LETTER

Target Source Separation Based on Discriminative Nonnegative Matrix Factorization Incorporating Cross-Reconstruction Error

Kisoo KWON[†], **Jong Won SHIN**[††], *Nonmembers, and* **Nam Soo KIM**[†a)], *Member*

SUMMARY Nonnegative matrix factorization (NMF) is an unsupervised technique to represent nonnegative data as linear combinations of nonnegative bases, which has shown impressive performance for source separation. However, its source separation performance degrades when one signal can also be described well with the bases for the interfering source signals. In this paper, we propose a discriminative NMF (DNMF) algorithm which exploits the reconstruction error for the interfering signals as well as the target signal based on target bases. The objective function for training the bases is constructed so as to yield high reconstruction error for the interfering source signals while guaranteeing low reconstruction error for the target source signals. Experiments show that the proposed method outperformed the standard NMF and another DNMF method in terms of both the perceptual evaluation of speech quality score and signal-to-distortion ratio in various noisy environments.

key words: *nonnegative matrix factorization, discriminative basis, cross-reconstruction error*

1. Introduction

Over the last few years, audio source separation has been one of the interesting topics in audio signal processing such as speech enhancement, speech recognition, music signal processing, and so on [1]–[9]. Data-representation methods and template-based approaches have been widely applied to audio source separation. They make the representation models or statistics from *a priori* information possibly available from a training database (DB) [1]–[9]. One notably successful techniques is based on nonnegative matrix factorization (NMF) [10]. NMF is a dimensionality reduction technique that usually leads to a part-based representation and has been shown to be effective for audio signals.

After being proposed by Lee and Seung [10], NMF was successfully applied to audio magnitude or power spectra analysis and has shown certain benefits over similar factorization schemes such as independent component analysis (ICA) and principal component analysis (PCA) [10], [11]. One of the possible reasons for these benefits is that NMF provides a framework for learning parts of dataset, and audio signal is suitable for a part-based representation in certain domains [12]. In NMF analysis, the input vector is represented by a linear combination of nonnegative basis vectors with nonnegative weights. After [10] was published, a number of attempts have been made to improve NMF in various conditions, which include sparse NMF [13], Itakura Saito-NMF [11], and convolutive NMF [14].

In the source separation task, the performance of the NMF-based techniques is limited when the subspaces that the bases for different sources span overlap. One possible interpretation for this is that the bases for each source are trained separately to reconstruct the individual signal faithfully without considering the source separation capability. To alleviate this problem, previous works tried to either modify the criterion of the NMF algorithm [6]–[8], [15], [16] or estimate the weights in such a way to consider the effect of mixed sources [17].

The former approaches are often called discriminative NMF (DNMF) [6]–[8], [15], [16]. Though sophisticatedly different from each other, these approaches generally aim to construct the basis vectors of a target source such that they can reconstruct the target signal even when the target source is mixed with interfering signals. In [6], the basis vectors of the target source are obtained with the constraint that they should be orthogonal to the basis vectors of the interfering sources. However, this orthogonality constraint may result in a high reconstruction error for the target source signals. In [7], the basis vectors of each source are separately updated by respective reconstruction error, while the encoding vectors are updated by the whole reconstruction error. In [8], the clean target source signal and the signal mixed with the other source signal are used during the training phase.

In this paper, we propose a novel approach to DNMF for which the criterion function for NMF training includes a term rewarding high reconstruction error for the interfering source signals in conjunction with a term for low reconstruction error for the target source signal. The proposed DNMF algorithm with cross-reconstruction error was applied to speech enhancement and showed improved performance in terms of both the ITU-T Recommendation P.862 perceptual evaluation of speech quality (PESQ) [18] and signal-to-distortion ratio (SDR) [19].

2. NMF-Based Audio Source Separation

When NMF is applied to audio source separation, it approximates the magnitude or power spectra of a given mixture $\mathbf{V} \in \mathbb{R}^{M\times N}$ as the product of a basis matrix $\mathbf{W} \in \mathbb{R}^{M\times r}$, and

Manuscript received May 13, 2015.
Manuscript revised July 29, 2015.
Manuscript publicized August 19, 2015.

[†]The authors are with the Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 151–742, Korea.

[††]The author is with the School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Korea..

a) E-mail: nkim@snu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2015EDL8114

an encoding matrix $\mathbf{H} \in \mathbb{R}^{r \times N}$ ($\mathbf{V} \approx \mathbf{WH}$) where M, N, and r denote the number of frequency bins, short-time frames, and the number of basis vectors, respectively. In this paper, we consider a simple speech enhancement scenario where the target source signal is speech and the interfering signal is the background noise. In this case, the basis matrix $\mathbf{W}$ is considered as a concatenation of the speech and noise basis matrices, $\mathbf{W}_S \in \mathbb{R}^{M \times r_s}$ and $\mathbf{W}_N \in \mathbb{R}^{M \times r_n}$ where r_s and r_n indicate the number of speech and noise basis vectors, respectively. $\mathbf{W}_S$ and $\mathbf{W}_N$ are usually trained separately with clean speech and noise DBs, respectively. If the Kullback-Leibler divergence (KL-divergence) and multiplicative update rules are used as a distance measure and an optimization method, the update rules for the encoding and basis matrices during the training phase are given as [10]

$$\mathbf{H}_i \leftarrow \mathbf{H}_i \otimes \frac{\mathbf{W}_i^T \frac{\mathbf{V}_i}{\mathbf{W}_i\mathbf{H}_i}}{\mathbf{W}_i^T \mathbf{1}}, \tag{1}$$

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \otimes \frac{\frac{\mathbf{V}_i}{\mathbf{W}_i\mathbf{H}_i} \mathbf{H}_i^T}{\mathbf{1}\mathbf{H}_i^T}. \tag{2}$$

where ^{T} denotes matrix transposition, and subscript i denotes either speech or noise signal, $\mathbf{V}_i \in \mathbb{R}^{M \times N_i}$ is the magnitude spectrogram of the training signal where N_i is the total number of short-time frames in the training data for source i, $\bigotimes$ and $\frac{a}{b}$ denote the element-wise multiplication and division of matrices, and $\mathbf{1}$ is a matrix of a proper size with all elements equal to one. $\mathbf{H}_i$ and $\mathbf{W}_i$ are obtained by iterative application of the update rules (1) and (2) for a fixed number of iterations.

In the separation phase, a noisy magnitude spectrum $|Y(t)|$ is approximated as $|Y(t)| \approx \mathbf{W}H(t)$ for each frame with the fixed basis matrix $\mathbf{W} = [\mathbf{W}_S \; \mathbf{W}_N]$ obtained during the training phase where $H(t) = [H_S(t)^T \; H_N(t)^T]^T \in \mathbb{R}^{(r_s+r_n)\times 1}$ denotes the encoding vector of the mixed signal in the t-th frame, $Y(t)$ is the short-time Fourier transform (STFT) coefficients of the noisy input, and $|\cdot|$ denotes taking element-wise magnitude. Keeping $\mathbf{W}$ fixed, $H(t)$ is computed by iterating (1) for a fixed number of times, in which $H_S(t)$ and $H_N(t)$ are initialized to nonnegative random numbers. After a fixed number of iterations, the magnitude spectra of the speech and noise signals are estimated as follows:

$$|\hat{S}(t)| = \mathbf{W}_S H_S(t), \qquad |\hat{N}(t)| = \mathbf{W}_N H_N(t). \tag{3}$$

Instead of directly using the estimated magnitude spectra in (3), a spectral gain function similar to the Wiener filter is adopted in [12] and [9]. In this scheme, the gain function is given by

$$G(t) = \frac{|\hat{S}(t)|^2}{|\hat{S}(t)|^2 + |\hat{N}(t)|^2}. \tag{4}$$

Finally, the STFT coefficients of the speech signal at the t-th frame are obtained according to $\hat{S}^{final}(t) = G(t) \bigotimes Y(t)$.

3. Discriminative NMF Incorporating Cross-Reconstruction Error

When the bases for each source are trained separately, the subspaces spanned by the bases for individual sources are not guaranteed to be disjoint. It implies that some data vectors from one source can be possibly well reconstructed by the bases for interfering sources. This may result in a degraded performance in the source separation.

One way to alleviate this issue is to modify the cost function of the NMF training. Although the modification may result in increased reconstruction error for each source, the source separation performance can be enhanced especially when there exists severe overlap between the subspaces spanned by different source bases.

In order to obtain discriminative bases, we propose an objective function for basis estimation in such a way that the reconstruction error of the target source computed from the interfering bases is also incorporated in conjunction with the conventional reconstruction error derived from the target bases. The reconstruction error of noise (interfering) signal based on the speech (target) bases may be considered as a measure for residual noise in speech enhancement. On the other hand, if the noise (interfering source) basis matrix is trained using the cross-reconstruction error along with the conventional reconstruction error, it can reduce the speech distortion caused by active noise bases. The objective function of the proposed method to train the basis matrix $\mathbf{W}_i$ where i indicates either speech or noise is given by

$$\begin{aligned} f(\mathbf{W}_i, \mathbf{H}_i, \mathbf{C}_j) &= \\ & D(\mathbf{V}_i \parallel \mathbf{W}_i\mathbf{H}_i) - \gamma_i D(\mathbf{V}_j \parallel \mathbf{W}_i\mathbf{C}_j) \\ & \gamma_i = \lambda_i \frac{\|\mathbf{V}_i\|_1}{\|\mathbf{V}_j\|_1} \end{aligned} \tag{5}$$

where $\mathbf{V}_i \in \mathbb{R}^{M \times N_i}$, $\mathbf{W}_i$, and $\mathbf{H}_i$ are the magnitude spectra of N_i frames, basis, and encoding matrix for the source signal for which we want to train a basis matrix, and $\mathbf{V}_j \in \mathbb{R}^{M \times N_j}$ and $\mathbf{C}_j$ are the magnitude spectrogram and encoding matrix for the other source signal of N_j frames. $D(a \parallel b)$ denotes the distance function between a and b, for which the KL-divergence is chosen, and $\|\cdot\|_1$ is an $l1$-norm of the vector constructed by concatenating the rows of the matrix. In the (5), γ_i makes a tradeoff between the self-reconstruction and cross-reconstruction errors after balancing the amount of each data set. It is noted that $\lambda_S = \lambda_N = 0$ corresponds to the standard NMF while $\lambda_S > 0$ and $\lambda_N > 0$ may enhance the source separation performance. The bases obtained with $\lambda_S > 0$ will cause less residual noise while the bases obtained with $\lambda_N > 0$ will result in less speech distortion than the standard NMF.

The update rules for $\mathbf{W}_i$, $\mathbf{H}_i$ and $\mathbf{C}_j$ can be derived in a similar way to (1) and (2) as follows:

$$\mathbf{H}_i \leftarrow \mathbf{H}_i \otimes \frac{\mathbf{W}_i^T \frac{\mathbf{V}_i}{\mathbf{W}_i\mathbf{H}_i}}{\mathbf{W}_i^T \mathbf{1}}, \quad \mathbf{C}_j \leftarrow \mathbf{C}_j \otimes \frac{\mathbf{W}_i^T \frac{\mathbf{V}_j}{\mathbf{W}_i\mathbf{C}_j}}{\mathbf{W}_i^T \mathbf{1}}, \tag{6}$$

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \otimes \frac{\frac{\mathbf{V}_i}{\mathbf{W}_i\mathbf{H}_i}\mathbf{H}_i^T}{\mathbf{1}\mathbf{H}_i^T + \gamma_i\left(\frac{\mathbf{V}_j}{\mathbf{W}_i\mathbf{H}_j}\mathbf{H}_j^T - \mathbf{1}\mathbf{C}_j^T\right)} \qquad (7)$$

A prominent difference of this training algorithm from the conventional NMF training is that both the speech and noise DBs are used to train the speech bases and vice versa.

In the separation phase, the speech and noise basis matrices obtained from the proposed method are used, and STFT coefficients of the speech source are finally estimated in the same way as in Sect .2.

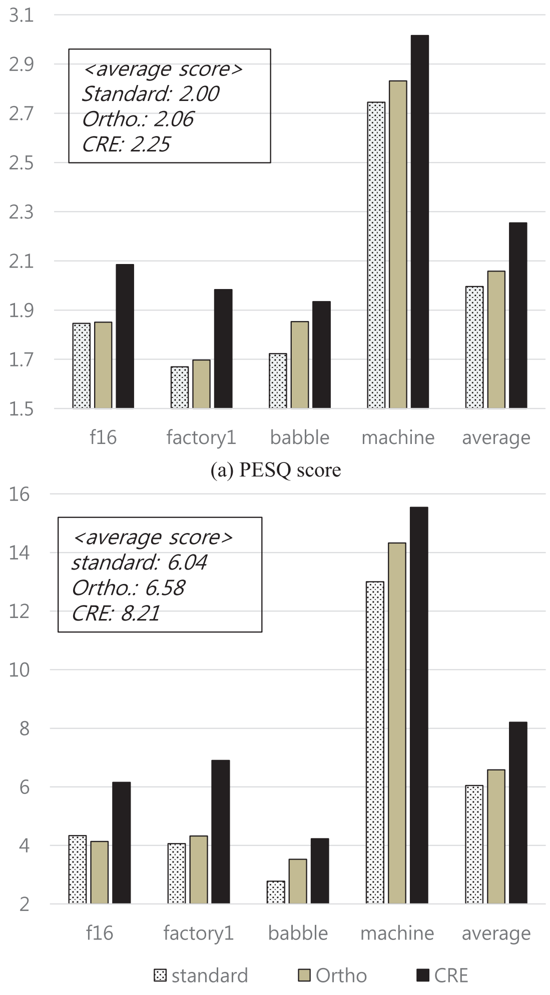
4. Experiments

To evaluate the performance of the proposed algorithm, speech enhancement was performed in a variety of noisy conditions. Speech and noise samples were selected from TIMIT and NOISEX-92 DBs, respectively, with a sampling rate of 16 kHz. A 512-point discrete Fourier transform with 75% overlap was used to form the spectrogram. The basis matrix for each noise type was obtained from about 120-second long noise waveforms which was not included in the test data, and the speech DB for training was 130-second long, which was spoken by 56 different speakers. The speech test data set consisted of 32 sentences from 32 different speakers. We tested 4 different types of noises including *F-16*, *factory1*, *babble* and *machinegun* noises. The number of bases r for each source was set to 128, which provided a good trade-off between the reconstruction error and the computational complexity.

The performance of the proposed method was evaluated in terms of PESQ [18] and SDR [19]. To demonstrate the performance improvement achieved by the proposed objective function, three speech enhancement systems for which only the basis matrices were trained in different ways were compared:

- *Standard*: the standard NMF training with KL-divergence and a multiplicative update rule without any additional penalty term [10]
- *Ortho*: the DNMF in [6] which tries to make bases for different sources orthogonal.
- *CRE*: the proposed method using the cross-reconstruction error.

To find proper values of λ_S and λ_N in (5), 4 speech utterances and noise signals which were not included in the test and training data were used for the validation process in which the performances were compared in terms of PESQ scores and SDRs. The optimal values of λ_i through this validation process were different depending on the types of sources. The parameter used for the constraint of *Ortho* was also experimentally chosen to get the best performance. Figure 1 shows the PESQ scores and SDRs for which the input signal-to-noise ratio (SNR) was 0 dB. For all of the four noises, the proposed algorithm outperformed other methods in terms of both the PESQ score and SDR. On average, the PESQ score improvements over the standard NMF and *Ortho* [6] were 0.25 and 0.19, respectively, and the SDR improvements over the competitors were 2.17 dB and 1.63 dB, respectively.



(a) PESQ score

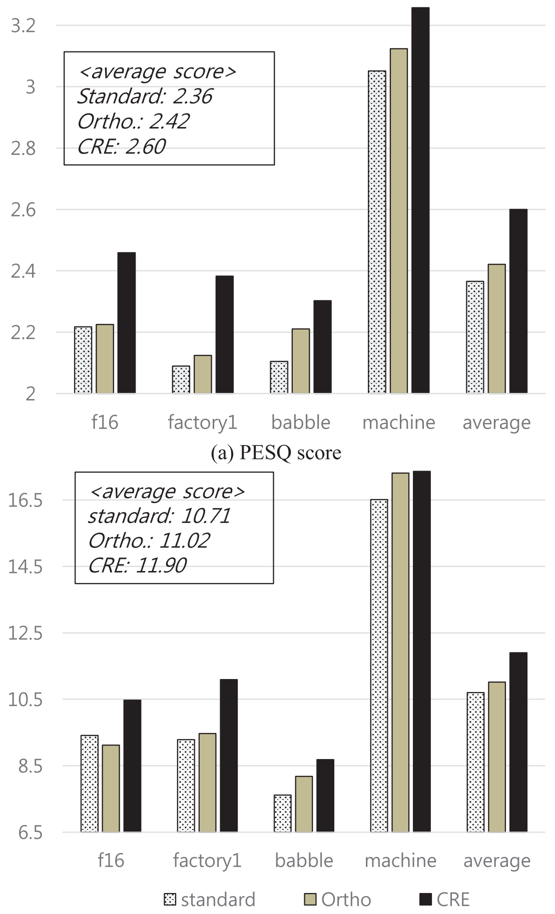
(b) SDR

Fig. 1 The PESQ scores and SDRs for various noises at 0 dB SNR.

The experimental results for the input SNR of 5 dB are illustrated in Fig. 2. The proposed algorithm outperformed other algorithms for all noise types at 5 dB SNR, too. The performance improvements were 0.24 and 0.18 in terms of the PESQ score and 1.19 dB and 0.88 dB in terms of SDR over the standard NMF and *Ortho*, respectively. These experimental results confirm that the proposed objective function incorporating cross-reconstruction error can enhance the performance of source separation not only in terms of an objective distortion measure but also in terms of subjective quality. This may lead us to the conclusion that the cross-reconstruction error term helps to reduce both the speech distortion and the residual noise.

5. Conclusions

This paper proposed a discriminative NMF algorithm incorporating using the cross-reconstruction error. The objective



(a) PESQ score

(b) SDR

Fig. 2 The PESQ scores and SDRs for various noises at 5 dB SNR.

function to train a basis matrix for a source is constructed to show high reconstruction error for the interfering source signals in addition to low self-reconstruction error, which may reduce the residual interference and the target source distortion. Experiments demonstrated that the proposed algorithm outperformed the standard NMF and the DNMF using orthogonality.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.2012R1A2A2A01045874), and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] P. Smaragdis, C. Fevotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorization: A unified view," IEEE Signal Process. Mag., vol.31, no.3, pp.66–75, 2014.

[2] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," Neural Comput., vol.13, no.4, pp.863–882, 2001.

[3] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," IEEE Trans. Audio, Speech, Language Process., vol.14, no.1, pp.191–199, 2006.

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

[5] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. Audio, Speech, Language Process., vol.18, no.3, pp.550–563, 2010.

[6] E.M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using crosscoherence penalties for single channel source separation," INTERSPEECH, pp.808–812, 2013.

[7] F. Weninger, J.L. Roux, J.R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," Proc. ISCA Interspeech, 2014.

[8] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3749–3753, 2014.

[9] K. Kwon, J.W. Shin, and N.S. Kim, "NMF-based speech enhancement using bases update," IEEE Signal Process. Lett., vol.22, no.4, pp.450–454, April 2015.

[10] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, pp.788–791, 1999.

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

[12] K.W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," INTERSPEECH, pp.411–414, 2008.

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

[14] P.D. O'Grady and B.A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," Proc. 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp.427–432, 2006.

[15] N. Guan, D. Tao, Z. Lwo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," IEEE Trans. Image Process., vol.20, no.7, pp.2030–2048, 2011.

[16] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation," ISMIR, pp.205–210, 2012.

[17] T.G. Kang, K. Kwon, J.W. Shin, and N.S. Kim, "NMF-based speech enhancement incorporating deep neural network," INTERSPEECH, pp.2843–2846, Sept. 2014.

[18] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol.2, pp.749–752, 2001.

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