TRANSDUCTIVE NONNEGATIVE MATRIX FACTORIZATION FOR SEMI-SUPERVISED HIGH-PERFORMANCE SPEECH SEPARATION

Naiyang Guan[1], *Long Lan*[1], *Dacheng Tao*[2], *Zhigang Luo*[1], *and Xuejun Yang*[3]

[1]Science and Technology on Parallel and Distributed Processing Laboratory, School of Computer Science, National University of Defense Technology, Changsha, Hunan, 410073 China.
[2]Centre for Quantum Computation and Intelligent Systems, FEIT, University of Technology, Sydney, Sydney, NSW 2007, Australia.
[3]State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, Hunan, 410073 China.

ABSTRACT

Regarding the non-negativity property of the magnitude spectrogram of speech signals, nonnegative matrix factorization (NMF) has obtained promising performance for speech separation by independently learning a dictionary on the speech signals of each known speaker. However, traditional NMF fails to represent the mixture signals accurately because the dictionaries for speakers are learned in the absence of mixture signals. In this paper, we propose a new transductive NMF algorithm (TNMF) to jointly learn a dictionary on both speech signals of each speaker and the mixture signals to be separated. Since TNMF learns a more descriptive dictionary by encoding the mixture signals than that learned by NMF, it significantly boosts the separation performance. Experiments results on a popular TIMIT dataset show that the proposed TNMF-based methods outperform traditional NMF-based methods for separating the monophonic mixtures of speech signals of known speakers.

Index Terms— Nonnegative matrix factorization, transductive learning, speech separation

1. INTRODUCTION

Speech separation aims at recovering sounds of known speakers from monophonic mixture signals when some speech signals of these speakers are provided for training. It has attracted extensive attention and has been widely used in several speech analysis tasks such as hearing aids [1], speaker recognition [2], and telephonic communications [3]. Existing methods first extract the magnitude spectrograms of both training speech signals and mixture signals, and then independently learn several spectral bases on the training speech signals of each speaker for separation. The learned spectral bases are concatenated to form a dictionary on which the mixture signals are represented, and the sounds of each speaker are recovered by collecting the components belonging to the corresponding spectral bases.

Considering the non-negativity property of the magnitude spectrogram of speech signals, nonnegative matrix factorization (NMF) shows great effectiveness to learn the spectral bases [4]. Recently, based on the NMF framework[16][6][17], many approaches [7][8][9] have been proposed for speech separation. Schmidt *et al.* [7] proposed nonnegative sparse coding (NNSC) to pre-compute spectral bases from wind noise and regard them as a part of the entire dictionary in decomposing the speech signals interfered by wind noise. To more accurately model the distribution of noise in speech signals, Fevotte *et al.* [10] proposed Itakura-Saito divergence based NMF (IS-NMF) for speech separation. To consider the dependencies across successive signals in speech, Smaragdis [8] proposed a convolutive NMF for speech separation that learns spectral bases on time-domain speech signals. Although the aforementioned approaches perform well in their tasks, the spectral bases are learned without considering the mixture signals to be separated, and thus the dictionary concatenated by the learned bases inaccurately recovers the sound of each speaker for the mixture signals.

Transductive learning methods [11] exploit available test examples to enhance the descriptive power of the learned model. In this regard, we propose a transductive NMF algorithm (TNMF) for semi-supervised high-performance speech separation. By contrast to conventional supervised speech separation, TNMF jointly learns a dictionary on both training speech signals from different speakers and the mixture signals to be separated. In particular, TNMF has two objectives: 1) TNMF independently minimizes the distance between the magnitude spectrogram of the training speech signal and the product of the corresponding spectral base and activation for each speaker, and 2) TNMF minimizes the distance between the magnitude spectrogram of the mixture signals and the

This work has been supported by the Scientific Research Plan Project of NUDT (No. JC13-06-01) and ARC DP-140102164.

product of a dictionary concatenated by these spectral bases and an activation matrix. TNMF combines the two objectives and utilizes a multiplicative update rule (MUR) to learn both the dictionary and the corresponding activations. Intuitively, since the dictionary learned by TNMF contains the phonemic features from both training speech signals and mixture signals, it can more accurately recover the speech of each speaker, and thus boost the separation performance. Experimental results on the popular TIMIT dataset show that TNMF outperforms the traditional NMF based methods.

The paper is organized as follows. Section 2 describes related works on NMF based methods for supervised speech separation, and Section 3 introduces the proposed TNMF algorithm. We present the experimental results in Section 4 and conclude the paper in Section 5.

2. RELATED WORKS

NMF [12] decomposes a given nonnegative matrix $V \in R_+^{m\times n}$ into the product of two lower-rank nonnegative matrices $W \in R_+^{m\times r}$ and $H \in R_+^{r\times n}$ by minimizing the following objective function

$$\min_{W\geq 0, H\geq 0} \|V - WH\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ signifies the Frobenius norm. The squared Frobenius in (1) measures the loss of decomposing V into WH.

2.1. NMF-based Speech Separation

Since the speech signals from different speakers are assumed to be additive, it is reasonable to use NMF to separate the mixture signals. However, NMF is unsuitable for time-domain signals because they contain negative entries. We therefore transform time-domain signals to the frequency-domain by using the short-time Fourier transform (STFT). Given a sequence of time-domain signal $v(t)$, its magnitude spectrogram is

$$V = |Y|, \tag{2}$$

where Y denotes the frequency-domain signal obtained by STFT on $v(t)$, and the operator $|\cdot|$ outputs the modulus.

It is obvious that V is nonnegative and NMF can be conducted to decompose the magnitude spectrogram of $v(t)$, i.e., $V \approx WH$, where the spectral basis W represents phonemic features including pitch inflections and consonant sounds, and H signifies the activations [8]. For supervised speech separation [13], NMF is conducted independently on the magnitude spectrogram of the training speech signals of each speaker, and the obtained spectral bases are concatenated to form a dictionary for separating the mixture signals.

Given $p \geq 2$ speakers and their training speech signals, NMF is utilized to learn several spectral bases for each speaker. Let $V_k \in R_+^{m\times n_k}$ denote the magnitude spectrogram of the training speech signals of the k-th speaker, by conducting NMF independently on each V_k, we have

$$V_k \approx W_k H_k, \tag{3}$$

where $W_k \in R_+^{m\times r}$ signifies the spectral basis learned for the k-th speaker. Let $v^m(t)$ denote the time-domain mixture signals and Y^m denote its frequency-domain signals, and $V^m \in R_+^{m\times n}$ denotes the magnitude spectrogram, a traditional NMF-based method decomposes it by

$$V^m \approx W^m H^m, \tag{4}$$

where $W^m = [W_1, \cdots, W_p]$ is constructed by concatenating spectral bases of all speakers and $H^m \in R_+^{rp\times n}$ signifies the obtained activations.

In the separation stage, by decomposing H^m into $H^m = [{H_1^m}^T, \cdots, {H_p^m}^T]^T$ according to the construction of W^m, it is easy to verify that

$$V^m \approx \sum_{k=1}^{p} V_k^m, \tag{5}$$

where $V_k^m = W_k H_k^m$. Eq. (5) means that the mixture signals are decomposed into p components, each of which corresponds to the magnitude spectrogram of one speaker. According to [14], the frequency-domain signals for the k-th speaker are recovered by

$$Y_k^m = \frac{V_k^m}{\sum_{k=1}^{p} V_k^m} \circ Y^m, \tag{6}$$

where the operator $\circ$ denotes the dot product.

Based on the NMF framework, Joder *et al*. [15] proposed to efficiently separate speech from noisy mixture signals with a hybrid method, i.e., $V \approx [W_s, W_n]H + (\vec{1} \times h) \circ B$, where W_s is estimated a-priori from training speech signals, and B is a noise estimate computed a-priori. By iteratively updating H, W_n, and h with multiplicative update rules, such a hybrid method successfully removes both stationary and time-varying noises from V.

3. TRANSDUCTIVE NONNEGATIVE MATRIX FACTORIZATION

Since the training stage of learning the spectral bases (3) and the test stage of separating the mixture signals (4) are independent, NMF cannot accurately represent the mixture signals. By contrast, TNMF connects these two stages by simultaneously achieving two objectives, i.e., (3) and (4). The objective function of TNMF is

$$\min_{\forall 1\leq k\leq p, W_k\geq 0, H_k\geq 0, H^m\geq 0} \{\sum_{k=1}^{p} \|V_k - W_k H_k\|_F^2 + \lambda \|V^m - W^m H^m\|_F^2\}, \tag{7}$$

where $W^m = [W_1, \cdots, W_p]$, and λ is a positive constant that balances these two objectives.

TNMF jointly learns a dictionary on both training speech signals V_k from different speakers and the mixture signals V^m to be separated. Since TNMF transduces the mixture signals to the learned dictionary by incorporating the second term in (7), it represents the mixture signals more accurately and overcomes the deficiency of NMF. Experimental results confirm that TNMF greatly boosts the separation performance.

Although the objective function of TNMF is jointly non-convex with respect to all variables $\{W_1, \cdots, W_p, H_1, \cdots, H_p, H^m\}$, it is convex with respect to each of them separately. According to [12][5][19], we utilized the majorization minimization (MM) method to derive a multiplicative update rule (MUR) for solving (7). MUR updates W_k, H_k, and H^m by

$$W_k \leftarrow W_k \circ \frac{V_k H_k^T + \lambda V^m H_k^{mT}}{W_k H_k H_k^T + \lambda W^m H^m H_k^{mT}}, \tag{8}$$

$$H_k \leftarrow H_k \circ \frac{W_k^T V_k}{W_k^T W_k H_k}, \tag{9}$$

and

$$H^m \leftarrow H^m \circ \frac{W^{mT} V^m}{W^{mT} W^m H^m}, \tag{10}$$

respectively, until they do not change the objective value (7). It is easy to prove that (8), (9), and (10) decrease the objective function [12]. We omit their proofs here for saving space. We can easily recover the sounds of each speaker by (6) from the solution obtained by MUR.

TNMF provides a flexible framework for semi-supervised high-performance speech separation which learns a phonemic dictionary from both training speech signals and mixture signals to be separated. It is suggested to address its extensions in future work.

4. EXPERIMENTS

In this section, we verify the effectiveness of the proposed TNMF-based semi-supervised speech separation method on the TIMIT dataset [18] by comparing with a traditional NMF-based supervised speech separation method. To evaluate both methods, we generated the mixture signals by synthetically summing two different roughly equal length speech segments from two speakers, i.e., a male (MDAB0) and a female (FAKS0). Another two speech segments from them were used for training. The training speech of each speaker is about 25 seconds, and the mixture speech is about 3 seconds long. All sounds are sampled at a rate of 16 kHz.

4.1. Evaluation Metrics

According to [8][19], we compared the performance of TNMF and NMF in terms of correlation-based measurements, i.e., similarity index (SI) and speaker ratio (SR). SI measures how much the recovered sound resembles the desired sound, and SR measures how much the signals of undesired speakers have been suppressed. Let $v_k^m(t)$ denote the recovered time-domain sound and $v_k(t)$ denote the ground-truth, SI is defined as

$$SI_k = 10 \log_{10} corr(v_k^m(t), v_k(t)), \tag{11}$$

where $corr(x, y)$ signifies the correlation between x and y. The higher SI_k, the more similar to the desired sounds of the k-th speaker. In this experiment, we utilized an inner product based correlation, i.e., $corr(x, y) = \frac{\langle x-\bar{x}, y-\bar{y}\rangle}{\|x-\bar{x}\|_2 \|y-\bar{y}\|_2}$, where $\bar{x}$ is the mean of x, $\langle \cdot, \cdot \rangle$ is the inner product, and $\| \cdot \|_2$ is the l_2-norm. Based on this correlation, SR is defined as

$$SR_k = 10 \log_{10} \frac{corr(v_k^m(t), v_k(t))}{\sum_{i \neq k} corr(v_k^m(t), v_i(t))}. \tag{12}$$

The higher SR_k, the better recovery of the k-th speaker.

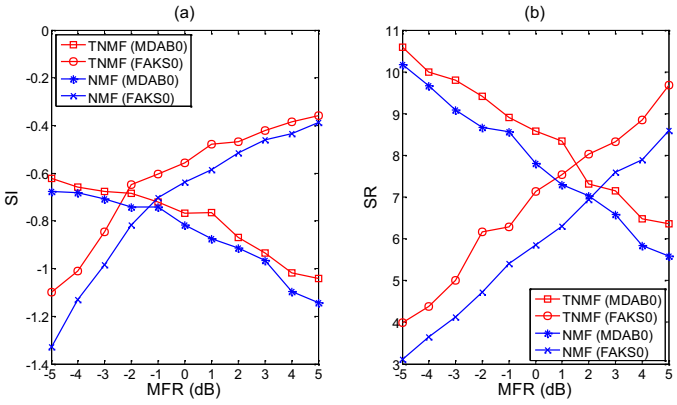

Fig. 1. The speech separation performance of TNMF and NMF in terms of SI (a) and SR (b) with the MFR varying from -5 dB to 5 dB.

4.2. Speech Separation

We compared TNMF with NMF in terms of both SI and SR and mixed the speech segments of both speakers at different MDAB0-to-FAKS0 ratios (MFR)[1] in the range of -5 dB to 5 dB to see how the algorithms work under different conditions. Figure 1(a) shows that the SIs of both speakers of TNMF are higher than those of NMF. It confirms that TNMF better recovers the speech segments of both speakers than NMF. Figure 1(b) shows that the SRs of both speakers of TNMF are significantly higher than those of NMF. That is because TNMF transduces the mixture signals to the learned dictionary and clearly recovers the desired sound for each speaker without interfering of another speaker.

[1]Similar to the signal-to-noise ratio (SNR) used in [7], we defined MDAB0-to-FAKS0 ratio (MFR) as quotient of the strength of the sentence of MDAB0 divided by the strength of the sentence of FAKS0 to simulate the condition of signal mixtures.

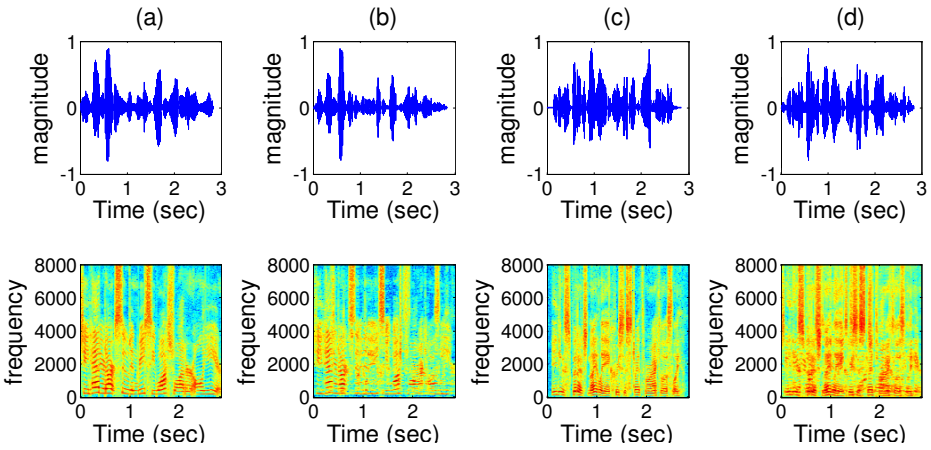

Fig. 2. The magnitudes of the time-domain signals (top) and frequency-domain signals (bottom) of the original speech (column a) and recovered speech (column b) of FAKS0, and the original speech (column c) and recovered speech (column d) of MDAB0 by TNMF.

To further illustrate the effectiveness of TNMF, Figure 2 shows the magnitudes of both pure speech and recovered speech by TNMF for both speakers in time-domain and frequency-domain. Figure 2 shows that TNMF almost perfectly recovers the desired speech for both speakers.

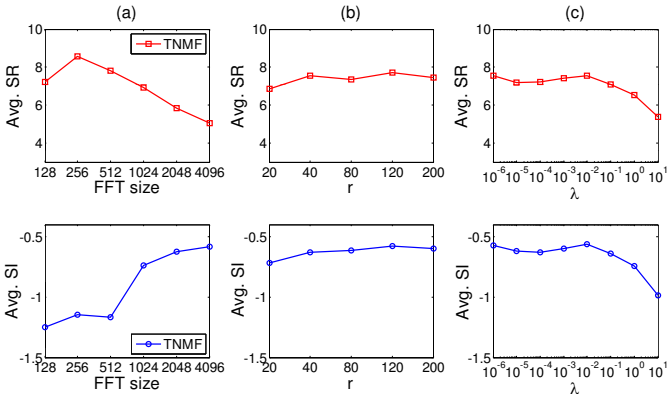

Fig. 3. The magnitudes of the time-domain signals (top) and frequency-domain signals (bottom) of the pure speech (column a) and recovered speech (column b) of FAKS0, and the pure speech (column c) and recovered speech (column d) of MDAB0 by TNMF.

4.3. Parameter Selection

There are three main parameters in the proposed TNMF-based semi-supervised high-performance speech separation framework. They are the FFT size, number of spectral bases r and the trade-off parameter λ. In this experiment, we evaluated their influences on the separation performance in terms of average SI and average SR of the recovered speech segments for both speakers by TNMF with the FFT size, r, and λ varying in the ranges of $\{2^{7+i}|i = 0, \cdots, 5\}$, $\{20 \times i|i = 1, 2, 4, 6, 10\}$, and $\{10^i|i = -6, \cdots, 1\}$, respectively.

Figure 3 gives the cross-validation results. The column (a) shows that it is reasonable to set the FFT size to 1024, and we kept this setting in the following experiments. The column (b) shows that the TNMF model is stable when the number of spectral bases $r > 40$ and the peak is reached when $r = 120$. The column (c) shows that the proposed TNMF model is stable when λ varies in a wide range from 10^{-6} to 0.1. Interestingly, the performance is significantly worsened when λ becomes larger than 0.1. That is because the mixture signals term in (7) might contaminate the learned spectral bases of one speaker by the phonemic features of another speaker in this case. This experiment shows that it is easy to determine the parameters in the TNMF framework.

5. CONCLUSION

This paper proposed transductive nonnegative matrix factorization (TNMF) for semi-supervised high-performance speech separation. TNMF surmises that the content information of the mixture speech is useful for speech separation, and thus it jointly decomposes the training speech segments and mixed speech to transfer the information and obtain more meaningful dictionaries. We apply TNMF in separating two known different gender speakers. Experimental results show that TNMF effectively recovers the original speech and achieves a higher separation performance than NMF.

6. REFERENCES

[1] Sprietand S. Jansen N. Madhuand A., Koning R., and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 63–72, 2013.

[2] Ji Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[3] C. Joder, Weninger F., Eyben F., Virette D., and Schuller B., "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proceedings 10th International Conference on Latent Variable Analysis and Signal Separation (LVA ICA 2012), Special Session Real-world Constraints and Opportunities in Audio Source Separation, Tel Aviv, Israel*, 2012, vol. 7191, pp. 323–329.

[4] Noboru Murata, Shiro Ikeda, and Andreas Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[5] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Non-negative patch alignment framework," *Neural Networks, IEEE Transactions on*, vol. 22, no. 8, pp. 1218–1230, 2011.

[6] Naiyang Guan, Dacheng Tao, Zhigang Luo, and John Shawe-Taylor, "Mahnmf: Manhattan non-negative matrix factorization," *arXiv preprint arXiv:1207.3438*, 2012.

[7] Schmidt M.N., J. Larsen, and F.T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Workshop on Machine Learning for Signal Processing*, 2007, pp. 431–436.

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

[9] Fevotte C., "Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization," 2011.

[10] Fevotte C., Bertin N., and Durrieu J.L., "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[11] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *International Conference on Machine Learning*, 1999, vol. 99, pp. 200–209.

[12] Lee D.D. and Seung H.S., "Algorithm for non-negative matrix factorization," in *Advance in Neural Information Proceeding Systems*, 2001, vol. 13.

[13] Weninger F. and Schuller B., "Supervised and semi-supervised supression of background music in monaural speech recordings," in *Proceeding 37th International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan*. 2012, pp. 61–64, IEEE.

[14] Mohammadiha N., Smaragdis P., and Leijon A., "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.

[15] Weninger C., Joderand F., Virette D., and Schuller B., "Integrating noise estimation and factorization-based speech separation: A novel hybrid approach," in *Proceedings 38th IEEE International Conference on Acoustics, Speech,and Signal Processing, ICASSP 2013, Vancouver, Canada*, May 2013, pp. 131–135.

[16] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Nenmf: an optimal gradient method for non-negative matrix factorization," *Signal Processing, IEEE Transactions on*, vol. 60, no. 6, pp. 2882–2898, 2012.

[17] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 7, pp. 1087–1099, 2012.

[18] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[19] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Manifold regularized discriminative non-negative matrix factorization with fast gradient descent," *Image Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2030–2048, 2011.