Joint Filtering and Factorization for Recovering Latent Structure from Noisy Speech Data

Colin Vaz, Vikram Ramanarayanan, and Shrikanth Narayanan

Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089
<cvaz,vramanar>@usc.edu, shri@sipi.usc.edu

Abstract

We propose a joint filtering and factorization algorithm to recover latent structure from noisy speech. We incorporate the minimum variance distortionless response (MVDR) formulation within the non-negative matrix factorization (NMF) framework to derive a single, unified cost function for both filtering and factorization. Minimizing this cost function jointly optimizes three quantities – a filter that removes noise, a basis matrix that captures latent structure in the data, and an activation matrix that captures how the elements in the basis matrix can be linearly combined to reconstruct input data. Results show that the proposed algorithm recovers the speech basis matrix from noisy speech significantly better than NMF alone or Wiener filtering followed by NMF. Furthermore, PESQ scores show that our algorithm is a viable choice for speech denoising.

Index Terms: NMF, MVDR, denoising, filtering.

1. Introduction

Observation of latent structure in data provides researchers with a tool for data analysis and interpretation. Non-negative matrix factorization (NMF) is a widely-used method for observing the latent structure in a signal of interest. First proposed by Paatero and Tapper [1, 2] and developed further by Lee and Seung [3], NMF has been employed in a variety of areas, from analyzing molecular structure [4] to speech enhancement [5, 6]. The drawback to NMF is that it is sensitive to outliers in the data, because the NMF formulation minimizes the *square* residual error. Researchers have proposed several techniques to overcome this drawback. Kong et al. derived update equations that minimize the $L_{2,1}$ norm rather than the Frobenius norm, and achieved better image clustering results using the modified metric [7]. Another approach is to induce sparsity on the activation matrix to control the number of basis elements that are simultaneously activated, which can reduce the influence of outlier in the data in the factorization process [8, 9]. We propose a method to filter the data during the factorization process to try to overcome outliers and noise in the data.

We model the filter on the minimum variance distortionless response (MVDR) filter. This filter was first proposed by Capon for beamforming in array signal processing [10], and then later adapted for spectral estimation by Rao et al. [11, 12]. The MVDR filter computes a power spectrum that estimates the spectral envelope of a signal with the property that it does not distort the spectrum. It does so by computing a bank of filters, each of which try to pass a specific frequency of the signal undistorted while suppressing the output at other frequencies. This formulation leads to a smoother estimate of the spectrum that is less sensitive to noise compared to the Fourier transform of the signal. This is a desirable property for improving the performance of NMF when factoring noisy data. Thus, we will rewrite the MVDR formulation and use it in the NMF framework to perform filtering of the noisy data during the factoring operation.

The paper is organized as follows. Section 2 describes the NMF and MVDR algorithms and describes our proposed approach to combine these two methods to achieve joint filtering and factorization of a noisy signal. Section 3 compares the performance of the proposed algorithm to NMF as well as NMF of a Wiener-filtered input. In Section 4, we discuss our experiments and point out the conditions in which our algorithm performs well and where it fails. Finally, we state our conclusions and future work in Section 5.

2. Joint filtering–factorization formulation

NMF factors a $M \times N$ non-negative matrix V into a $M \times K$ non-negative basis matrix W and $K \times N$ non-negative activation matrix H by minimizing $\|V - WH\|_F^2$. Since all the matrices are non-negative, WH is a parts-based representation of V that discovers latent structure in the data. The columns of W contain the fundamental units of this structure while H describes the level of activation for each of these fundamental units. For example, if V is a spectrogram of speech, then W will capture the phones in that speech, while H tells how much each phone was activated during the speech segment. However, if there is noise in the speech, then it is possible for some columns of W to capture properties of the noise, which can obscure the speech structure.

To overcome noise in the data, we design a filter modeled on the MVDR filter because it has the desirable property of preserving salient peaks in the spectrum. Thus, it can be used to preserve speech, which has spectral peaks at the formant frequencies. Vaz et al. showed in [13] that you can relax the distortionless constraint to improve the spectral estimation of vowels in noisy conditions. The relaxed constraint prevents the MVDR filter from preserving peaks outside the frequency range of interest. We aim to incorporate the MVDR formulation within the NMF framework to perform joint filtering and factorization of noisy data. More specifically, we will use the data in V to derive a set of filters that optimally estimates the spectrum of the desired signal corrupted with noise, and use this filtered data to calculate an improved basis matrix estimate. The MVDR formula for a filter $\boldsymbol{g}_k$ that passes a frequency ω_k undistorted is

$$\hat{\boldsymbol{g}}_k = \arg\min_{\boldsymbol{g}_k} \boldsymbol{g}_k^H R \boldsymbol{g}_k \quad \text{s.t.} \quad \left|\boldsymbol{e}^H(\omega_k)\boldsymbol{g}_k\right| = 1, \qquad (1)$$

Work supported by NIH Grant DC007124.

where R is a Toeplitz autocorrelation matrix of the input signal $x[n]$ and $\boldsymbol{e}(\omega_k) = \left[1 \; e^{j\omega_k} \cdots e^{j\omega_k(N-1)}\right]^T$. By relaxing the distortionless constraint, as in [13], we can rewrite this as:

$$\hat{\boldsymbol{g}}_k = \arg\min_{\boldsymbol{g}_k} \boldsymbol{g}_k^H R \boldsymbol{g}_k \quad \text{s.t.} \quad \left|\boldsymbol{e}^H(\omega_k)\boldsymbol{g}_k\right| = \alpha_k, \tag{2}$$

where α_k is the desired frequency response at the kth frequency. Using Parseval's Theorem, we can write a roughly equivalent formulation in the frequency domain as

$$\hat{\boldsymbol{G}}_k = \arg\min_{\boldsymbol{G}_k} \|\boldsymbol{G}_k \otimes \boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad G_{kk}X_k = \alpha_k X_k, \tag{3}$$

where $\boldsymbol{G}_k$ is the frequency response of $\boldsymbol{g}_k$, G_{kk} is the value of the frequency response of $\boldsymbol{G}_k$ at the kth frequency, $\boldsymbol{X}$ is the frequency response of the input data, and $\otimes$ denotes element-wise multiplication. Equation 3 computes the frequency response that has a pre-determined fixed value at the kth frequency and has minimal amplitude at the other frequencies. For maximum minimization, the frequency response that the other frequencies should be 0. To jointly compute a bank of filters, we solve

$$\hat{\boldsymbol{G}} = \arg\min_{\boldsymbol{G}} \|\boldsymbol{G} \otimes \boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{G} \otimes \boldsymbol{X} = \boldsymbol{A} \otimes \boldsymbol{X}, \tag{4}$$

where $\boldsymbol{A} = [\alpha_1 \;\; \alpha_2 \cdots \alpha_M]^T$ is a vector of the desired frequency response for all frequencies. To achieve joint filtering and factoring, we incorporate Equation 4 in the NMF framework. We formulate the cost function as

$$\begin{aligned} J = \|G \otimes V - WH\|_F^2 + \lambda_1 \|G \otimes (WH)\|_F^2 + \\ \lambda_2 \|G \otimes (WH) - A \otimes (WH)\|_F^2. \end{aligned} \tag{5}$$

where V is the spectrogram of the input speech data. The first term in the cost function performs NMF on the filtered input data, λ_1 controls the level of filtering, and λ_2 controls the extent to which G is constrained by the desired frequency response A.

2.1. Update equations

Computing the gradient of J with respect to G and setting it to zero allows us to obtain a closed-form solution for the filter G:

$$G = \frac{(WH) \otimes V + \lambda_2 A \otimes (WH)^2}{V^2 + (\lambda_1 + \lambda_2)(WH)^2} \tag{6}$$

where the division and square operators are element-wise. Since G depends on W and H, the filter is updated at every iteration during the algorithm. The iterative update equations for the basis matrix W and time-activation matrix H are as follows:

$$\begin{aligned} W_{ab} &\leftarrow W_{ab} + \eta_{ab}\frac{\partial J}{\partial W_{ab}} \\ &\leftarrow W_{ab} + \eta_{ab}\Big(\sum_{j=1}^{N} G_{aj}V_{aj}H_{bj} - W_{ab}\sum_{j=1}^{N} H_{bj}^2 - \\ &\lambda_1 W_{ab}\sum_{j=1}^{N} G_{aj}^2 H_{bj}^2 - \lambda_2 W_{ab}\sum_{j=1}^{N}(G_{aj} - A_{aj})^2 H_{bj}^2\Big) \\ G_{ab} &\leftarrow G_{ab} + \gamma_{ab}\frac{\partial J}{\partial G_{ab}} \\ &\leftarrow G_{ab} + \gamma_{ab}\Big(\sum_{i=1}^{M} G_{ib}V_{ib}W_{ia} - H_{ab}\sum_{i=1}^{M} W_{ia}^2 - \\ &\lambda_1 H_{ab}\sum_{i=1}^{M} G_{ib}^2 W_{ia}^2 - \lambda_2 H_{ab}\sum_{i=1}^{M}(G_{ib} - A_{ib})^2 W_{ia}^2\Big) \end{aligned} \tag{7}$$

By setting

$$\begin{aligned} \eta_{ab} &= \frac{W_{ab}}{\begin{aligned} W_{ab}\sum_{j=1}^{N} H_{bj}^2 + \lambda_1 W_{ab}\sum_{j=1}^{N} G_{aj}^2 H_{bj}^2 + \\ \lambda_2 W_{ab}\sum_{j=1}^{N}(G_{aj} - A_{aj})^2 H_{bj}^2 \end{aligned}} \\ \gamma_{ab} &= \frac{H_{ab}}{\begin{aligned} H_{ab}\sum_{i=1}^{M} W_{ia}^2 + \lambda_1 H_{ab}\sum_{i=1}^{M} G_{ib}^2 W_{ia}^2 + \\ \lambda_2 H_{ab}\sum_{i=1}^{M}(G_{ib} - A_{ib})^2 W_{ia}^2 \end{aligned}}, \end{aligned} \tag{8}$$

one can obtain multiplicative updates for W and H as

$$\begin{aligned} W &\leftarrow W \otimes \frac{(G \otimes V)H^T}{WHH^T + \lambda_1 W \otimes C + \lambda_2 W \otimes D} \\ H &\leftarrow H \otimes \frac{W^T(G \otimes V)}{W^T WH + \lambda_1 E \otimes H + \lambda_2 F \otimes H} \end{aligned} \tag{9}$$

where the division is element-wise, and

$$C_{ab} = (\boldsymbol{G}_{a,:} \otimes \boldsymbol{H}_{b,:})(\boldsymbol{G}_{a,:} \otimes \boldsymbol{H}_{b,:})^T \tag{10}$$

$$D_{ab} = ((\boldsymbol{G}_{a,:} - \boldsymbol{A}_{a,:}) \otimes \boldsymbol{H}_{b,:})((\boldsymbol{G}_{a,:} - \boldsymbol{A}_{a,:}) \otimes \boldsymbol{H}_{b,:})^T \tag{11}$$

$$E_{ab} = (\boldsymbol{W}_{:,a} \otimes \boldsymbol{G}_{:,b})^T(\boldsymbol{W}_{:,a} \otimes \boldsymbol{G}_{:,b}) \tag{12}$$

$$F_{ab} = (\boldsymbol{W}_{:,a} \otimes (\boldsymbol{G}_{:,b} - \boldsymbol{A}_{:,b}))^T(\boldsymbol{W}_{:,a} \otimes (\boldsymbol{G}_{:,b} - \boldsymbol{A}_{:,b})) \tag{13}$$

2.2. Alternative interpretation of G

Recall that our input data matrix, $\mathbf{V}$, is computed as the spectrogram of the input speech signal $x[n]$, or in other words, the magnitude-squared of the discrete short-term Fourier transform (STFT) of $x[n]$, which may be expressed as

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]\mathrm{e}^{-j\omega n} \tag{14}$$

where $w[n-m]$ is a L-length window function that "selects" L samples of $x[n]$ at shift m to be Fourier-transformed. If $x[n]$ is a noisy signal, then we can compute a filter $g_m[n]$ at each shift m to optimally filter the noise for that windowed segment. Thus, we can filter the windowed $x[n]$ at shift m with $g_m[n]$ to get a denoised signal $y_m[n]$. The discrete-time Fourier transform (DTFT) of $y_m[n]$ is

$$Y(m, \omega) = \sum_{n=-\infty}^{\infty} y_m[n]\mathrm{e}^{-j\omega n} \tag{15}$$

$$= \sum_{n=-\infty}^{\infty} \left(x[n]w[n-m] * g_m[n]\right)\mathrm{e}^{-j\omega n}, \tag{16}$$

Using Parseval's Theorem, we can rewrite the right hand side of Equation 15 as

$$Y(m, \omega) = X(m, \omega)G(m, \omega), \tag{17}$$

where $G(m, \omega)$ is the DTFT of $g_m[n]$. The right hand side of Equation 17 is essentially the same as what is expressed in Equation 4. Therefore, one can think of Equation 4 as computing an optimal filter $g_m[n]$ for a windowed segment of the noisy signal.

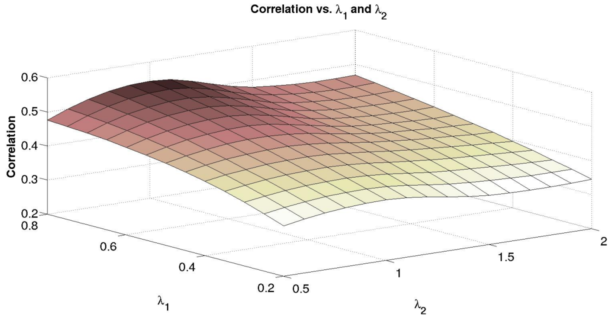


Figure 1: Mean correlation for different values of λ_1 and λ_2 with $K = 40$. This plot was generated with bicubic interpolation.

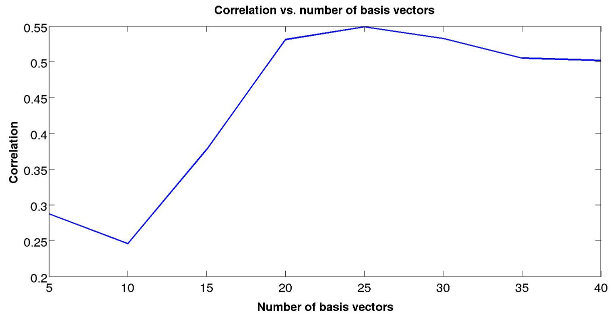


Figure 2: Mean correlation for different values of K with $\lambda_1 = 0.8$ and $\lambda_2 = 1$.

3. Experiments and Results

3.1. Parameter settings

We first optimized the parameters λ_1 and λ_2 and the number of basis vectors K. We randomly sampled 40 sentences from the MOCHA-TIMIT corpus [16], with 20 sentences spoken by male speakers and the other 20 sentences spoken by female speakers. The sentences are recorded in clean conditions. We added white noise to these sentences at 5 dB SNR level. We used our algorithm, with K fixed to 40[1], $\lambda_1 \in \{0.2, 0.4, 0.6, 0.8\}$, and $\lambda_2 \in \{0.5, 1, 1.5, 2\}$, to obtain basis matrices for the clean and noisy signals. We calculated the correlation between the basis matrix of the clean signal and the basis matrix of its corresponding noisy version to find the optimum λ_1 and λ_2. Figure 1 shows the mean of the correlations for each combination of λ_1 and λ_2. The correlation between the clean and noisy basis matrices is calculated using

$$\rho = \frac{1}{K} \sum_{k=1}^{K} \frac{|\boldsymbol{W}_{\text{clean}}^T(:,k)\boldsymbol{W}_{\text{noisy}}(:,k)|}{\|\boldsymbol{W}_{\text{clean}}(:,k)\|_2 \|\boldsymbol{W}_{\text{noisy}}(:,k)\|_2}. \tag{18}$$

Equation 18 computes the mean of the cosine of the angle between a vector in the clean basis and the corresponding vector in the noisy basis. The correlation measure ranges from 0 to 1, with higher values indicating better correlation between the clean basis and noisy basis. Hence, the correlation is a measure of the basis recovery of the noisy signal compared to the clean signal. We sorted the columns of the basis matrices in ascending order of center of gravity prior to computing the correlation to make the computation meaningful. Using the optimum λ_1 and λ_2, we re-ran our algorithm with $K \in \{5, 10, 15, 20, 25, 30, 35, 40\}$ to find the K that maximizes the correlation[2]. Figure 2 shows the mean of the correlations for each K. The parameter settings that maximized the correlation are $\lambda_1 = 0.8$, $\lambda_2 = 1$, and $K = 25$.

Since we are doing joint filtering and factorization of speech, we want to set the filter constraint A to something meaningful. Generally, the first three formants (range typically from 200 Hz to 3000 Hz [14]) are important for speech intelligibility. Therefore, we set each column of A to be the frequency response of a 24-order equiripple bandpass FIR filter with a passband of 150 Hz to 4500 Hz. We set the higher part of the passband range to be higher than 3000 Hz to account for fricatives, which have high frequency components. We note, however, that the setting of A is data-specific that enforces prior knowledge about the data (in this case, typical frequency range of speech) on G. If there is no prior knowledge, or the signal of interest exists at all frequencies, then an allpass filter (matrix of 1s) can be used for A.

3.2. Evaluation

Using the optimum parameter settings, we compared the performance of the proposed algorithm to standard NMF as well as Wiener filter denoising followed by standard NMF (henceforth Wiener filter + NMF). We randomly sampled a different set of 100 sentences from the MOCHA-TIMIT corpus [16], with 50 sentences spoken by male speakers and the other 50 sentences spoken by female speakers. We added white noise, pink noise, speech babble, and factory floor noise from the NOISEX database [18] to these sentences at 5 dB and 10 dB SNR levels. For each noise and SNR level, we computed correlation for the basis matrices returned by our algorithm, standard NMF, and Wiener filter + NMF. Figure 3a shows the correlations for each algorithm in each noise conditions. The values shown are the average of the correlation values of the 100 sentences in each noise condition. Figure 4 shows the basis matrices recovered from the different noises at 10 dB SNR for one sentence.

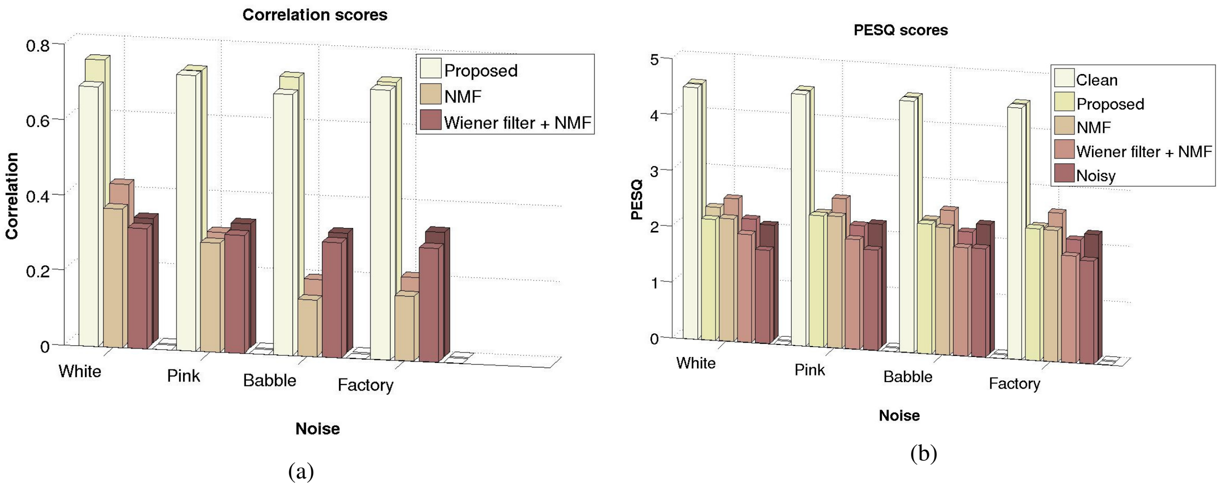
To evaluate the performance of the filtering in the proposed algorithm, we calculated the Perceptual Evaluation of Speech Quality (PESQ) score of the denoised speech [15]. We reconstructed the denoised speech from the recovered noisy basis and activation matrices by applying the formula $\hat{V}_{\text{denoised}} = V_{\text{noisy}} \otimes (W_{\text{noisy}} H_{\text{noisy}})$ and then computing the inverse Fourier transform of $\hat{V}_{\text{denoised}}$. Figure 3b shows the PESQ scores for the reconstructed signals from our algorithm and standard NMF, and the denoised signal from the Wiener filter in the different noise conditions. As with the correlation metric, these scores are averaged over the 100 sentences.

4. Discussion

We used the Wilcoxon rank-sum test, a non-parametric version of the Student's T-test, to evaluate the statistical significance of our results. The correlation values of our proposed algorithm is significantly better than NMF's and Wiener filter + NMF's correlation values at the 99% level in all noise conditions. This suggests that the basis matrices recovered from noisy data by our algorithm better represents the underlying structure of the signal of interest compared to using standard NMF or filtering the signal prior to performing NMF. This holds true across stationary and non-stationary noises, and wideband and narrowband noises. Overall, it appears that the joint filtering and factoring

[1] An empirical initial choice, roughly equal to the number of phones in English, so that we get a basis for each phone.

[2] We also considered using the Akaike Information Criterion or AIC [17] to find the optimal K. However, this criterion is not be very useful in our case as the model complexity term outweighs the data log-likelihood term significantly. In other words, the criterion prefers less complex models, i. e., smaller values of K.



(a)

(b)

Figure 3: (a) Correlation scores and (b) PESQ scores for proposed algorithm, standard NMF, and Wiener filter + NMF in 5 dB (front) and 10 dB (behind) SNR levels.

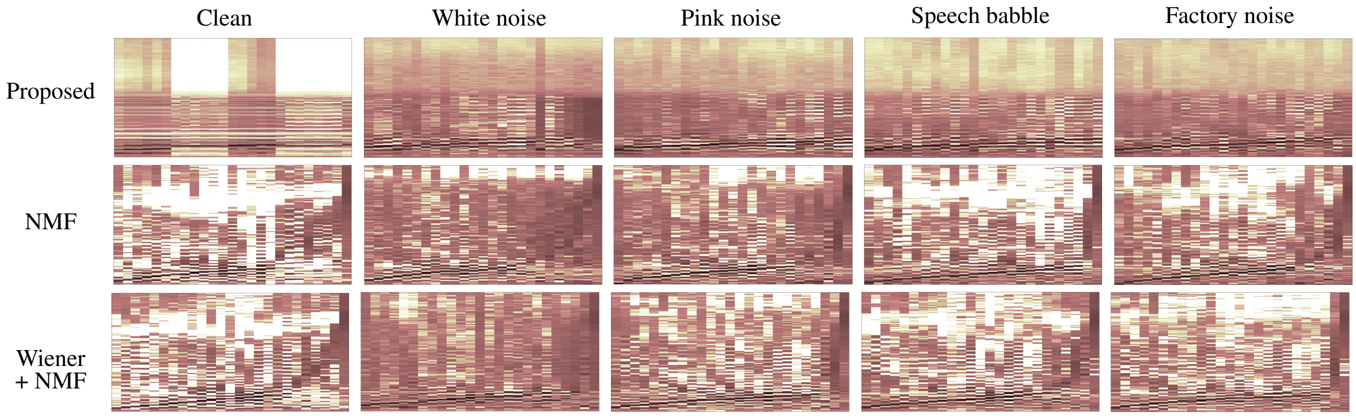


Figure 4: Basis matrices for one TIMIT sentence recovered from the proposed algorithm (top row), NMF (middle row), and Wiener filter + NMF (bottom row) under different noise conditions.

approach performs the best on wideband stationary noise, such as white and pink noises. This is because the bandpass filter in A helps G to remove a lot of the out-of-band noise, while the MVDR-like formulation in our algorithm helps G preserve the speech, which appears as peaks in the passband region of the spectrum. On the other hand, the proposed algorithm's performance degrades in narrowband noise. If the noise lies outside the frequencies of interest, then it will be likely suppressed by the stopband imposed by A. However, if the noise is within the passband, there are no such guarantees.

From looking at the correlation scores in Figure 3a, one can see that the joint filtering and factorization approach consistently outperforms the filtering followed by factorization approach. This suggests that there are benefits to filtering during the factorization operation rather than prior to factorization. One such benefit is that the filter can adapt to the factored outputs W and H and reweight the frequency response to further boost frequencies of interest while suppressing undesirable frequencies. Another benefit is relatively consistent basis recovery in different kinds of noise. One can see in Figure 4 that the proposed algorithm returns similar basis matrices in the different noise conditions more consistently as compared to the other algorithms.

From the PESQ scores in Figure 3b, one can see that the proposed algorithm's denoising performance is on par with NMF and Wiener filtering methods. The difference in PESQ scores between our algorithm and NMF is statistically significant only in the 10 dB noise conditions, but the Wiener filter scores are significantly worse than our algorithm and NMF in all noise conditions. This suggests that in addition to recovering improved basis matrices with the proposed algorithm, one can also reconstruct a denoised signal with a quality comparable to other denoising methods. We note that the parameters we used were optimized on the correlation metric, so a different set of parameter could improve our algorithm's denoising performance.

5. Conclusion

We have proposed a joint filtering and factorization approach for recovering latent structure in a signal of interest that is corrupted with noise. Results show that the basis matrices recovered by our proposed algorithm represent structural information better than using NMF factorization alone or performing filtering prior to NMF factorization. Furthermore, we found that the quality of denoised signals reconstructed from our algorithm is comparable to the quality when using NMF or Wiener filtering for denoising, making our algorithm a viable alternative fo r signal denoising.

In the future, we would like to broaden the applicability of our method by incorporating more generalized divergence metrics into the cost function. We will also explore probabilistic extensions, similar to how probabilistic latent component analysis (PLCA) is a probabilistic interpretation of NMF [19].

6. References

[1] P. Paatero and U. Tapper, “Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[2] P. Paatero, “Least squares formulation of robust non-negative factor analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[3] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Adv. in Neu. Info. Proc. Sys. 13*, 2001, pp. 556–562.

[4] Y. Gao and G. Church, “Improving molecular cancer class discovery through sparse non-negative matrix factorization,” *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[5] T. Virtanen, “Sound source separation using sparse coding with temporal continuity objective,” in *Proc. Int. Computer Music Conference*, 2003, pp. 231–234.

[6] C. Vaz, V. Ramanarayanan, and S. Narayanan, “A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis”, in *Proc. InterSpeech*, Lyon, France, 2013, pp. 1312–1315.

[7] D. Kong, C. Ding, and H. Huang, “Robust Nonnegative Matrix Factorization using L21-norm,”

[8] F. J. Theis, K. Stadlthanner, and T. Tanaka, “First Results on Uniqueness of Sparse Non-negative Matrix Factorization,”

[9] V. Ramanarayanan, L. Goldstein, and S. Narayanan, “Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation,” *J. Acoustical Soc. America*, vol. 134, no. 2, pp. 1378–1394, 2013.

[10] J. Capon, “High-Resolution Frequency-Wavenumber Spectrum Analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[11] S. Dharanipragada and B. D. Rao, “MVDR based feature extraction for robust speech recognition,” in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001, pp. 309–312.

[12] M. N. Murthi and B. D. Rao, “Minimum Variance Distortionless Response (MVDR) Modeling of Voiced Speech,” in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 1687–1690.

[13] C. Vaz, A. Tsiartas, and S. Narayanan, “Energy-Constrained Minimum Variance Response Filter for Robust Vowel Spectral Estimation,” in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[14] R. W. Schafer and L. R. Rabiner, “System for Automatic Formant Analysis of Voiced Speech,” *J. Acoustical Society of America*, vol. 47, no. 2B, pp. 634–648, 1970.

[15] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[16] A. A. Wrench, “A multi-channel/multi-speaker articulatory database for continuous speech recognition research,” *Workshop on Phonetics and Phonology in Automatic Speech Recognition, Saarbrucken, Germany*, 2000.

[17] H. Akaike, “Likelihood of a model and information criteria,” *Journal of Econometrics*, vol. 16, no. 1, pp. 3–14, 1981.

[18] A. Varga, and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] Z. Duan, G. J. Mysore, and P. Smaragdis, “Online PLCA for Real-time Semi-supervised Source Separation,” in *Proc. Int. Conf. Latent Variable Analysis/Independent Component Analysis*, Tel-Aviv, Israel, 2012, pp. 34–41.