Stopping Criteria for Non-negative Matrix Factorization Based Supervised and Semi-Supervised Source Separation

François G. Germain, *Student Member, IEEE,* and Gautham J. Mysore, *Member, IEEE*

***Abstract*—Numerous audio signal processing and analysis techniques using non-negative matrix factorization (NMF) have been developed in the past decade, particularly for the task of source separation. NMF-based algorithms iteratively optimize a cost function. However, the correlation between cost functions and application-dependent performance metrics is less known. Furthermore, to the best of our knowledge, no formal heuristic to compute a stopping criterion tailored to a given application exists in the literature. In this paper, we examine this problem for the case of supervised and semi-supervised NMF-based source separation and show that iterating these algorithms to convergence is not optimal for this application. We propose several heuristic stopping criteria that we empirically found to be well correlated with source separation performance. Moreover, our results suggest that simply integrating the learning of an appropriate stopping criterion in a sweep for model size selection could lead to substantial performance improvements with minimal additional effort.**

***Index Terms*—Non-negative Matrix Factorization, Source Separation.**

I. Introduction

In THE past decade, methods based on non-negative matrix factorization (NMF) have become popular in a variety of fields, and in particular for audio signal processing tasks such as source separation. From its original formulation [1] and application to audio [2], multiple variants of the algorithm have been proposed to improve performance in different scenarios, through the use of cost functions such as the Euclidean distance, the generalized Kullback-Leibler (KL) divergence [1], the Itakura-Saito divergence [3] and the more general Beta divergence [4]. Other developments include temporal modeling [5]–[8] and various sparsity constraints [5], [9]–[12].

Source separation methods are commonly evaluated using the BSS evaluation metrics [13]. They consist of three scores: Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR) and Source-to-Artifact Ratio (SAR) which measure the overall quality of the separation, the attenuation of the interfering sources, and the degradation of the target signal due to artifacts respectively. Recently, four new evaluation metrics based on a nonlinear mapping of signal-related quantities to the results of perceptual experiments were presented as the PEASS metrics [14], in order to improve matching of the perceptual quality and BSS evaluation scores. The PEASS metric that is used to measure the overall quality of the separation is Overall Perceptual Score (OPS). For speech separation, metrics such as the Short-Time Objective Intelligibility Measure (STOI) [15] that measure speech intelligibility degradation can be valuable as well.

While the number of NMF-based methods are growing, many questions remain on how to choose the best technique, with the best model size for a given task. A recent study [16] addresses the question of optimizing some of the parameters of NMF-based source separation. Another issue arises from the fact none of the evaluation metrics are formally related to the algorithm cost function. Although consecutive iterations of the algorithm are guaranteed to monotonically improve the cost function value, there is no guarantee that they will improve the performance with respect to the evaluation metric of interest. A recent study [17] shows the mismatch between PEASS scores and cost optimization for a class of NMF-based source separation algorithms. It does not address supervised and semi-supervised source separation or the mismatch between cost, BSS and STOI metrics. Moreover, no guideline is provided as to how to choose a stopping criterion to alleviate this issue.

In this paper, we examine the correlation between the cost optimization of NMF-based supervised and semi-supervised source separation and the BSS, PEASS, and STOI scores. Then, we propose several heuristic stopping criteria to alleviate the mismatch between those quantities. In Section II, we present the algorithms. In Section III, we present the results of empirical convergence analysis of these algorithms and the associated performance metrics. In Section IV, we propose our criteria and present the results of validation experiments.

II. Algorithms

NMF-based source separation algorithms take advantage of the non-negative nature of the spectrogram $\mathbf{X} = |\mathcal{X}|^{\alpha}$, with $\mathcal{X}$ the signal short-time Fourier transform (STFT) and $\alpha > 0$, to approximate it as $\mathbf{X} \approx \mathbf{WH}$ through the optimization:

$$\text{argmin}_{\mathbf{W},\mathbf{H}\geq 0}\ D(\mathbf{X}||\mathbf{WH}) \quad (1)$$

with $\mathbf{W}$ and $\mathbf{H}$ non-negative matrices. The columns of $\mathbf{W}$ can typically be interpreted as the spectral basis vectors of the sources in the spectrogram. The matrix $\mathbf{H}$ can then be

F. G. Germain is with the Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA 94305 (email: fgermain@stanford.edu). This work was performed while interning at Adobe Research.

G. J. Mysore is with Adobe Research, San Francisco, CA 94103 (email: gmysore@adobe.com).

interpreted as the activity of each vector in a given time frame. Here, we use the magnitude spectrogram ($\alpha = 1$) and the generalized KL divergence as cost function as it is commonly used in source separation. For $\hat{\mathbf{X}} = \mathbf{WH}$, it is defined as:

$$D(\mathbf{X}||\hat{\mathbf{X}}) = \sum_{i,j} \mathbf{X}_{i,j} \log \left(\mathbf{X}_{i,j} / \hat{\mathbf{X}}_{i,j} \right) - \mathbf{X}_{i,j} + \hat{\mathbf{X}}_{i,j} \quad (2)$$

The typical pipeline to perform NMF-based source separation in the presence of two sources, say speech and noise, follows the process detailed in [18], (derived from the equivalent perspective of Probabilistic Latent Component Analysis):

1) Compute the spectrograms $\mathbf{X}_S$ and $\mathbf{X}_N$ from the speech and noise training data, as well as the spectrogram $\mathbf{X}$ of the test mixture signal.
2) Factorize the spectrograms $\mathbf{X}_i \approx \mathbf{W}_i \tilde{\mathbf{H}}_i$, for $i = S, N$ using NMF and form the matrix $\mathbf{W} = \begin{bmatrix} \mathbf{W}_S & \mathbf{W}_N \end{bmatrix}$.
3) Learn the activations $\mathbf{H}$ from the test mixture spectrogram while keeping $\mathbf{W}$ fixed: $\mathbf{X} \approx \mathbf{WH}$.
4) Partition the activations as $\mathbf{H} = \begin{bmatrix} \mathbf{H}_S \\ \mathbf{H}_N \end{bmatrix}$, and construct estimated spectrograms $\hat{\mathbf{X}}_i = \mathbf{W}_i \mathbf{H}_i$.
5) Construct two time-frequency masks from the $\hat{\mathbf{X}}_i$ and extract estimated STFTs of each source through Wiener-like filtering of the mixture STFT $\mathcal{X}$:

$$\hat{\mathcal{X}}_S = \frac{\hat{\mathbf{X}}_S}{\hat{\mathbf{X}}_S + \hat{\mathbf{X}}_N} \mathcal{X} \quad \text{and} \quad \hat{\mathcal{X}}_N = \frac{\hat{\mathbf{X}}_N}{\hat{\mathbf{X}}_S + \hat{\mathbf{X}}_N} \mathcal{X} \quad (3)$$

6) Compute the inverse STFT of $\hat{\mathcal{X}}_i$ to get an estimate of each source signal.

This algorithm corresponds to **supervised** separation, where training data is available for both sources. In the case in which training data is available for only one of the sources (for example speech, as is often the case for speech denoising), **semi-supervised** separation can be performed, by modifying the pipeline such that only $\mathbf{W}_S$ is learned in step 2, while $\mathbf{W}_N$ and $\mathbf{H}$ are learned simultaneously from $\mathbf{X}$ in step 3.

III. Empirical Convergence Analysis

A. Task

For the purpose of analyzing the convergence properties of both supervised and semi-supervised NMF-based source separation, we examine the case of speech denoising. In this application, we are given a mixture of speech with background noise, on which we apply the source separation algorithm, with a particular focus on 1) a clean reconstruction of the speech, and 2) a significant reduction of the noise level. In both cases, speaker-dependent training data is available, while noise training data is available only in the supervised case. To evaluate the results, we focus on the metrics relative to the estimated speech signal and its ground truth.

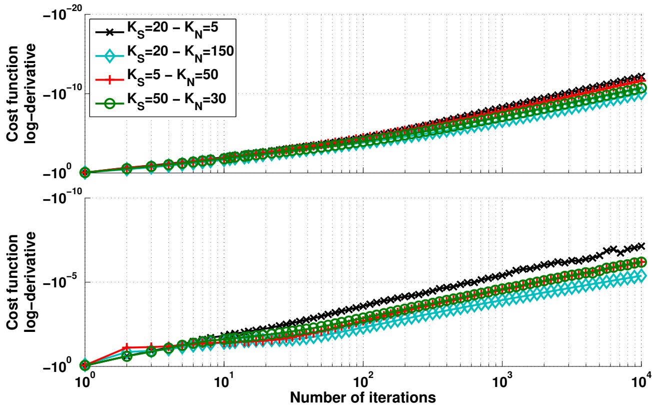

Fig. 1. Average cost function log-derivative value for the supervised (top) and the semi-supervised (bottom) algorithm with 4 sets of model sizes.

Spectrograms were computed using a 1024-sample Hann window with 75% overlap. For both supervised and semi-supervised algorithms, we trained speaker- and noise-dependent models $\mathbf{W}_S$ and $\mathbf{W}_N$ from the training data with the following number of vectors (K_S for speech, K_N for noise): $(K_S, K_N) = \{(20, 5), (20, 150), (5, 50), (50, 30)\}$. We motivate those choices in Section IV. Models of individual sources were trained (from isolated training data) until the cost D_n (at iteration n) verifies $|D_{n+1} - D_n| < 10^{-4}|D_n|$. The separation performance scores on the mixtures were computed along a logarithmic grid of iteration numbers accounting for the fact that the variations of the studied quantities become smaller as we reach large iteration numbers.

B. Data

For speech data, we use the utterances from 600 speakers of the TIMIT dataset [19] recorded at a sampling rate of 16kHz. The utterances for a given speaker are divided into training and testing segments. For noise data, we use samples from two datasets, for a total of 30 different noise types: The NOISEX-92 dataset [20] contains examples of quasi-stationary noises (e.g., factory, helicopter, jet aircraft). The second dataset [21] contains examples of non-stationary noises (e.g., frogs, keyboard, ringtones). The data for a given noise type is divided into training and testing segments (the training data was not used in the case of semi-supervised separation). A test utterance for each speaker is mixed with a single type of test noise for a total of 600 test mixtures (20 mixtures per noise type). Each test mixture is 5 seconds long, and is associated with at least 17 seconds of training data for speech and noise.

C. Results

In setting stopping criteria for an iterative optimization algorithm, we often use $D(\mathbf{X}||\mathbf{WH})$ as only available measurable quantity in practical scenarios. It is common to set a threshold $\epsilon > 0$ on the relative variation of the cost function value instead of the absolute variation for the criterion to become scale-independent as in $|D_{n+1} - D_n| < \epsilon|D_n|$. This corresponds to a discretization of a threshold on the cost function log-derivative value $\frac{1}{D}\frac{dD}{dn}$ with respect to the number of algorithm iterations n (for conciseness, we omit here the absolute value). Our results show that this quantity is roughly a linear function of the number of iterations in the log-log domain for both the supervised and the semi-supervised algorithms (Fig. 1). They also show that the log-derivative

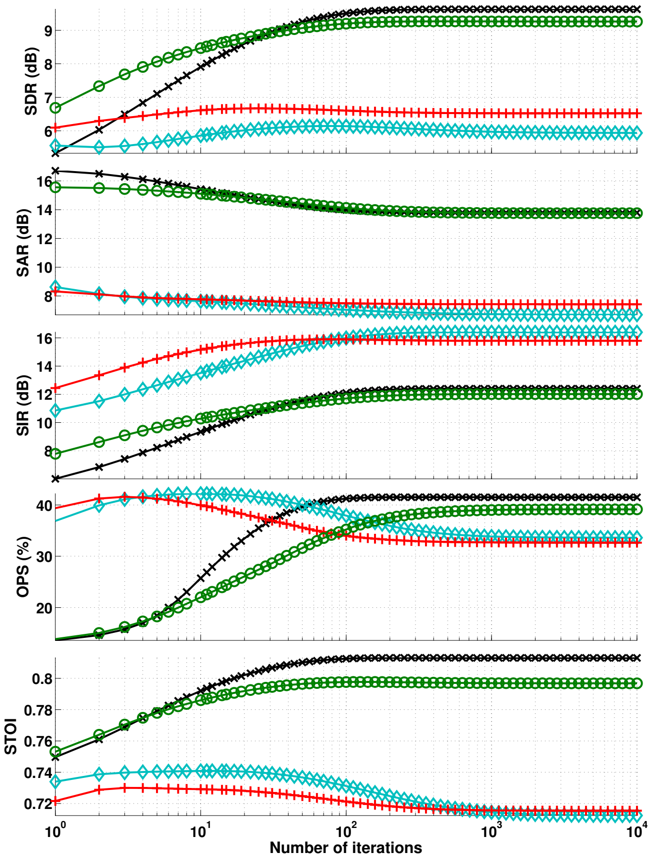

Fig. 2. From top to bottom: Average SDR, SAR, SIR, OPS and STOI scores for the supervised algorithm (legend: see Fig. 1).

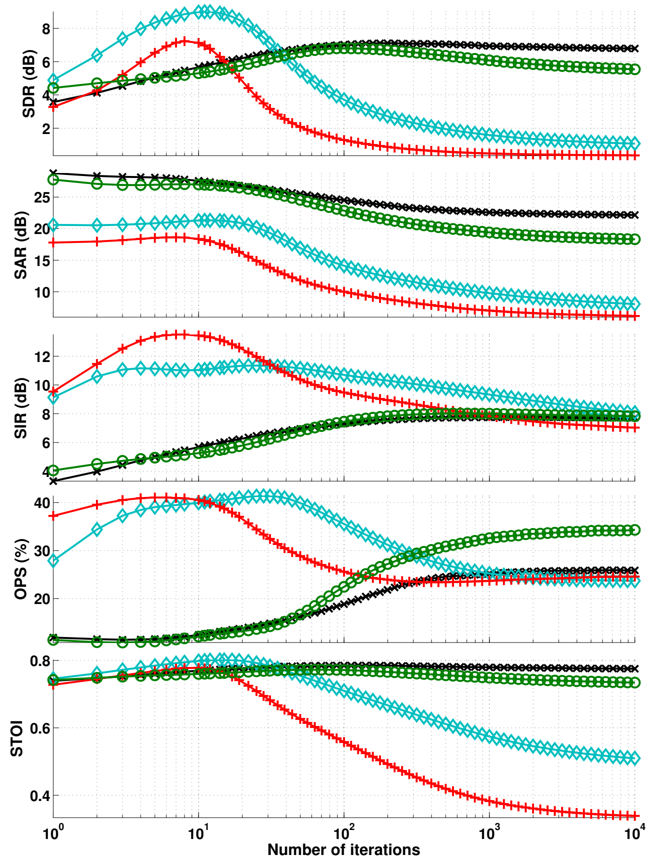

Fig. 3. From top to bottom: Average SDR, SAR, SIR, OPS and STOI scores for the semi-supervised algorithm (legend: see Fig. 1).

value behavior at a given iteration is mostly independent of the model sizes K_S and K_N.

While the variation of the average cost function log-derivative value is roughly monotonic, this does not match the behavior of the overall scores in the source separation metrics. For the supervised algorithm, the average SDR, OPS and STOI scores reach their maximum before slowly decreasing at each additional iteration, but the degradation of the score is limited (Fig. 2). Except for the OPS score of $(K_S, K_N) = (50, 30)$, we can make a similar observation for the semi-supervised algorithm. However, the scores now decrease more significantly for each additional iteration beyond the maximum (Fig. 3).

The SAR appears to be monotonically decreasing for both algorithms after a few iterations, but that degradation is much faster for the semi-supervised algorithm. The SIR appears to increase monotonically and then in most cases degrades after reaching a maximum value, similar to the SDR. This is more pronounced in the semi-supervised algorithm. This suggests that the noise model starts learning parts of the speech patterns.

We notice that the performance ranking between model sizes changes significantly over iterations for SDR, OPS and STOI, and that, in general, larger K_S and smaller K_N seem to require more iterations to reach the maximum average score. The fact that the variation of the cost function log-derivative value is independent of the model sizes K_i (Fig. 1), while the scores are strongly K_i-dependent (Figs. 2, 3), suggests that arbitrary stopping criteria are likely to lead to sub-optimal results, and that no simple K_i-independent relationship exists between the optimal interation number and the cost function value.

IV. Proposed stopping criteria

We cannot measure the performance scores without ground truth data, which is not available in real world scenarios. Therefore, considering the non-monotonic relation between the convergence score and performance scores, it seems crucial to find a stopping criterion based on a measurable quantity that would maximize performance scores. It is also an important issue to consider when searching for the best set of model sizes for a given application, and could be considered as an integral part of the sweep for model size selection when looking for the most performant setup.

We explore the performance in SDR and OPS of two heuristics based on the cost function log-derivative values, and three heuristics based on the number of iterations. To do so, we run a 30-fold cross-validation on the test data by using 1 single type of noise for each fold (20 mixtures). We take one fold as the evaluation set and the other 29 folds as development set. By construction, the speakers in those two sets are always different. For each development set mixture, we record the iteration number and the associated cost function log-derivative value at which the best performance score is measured. We then look at the average of the distribution of the following quantities at these optimal values on the development set results:

	K_S	K_N	Optimal	$\frac{dD}{D}$	$\log\left(\frac{dD}{D}\right)$	Iter.	log-Iter.	Max. aver.
SDR	20	5	9.64	9.33	**9.63**	**9.63**	**9.63**	**9.63**
	20	150	6.47	5.51	6.05	5.94	6.05	**6.14**
	5	50	7.05	6.29	6.61	6.52	6.62	**6.67**
	30	50	9.32	8.86	**9.26**	**9.26**	**9.26**	**9.26**
OPS	20	5	42.73	23.14	41.42	**41.49**	**41.49**	41.41
	20	150	44.38	39.68	41.79	33.6	41.88	**42.06**
	5	50	43.47	41.11	41.37	32.81	41.44	**41.58**
	30	50	39.75	20.88	39.07	**39.14**	39.06	**39.14**
SDR	20	5	7.57	6.94	6.83	6.76	6.92	**7.09**
	20	150	9.26	8.89	8.69	8.85	8.96	**8.98**
	5	50	7.4	6.16	5.95	6.97	7.13	**7.23**
	30	50	7.36	6.7	6.58	5.42	6.66	**6.75**
OPS	20	5	27.8	19.88	25.72	**25.9**	25.69	25.82
	20	150	43.64	40.61	41.18	38.76	40.97	**41.33**
	5	50	43.14	39.53	40.46	25.08	**41.05**	40.89
	30	50	36.88	12.22	33.9	**34.57**	33.92	**34.57**

TABLE I

Comparison of performance metrics at the iteration number that yields the highest SDR and OPS scores (Optimal) and the iteration numbers obtained from the proposed five heuristics (see Section IV) for the supervised (top) and semi-supervised algorithms (bottom). The metrics obtained from the 4th and the 5th heuristics are typically close to optimal.

$K_S \backslash K_N$		5	10	20	30	50	70	100	150	200
5	Best	8.7	8.8	8.5	8.2	7.7	7.3	6.8	6.2	5.8
	No.	10	10	10	7	7	7	7	7	7
	Fixed	*5.9*	*3.7*	*2.1*	*1.3*	*0.6*	*0.3*	*0.0*	*-0.2*	*-0.2*
10	Best	8.5	8.9	9.1	9.2	9.2	9.1	9.0	8.8	8.5
	No.	30	15	10	10	10	10	10	10	10
	Fixed	*7.7*	*6.2*	*4.6*	*3.8*	*2.8*	*2.3*	*1.9*	*1.6*	*1.5*
20	Best	8.1	8.3	8.6	8.8	9.1	9.2	9.3	**9.4**	**9.4**
	No.	100	30	20	20	15	15	15	**10**	**10**
	Fixed	***8.1***	*7.7*	*6.7*	*6.0*	*5.1*	*4.6*	*4.1*	*3.7*	*3.5*
30	Best	7.7	7.9	8.2	8.4	8.7	8.8	9.0	9.2	9.3
	No.	200	50	30	30	20	20	15	15	15
	Fixed	*7.7*	*7.8*	*7.4*	*7.0*	*6.3*	*5.8*	*5.4*	*4.9*	*4.7*
50	Best	7.1	7.4	7.6	7.7	8.0	8.2	8.4	8.6	8.8
	No.	500	200	70	50	30	30	30	20	20
	Fixed	*6.8*	*7.3*	*7.5*	*7.5*	*7.2*	*7.0*	*6.7*	*6.4*	*6.2*

TABLE II

SDR scores recorded at the iteration number for the proposed maximum average score heuristic (Best) and the corresponding iteration number (No.) compared to the SDR score when using a fixed number of 100 iterations (Fixed).

- cost function log-derivative value
- logarithm of the cost function log-derivative value
- number of iterations
- logarithm of the number of iterations

Additionally, we consider a fifth stopping criterion where we use the iteration number corresponding to the maximum average score on the development set as stopping criteria.

We then use these values as stopping criterion (either on the number of iterations or the cost function log-derivative value) on the evaluation samples, and record the score obtained when the algorithm is stopped. The average scores for both the supervised and the semi-supervised algorithms and for both the BSS and PEASS overall metrics are then compared to the average true optimal score on each individual mixture.

For the supervised algorithm (Table I - top), the 2nd, 4th and 5th heuristics consistently achieve scores close to optimality. By using the 5th criterion (iteration number of maximum average score), the performance loss is on average 0.20dB for the SDR and 1.54% for the OPS compared to optimality.

For the semi-supervised algorithm (Table I - bottom), the 4th and 5th heuristics achieve scores close to optimality. With the 5th criterion, the average performance loss of 0.39dB for the SDR and 2.21% for the OPS compared to optimality.

Based on our experiments, we believe that a general method to perform model size selection in an NMF-based source separation problem, taking stopping criteria into account, is:

1) Gather a dataset from sources you would like to separate and build a training set of with isolated segments and an independent development set with synthesized mixtures.
2) Train models of different sizes K_i on the training set, then run the separation algorithm on the development mixtures for 500 iterations. Compute the cost and metric of interest at every iteration (or once every few iterations), and find the optimal iteration number and its cost function value for each mixture with each set of K_i.
3) Compute the value associated with the chosen stopping criterion either as the average of the associated quantity at the optimal iteration for each mixture, or as the number of iterations where the best average score was found. The optimal model sizes and stopping criterion will then be the one with the highest average performance.
4) Use the above optimal model sizes and stopping criterion when performing separation of test samples.

In the case of semi-supervised NMF, we can see in Fig. 3 that an arbitrary choice of number of iterations or cost can lead to a suboptimal choice of model sizes if the iteration number is not optimized as well. We run this algorithm on our test data set (600 mixtures) and record the SDR scores at iterations $\{3, 5, 7, 10, 15, 20, 30, 50, 70, 100, 150, 200, 300, 500\}$. We compare the best average score at the given iteration number (the 5th criterion) to the average score recorded at a fixed number of 100 iterations as this is commonly used in the literature (Table II). From this table, we see that the performance at 100 iterations for given model sizes is often much lower than the one at the best number of iterations. Additionally, the best model sizes are different if the iteration number is included as variable, with a gain of 1.3dB SDR for only 10 iterations. From this results, we choose the model sizes (K_S, K_N) used in the plots of Section III as the best set at 100 iterations $(20, 5)$, the best set overall $(20, 150)$, and two other sets $(5, 50)$ and $(50, 30)$ for demonstration purposes.

V. Conclusion and Future Work

We have shown a clear mismatch between the optimization of the cost function and the optimization of common performance metrics in NMF-based supervised and semi-supervised source separation algorithms. We also proposed empirical stopping criteria that were more closely correlated with the optimal value of the performance metrics of interest, and validated these criteria with speech denoising experiments. Our approach is likely to be useful for a wide variety of applications that use iterative algorithms in which the performance metric of interest is not directly optimized.

References

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

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

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

[4] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proceedings of the 9th international conference on Latent variable analysis and signal separation*, ser. LVA/ICA'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 140–148.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 873–877.

[8] C. Févotte, J. Le Roux, and J. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3158–3162.

[9] B. Gao, W. Woo, and S. Dlay, "Variational regularized 2-d nonnegative matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 703–716, May 2012.

[10] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[11] O. Zoidi, A. Tefas, and I. Pitas, "Multiplicative update rules for concurrent nonnegative matrix factorization and maximum margin classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 422–434, March 2013.

[12] V. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, July 2013.

[13] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010, pp. 4214–4217.

[16] B. King, C. Fevotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[17] D. Fitzgerald and R. Jaiswal, "On the use of masking filters in sound source separation," in *Proceedings of the 15th Internation Conference on Digital Audio Effects*. Dublin Institute of Technology, 2012.

[18] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, ser. ICA'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 414–421.

[19] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*. National Institute of Standards and Technology, NISTIR 4930, 1993.

[20] A. Varga and H. 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.

[21] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments." in *Proceedings of INTERSPEECH 2012*. ISCA, 2012.