SPEECH DENOISING USING NONNEGATIVE MATRIX FACTORIZATION WITH PRIORS

Kevin W. Wilson[*]*, Bhiksha Raj*[*]*, Paris Smaragdis*[†]*, Ajay Divakaran*[*]

[*] Mitsubishi Electric Research Lab, Cambridge, MA
[†] Adobe Systems, Newton, MA

ABSTRACT

We present a technique for denoising speech using nonnegative matrix factorization (NMF) in combination with statistical speech and noise models. We compare our new technique to standard NMF and to a state-of-the-art Wiener filter implementation and show improvements in speech quality across a range of interfering noise types.

Index Terms— Speech enhancement, Speech processing

1. INTRODUCTION

This paper presents a regularized version of nonnegative matrix factorization (NMF) and demonstrates its usefulness for the denoising of speech in nonstationary noise. Speech denoising in nonstationary noise is an important problem with increasingly broad applications as cellular phones and other telecommunications devices make electronic voice communication more common in a wide range of challenging environments, from urban sidewalk to construction site to factory floor. Standard approaches such as spectral subtraction and Wiener filtering require signal and/or noise estimates and therefore are typically restricted to stationary or quasi-stationary noise in practice.

Nonnegative matrix factorization, popularized by Lee and Seung [1], finds a locally optimal choice of W and H to solve the matrix equation $V \approx WH$. This provides a way of decomposing a signal into a convex combination of nonnegative building blocks. When the signal, V, is a spectrogram and the building blocks, W, are a set of specific spectral shapes, Smaragdis [2] showed how NMF can be used to separate single-channel mixtures of sounds by associating different sets of building blocks with different sound sources. In Smaragdis's formulation, H becomes the time-varying activation levels of the building blocks. The building blocks in W constitute a model of each source, and because H allows activations to vary over time, this decomposition can easily model nonstationary noises. ([2] refers to its algorithm as probabilistic latent semantic analysis (PLSA). Under proper normalization and for the KL objective function used in this paper, NMF and PLSA are numerically equivalent [3], so the results in [2] are equally relevant to NMF or PLSA.)

NMF works well for separating sounds when the building blocks for different sources are sufficiently distinct. For example, if one source, such as a flute, generates only harmonic sounds and another source, such as a snare drum, generates only nonharmonic sounds, the building blocks for one source will be of little use in describing the other. In many cases of practical interest, however, there is much less separation between sets of building blocks. In particular, human speech consists of harmonic sounds (possibly at different fundamental frequencies at different times) and nonharmonic sounds, and it can have energy across a wide range of frequencies. For these reasons, many interfering noises can be represented, at least partially, by the speech building blocks. In a speech denoising application, where one "source" is the desired speech and the other "source" is interfering noise, this overlap between speech and noise models will degrade performance.

There is additional structure in speech and many other sounds, however. For example, a human speaker will never generate a simultaneous combination of two harmonic sounds with harmonically unrelated pitches. Using the standard NMF reconstruction, however, this combination would be allowed. By enforcing what we know about the co-occurrence statistics of the basis functions for each source, we can potentially improve the performance of NMF.

This paper makes two contributions. First, we present a regularized version of NMF that encourages the denoised output signal to have statistics similar to the known statistics of our source model. Second, we evaluate the speech denoising performance of NMF and regularized NMF and compare them to a state-of-the-art Wiener filter implementation.

2. ALGORITHM

Our technique for speech denoising consists of a training stage and an application (denoising) stage. During training, we assume availability of a clean speech spectrogram, V_{speech}, of size $n_f \times n_{st}$, and a clean (speech-free) noise spectrogram, V_{noise}, of size $n_f \times n_{nt}$, where n_f is the number of frequency bins, n_{st} is the number of speech frames, and n_{nt} is the number of noise frames. Different objective functions lead to different variants of NMF, a number of which are described in [4]. Kullback-Leibler (KL) divergence between V and WH, denoted $D(V||WH)$, was found to work well for audio source separation in [2], so we will restrict ourselves to KL divergence in this paper. Generalization to other objective functions using the techniques described in [4] is straightforward.

During training, we separately perform standard NMF on the speech and the noise, minimizing $D(V_{speech}||W_{speech}H_{speech})$ and $D(V_{noise}||W_{noise}H_{noise})$, respectively. W_{speech} and W_{noise} are each of size $n_f \times n_b$, where n_b is the number of basis vectors chosen to represent each source. Each column of W is therefore one of the spectral "building blocks" we referred to earlier. H_{speech} and H_{noise} are of size $n_b \times n_{st}$ and $n_b \times n_{nt}$, respectively, and represent the time-varying activation levels of the basis vectors.

Also as part of the training phase, we estimate the statistics of H_{speech} and H_{noise}. Specifically, we compute the empirical means and covariances of their log values, yielding μ_{speech}, μ_{noise}, Λ_{speech}, and Λ_{noise} where each μ is a length n_b vector and each Λ is an $n_b \times n_b$ covariance matrix. We choose this implicitly Gaussian representation for computational convenience, and we choose to operate in the logarithmic domain because preliminary experiments showed better results for the log domain than the linear do-

Paris Smaragdis performed this work while working at Mitsubishi Electric Research Lab

main. This is consistent with the fact that the nonnegative constraint on H means that a Gaussian, which has support for both positive and negative values, will probably be a poor fit to the true distribution.

In the denoising stage, we fix W_{speech} and W_{noise} and assume that they will continue to be good basis functions for describing speech and noise. We concatenate the two sets of basis vectors to form W_{all} of size $n_f \times 2n_b$. This combined set of basis functions can then be used to represent a signal containing a mixture of speech and noise. Assuming the speech and noise are independent, we also concatenate to form $\mu_{all} = [\mu_{speech}; \mu_{noise}]$ and $\Lambda_{all} = [\Lambda_{speech}\, 0; 0\, \Lambda_{noise}]$. A main contribution of this paper is then to find an H_{all} to minimize the following regularized objective function:

$$D_{reg}(V||WH) = \sum_{ik}(V_{ik}\log\frac{V_{ik}}{(WH)_{ik}} + V_{ik} - (WH)_{ik}) - \alpha L(H) \quad (1)$$

$$L(H_{all}) = -\frac{1}{2}\sum_k\{(\log H_{all_{ik}} - \mu_{all})^{\mathsf{T}}\Lambda_{all}^{-1}(\log H_{all_{ik}} - \mu_{all}) - \log[(2\pi)^{2n_b}|\Lambda|]\} \quad (2)$$

When α is zero, this is equal to the standard KL divergence objective function [4]. For nonzero α, there is an added penalty proportional to the negative log likelihood under our jointly Gaussian model for $\log H$. This term encourages the resulting H_{all} to be consistent with the statistics of H_{speech} and H_{noise} as empirically determined during training. Varying α allows us to control the tradeoff between fitting the observed spectrogram of mixed speech and noise, V_{mix} and achieving high likelihood under our prior model. Following [4], the multiplicative update rule for H_{all} is

$$H_{all_{a\mu}} \leftarrow H_{all_{a\mu}}\frac{\sum_i W_{all_{ia}}V_{mix_{i\mu}}/(W_{all}H_{all})_{i\mu}}{[\sum_k W_{all_{ka}} + \alpha\varphi(H_{all})]_\varepsilon} \quad (3)$$

$$\varphi(H_{all_{a\mu}}) = -\frac{\partial L(H_{all})}{\partial H_{all_{a\mu}}} = -\frac{(\Lambda_{all}^{-1}\log H_{all})_{a\mu}}{H_{all_{a\mu}}}$$

where $[\,]_\varepsilon$ indicates that any values within the brackets less than the small positive constant ε should be replaced with ε to prevent violations of the nonnegativity constraint and avoid divisions by zero.

Finally, to reconstruct the denoised spectrogram, we compute $\hat{V}_{speech} = W_{speech}H_{all_{1:n_b}}$, using the speech basis functions and the top n_b rows of H_{all} to approximate the target speech.

Figure 1 gives a simple toy example of separating with and without a prior distribution. Here we set $n_f = n_b = 2$, and we assume that for both speech and noise, one basis function represents the high frequency and the other represents the low frequency. The original signals are in the left column, the unregularized NMF reconstructions are in the center column, and the regularized NMF reconstructions are in the right column. Because the basis functions for the speech and noise are the same, unregularized NMF is completely unable to reconstruct the individual sources. Note however that its chosen reconstructions do sum to accurately model the mixture signal, indicating that it successfully minimized $D(V||WH)$. The regularized NMF is able to exploit the fact that high and low frequencies are perfectly correlated in Source 1 and negatively correlated in Source 2 to accurately reconstruct the two signals given only the mixture signal and their statistical models. This example is extreme in that the "speech" and "noise" bases are identical while their statistics are quite different, but it makes the potential of the approach clear. We show in the following section that incorporating this regularizing prior term does improve speech denoising in practice.

3. RESULTS

We tested NMF and regularized NMF on a variety of speakers and with four different types of nonstationary background noise (jackhammer noise, bus/street noise, combat noise, and speech babble noise). All parameters remained at fixed values across all experiments. We used 16 kilohertz audio with $n_f = 513$, $n_b = 80$, and $\alpha = 0.25$. (The numerical value of α is meaningless without knowing the magnitude of the spectrogram values, but we want to emphasize that α remained fixed throughout.) We used speech from the TIMIT database [6], testing two sentences from each of ten speakers in each of our four chosen types of background noise. We normalized speech and noise so that the average signal-to-noise ratio (SNR) for each mixture was 0 dB.

We trained a separate noise model for each of the four noise types, and we trained two different types of speech model. One model, which we call the "group" model, was trained on a mixed group of male and female speakers, none of which were in our test set. This single model was then used to denoise noisy signals from a variety of test speakers. For the other type of model, the "self" model, we train a speaker-specific model for each speaker in the test set using sentences from outside the test set. Comparing "group" to "self" lets us see how much we gain from speaker-specific models.

Our results are shown in Figure 2. All results are shown as improvement relative to the score of the unprocessed 0 dB SNR mixture, and each bar represents an average value over ten speakers. To quantify our results, we use segmental SNR, a simple metric which has been found to correlate reasonably well with perceived quality [7], and the ITU Perceptual Evaluation of Speech Quality (PESQ) [8], a more sophisticated metric specifically designed to match mean opinion scores of perceptual quality. PESQ scores range from 1 through 5, and PESQ improvements on the order of 0.5, which we achieve in many cases, are quite noticeable.

In addition to NMF and regularized NMF, we processed each example with the ETSI Aurora front end's Wiener filtering [5], a European telecommunications standard which has been carefully tuned for good performance in denoising speech. It is important to note that, in contrast to the ETSI Wiener filter, all of our NMF variants use both a training and a testing stage, so they benefit from environment-specific noise models. They have also been specifically designed to work on nonstationary noise. The ETSI Wiener filter has no training stage, so its noise model must be estimated online using a voice activity detector and assumptions about the stationarity of the noise. However, the ETSI Wiener filter has an advantage as long as its voice activity detector works properly because it can then completely silence intervals with no speech activity, yielding very good denoising in those intervals. Because of the major differences between the two types of denoising, detailed comparisons of the results are of limited use, but we feel that it is important to compare to an established baseline and that some general conclusions are possible. The PESQ scores for both regularized and unregularized NMF are almost always greater than for the ETSI Wiener filter, and in many cases are substantially greater. Segmental SNR results are not as impressive

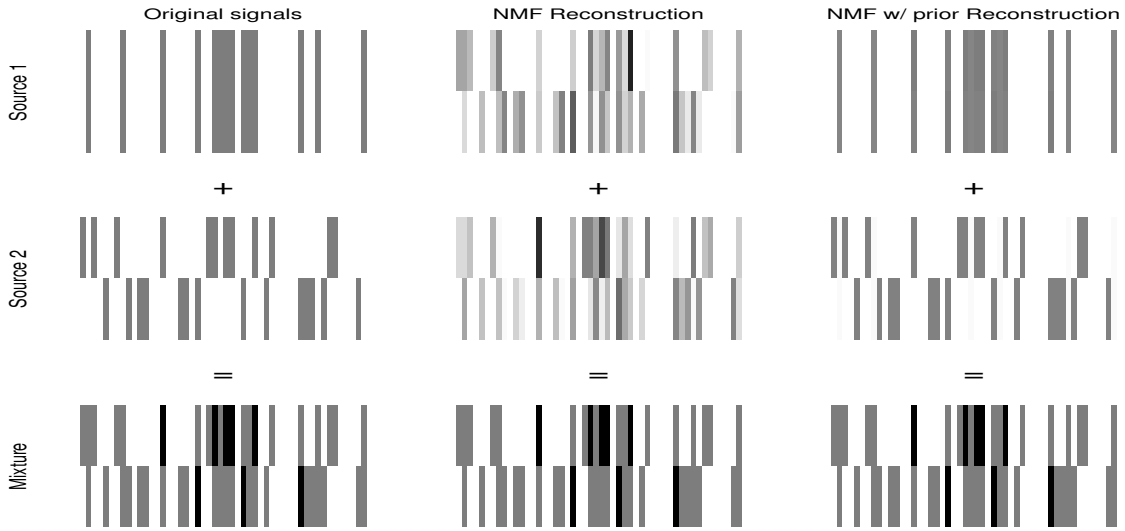
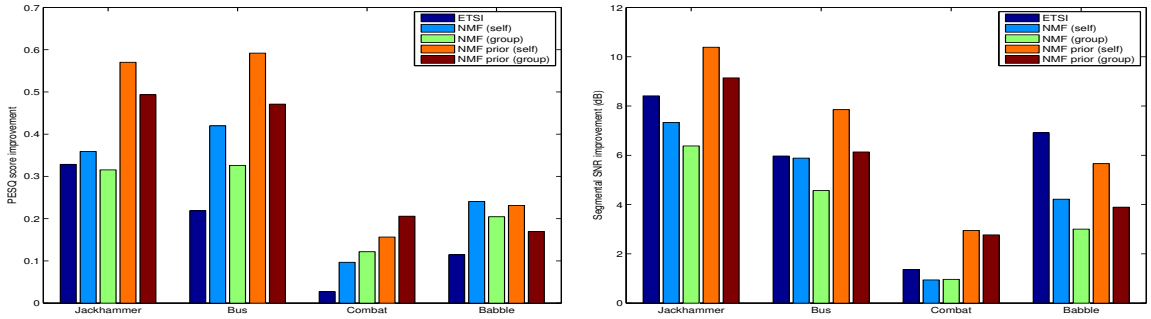
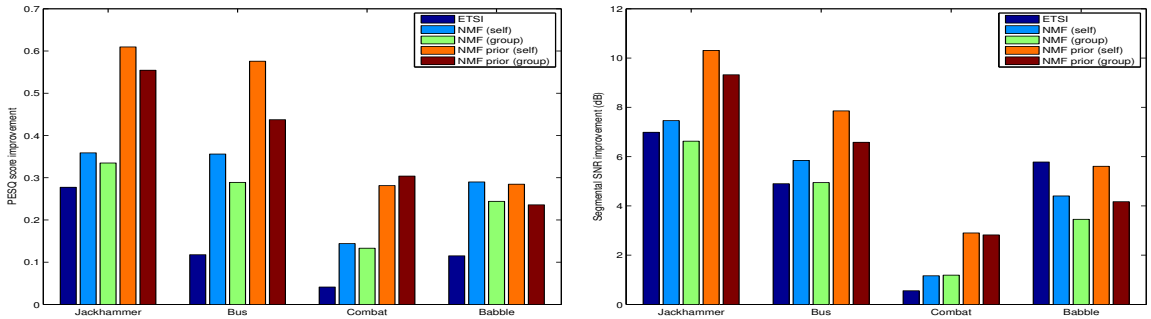


Fig. 1. A toy example showing the advantage of regularizing with the log likelihood. In each panel, the horizontal axis represents time and the vertical axis represents frequency. Darker colors represent higher intensity. The leftmost column shows the original signals. For source 1, high frequencies and low frequencies are perfectly correlated. For source 2, high frequencies and low frequencies are negatively correlated. In the middle column, unregularized NMF finds a reconstruction that perfectly models the mixture signal, but each individual source is poorly reconstructed. In the rightmost column, near-perfect reconstruction of individual sources is achieved by enforcing the prior.



(a) PESQ Male

(b) Segmental SNR Male



(c) PESQ Female

(d) Segmental SNR Female

Fig. 2. Speech denoising performance for our chosen noise types. "ETSI" is the front end Wiener filtering described in [5]. "NMF" is applying the iterative update in Equation 3 with $\alpha = 0$. "NMF prior" is applying Equation 3 with $\alpha = 0.25$. "Self" denotes results for speaker-specific speech models. "Group" denotes results for a non-speaker-specific speech model trained on a mixed gender group.

compared to ETSI, but even there regularized NMF is almost always superior.

Overall, there is little difference between the results for male speakers and female speakers. In particular, the fact that this holds true for the "group" speech model suggests that the mixed-gender group model is not strongly biased toward either gender. Next, note that the regularized NMF results are almost always substantially better than the corresponding unregularized NMF results. This shows that the additional structure imposed by the prior consistently improves the denoising performance across a variety of background noises. The "self" models almost always outperform the "group" models, which is not surprising since they are more specifically targeted to each individual. However, it is interesting to note that regularized NMF with the "group" model almost always outperforms unregularized NMF, even when the unregularized NMF uses the speaker-specific "self" basis functions.

The aforementioned trends are relatively consistent across three of the four noise types, but performance on "babble" noise departs from these trends, especially as measured by segmental SNR. For "babble," it appears that regularization is not as helpful, and the ETSI Wiener filter outperforms all NMF variants in segmental SNR. We speculate that one reason for the impressive relative performance of the Wiener filter is that the babble noise is the closest of the four to stationary noisy, in the sense that it is a near constant drone of indistinct speech-like noise. It is possible that the priors are not as useful for babble because the prior for the speech-like babble noise is very similar to the prior for speech itself, although further analysis is needed to confirm this hypothesis.

4. CONCLUSION

We have shown that NMF can be used to denoise speech in the presence of nonstationary noise, and we have shown that by regularizing NMF based on a prior model of speech and noise, we can exploit additional signal structure to improve performance. Our results equal or surpass results from a state-of-the-art Wiener filter implementation on a range of noise types.

There are a number of interesting directions for future work. This work complements work on explicit control of sparseness for source-separation [9], and combining the two approaches may improve results further. Determining a useful way of incorporating regularization into the training stage and/or incorporating temporal dynamics could also improve performance. Finally, we would like to better characterize how well NMF and regularized NMF can be expected to work on a given problem, presumably depending on the distinctness of the basis sets and the divergences between prior distributions.

5. REFERENCES

[1] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[2] P. Smaragdis, "From learning music to learning to separate," in *Forum Acusticum*, 2005.

[3] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[4] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, vol. 5, pp. 621–625.

[5] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050 V1.1.3, 2003.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom documentation," Tech. Rep. PB93-173938, National Technical Information Service, 1993.

[7] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective measures of speech quality*, Prentice Hall, 1988.

[8] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.

[9] M. Shashanka, B. Raj, and P. Smaragdis, "parse overcomplete decomposition for single channel speaker separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, vol. 2, pp. 641–644.