INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Discriminative Layered Nonnegative Matrix Factorization for Speech Separation

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

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan
chien.cm97g@nctu.edu.tw, tschi@mail.nctu.edu.tw, jtchien@nctu.edu.tw

Abstract

This paper proposes a discriminative layered nonnegative matrix factorization (DL-NMF) for monaural speech separation. The standard NMF conducts the parts-based representation using a single-layer of bases which was recently upgraded to the layered NMF (L-NMF) where a tree of bases was estimated for multi-level or multi-aspect decomposition of a complex mixed signal. In this study, we develop the DL-NMF by extending the generative bases in L-NMF to the discriminative bases which are estimated according to a discriminative criterion. The discriminative criterion is conducted by optimizing the recovery of the mixed spectra from the separated spectra and minimizing the reconstruction errors between separated spectra and original source spectra. The experiments on single-channel speech separation show the superiority of DL-NMF to NMF and L-NMF in terms of the SDR, SIR and SAR measures.
Index Terms: dictionary learning, discriminative learning, nonnegative matrix factorization, speech separation

1. Introduction

Nonnegative matrix factorization (NMF) is known a popular approach to data representation which decomposes a given nonnegative matrix $\mathbf{X}$ into a nonnegative dictionary (or basis) matrix $\mathbf{W}$ and a nonnegative weight matrix $\mathbf{H}$ via $\mathbf{X} \approx \mathbf{WH}$. Inspired from the visual perception, NMF attempts to learn the hidden representation of the part information of the objects [1]. It has been successfully developed in many research fields, such as image processing [2], blind source separation [3], document clustering [4], and computational biology [5].

Over the past years, various extensions of NMF have been proposed. For instance, the sparse NMF was proposed to learn the sparse representation of data for solving the overcomplete problem [6]. The graph regularized NMF was proposed by taking the intrinsic geometric structure of data into consideration [7]. For audio signals, NMF can be directly applied for the Fourier magnitude spectrogram or its variants (e.g., the mel-spectrogram). For supervised speech separation, the convolutive NMF (CNMF) [8] was proposed to discover the phoneme-like bases by considering the temporal dependencies of the magnitude spectrogram across several consecutive frames. The two-dimensional CNMF was also proposed to identify musical notes by further decoding the spectral dependencies (e.g., the harmonic structure) of the magnitude spectrogram for blind music separation [9].

However, the standard NMF model and the extensions mentioned above do not generate hierarchical features. In recent years, various deep NMF algorithms were proposed by incorporating the hierarchical architecture into the standard NMF. For instance, the multi-layer NMF was proposed as a sequential factorization for different NMF variants [10]. However, its reconstruction error increases with the increasing number of layers due to the lack of error correction procedure through the layers. In [11], the deep semi-NMF with error correction was proposed to learn a hierarchical representation of features from an image dataset for attribute-based clustering. But, this method did not impose the nonnegative constraint so that the additive combination of bases was not possible for speech separation applications. Extended from [12], the iterative inference procedure of the standard NMF was unfolded to mimic the architecture of a deep neural network. The architecture can be thought of as performing a joint optimization on the generative-discriminative hybrid model [13]. However, the underlying structure is still single-layer. In our previous work [14], a layered nonnegative matrix factorization (L-NMF) algorithm was proposed to learn the hierarchical bases for speech separation. Deep neural networks (DNNs) have emerged as a powerful machine learning approach and produced state-of-the-art results in many research fields such as speech recognition [15], speech enhancement [16] and source separation [17]. DNNs, as deep discriminative models, hold the promise of the performance if given a large amount of labelled data. But sometimes it is not trivial to get a large amount of labelled data.

The NMF and most of its variants all tend to learn hidden bases, which link explicitly or implicitly the original data to the corresponding representations, of a given dataset. These models behave like generative models. For supervised speech separation problem, NMF-based methods are first applied to learn the nonnegative bases for each individual speaker. After learning, the speaker-dependent bases are used to separate the mixed spectrogram. However, the bases of each speaker are learned without considering the *interfering effect* from the other speakers during training. Therefore, a mismatch exists between training and test conditions such that the separation result is not truly optimized. To address this problem, the discriminative NMF (D-NMF) methods were proposed to directly optimize the separation objective, having an accurate reconstruction of the mixture using the separated spectra while keeping the separated spectra as close to the original spectra as possible [12][18]. In the same spirit of NMF to D-NMF, we propose a discriminative L-NMF (DL-NMF) algorithm for single-channel speech separation. After learning the hierarchical bases for each speaker independently by L-NMF, we introduce a discriminative criterion on a small subset of data, which adapt the bases by enforcing an optimal recovery of the mixed spectra from separated spectra and minimizing the total distance between the separated spectra and the original spectra so as to further improve the separation performance.

http://dx.doi.org/10.21437/Interspeech.2016-415

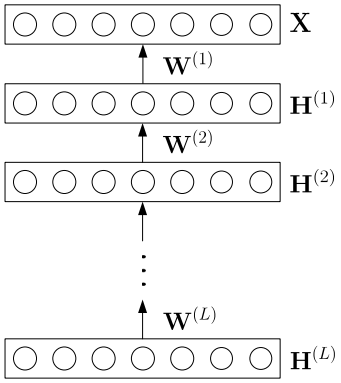

Figure 1: *Layered NMF model.*

The rest of the paper is organized as follows. Section II gives a brief review of the standard NMF and our previous work, L-NMF. In section III, we propose the DL-NMF and show its update equations. Section IV demonstrates the experimental result. We end in section V with some conclusions and future work.

2. NMF and Layered NMF

2.1. NMF

Given a nonnegative data matrix $\mathbf{X} \in \mathcal{R}_+^{M\times N}$, NMF aims to decompose this data matrix into a product of two nonnegative matrices $\mathbf{W} \in \mathcal{R}_+^{M\times K}$ and $\mathbf{H} \in \mathcal{R}_+^{K\times N}$ with their entries related as

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

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 the Euclidean distance, Kullback-Leibler divergence, Itakura-Saito divergence, and so on. This model could be solved by performing the alternating minimization. Multiplicative update rules are simple and efficient in inferring the model parameters $\{\mathbf{W}, \mathbf{H}\}$ as follows

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{[\nabla_{\mathbf{W}}\mathcal{D}]^-}{[\nabla_{\mathbf{W}}\mathcal{D}]^+} \tag{3}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{[\nabla_{\mathbf{H}}\mathcal{D}]^-}{[\nabla_{\mathbf{H}}\mathcal{D}]^+} \tag{4}$$

where $\otimes$ denotes an element-wise multiplication and 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

In [14], the layered NMF with L layers is performed by

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

This approximation is performed based on a hierarchical architecture with L layers as shown in Fig. 1. In training procedure, the 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 $\mathbf{W}^{(1)} \in \mathcal{R}_+^{M\times K_1}$ and $\mathbf{H}^{(1)} \in \mathcal{R}_+^{K_1\times N}$. Then the same factorization is performed on the result 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 procedure to pre-train all layers until $\mathbf{H}^{(L-1)} \approx \mathbf{W}^{(L)}\mathbf{H}^{(L)}$, where $\mathbf{W}^{(L)} \in \mathcal{R}_+^{K_{L-1}\times K_L}$ and $\mathbf{H}^{(L)} \in \mathcal{R}_+^{K_L\times N}$. After the initialization, we fine-tune the parameters of all layers, $\{\mathbf{W}^{(l)}\}$ and $\mathbf{H}^{(L)}$, to reduce the total reconstruction error via

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

The multiplicative update rules for all layers can be derived as

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

$$\mathbf{H}^{(L)} \leftarrow \mathbf{H}^{(L)} \otimes \frac{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^-}{[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^+} \tag{8}$$

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 $\mathbf{W}^{(l)}$ and $[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^+$ and $[\nabla_{\mathbf{H}^{(L)}}\mathcal{D}]^-$ denote the positive and negative parts of the gradient with respect to $\mathbf{H}^{(L)}$, respectively. As indicated in [14], comparing with the standard NMF, the L-NMF can realize more complex bases via the hierarchical structure by combining sparse parts-based bases extracted by the single layer NMF to interpret the data differently, hence to improve separation performance.

3. Discriminative Layered NMF

Both NMF and L-NMF are seen as the generative models, and each of which provides a way to represent the given data. However, a good general representation does not guarantee the satisfied performance in a specific application. Therefore, we incorporate a discriminative cost function into the L-NMF. Here, we consider the task of supervised speech separation between two speakers. In this task, first, the hierarchical bases of each individual speaker ($\mathbf{W}_{s_1}^{(1)}, \mathbf{W}_{s_1}^{(2)}, \cdots, \mathbf{W}_{s_1}^{(L)}$ and $\mathbf{W}_{s_2}^{(1)}, \mathbf{W}_{s_2}^{(2)}, \cdots, \mathbf{W}_{s_2}^{(L)}$) are separately learned from his/her clean sentences using the L-NMF algorithm. Then, using the input mixed spectrograms $\mathbf{X}_{mix}$, the parameters $\mathbf{H}_{s_1}^{(L)}$ and $\mathbf{H}_{s_2}^{(L)}$ are obtained by minimizing the following cost function with the fixed basis parameters ($\mathbf{W}_{s_1}^{(1)}, \mathbf{W}_{s_1}^{(2)}, \cdots, \mathbf{W}_{s_1}^{(L)}$ and $\mathbf{W}_{s_2}^{(1)}, \mathbf{W}_{s_2}^{(2)}, \cdots, \mathbf{W}_{s_2}^{(L)}$) as

$$\min_{\{\mathbf{H}_{s_1}^{(L)},\mathbf{H}_{s_2}^{(L)}\}\geq 0} \mathcal{D}\left(\mathbf{X}_{mix}\middle\|(\mathbf{I},\mathbf{I})\begin{pmatrix}\mathbf{W}_{s_1}^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(1)}\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\mathbf{W}_{s_1}^{(2)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(2)}\end{pmatrix}\cdots\begin{pmatrix}\mathbf{W}_{s_1}^{(L)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(L)}\end{pmatrix}\begin{pmatrix}\mathbf{H}_{s_1}^{(L)}\\ \mathbf{H}_{s_2}^{(L)}\end{pmatrix}\right) \tag{9}$$

where $\mathbf{I}$ and $\mathbf{0}$ are identity and zero matrices with proper sizes, respectively. For ease of expression, we can rewrite this cost

function by using the compound matrices for each individual matrices in Eq. (9) in a form of

$$\min_{\mathbb{H}^{(L)}\geq 0} \mathcal{D}\left(\mathbf{X}_{mix}\middle\|\mathbb{I}\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\mathbb{H}^{(L)}\right). \quad (10)$$

In this paper, the Kullback-Leibler divergence is selected as the cost function in all steps. Therefore, similar to Eq. (8) with additional consideration of matrix $\mathbb{I}$, the update equation of $\mathbb{H}^{(L)}$ can be obtained as

$$\mathbb{H}^{(L)} \leftarrow \mathbb{H}^{(L)} \otimes \frac{\left(\mathbb{I}\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\right)^{\mathrm{T}}\left(\frac{\mathbf{X}_{mix}}{\hat{\mathbf{X}}_{mix}}\right)}{\left(\mathbb{I}\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\right)^{\mathrm{T}}\mathbf{1}} \quad (11)$$

where $\hat{\mathbf{X}}_{mix}$ is the reconstructed mixed spectra and $\mathbf{1}$ is a matrix of the proper size with all elements equal to one. This step is actually the same as the separation stage in L-NMF [14]. However, we introduce a discriminative cost function to further adapt the learned hierarchical bases and make them more discriminative between speakers. With the fixed weight matrices ($\mathbf{H}_{s_1}^{(L)}$ and $\mathbf{H}_{s_2}^{(L)}$, which are learned from Eq. (11)), the discriminative criterion modifies the hierarchical basis matrices of two speakers ($\mathbf{W}_{s_1}^{(1)}, \mathbf{W}_{s_1}^{(2)}, \cdots, \mathbf{W}_{s_1}^{(L)}$ and $\mathbf{W}_{s_2}^{(1)}, \mathbf{W}_{s_2}^{(2)}, \cdots, \mathbf{W}_{s_2}^{(L)}$) by minimizing the total errors between the reconstructed spectra and the original spectra by

$$\min_{\{\mathbf{W}_{s_1}^{(l)},\mathbf{W}_{s_2}^{(l)}\}\geq 0} \mathcal{D}\left(\begin{pmatrix}\mathbf{S}_1\\ \mathbf{S}_2\end{pmatrix}\middle\|\begin{pmatrix}\mathbf{W}_{s_1}^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(1)}\end{pmatrix}\begin{pmatrix}\mathbf{W}_{s_1}^{(2)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(2)}\end{pmatrix}\cdots\begin{pmatrix}\mathbf{W}_{s_1}^{(L)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{s_2}^{(L)}\end{pmatrix}\begin{pmatrix}\mathbf{H}_{s_1}^{(L)}\\ \mathbf{H}_{s_2}^{(L)}\end{pmatrix}\right), \quad \forall\, l=1,\cdots,L \quad (12)$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are the original spectra, i.e., the target of the separated spectra of two speakers. Again, we rewrite this cost function by using compound matrices as

$$\min_{\{\mathbb{W}^{(l)}\}\geq 0} \mathcal{D}\left(\mathbb{S}\middle\|\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\mathbb{H}^{(L)}\right),\ \forall\, l=1,\cdots,L. \quad (13)$$

Similar to Eq. (7), the update equations for the bases of two speakers in different layers can be obtained by

$$\mathbb{W}^{(1)} \leftarrow \mathbb{W}^{(1)} \otimes \frac{\left(\frac{\mathbb{S}}{\hat{\mathbb{S}}}\right)\left(\left(\prod_{m=2}^{L}\mathbb{W}^{(m)}\right)\mathbb{H}^{(L)}\right)^{\mathrm{T}}}{\mathbf{1}\left(\left(\prod_{m=2}^{L}\mathbb{W}^{(m)}\right)\mathbb{H}^{(L)}\right)^{\mathrm{T}}} \quad (14)$$

$$\mathbb{W}^{(l)} \leftarrow \mathbb{W}^{(l)} \otimes \frac{\left(\prod_{m=1}^{l-1}\mathbb{W}^{(m)}\right)^{\mathrm{T}}\left(\frac{\mathbb{S}}{\hat{\mathbb{S}}}\right)\left(\left(\prod_{m=l+1}^{L}\mathbb{W}^{(m)}\right)\mathbb{H}^{(L)}\right)^{\mathrm{T}}}{\left(\prod_{m=1}^{l-1}\mathbb{W}^{(m)}\right)^{\mathrm{T}}\mathbf{1}\left(\left(\prod_{m=l+1}^{L}\mathbb{W}^{(m)}\right)\mathbb{H}^{(L)}\right)^{\mathrm{T}}}, \quad \forall\, l=2,\cdots,L-1 \quad (15)$$

$$\mathbb{W}^{(L)} \leftarrow \mathbb{W}^{(L)} \otimes \frac{\left(\prod_{m=1}^{L-1}\mathbb{W}^{(m)}\right)^{\mathrm{T}}\left(\frac{\mathbb{S}}{\hat{\mathbb{S}}}\right)\left(\mathbb{H}^{(L)}\right)^{\mathrm{T}}}{\left(\prod_{m=1}^{L-1}\mathbb{W}^{(m)}\right)^{\mathrm{T}}\mathbf{1}\left(\mathbb{H}^{(L)}\right)^{\mathrm{T}}} \quad (16)$$

where $\mathbb{S}=[\mathbf{S}_1;\mathbf{S}_2]$ and $\hat{\mathbb{S}}=[\hat{\mathbf{S}}_1;\hat{\mathbf{S}}_2]$ represent the original and the reconstructed spectra, respectively. In summary, the model

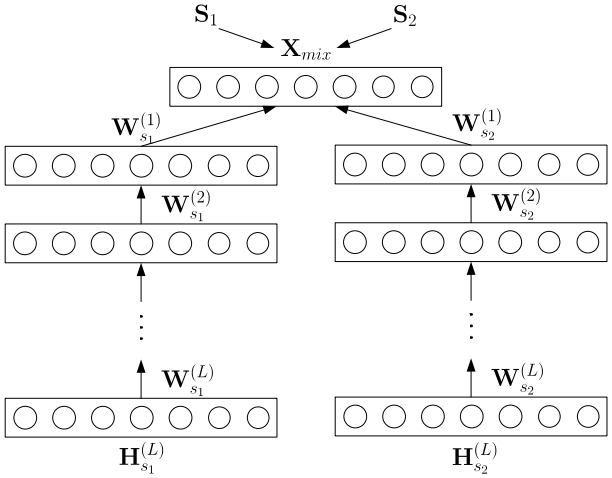

Figure 2: *Discriminative Layered NMF model.*

structure of DL-NMF is shown in Fig. 2 and the bases of DL-NMF are optimally trained by

$$\{\hat{\mathbb{W}}^{(l)}\} = \underset{\{\mathbb{W}^{(l)}\}\geq 0}{\operatorname{argmin}} \mathcal{D}\left(\mathbb{S}\middle\|\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\hat{\mathbb{H}}^{(L)}\right), \quad \forall\, l=1,\cdots,L \quad (17)$$

where

$$\hat{\mathbb{H}}^{(L)} = \underset{\mathbb{H}^{(L)}\geq 0}{\operatorname{argmin}} \mathcal{D}\left(\mathbf{X}_{mix}\middle\|\mathbb{I}\left(\prod_{l=1}^{L}\mathbb{W}^{(l)}\right)\mathbb{H}^{(L)}\right). \quad (18)$$

In practice, the procedure of updating $\hat{\mathbb{H}}^{(L)}$ and $\{\hat{\mathbb{W}}^{(l)}\}$ would be repeated several times.

4. Experiments and Results

In this paper, we followed the experimental procedure of the L-NMF in [14] and demonstrate the benefit of the discriminative learning in a supervised speaker-dependent speech separation task. The speech mixtures were generated by combining sentences from one female (FA) and one male (MC) speakers from TSP corpus [19] to evaluation the proposed DL-NMF algorithm. There are 60 sentences spoken by each of the speakers. The TSP corpus contains over 1400 sentences spoken by 25 speaker. The 1024-point short-term Fourier transform (STFT) with a 64-ms frame length and a 16-ms frame shift was calculated to obtain the Fourier magnitude spectrogram. First, the hierarchical bases of each speaker were learned from clean training data using L-NMF. Then, the discriminative criterion was used to adapt the layered basis matrices. After the adaptation, for the spectrogram of a new test mixture, the weight matrices $\hat{\mathbb{H}}^{(L)} = [\hat{\mathbf{H}}_{s_1}^{(L)};\hat{\mathbf{H}}_{s_2}^{(L)}]$ were obtained using Eq. (11) and the separated spectrograms were then calculated by applying the Wiener gain as follows

$$\hat{\mathbf{X}}_{s_1} = \mathbf{X}_{mix} \otimes \frac{\left(\prod_{l=1}^{L}\hat{\mathbf{W}}_{s_1}^{(l)}\right)\hat{\mathbf{H}}_{s_1}^{(L)}}{\left(\prod_{l=1}^{L}\hat{\mathbb{W}}^{(l)}\right)\hat{\mathbb{H}}^{(L)}} \quad (19)$$

$$\hat{\mathbf{X}}_{s_2} = \mathbf{X}_{mix} \otimes \frac{\left(\prod_{l=1}^{L}\hat{\mathbf{W}}_{s_2}^{(l)}\right)\hat{\mathbf{H}}_{s_2}^{(L)}}{\left(\prod_{l=1}^{L}\hat{\mathbb{W}}^{(l)}\right)\hat{\mathbb{H}}^{(L)}}. \quad (20)$$

Table 1: *Separation performance of NMF, L-NMF, and DL-NMF in terms of SDR, SIR and SAR measures (dB).*

Methods	Numbers of bases	Female (FA)			Male (MC)		
		SDR	SIR	SAR	SDR	SIR	SAR
NMF	K=10	6.76	10.34	9.94	6.28	8.92	10.51
	K=20	6.83	10.05	10.27	6.34	8.82	10.81
	K=30	6.71	9.88	10.15	6.10	8.30	10.94
L-NMF	K_1=180, K_2=10	7.26	11.09	10.14	6.75	9.46	10.78
	K_1=180, K_2=20	7.38	10.94	10.43	6.70	9.16	11.21
	K_1=180, K_2=30	6.88	10.17	10.19	6.26	8.48	11.04
DL-NMF	K_1=180, K_2=10	7.42	11.11	10.36	6.87	9.36	11.19
	K_1=180, K_2=20	**7.87**	**11.33**	**11.04**	**7.48**	**10.30**	**11.38**
	K_1=180, K_2=30	7.65	10.92	10.99	7.37	10.05	11.33

In the implementation, we randomly selected 70% of each speaker's sentences (42 sentences) for training his/her hierarchical bases. Then, we used another 20% (12 mixtures and their corresponding clean utterances) for the discriminative learning of DL-NMF. The remaining 10% (6 test mixtures) were used for testing. We repeated the experiments 15 times using different random selections. Totally, 90 test mixtures were created from the FA-MC speaker pair. For both L-NMF and DL-NMF, the layer-related parameters were set as $L = 2$, $K_1 = 180$ and $K_2 = [10, 20, 30]$. The performance of the proposed algorithms was assessed in terms of the source-to-distortion ratio (SDR), the source-to-interference ratio (SIR) and source-to-artifacts ratio (SAR) [20]. The separation results of the standard NMF, the L-NMF and the proposed DL-NMF are shown in Table 1. The result in Table 1 shows the average SDR, SIR and SAR over 90 test mixtures. Clearly, L-NMF outperforms the standard NMF in both SDR and SIR measures while the proposed DL-NMF improves the SDR, SIR and SAR further by incorporating the discriminative cost function.

5. Conclusions

In our previous work, we demonstrated that L-NMF algorithm can realize more complex bases by combining sparse parts-based bases extracted by the single-layer standard NMF and interpret data differently. We also showed that L-NMF outperforms the standard NMF in terms of the SDR measure in speech separation experiments. However, the NMF and L-NMF are both generative models. These generative methods do not directly optimize the performance for specific applications. In contrast, the discriminative learning is able to boost performance when dealing with the supervised learning problem. Accordingly, we proposed the DL-NMF, based on the L-NMF method, by incorporating the discriminative dictionary learning technique to learn *discriminative* hierarchical bases for the application of speech separation. Simulation results show that the proposed DL-NMF outperforms the standard NMF and L-NMF in terms of SDR, SIR and SAR measures in the speech separation tasks.

In this study, we set the numbers of layer in DL-NMF and L-NMF as $L = 2$ because the variety in one-frame spectrum is not too much in speech samples from TSP corpus such that the performance is not significantly different by increasing L. One way to increase the variety of the training spectrum is to consider temporal variations by extracting multiple-frame spectra. In this case, the observation at time t contains several consecutive one-frame spectra, which is unfolded into a column vector [12]. In other words, we can apply our method to analyze spectral patches of the spectrograms for speech separation. Another future work is to incorporate sparsity constraints into the proposed method. The sparsity constraints have been shown to help the SIR score but decrease the SAR score [21]. Therefore, the degree of sparseness should be carefully designed based on the requirements of the application. Furthermore, all size parameters of the hierarchical architecture of the proposed DL-NMF algorithm are pre-determined in advance currently. In the future, we will extend our method to a Bayesian approach [22], which can regularize the model and automatically select model parameters by given data. In addition, exploring benefits from using a nonlinear function between layers would be also pursued.

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] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2030–2048, 2011.

[3] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1171–1185, 2014.

[4] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. of Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, pp. 267–273.

[5] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLOS Computational Biology*, vol. 4, no. 7, p. e1000029, 2008.

[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] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation." in *Independent Component Analysis and Blind Signal Separation*, 2006, pp. 700–707.

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

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

[12] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. of Annual Conference of International Speech Communication Association*, 2014, pp. 865–869.

[13] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 66 – 70.

[14] C.-C. Hsu, J.-T. Chien, and T.-S. Chi, "Layered nonnegative matrix factorization for speech separation," in *Proc. of Annual Conference of International Speech Communication Association*, 2015, pp. 628–632.

[15] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. 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.

[16] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[17] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 3734–3738.

[18] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3749–3753.

[19] P. Kabal, "TSP speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[20] 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.

[21] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 141–145.

[22] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation." in *Proc. of International Conference on Music Information Retrieval*, 2014, pp. 507–512.