Robust Non-negative Matrix Factorization with β-Divergence for Speech Separation

Yinan Li, Xiongwei Zhang, and Meng Sun

This paper addresses the problem of unsupervised speech separation based on robust non-negative matrix factorization (RNMF) with β-divergence, when neither speech nor noise training data is available beforehand. We propose a robust version of non-negative matrix factorization, inspired by the recently developed sparse and low-rank decomposition, in which the data matrix is decomposed into the sum of a low-rank matrix and a sparse matrix. Efficient multiplicative update rules to minimize the β-divergence-based cost function are derived. A convolutional extension of the proposed algorithm is also proposed, which considers the time dependency of the non-negative noise bases. Experimental speech separation results show that the proposed convolutional RNMF successfully separates the repeating time-varying spectral structures from the magnitude spectrum of the mixture, and does so without any prior training.

Keywords: Robust non-negative matrix factorization, Speech separation, Sparse and low-rank decomposition, β-divergence, Convolutional bases.

I. Introduction

Non-negative matrix factorization (NMF) is an efficient tool for extracting perceptually meaningful components from mixtures [1], and has been extensively investigated within the context of several speech processing tasks such as speech separation [2], [3] and speech enhancement [4], [5]. To extract features for the involved constituent sources, a prior learning procedure to obtain the NMF bases is usually required. However, training data for the encountered speech and noise are not always available beforehand, which significantly limits the practicality of this approach. In this paper, we investigate an unsupervised approach to speech separation, capable of alleviating the reliance on prior training.

An emerging technique for sparse and low rank decomposition, robust principal component analysis (RPCA) [6], has recently gained much attention. This technique decomposes the input data matrix into a sum of a low-rank matrix and a sparse matrix, in a well-behaved convex optimization framework. This mathematical model is particularly suitable for designing systems for unsupervised source separation, as it requires neither prior training nor handcrafted features. An RPCA was applied to separate voices from the background musical accompaniment in [7], by assuming that the musical accompaniment lies in a low-rank subspace because of its repeating nature—whereas a singing voice has rich variations and a harmonic structure that makes its spectrum relatively sparse. Based on these assumptions, the problem of source separation can be formulated as a decomposition of the mixed magnitude spectrum via RPCA, to estimate its underlying low-rank and sparse matrices. Several modifications to this basic approach—sharing the same principles—have been investigated, to further improve source separation performance [8], [9].

However, there is no guarantee in RPCA that all entries of the

Manuscript received Feb. 9, 2015; revised Oct. 4, 2016; accepted Nov. 3, 2016. This work was supported by the National Natural Science Foundation of China (grants 61402519 and 61471394) and the National Science Foundation of Jiangsu Province, China (grants BK2012510, BK20140071, and BK20140074).

Yinan Li (corresponding author, audiolyn@163.com), Xiongwei Zhang (xwZhang@gmail.com), and Meng Sun (sunmengccjs@gmail.com) are with the Lab of Intelligent Information Processing, PLA University of Science and Technology, Nanjing, China.

This is an Open Access article distributed under the term of Korea Open Government License (KOGL) Type 4: Source Indiction + Commercial Use Prohibition + Change Prohibition (http://www.kogl.or.kr/news/dataFileDown.do?dataIdx=71&dataFileIdx=2).

decomposed matrices will be non-negative. Eventually negative entries are difficult to interpret when trying to relate them to the underlying speech or noise structures. Considering the non-negativity of the input magnitude spectrum, it would be better if the entries of the constituent matrices could be forced to be non-negative; this is the first motivation for this paper.

The existing methods based on sparse and low-rank approximations—such as the robust non-negative matrix factorization (RNMF)—usually optimize a cost function defined by the squared error or Euclidean distance [10]. However, the Euclidean distance tends to overemphasize the reconstruction accuracy of large values—usually appearing in the low frequency regions—causing relatively large reconstruction errors in the higher frequency regions—which have relatively small values, but are perceptually important [11]. Divergence-based measures such as the generalized Kullback–Leibler (KL) divergence and the Itakura–Saito (IS) divergence have been found more appropriate in the context of speech separation [12], [13], because the divergence is sensitive to small values in the spectral approximation. However, as far as we know, RNMF was rarely investigated for these divergence metrics; this is the second motivation for this paper.

To tackle the issues mentioned above, we propose an RNMF method with β-divergence, which is capable of elegantly preserving the non-negativity of the matrices resulting from the decomposition. Moreover, by considering the time dependency of the background noise, we introduce a convolutional extension of RNMF to model the repeating time-varying spectra: the convolutional RNMF.

II. Robust Non-negative Matrix Factorization

In this section, we will first address the issue of robust non-negative matrix factorization, and subsequently derive multiplicative updating rules for RNMF. The convergence of the proposed algorithm is also considered. Finally, a convolutional version of RNMF is provided as an extension to conventional RNMF.

1. Optimization Problem Formulation

The goal of RNMF is to minimize the β-divergence between the non-negative matrix $\boldsymbol{Y}$ and its reconstruction $\boldsymbol{WH} + \boldsymbol{S}$, where $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{S}$ are constrained to be non-negative. The problem of RNMF can therefore be formulated as follows:

$$\begin{aligned} &\arg\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{S}} D_\beta\left(\boldsymbol{Y}\|\boldsymbol{WH}+\boldsymbol{S}\right)+\lambda\|\boldsymbol{S}\|_1 \\ &\text{s.t.}\quad \boldsymbol{Y}\in\mathbb{R}_{\geq 0}^{m\times n},\ \boldsymbol{W}\in\mathbb{R}_{\geq 0}^{m\times r},\ \boldsymbol{H}\in\mathbb{R}_{\geq 0}^{r\times n},\ \boldsymbol{S}\in\mathbb{R}_{\geq 0}^{m\times n}. \end{aligned} \tag{1}$$

The product of matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ is the low-rank approximation of $\boldsymbol{Y}$, protected by $\boldsymbol{S}$ from outlier corruption. Matrix $\boldsymbol{S}$ is constrained to be non-negative and sparse; parameter λ controls its sparsity. The error function $D_\beta(\cdot\|\cdot)$ is defined by the β-divergence, which covers a variety of cost functions that reflect the difference between the input and the reconstruction. The mathematical definition of β-divergence is:

$$D_\beta\left(x\|y\right)=\begin{cases}\dfrac{x^\beta+(\beta-1)y^\beta-\beta xy^{\beta-1}}{\beta(\beta-1)} & \beta\in\mathbb{R}\setminus\{0,1\},\\ x(\log x-\log y)+(y-x) & \beta=1,\\ \dfrac{x}{y}-\log\dfrac{x}{y}-1 & \beta=0.\end{cases} \tag{2}$$

The squared Euclidean distance, generalized KL divergence, and IS divergence correspond to the special cases of $\beta = 0$, $\beta = 1$, and $\beta = 2$, respectively.

2. Updating Rules for RNMF with β-Divergence

Inspired by the work of Lee and Seung [14], we derive multiplicative update rules for RNMF that promote convergence to the stationary points of (1), while ensuring the non-negativity of the parameter updates.

The derivative of $D_\beta(x\|y)$ with respect to y is continuous in β, and can be written as:

$$\nabla_y D_\beta(x\|y)=y^{\beta-2}(y-x)\,. \tag{3}$$

From this equation, the cost function gradients can be simply expressed as:

$$\begin{aligned}&\nabla_{\boldsymbol{W}} D_\beta\left(\boldsymbol{Y}\|\boldsymbol{WH}+\boldsymbol{S}\right)\\ &\quad=\left\{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{(\beta-2)}\odot\left(\boldsymbol{WH}+\boldsymbol{S}-\boldsymbol{Y}\right)\right\}\boldsymbol{H}^{\mathrm{T}},\end{aligned} \tag{4}$$

$$\begin{aligned}&\nabla_{\boldsymbol{H}} D_\beta\left(\boldsymbol{Y}\|\boldsymbol{WH}+\boldsymbol{S}\right)\\ &\quad=\boldsymbol{W}^{\mathrm{T}}\left\{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{(\beta-2)}\odot\left(\boldsymbol{WH}+\boldsymbol{S}-\boldsymbol{Y}\right)\right\},\end{aligned} \tag{5}$$

and

$$\begin{aligned}&\nabla_{\boldsymbol{S}} D_\beta\left(\boldsymbol{Y}\|\boldsymbol{WH}+\boldsymbol{S}\right)+\lambda\|\boldsymbol{S}\|_1\\ &\quad=\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{(\beta-1)}+\lambda\boldsymbol{I}-\boldsymbol{Y}\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{(\beta-2)},\end{aligned} \tag{6}$$

where T denotes transposition, and $\odot$ is the notation for the Hadamard product, the element-wise product of two matrices of the same size. $\boldsymbol{I}$ is an all-ones matrix with the same size of $\boldsymbol{Y}$. The exponentiations are also carried out element-wise.

Let $C(\theta)$ denote the cost as a function of θ; its gradient can be divided into a positive and a negative part, as follows:

$$\nabla C(\theta)=\nabla^+C(\theta)-\nabla^-C(\theta)\,, \tag{7}$$

where both $\nabla^+C(\theta)$ and $\nabla^-C(\theta)$ are non-negative. The update algorithms can be transformed into their multiplicative forms as $\theta\leftarrow\theta\odot\nabla^-C(\theta)./\nabla^+C(\theta)$ (the " $./$ " operator stands for element-wise division) to satisfy the non-negativity

constraint of θ [15], [16].

Following this principle, we can derive the multiplicative update rules for RNMF based on the results of (4) to (6):

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\left\{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-2)} \odot \boldsymbol{Y}\right\} \boldsymbol{H}^{\mathrm{T}}}{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-1)} \boldsymbol{H}^{\mathrm{T}}}, \tag{8}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^{\mathrm{T}}\left\{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-2)} \odot \boldsymbol{Y}\right\}}{\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-1)}}, \tag{9}$$

and

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{Y}\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-2)}}{\left(\boldsymbol{WH}+\boldsymbol{S}\right)^{\cdot(\beta-1)}+\lambda \cdot \boldsymbol{I}}, \tag{10}$$

where all divisions are carried out element-wise.

Note that all the multiplied elements are non-negative—which ensures the non-negativity of the updated matrices—and that the updating rules are conveniently implemented, because there is no need for users to define the updating rate.

3. Convergence of the Proposed Algorithm

Definition. $G(\theta \mid \hat{\theta})$ is an auxiliary function for $C(\theta)$ if it satisfies the condition

$$G(\theta|\hat{\theta}) \geq C(\theta), \quad G(\hat{\theta}|\hat{\theta}) = C(\hat{\theta}). \tag{11}$$

Different auxiliary functions for different selections of β can be found in [17].

Theorem. By updating the three matrices with (8) through (10), the cost function of RNMF will monotonically decrease, until it converges to a local minimum.

Proof. Let t denote the individual iterations; the decomposition of cost function $C^{(t)}$ after updating (8) to (10) will be denoted by $C_1^{(t)}$, $C_2^{(t)}$, and $C_3^{(t)}$, respectively.

The optimization procedure when fixing $\boldsymbol{H}$ and $\boldsymbol{S}$ to update $\boldsymbol{W}$, and when fixing $\boldsymbol{W}$ and $\boldsymbol{S}$ to update $\boldsymbol{H}$ are essentially the same; therefore, we will again use θ to represent either of them.

Given that there exists such an auxiliary function of $C(\theta)$, and the optimization of $G(\theta|\hat{\theta})$ over θ while fixing $\hat{\theta}$ is simple, we can simplify the optimization of $C(\theta)$ by replacing it with a simpler optimization $\theta = \arg_{\theta} \min G(\theta|\hat{\theta})$. The minimization of $C(\theta)$ can be conducted iteratively, because

$$C\left(\theta^{(t+1)}\right) \leq G\left(\theta^{(t+1)}|\theta^{(t)}\right) \leq G\left(\theta^{(t)}|\theta^{(t)}\right) = C\left(\theta^{(t)}\right). \tag{12}$$

From (12), we know that $C_3^{(t)} \geq C_1^{(t+1)} \geq C_2^{(t+1)}$. On the other hand, the global optimality of $\boldsymbol{S}^{(t)}$ ensures that $C_2^{(t)} \geq C_3^{(t)}$. Therefore, the objective function keeps decreasing as the sequence of iterations progresses:

$$C_1^{(1)} \geq C_2^{(1)} \geq C_3^{(1)} \geq C_1^{(2)} \geq \cdots \geq C_1^{(t)} \geq C_2^{(t)} \geq C_3^{(t)} \geq \cdots. \tag{13}$$

When $C^{(t+1)} = C^{(t)}$, the inequalities become equalities, and a local minimum is reached. ■

4. Convolutional RNMF

RNMF constitutes a useful tool for unsupervised speech separation. However, it ignores the potential time dependencies across successive frames of the input magnitude spectrum. Such dependencies are not uncommon, especially when a few repeating time-varying patterns span over multiple frames of the entire sequence. We represent a single pattern by a sequence of consecutive column vectors, defined as a non-negative basis function that spans the pattern length, as presented in [18].

The reconstruction approach in conventional NMF, $\boldsymbol{Y} \approx \boldsymbol{WH}$, was extended by the convolutional non-negative matrix factorization approach [19]–[21] as follows:

$$\boldsymbol{Y} \approx \boldsymbol{\Lambda} = \sum_{t=0}^{T-1} \boldsymbol{W}(t) \overset{t\rightarrow}{\boldsymbol{H}}, \tag{14}$$

where $\boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{m \times n}$ is the non-negative matrix to be approximated, $\boldsymbol{W}(t) \in \mathbb{R}_{\geq 0}^{m \times r}$ is a set of bases and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{r \times n}$ stores the weights of the bases. $\overset{t\rightarrow}{\boldsymbol{H}}$ shifts t columns of $\boldsymbol{H}$ to the right. Zeros are padded once the shifted columns reach the matrix boundaries. Analogously, $\overset{\leftarrow t}{\boldsymbol{H}}$ designates a t-column shift of $\boldsymbol{H}$ to the left. Therefore, with the above RNMF configuration, the objective function can be rewritten as:

$$\begin{aligned} &\arg\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{S}} D_{\beta}\left(\boldsymbol{Y} \| \sum_{t=0}^{T-1} \boldsymbol{W}(t) \overset{t\rightarrow}{\boldsymbol{H}} + \boldsymbol{S}\right) + \lambda \|\boldsymbol{S}\|_1 \\ &\text{s.t.} \quad \boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{m \times n}, \boldsymbol{W}(t) \in \mathbb{R}_{\geq 0}^{m \times r}, \boldsymbol{H} \in \mathbb{R}_{\geq 0}^{r \times n}, \boldsymbol{S} \in \mathbb{R}_{\geq 0}^{m \times n}. \end{aligned} \tag{15}$$

This objective function can be seen as an extension of its RNMF counterpart. It can be optimized by updating the time-sliced $\boldsymbol{W}(t)$ and the shifted approximation to $\boldsymbol{H}$ with the following rules:

$$\boldsymbol{W}(t) \leftarrow \boldsymbol{W}(t) \odot \frac{\left\{\left(\boldsymbol{\Lambda}+\boldsymbol{S}\right)^{\cdot(\beta-2)} \odot \boldsymbol{Y}\right\} \overset{t\rightarrow}{\boldsymbol{H}}^{\mathrm{T}}}{\left(\boldsymbol{\Lambda}+\boldsymbol{S}\right)^{\cdot(\beta-1)} \overset{t\rightarrow}{\boldsymbol{H}}^{\mathrm{T}}}, \tag{16}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}(t)^{\mathrm{T}}\left\{\left(\overset{\leftarrow t}{\boldsymbol{\Lambda}}+\overset{\leftarrow t}{\boldsymbol{S}}\right)^{\cdot(\beta-2)} \odot \overset{\leftarrow t}{\boldsymbol{Y}}\right\}}{\boldsymbol{W}(t)^{\mathrm{T}}\left(\overset{\leftarrow t}{\boldsymbol{\Lambda}}+\overset{\leftarrow t}{\boldsymbol{S}}\right)^{\cdot(\beta-1)}}, \tag{17}$$

and

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{Y}\left(\boldsymbol{\Lambda}+\boldsymbol{S}\right)^{\cdot(\beta-2)}}{\left(\boldsymbol{\Lambda}+\boldsymbol{S}\right)^{\cdot(\beta-1)}+\lambda \cdot \boldsymbol{I}}. \tag{18}$$

The fact that each pattern $\boldsymbol{W}(t)$ shares the same $\boldsymbol{H}$ will lead to a biased estimate. To eliminate this effect, we can take the average of all the updates of $\boldsymbol{H}$.

Note that the proposed convolutional RNMF (CRNMF) is a

convolutional extension of RNMF; when $T = 1$, it will result in the standard RNMF described above. As will be seen in Section IV, this type of configuration of the bases model is particular good at detecting the potentially repeating (but still time-varying) patterns.

III. Unsupervised Speech Separation

The overall framework for unsupervised speech separation using CRNMF is illustrated in Fig. 1. The involved procedures can generally be divided into two categories: sparse and low-rank decomposition, and post-processing.

We first store the noisy speech into a buffer and calculate the noisy magnitude spectrogram using a short-time Fourier transformation (STFT). The phase of the noisy speech (denoted by $\angle\cdot$) is stored for posterior clean speech synthesis.

Either RNMF or CRNMF is then adopted to decompose the spectrogram into three components: a low-rank non-negative matrix resulting from the product of two low-rank non-negative matrices, a sparse non-negative matrix representing clean speech, and a residual noise matrix. This type of decomposition is chosen because [22] has shown that, for the Euclidean distance metric, decomposing the noisy spectrum into three sub-matrices can provide better results than the use of the two sub-matrices alternative. However, only seldom has the issue of determining which divergence metric and residual noise assumptions are the most suitable in the context of speech separation. We will mainly focus on this issue, and explore it in the experimental section below.

Note that the cost functions of RNMF and CRNMF are both defined by β-divergence. The assumptions imposed on the residual noise distributions will vary when the cost function changes. In particular, when $\beta = 0$, 1, and 2, the residual noise is assumed to be Gamma, Poisson, and Gaussian, respectively.

Finally, to further boost the separation performance, the estimated clean speech and noise spectrograms are obtained by Wiener type filtering, a standard and widely used post-processing strategy for NMF-based speech separation.

$$\hat{\boldsymbol{S}} = \frac{\boldsymbol{S}}{\sum_{t=0}^{T-1} \boldsymbol{W}(t)\overrightarrow{\boldsymbol{H}}^{t} + \boldsymbol{S}} \odot \boldsymbol{Y}, \quad \hat{\boldsymbol{N}} = \frac{\sum_{t=0}^{T-1} \boldsymbol{W}(t)\overrightarrow{\boldsymbol{H}}^{t}}{\sum_{t=0}^{T-1} \boldsymbol{W}(t)\overrightarrow{\boldsymbol{H}}^{t} + \boldsymbol{S}} \odot \boldsymbol{Y}. \quad (19)$$

The clean speech and background noise waveforms can be estimated using the noisy phase $\angle\cdot$ and the inverse STFT of $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{N}}$.

It should be noted that the diagram in Fig. 1 is only describing a basic separation system to help focus on the selection of the divergence cost function to be used under the sparse and low-rank framework. The obtained performance can, however, be further improved through techniques such as adopting a universal speech dictionary [23], imposing temporal continuity to the sparse matrix [24], using an information fusion strategy [25], or a combination with autocorrelation [26]. The use of these techniques for performance improvement is beyond the scope of this paper, and will be explored in future works.

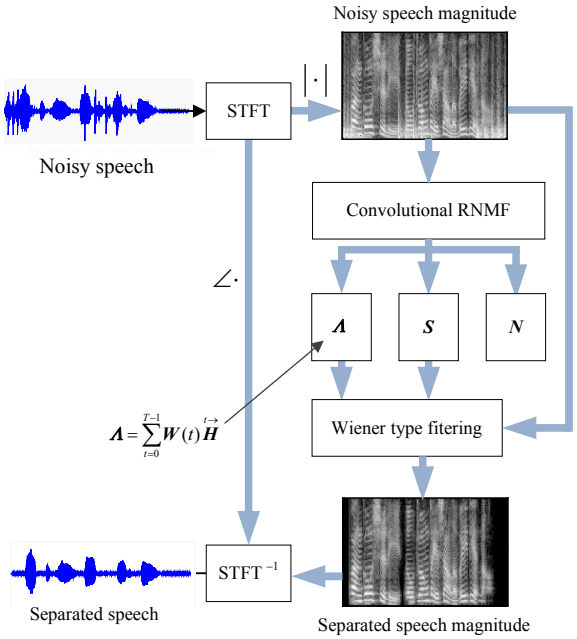


Fig. 1. Diagram of the proposed unsupervised framework for speech separation based on CRNMF. When the pattern length shrinks to one, CRNMF will reduce to the standard RNMF.

IV. Experiments

We conducted some experiments to evaluate the proposed algorithms in unsupervised speech separation. The noisy speech was synthesized by adding typical noises—including *f16*, *babble*, *machinegun* and *factory1*—to clean speech (ten sentences including five male and five female) at different input signal-to-noise ratios (SNRs), ranging from –5 dB to 10 dB in 5 dB steps. We resampled clean speech from the TIMIT database and noises from the Noisex-92 database to a sampling rate of 8 kHz. For the STFT, we used 512-sample Hamming windows with 128-point shifts. Preliminary experiments showed that, for most kind of noises, the optimal number of noise bases was one. Even though a larger number of bases would enable a more accurate description of the noise component, it would also result in higher speech distortion, because speech components could leak into the low-rank noise representation. Therefore, in the following experiments, the number of noise bases was always set to one. For details on automatic parameter setting, please refer to [27].

We also studied the selection of the regularization parameter λ, which controls the sparsity of $\boldsymbol{S}$. This parameter can also be seen as a control of the trade-off between speech distortion and noise reduction. A smaller value will result in optimal separation results when the input SNR is high, whereas for lower input SNRs, a larger value of λ tends to be more effective in extracting speech components while suppressing noise.

The separation performance was evaluated by computing the signal-to-distortion-ratio (SDR) values using BSS-EVAL [28]. To evaluate the performance of the different divergence measures, we also used the segment SNR (SegSNR) and perceptual evaluation of speech quality (PESQ) as performance metrics. In the experimental setting, we first compared the performance of RNMF with different cost functions, to show the necessity of using divergence—rather than Euclidean distance—as a metric. Subsequently, the separation performance of RNMF under different input SNRs and noise types was investigated, to determine which divergence measure is most suitable in the context of speech enhancement. Finally, the superiority of CRNMF when dealing with time-varying noise patterns is demonstrated through a comprehensive experiment.

1. Performance for Different Divergence Measures

Extensive experiments show that the generalized KL divergence achieves the best results (in terms of SDR) in speech separation tasks. In Fig. 2, we show two groups of experimental results, for clean speech degraded by machinegun noise (Fig. 2(a)) and f16 (Fig. 2(b)), at 5 dB input SNR.

As shown, the generalized KL divergence reaches the highest SDR value in both situations, and does so with a moderate choice of the regularization parameter. Using either IS divergence or Euclidean distance results in a significant performance loss when compared with KL divergence. In the machinegun noise case (top panel), IS divergence produced better results than Euclidean distance; the bottom panel, however, shows that a dramatically different behavior occurs in the f16 noise case. This is because the speech spectrogram contaminated by machinegun noise has a larger dynamic range than the spectrogram contaminated by f16 noise, and the IS cost function is less sensitive to large dynamic ranges than the Euclidean distance, because of its ratio term. As shown, the KL divergence can produce better results, as it provides relatively good compromises throughout the various noise types.

2. Speech Separation Performance of RNMF

To evaluate the speech separation performance of RNMF, we calculate the average SDR for all noise types at several input SNRs, ranging from –5 dB to 10 dB. The results obtained

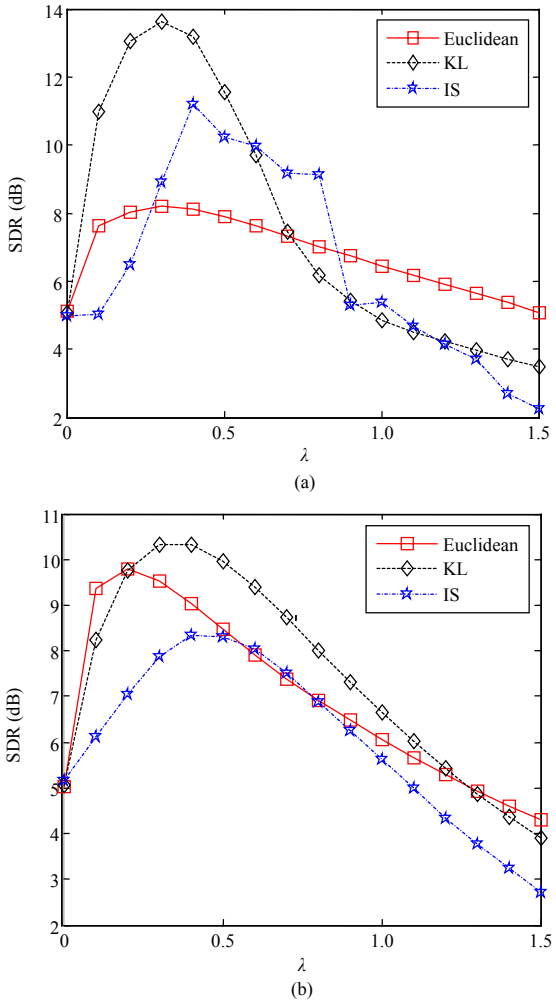


Fig. 2. SDR values of the separation results for the three cost functions: (a) signal degradation due to machinegun noise at 5 dB of input SNR and (b) signal degradation due to f16 noise, with the same input SNR.

with RPCA [7] are taken as a baseline.

As shown in Fig. 3, the proposed RNMF approach outperformed the recently proposed RPCA method, as a result of the non-negative constraints imposed on the component matrices. Once more, the KL divergence exhibited the best performance among the three cost functions, with 1.60 dB and 1.55 dB SDR improvements over the Euclidean distance and IS divergence, respectively. The IS divergence yielded results comparable with those of the Euclidean distance, as seen by their SDR curves in Fig. 3.

As mentioned before, we also calculated the SegSNR and PESQ improvements obtained with different divergence measures; the obtained results are presented in Table 1. In this table, we use ELD, KLD, and ISD to represent the Euclidean distance, generalized KL, and IS divergence, respectively. The NMF results (a supervised method) can be viewed as the upper bound of performance for the unsupervised methods, which

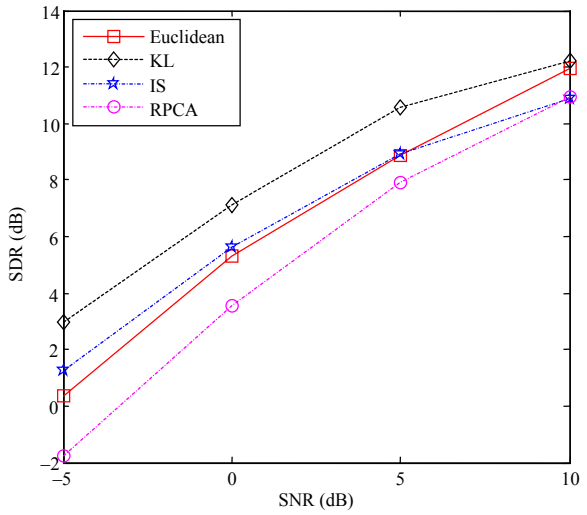


Fig. 3. Average SDR (over all noise types), at different SNR levels.

Table 1. SegSNR and PESQ improvements.

Metrics	SegSNR improv. (dB)				PESQ improv. (dB)			
SNR	–5	0	5	10	–5	0	5	10
NMF	4.8	4.5	3.9	3.6	0.31	0.40	0.44	0.47
RPCA	3.1	3.0	2.7	2.2	0.09	0.13	0.21	0.26
ELD	3.6	3.7	2.7	2.4	0.16	0.21	0.28	0.30
KLD	3.8	3.5	3.3	2.9	0.19	0.24	0.28	0.36
ISD	3.4	3.2	2.8	2.2	0.14	0.23	0.24	0.25

results from the fact that it uses prior knowledge not available to the unsupervised methods. From the results in Table 1, we can draw a conclusion similar to the ones obtained from the previous experiments: the generalized KL divergence seems to exhibit the best performance in the context of speech enhancement.

3. CRNMF and Time-Varying Noise Patterns

To evaluate the performance of the proposed convolutional RNMF, we conducted experiments on synthesized data, which was produced by first generating periodic background noise, and then adding it to clean speech, at 5 dB input SNR; a spectrogram of the resulting mixture is shown in Fig. 4(a). From this figure, we can see there is a large overlap between the noise and speech spectrograms. Moreover, the noise is time-varying, and repeats during the whole time interval. These properties make the unsupervised speech separation problem a challenging one. Given that the background noise is not exactly low-rank (it cannot be represented by a linear combination of only a few vectors), conventional methods based on sparse and low-rank decomposition—such as RPCA and RNMF—would fail to cope with this situation.

Considering the time continuity of the noise spectrogram,

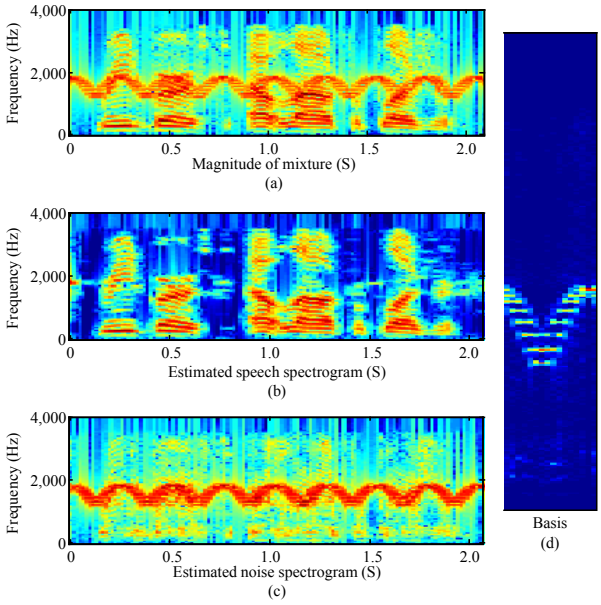


Fig. 4. CRNMF performance with time-varying background noise: (a) spectrogram of the mixture signal, in which a periodic time-varying noise signal (each period spans 16 frames of the spectrogram) is added to the speech signal, at 5 dB input SNR, (b) speech spectrogram, (c) noise spectrogram, and (d) non-negative basis learned by CRNMF.

CRNMF was applied to this mixture signal and evaluated; its source separation performance is shown in Figs. 4(b) and 4(c). Given the analysis results obtained in Sections IV.1 and IV.2, KL divergence was chosen as a cost function (rather than IS divergence or Euclidean distance).

The obtained experimental results show that the time span of the noise pattern (T in Section II.3) has some impact on the performance of convolutional RNMF, with the best performance being observed when T assumes the exact value of the noise pattern duration.

V. Conclusion

In this paper, we investigated the robust non-negative matrix factorization (RNMF) approach with β-divergence. By imposing non-negativity constrains, the proposed method outperformed the state-of-the-art unsupervised baseline method in speech separation tasks. Experimental results showed that the Kullback-Leibler divergence seems more suitable to be used as a cost function than both the Itakura-Saito divergence and the Euclidean distance. Moreover, we extended the idea of RNMF to a convolutional version, which is capable of describing the time-varying spectral characteristics of dynamic noises. Experiments showed the effectiveness of the proposed convolutional RNMF algorithm.

Appendix. Derivation of CRNMF with Generalized KL Divergence

Let us review the objective function of (15), and focus on the special case of generalized KL divergence. The objective function can be rewritten as:

$$\begin{aligned} C_{\mathrm{KL}} &= \left\| \boldsymbol{Y} \odot \ln \frac{\boldsymbol{Y}}{\boldsymbol{\Lambda}+\boldsymbol{S}} - \boldsymbol{Y} + \boldsymbol{\Lambda} + \boldsymbol{S} \right\|_F + \lambda \|\boldsymbol{S}\|_1 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{i,j} \ln \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} - Y_{i,j} + \Lambda_{i,j} + (1+\lambda)\cdot S_{i,j}, \end{aligned} \tag{A1}$$

where matrices are denoted with bold capital letters, and their respective elements by corresponding lowercase letters. Note that

$$\begin{aligned} \frac{\partial \Lambda_{i',j}}{\partial W_{i',k'}(t')} &= \frac{\partial}{\partial W_{i',k'}(t')} \sum_{t=0}^{T-1}\sum_{k=1}^{K} W_{i',k}(t) H_{k,j-t} \\ &= H_{k',j-t'} \end{aligned} \tag{A2}$$

and

$$\begin{aligned} \frac{\partial \Lambda_{i,j'}}{\partial H_{k',j'}} &= \frac{\partial}{\partial H_{k',j'}} \sum_{t=0}^{T-1}\sum_{k=1}^{K} W_{i,k}(t) H_{k,j-t} \\ &= W_{i,k'}(j-j'). \end{aligned} \tag{A3}$$

Therefore, the derivative of the objective function with respect to $W_{i',k'}(t')$, $H_{k',j'}$, and $S_{i',j'}$ can be calculated as follows:

$$\begin{aligned} \frac{\partial C_{\mathrm{KL}}}{\partial W_{i',k'}(t')} &= \frac{\partial}{\partial W_{i',k'}(t')} \sum_{i=1}^{I}\sum_{j=1}^{J} \left(Y_{i,j} \ln \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} - Y_{i,j} + \Lambda_{i,j} \right) \\ &= \sum_{j=1}^{J} \left(1 - \frac{Y_{i',j}}{\Lambda_{i',j}+S_{i',j}} \right) \frac{\partial \Lambda_{i',j}}{\partial W_{i',k'}(t')} \\ &= \sum_{j=1}^{J} \left(1 - \frac{Y_{i',j}}{\Lambda_{i',j}+S_{i',j}} \right) H_{k',j-t'}, \end{aligned} \tag{A4}$$

$$\begin{aligned} \frac{\partial C_{\mathrm{KL}}}{\partial S_{i',j'}} &= \frac{\partial}{\partial S_{i',j'}} \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{i,j} \ln \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} + (\lambda+1)\cdot S_{i,j} \\ &= \lambda + 1 - \frac{Y_{i',j'}}{S_{i',j'}+\Lambda_{i',j'}}, \end{aligned} \tag{A5}$$

and

$$\begin{aligned} \frac{\partial C_{\mathrm{KL}}}{\partial H_{k',j'}} &= \frac{\partial}{\partial H_{k',j'}} \sum_{i=1}^{I}\sum_{j=1}^{J} \left(Y_{i,j} \ln \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} - Y_{i,j} + \Lambda_{i,j} \right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \left(1 - \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} \right) \frac{\partial \Lambda_{i,j}}{\partial H_{k',j'}} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \left(1 - \frac{Y_{i,j}}{\Lambda_{i,j}+S_{i,j}} \right) W_{i,k'}(j-j') \\ &= \sum_{i=1}^{I}\sum_{t=0}^{T-1} \left(1 - \frac{Y_{i,j'+t}}{\Lambda_{i,j'+t}+S_{i,j'+t}} \right) W_{i,k'}(t). \end{aligned} \tag{A6}$$

Equations (A4) to (A6) can be split into positive and negative parts, and the following update rules can hence be obtained:

$$W_{i',k'}(t') \leftarrow W_{i',k'}(t') \frac{\sum_{j=1}^{J} \frac{Y_{i',j}}{\Lambda_{i',j}+S_{i',j}} H_{k',j-t'}}{\sum_{j=1}^{J} H_{k',j-t'}}, \tag{A7}$$

$$H_{k',j'} \leftarrow H_{k',j'} \frac{\sum_{i=1}^{I}\sum_{t=0}^{T-1} \frac{Y_{i,j'+t}}{\Lambda_{i,j'+t}+S_{i,j'+t}} W_{i,k'}(t)}{\sum_{i=1}^{I}\sum_{t=1}^{T-1} W_{i,k'}(t)}, \tag{A8}$$

and

$$S_{i',j'} \leftarrow S_{i',j'} \frac{Y_{i',j'}}{(\lambda+1)\cdot\left(S_{i',j'}+\Lambda_{i',j'}\right)}. \tag{A9}$$

These equations can be rewritten in matrix form as:

$$\boldsymbol{W}(t) \leftarrow \boldsymbol{W}(t) \odot \frac{\frac{\boldsymbol{Y}}{\boldsymbol{\Lambda}+\boldsymbol{S}} \overset{t\rightarrow}{\boldsymbol{H}}^{\mathrm{T}}}{\mathbf{1}\cdot \overset{t\rightarrow}{\boldsymbol{H}}^{\mathrm{T}}}, \tag{A10}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}(t)^{\mathrm{T}} \left[\frac{\overset{\leftarrow t}{\boldsymbol{Y}}}{\boldsymbol{\Lambda}+\boldsymbol{S}} \right]}{\boldsymbol{W}(t)^{\mathrm{T}}\cdot\mathbf{1}}, \tag{A11}$$

and

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{Y}}{(\lambda+1)\cdot(\boldsymbol{S}+\boldsymbol{\Lambda})}. \tag{A12}$$

Note that (16) to (18) are the same as (A10) to (A12) when $\beta = 1$. The results for the Euclidean distance and Itakura-Saito divergence can be similarly derived.

References

[1] P. Smaragdis et al., "Static and Dynamic Source Separation Using Nonnegative Matrix Factorizations: a Unified View," *IEEE Signal Process. Mag.*, vol. 31, no. 3, May 2014, pp. 66–75.

[2] T. Virtanen, "Monaural Sound Source Separation by Non-negative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, Mar. 2007, pp. 1066–1074.

[3] Z. Duan, GJ. Mysore, and P. Smaragdis, "Online PLCA for Real-Time Semi-supervised Source Separation," *Proc. Latent Variable Anal. Signal Separation*, Tel Aviv, Israel, Mar. 12–15, 2012, pp. 34–41.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Non-negative Matrix Factorization," *IEEE Trans. Audio, Speech, Language Process.*,

vol. 21, no. 10, Oct. 2013, pp. 2140–2151.

[5] K.W. Wilson, B. Ray, and P. Smaragdis, "Regularized Non-negative Matrix Factorization with Temporal Dependencies for Speech Denoising," *Proc. Interspeech*, Jan. 2008, pp. 411–414.

[6] E.J. Candès et al., "Robust Principle Component Analysis?," *J. ACM*, vol. 58, no. 3, 2011, pp. 11:1–11:37.

[7] P. Huang et al., "Sing-Voice Separation from Monaural Recording Using Robust Principal Component Analysis," *Proc. IEEE Conf. Acoustics, Speech, Signal, Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 57–60.

[8] Z. Chen and D.P.W. Eills, "Speech Enhancement by Sparse, Low-Rank, and Dictionary Spectrogram Decomposition," *Proc. Workshop Appli. Signal Process. Audio Acoustics*, New Paltz, NY, USA, Oct. 20–23, 2013, pp. 1–4.

[9] C. Sun, Q. Zhu, and M. Wan, "A Novel Speech Enhancement Method Based on Constrained Low-Rank and Sparse Matrix Decomposition," *Speech Commun*., vol. 60, May 2014, pp. 44–55.

[10] L. Zhang et al., "Robust Non-negative Matrix Factorization," *Frontiers Electric. Electron. Eng. China*, vol. 6, no. 2, June 2011, pp. 192–200.

[11] T. Virtanen et al., "Compositional Models for Audio Processing: Uncovering the Structure of Sound Mixtures," *IEEE Signal Process. Mag*., vol. 32, no. 2, Mar. 2015, pp. 125–144.

[12] J.J. Carabias-Orti et al., "Constrained Non-negative Sparse Coding Using Learnt Instrument Templates for Real Time Music Transcription," *Eng. Appli. Artificial Intell.*, vol. 26, no. 7, Aug. 2013, pp. 1671–1680.

[13] C. Févotte and J. Idier, "Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence," *Neural Comput.*, vol. 23, no. 9, Sept. 2011, pp. 2421–2456.

[14] D.D. Lee and H.S. Seung, "Learning the Parts of Objects with Nonnegative Matrix Factorization," *Nature*, vol. 401, Oct. 1999, pp. 788–791.

[15] H. Li, Y. Shen, and J. Wnag, "An Improved Multiplicative Updating Algorithm for Non-negative Independent Component Analysis," *ETRI J.*, vol. 35, no. 2, Apr. 2013, pp. 193–199.

[16] M. Sun and H. Van Hamme, "Large Scale Graph Regularized Non-negative Matrix Factorization with l_1 Normalization Based on Kullback-Leibler Divergence," *IEEE Trans. Signal Process.*, vol. 60, no. 7, July 2012, pp. 3876–3880.

[17] V.Y.F. Tan, "Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, July 2013, pp. 1592–1605.

[18] P. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learn. Res.*, vol. 5, 2004, pp. 1457–1469.

[19] W. Wang, A. Cichocki, and J.A. Chamners, "A Multiplicative Algorithm for Convolutive Non-negative Matrix Factorization Based on Squared Euclidean Distance," *IEEE Trans. Signal Process.*, vol. 57, no. 7, July 2009, pp. 2858–2864.

[20] P. Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, Jan. 2007, pp. 1–12.

[21] D. Wang et al., "Online Non-negative Convolutive Pattern Learning for Speech Signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 61, no. 1, Jan. 2013, pp. 44–56.

[22] J. Huang et al., "Speech Denoising Via Low-Rank and Sparse Matrix Decomposition," *ETRI J.*, vol. 36, no. 1, Feb. 2014, pp. 167–170.

[23] F.G Germain and G.J. Mysore, "Speaker and Noise Independent Online Single-Channel Speech enhancement," *Proc. IEEE Conf. Acoustics Speech Signal Process.*, Queensland, Australia, Apr. 19–24, 2015, pp. 71–75.

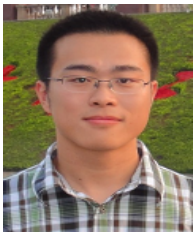
[24] Y. Li et al., "Speech Enhancement Using Non-negative Matrix Low-Rank Modeling with Temporal Continuity and Sparseness Constraints," *Proc. PCM*, 2016, accepted.

[25] M. Sun et al., "Speech Enhancement Under Low SNR Conditions Via Noise Estimation Using Sparse and Low-Rank NMF with Kullback-Leibler Divergence," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 7, July 2015, pp. 1233–1242.

[26] Y. Li et al., "Adaptive Extraction of Repeating Non-negative Temporal Patterns for Single Channel Speech Enhancement," *Proc. IEEE Conf. Acoustics Speech Signal Process.*, Shanghai, China, Mar. 20–25, 2016, pp. 494–498.

[27] Y. Li, et al., "Automatic Model Order Selection for Convolutive Non-negative Matrix Factorization," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E99-A, no. 10, 2016, pp. 1867–1870.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, July 2006, pp. 1462–1469.



Yinan Li received his MS degree from the Lab of Intelligent Information Processing of the PLA University of Science and Technology, Nanjing, China. He is pursuing a PhD degree in the same university. He is currently an international scholar with the Department of Electrical Engineering, Katholieke Universiteit, Leuven, Belgium, with the support of the China Scholarship Council. His research interests lie in the areas of speech enhancement and sparse representations of audio signals.

Xiongwei Zhang received his PhD degree from the PLA University of Science and Technology, Nanjing, China. He is now a professor at the Lab of Intelligent Information Processing in the same university. His research interests lie in the areas of speech coding, speech enhancement, and image processing.

Meng Sun received his PhD degree from the Department of Electrical Engineering (ESAT), Katholieke Universiteit, Leuven, Belgium in November 2012. He is now a researcher at the Lab of Intelligent Information Processing of the PLA University of Science and Technology, Nanjing, China. His research interests lie in the areas of speech processing, unsupervised/semi-supervised machine learning, and sequential pattern recognition.