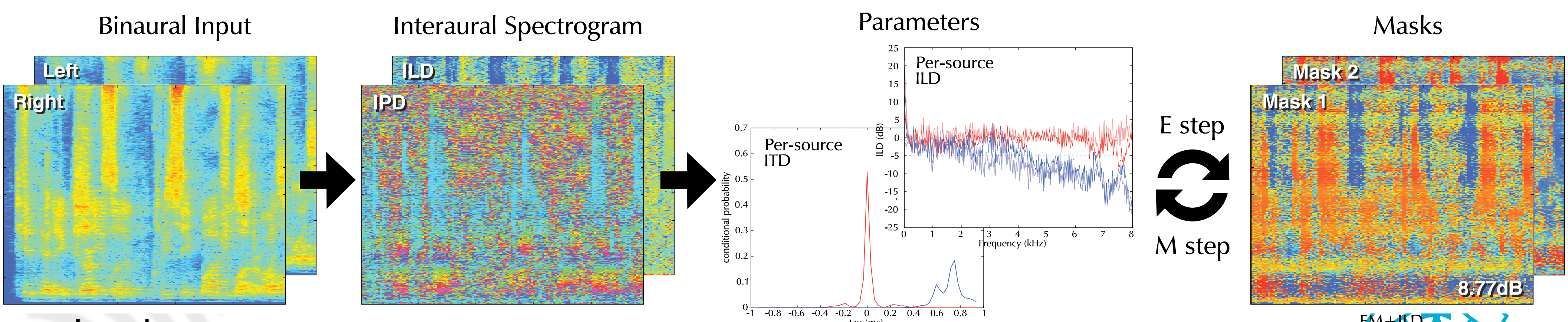
Spatial Estimation in Reverb

Mandel & Ellis '07

- Model interaural spectrum of each source as stationary level and time differences:

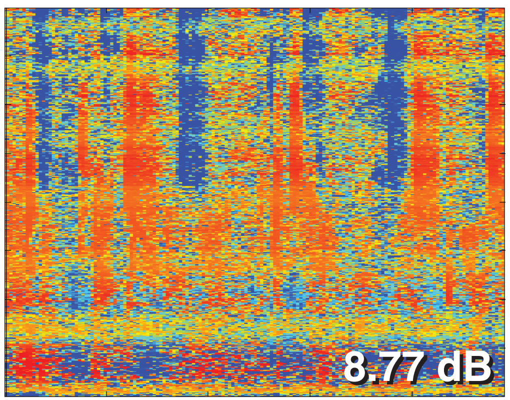
$$\frac{L(\omega,t)}{R(\omega,t)} = a(\omega)e^{j\omega\tau}N(\omega,t)$$

  - converge via EM to $a()$, τ for each source
  - mask is $\Pr(X(t,\omega)$ dominated by source $i)$

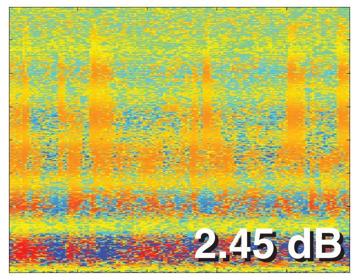


Spatial Estimation Results

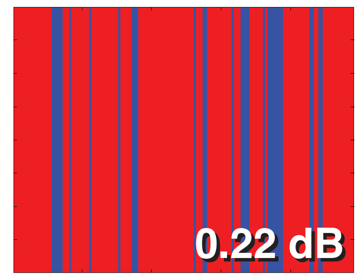
- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness



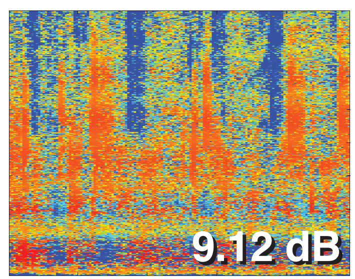
EM+ILD



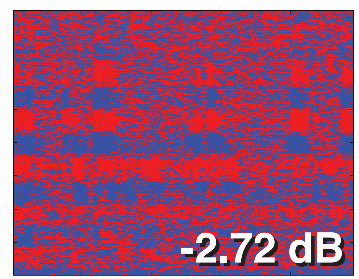
EM-ILD (only IPD)



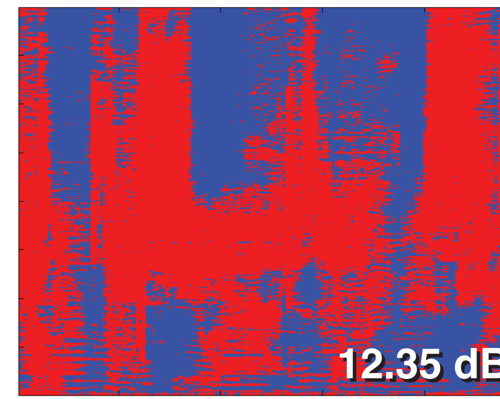
PHAT-histogram



EM+1ILD (tied means)



DUET



Ground Truth

Combining Spatial + Speech Model

- Interaural parameters give

$$ILD_i(\omega),\ \ ITD_i,\ \ \Pr\big(X(t,\omega) = S_i(t,\omega)\big)$$

- Speech source model can give

$$\Pr\big(S_i(t,\omega) \text{ is speech signal}\big)$$

- Can combine into one big EM framework...

E-step

$$p(u|\Theta^{(n)}) = p(x,u|\Theta^{(n)})/p(x|\Theta^{(n)})$$

M-step

$$\Theta^{(n+1)} = \underset{\Theta}{\operatorname{argmax}}\ E_{p(u|\Theta^{(n)})}p(x,u|\Theta)$$

u is: Pr(cell from source i)
phoneme sequence

Θ is: interaural params
speaker params

Summary & Conclusions

- Inferring model parameters is very general
  - .. and very difficult, in general
- Speech models can separate single channels
  - .. better match to individual → better results
- Spatial cues can separate binaural signals
  - .. but account for uncertainty from e.g. reverb
- EM-type approach can integrate them both