Estimating Single-Channel Source Separation Masks: Relevance Vector Machine Classifiers vs. Pitch-Based Masking

Ron J. Weiss and Daniel P. W. Ellis

LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
{ronw,dpwe}@ee.columbia.edu

Abstract

Audio sources frequently concentrate much of their energy into a relatively small proportion of the available time-frequency cells in a short-time Fourier transform (STFT). This *sparsity* makes it possible to separate sources, to some degree, simply by selecting STFT cells dominated by the desired source, setting all others to zero (or to an estimate of the obscured target value), and inverting the STFT to a waveform. The problem of source separation then becomes identifying the cells containing good target information. We treat this as a classification problem, and train a Relevance Vector Machine (a probabilistic relative of the Support Vector Machine) to perform this task. We compare the performance of this classifier both against SVMs (it has similar accuracy but is not as efficient as RVMs), and against a traditional Computational Auditory Scene Analysis (CASA) technique based on a noise-robust pitch tracker, which the RVM outperforms significantly. Differences between the RVM- and pitch-tracker-based mask estimation suggest benefits to be obtained by combining both.

1. Introduction

The problem of single channel source separation involves decomposing a mixture of two or more sources into its constituent clean signals. This problem is under-determined since we want to be able to extract two or more signals when only one signal is given. Therefore techniques such as indepedent component analysis will not work directly. However, due to the sparsity of the short-time Fourier transform (STFT) representation for most audio signals, only one source is likely to have a significant amount of energy in any given time-frequency cell. This motivates the approach of attempting to identify the regions of the mixed signal that are dominated by each source and treating these regions as independent signals (i.e. refiltering [1, 2]).

Many recent approaches to single channel source separation, such as [2, 3], require prior knowledge of the nature of the signals present in the mixed signal. Each source is modeled by clustering spectral slices from the STFT using a Gaussian mixture model (GMM). Inference involves the creation of binary masks that indicate which STFT cells are dominated by each source. This approach requires explicit models for each of the interfering signals and a factorial search over all possible combinations of frames of each signal.

An alternative approach to mask generation is given in [4] which does not require a factorial search. A simple maximum likelihood Gaussian classifier is used to generate these masks. This approach was shown to generalize well over many different kinds of interference.

Given these masks, a GMM signal model can be used to fill in the missing spectral regions that were labelled as unreliable and reconstruct the clean signal as in [5, 6, 7].

In this paper we present a system that is able to recover a speech signal in the presence of additive non-stationary noise through a combination of the classification approach to mask estimation and the use of signal models for reconstructing the parts of the speech signal that are obscured by the interference. We also compare this classifier-based approach to an alternative approach, frequently referred to as Computational Auditory Scene Analysis (CASA), which attempts to identify the pitch track of target speech, then to build an STFT mask to select cells reflecting that pitch.

Section 2 reviews relevance vector machine classifiers which we use to generate the masks. Section 3 reviews techniques for reconstructing the unreliable dimensions of the mixed signal using missing data masks. In section 4, we briefly describe our contrast, CASA-based mask generation system. Section 5 presents some experimental results, followed by conclusions in section 6.

2. The Relevance Vector Machine

The relevance vector machine [8] is a kernel classifier similar to the support vector machine, but derived using a Bayesian approach. As with the SVM, the RVM forms a linear classifier in a high dimensional kernel space defined by some kernel function.

Like an SVM, the RVM makes predictions using a function of the following form:

$$y(\mathbf{z}|\mathbf{w},\mathbf{v}) = \sum_n w_n K(\mathbf{z},\mathbf{v}_n) + w_0 \tag{1}$$

where $\mathbf{z}$ is the data point to be classified, $\mathbf{v}_n$ is the nth support vector with associated weight w_n, and K is some kernel function.

For classification, the probability of the data point $\mathbf{z}$ being in the positive class is given by wrapping eqn. (1) in a sigmoid squashing function:

$$P(t=1|\mathbf{z},\mathbf{w},\mathbf{v}) = \frac{1}{1+e^{-y(\mathbf{z}|\mathbf{w},\mathbf{v})}} \tag{2}$$

Instead of attempting to produce a classifier with maximum margin, as in the SVM case, the RVM approach attempts to produce a sparse model (i.e. one with very few support vectors). This

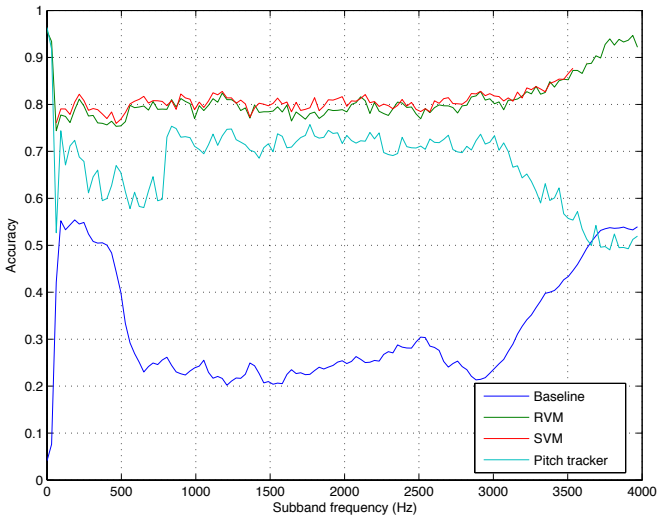

Figure 1: Mask generation accuracy for each frequency band on held out testing data. The baseline performance is the percentage of positive labels in the test data. The SVM performs slightly better than the RVM on average.

is accomplished by learning the weights, $\mathbf{w}$, in a probabilistic manner and defining hyperparameters over each weight w_n so that a different hyperparameter is associated with each support vector. As noted in [8], in practice the posterior distributions over the weights become infinitely peaked at zero. The weights associated with uninformative support vectors, i.e. those that do not help predict class labels, go to zero. Therefore, the support vectors associated with those weights can effectively be removed from the RVM model. In the interest of space, the details of the learning algorithm are omitted. They can be found in [8].

2.1. Advantages of the RVM

The RVM approach has a number of advantages over the SVM, including a significant improvement in sparsity over the equivalent SVM, as mentioned above. The RVM generally uses about 10% as many support vectors as the SVM on this task; for our experiments where we trained on 500 frames, the RVM used around 50 examples in the classifier, whereas the SVM consistently used virtually all of them. In addition, the RVM does not restrict the set of allowable kernels to those that obey Mercer's condition. There is also no need to estimate the "nuisance" parameter C. Finally, the RVM does more than just discriminate: Eqn. (2) gives an estimate of posterior probability of class membership. Tipping argues in [8] that unlike methods to obtain posterior probability estimates from the distance from SVM classifier boundaries, the estimate of the posterior given by the RVM more closely approximates the actual posterior.

2.2. Comparison to the SVM

In order to evaluate the efficacy of RVMs as compared to SVMs on this task, both types of classifiers were trained to predict reliable data masks on speech signals corrupted by various types of noise, similar to [4], but using plain STFT magnitudes as features. Separate classifiers were used for each frequency bin in the STFT. In total, 129 subband classifiers were used. The inputs to each classifier were drawn from all frequency bands (not just the band being classified) over several time frames.

2.2.1. Data

Training and testing data were generated by digitally mixing speech and corrupting noise in MATLAB. Since the clean versions of the underlying signals are available, it is easy to generate ground truth mask labels for mixed signals: An STFT cell in the mixed signal is said to be dominated by the speech signal if the same cell in the clean speech signal has more energy than the same cell in the noise signal. The speech signal was taken from an audiobook recording (male speaker), known to be recorded in clean conditions. The noise signals used were excerpts from the NOISEX database, including: babble noise, car noise ("volvo"), and two different recordings of background factory noise, all of which are non-stationary. In addition, simple stationary signals, including white noise, pink noise, and speech shaped noise (white noise filtered to have the same average spectral envelope as the speech signal) were generated in MATLAB.

The training data consisted of 20 s of speech mixed with 20 s of each of the noise signals at signal to noise ratios varying between -5 dB and 20 dB in increments of 5 dB. Testing was performed using 10 s mixtures with held out sections of the same signals under the same SNRs. The same speaker and noise types were used, but the testing signals consisted of later sections of the sound files that were not used in training.

All signals used were sampled at 8 kHz. STFTs were generated using a 256 point FFT with a 256 point (32 ms) Hanning window and a 64 point (8 ms) hopsize.

2.2.2. Features

The same features were used for each of the subband classifiers. They consisted of the STFT power measured in decibels of the current frame and the previous 5 frames of context, for a total of $6 \times 129 = 774$ feature dimensions. We observed empirically that adding context improved classification accuracy by a few percent. This follows our expectation because speech signals are locally stationary, so knowing that there was significant speech energy at time $t - 30$ ms will usually imply that there is still significant speech energy present at time t.

2.2.3. Cross validation

To obtain the best performance, cross validation was performed to select the best kernel type and kernel parameters for both the RVM and SVM classifiers. Evaluated kernels included linear, polynomial (order 2 and 3), and radial basis function (variance varied between 1 and 16) kernels for both RVM and SVM. In addition, a few exponential family variants of the RBF kernel, including Laplace and Cauchy kernels, were evaluated for the RVM classifiers only. Finally, another level of cross validation had to be performed for the SVM to obtain a good value of C. The parameters that had the highest mean accuracy across all frequency subbands on the test data were chosen as the best.

The best performing SVM used a Gaussian kernel with a variance of 8 and $C = 256$. The best performing RVM used a Cauchy kernel with parameter 8. Use of the Cauchy kernel resulted in only one or two percentage point increases in accuracy over the Gaussian kernel for the RVM.

2.2.4. Results

As seen in fig. 1, the SVM classifiers generally performed slightly better on the test data than the RVM classifiers in most frequency bands. In both cases, the mean accuracy of the 129 subband classifiers was just over 80%. This is a significant improvement over baseline performance of about 31% where every cell is labeled as reliable, i.e. all classifiers output 1 all the time. A more realistic baseline (not pictured) would be one in which each subband classifier always labeled the input with the label that is most common in that subband in the data, giving each classifier at least 50% accuracy. In this case, the mean accuracy is still significantly below that of the SVM and RVM.

The primary difference between the RVM and SVM becomes apparent when looking at the number of support vectors used by each of the classifiers. The number of support vectors used for each subband classifier is roughly constant across all frequency bands for both the SVM and RVM. But the RVM classifiers consistently use a only small fraction (about 10%) of the number of support vectors used by the SVM classifiers. This leads to a corresponding increase in classification speed since the RVM requires fewer inner product/kernel function computations.

3. Missing Feature Reconstruction

Using the RVM subband classifiers described in section 2, a good estimate of the frequency bands of each observed audio frame that are dominated by speech (reliable) or not (unreliable/missing) can be obtained. The RVM goes a step further and gives the probability that each frequency bin is reliable for each observed audio frame.

If much of the observation is missing (e.g. if lowpass noise obscured everything below 200 Hz), these dimensions must be reconstructed in order to obtain a good estimate of the underlying clean signal. This can be accomplished using a prior GMM model of the clean signal to create a minimum mean squared error (MMSE) estimator to reconstruct the missing dimensions given the observed ones. The soft mask reconstruction process is described in [7].

4. CASA Pitch-based masking

Much of the energy in speech is associated with pseudo-periodic segments of vowels and similar sounds, and human listeners appear to be well able to separate and track speech by following the pitch percept that arises from this local periodicity. This has led to several so-called Computational Auditory Scene Analysis systems that attempt to effect signal separation by mimicking the processing of the auditory system. We use an implementation of the system described by [9] which is able to track the pitch of target speech despite high levels of interfering noise. It operates by extracting envelopes from many band-pass signals roughly corresponding to the separate frequency bands used by the ear. The short-time autocorrelation of each envelope is checked for strong periodicities, and the normalized autocorrelations of all such channels are summed to find a single, dominant periodicity. Channels whose individual autocorrelation indicated energy at this period are then added to the target mask for that time step as being dominated by the inferred target. Our work with this pitch tracker is described in more detail in [10].

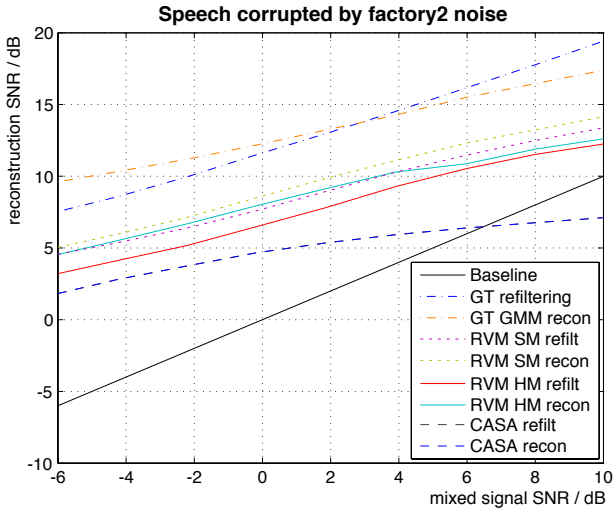

Figure 2: Comparison of the different reconstruction techniques using different masks on speech corrupted by factory noise. Performance using ground truth masks present an upper bound on performance using estimated masks. For each type of mask signal, the GMM reconstrucion performs better. RVM reconstruction using soft masks performs better than reconstruction using hard (binary) masks.

5. Experiments

The data described in section 2.2.1 was used to evaluate the performance of RVM mask generation. However the RVM classifiers were only trained on a subset of the noise signals (speech shaped noise, babble noise, factory noise 1) to evaluate how well the classifiers could generalize to unseen types of noise. Evaluation on out-of-model noise was performed on car noise, different factory noise, white noise and highly nonstationary instrumental music.

The RVM subband classifiers were trained using the kernel and parameters as in section 2.2.3. A random sample of 2000 frames of the training data was used for training. To evaluate performance of MMSE reconstruction, a GMM with 512 mixture components was trained on 80 s of clean speech.

Evaluation was performed on data that was not used to train any of the models used. Four kinds of masks were evaluated: ground truth masks (GT) consisting of binary labels corresponding to a priori knowledge of where the speech signal dominates the mixture, RVM hard masks (HM) consisting of binary labels predicted by the RVM subband classifiers (i.e. $P(r_d) \geq .5$)), and RVM soft masks (SM) consisting of the RVM posterior probability estimates ($P(r_d)$). Finally, performance of the RVM mask generation system is compared to that of the CASA mask generation system described above.

Reconstruction was performed by refiltering as in [2], where each cell of the mixed signal STFT is multiplied by the corresponding cell in the mask, and by MMSE reconstruction as in [7].

All SNR measurements listed in the evaluation are magnitude SNRs measured on the magnitude of the reconstructed STFTs.

5.1. Results

Fig. 2 shows the performance of different reconstruction techniques on speech corrupted by a non-stationary noise signal. SVM

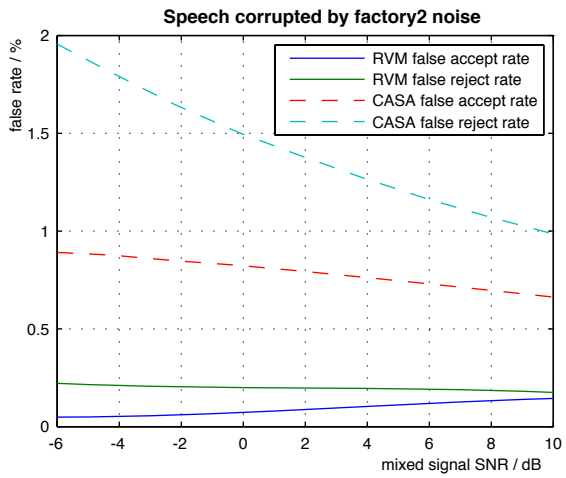

Figure 3: Comparison of errors made by RVM and pitch tracker masks.

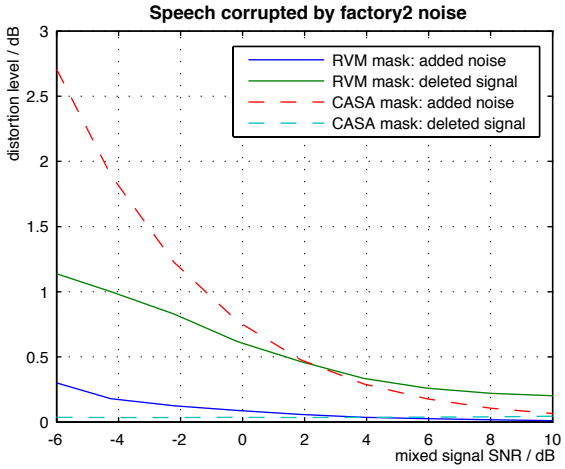

Figure 4: Amount of noise energy added by false positive mask cells and the amount of signal energy deleted by false negative mask cells.

hard mask reconstruction results are omitted since they were very similar to those of RVM hard masks.

GMM reconstruction performs better than simple masked refiltering. The exception to this is refiltering with the ground truth mask at higher SNRs where less data is missing and the GMM reconstruction fills in cells with energy that exceeds that of the clean speech signal. In all other cases refiltering performs worst since it leaves gaps in the signal wherever noise obscures the speech signal, and MMSE inference puts energy in these gaps that is at least somewhat closer to the original. However, for the CASA masks, the difference between refiltering and reconstruction is very small because in many cases the pattern of present-data returned by the pitch tracker, which included many falsely-accepted noisy dimensions, returned no meaningful inference from the GMMs.

Finally, it is clear that the use of soft masks where applicable gives approximately a 1 dB improvement over the same reconstruction method using hard masks across all SNRs. However, there is still room for improvement in mask estimation as evidenced by the big gap in reconstruction SNR between the use of ground truth masks and RVM masks. Part of this is due to the fact that time and memory constraints limited the amount of data that could be used to train the RVMs.

It is important to note that reconstruction SNR is not necessarily the best evaluation metric. Much of the noise present in the reconstructed signal is due to mismatches between the signal model and the actual clean signal, not to the presence of noise in sections of the signal where there is no speech present. This is especially true when the mixed signal is at higher SNRs. The exceptions to this are instances when the mask mistakenly labels noise dominated cells as reliable.

Figs. 3 and 4 break the mask errors down into false accept/insertion errors where the mask mistakenly labels noise-dominated cells as reliable and false reject/deletion errors where the mask mistakenly labels speech-dominated cells as noise. The false positive rate of the pitch tracker mask is much higher than that of the RVM mask. This is a result of the fact that the pitch tracker masks tend to be very inaccurate at high frequencies.

Fig. 6 compares the mutual information between the ground-truth STFT cell labels and the masks based on RVM classifier and CASA pitch tracker. Mutual information is somewhat independent of the false alarm / false reject tradeoff, to allow a comparison that does not depend so strongly on threshold. The RVM mask is significantly more informative about the true mask label than the CASA-based pitch track mask, but the joint MI between the ground truth and both masks is higher still, indicating that there is some information in the CASA mask not capture by the RVM, and hence there could be some value in combining them.

From fig. 5 it is clear that performance is best on the same kind of noise signals that were used to train the RVM classifiers. Despite this there is a clear a boost in SNR on all noise signals when the mixed signal is below 8 dB SNR using RVM masks. The worst performance occurs on the music noise. The estimated RVM masks on this signal are often wrong because it is highly nonstationary with highly harmonic sections, unlike any of the signals used to train the RVM.

Fig. 7 shows specific examples of the mask estimation and different types of reconstruction. Problems with RVM mask prediction are evidenced by the false negatives in the first 0.5 s. When the masks are wrong, there is no way for the MMSE reconstruction to properly recover the missing data. Even though MMSE reconstruction does not give a huge boost in SNR, it does much to fill in the blanks (e.g. in the vowel at about 1 second). As noted earlier, the biggest failing of the CASA mask generation lies in the prevalence of false positives. When a lot of noisy data is labelled as being reliable, the MMSE reconstruction is unable to get a good estimate of the underlying speech signal. We note in passing that the pitch-track based CASA mask has no way to identify correct masks for unpitched speech sounds (fricatives), limiting its potential performance.

We also note in passing that the CASA pitch tracking system fares much better in the low frequency regions below about 1 kHz where pitch harmonics are strongest. Our measures such as detection rate and mutual information count individual STFT cells of fixed bandwidth; a division of time-frequency using a more perceptual frequency axis (e.g. Mel or Bark scale) would increase the relative significance of these low-frequency bands, and would show the CASA system in a more favorable light.

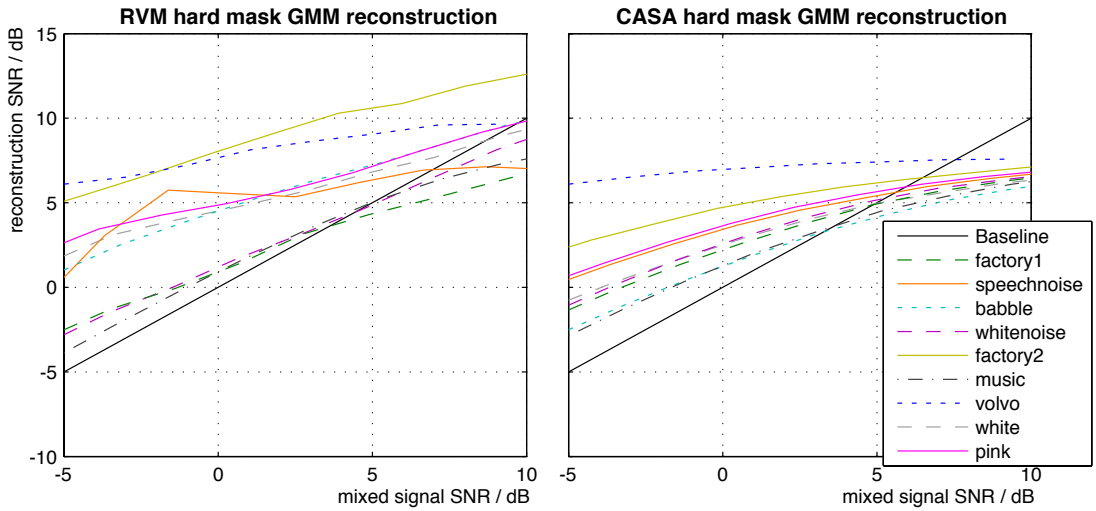

Figure 5: SNR of GMM MMSE reconstruction using missing data masks versus SNR of mixed signal.

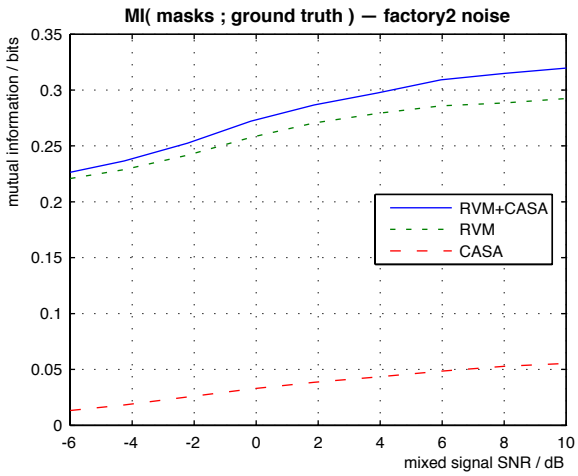

Figure 6: Mutual information between the generated missing data masks and ground truth masks.

6. Conclusions

A system for inferring a clean speech signal from a noisy signal that does not depend on explicit noise models was presented. RVM classifiers were evaluated and shown to have a clear advantage over SVMs in terms of model sparsity, without a large cost in accuracy. Sparsity has a large effect on computational complexity of actual classification since the run time scales with the number of support vectors.

The performance of RVM masks was also shown to be superior to that of masks generated by a pitch tracking CASA approach. Poor mask estimation where many noisy cells are labelled as being reliable, as is the abundance of false positives using the pitch tracker mask, poses significant problems to the feature reconstruction process. Because of this, the false negative errors made by the RVM mask are actually less detrimental than the false positives made by the pitch tracker mask.

The biggest drawback to this system is the computational complexity of the RVM training algorithm. The amount of data used to train the RVMs was limited since the run time of the training algorithm is cubic in the number of training examples. Use of the fast training algorithm described in [11] would mitigate this.

Our analysis showed large differences between the RVM-based masks and masks from a traditional CASA pitch-tracking system. However, although the RVM system was superior, the mutual information results indicate that there is benefit to be had by combining both systems. One natural approach to this would be to include pitch-related information as features for the RVM classifier.

Finally, as hinted at in [4], the subband classifiers might be able to generalize better across different types of interference if they used features that are less dependant on the type of noise. These might include broad spectral shape features such as spectral flatness and spectral centroid or perceptually motivated features such as MFCCs.

7. Acknowledgments

Many thanks to Keansub Lee who implemented the noisy pitch tracker used to generate the CASA masks. This work is supported by the National Science Foundation (NSF) under Grants No. IIS-0238301 and IIS-05-35168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

8. References

[1] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

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

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

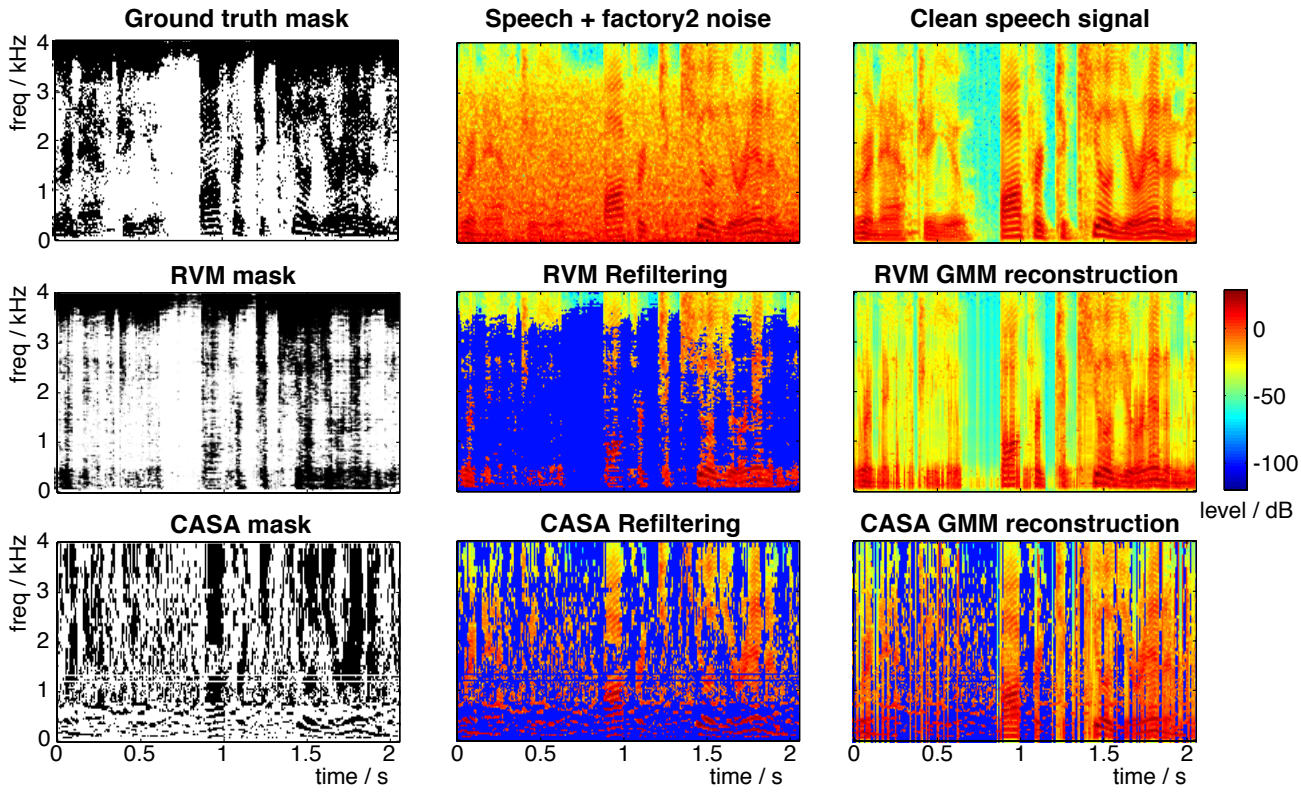

Figure 7: Example of refiltering and GMM MMSE reconstruction of speech corrupted by factory noise ($\sim$ .2 dB SNR) using both RVM and CASA-based masks. Despite the fact that the RVM mask does not entirely match the ground truth a priori mask, it filters out most of the interference. Simple refiltering improves the SNR by about 7 dB. MMSE reconstruction helps even more by filling in the missing parts of the refiltered signal, yielding an overall SNR increase of about 8.5 dB. The pitch tracker does a reasonable job tracking the speech harmonics that appear above the noise floor, but it also adds a lot of noisy cells. The MMSE reconstruction is particularly poor in this case due to the false positive errors in the missing data mask.

[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] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proceedings of ICASSP*, 2004.

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

[7] 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*, November 2005, pp. 27–32.

[8] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[9] M. Wu, D.L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.

[10] 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.

[11] M. E. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse bayesian models," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.