PREDICTION BASED FILTERING AND SMOOTHING TO EXPLOIT TEMPORAL DEPENDENCIES IN NMF

Nasser Mohammadiha[*†] *Paris Smaragdis*[‡] *Arne Leijon*[†]

[†] KTH Royal Institute of Technology
Sound and Image Processing Lab
Stockholm, Sweden
[‡]University of Illinois at Urbana-Champaign
Dept. of Computer Science
Dept. of Electrical and Computer Engineering
Adobe Systems Inc.

ABSTRACT

Nonnegative matrix factorization is an appealing technique for many audio applications. However, in it's basic form it does not use temporal structure, which is an important source of information in speech processing. In this paper, we propose NMF-based filtering and smoothing algorithms that are related to Kalman filtering and smoothing. While our prediction step is similar to that of Kalman filtering, we develop a multiplicative update step which is more convenient for nonnegative data analysis and in line with existing NMF literature. The proposed smoothing approach introduces an unavoidable processing delay, but the filtering algorithm does not and can be readily used for on-line applications. Our experiments using the proposed algorithms show a significant improvement over the baseline NMF approaches. In the case of speech denoising with factory noise at 0 dB input SNR, the smoothing algorithm outperforms NMF with 3.2 dB in SDR and around 0.5 MOS in PESQ, likewise source separation experiments result in improved performance due to taking advantage of the temporal regularities in speech.

Index Terms— Nonnegative matrix factorization (NMF), Probabilistic latent component analysis (PLCA), Prediction, Temporal dependencies.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [1] is a technique that decomposes a nonnegative matrix into a product of two nonnegative matrices such that one contains basis vectors and the other contains activations. NMF can be seen as a feature extraction method that discovers a low-dimensional representation in terms of a set of basis vectors. When applied to speech or music spectrograms, NMF has been shown to produce promising results in different applications [2–5].

Since the basic NMF model ignores temporal correlations, different approaches have been used in the past to enhance the decomposition to model time dependencies for audio signals. For example, Virtanen [2] used a regularization term in NMF, motivated by the temporal dependencies of speech signals, to develop a monaural sound source separation algorithm. A regularized NMF was also used in [6] where a heuristic regulation term was added to the NMF cost function that enforced temporal constraints as part of a noise reduction scheme. Another regularized NMF was proposed in [7] in which an l_2-norm penalty term was constructed and added to the NMF cost function to encourage temporal smoothness between the NMF coefficients.

In a recently developed class of approaches, NMF and the hidden Markov model (HMM) are combined to model the temporal aspects in the NMF [3, 8, 9]. In order to develop a blind source separation or speech enhancement algorithm in this case, the models for the two considered signals should be combined to form a factorial HMM. Therefore, even though these approaches are quite successful in modeling temporal dependencies, they are too computationally expensive for an on-line algorithm. Moreover, the temporal modeling in these methods cannot go beyond the first order Markov chain because of computational issues.

Bayesian NMF approaches can also provide an alternative way to derive more meaningful factorizations for audio signals. A linear minimum mean square error (LMMSE) estimator was proposed in [10] for speech enhancement where the temporal dynamics were used in filter construction. In [11], an on-line speech enhancement algorithm was proposed in which temporal aspects of the data were used to obtain informative prior distributions to be applied in a Bayesian NMF framework.

In this paper, we propose filtering and smoothing algorithms for NMF strategies that are motivated by Kalman filtering and smoothing. We assume that the NMF coefficients are stochastic processes, and that they evolve through a vector autoregressive (VAR) model over time. Therefore, in addition to the basis matrix, there will be some regression parameters associated with each signal. The proposed algorithm (for both filtering and smoothing) has two steps. First, we predict the current frame's NMF coefficients given either past observations (in filtering) or both past and future observations (in smoothing), and second, we update the estimates given the current observation. We propose a multiplicative update step of the estimates that can be interpreted using the HMM terminology. The proposed scheme introduces a new way of thinking about the problem that has not been considered in the current literature. We demonstrate the strength of our method using both synthetic examples and real applications including denoising and speech source separation.

2. PROPOSED METHOD

In this section, we present the proposed approach for a probabilistic NMF in the context of probabilistic latent component analysis (PLCA) [12]. In Subsection 2.1, we review the basic PLCA model

[*]This work was performed while at the University of Illinois.

and define the required notations. The proposed approach is given in Subsection 2.2 for the filtering and in 2.3 for the smoothing problems, and finally, Subsection 2.4 illustrates how we can process a mixed signal with these techniques.

2.1. Background

PLCA is a probabilistic formulation of NMF in which the distribution of an input vector is approximated as a convex combination of some weighted marginal distributions. A latent variable is defined to refer to the index of the underlying mixture component that has generated an observation, and the probabilities of different outcomes of this latent variable determine the weights in the mixture.

We denote the magnitude spectrogram of the speech by a random matrix $\mathbf{X}$ with elements X_{ft} where f is the frequency index and t is the time index, and the realizations by $\mathbf{x} = [x_{ft}]$. Also, we refer to the t-th column of $\mathbf{X}$ by $\mathbf{X}_t$. The random vector $\mathbf{X}_t$ is assumed to be distributed according to a multinomial distribution [13] whose parameter vector is denoted by $\boldsymbol{\theta}_t$, with the expected value given as: $E(\mathbf{X}_t) = \gamma_t \boldsymbol{\theta}_t$. Here, $\gamma_t = \sum_f x_{ft}$ is the total number of draws from the distribution at time t. The f-th element of $\boldsymbol{\theta}_t$ (θ_{ft}) indicates the probability that f-th row of $\mathbf{X}_t$ will be chosen in a particular draw from the multinomial distribution.

Let us define the scalar random variable Φ_t that can take one of the F possible frequency indices $f = 1, \dots F$ as its outcome. The f-th element of $\boldsymbol{\theta}_t$ is now given as: $\theta_{ft} = p(\Phi_t = f)$. Also, let V_t denote a scalar random latent variable that can take one of the I possible discrete values $i = 1, \dots I$. Using the conditional probabilities, $p(\Phi_t = f)$ is given by

$$\theta_{ft} = p\left(\Phi_t = f\right) = \sum_{i=1}^{I} p\left(\Phi_t = f \mid V_t = i\right) p\left(V_t = i\right). \quad (1)$$

We define a coefficient matrix $\mathbf{v}$ with elements $v_{it} = p(V_t = i)$, and a basis matrix $\mathbf{b}$ with elements $b_{fi} = p(\Phi_t = f \mid V_t = i)$. In principle, $\mathbf{b}$ is time-invariant and includes the possible spectral structures of the speech signal. Eq. (1) is now equivalently written as: $\boldsymbol{\theta}_t = \mathbf{b}\mathbf{v}_t$.

An observed spectrogram $\mathbf{x}$ can be approximated as the expected value of the underlying multinomial distribution as $\mathbf{x}_t \approx E(\mathbf{X}_t) = \gamma_t \boldsymbol{\theta}_t$. Consequently, the nonnegative factorization is written as: $\mathbf{x}_t \approx \gamma_t \mathbf{b}\mathbf{v}_t$ or $\mathbf{x}_t = \gamma_t \mathbf{b}\mathbf{v}_t + \mathbf{w}_t$ where $\mathbf{w}_t$ is an additive noise.

The basis and coefficient matrices ($\mathbf{b}$ and $\mathbf{v}$) can be estimated using the expectation-maximization (EM) algorithm [13]. The iterative update rules are given by:

$$v_{it} \leftarrow \frac{v_{it} \sum_f b_{fi} \left(x_{ft} / \hat{x}_{ft}\right)}{\sum_i v_{it} \sum_f b_{fi} \left(x_{ft} / \hat{x}_{ft}\right)} \ , \quad (2)$$

$$b_{fi} \leftarrow \frac{b_{fi} \sum_t v_{it} \left(x_{ft} / \hat{x}_{ft}\right)}{\sum_f b_{fi} \sum_t v_{it} \left(x_{ft} / \hat{x}_{ft}\right)} \ , \quad (3)$$

where $\hat{\mathbf{x}}_t = \gamma_t \mathbf{b}\mathbf{v}_t$ is the model approximation that is updated after each iteration. Note that, given the basis matrix $\mathbf{b}$ in (2), the update equation of $\mathbf{v}_t$ is independent of all the other time instances. Therefore, the time dependencies can not be modeled using (2).

2.2. Filtering

The goal of the proposed filtering approach is to develop an on-line algorithm to estimate a coefficient vector $\mathbf{v}_t$ given all the current and past observations, which are denoted by $\mathbf{x}_1^t = \{\mathbf{x}_1, \dots \mathbf{x}_t\}$. Here, we assume that the basis matrix $\mathbf{b}$ is obtained using some training data and is kept fixed thereafter. We assume that the coefficient vectors are modeled by an M-th order vector autoregressive (VAR) model as:

$$\mathbf{v}_t = \sum_{m=1}^{M} A_m \mathbf{v}_{t-m} + \mathbf{u}_t \ , \quad (4)$$

$$\mathbf{x}_t = \gamma_t \mathbf{b}\mathbf{v}_t + \mathbf{w}_t \ , \quad (5)$$

where A_m is the $I \times I$ autoregressive coefficient matrix associated with m-th lag, $\mathbf{u}_t$ is the process noise, and $\mathbf{w}_t$ is the observation noise in the model.

Even though (4) and (5) represent a complete state-space model that can be easily converted to a first order VAR model, nonnegativity of $\mathbf{v}_t$ and $\mathbf{x}_t$ prohibits the direct application of Kalman filtering. Following, we present an alternative approach that has a prediction and an update step as with Kalman filtering. The prediction of the coefficient vector $\mathbf{v}_t$, given $\mathbf{x}_1^{t-1}$, is denoted by $\hat{\mathbf{v}}_{t|t-1}$ and is simply obtained as:

$$\hat{\mathbf{v}}_{t|t-1} = \sum_{m=1}^{M} A_m \hat{\mathbf{v}}_{t-m|t-m} \ , \quad (6)$$

where $\hat{\mathbf{v}}_{t-m|t-m}$ is the updated estimate of $\mathbf{v}_{t-m}$ given $\mathbf{x}_1^{t-m}$. As the update step, the basic PLCA model is applied (by iterating (2)) to obtain the correction term that is denoted by $\tilde{\mathbf{v}}_t$. Now, we update the estimate of $\mathbf{v}_t$ as

$$\hat{\mathbf{v}}_{t|t} = \frac{\left(\hat{\mathbf{v}}_{t|t-1}\right)^{\beta} \odot \tilde{\mathbf{v}}_t}{\sum_i \left(\hat{\mathbf{v}}_{t|t-1}\right)^{\beta} \odot \tilde{\mathbf{v}}_t} \ , \quad (7)$$

where $(\cdot)^{\beta}$ and $\odot$ denote element-wise power and product operators, respectively, β is the prior strength and might not be equal one, and the normalization is performed to ensure that $\hat{\mathbf{v}}_{t|t}$ is a probability vector. $\tilde{\mathbf{v}}_t$ is a probability vector where each of its elements is proportional to the similarity between the corresponding basis vector and the observation $\mathbf{x}_t$. The multiplicative update in (7) is similar to the forward algorithm in an HMM, where the observation likelihood is replaced with $\tilde{\mathbf{v}}_t$. Therefore, $\hat{\mathbf{v}}_{t|t}$ can also be seen as the posterior probability of the latent variables (hidden states in the HMM).

The VAR coefficients $A_m, m = 1, \dots M$, can be estimated in different ways (e.g., [14, ch. 11]). In this paper, we carry out a suboptimal approach to estimate these matrices for simplicity. Let $\mathbf{v}^{(m)}$ denote the matrix $\mathbf{v}$, in which the columns are shifted by m, i.e. $v_{i,t}^{(m)} = v_{i,t+m}$. Then, A_m is estimated as $A_m = \mathbf{v}^{(m)}\mathbf{v}^{\top}$ where $\top$ represents the matrix transpose. The columns of A_m are then normalized to sum to one, and hence, $A_m^{\top}$ can also be interpreted as a transition matrix in a multimatrix mixture transition distribution (MTD) model [15].

2.3. Smoothing

The smoothing problem arises when we want to estimate a coefficient vector $\mathbf{v}_t$ given both past and future data, i.e. $\mathbf{x}_1^T = \{\mathbf{x}_1, \dots \mathbf{x}_t, \mathbf{x}_{t+1}, \dots \mathbf{x}_T\}$, where T is the total number of observations. This estimate is referred to as $\hat{\mathbf{v}}_{t|T}$ (in contrast, the estimate using filtering was denoted by $\hat{\mathbf{v}}_{t|t}$ in (7)).

For this purpose, the PLCA algorithm is applied to $\mathbf{x}_1^T$ to find the coefficient matrix $\tilde{\mathbf{v}}$. Then, a forward prediction matrix with columns given by $\hat{\mathbf{v}}_{t|t-1}$ and a backward prediction matrix with

columns given by $\hat{\lambda}_{t|T}$ are obtained as:

$$\hat{\mathbf{v}}_{t|t-1} = \sum_{m=1}^{M} A_m \tilde{\mathbf{v}}_{t-m} \,, \tag{8}$$

$$\hat{\lambda}_{t|T} = \sum_{m=1}^{M} A_m^{\top} \tilde{\mathbf{v}}_{t+m} \,. \tag{9}$$

In principle, to evaluate (8) and (9) it suffices to have access to observations from $t - M$ through $t + M$. Therefore, the algorithm will introduce a delay of M short time frames. Since our estimation approach of the VAR model parameters makes $A_m^{\top}$ similar to a transition matrix, (9) can be seen as an adaption of the HMM backward algorithm [16]. The updated estimate of $\mathbf{v}_t$ is now given as:

$$\hat{\mathbf{v}}_{t|T} = \frac{\left(\hat{\lambda}_{t|T} \odot \hat{\mathbf{v}}_{t|t-1}\right)^{\beta} \odot \tilde{\mathbf{v}}_t}{\sum_i \left(\hat{\lambda}_{t|T} \odot \hat{\mathbf{v}}_{t|t-1}\right)^{\beta} \odot \tilde{\mathbf{v}}_t}. \tag{10}$$

2.4. Source Separation Using the Proposed Method

To separate unknown sources from a given mixture, we can learn the basis matrices and VAR coefficient matrices for all the involved sources off-line, and then concatenate them properly to model the mixed signal.

Denote the coefficient vector of the mixed signal by $\mathbf{v}_t$, which is estimated using (7) or (10). Let $\mathbf{x}_t \approx \sum_k \mathbf{s}_{k,t}$ be the observed mixture, where $\mathbf{s}_{k,t}$ represents the t-th column of the k-th source's spectrogram. The spectrogram of each source is estimated by

$$\hat{\mathbf{s}}_{k,t} = \frac{\mathbf{b}_{s_k} \mathbf{v}_{k,t}}{\sum_k \mathbf{b}_{s_k} \mathbf{v}_{k,t}} \odot \mathbf{x}_t \,, \tag{11}$$

where division is performed element-wise, $\mathbf{b}_{s_k}$ is the basis matrix of the k-th source, and $\mathbf{v}_{k,t}$ is a coefficient vector that includes a subset of the elements of $\mathbf{v}_t$ that are associated with $\mathbf{b}_{s_k}$. Eq. (11) is known as the Wiener reconstruction and is widely used with NMF-based source separation (e.g., [5]).

3. EXPERIMENTS AND RESULTS

The proposed filtering/smoothing and the basic PLCA algorithms were applied to three different problems. In this section, we present the results and discuss the effect of different model parameters on the performance. We used the magnitude spectrogram of speech and noise signals as the input to the algorithms. The separated/enhanced time-domain signals were obtained using the phase of the mixed input signal and the overlap-add procedure. In our experiments here we consider three tasks: the separation of structured speech signals, speech denoising, and source separation.

3.1. Separation of Speech and Its Time-reversed Version

We applied the smoothing algorithm (10) to a mixed signal where the mixture was obtained as the sum of a temporally structured speech signal (see Fig. 1) and its time-reversed version at a sampling rate of 8 kHz. The discrete Fourier transform (DFT) with a frame length of 128 ms, 75% overlap, and a Hann window was applied to obtain the magnitude spectrogram of the signals as the input to the NMF algorithms. 60 basis vectors were trained for each source and were used in PLCA and the proposed algorithm.

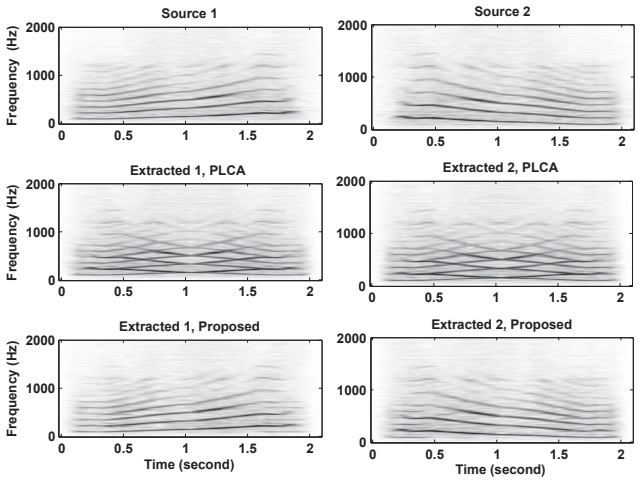

Fig. 1: Magnitude spectrograms of the original inputs (top row), the separated sources using PLCA (middle row) and the separated sources using the proposed algorithm (bottom row). For legibility reasons we only show the frequency range $0 \sim 2$ kHz.

The top panels of Fig. 1 show the spectrogram of the original signals. Since the basis matrices for the two source signals are effectively similar, basic PLCA or any other standard NMF algorithm will not be able to separate the sources. We see that by observing the separated sources which unfortunately closely resemble the mixture signal (see second row panels in Fig. 1). The bottom panels of Fig. 1 show the extracted source spectrograms using (10), which are obtained using parameters $M = 4$ and $\beta = 1$. Because there is a specific temporal structure that the two sources have (either ascending or descending pitch), we can tell the two sounds apart despite the fact that they have spectrally identical basis matrices. This experiment verifies the benefit of temporal modeling in a difficult separation task. The separation performance in this case is around 11 dB improvement in source to distortion ratio (SDR) [17], while the basic PLCA leads to only 0.5 dB improvement, which is effectively no separation.

3.2. Speech Denoising

We consider a noise reduction application where the desired speech signal is corrupted by an additive noise. A speaker-dependent approach is followed here in which a separate basis matrix is trained for each speaker and each noise type beforehand. The experiment was done for 100 randomly chosen speakers with different genders from the TIMIT database [18], where 9 out of the 10 available sentences were used for training speech model and the other sentence was used for testing.

The denoising algorithms were evaluated for two babble and factory noises taken from the NOISEX-92 database [19]. All the signals were down-sampled to 16 kHz. The frame length and overlap length in the DFT analysis were set to 64 and 60 ms, respectively. We learned 60 basis vectors for speech and 20 and 30 basis vectors for babble and factory noises, respectively.

First, we start by presenting an overall result of the denoising performances for both the smoothing and filtering algorithms. Since speech and noise signals have different temporal characteristics, we chose to use different powers (β) in (7) or (10) for speech (β_{speech}) and noise (β_{noise}) coefficients. These should be set experimentally,

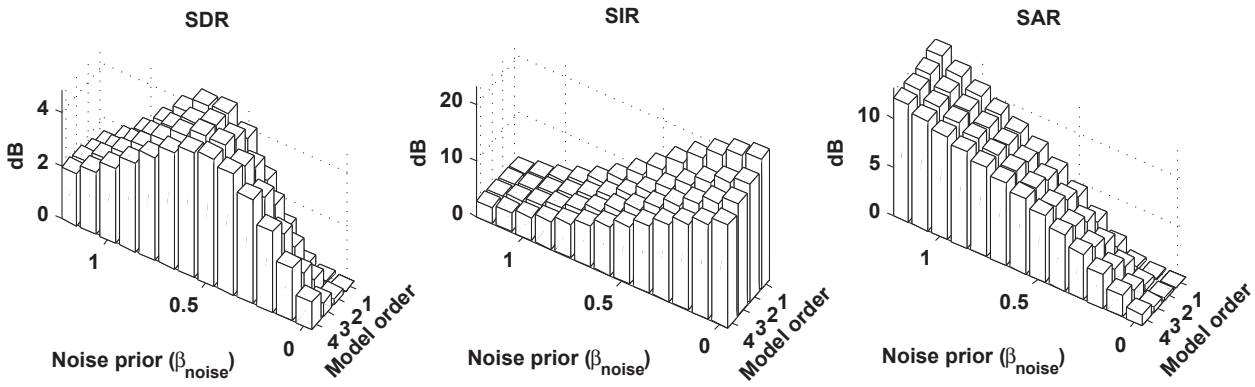

Fig. 2: Effect of the VAR model order and noise prior strength on the performance of speech denoising with the smoothing algorithm.

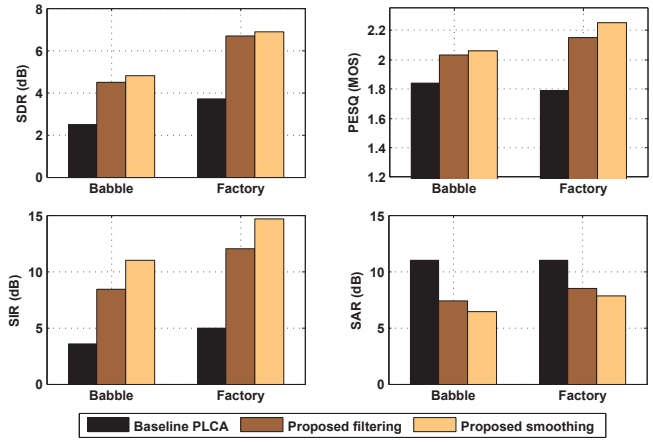

Fig. 3: Performance of denoising algorithms for a noisy signal at a 0 dB input SNR.

and we will discuss it shortly using Fig. 2. The performance is measured using SDR, source to interference ratio (SIR), and source to artifact ratio (SAR) [17]. We also evaluated the perceptual quality of the enhanced speech using PESQ [20]. Fig. 3 presents the results for a noisy signal at a 0 dB input signal to noise ratio (SNR), where we have used $M = 1$, $\beta_{\text{speech}} = 0.5$, $\beta_{\text{noise}} = 0.2$ for filtering, and $\beta_{\text{speech}} = 0.9$, $\beta_{\text{noise}} = 0.6$ for smoothing.

The results show a significant improvement in SDR, which results in better overall quality of demised speech, as compared to the baseline PLCA. Moreover, the evaluation shows that applying the temporal dynamics has increased the SIR whereas the SAR was reduced compared to the baseline. In fact, the algorithms have led to a fair trade-off between removing noise and introducing artifacts in the enhanced signal. The PESQ values also confirm a very good quality improvement using the proposed algorithms. Specifically in the case of the factory noise and with the smoothing algorithm, PESQ is improved by around 0.5 MOS compared to the baseline. Additionally, the figure illustrates that the smoothing algorithm has produced slightly better SDR and PESQ values than the filtering approach.

Finally, let us consider the smoothing approach applied to the babble case and study the effect of the model order (M) and prior strength (β) on the performance. Fig. 2 shows three objective measures as functions of the model order ($M = 1, 2, 3, 4$) and noise prior strength (β_{noise}) while $\beta_{\text{speech}} = 0.9$. As the figure shows, increasing the model order from 1 to 4 has not changed the peak performance. However, it has made the algorithm more robust to the value of β_{noise}. Also, the previously used $\beta_{\text{noise}} = 0.6$ falls into the optimal range of β_{noise}.

Table 1: Performance of the algorithms for speech source separation.

Algorithm	SDR (dB)	SIR (dB)	SAR (dB)
Baseline PLCA	4.8	8.5	**8.2**
Filtering	5.5	11	7.8
Smoothing	**5.7**	**12.5**	7.5

3.3. Speech Source Separation

The last application we consider here is monaural speech source separation. We applied the proposed algorithms to 50 mixture signals for randomly-chosen different-gender speaker pair 0dB mixtures from the TIMIT database. The DFT analysis and the setting of model parameters including the number of speech basis vectors, M, and β_{speech} were done as described in Subsection 3.2.

Table 1 summarizes the results in terms of BSS-EVAL measures [17]. Including the temporal dynamics has increased SIR but reduced SAR compared to the baseline. This is consistent with what was also observed in noise reduction in 3.2. In this case, the reduction in SAR is small and almost negligible while the SIR improvement is significant. Considering the SDR as a measure of overall speech quality, the evaluation shows that the performance has increased up to 0.9 dB due to the smoothing algorithm.

4. CONCLUSION

In this paper we introduced an approach to take advantage of temporal dependencies of sounds when performing NMF-style denoising and separation. Although we developed the algorithm using the PLCA terminology, adaption of the scheme to NMF and its variants is straightforward. The proposed two-step estimation approach for the NMF coefficients makes use of both temporal continuity and fidelity of an observation at a given time instant. We demonstrated the improvements that we obtained by the developed method in various applications using experimental means. Noticeably, we showed that our method can lead to improved results in source separation even when the basis matrices of the two underlying sources are practically the same. This allows us to attack mixture problems with sources that can be very similar in spectral characteristics and discernible only through their temporal structure.

5. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Process. Systems Conf. (NIPS)*, 2000, pp. 556–562.

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

[3] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, may 2013.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, apr. 2008, pp. 2069–2072.

[5] 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, pp. 793–830, 2009.

[6] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 411–414.

[7] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new approach for speech enhancement based on a constrained nonnegative matrix factorization," in *IEEE Int. Symp. on Intelligent Signal Process. and Communication Systems (ISPACS)*, dec. 2011, pp. 1–5.

[8] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2009, pp. 121–124.

[9] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may. 2011, pp. 17–20.

[10] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, 2011, pp. 45–48.

[11] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2012, pp. 4561–4564.

[12] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Advances in models for acoustic processing workshop, NIPS*, 2006.

[13] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Proc. Neural Information Process. Systems Conf. (NIPS)*, 2007.

[14] J. D. Hamilton, *Time series analysis*. New Jersey: Princeton University Press, 1994.

[15] A. Berchtold and A. E. Raftery, "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series," *Statistical Science*, vol. 17, no. 3, pp. 328–356, 2002.

[16] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," U.C. Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.

[17] 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.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus." Philadelphia: Linguistic Data Consortium, 1993.

[19] A. Varga and H. J. 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, jul. 1993.

[20] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.