Enhancing Audio Source Separability Using Spectro-Temporal Regularization with NMF

Colin Vaz[1], *Dimitrios Dimitriadis*[2], *and Shrikanth Narayanan*[1]

[1]Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089
[2]AT&T Labs–Research
Bedminster, NJ 07932

cvaz@usc.edu, ddim@research.att.com, shri@sipi.usc.edu

Abstract

We propose a spectro-temporal regularization approach for NMF that accounts for a source's spectral variability over time. The regularization terms allow NMF to adapt the spectral basis matrices optimally to reduce mismatch between the spectral characteristics of sources observed during training and encountered during separation. We first tested our algorithm on a simulated source separation task. Preliminary results show significant improvement of SAR, SDR, and SIR values over some current NMF methods. We also tested our algorithm on a speech enhancement task and were able to show a modest improvement of the PESQ scores of the recovered speech.

Index Terms: dictionary learning, NMF, speech enhancement, source separation.

1. Introduction

The goal of source separation algorithms is to extract individual audio components from a mixture of signals. For example, source separation can be used to separate a mixture of a speaker and background traffic audio into speaker-only and traffic-only sounds. It is used in a wide variety of tasks, such as speech enhancement [1, 2] and automatic music transcription [3], improving the overall performance.

One popular method for source separation is non-negative matrix factorization (NMF). First proposed by Paatero and Tapper [4, 5] and further developed by Lee and Seung [6], NMF takes a non-negative matrix representation of the mixture signal and factors it into basis and time activation matrices by minimizing a cost function. Convolutive NMF was proposed in [7] to represent spectral variability of a source over time by using multiple basis matrices. Sparsity constraints in the NMF formulation have been shown to reduce the chance of having more than one activation pattern being activated at the same time [8, 9]. Researchers have also developed different cost functions to tailor NMF to specific problems. For example, Cichocki et al. proposed cost functions based on Csiszár's φ-divergence to increase robustness to noise [10], and Guillamet et al. incorporated a diagonal weight matrix in the cost function to reduce redundancy in the basis matrix [11].

Source separation with NMF can also be improved by adding regularization terms to the cost function. Févotte et al. implemented temporal smoothing by using a Markov chain regularization term that enforces smoothness over the rows of the time-activation matrix [12]. Recently, Wilson et al. proposed temporal regularization for separating audio sources by incorporating temporal statistics of the sources in the regularization term [13]. This is useful for separating, for example, sources corresponding to fast and slow talkers, which have similar spectral characteristics but different temporal characteristics. Their approach, however, assumes that the basis matrices stay fixed, which degrades separability when the spectral characteristics evolve over time and no longer match the trained basis matrices. We propose a reformulation of their algorithm that accounts for spectral variability over time. This allows us to adapt our basis matrices to changing conditions in the sources.

The paper is organized as follows. Section 2 describes the formulation of the spectro-temporal regularization. We test our algorithm on a simulated example source separation task and a real speech enhancement task and report the results in Section 3. In Section 4, we analyze these results and discuss areas in which we can improve our algorithm. Finally, we state our conclusions and future work in Section 5.

2. Algorithm

NMF takes a $M \times N$ non-negative representation V of the source and factors it into a $M \times K$ basis matrix W and $K \times N$ time-activation matrix H such that $V \approx WH$. The product WH has rank at most K. Typically, K is chosen such that $K < \min(M, N)$, giving a lower-rank approximation of V. The non-negative representation for audio signals is usually the magnitude spectrogram and the basis matrix contains the spectral characteristics of the source. NMF performs the decomposition by minimizing a cost function, which is usually a distance metric between V and WH, such as the Frobenius norm, KL divergence, or Itakura-Saito divergence. We add a spectro-temporal regularization term to the NMF cost function, in a similar manner to Wilson et al. [13]. The regularization term constrains the updates of the basis matrix and time-activation matrices to better match the spectro-temporal characteristics of the source.

The proposed source separation algorithm consists of two stages: training and separation. In the training stage, we assume we have access to the J source signals in the mixture. We take the $M \times N$ magnitude spectrogram V_i of the ith source ($i \in \{1, \ldots, J\}$) and use standard NMF to learn a $M \times K$ basis matrix W_i and $K \times N$ time-activation matrix H_i. We replicate the W_i matrix N times and stack it into a $M \times K \times N$ tensor $W_{i,rep}$. Similarly, we replicate the H_i matrix M times and stack it into a $M \times K \times N$ tensor $H_{i,rep}$. We then do an

Work supported by NIH Grant DC007124. We would like to thank Mary Francis for her devotion and help in all SAIL research efforts.

element-wise multiplication of $W_{i,rep}$ and $H_{i,rep}$ to produce a new tensor A_i. Figure 1 illustrates this procedure. A_i captures the value of each element in the basis as a function of time for source i.

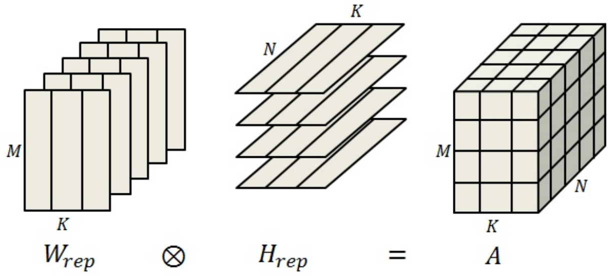


Figure 1: Diagram illustrating the calculation of tensor A. In this example, $M = 4$, $K = 3$, and $N = 5$. ⊗ denotes element-wise multiplication.

We assume that the elements in the basis matrix W_i and time-activation matrix H_i were drawn from a log-normal distribution and are independent of each other. The log-normal distribution is constrained to non-negative values, and the log of the distribution is a normal distribution, which is fully characterized by the mean vector and covariance matrix. Hence, we can characterize the statistics of W_i and H_i by calculating the means and covariances of $\log(W_i)$ and $\log(H_i)$. Furthermore, the multiplication of two independent log-normal random variables is also a log-normal random variable. Thus, the tensor A_i also has a log-normal distribution, which can be fully described by the mean and covariance of $\log(A_i)$.

We now learn spectro-temporal statistics about the ith source from A_i. We characterize the time variation of each frequency bin by calculating the $K \times 1$ mean vector $\mu_i(m)$ and $K \times K$ covariance matrix $\Sigma_i(m)$ of $\log(A_i(m, :, :))$, for $m \in [1, M]$. This gives us the mean activation of each element in the basis in the mth frequency bin and the covariance of those elements for each frequency bin. These statistics characterize the activation of each element in the basis. Hence, we will use these statistics to regularize the update equations for the basis and time activation matrices of the mixture signal. We note that we constructed A_i from the ground-truth source, so our experiments used supervised source separation. It is possible to use an unsupervised or semi-supervised approach, such as Probabilistic Latent Component Analysis (PLCA) [14], to construct A_i.

After the training stage, we run the separation stage. We obtain the spectrogram of the mixture V and calculate the best W and H that approximates this spectrogram while taking into consideration the spectro-temporal statistics of the sources. The proposed NMF algorithm includes an additional penalty term in its cost function to regularize the NMF updates. The cost function minimizes the KL divergence between the element in the mth row and nth column of V and the corresponding element in WH:

$$C(V, W, H) = \sum_{m=1}^{M} \sum_{n=1}^{N} [V(m,n) \log \frac{V(m,n)}{W(m,:)H(:,n)} + V(m,n) - W(m,:)H(:,n)] - \alpha L(A) \tag{1}$$

where

$$\begin{aligned} L(A) &= -\frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{[\log A(m,:,n) - \mu(m)]^T \Sigma^{-1}(m) \\ &\qquad [\log A(m,:,n) - \mu(m)] - \log((2\pi)^K |\Sigma(m)|)\} \\ &= -\frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{[\log W(m,:) + \log H(:,n) - \mu(m)]^T \Sigma^{-1}(m) \\ &\qquad (\log W(m,:) + \log H(:,n) - \mu(m)] - \log((2\pi)^K |\Sigma(m)|)\} \end{aligned} \tag{2}$$

where $A(m, :, n) = W(m, :) \otimes H(:, n)$, ⊗ means element-wise multiplication, $\mu(m) = [\mu_1^T(m)\ \mu_2^T(m) \cdots \mu_J^T(m)]^T$, and $\Sigma(m) = \text{blockdiagonal}(\Sigma_1(m), \Sigma_2(m), \ldots, \Sigma_J(m))$. $L(A)$ regularizes each frequency bin in the basis based on the statistics of the time-activation of that frequency bin. Following the method in [15], the multiplicative update rules for W and H are

$$\begin{aligned} W &\leftarrow W \otimes \frac{\frac{V}{WH} H^T}{[\mathbf{1} H^T + \alpha \phi(A)]_\varepsilon} \\ H &\leftarrow H \otimes \frac{W^T \frac{V}{WH}}{[W^T \mathbf{1} + \alpha \varphi(A)]_\varepsilon} \end{aligned} \tag{3}$$

where $\mathbf{1}$ denotes a matrix of ones, $[\cdot]_\varepsilon$ indicates that values less than a small positive constant ε should be replaced by ε to preserve non-negativity, and

$$\begin{aligned} \phi(A) &= \frac{\partial L(A)}{\partial W(a,b)} \\ &= \sum_{n=1}^{N} \frac{[\Sigma^{-1}(a)(\log W(a,:) + \log H(:,n) - \mu(a))]_b}{W(a,b)} \\ \varphi(A) &= \frac{\partial L(A)}{\partial H(a,b)} \\ &= \sum_{m=1}^{M} \frac{[\Sigma^{-1}(m)(\log W(m,:) + \log H(:,b) - \mu(m))]_a}{H(a,b)} \end{aligned}$$

In these equations, ⊗ means element-wise multiplication, division is element-wise, and $[\boldsymbol{v}]_a$ means the ath component of vector $\boldsymbol{v}$. W is initialized with $[W_1\ W_2 \cdots W_J]$ and H is initialized with a random matrix. To reconstruct the spectrogram for source i, we compute $\hat{V}_i = V \otimes \frac{\hat{W}_i \hat{H}_i}{WH}$, where $\hat{W}_i$ refers to the columns of W and $\hat{H}_i$ refers to the rows of H corresponding to source i.

3. Experiments

3.1. Setup

We ran a simulated source separation experiment and a speech enhancement experiment using our proposed algorithm, the algorithm presented in [13] (we will refer to this as Wilson's algorithm), and a standard NMF implementation that minimized the KL divergence. We used the standard NMF to learn the basis matrices for the sources during training. For the ith source signal, we computed its magnitude spectrogram V_i using a 20 ms Hamming window with a 10 ms shift. We initialized the basis matrix W_i with randomly-chosen columns of V_i. The NMF algorithm decomposed the spectrogram for each source into basis and time-activation matrices, yielding W_i and H_i. During the source separation stage, we computed the magnitude spectrogram for the mixture signal and initialized the basis with

$W = [W_1 \; W_2 \cdots W_J]$ for J sources in the mixture. We then used a sliding window on the spectrogram of 64 frames long, shifted every 32 frames, as input to the three NMF algorithms. We initialized the time-activation matrix H with a random matrix from the standard uniform distribution. For consistency, we used this same initial W and H for all three NMF algorithms. Additionally, we set the maximum number of update iterations to 100 for all the algorithms.

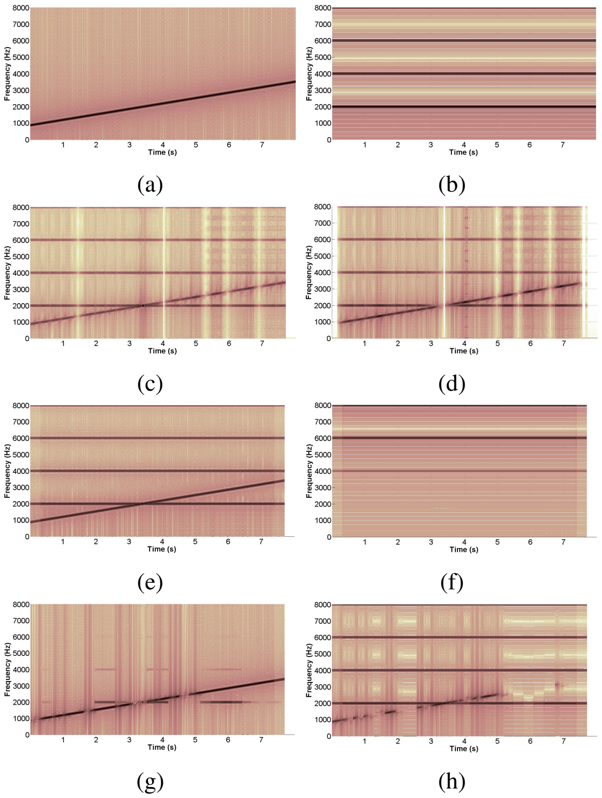
To optimize α, we mixed together chirp and sawtooth signals, which were used for the source separation experiment, and separated the mixture signal using our proposed algorithm. We calculated the SAR, SDR, and SIR of the separated sources using the BSSEval Toolbox [16]. We found the optimum α by doing a grid search from 0 to 1 with a step size of 0.1 and finding the α that maximized the SAR, SDR, and SIR values. Oftentimes, a certain parameter combination did not maximize all three measures simultaneously. So, we maximized the value of $\bar{sar} + \bar{sdr} + \bar{sir}$, where $\bar{(\cdot)}$ indicates the value is normalized to between 0 and 1 relative to values we observed during the grid search. For this particular mixture of chirp and sawtooth signals, we found the optimum α to be 0.2. We held these values fixed for all of our experiments. However, for optimum source separation performance, α should be tuned to the particular mixture signal that is being separated.

3.2. Separation Experiment

To test the performance of our algorithm, we ran a source separation task. We created a chirp signal with frequencies that swept from 880 Hz to 3520 Hz and a sawtooth wave at a constant 2000 Hz. The signals were 8 seconds long and had a sampling rate of 16 kHz. Figures 2a and 2b show the spectrograms of these signals. We trained a 1-component basis for the chirp and sawtooth signals and computed the statistics for Wilson's algorithm and our proposed algorithm. We added these signals together at 0 dB SNR to create the mixture signal. We passed the mixture signal through our proposed algorithm, Wilson's algorithm, and standard NMF to perform source separation. Figures 2c–2h show the spectrograms of the recovered signals. We ran this experiment 50 times, each time calculating the SAR, SDR, and SIR of the recovered signals. Figures 3a and 3b show the mean of these values for the recovered chirp and sawtooth signals.

3.3. Speech Enhancement Experiment

We also tested our algorithm on a speech enhancement task, where we separate speech from background noise. We took 48 sentences from the Wall Street Journal database, learned 20-component basis matrices, and computed spectro-temporal statistics for those sentences. We then took white noise, pink noise, F16 noise, and speech babble from the NOISEX database [17]. For each noise, we learned 20-component basis matrices, calculated spectro-temporal statistics, and added them to the speech at 0 dB SNR. We ran our proposed algorithm to separate the noisy signal into the speech and noise sources. We calculated the PESQ score, a quantitative measure of the perceptual quality of speech, of the recovered speech signal [18]. Again, we compared our performance to the standard NMF and Wilson's algorithm. Figure 4 shows the PESQ scores for this experiment in the different noise conditions.



(a) (b)

(c) (d)

(e) (f)

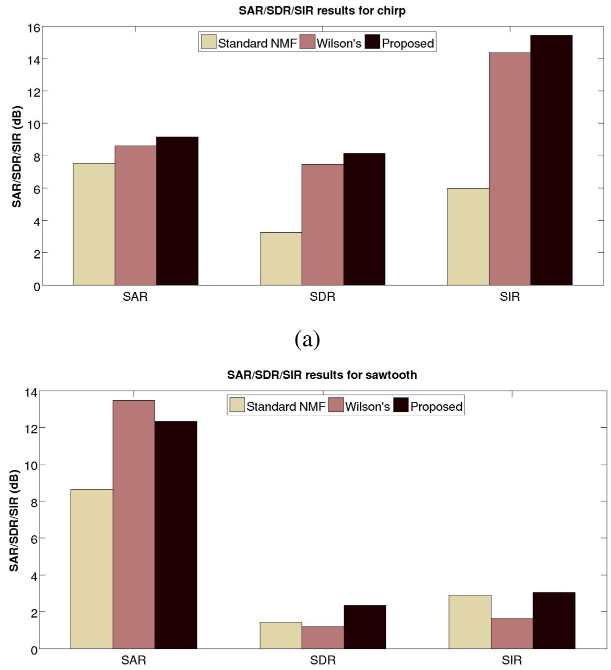
(g) (h)

Figure 2: Spectrograms for (a) chirp, (b) sawtooth, (c) chirp recovered by standard NMF, (d) sawtooth recovered by standard NMF, (e) chirp recovered by Wilson's algorithm, (f) sawtooth recovered by Wilson's algorithm, (g) chirp recovered by proposed algorithm, and (h) sawtooth recovered by proposed algorithm.

One advantage our algorithm has over Wilson's algorithm is the ability to update the basis matrix at each iteration. This means that the basis can change to compensate for any spectral mismatch of the source between the training and separation stages. The regularization terms in our algorithm help the basis matrix to update based on the spectro-temporal statistics of the source. To test this, we ran another speech enhancement task, this time training the speech source on male speakers but doing separation of noisy female speech. The spectral mismatch between male and female speakers would require the basis to adapt during separation to better match the female speaker. Figure 5 shows the PESQ scores of this experiment.

4. Discussion

We used the Wilcoxon rank-sum statistical test, a nonparametric version of the Student's T-test, to determine the statistical significance of the results. For the source separation experiment, we determined that the improvement of the SAR, SDR, and SIR results for the chirp over the other two NMF methods was statistically significant at the 95% level. This improvement can be attributed to the basis update and the spectro-temporal statistics for the chirp. The chirp's frequency varied with time, so it required the basis, which had only one component, to change over time. The spectro-temporal statistics guided the basis update as the chirp's frequency changed. Thus, it can be seen in Figure 2 that our proposed algorithm recovers the chirp better than the other methods. Similar results hold for the sawtooth signal, though the differences in performance between the algorithms are much less than the results for chirp (our algorithm is only significantly better for SDR at the 95% level). Sawtooth has a constant frequency over time, so in this case, there is no particular advantage of our algorithm over the



(a)

(b)

Figure 3: Mean SAR, SDR, and SIR values for (a) chirp and (b) sawtooth.

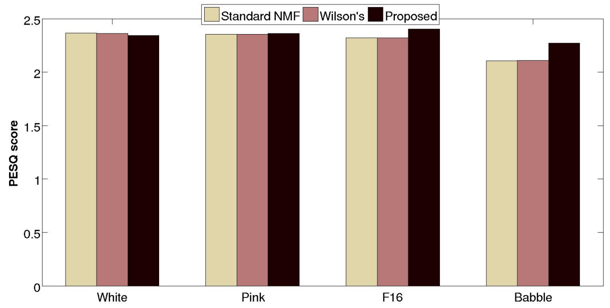


Figure 4: PESQ scores for the speech recovered by standard NMF, Wilson's algorithm, and proposed algorithm in different noises matched training data.

others.

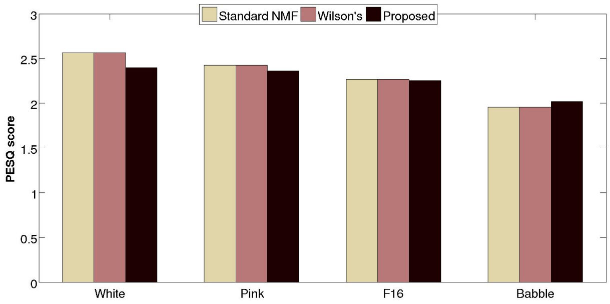


Figure 5: PESQ scores for the speech recovered by standard NMF, Wilson's algorithm, and proposed algorithm in different noises with mismatched training data.

For the speech enhancement experiment, our proposed algorithm performs significantly better than Wilson's algorithm at the 95% level in F16 and speech babble noises. The performance for all three algorithms are similar for white and pink noise. Because these are stationary noises, they do not have much temporal variability in their spectra. Hence, there is not much gained by using spectro-temporal regularization, as in the case of the sawtooth signal. The results are similar for the mismatched speech enhancement experiment, with the proposed algorithm performing better than the others in non-stationary noise. It appears, though, that all of the algorithms perform similarly well in the mismatched case, contrary to our hypothesis that a spectral mismatch would degrade the performance of standard NMF and Wilson's algorithm. To investigate this, we observed the basis matrices of the male-spoken sentences and female-spoken sentences and found many similar basis vectors between the male sentences and female sentences. Hence, there isn't an overwhelming spectral mismatch to affect the performance of the algorithms. Surprisingly, standard NMF performed very well in the speech enhancement experiments. To investigate this, we computed the SAR, SDR, and SIR values of the recovered speech. We found that the recovered speech from standard NMF had a very high SIR and low SAR and SDR. This means that standard NMF separated the speech and noise rather well but at the expense of introducing distortions and artifacts into the recovered signal. On the other hand, Wilson's algorithm and our algorithm had more balanced SAR, SDR, and SIR values. This suggests that the regularization term balances the trade-off between good separation and high distortion.

We note that α was not optimized for the speech enhancement experiments. Different background noises and SNR levels will most likely require different levels of spectro-temporal regularization. This would be especially true depending on whether the noise is stationary or not, because this affects how much the basis and time-activation matrices should change at each iteration. Thus, we can improve our algorithm by calculating α based on the statistics of the sources, which would allow α to account for the spectral, temporal, and energy characteristics of the sources.

5. Conclusion

We have proposed a spectro-temporal regularization for NMF that uses statistics of the temporal evolution of the sources' spectra. Unlike some NMF approaches that assume a fixed basis during separation, our approach allows updating of the basis to adapt to unseen spectral characteristics in the mixture signal. Preliminary results from a source separation experiment and speech enhancement tasks were promising for dealing with sources with non-stationary spectral characteristics.

We will revise our spectro-temporal regularization terms to better handle stationary sources and find a way to automatically calculate the optimum α. Additionally, we will apply our regularization approach to convolutive NMF because convolutive NMF is designed to deal with sources that have a time-varying basis. After implementing these improvements, we will evaluate the algorithm's performance on more standard source separation databases, such as CHiME. We will also work on a computationally efficient implementation of the regularization so that it can be more useful for practical applications, such as real-time speech enhancement. More specifically, we will evaluate ASR performance on noisy speech when using our proposed algorithm on the front-end of an ASR system.

6. References

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

[2] 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.

[3] S. A. Abdallah and M. D. Plumbley, “Polyphonic trascription by non-negative sparse coding of power spectra,” in *Proc. 5th Int. Conf. Music Information Retrieval*, 2004, pp. 318–325.

[4] 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.

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

[6] 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.

[7] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Fifth Int. Conf. Independent Component Analysis*, Granada, Spain, 2004, pp. 494–499.

[8] P. D. O’Grady and B. A. Pearlmutter, “Convolutive non-negative matrix factorisation with a sparseness constraint,” in *Proc. IEEE Signal Processing Society Machine Learning for Signal Processing*, Arlington, VA, 2006, pp. 427–432.

[9] M. Schmidt and R. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *Proc. InterSpeech*, Pittsburgh, PA, 2006.

[10] A. Cichocki, R. Zdunek, and S. Amari, “Csiszár’s divergences for non-negative matrix factorization: family of new algorithms,” in *Proc. Sixth Int. Conf. Independent Component Analysis and Blind Signal Separation*, Charleston, SC, 2006, pp. 32–39.

[11] D. Guillamet, M. Bressan, and J. Vitrià, “A weighted non-negative matrix factorization for local representations,” in *Proc. 2001 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Kavai, HI, 2001, pp. 942–947.

[12] C. Févotte, N. Bertin, and J. L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[13] K. W. Wilson, B. Raj, and P. Smaragdis, “Regularized non-negative matrix factorization with temporal dependencies for speech denoising,” in *Proc. InterSpeech*, Brisbane, Australia, 2008, pp. 411–414.

[14] 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.

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

[16] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, and Language Process.*,” vol. 14, no. 4, pp. 1462–1469, 2006.

[17] 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.

[18] 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.