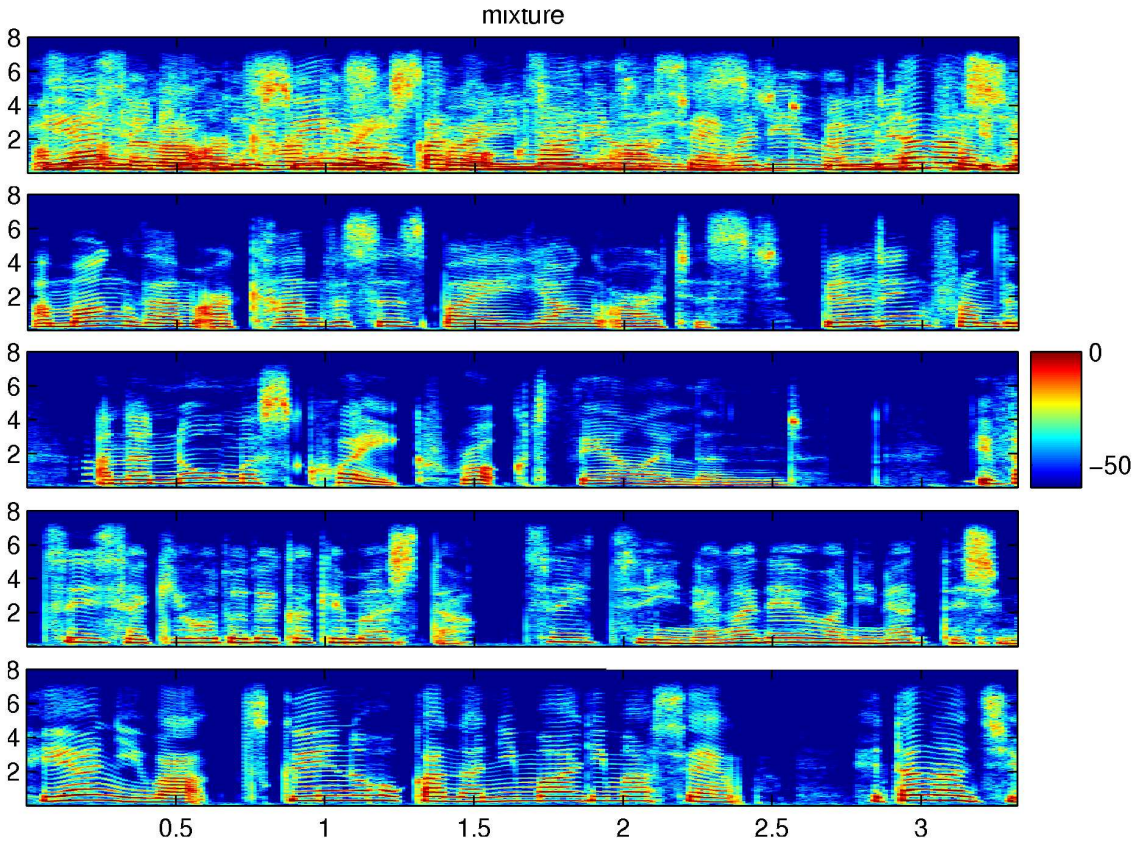
Signal Models for Monaural Source Separation

Ron Weiss

LabROSA, Columbia University

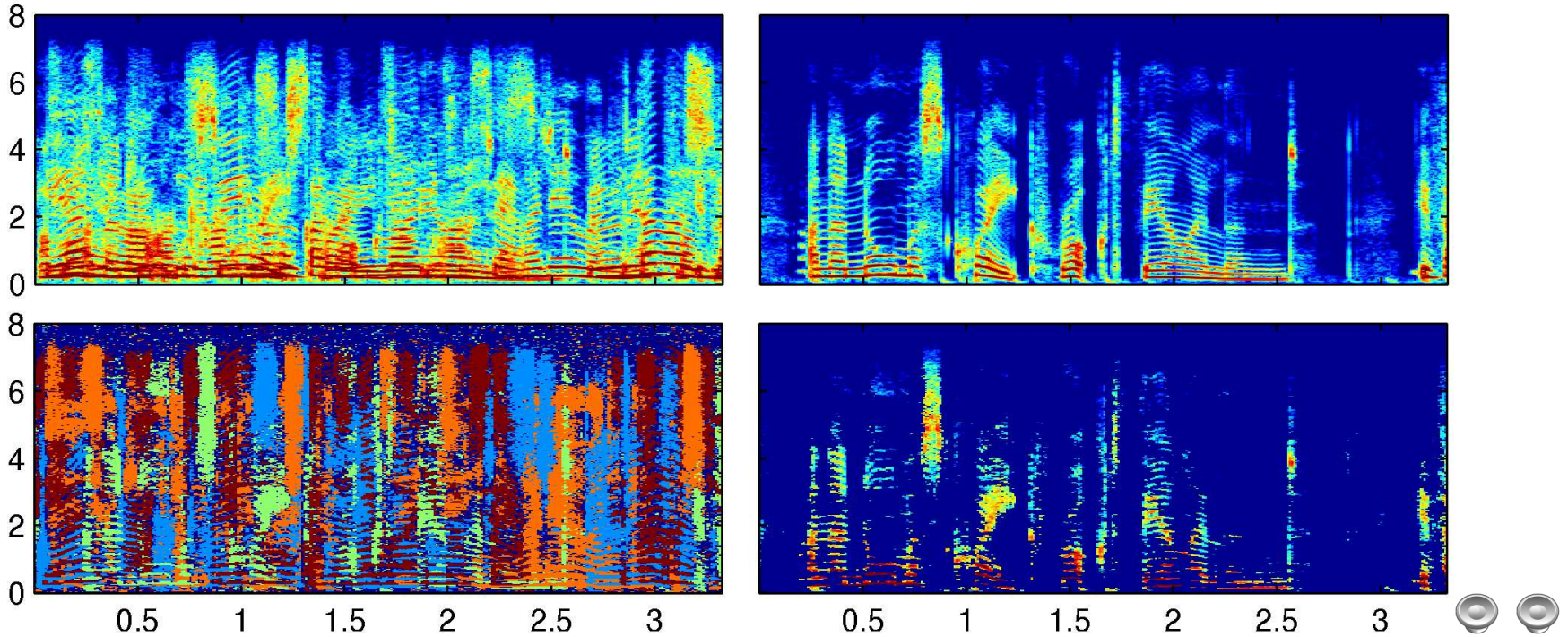
April 27, 2007

Source separation



- Most 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

What makes monaural separation possible



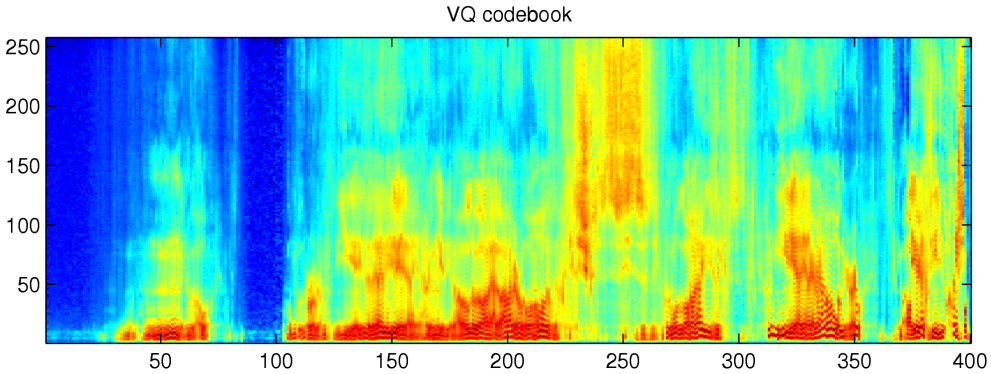
- Natural sounds tend to be sparse in time and frequency - 10% of T-F cells contain 78% of the energy
- And redundant - still intelligible when 22% of the source energy is masked
- "Glimpses" of clean signal even in dense mixtures

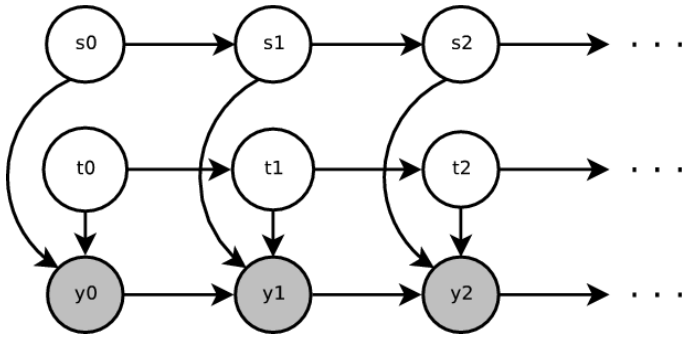
Previous work

- Use constraints on the source signals to guide separation
- Multiple-channel case
  - Independence constraints (e.g. ICA) - perfect separation possible
  - Spatial constraints (e.g. beamforming)
- Single-channel
  - Under-determined - more unknowns (sources) than observations
  - Use perceptual cues similar to lower level processes in human auditory system (CASA)
    - Segment STFT into glimpses of each source
    - By harmonicity, common onset, etc.
    - Sequential grouping heuristics
    - Gaps...
  - Inference based on prior source models
    - Generative models (VQ/GMMs, HMMs)

$$P(\{S_i\}|Y) = P(Y|\{S_i\})P(\{S_i\})$$

Model-based separation





- Log spectral features
- Factorial inference
- Efficient inference using max approximation [Roweis, 2003] :

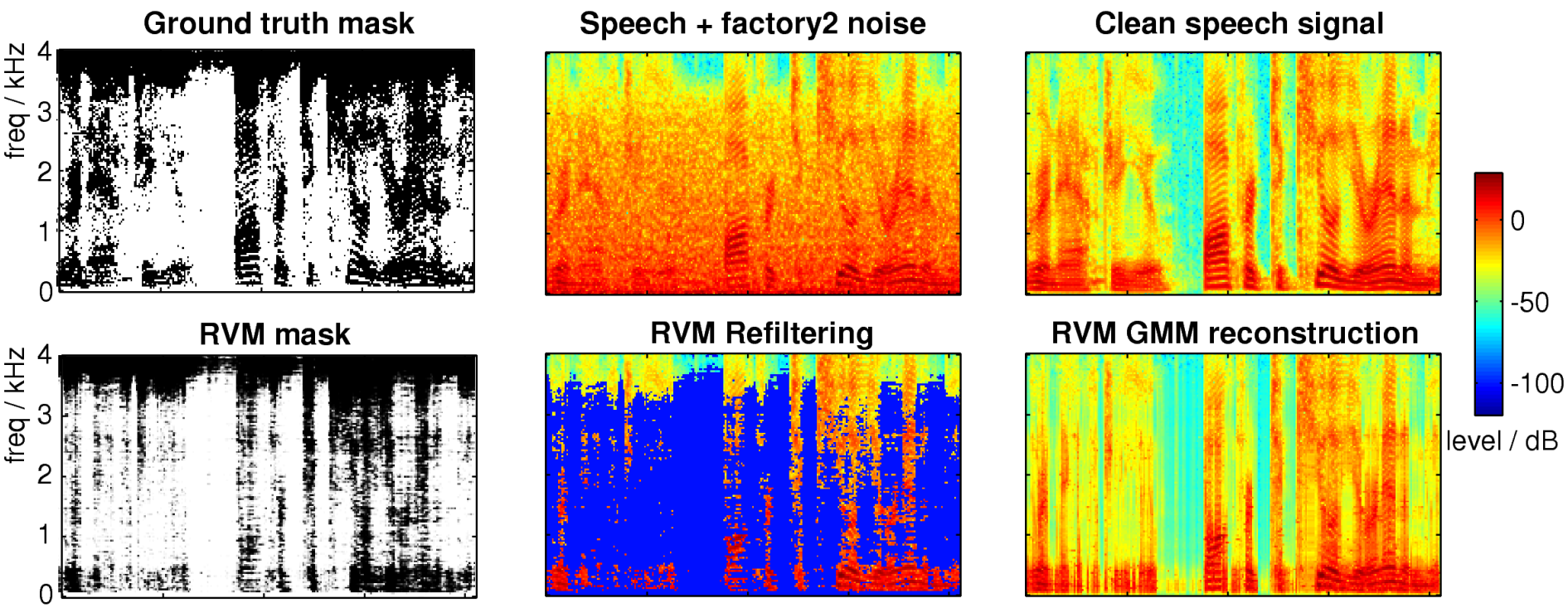
$$\log(s_1 + s_2) \approx \max(\log s_1, \log s_2)$$

- Generate time-frequency masks for each source:

$$M_1 = \begin{cases} 1 & \text{if } s_1 \geq s_2 \\ 0 & \text{if } s_1 < s_2 \end{cases}$$

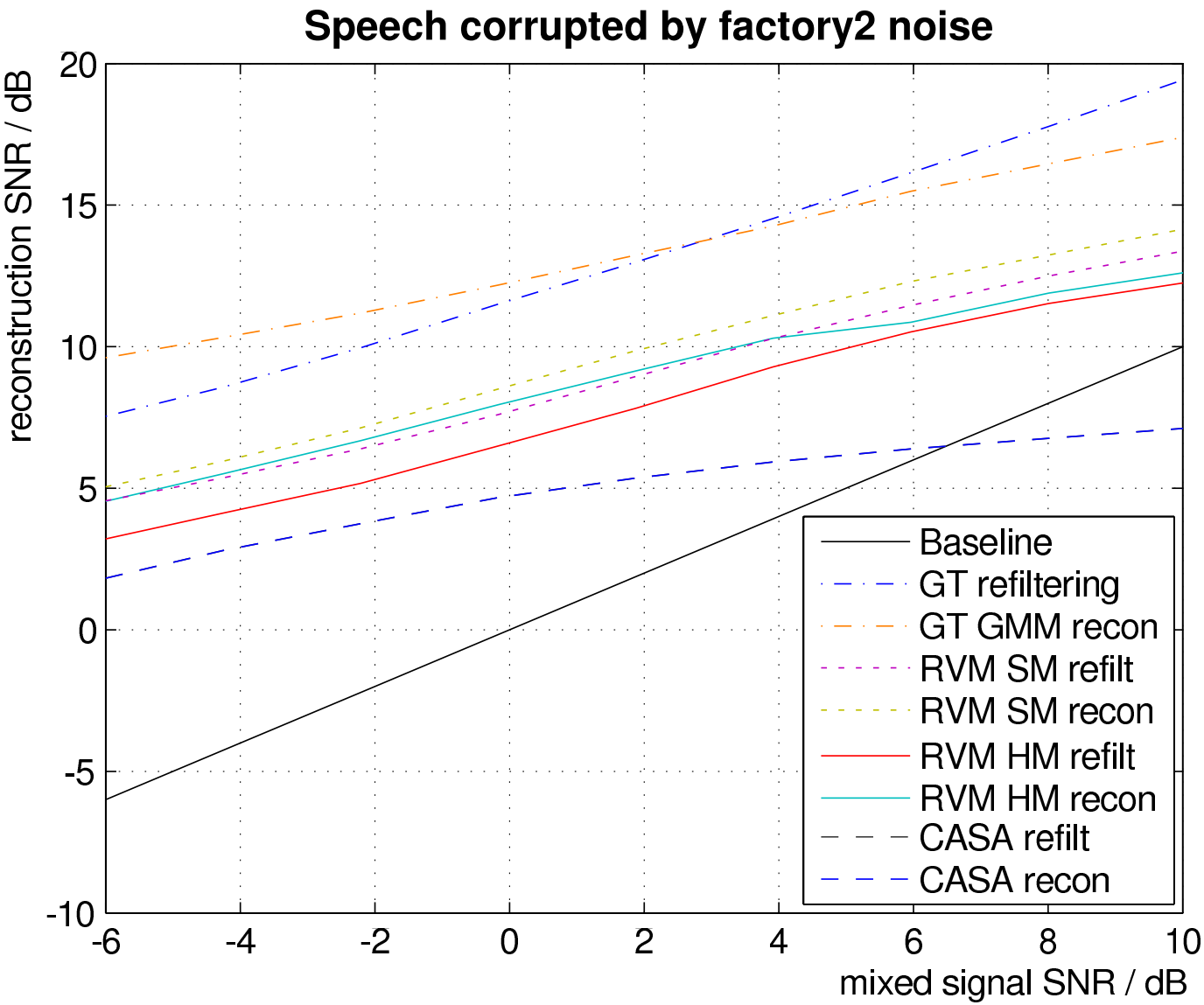
- MMSE reconstruction - fill in the gaps
- What constraints are necessary?
  - Source-dependent or source-adapted models?
  - How important are dynamics?

Preliminary results - T-F masking [Weiss and Ellis, 2006]

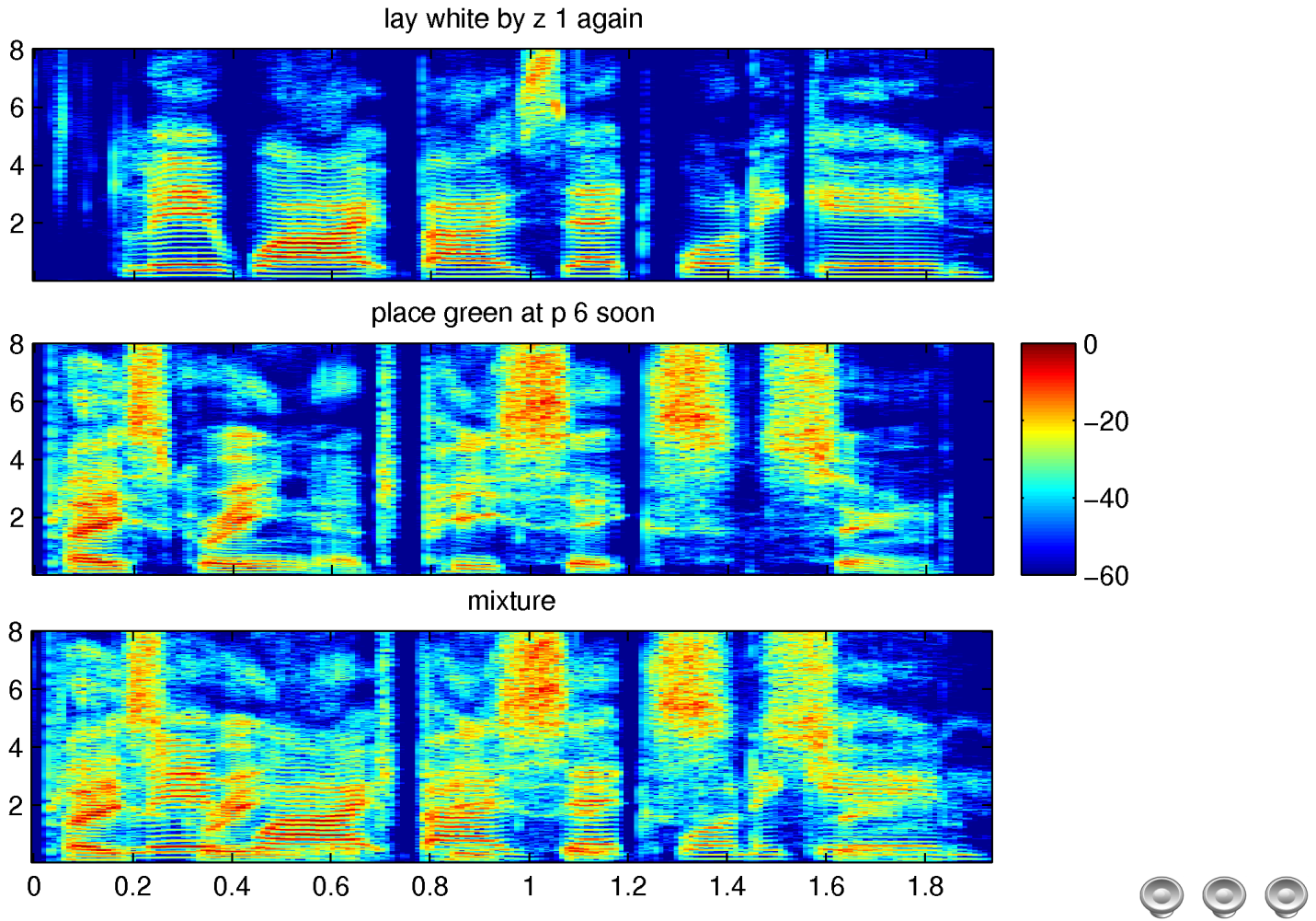


- Treat foreground/background segregation as a classification task
- Use discriminative classifier trained on noisy speech to compute time-frequency masks
- Missing data inference using prior signal models for MMSE reconstruction

Preliminary results - T-F masking performance

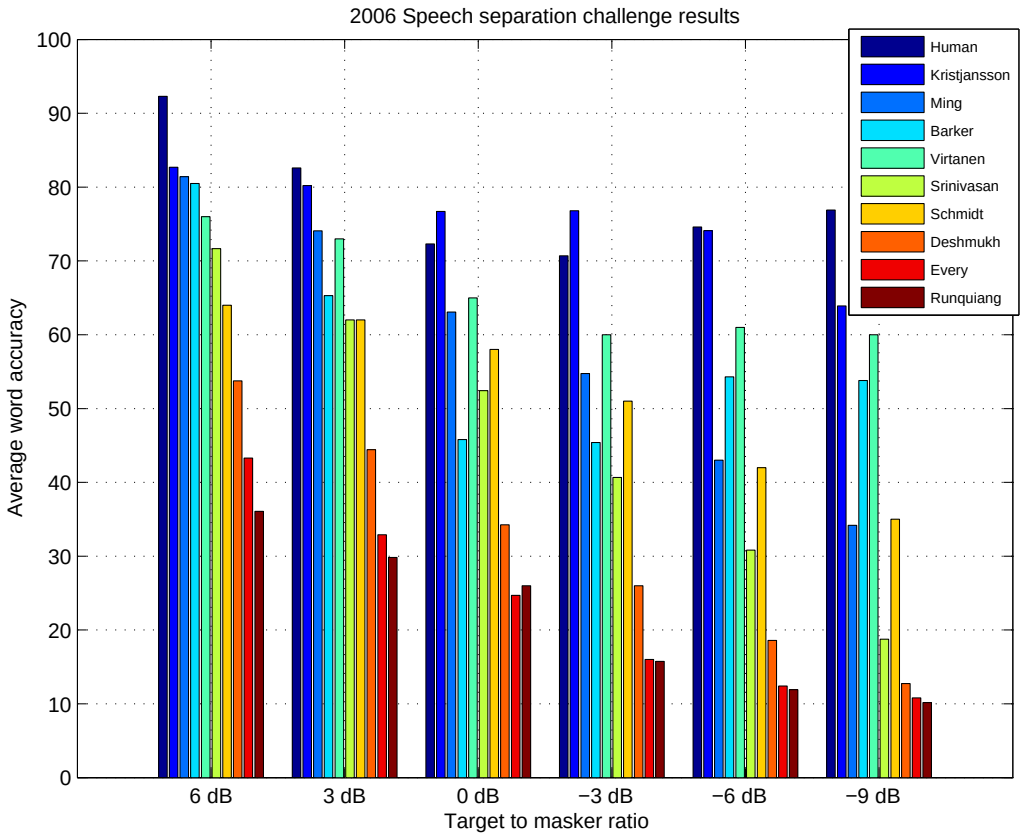


2006 Speech separation challenge [Cooke and Lee, 2006]



- Single channel mixtures of utterances from 34 different speakers
- Constrained grammar: `command(4) color(4) preposition(4) letter(25) digit(10) adverb(4)`

2006 Speech separation challenge - results



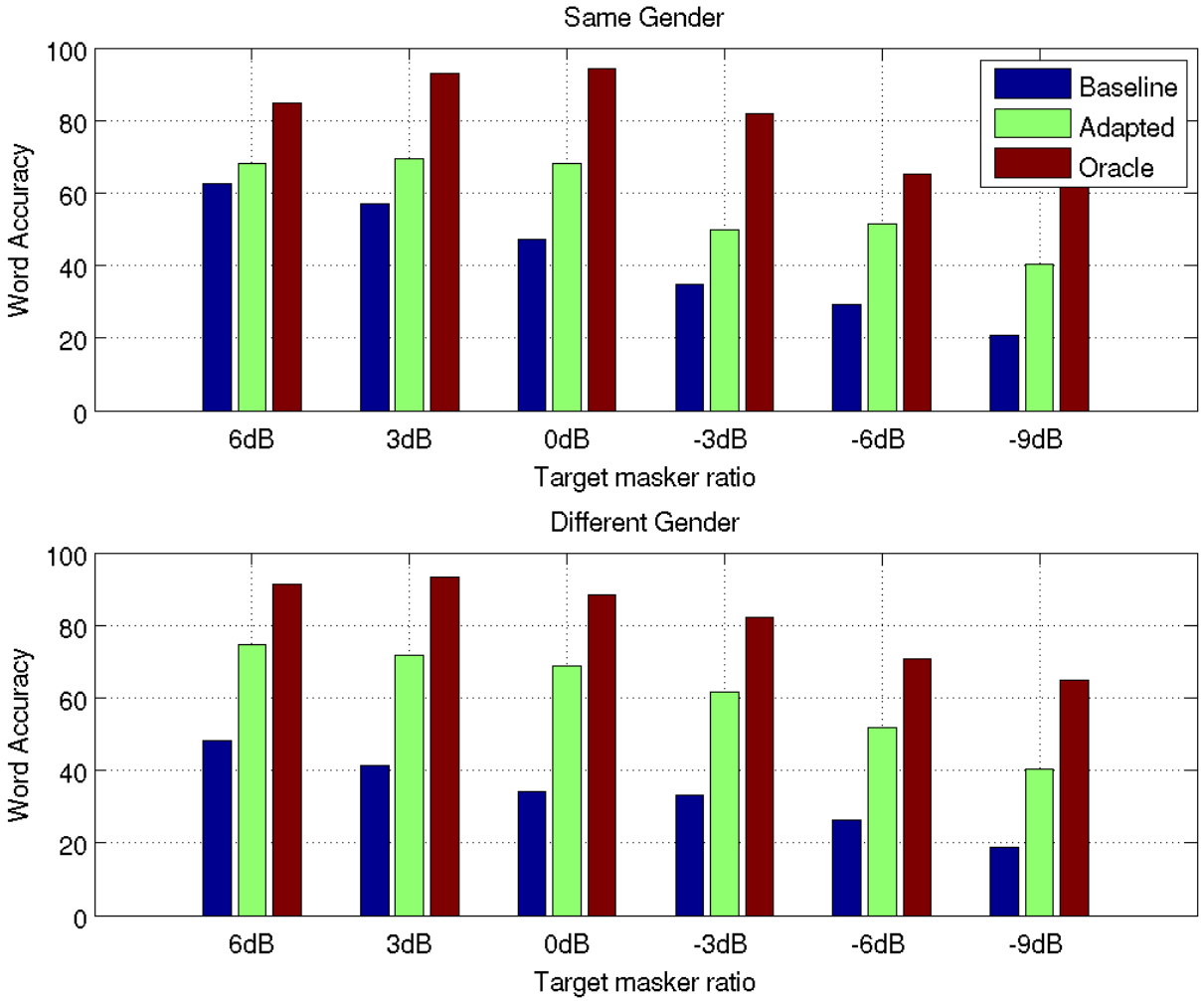
- Model-based separation systems worked best
- Best results from systems highly tuned to the challenge parameters
  - e.g. Iroquois [Kristjansson et al., 2006] - Speaker dependent models with acoustic dynamics (fully connected HMMs) and grammar constraints - "separation by recognition"

Current work - model adaptation

- What if task isn't so well defined?
  - No a priori knowledge of speakers or grammar
- Speaker independent speech model
- Need strong dynamic constraints for good separation - permutations
  - "place white by t 4 now" mixed with "lay green with p 9 again"
  - "place white by p 9 again"
- Adapt SI model to sources present in mixture [Ozerov et al., 2005]
- How to reliably infer adaptation parameters from a single mixture?
  - Use PCA to reduce number of parameters - "Eigenvoices" [Kuhn et al., 2000]
  - Iterative parameter inference

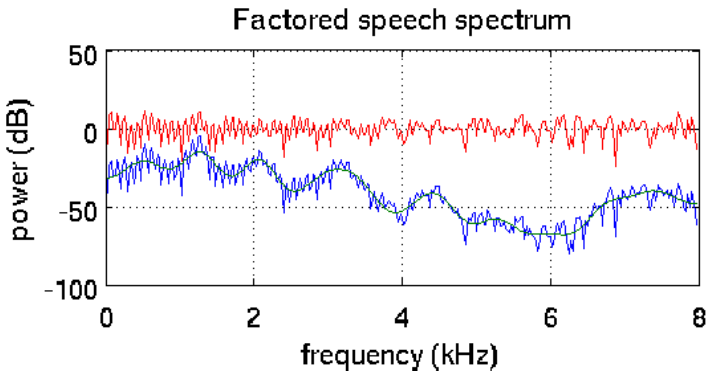
- Result is a compact parametric speech model that can be adapted to unknown speakers

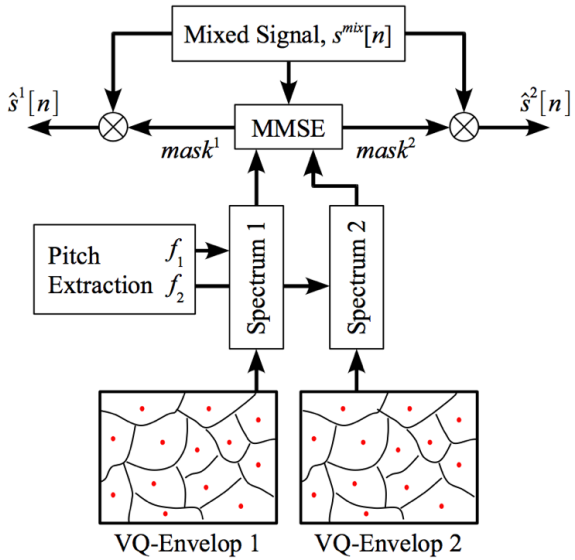
Preliminary results - Speech separation with source adapted models



Future work - Efficient inference using factored speech representation

- Models need to be quite large for good separation [Ellis and Weiss, 2006] - factorial search expensive
- Factor models to treat pitch and formants independently [Radfar et al., 2006]
  - Smaller models to capture the same information ($n + m$ instead of nm codewords) - need less training data
  - Read pitch directly from mixed signal, only need to adapt formants





Summary

- Explicit prior signal models for under-determined source separation
- Contributions
  - Classification system for foreground/background segregation and signal models for MMSE reconstruction
  - Ongoing work - speech separation using speaker adapted models
  - Future work - efficient inference using pitch-factored speech models
- Tentative timeline:

Spring 2007	Speaker adapted models
Summer 2007	Summer internship
Fall 2007 - Winter 2008	Pitch-factored speech models
Spring 2008 - Fall 2008	Write thesis
Fall 2008	Thesis defense

References

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

Ellis, D. P. W. and Weiss, R. J. (2006).
Model-based monaural source separation using a vector-quantized phase-vocoder representation.
In *Proc. of ICASSP*, Toulouse, France.

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 *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustice*.

Radfar, M. H., Dansereau, R. M., and Sayadiyan, A. (2006).
Performance evaluation of three features for model-based single channel speech separation.
In *Proceedings of InterSpeech*.

Roweis, S. T. (2003).
Factorial models and refiltering for speech separation and denoising.
In *Proceedings of EuroSpeech*.

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. of SAPA*, Pittsburgh, PA.