Layered Nonnegative Matrix Factorization for Speech Separation

Chung-Chien Hsu, Jen-Tzung Chien, and Tai-Shih Chi

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

hsu.chung.chien@gmail.com, jtchien@nctu.edu.tw, tschi@mail.nctu.edu.tw

Abstract

This paper proposes a layered nonnegative matrix factorization (L-NMF) algorithm for speech separation. The standard NMF method extracts parts-based bases out of nonnegative training data and is often used to separate mixed spectrograms. The proposed L-NMF algorithm comprises of several layers of standard NMF blocks. During training, each layer of the L-NMF is initialized separately and then fine-tuned by minimizing the propagated reconstruction error. More complicated bases of the training data are emerged in deeper layers of the L-NMF by progressively combining parts-based bases extracted in the first layer. In other words, these complicated bases contain collective information of the parts-based bases. The bases deciphered by all layers are then used to separate spectrograms in the conventional NMF way. Simulation results show the proposed L-NMF outperforms the standard NMF in terms of the source-to-distortion ratio (SDR).

Index Terms: Layered NMF, dictionary learning, NMF, speech separation

1. Introduction

The nonnegative matrix factorization (NMF) decomposes a nonnegative matrix X into a product of a nonnegative dictionary (or basis) matrix W and a nonnegative weight (activation) matrix H, such that $X \approx WH$. The NMF is motivated by the visual perception and its computational model produces a parts-based representation [1]. It has been successfully utilized in a wide range of research areas including, but not limited to, face recognition [2], blind source separation [3], document clustering [4], and EEG classification [5].

Extended from the basic NMF [1], many NMF variants have been presented. For instance, the sparse NMF (SNMF) was proposed to learn sparse representation for solving overcomplete problem [6]. Recently, the graph regularized NMF (GNMF) was proposed to take the intrinsic geometric structure into consideration [7]. For audio signals, NMF can be directly applied to the Fourier magnitude spectrogram or its variants. For supervised speech separation, the convolutive NMF (CNMF) [8] and was proposed to decipher the phone-like bases by considering the dependencies across adjacent columns of input magnitude spectrogram. Furthermore, the NMF was optimized in a discriminative way for source separation [9]. To decode the harmonic structure, the 2-dimensional CNMF was proposed to identify musical notes for blind music separation [10].

From a probabilistic perspective, the nonnegative elements can be considered drawn from an underlying probability distribution. By using expectation-maximization (EM) algorithm, the model of probabilistic latent variables can decompose the probability of given nonnegative data into a product of two conditional probabilities given latent variables [11]. More generally, a Bayesian NMF was proposed for image feature extraction by assuming Gaussian likelihood and Exponential prior. The model inference is then approximated via Gibbs sampling [12]. Another Bayesian NMF based on Poisson likelihood and Exponential prior was proposed for speech-music separation and the Bayesian inference was approximated by using variational Bayesian EM (VBEM) algorithm [13].

Deep learning has emerged as a powerful machine learning approach in recent years. Furthermore, it produces state-of-the-art results in many research fields such as speech recognition [14] and image recognition [15]. Deep learning uses a hierarchical architecture to grasp high-level information [16] in data for various classification and regression tasks. In recent years, there are attempts to incorporate the hierarchical architecture into the standard NMF, which has the single-layer architecture. For instance, the multi-layer NMF was proposed as a sequential factorization for any NMF variants [17]. However, the reconstruction error will increase with increasing number of layers due to the lack of error correction procedures through all layers. In [18], the deep semi-NMF was proposed to learn a hierarchical representation of features from an image dataset for attribute-based clustering. However, the nonnegative constraint, which only allows additive combinations of bases and is not enforced in [18], might be needed for speech separation. In [19], extended from [9], the iterative inference procedures of the NMF were unfolded to mimic the architecture of the deep neural network (DNN), however, its underlying structure is still single-layer. Nevertheless, these works [17][18][20] tried to generalize the standard NMF to decompose a nonnegative matrix into a product of a series of matrices.

In the same spirit but with different details, we propose a layered nonnegative matrix factorization (L-NMF) algorithm for single-channel speech separation in this paper. The proposed algorithm consists of several layers of NMFs and is implemented by first stacking standard NMFs and then correcting the propagated reconstruction error. The L-NMF provides a way for realizing more complex bases by combining sparse parts-based bases extracted by the single layer NMF. We assume these more complex bases will give a better description to the magnitude spectrogram and hence improve the separation results.

The rest of the paper is organized as follows. Section 2 gives a brief review of the standard NMF and describes our proposed L-NMF. In section 3, we demonstrate the speech bases learned from the proposed L-NMF with different numbers of layers and the corresponding performance in speech separation. We end in section 4 with some conclusions and future work.

2. Proposed Method

In this section, we review the standard NMF and introduce our proposed L-NMF.

2.1. Standard NMF

NMF was proposed in the year 2000 [1] and has been successfully used in many applications. Given a nonnegative data matrix $\mathbf{X} \in \mathcal{R}_+^{M\times N}$, NMF aims to decompose the data matrix into a product of two nonnegative matrices $\mathbf{W} \in \mathcal{R}_+^{M\times K}$ and $\mathbf{H} \in \mathcal{R}_+^{K\times N}$ as follows

$$X_{mn} \approx [\mathbf{WH}]_{mn} = \sum_k W_{mk}H_{kn} \tag{1}$$

The NMF decomposition is optimized by minimizing the reconstruction error between the observed data $\mathbf{X}$ and its reconstruction $\mathbf{WH}$ as follows

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \mathcal{D}\left(\mathbf{X} \parallel \mathbf{WH}\right) \tag{2}$$

where $\mathcal{D}$ is the defined cost function, which can be Euclidean distance, Kullback-Leibler (KL) divergence , Itakura-Saito (IS) divergence, and so on. Many algorithms were proposed by performing alternating minimization. Multiplicative update rules are simple and efficient in inferring the model parameters $\{\mathbf{W}, \mathbf{H}\}$ as follows

$$\begin{aligned} \mathbf{W} &\leftarrow \mathbf{W} \otimes \frac{[\nabla_{\mathbf{W}}\mathcal{D}]^-}{[\nabla_{\mathbf{W}}\mathcal{D}]^+} \\ \mathbf{H} &\leftarrow \mathbf{H} \otimes \frac{[\nabla_{\mathbf{H}}\mathcal{D}]^-}{[\nabla_{\mathbf{H}}\mathcal{D}]^+} \end{aligned} \tag{3}$$

where $\otimes$ denotes a element-wise multiplication; the division is also element-wise; $[\nabla_{\mathbf{W}}\mathcal{D}]^+$ and $[\nabla_{\mathbf{W}}\mathcal{D}]^-$ indicate the positive and negative parts of the gradient with respect to $\mathbf{W}$, respectively. Similarly, $[\nabla_{\mathbf{H}}\mathcal{D}]^+$ and $[\nabla_{\mathbf{H}}\mathcal{D}]^+$ are the positive and negative parts of the gradient with respect to $\mathbf{H}$, respectively.

2.2. Layered NMF

The generalized form of the factor analysis can be written as [21]

$$\mathbf{X} \approx \hat{\mathbf{X}} = g\left(\mathbf{W}^{(1)}g\left(\mathbf{W}^{(2)}g\left(\cdots g(\mathbf{W}^{(L)}\mathbf{H}^{(L)})\right)\right)\right) \tag{4}$$

where $g(\cdot)$ is a non-linear function. The standard NMF can then be thought as a simple and shallow version of the factor analysis with the nonnegative constraints and a linear function $g(\cdot)$. The layered NMF with L layers can be written as follows

$$\mathbf{X} \approx \hat{\mathbf{X}} = \left(\prod_{l=1}^{L}\mathbf{W}^{(l)}\right)\mathbf{H}^{(L)} \tag{5}$$

This formulation depicts a hierarchical architecture with L layers as shown in Fig. 1. To construct the model, the model factors $\mathbf{W}^{(l)}$ and $\mathbf{H}^{(L)}$ are initialized layer by layer. In the first step, the standard (single-layer) NMF is performed to factorize $\mathbf{X} \approx \mathbf{W}^{(1)}\mathbf{H}^{(1)}$, where $\mathbf{X} \in \mathcal{R}_+^{M\times N}$, $\mathbf{W}^{(1)} \in \mathcal{R}_+^{M\times K_1}$, and $\mathbf{H}^{(1)} \in \mathcal{R}_+^{K_1\times N}$. Then the same factorization is performed on the results obtained from the first step as $\mathbf{H}^{(1)} \approx \mathbf{W}^{(2)}\mathbf{H}^{(2)}$, where $\mathbf{W}^{(2)} \in \mathcal{R}_+^{K_1\times K_2}$ and $\mathbf{H}^{(2)} \in \mathcal{R}_+^{K_2\times N}$. We continue the same procedure to pre-train all layers. After the initialization, we have to fine-tune the parameters of all layers, $\mathbf{W}^{(l)}$ and $\mathbf{H}^{(L)}$, to reduce the total reconstruction error as follows

$$\min_{\mathbf{W}^{(l)},\mathbf{H}^{(L)}\geq 0} \mathcal{D}\left(\mathbf{X} \parallel \left(\prod_{l=1}^{L}\mathbf{W}^{(l)}\right)\mathbf{H}^{(L)}\right), \ \forall\, l = 1, \cdots, L \tag{6}$$

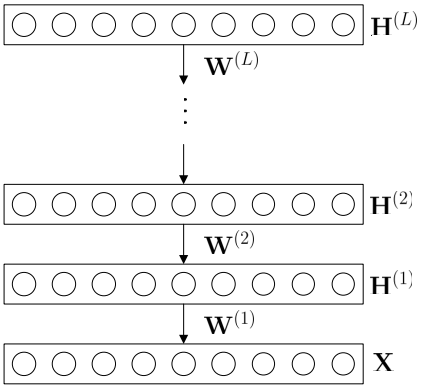

Figure 1: *Layered NMF (L-NMF) model.*

Similar to the standard NMF, the multiplicative update rules for all layers are derived as follows

$$\begin{aligned} \mathbf{W}^{(l)} &\leftarrow \mathbf{W}^{(l)} \otimes \frac{[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^-}{[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^+}, \ \forall\, l = 1, \cdots, L \\ \mathbf{H}^{(L)} &\leftarrow \mathbf{H}^{(L)} \otimes \frac{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^-}{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^+} \end{aligned} \tag{7}$$

where $[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^+$ and $[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^-$ denote the positive and negative parts of the gradient with respect to each layer $\mathbf{W}^{(l)}$; $[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^+$ and $[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^-$ denote the positive and negative parts of the gradient with respect to $\mathbf{H}^{(L)}$.

For instance, if the Euclidean distance is chosen as the cost function and the number of layers L is set to 2, the multiplicative update rules are

$$\begin{aligned} \mathbf{W}^{(1)} &\leftarrow \mathbf{W}^{(1)} \otimes \frac{\mathbf{X}\mathbf{H}^{(2)\mathrm{T}}\mathbf{W}^{(2)\mathrm{T}}}{\mathbf{W}^{(1)}\mathbf{W}^{(2)}\mathbf{H}^{(2)}\mathbf{H}^{(2)\mathrm{T}}\mathbf{W}^{(2)\mathrm{T}}} \\ \mathbf{W}^{(2)} &\leftarrow \mathbf{W}^{(2)} \otimes \frac{\mathbf{W}^{(1)\mathrm{T}}\mathbf{V}\mathbf{H}^{(2)\mathrm{T}}}{\mathbf{W}^{(1)\mathrm{T}}\mathbf{W}^{(1)}\mathbf{W}^{(2)}\mathbf{H}^{(2)}\mathbf{H}^{(2)\mathrm{T}}} \\ \mathbf{H}^{(2)} &\leftarrow \mathbf{H}^{(2)} \otimes \frac{\mathbf{W}^{(2)\mathrm{T}}\mathbf{W}^{(1)\mathrm{T}}\mathbf{X}}{\mathbf{W}^{(2)\mathrm{T}}\mathbf{W}^{(1)\mathrm{T}}\mathbf{W}^{(1)}\mathbf{W}^{(2)}\mathbf{H}^{(2)}} \end{aligned} \tag{8}$$

If a nonlinear function $g(\cdot)$ is used in the model, the update rules can be easily modified by applying the chain rule of the gradient. The pseudo code of the proposed L-NMF is given in Algorithm 1.

3. Experiments and Results

3.1. Learning Bases

In the first experiment, we examined the speech bases derived by the standard NMF and the proposed L-NMF algorithms. The Fourier magnitude spectrograms of the sentences spoken by a female speaker in TIMIT corpus [22] were concatenated for learning the speech bases $\mathbf{W}$. Fig. 2(a) depicts the thirty bases learned by the standard NMF (as the L-NMF with the parameters of $L = 1$ and $K = 30$). Similar to results demonstrated in the literature, we can observe that $\mathbf{W}$ contains parts-based information of speech spectra. Fig. 2(b) and (c) respectively show the basis matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(1)}\mathbf{W}^{(2)}$ learned by the proposed L-NMF with parameters of $L = 2$, $K_1 = 180$ and $K_2 = 30$. Comparing Fig. 2(c) with Fig. 2(a), the thirty bases deciphered by $\mathbf{W}^{(1)}\mathbf{W}^{(2)}$ contains more complex structures, which are formed by linearly combining parts-based

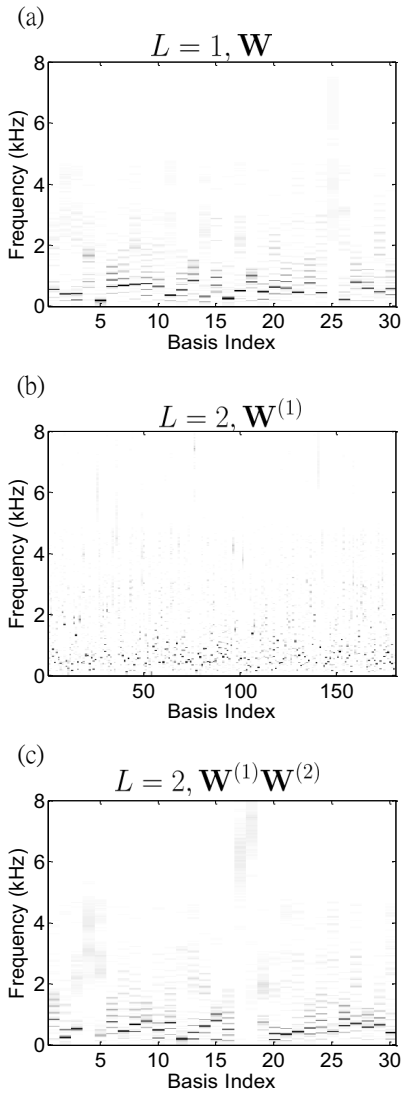

Figure 2: *Speech bases learned from NMF and L-NMF; (a)* $\mathbf{W}$*; (b)* $\mathbf{W}^{(1)}$*; (c)* $\mathbf{W}^{(1)}\mathbf{W}^{(2)}$.

bases $\mathbf{W}^{(1)}$ using the factor $\mathbf{W}^{(2)}$. For examples, the 4th, 17th and the 18th bases shown in Fig. 2(c) contain clear structures only above 4 kHz, resembling the spectra of three types of consonants. In contrast, only one consonant basis above 4 kHz is obtained by the standard NMF as the 25th basis in Fig. 2(a). Furthermore, we also tried L-NMF with the parameter $L = 3$. Unfortunately, it didn't produce bases with significant difference but required a great deal of computations. Therefore, we used the L-NMF with the parameter $L = 2$ for speech separation tasks.

3.2. Speech Separation

The second experiment was a supervised speaker-dependent speech separation task. Two sets of mixtures, which contain sentences from one female and one male speakers in TIMIT corpus (fcjf0 & mcpm0, fdaw0 & mdac0), were selected as test materials to evaluate the proposed L-NMF algorithm. The 1024-point short-term Fourier transform (STFT) with a 40-ms

Algorithm 1 Layered NMF

Require: $\mathbf{X} \in \mathcal{R}_{+}^{M\times N}$, number of layers ($L$), and sizes of each layer ($K_1, \cdots, K_L$)

Initialize layers
for $l = 1$ to $L-1$ **do**
  if $l = 1$ **then**
    $\mathbf{W}^{(l)}, \mathbf{H}^{(l)} = \text{argmin}_{\mathbf{W}^{(l)},\mathbf{H}^{(l)}\geq 0}\mathcal{D}\left(\mathbf{X} \parallel \mathbf{W}^{(l)}\mathbf{H}^{(l)}\right)$
  else
    $\mathbf{W}^{(l+1)}, \mathbf{H}^{(l+1)} =$
      $\text{argmin}_{\mathbf{W}^{(l+1)},\mathbf{H}^{(l+1)}\geq 0}\,\mathcal{D}\left(\mathbf{H}^{(l)} \parallel \mathbf{W}^{(l+1)}\mathbf{H}^{(l+1)}\right)$
  end if
end for

Fine-tune across all layers
repeat
for $l = 1$ **to** L **do**
  $\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} \otimes \frac{[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^{-}}{[\nabla_{\mathbf{W}^{(l)}}\mathcal{D}]^{+}},\ \forall\, l = 1, \cdots, L$
  $\mathbf{H}^{(L)} \leftarrow \mathbf{H}^{(L)} \otimes \frac{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^{-}}{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^{+}}$
end for
until Converge
return $\mathbf{W}^{(l)}, \mathbf{H}^{(L)}$

frame duration and a 10-ms frame shift was calculated to obtain the Fourier magnitude spectrograms. For speech separation using the standard NMF, the NMF was applied to the magnitude spectrograms of training data for finding bases of the female speaker $\mathbf{W}_f$ and the male speaker $\mathbf{W}_m$ via $\mathbf{X}_f^{train} \approx \mathbf{W}_f\mathbf{H}_f$ and $\mathbf{X}_m^{train} \approx \mathbf{W}_m\mathbf{H}_m$, respectively. In the test phase, we aimed to separate the mixed spectrogram $\mathbf{X}_{mix}$ by estimating the weight matrix $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_f; \hat{\mathbf{H}}_m]$ with the trained bases $\mathbf{W}_f$ and $\mathbf{W}_m$ such that $\mathbf{X}_{mix} \approx [\mathbf{W}_f, \mathbf{W}_m]\hat{\mathbf{H}}$. The separated spectrograms were then obtained by applying the Wiener gain as follows

$$\begin{aligned}\hat{\mathbf{X}}_f &= \mathbf{X}_{mix} \otimes \frac{\mathbf{W}_f\hat{\mathbf{H}}_f}{[\mathbf{W}_f, \mathbf{W}_m]\hat{\mathbf{H}}} \\ \hat{\mathbf{X}}_m &= \mathbf{X}_{mix} \otimes \frac{\mathbf{W}_m\hat{\mathbf{H}}_m}{[\mathbf{W}_f, \mathbf{W}_m]\hat{\mathbf{H}}}\end{aligned} \tag{9}$$

Finally, the separated speech signals were obtained by inverting the separated spectrograms using the inverse STFT with the overlap-and-add technique. In our experiments, the numbers of trained bases of each speaker ranged from 10 to 50 [23].

In the simulations, each mixture was generated by adding one sentence from a female speaker and one different sentence from a male speaker. There were 20 test mixtures created from two female-male pairs (fcjf0 & mcpm0, fdaw0 & mdac0). All sentences were normalized to be with equal power. The remaining 9 sentences of each speaker were used for training. In our method, we set K_2 is six times of K_1 to extract the basis matrices of two layers for each speaker. Then $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ were multiplied together to form the final bases of the corresponding speaker. After bases were learned, the separation was done in the same way as by the conventional NMF-based method. The separation performance of the proposed algorithm was assessed using the source-to-distortion ratio (SDR) [24] and results of the standard NMF and the proposed L-NMF against numbers of bases are shown in Fig. 3. Five different random initializations to our L-NMF were performed when separating each

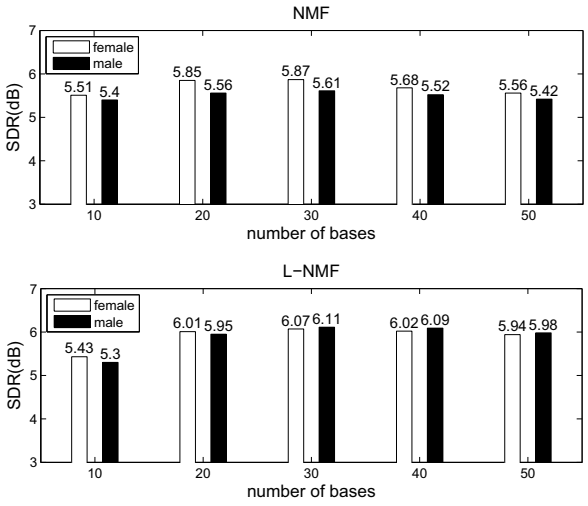

Figure 3: *Separation performance of NMF and L-NMF against number of bases in terms of the SDR measure.*

test mixture. Therefore, each SDR value in Fig. 3 was averaged over 100 test conditions (20 mixtures $\times$ 5 initializations). The number of bases of L-NMF is the number of columns in $\mathbf{W}^{(1)}\mathbf{W}^{(2)}$. As shown in Fig. 3, the proposed L-NMF beats the standard NMF by a fair margin on each of the separated speech signals of male and female speakers when the number of bases is larger than 20. In contrast, these two methods produce comparable SDR in $K = K_2 = 10$ condition, where the bases learned by both methods are quite similar.

4. Conclusions and Future Work

In this paper, we propose a layered nonnegative matrix factorization algorithm. Unlike the standard NMF, our method can realize more complex bases by combining sparse parts-based bases extracted by the single layer NMF to interpret the data differently. In speech separation experiments, our proposed method outperforms the standard NMF in terms of the source-to-distortion ratio. Comparing with the standard NMF, the proposed L-NMF consumes more computation time in training the bases. After bases are learned, both methods share the same computation time in the test phase.

In this paper, the size parameters (e.g. $K_1, K_2, \cdots, K_L$) of the hierarchical architecture of the proposed L-NMF algorithm are pre-determined in advance. In the future, we will extend our method to a Bayesian approach [13], which can regularize the model and automatically select the model parameters by given data. In addition, we believe the performance of speech separation can be further improved by adopting the discriminative dictionary learning criterion [9]. We are also interested in optimizing size parameters of the proposed L-NMF for audio recognition problems [25][26].

5. Acknowledgements

This research is supported by the Ministry of Science and Technology, Taiwan under Grant MOST 103-2220-E-009- 003 and the Biomedical Electronics Translational Research Center, National Chiao Tung University.

6. References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, pp. 556–562.

[2] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683–695, 2006.

[3] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[4] F. Shahnaz, M. W. Berry, V. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373 – 386, 2006.

[5] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 320–327.

[6] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[7] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[8] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[9] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. of ISCA Interspeech*, 2014, pp. 865–869.

[10] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation." in *ICA*, 2006, pp. 700–707.

[11] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, vol. Article ID 947438, 2008.

[12] M. Schmidt, O. Winther, and L. Hansen, "Bayesian non-negative matrix factorization," in *Independent Component Analysis and Signal Separation*, 2009, pp. 540–547.

[13] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian factorization and selection for speech and music separation," in *Proc. of ISCA Interspeech*, 2014, pp. 998–1002.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[16] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 1096–1104.

[17] A. Cichocki and R. Zdunek, "Multilayer nonnegative matrix factorization," *Electronics Letters*, vol. 42, pp. 947–948, 2006.

[18] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. Schuller, "A deep semi-NMF model for learning hidden representations," in *Proceedings of the International Conference on Machine Learning*, 2014.

[19] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. of ICASSP*, 2015, pp. 66 – 70.

[20] S. Lyu and X. Wang, "On algorithms for sparse multi-factor NMF," in *Advances in Neural Information Processing Systems*, 2013, pp. 602–610.

[21] J.-H. Oh and H. S. Seung, "Learning generative models with the up propagation algorithm," in *Advances in Neural Information Processing Systems*, 1998, pp. 605–611.

[22] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. of ICASSP*, 2014, pp. 1562–1566.

[24] 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.

[25] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 424–434, 2008.

[26] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proc. of EUSIPCO*, 2013, pp. 1–5.