Estimating Single-Channel Source Separation Masks

Relevance Vector Machine Classifiers vs. Pitch-Based Masking

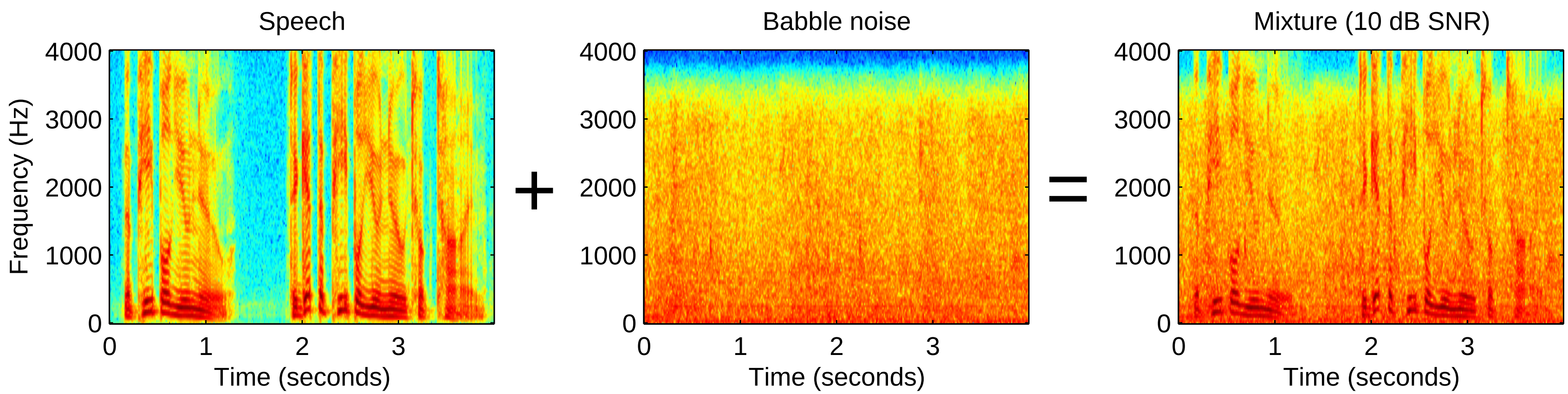
Ron J. Weiss, Daniel P. W. Ellis

{ronw,dpwe}@ee.columbia.edu

LabROSA, Columbia University

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

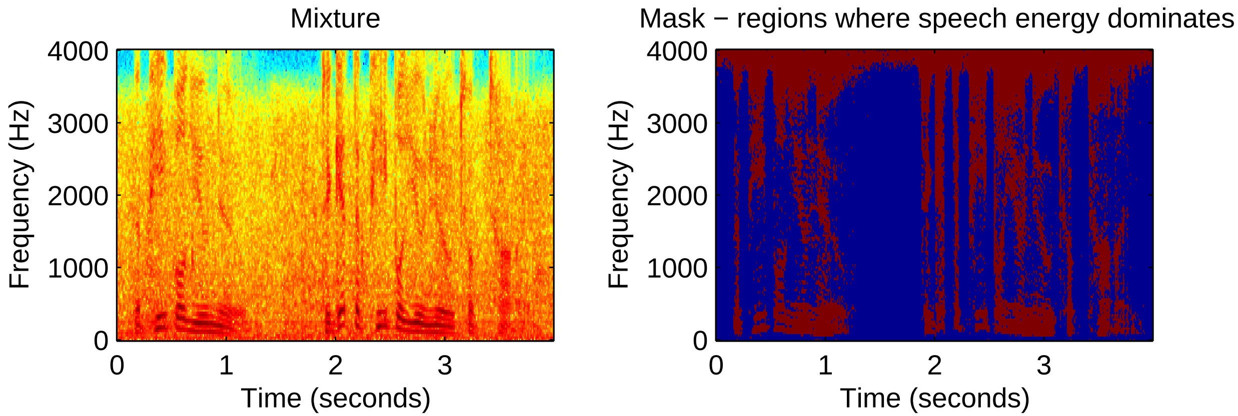
Single Channel Source Separation



- Given 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

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

Missing Data Masks



- Leverage the sparsity of audio sources - 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), we can filter out noise dominated cells ("refiltering" [3])
- Create a binary mask that labels each cell of the spectrogram as missing or reliable

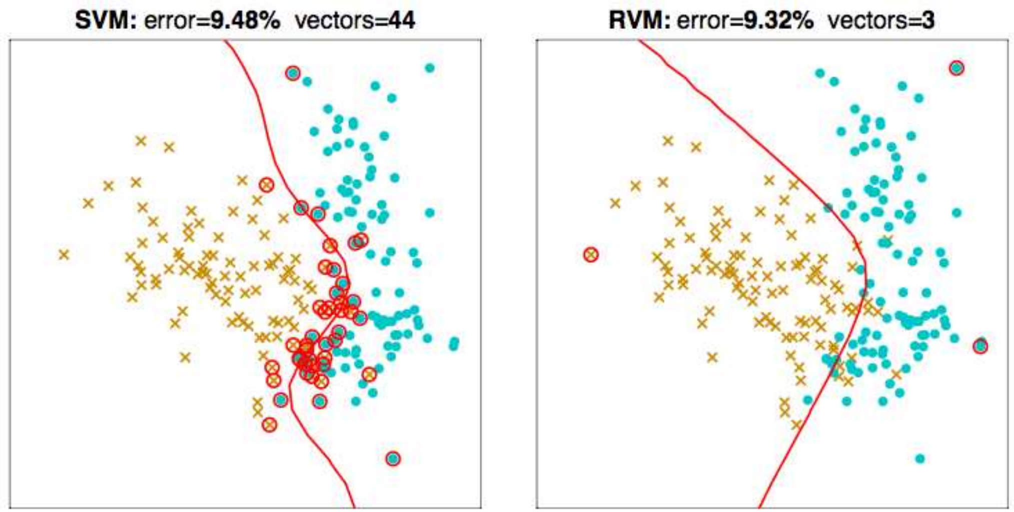
Laboratory for the Recognition and Organization of Speech and Audio

Mask Estimation As Classification [4]

- 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

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

The Relevance Vector Machine [5]



- Bayesian treatment of the SVM
- Kernel classifier of the form:

$$y(\mathbf{z}|\mathbf{w},\mathbf{v}) = \sum_n w_n K(\mathbf{z},\mathbf{v}_n) + w_0$$

  - $\mathbf{z}$ = data point to be classified
  - $\mathbf{v}_n$ = nth support vector
  - w_n = weight associated with the nth support vector

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

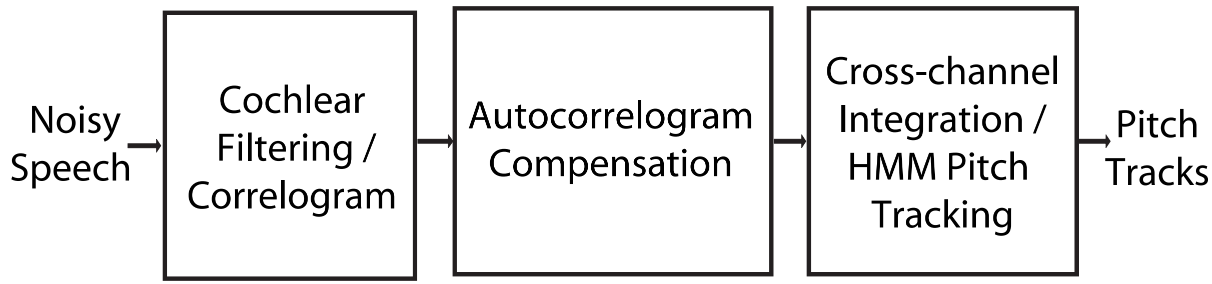
RVM Versus SVM

- Pros
  - Huge improvement in sparsity over SVM (~ 50 rvs vs. ~ 450 svs per classifier on this task) - faster classification
  - Wrap y in a sigmoid squashing function to estimate posterior probability of class membership.

$$P(t = 1|\mathbf{z}, \mathbf{w}, \mathbf{v}) = \frac{1}{1 + e^{-y(\mathbf{z}|\mathbf{w},\mathbf{v})}}$$

  - Masks are no longer strictly binary. Can use RVM to estimate the probability that each spectrogram cell is reliable.
- Cons
  - RVM training is significantly slower

CASA Pitch-based Masking [1]



- Most energy in speech signals is associated with the pseudo-periodic segments of vowel sounds
- Get envelopes of auditory filter outputs
- Find strong periodicities in short-time autocorrelation of each envelope
- Sum each channel to find single dominant periodicity
- Channels whose autocorrelation indicated energy at this period are added to the target mask

Missing Data Reconstruction [2]

- What if a significant part of the signal is missing?
- Want to fill in the blanks in spectrogram of mixed signal
- Do MMSE reconstruction on missing dimensions using signal model of spectrogram frames - GMM trained on clean speech
- Marginalize over missing dimensions to do inference

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

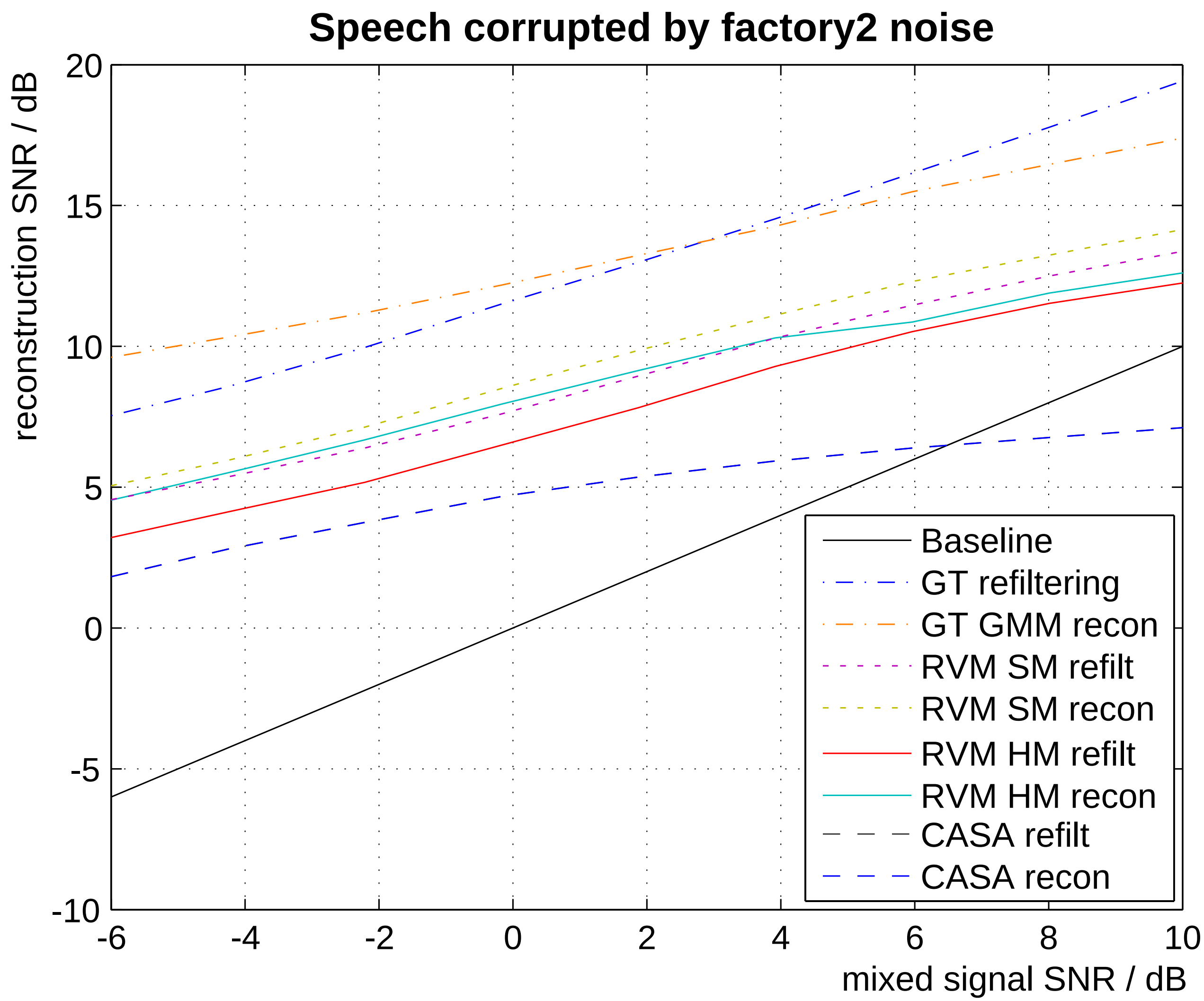
- MMSE estimator reconstructs by mixing the observed signal and GMM reconstruction based on the probability that each cell is reliable:

$$x_d = E[x_d|z] = P(r_d)z_d + (1 - P(r_d)) \sum_k P(k|z)\mu_{k,d}$$

Experiments

- Speech signal: single male speaker from audio book recording
- Training noise signals: Babble noise, speech shaped noise, factory noise 1
- Out of model noise signals used for testing: car noise, white noise, factory noise 2, music
- RVM trained on 20s of speech + noise
- 512 component GMM trained on 80s of clean speech

Experiments - Results

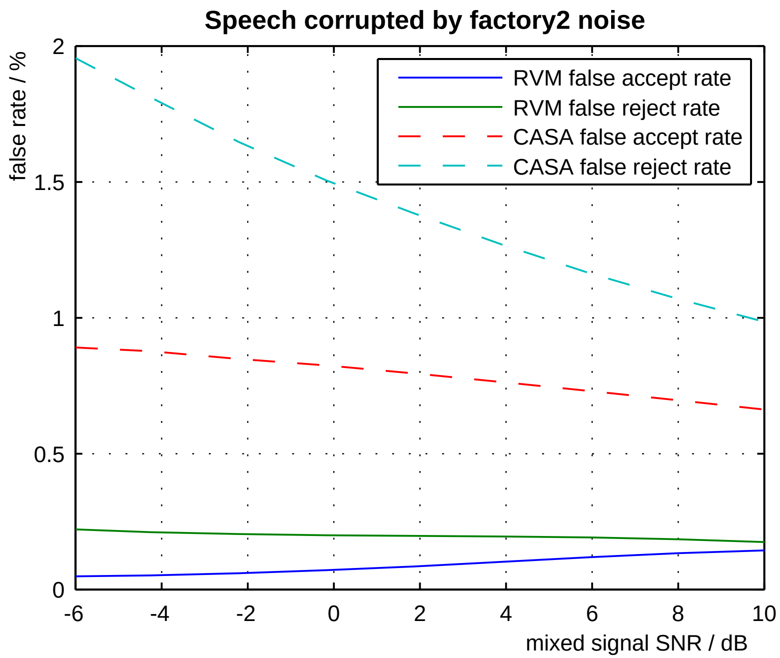


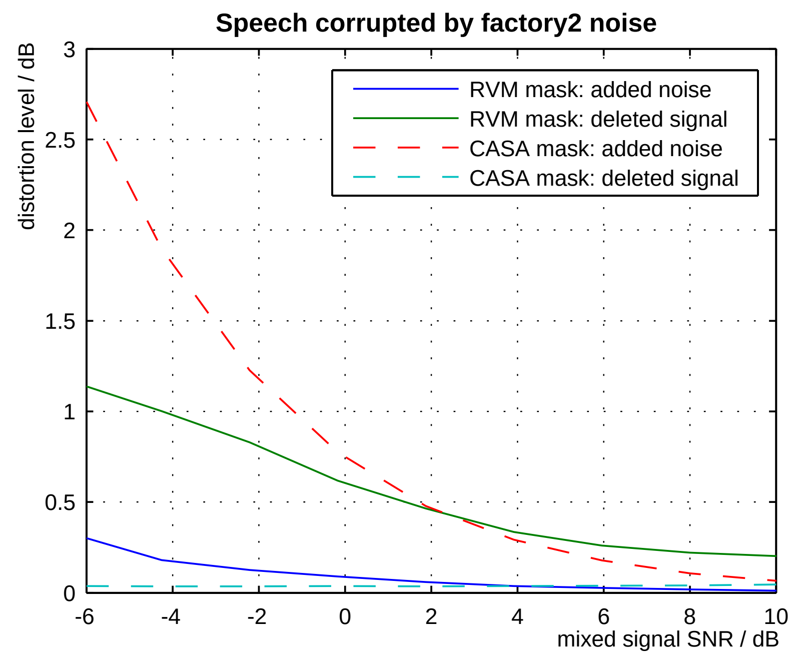
Laboratory for the Recognition and Organization of Speech and Audio

Experiments - Results

- GMM reconstruction outperforms simple refiltering since the GMM reconstruction can fill in the blanks
- Soft masks give about 1 dB improvement over hard masks
- CASA masks not as good as RVM masks
- Still room for improvement in mask estimation based on performance using ground truth masks

Experiments - Results

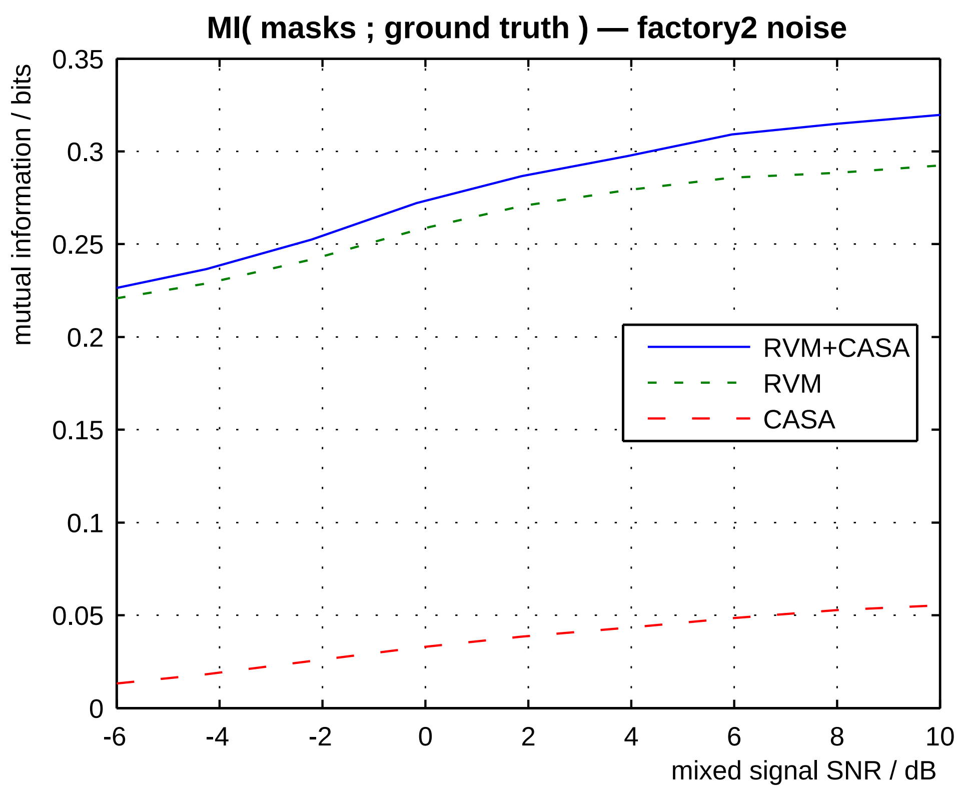




- False positive rate of CASA masks is much higher than that of RVM masks.
- Major problem with CASA mask is added noise. Deleted signal is not very significant in terms of signal energy
- RVM mask deletes a significant amount of signal energy

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

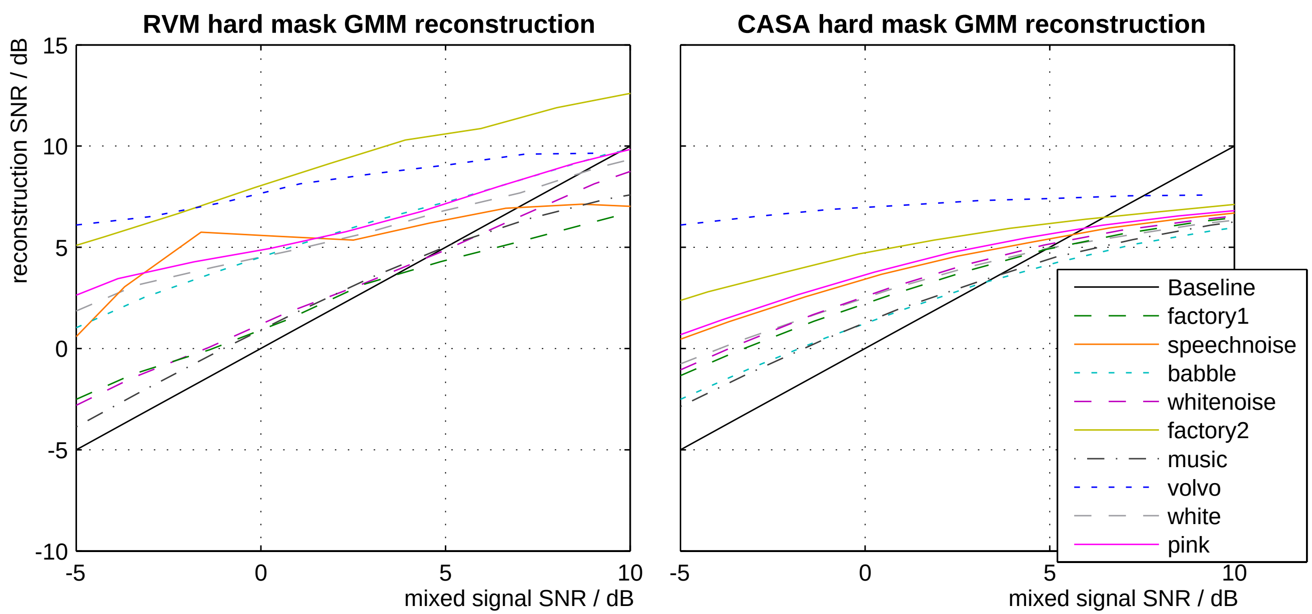
Experiments - Results



- RVM mask is significantly more informative about ground truth mask than CASA mask
- Some information in CASA mask is not captured by RVM mask

LabROSA
Laboratory for the Recognition and Organization of Speech and Audio

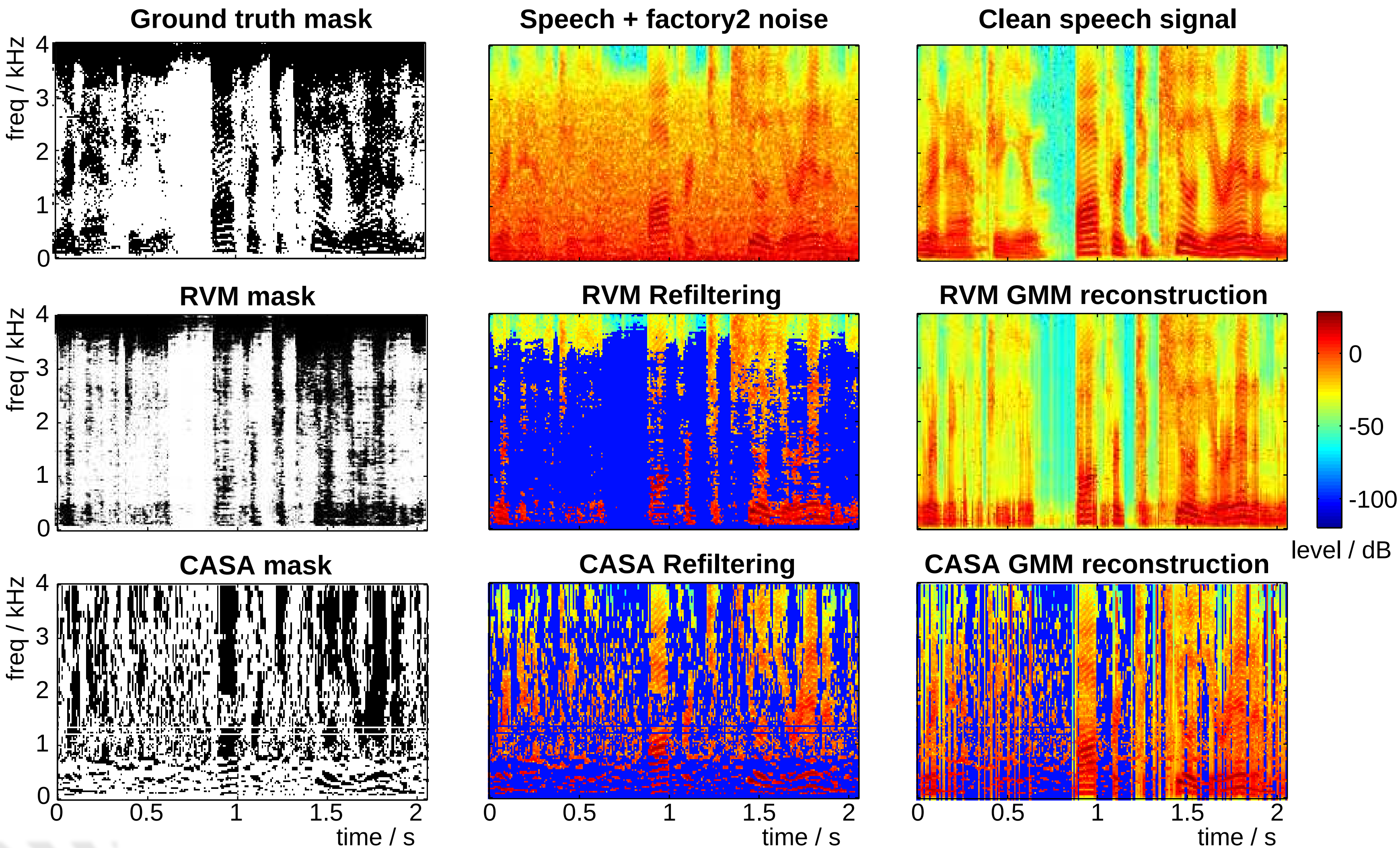
Experiments - Results



- Clear SNR boost when mixed signal at low SNR
- RVM clearly outperforms CASA system
- Both systems perform poorly on music noise
  - RVM not trained on highly pitched interference
  - CASA system can't distinguish between voiced speech and musical instrument harmonics

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

Spectrograms



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

References

[1] K. S. Lee and D. P. W. Ellis. Voice activity detection in personal audio recordings using autocorrelogram compensation. In *Proc. Interspeech ICSLP-06*, Pittsburgh PA, 2006. submitted.

[2] B. Raj and R. Singh. Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 27–32, November 2005.

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

[4] 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.

[5] 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.