Discriminative Nonnegative Matrix Factorization Using Cross-Reconstruction Error for Source Separation

Kisoo Kwon[1], *Jong Won Shin*[2], *Hyung Yong Kim*[1], *Nam Soo Kim*[1]

[1]Dept. of Electrical and Computer Engineering and the INMC,
Seoul National University, Seoul, Korea
[2]School of Information and Communications,
Gwangju Institute of Science and Technology, Gwangju, Korea.
kskwon@hi.snu.ac.kr, jwshin@gist.ac.kr, {hykim@hi.snu, nkim@snu}.ac.kr

Abstract

Non-negative matrix factorization (NMF) is a dimensionality reduction method that usually leads to a part-based representation, and it is shown to be effective for source separation. However, the performance of the source separation degrades when one signal can be described with the bases for the other source signals. In this paper, we propose a discriminative NMF (DNMF) algorithm which exploits the reconstruction error for the other signals as well as the target signal based on target bases. The objective function to train the basis matrix is constructed to reward high reconstruction error for the other source signals in addition to low reconstruction error for the signal from the corresponding source. Experiments showed that the proposed method outperformed the standard NMF by about 0.26 in perceptual evaluation of speech quality score and 1.95 dB in signal-to-distortion ratio when it is applied to speech enhancement at input SNR of 0 dB.

Index Terms: non-negative matrix factorization, discriminative basis, cross-reconstruction error

1. Introduction

Audio source separation is one of the main topics in the audio signal processing including music signal processing, speech enhancement and speech recognition [1]-[9]. Template-based approaches and data-representation methods have been widely applied to audio source separation, which make the statistics or representation models from the training database (DB) [1]-[9]. One of these notably successful techniques is based on non-negative matrix factorization (NMF) [10]. NMF is an unsupervised technique to discover part-based representations underlying non-negative data.

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

When applied to 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 reason is that the bases for each source are trained separately to reconstruct the corresponding signal faithfully without considering the source separation capability. To alleviate this difficulty, one can try to either modify the criterion of the NMF algorithm [6]-[8], [15], and [16] or estimate the weights for bases in a way to consider the effect of mixed sources [17].

The former approach was taken in many papers with the name of discriminative NMF (DNMF) [6]-[8], [15], and [16]. Although the detailed methods are different from each other, these works aim to make the basis vectors of a target source reconstruct only the target source by utilizing the other source DBs or DBs mixed with a target source DB. 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 other source. However, the above orthogonality constraint may result in a high reconstruction error because each basis is apt to represent a narrow frequency band. In [7], the basis vectors of each source are updated by the reconstruction error of the source, 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 discriminative NMF for which the objective function for NMF training includes a term rewarding high reconstruction error for the other source signals in addition to a term for low reconstruction error for the target signal. The proposed DNMF algorithm with cross-reconstruction error was applied to speech enhancement and showed improved performance in terms of the perceptual evaluation of speech quality (PESQ) [18] and the signal-to-distortion ratio (SDR) [19].

2. NMF-based audio source separation

When applied to audio source separation, NMF approximates the magnitude or power spectrogram of mixture $V \in \mathbb{R}^{M \times N}$ as the product of a basis matrix $W \in \mathbb{R}^{M \times r}$, and an encoding matrix $H \in \mathbb{R}^{r \times N}$ ($V \approx WH$) where M, N, and r denote the number of frequency bins, short-time frames, and basis vectors, respectively. W_S and W_N are usually trained separately with clean signal 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, respectively, the update rules for the encoding

and basis matrices during the training phase are given as [10]:

$$H_i \leftarrow H_i \otimes \frac{W_i^T \frac{V_i}{W_i H_i}}{W_i^T \mathbf{1}}, \quad (1)$$

$$W_i \leftarrow W_i \otimes \frac{\frac{V_i}{W_i H_i} H_i^T}{\mathbf{1} H_i^T}. \quad (2)$$

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

In the test phase, the noisy magnitude spectrum $|Y(t)|$ is approximated as $|Y(t)| \approx WH(t)$ for each frame with the fixed basis matrix $W = [W_S\ W_N]$ 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 noisy input, and $|\cdot|$ denotes element-wise absolute value. Keeping 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 non-negative random numbers. After the iteration, the magnitude spectra of the target and noise signals are reconstructed as:

$$|\hat{S}(t)| = W_S H_S(t), \qquad |\hat{N}(t)| = W_N H_N(t). \quad (3)$$

Instead of directly using the reconstructed magnitude spectra in (3), a spectral gain function similar to the Wiener filter is adopted in [12] and [9] based on the estimated magnitude spectra in (3). The gain function is given as:

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

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

3. Discriminative NMF using cross-reconstruction error

When the bases for each source are trained separately, the subspaces that the bases for individual sources span are not guaranteed to be disjoint. It implies that some data vectors from one source can be reconstructed by one or more bases for other source along with the bases for the source. Consequently, it can degrade the performance of 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, final source separation performance can be enhanced especially when there are multiple basis matrices which produce similar reconstruction errors but different source separation performances.

In order to obtain discriminative bases, we propose an objective function to estimate the basis matrix that utilizes the reconstruction error of the other source signals based on the bases along with conventional reconstruction error of the original signal. In this paper, we assume that input signal consists of two kinds of sources, the target and interfering sources. The reconstruction error of the interfering source signals based on the target bases may be considered as a measure of the performance degradation caused by the residual interferences incurred by misuse of the target bases for the interference signal. On the other hand, if the interfering signal basis matrix is trained using the cross-reconstruction error on top of the conventional reconstruction error, it can reduce the distortion of the target source caused by the usage of interference bases for the target source signal. The objective function of the proposed method to train the basis matrix W_i where i indicates either the target or the interference is defined by

$$\begin{aligned} f(W_i, H_i, C_j) = & \\ & D(V_i \parallel W_i H_i) - \gamma_i D(V_j \parallel W_i C_j) \\ & \gamma_i = \lambda_i \frac{\|V_i\|_1}{\|V_j\|_1} \end{aligned} \quad (5)$$

where $V_i \in \mathbb{R}^{M \times N_i}$, W_i, and H_i are the magnitude spectrogram of N_i frames, basis, and encoding matrix for the source signal that we want to train a basis matrix, and $V_j \in \mathbb{R}^{M \times N_j}$ and C_j are the magnitude spectrogram of N_j frames and encoding matrix for the other source signal. $D(a \parallel b)$ is the distance function between a and b, for which the KL-divergence is chosen, and $\|\cdot\|_1$ is an l_1-norm of the vector constructed by concatenating the rows of the matrix. γ_i is introduced to control the weight of the cross-reconstruction error by λ_i while the length of the training data are normalized. $\lambda_S = \lambda_N = 0$ corresponds to the standard NMF while $\lambda_S > 0$ and $\lambda_N > 0$ may enhance the source separation performance.

The update equations of W_i, H_i and C_j are obtained in a similar way to what is shown in Sec. 2. as follows:

$$H_i \leftarrow H_i \otimes \frac{W_i^T \frac{V_i}{W_i H_i}}{W_i^T \mathbf{1}}, \quad C_j \leftarrow C_j \otimes \frac{W_i^T \frac{V_j}{W_i C_j}}{W_i^T \mathbf{1}}, \quad (6)$$

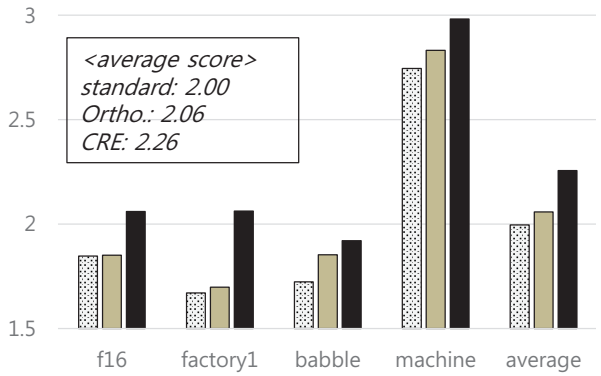
$$W_i \leftarrow W_i \otimes \frac{\frac{V_i}{W_i H_i} H_i^T}{\mathbf{1} H_i^T + \gamma_i (\frac{V_j}{W_i H_j} H_j^T - \mathbf{1} C_j^T)} \quad (7)$$

It is noted that both the target and interference DB are needed to train each of the basis matrix. Each basis matrix is trained by iteratively applying (6) and (7) for a fixed number of times with random initialization.

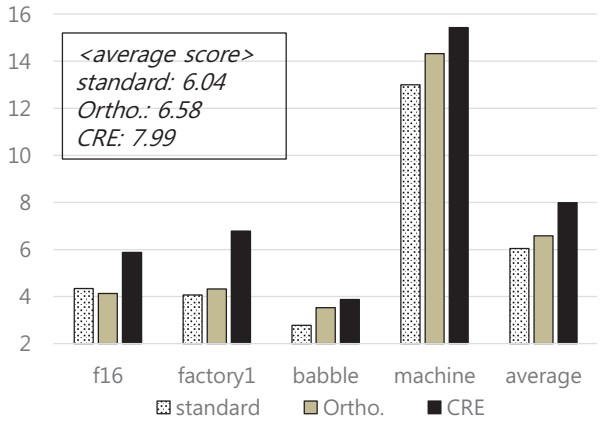
In the source separation phase, the target and noise basis matrices from the proposed method are used, and STFT coefficients of the target source are finally estimated in the same way as in Sec.2.

4. Experiment

To evaluate the performance of the proposed algorithm, it was applied to the task of speech enhancement. 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 signal which is not included in the test data, and the speech DB for the training was 130-second long spoken by 56 different speakers. The speech test data set consisted of 32 sentences from 32 different speakers. We have tested 4 different types of noises including *F-16*, *factory1*, *babble* and *machinegun* noise. 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.

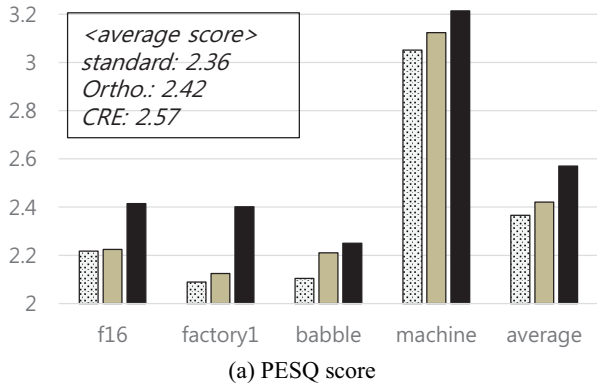


(a) PESQ score

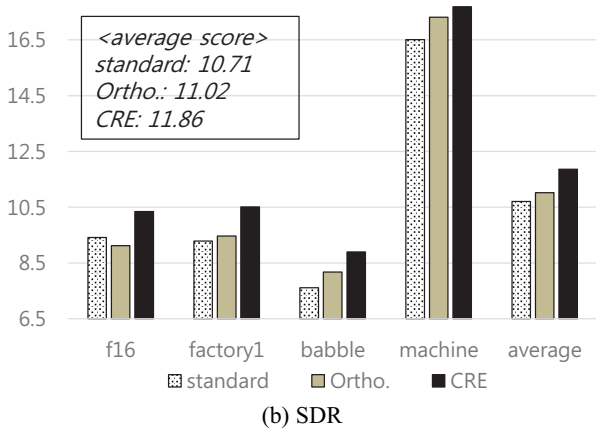


(b) SDR

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



(a) PESQ score



(b) SDR

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

The performance of the proposed method was evaluated using the ITU-T Recommendation P.862 perceptual evaluation of speech quality (PESQ) [18] and the signal-to-distortion ratio (SDR) [19]. To demonstrate the performance improvement achieved by the proposed objective function, three speech enhancement systems of 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 was performed without any additional penalty term [10]
- *Ortho.*: the DNMF in [6] which tries to make basis matrices for different sources orthogonal.
- *CRE*: the proposed method using the cross-reconstruction error.

The best separation performance of the proposed algorithm was obtained when $0.1 \leq \lambda_i \leq 0.5$, and λ_i is chosen from $\{0.1, 0.3, 0.35, 0.4, 0.45, 0.5\}$. The optimal values of λ_i were different according to the types of sources. The parameter used for the constraint of *Orhtho.* was chosen by experiments to get the best performance. Fig.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 the SDR. On average, the PESQ score improvements over the standard NMF and [6] were 0.26 and 0.20, respectively, and the SDR improvements over the competitors were 1.95 dB and 1.41 dB, respectively.

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 for 5 dB SNR were 0.21 and 0.15 in terms of the PESQ score and 1.15 dB and 0.84 dB in terms of the SDR over the standard NMF and [6], respectively. Experimental results may imply that the proposed objective function with cross-reconstruction error can enhance the performance of source separation not only in terms of an objective quality measure but also in terms of an objective measure of subjective quality. It may be because the cross-reconstruction error term helps to reduce both the speech distortion and the residual noise.

5. Conclusions

This paper proposed a discriminative NMF using the cross-reconstruction error. The objective function to train a basis matrix for a source is constructed to reward high reconstruction error for the other source signals in addition to low reconstruction error for the given source, which may reduce the residual interference and the target source distortion. Experimental results demonstrated that the proposed algorithm outperformed the standard NMF and the DNMF using orthogonality.

6. Acknowledgements

This research was supported in part by the A3 Foresight Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology 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).

7. References

[1] P. Smaragdis, C. Fvotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using npnnegative factorization," IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 66-75, 2014.

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

[3] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *Audio, Speech, and Language Processing, IEEE Transactions on*, 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066-1074, 2007.

[5] A. Ozerov and C. Fvotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, 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. of ISCA Interspeech*, 2014.

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

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

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

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

[12] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative 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," *Machine Learning for Signal Processing 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on. IEEE*, pp. 427-432, 2006.

[15] N. Guan, D. Tao, Z. Lwo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *Image Processing, IEEE Transactions on*, 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, Sep. 2014.

[18] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.

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