DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION FOR SINGLE-CHANNEL SPEECH SEPARATION

Zi Wang

Department of Computer Science and Technology
Tsinghua University
Beijing 100084, China

Fei Sha

Department of Computer Science
University of Southern California
Los Angeles, CA 90089

ABSTRACT

Non-negative matrix factorization (NMF) has emerged as a promising approach for single-channel speech separation. In this paper, we propose a new method of *discriminative learning* of NMF. In contrast to conventional approaches where the basis vectors are learned independently on clean signals from each speaker, our approach optimizes all basis vectors jointly to reconstruct both clean signals and mixed signals well. Our empirical studies validated our approach. Specifically, discriminative NMF outperforms standard methods by a large margin in improving signal-to-noise ratio for reconstructing signals.

Index Terms— non-negative matrix factorization, discriminative training, speech separation

1. INTRODUCTION

Sound source separation is a classical problem in auditory scene analysis [1]. In particular, single-channel speech separation – extracting individual streams of speech from a mixed signal of several speakers – is a challenging task with many applications in robust automatic speech recognition, speech enhancement, and others.

Recently, non-negative matrix factorization (NMF) has been extensively investigated and has since emerged as a promising approach for this task [2, 3, 4, 5, 6, 7]. Specifically, the magnitude spectrogram of the mixed signal is modeled as the *additive* superposition of the spectrograms of individual speakers. Moreover, each individual speaker's spectrograms are modeled as non-negative combinations of basis vectors. The basis vectors are speaker-dependent and learned in unsupervised manner from those speakers' speech samples. Once learned, NMF identifies the optimal combination coefficients for any mixed signal and use them to reconstruct each sound source. Note that learning the basic vectors for one speaker is completely independent of learning for another.

One drawback of this type of independent learning is that the learned basis vectors excel at the task of reconstructing signals but *not* at lifting the corresponding signals out of mixed ones. In particular, since learning the basis vectors for one speaker occurs without any knowledge (or interference) of other speakers, the basis vectors for this speaker are not *discriminative* enough to tell apart which elements in the mixed signals are the most relevant. For instance, it is possible that the basis vectors learned for one speaker also models well other speakers' speech signals [7].

In this paper, we propose a new approach to learn discriminative basis for NMF. Given a set of speakers and their speech samples, we construct discriminative tasks by mixing speech samples from different speakers. Our discriminative method optimizes basis vectors such that they can be used to reconstruct speech signals well even when the target signals are mixed with others. One distinctive property of the proposed method is that the basis vectors for all speakers are learned *jointly*.

The numerical optimization procedure is an instance of coordinate-descent where at each iteration, we select one speaker and optimize the corresponding basis vectors while holding the basis vectors from other speakers fixed. The updates inherit the multiplicative form of NMF and converge monotonically to a local optimum.

We evaluate the proposed method for speech separation using the GRID speech corpus with sparse NMF (SNMF) as baseline. The empirical study validates our approach, which improves the evaluation metric of signal-to-noise ratio significantly over the baseline for reconstructing signals.

Related work [6] is similar to our work in spirit, but they only learn the basis vectors for one sound source discriminatively. [7] minimizes the overlapping among basis vectors to prevent one speaker's basis vectors to model other speakers well. [8] learns the basis vector in a supervised manner by assuming the knowledge ideal combination coefficient.

2. NMF AND ITS VARIANTS

The main idea of non-negative matrix factorization (NMF) is to model data with additive and non-negative parts. In the context of modeling speech data with NMF, we represent speech signals with the non-negative magnitudes of their

spectra – in this paper, we use mel spectra, the power spectra mapped onto the mel scale. In what follows, we describe NMF first then sparse NMF.

2.1. Non-negative matrix factorization

Let $\boldsymbol{S}_a \in \mathbb{R}^{\mathsf{D}\times\mathsf{T}}$ denote the mel-spectra computed from the a-th speaker's speech samples. D denotes the number of mel scale and T denote the number of analysis windows (i.e., frames). For notational simplicity, we assume all speakers have the same T.

We model $\boldsymbol{S}_a$ as the non-negative combination of the speaker's basis vectors $\boldsymbol{W}_a \in \mathbb{R}^{\mathsf{D}\times\mathsf{K}}$ where K is the number of basis vectors, assumed to be the same for all speakers too. The combination coefficients are denoted by $\boldsymbol{H}_a \in \mathbb{R}^{\mathsf{K}\times\mathsf{T}}$. We constrain both the basis vectors and the coefficients to be non-negative, modeling data as superpositions of parts [2].

The aim of NMF is to approximate $\boldsymbol{S}_a$ as much as possible under those constraints, using the reconstruction computed from linear combination

$$\boldsymbol{R}_a = \boldsymbol{W}_a\boldsymbol{H}_a \approx \boldsymbol{S}_a \tag{1}$$

In this paper, we use generalized KL divergence to measure the approximation errors so as to exemplify the NMF algorithm, though other choices are possible too [9]:

$$J(\boldsymbol{S}_a, \boldsymbol{R}_a) = \mathrm{Tr}[(\boldsymbol{S}_a \odot \log(\boldsymbol{S}_a \oslash \boldsymbol{R}_a) - \boldsymbol{S}_a + \boldsymbol{R}_a)\mathbf{1}_{\mathsf{T},\mathsf{D}}] \tag{2}$$

where $\odot$ and $\oslash$ stand for Hadamard product and division, respectively. $\mathbf{1}_{\mathsf{P},\mathsf{Q}}$ stands for an all-one element matrix with dimensions of $\mathsf{P}\times\mathsf{Q}$. When the context is clear, we omit the dimensions of those special matrices. For instance, the trace operator Tr applies only when the argument is a square matrix, which can be used to infer the dimensions. The $\log$ is defaulted to be element-wise logarithm. Throughout the rest of the paper, we impose the order of algebraic evaluations is to associate from the left to the right, except where the parentheses and functions (such as $\log$) take precedence.

2.2. Sparse non-negative matrix factorization

The original NMF described in the previous section often achieves sparse solutions where the basis vectors correspond to distinctive parts, such as localized facial structures (e.g., nose, eyebrows) [2]. This means we only need a subset of basis vectors to represent unknown signals and these basis vectors are necessarily redundant.

Sparse NMF further improves the spareness in the solutions by learning basis vectors while enforcing a sparsity regularization on the combination coefficients [10]. Concretely, the method minimizes the following reconstruction error

$$J_\lambda(\boldsymbol{S}_a, \boldsymbol{R}_a) = J(\boldsymbol{S}_a, \boldsymbol{R}_a) + \lambda\,\mathrm{Tr}[\boldsymbol{H}_a\mathbf{1}] \tag{3}$$

The ℓ_1-norm of the coefficients promotes sparseness in $\boldsymbol{H}_a$, similar to its role in compressive sensing, thus forcing learning a set of basis vectors that are overcomplete.

In practice, however, having a small-valued regularizer $\boldsymbol{H}_a$ can be achieved by scaling $\boldsymbol{H}_a$ to arbitrarily small values while compensating by scaling up the basis vectors. This type of scaling does not promote sparseness. Instead, we need to hold the the basis vectors at a fixed scale such that minimizing the objective function $J_\lambda(\boldsymbol{S}_a, \boldsymbol{R}_a)$ can be driven by making $\boldsymbol{H}_a$ sparse.

To this end, the reconstruction $\boldsymbol{R}_a$ is computed using the column-wise normalized basis vector

$$\boldsymbol{R}_a = \overline{\boldsymbol{W}}_a\boldsymbol{H}_a, \text{ with } \overline{\boldsymbol{W}}_{adk} = \frac{W_{adk}}{\sqrt{\sum_d W_{adk}^2}} \tag{4}$$

Learning $\boldsymbol{W}_a$ and $\boldsymbol{H}_a$ can be cast in multiplicative updates, though the monotonic convergence property is lost,

$$\begin{aligned}
\boldsymbol{H} &\leftarrow \boldsymbol{H}\odot(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V} \oslash (\boldsymbol{W}^{\mathrm{T}}\mathbf{1}_{\mathsf{D},\mathsf{T}} + \lambda)) \\
\boldsymbol{W} &\leftarrow \boldsymbol{W}\odot\Big\{[\boldsymbol{V}\boldsymbol{H}^{\mathrm{T}} + \mathbf{1}_{\mathsf{D},\mathsf{D}}(\mathbf{1}_{\mathsf{D},\mathsf{T}}\boldsymbol{H}^{\mathrm{T}} \odot \boldsymbol{W}) \odot \boldsymbol{W}] \\
&\qquad \oslash [\mathbf{1}_{\mathsf{D},\mathsf{T}}\boldsymbol{H}^{\mathrm{T}} + \mathbf{1}_{\mathsf{D},\mathsf{D}}(\boldsymbol{V}\boldsymbol{H}^{\mathrm{T}} \odot \boldsymbol{W}) \odot \boldsymbol{W}]\Big\}
\end{aligned} \tag{5}$$

where $\boldsymbol{V} = \boldsymbol{S} \oslash \boldsymbol{R}$. We have dropped the subscript a to avoid notation cluttering. Essentially, those updates converge to stationary points of the objective function where the gradients are zeros. The derivation is similar to what is in [11].

3. DISCRIMINATIVE NMF

The methods of NMF and its variants, as described in the previous section, learn each speaker's basis vectors independently. They are optimized to reconstruct clean signals when there is no interference from other speakers. Thus, while our goal is to separate mixed speech, such learning strategy does not lead to basis vectors that optimize the quality of reconstructed signals under interference. For example, if the basis vectors for one of the two speakers can also model the other speaker well, then it is difficult to see how we can reliably extract both speakers' speech well.

In what follows, we describe our approach of discriminative NMF for tackling this problem. The key idea is to learn basis vectors such that we have good reconstructions for both clean and mixed signals.

Concretely, given a training corpus of speech signals of several speakers, we mix those signals in pairwise and artificially construct many speech separation tasks. Specifically, assume we have N speakers and each speaker has M utterances. We further assume in the interest of simplifying notation, each utterance is equal in length thus yielding the same number of analysis frames T. We will construct a total $\mathsf{N}(\mathsf{N}-1)\mathsf{M}^2$ speech separation tasks.

Let $\boldsymbol{S}_{ai}$ denote the mel-spectra of the a-th speaker's i-th sentence and $\boldsymbol{S}_{abij}$ denote the mel-spectra of the signal from mixing the a-th speaker's i-th sentence with the b-th speaker's j-th sentence.

Our objective function for learning the basis vector consists of two major components. The first component is analogous to the regular NMF, where we desire the basis vectors

can reconstruct clean signals well for every speaker:

$$J_1 = \sum_{a,i} J(\boldsymbol{S}_{ai}, \boldsymbol{R}_{ai}) = \sum_{a,i} J(\boldsymbol{S}_{ai}, \overline{\boldsymbol{W}}_a \boldsymbol{H}_{ai}) \tag{6}$$

where $\boldsymbol{H}_{ai}$ is the reconstruction coefficients for the corresponding pair of speaker and sentence.

The second component is composed of the objective functions for the speech separation tasks we have created

$$\begin{aligned} J_2 &= \sum_a \sum_{b \neq a,i,j} J(\boldsymbol{S}_{abij}, \boldsymbol{R}_{abij}) \\ &= \sum_a \sum_{b \neq a,i,j} J(\boldsymbol{S}_{abij}, \overline{\boldsymbol{W}}_a \boldsymbol{H}_{ai} + \overline{\boldsymbol{W}}_b \boldsymbol{H}_{bj}) \end{aligned} \tag{7}$$

while we have imposed that the reconstruction for the mixed signals should be computed using the combination coefficients from the *clean* signals.

The form of J_2 couples J_1 together: it is not sufficient to minimize J_1 so as to reconstruct well only for clean signals, as in the regular NMF. The basis vectors need to be discriminatively optimized such that they can also be used to compute reconstruction signals to approximate the mixed signals.

We balance the two forces by optimizing a joint objective function, also combined with a sparsity regularizer,

$$J = \mathsf{M} J_1 + J_2 + \lambda\alpha \sum_{a,i} \mathrm{Tr}[\boldsymbol{H}_{ai}\mathbf{1}] \tag{8}$$

The pre factor M scales J_1 up to match J_2. Similarly, α is used to scale the regularizer up to match J_1 and J_2 so that we can compare to the regular sparse NMF using the same regularizer strength λ. In practice, we choose $\alpha = \mathsf{NM}$, reflecting the number of times any $\boldsymbol{R}_{ai}$ occurring in J.

Algorithmically, optimizing J entails jointly optimizing all basis vectors together because of the mutual dependency.We propose the following iterative strategy to optimize each speaker's basis vectors in turn.

The numerical optimization consists of two main loops. The outside loop selects (in round-robbin) a speaker. We then hold all other speakers' basis vectors and combination coefficients as fixed. The inside loop optimizes the basis vectors and the combination coefficients of the selected speaker. Note that J_1 now contains only one adjustable terms, while J_2 now contains only $(\mathsf{N}-1)\mathsf{M}^2$ terms – all related to the selected speaker. All these terms are in the form of generalized KL divergence, except their basis vectors and combination coefficients are tied.

$$\begin{aligned} \boldsymbol{H}_{ai} &\leftarrow \boldsymbol{H}_{ai} \odot \left\{ \boldsymbol{W}_a^{\mathrm{T}} \sum_{b,j} \boldsymbol{V} \oslash \left(\mathsf{NM}(\boldsymbol{W}_a^{\mathrm{T}}\mathbf{1} + \lambda) \right) \right\} \\ \boldsymbol{W}_a &\leftarrow \boldsymbol{W}_a \odot \left\{ \sum_{i,b,j} \left(\boldsymbol{V}\boldsymbol{H}_{ai}^{\mathrm{T}} + \mathbf{1}(\mathbf{1}\boldsymbol{H}_{ai}^{\mathrm{T}} \odot \boldsymbol{W}_a) \odot \boldsymbol{W}_a \right) \right. \\ &\qquad \left. \oslash \sum_{i,b,j} \left(\mathbf{1}\boldsymbol{H}_{ai}^{\mathrm{T}} + \mathbf{1}(\boldsymbol{V}\boldsymbol{H}_{ai}^{\mathrm{T}} \odot \boldsymbol{W}_a) \odot \boldsymbol{W}_a \right) \right\} \end{aligned} \tag{9}$$

where $\boldsymbol{V} = \boldsymbol{S}_{abij} \oslash \boldsymbol{R}_{abij}$ (We assume when $b = a$, $S_{abij} = S_{ai}, R_{abij} = R_{ai}$). $\boldsymbol{W}_a$ is normalized after each update. Intuitively, tied parameters lead to accumulating gradients from all related error terms.

4. EXPERIMENTS

We evaluate our method of discriminative NMF (DNMF) and contrast to (sparse) NMF (SNMF) for speech recognition.

4.1. Setup

We use the speech data in the Grid Corpus [12]. The corpus consists of 34 speakers, each speaking 1000 short sentences. We randomly select 8 out of 34 speakers and use the first 500 sentences for training and the rest half for evaluating.

The mel-spectra of the speech data was extracted using the package describe in [13], similar to what was used in previous works on this dataset [4]. Concretely, the signals were pre-emphasized with a FIR filter $1-0.95z^{-1}$ and analyzed with a 32-ms window (800 samples at 25KHz sampling rate) sliding at 16 milliseconds. The power spectra were then mapped to 80 mel scales covering from 0Hz to 4KHz. For DNMF, a long sentence containing 10 short sentences was grouped for each speaker – the sole purpose is to reduce the number of artificial speech separation tasks we need to create in order to discriminatively learn basis vectors.

We randomly initialize basis vectors and combination coefficients, for NMF, sparse NMF and DNMF. We stop the iterative procedures if the maximum number of iterations, which is 100, is reached or the improvement over successive iterations drops below the threshold of 0.001.

We examine how well signals are separated from mixed signals in two ways. Our first metric is the signal-noise-ratio (SNR) on mel-spectra. Specifically, given the spectra $\boldsymbol{S}$ of the clean signal and its reconstruction $\boldsymbol{R}$, this metric is defined as: $\mathrm{SNR}^{\mathrm{MEL}} = 10\log_{10}\frac{\sum_{d,t} S_{dt}^2}{\sum_{d,t}(S_{dt}-R_{dt})^2}$. Note that this metric keeps close track of our objective function.

Our second metric is $\mathrm{SNR}^{\mathrm{WAV}}$– SNR on reconstructed speech waveforms, inverted from mel-spectra. We first invert reconstructed mel-spectra into frequency domain with the phase of the mixed speech. Then the spectral mask was computed and applied to the spectrogram of the mixed signal. We extract each speaker's spectrogram and then invert the spectrogram to obtain the final estimate of the speech waveforms. This procedure is inspired by the recent work estimating spectral mask as an intermediate step of obtaining speech waveforms [14]. Note that this procedure tends to eclipse the differences of different methods.

For either metric, we sample 5 sentences from the evaluation set, for every pair of mixed speakers. We report averaged metrics over the total 140 speech separation tasks.

We learn models with the sparsity parameter λ ranging within $\{0, 10^{-4}, 10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$. The

	K \ λ	0	0.0001	0.001	0.01	0.1
		SNR$^{\text{MEL}}$				
DNMF	70	4.84	4.69	4.70	5.07	5.28
	140	5.43	5.43	5.40	5.44	5.72
	280	6.03	6.08	6.05	6.19	6.44
	560	6.51	6.43	6.41	6.52	**6.71**
SNMF	70	4.42	4.42	4.42	4.59	4.91
	140	4.56	4.55	4.61	5.13	5.69
	280	5.30	5.30	5.31	5.67	5.80
	560	5.68	5.68	5.76	**6.02**	5.99
		SNR$^{\text{WAV}}$				
DNMF	70	4.49	4.46	4.39	4.51	4.76
	140	4.54	4.53	4.53	4.67	4.93
	280	4.71	4.67	4.72	4.76	5.08
	560	4.82	4.76	4.83	4.87	**5.16**
SNMF	70	4.29	4.29	4.30	4.39	4.71
	140	4.25	4.25	4.27	4.48	4.87
	280	4.41	4.41	4.44	4.64	4.91
	560	4.52	4.52	4.58	4.72	**5.01**

Table 1. Contrast DNMF to SNMF with average signal-to-noise ratio (SNR) computed on both mel-spectra and speech waveforms.

number of basis vectors K ranges within $\{70, 140, 280, 560\}$.

4.2. Contrast DNMF to NMF and SNMF

Table 1 displays two SNR metrics for a few representative λ. Note that when $\lambda = 0$, SNMF reduces to the regular NMF. We observe that in both metrics, DNMF outperforms SNMF in almost all settings. The improvement is the most significant in reconstructing mel-spectra, attaining relatively 11.5%.

Another trend is that both SNMF and DNMF improve as λ increases. This is demonstrated in Fig. 1 where K is set to 560. SNMF seems to benefit more from a stronger sparse regularization (corresponding to a larger λ). However, as pointed in [4], too much sparsity does not lead to useful basis vectors. Thus, we follow the same advice by letting $\lambda \leq 0.1$.

4.3. Sex difference

It is also insightful to understand on which aspect DNMF improves the most than SNMF. To this end, we compare the two methods under two different settings in Table 2: mixing speakers of the same sex (SS) , or of opposite sexes (OS).

Note that DNMF improves over SNMF much more when mixing speakers of the same sex. The relative improvement is around 14%, while on opposite sex, the improvement is around 8.7%. This supports our intuition: for signals from the opposite sexes, the basis vectors from one speaker are unlikely to be able to model the other speaker well. Thus, discriminative training does not help much. On the other end, for mixed signals from the same sex, the basis vectors from two different speakers are likely to be able to model the other well. Thus, discriminative training is able to prevent basis vectors from generalizing too much (across speakers).

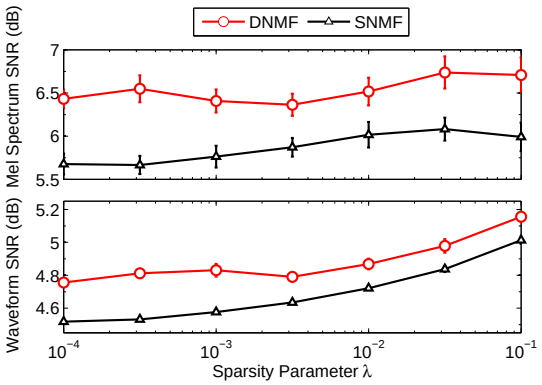

Fig. 1. Contrast DNMF to SNMF in SNR$^{\text{MEL}}$ under different sparsity regularization strength. The number of components is K = 560.

	λ	0	0.0001	0.001	0.01	0.1
DNMF	SS	4.79	4.77	4.71	4.91	**5.28**
SNMF	SS	4.31	4.30	4.34	4.50	**4.63**
DNMF	OS	7.80	7.67	7.68	7.72	**7.78**
SNMF	OS	6.71	6.71	6.83	**7.16**	7.02

Table 2. SNR$^{\text{MEL}}$ of DNMF and SNMF under different setting of mixing speakers. SS: same sex, OS: opposite sex. K = 560

5. CONCLUSIONS

We have developed a new method for discriminative learning of NMF for speech separation. The key idea is to learn speaker-specific basis vectors jointly, in contrast to the standard approach of learning them independently. The discriminative learned basis vectors are capable of reconstructing signals when they are clean *and* when they are mixed with interfering ones. Empirical studies validate our approach, which outperforms baselines of other NMF methods by a large margin of improvement in signal-noise-ratio. We plan to investigate alternative discriminative learning criteria and algorithms in the future.

6. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

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

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[4] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

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

[6] F. Weninger, J. Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012.

[7] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Internaltional Conference on Spoken Language Processing (INTERSPEECH)*, 2013.

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

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems (NIPS)*, vol. 13, pp. 556–562, 2001.

[10] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing*. IEEE, 2002.

[11] J. Eggert and E. Korner, "Sparse coding and nmf," in *Neural Networks*. IEEE, 2004, vol. 4, pp. 2529–2533.

[12] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421, 2006.

[13] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2006.

[14] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks," in *International Conference on Spoken Language Processing (INTERSPEECH)*, 2011.