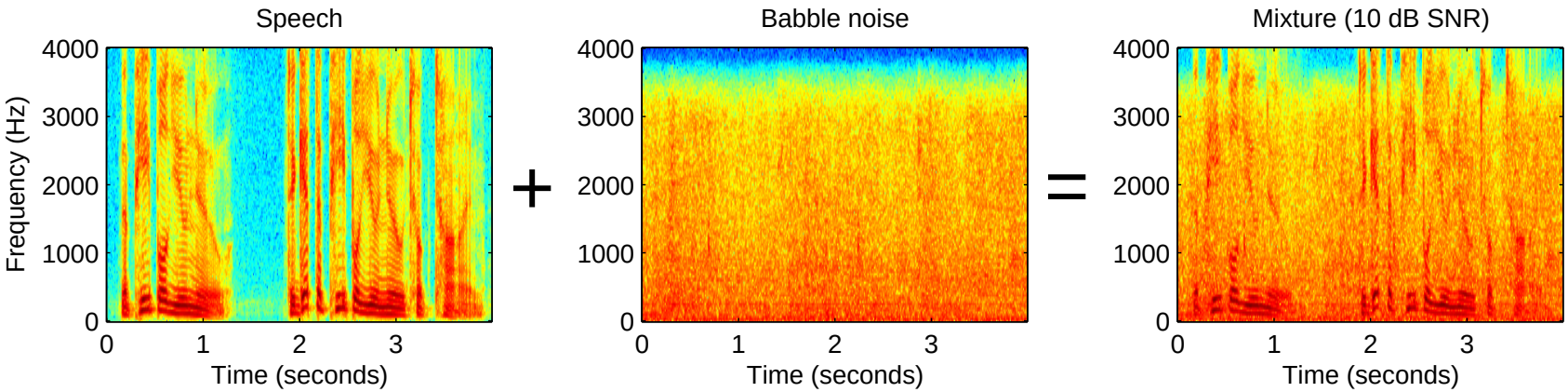
A Classification Approach to Single Channel Source Separation

CS 6772 Project

Ron Weiss

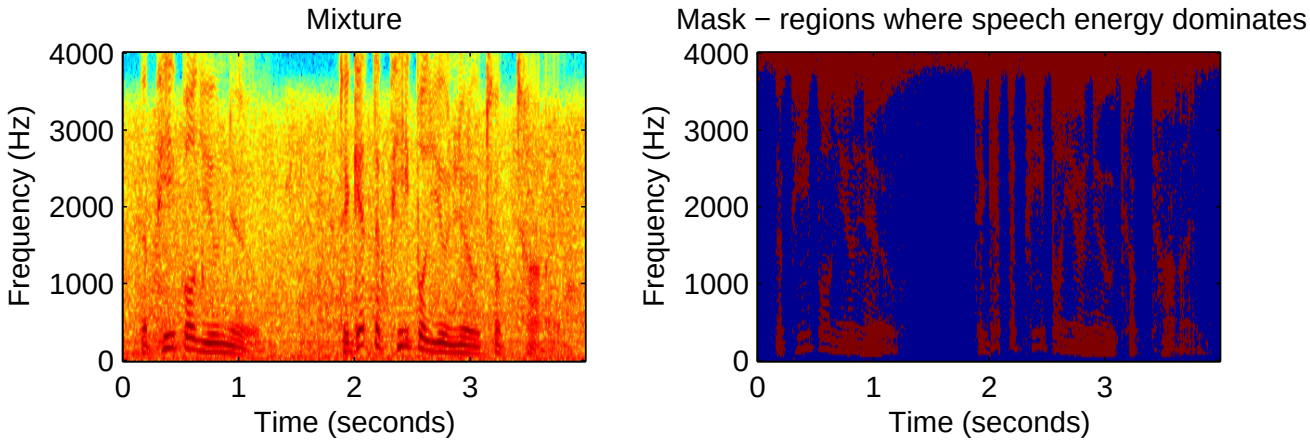
ronw@ee.columbia.edu

Single Channel Source Separation



- Have a monoaural signal composed of multiple sources
- e.g. multiple speakers, speech + music, speech + background noise
- Want to separate the constituent sources
- For noise robust speech recognition, hearing aids

What Data Is Reliable?

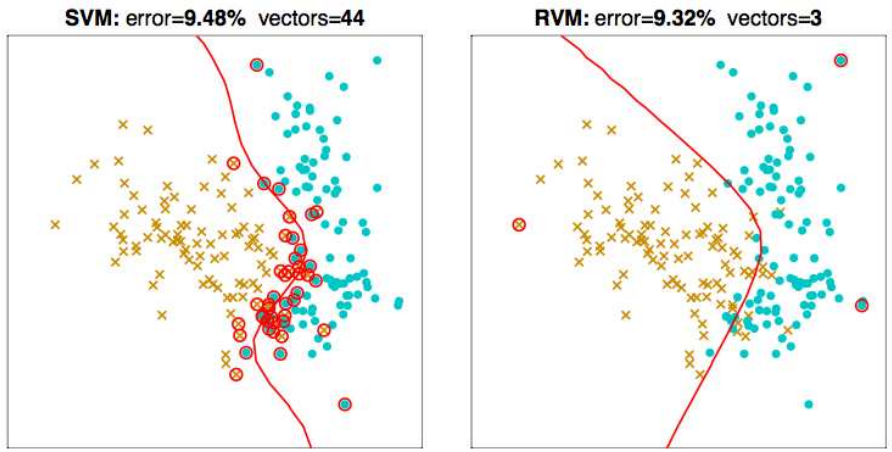


- Only one source is likely to have a significant amount of energy in any given time/frequency cell
- If we can decide which cells are dominated by the source of interest (i.e. has local SNR greater than some threshold), can filter out noise dominated cells ("refiltering" [5])

Binary Masks As Classification [6]

- Goal is to classify each spectrogram cell as being reliable (dominated by speech signal) or not.
- Separate classifier for each frequency band
- Train on speech mixed with a variety of different noise signals (babble noise, white noise, speech shaped noise, etc...) at a variety of different levels (-5 to 10 dB SNR)
- Features: raw spectrogram frames
  - current frame + previous 5 frames (~ 40 ms) of context

The Relevance Vector Machine [7]



- Bayesian treatment of the SVM
- Huge improvement in sparsity over SVM (~ 50 rvs vs. ~ 450 svs per classifier on this task)
- Does more than just discriminate - gives estimate of posterior probability of class membership
- So masks are no longer strictly binary. Can use RVM to estimate the probability that each spectrogram cell is reliable.

Missing Feature Signal Reconstructio

- What if significant part of the signal is missing?
- Want to fill in the blanks in spectrogram of mixed signal
- Do MMSE reconstruction on missing dimensions:

$$x_m = E[x_m|z] = \sum_k \mu_{k,m} P(k|z)$$

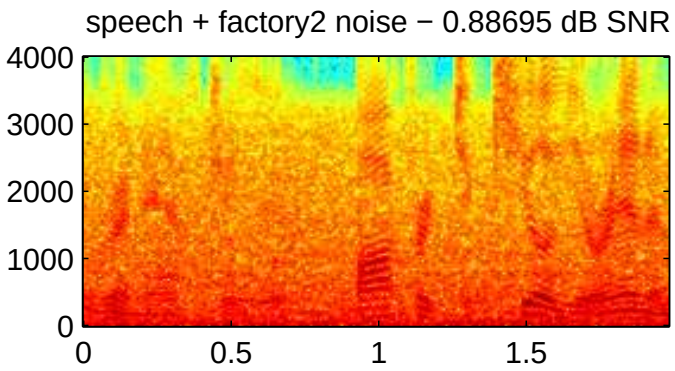
- Use signal model of spectrogram frames - GMM with diagonal covariance

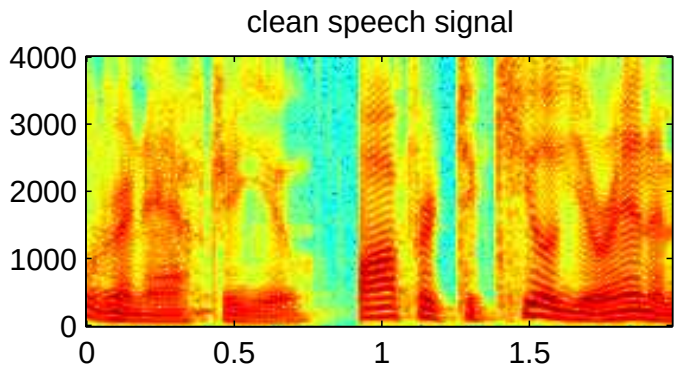
$$P(k|z) = P(k)P(z|k) = P(k)\prod_d P(z_d|k)$$

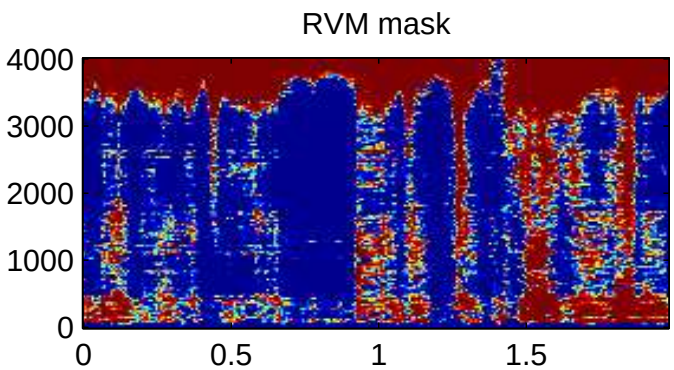
- Just marginalize over missing dimensions to do inference

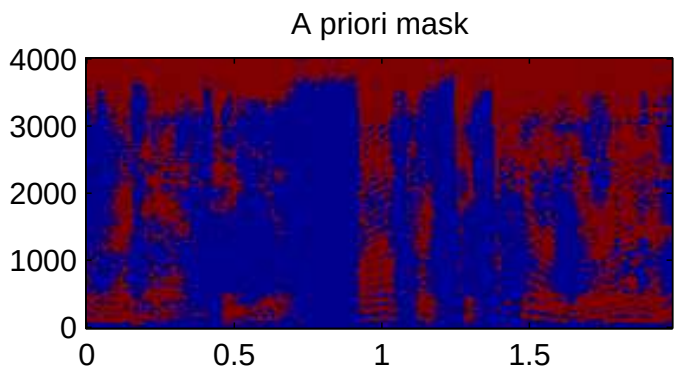
$$P(z_d|k) = P(r_d)\mathcal{N}(z_d|\mu_{k,d},\sigma_{k,d}) + (1-P(r_d))\int \mathcal{N}(z_d|\mu_{k,d},\sigma_{k,d})dz_d$$

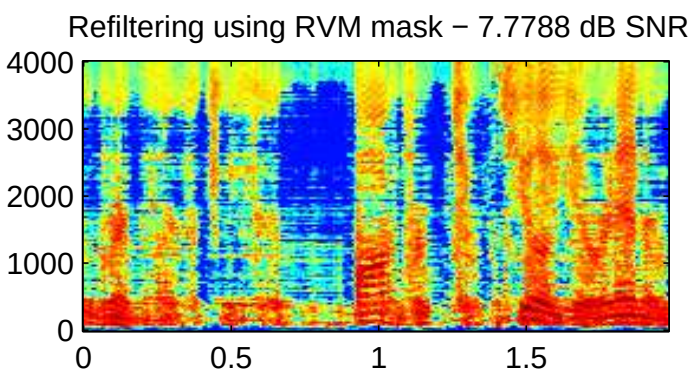
Example

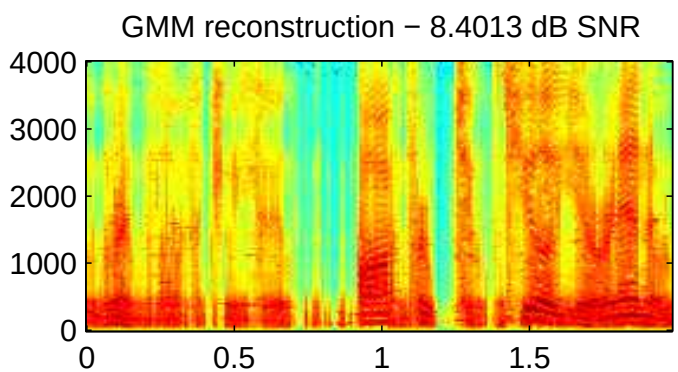

speech + factory2 noise − 0.88695 dB SNR
4000
3000
2000
1000
0
0
0.5
1
1.5


clean speech signal
4000
3000
2000
1000
0
0
0.5
1
1.5


RVM mask
4000
3000
2000
1000
0
0
0.5
1
1.5


A priori mask
4000
3000
2000
1000
0
0
0.5
1
1.5


Refiltering using RVM mask − 7.7788 dB SNR
4000
3000
2000
1000
0
0
0.5
1
1.5


GMM reconstruction − 8.4013 dB SNR
4000
3000
2000
1000
0
0
0.5
1
1.5

References

[1] J. Barker, P. Green, and M. Cooke. Linking auditory scene analysis and robust asr by missing data techniques. In *WISP*, pages 295–307, April 2001.

[2] M. P. Cooke, P. Green, L. B. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, May 2001.

[3] B. Raj, M. L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296, 2004.

[4] A. M. Reddy and B. M. Raj. Soft mask estimation for single channel source separation. In *SAPA*, 2004.

[5] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proceedings of EuroSpeech*, 2003.

[6] M. L. Seltzer, B. Raj, and R. M. Stern. Classifier-based mask estimation for missing feature methods of robust speech recognition. In *Proceedings of ICSLP*, 2000.

[7] M. Tipping. The relevance vector machine. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press, 2000.