EAI Endorsed Transactions
on Context-aware Systems and Applications
Research Article EAI.EU

Exploiting Nonnegative Matrix Factorization with Mixed Group Sparsity Constraint to Separate Speech Signal from Single-channel Mixture with Unknown Ambient Noise

Thanh Thi Hien Duong[1,2,*], Phuong Cong Nguyen[1,3], Cuong Quoc Nguyen[3]

[1]International Research Institute MICA, Hanoi University of Science and Technology, Vietnam
[2]Dept. of Information Technology, Hanoi University of Mining and Geology, Vietnam
[3]Dept. of Instrumentation and Industrial Informatic, Hanoi University of Science and Technology, Vietnam

Abstract

This paper focuses on solving a challenging speech enhancement problem: improving the desired speech from a single-channel audio signal containing high-level unspecified noise (possibly environmental noise, music, other sounds, etc.). Using source separation technique, we investigate a solution combining nonnegative matrix factorization (NMF) with mixed group sparsity constraint that allows exploiting generic noise spectral model to guide the separation process. The experiment performed on a set of benchmarked audio signals with different types of real-world noise shows that the proposed algorithm yields better quantitative results in term of the signal-to-distortion ratio than the previously published algorithms.

Received on 30 December 2017; accepted on 24 February 2018; published on 14 March 2018

Keywords: Speech enhancement, source separation, nonnegative matrix factorizarion (NMF), sparsity constraint, generic source spectral model

doi:10.4108/eai.14-3-2018.154342

1. Introduction

Speech enhancement is a process of removing unexpected audio signals (noise) from their mixture with a desired speech signal. This subject has been widely studied for decades as it brings huge impact in many different domains such as communication, speech-based control systems, medical surveillance, audio post-processing in movies and entertainments, etc., [1]. Recent scientific research [2–4] has shown that the performance of speech recognition systems in practical noisy and reverberant environments degraded dramatically. This situation demonstrates the need for improving speech quality in such noisy recordings. Popular approaches for speech enhancement includes beamforming [5, 6], spectral subtraction [7], and source separation [8–10].

Considering speech and noise as two independent sources to be separated, audio source separation technique can be used to isolate the desired speech from high level noise. Some recent work has developed methods for single-channel speech enhancement based on *e.g.,* NMF [11, 12], Gaussian mixture model (GMM) [13], or deep neural network [14, 15]. The two former methods first learn the characteristics of speech and noise signals, then such learned models were used to guide the signal separation process. The deep learning based approaches can learn the separation mask or the separation model by end-to-end training and gain a significant impact. However deep learning based systems require a lot of training data and processing power. For cases with only few training examples available, the work of Sun and Mysore [16] proposed the use of NMF [17] to establish the general spectral model for speech signals from some other voices. Studies of El Badawy et al. [18–20] employed the similar NMF-based spectral models learned from source examples obtained by a search engine to guide the separation algorithm.

In this paper, we focus on a slightly different setting compared to the existing works [16–18], where the speaker is assumed to be known but the noise signal is non deterministic. This speaker-dependent situation is very popular in practice.

*Corresponding author. Email: duongthihienthanh@humg.edu.vn

For instance, when speech is used to control robots or devices, the operator/speaker is often known so that his/her voice can be collected in advance for training the system. Concerning noise, it is highly non-stationary and if the operating environment is changed (different moments or different locations), it will vary accordingly. Therefore, noise should not be well-identified in the training process. From this intuition, we propose a novel approach that first constructs the general spectral noise model from some noise examples in advance. Such noise examples can beeasily pre-collected in some environments. Then the general noise model is used to guide the separation process. Within the considered NMF based approach, we investigate the combination of the existing *block sparsity* proposed in [16] and *component sparsity* proposed in [18] in order to improve the source separation performance. Developing further from our preliminary studies [21, 22], this paper presents more detail about the algorithm and extends the experiments using large test database containing various types of noise signals to confirm the effectiveness of the proposed approach. Furthermore, we report the investigation of the algorithm's convergence and stability.

The paper is organized into five sections. We first summarize the baseline audio source separation algorithm using the NMF model in Section 2. We then present the proposed approach in Section 3. Section 4 discusses experiment settings, algorithm analysis, and speech enhancement results. Finally, we conclude in Section 5.

2. Baseline Supervised NMF-based Speech Separation Method

To extract the desired speech signal from the single-channel noisy signal (referred to as *mixture*), we consider the mixture as a signal which created by mixing two audio sources: the desired speech and the noise. Noise can be environmental noise and any other unwanted sounds.
In general, the source separation processing is done in the time-frequency domain after the short-time Fourier transform (STFT) so that the 1D waveform is represented by the 2D spectrogram. Then this 2D spectrogram is modeled by the NMF, which is a widely used model in audio signal processing in general and in audio separation in particular [23–25].

Let $\mathbf{X} \in \mathbb{C}^{F\times M}$, $\mathbf{Y} \in \mathbb{C}^{F\times M}$, and $\mathbf{Z} \in \mathbb{C}^{F\times M}$ are the complex-valued matrices of the short-time Fourier transform (STFT) coefficients of the observed mixture signal, the speech signal, and the noise signal, respectively, where F is the number of frequency bins, M is the number of time frames, then the mixing model writes:

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}. \tag{1}$$

Denoting by $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ the power spectral matrix of the mixture signal, where $\mathbf{X}^{\cdot n}$ is the matrix whose elements are $[\mathbf{X}]_{il}^{n}$, NMF decomposes $\mathbf{V} \in \mathbb{C}_{+}^{F\times M}$ into two nonnegative matrices as

$$\mathbf{V} \approx \mathbf{B} * \mathbf{A}, \tag{2}$$

where $*$ is the normal matrix multiplication, $\mathbf{B} \in \mathbb{R}_{+}^{F\times K}$ is the spectral basis matrix whose column vectors are spectral characteristics appearing in $\mathbf{V}$, $\mathbf{A} \in \mathbb{R}_{+}^{K\times M}$ is the activation matrix whose row vectors are times of appearance of spectral components in $\mathbf{B}$, K is the number of spectral components to be synthesized. Depending on the applications and properties of input data, K is usually chosen such that $\mathbf{B}$ is able to represent most spectral characteristics of the input signal [26]. To estimate the latent matrices, $\mathbf{B}$ and $\mathbf{A}$ are initialized with random non-negative values and are updated in an iterative process such that the cost function (3) representing the divergence between $\mathbf{V}$ and $\mathbf{B} * \mathbf{A}$ is minimized:

$$D(\mathbf{V}||\mathbf{B} * \mathbf{A}) = \sum_{f=1}^{F}\sum_{m=1}^{M} d_{IS}(\mathbf{V}_{fm}||[\mathbf{B} * \mathbf{A}]_{fm}), \tag{3}$$

where f and m denote frequency bin index and time frame index, respectively, and

$$d_{IS}(x||y) = \frac{x}{y} - \log(\frac{x}{y}) - 1 \tag{4}$$

is the Itakura-Saito divergence. This divergence is commonly used in ausio source separation as it offers the scale-invariant property. In each iteration steps, $\mathbf{B}$ and $\mathbf{A}$ are updated via the well-known multiplicative update (MU) rules [26] as

$$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\left((\mathbf{B} * \mathbf{A})^{\cdot -2} \odot \mathbf{V}\right)\mathbf{A}^T}{(\mathbf{B} * \mathbf{A})^{\cdot -1}\mathbf{A}^T} \tag{5}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\mathbf{B}^T\left((\mathbf{B} * \mathbf{A})^{\cdot -2} \odot \mathbf{V}\right)}{\mathbf{B}^T(\mathbf{B} * \mathbf{A})^{\cdot -1}} \tag{6}$$

in which $\mathbf{C}^T$ is the transposition of matrix $\mathbf{C}$, $\odot$ denotes the element-wise Hadamard product, the power and the division is also element-wise.

Suppose that $\mathbf{B}_Y$ and $\mathbf{B}_Z$ are spectral basis matrices of speech and noise, respectively. In training process of the supervised approach, they are learned from the corresponding training examples by optimizing similar criterion as (11), then the spectral model for two sources $\mathbf{B}$ is obtained by

$$\mathbf{B} = [\mathbf{B}_Y, \mathbf{B}_Z]. \tag{7}$$

In the speech enhancement process, this spectral model $\mathbf{B}$ is fixed, and the time activation matrix $\mathbf{A}$ is estimated via the MU rule by iterating (5) and (6). Note that $\mathbf{A}$ also consists of two blocks as $\mathbf{A}_Y$ and $\mathbf{A}_Z$, which are block characterizing the time activations for speech and noise, respectively, as

$$\mathbf{A} = [\mathbf{A}_Y^T, \mathbf{A}_Z^T]^T. \tag{8}$$

After the parameters $\mathbf{B}$ and $\mathbf{A}$ are obtained, the speech STFT coefficients are determined by Wiener filtering as the following

$$\hat{\mathbf{Y}} = \frac{\mathbf{B}_Y * \mathbf{A}_Y}{\mathbf{B} * \mathbf{A}} \odot \mathbf{X}. \quad (9)$$

Finally, the estimated speech signal in time domain is obtained via the inverse STFT.

3. Proposed Method

In the unspecified noise scenario, clean speech example from a desired speaker is assumed to be available a priori for training but exact noise example is not available. However, some general noise examples could be collected easily from different noisy environments for training also. For example, in order to separate speech and environmental noise, we collect some environmental sounds such as wind sound, street noise, cafeteria, etc., for noise training. The global workflow of the proposed approach for speech separation is shown in Fig. 1. In the following, we first present the training for both speech spectral model $\mathbf{B}_Y$ and generic spectral noise model $\widetilde{\mathbf{B}}_Z$ in Section 3.1. We then describe the model fitting with the proposed mixed group sparsity constraint for the source separation process in Section 3.2

3.1. Training Spectral Models for Speech and Noise

(i) Speech spectral model
Let $\mathbf{V}_Y = |\mathbf{Y}|^{\cdot 2}$ is the spectrogram of a clean speech example obtained by the STFT transform. Speech spectral model $\mathbf{B}_Y$ is learned given $\mathbf{V}_Y$ by optimizing the divegence between $\mathbf{V}_Y$ and $\mathbf{B}_Y * \mathbf{A}_Y$ as

$$\min_{\mathbf{B}_Y \geq 0, \mathbf{A}_Y \geq 0} D(\mathbf{V}_Y \| \mathbf{B}_Y * \mathbf{A}_Y), \quad (10)$$

where $\mathbf{A}_Y$ is the time activation matrix.

(ii) Generic noise spectral model
Let P is the number of collected noise examples for training, $\mathbf{V}_Z^p = |\mathbf{Z}^p|^{\cdot 2}, 1 \leq p \leq P$ is the spectrogram of $p-th$ noise example. Firstly, $\mathbf{V}_Z^p$ is used to learn the NMF spectral model, denoting by $\mathbf{B}_Z^p$, by minimizing the criterion:

$$\min_{\mathbf{B}_Z^p \geq 0, \mathbf{A}_Z^p \geq 0} D(\mathbf{V}_Z^p \| \mathbf{B}_Z^p * \mathbf{A}_Z^p), \quad (11)$$

where $\mathbf{A}_Z^p$ is the time activation matrix.
After all spectral model $\mathbf{B}_Z^p, p = 1, \ldots, P$, are learned from noise examples, the generic noise spectral model, denoted by $\widetilde{\mathbf{B}}_Z$, is constructed as the following

$$\widetilde{\mathbf{B}}_Z = [\mathbf{B}_Z^1, \ldots, \mathbf{B}_Z^P]. \quad (12)$$

(iii) Spectral model for all sources
The spectral model for all speech and noise is computed by

$$\widetilde{\mathbf{B}} = [\mathbf{B}_Y, \widetilde{\mathbf{B}}_Z]. \quad (13)$$

In the speech enhancement phase, this spectral model $\widetilde{\mathbf{B}}$ is fixed, and the time activation matrix $\widetilde{\mathbf{A}}$ is estimated via the MU rule. Matrix $\widetilde{\mathbf{A}}$ includes the speech activation matrix $\mathbf{A}_Y$ and noise activation matrix $\widetilde{\mathbf{A}}_Z$ as

$$\widetilde{\mathbf{A}} = [\mathbf{A}_Y^T, \widetilde{\mathbf{A}}_Z^T]. \quad (14)$$

3.2. Proposed Mixed Group Sparsity-inducing penalty for Noise model fitting

The generic spectral model for noise $\widetilde{\mathbf{B}}_Z$ become a larger matrix when the number of noise examples P increases. Moreover, it is actually redundant when different examples share the similar spectral patterns [27–29]. Thus, in the NMF model fitting for the signal separation, sparsity constraint is naturally needed so as to fit only a subset of the large matrix $\widetilde{\mathbf{B}}_Z$ to the actual noise representing in the mixture [28]. In other words, the mixture spectrogram V is decomposed by solving the following optimization problem

$$\min_{\mathbf{A} \geq 0} D(\mathbf{V} \| \widetilde{\mathbf{B}} * \widetilde{\mathbf{A}}) + \lambda \Omega(\widetilde{\mathbf{A}}_Z) \quad (15)$$

where $\Omega(\widetilde{\mathbf{A}}_Z)$ denotes a penalty function imposing sparsity on the activation matrix $\widetilde{\mathbf{A}}_Z$, λ is a trade-off parameter determining the contribution of the penalty.
Recent work in audio source separation has considered two penalty functions. The first one is block sparsity-inducing penalty [16] formulated as the following

$$\Omega_1(\widetilde{\mathbf{A}}_Z) = \sum_{p=1}^{P} \log(\epsilon + \|\widetilde{\mathbf{A}}_Z^{(p)}\|_1), \quad (16)$$

where ϵ is a non-zero constant, $\widetilde{\mathbf{A}}_Z^{(p)}$ is a subset of $\widetilde{\mathbf{A}}_Z$ representing the activation coefficients for $p-th$ block, $\|.\|_1$ is ℓ_1-norm operator, and P denotes the total number of blocks. In this case, a block represents one training example and P is the total number of noise examples. This penalty enforces the activation for relevant examples only while omitting the poorly fitting examples since their corresponding activation block will likely converge to zero.
The second one is named component sparsity-inducing penalty [18] formulated as

$$\Omega_2(\widetilde{\mathbf{A}}_Z) = \sum_{k=1}^{K} \log(\epsilon + \|\widetilde{\mathbf{a}}_Z^{(k)}\|_1), \quad (17)$$

where $\widetilde{\mathbf{a}}_Z^{(k)}$ denotes $k-th$ row of $\widetilde{\mathbf{A}}_Z$. This penalty is motivated by the fact that only a part of the spectral model learned from an example may fit well with the targeted

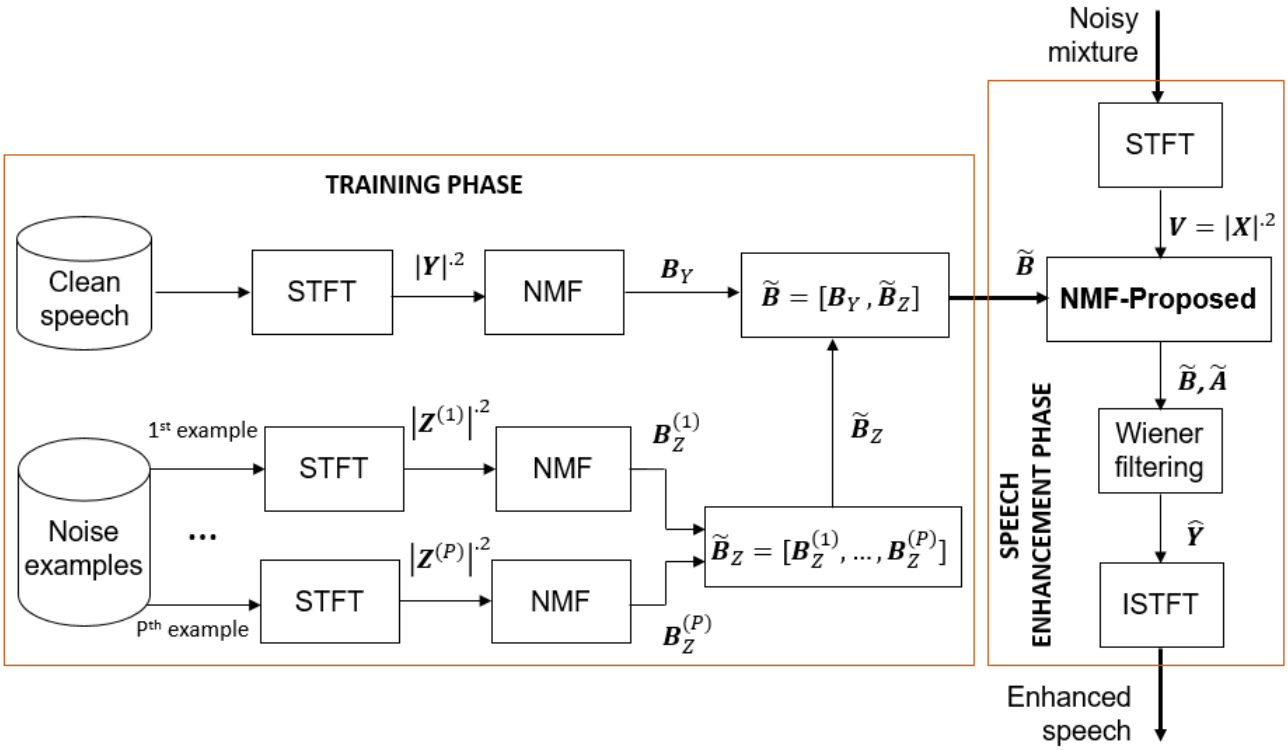

Figure 1. General workflow of the proposed speech enhancement approach.

source in the mixture, while the remaining components in the model do not. Thus instead of activating the whole block, this penalty allows selecting only the more likely relevant spectral components from $\widetilde{\mathbf{B}}_Z$.

However, the component sparsity-inducing penalty also quite slowly removes unsuitable parts, because it carefully considers each row in the large matrix. Inspired by the advantage of these two state-of-the-art penalty functions, in our recent works [21, 22], we proposed to combine them in a more general form as

$$\Omega(\widetilde{\mathbf{A}}_Z) = \alpha \sum_{p=1}^{P} \log(\epsilon + \|\widetilde{\mathbf{A}}_Z^{(p)}\|_1) + (1-\alpha) \sum_{k=1}^{K} \log(\epsilon + \|\widetilde{\mathbf{a}}_Z^{(k)}\|_1), \tag{18}$$

where the first term on the right hand side of the equation presents the block sparsity-inducing penalty, the second term presents the component sparsity-inducing penalty, and $\alpha \in [0, 1]$ weights the contribution of each term. Proposed penalty function (18) can be seen as the generalization of (16) and (17) in the sense that when $\alpha = 1$, (18) is equivalent to (16) and when $\alpha = 0$, (18) is equivalent to (17).
In order to derive the parameter estimation algorithm optimizing (15) with the proposed penalty function (18), one can rely on MU rules and the majorization-minimization algorithm. The proposed algorithm is summarized in Algorithm 1, where $\mathbf{E}_{(p)}$ is a uniform matrix of the same size as $\widetilde{\mathbf{A}}_Z^{(p)}$, and $\mathbf{g}_{(k)}$ a uniform row vector of the same size as $\widetilde{\mathbf{a}}_Z^{(k)}$.

4. Experiment

We start by describing the data set and parameter settings in Section 4.1. We then describe evaluation metrics in Section 4.2. The performance of the proposed speech enhancement algorithm and its sensitivity with respect to the choice of the hyper parameters are presented in Section 4.3.

4.1. Dataset and parameter settings

To validate the performance of the proposed approach, we select noise examples from DEMAND[1] dataset for training the generic noise spectral model, and perform the test on the benchmarked dataset from SISEC campaign[2]. These datasets were carefully designed by researchers in the audio source separation community and widely used.

Training speech example is five-second long and is spoken by the same person with speech in the tested mixtures. We use five types of environmental noise: kitchen sound, waterfall, metro, field sound, cafeteria to train the generic

[1] http://parole.loria.fr/DEMAND/
[2] http://sisec.wiki.irisa.fr.

Algorithm 1 Proposed NMF with mixed group sparsity constraint algorithm

Require: $\mathbf{V}$, $\widetilde{\mathbf{B}}$, λ, α
Ensure: $\widetilde{\mathbf{A}}$
  Initialize $\widetilde{\mathbf{A}}$ randomly with nonnegative values
  $\hat{\mathbf{V}} = \widetilde{\mathbf{B}} * \widetilde{\mathbf{A}}$
  repeat
    // Taking into account block sparsity-inducing penalty
    for $p = 1, ..., P$ **do**
      $\mathbf{E}_{(p)} \leftarrow \frac{1}{\epsilon + \|\widetilde{\mathbf{A}}_Z^{(p)}\|_1}$
    end for
    $\mathbf{E} = [\mathbf{E}_{(1)}^T, \ldots, \mathbf{E}_{(P)}^T]^T$
    // Taking into account component sparsity-inducing penalty
    for $k = 1, ..., K$ **do**
      $\mathbf{g}_k \leftarrow \frac{1}{\epsilon + \|\widetilde{\mathbf{a}}_Z^{(k)}\|}$
    end for
    $\mathbf{G} = [\mathbf{g}_1^T, \ldots, \mathbf{g}_K^T]^T$
    // Updating activation matrices
    $\mathbf{A}_Y \leftarrow \mathbf{A}_Y \odot \frac{\mathbf{B}_Y^T * (\hat{\mathbf{V}}^{\cdot -2} \odot \mathbf{V})}{\mathbf{B}_Y^T * (\hat{\mathbf{V}}^{\cdot -1})}$
    $\widetilde{\mathbf{A}}_Z \leftarrow \widetilde{\mathbf{A}}_Z \odot \left(\frac{\widetilde{\mathbf{B}}_Z^T * (\hat{\mathbf{V}}^{\cdot -2} \odot \mathbf{V})}{\widetilde{\mathbf{B}}_Z^T * (\hat{\mathbf{V}}^{\cdot -1}) + \lambda(\alpha \mathbf{E} + (1-\alpha)\mathbf{G})} \right)^{\cdot \frac{1}{2}}$
    //Updating $\hat{\mathbf{V}}$
    $\hat{\mathbf{V}} = \widetilde{\mathbf{B}} * \widetilde{\mathbf{A}}$
  until convergence

noise spectral model (see Section 3.1). They are extracted from DEMAND with duration varying from 5 to 15 seconds.

The performance of the proposed algorithm was evaluated over a test set containing 15 single-channel mixtures of two sources artificially mixed at 0 dB signal to noise ratio (SNR). Note that with this 15 mixtures with various types of noise could be sufficient to access the performance of the proposed algorithm. During the mixing process, we made sure that in all mixtures both sources appear all the time. The mixtures were sampled at 16000 Hz and their duration varies between 5 and 10 seconds. The speech samples include female speech and male speech in English, they were obtained from SiSEC data set. The noise samples were obtained from DEMAND from one channel out of the 16 channels. Some of them were mixed two noises, *e.g.,* traffic + wind sound, ocean waves + birdsong, restaurant + guitar, forest birds + car, square + music, ect.,.

The parameters were set as follow. The STFT was calculated using a sliding window with a frame length of 1024, 50% overlap. The number of NMF components were set to 32 and 16 for speech and noise, respectively. The number of iterations for MU updates was 100 for the training step and was tested with values from 1 to 100 in the testing in order to investigate the convergence of the algorithm.

To consider the sensitivity of the proposed algorithm to the trade-off parameter λ determining the contribution of the sparsity-inducing penalty and the contribution weighting of each penalty term α, we varied the values of these parameters as $\lambda = \{1, 10, 25, 50, 100, 200, 500\}$, $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

4.2. Evaluation method

We compare the separation performance obtained by proposed algorithm with several state-of-the-art algorithms as follows:

- Baseline NMF - without training: The NMF-based algorithm was described in Section 2. This test did not use training data, instead, the spectral models for both speech and noise were initialized with random non-negative values and were iteratively updated via (5) and (6).
- Baseline NMF - speech training: The algorithm based on NMF was described in Section 2. In this experiment, the spectral model for speech signal was learned by speech examples that were five-second long and were made by the same person with the speech in the tested mixtures. The spectral model for noise was initialized with random non-negative values and was iteratively updated via (5) and (6).
- NMF non-sparsity: The algorithm based on NMF was described in Section 2. The spectral model for speech was also learned by five-second long that was spoken by the same person with the speech in the tested mixtures. The noise spectral model was learned by one noisy file which was made by pairing five noise samples in the noise training set described in Section 4.1.
- NMF - Block sparsity: Proposed framework, combining NMF with block sparsity constraint by (16) [16].
- NMF - Component sparsity: Proposed framework, combining NMF with component sparsity constraint by (17) [18].

Separated speech results were evaluated using the source-to-distortion ratio (SDR) measuring overall distortion as well as the source-to-interference ratio (SIR) and the source-to-artifacts ratio (SAR). They were measured in dB and averaged over all sources where the higher is the better. These criteria, known as BSS-EVAL metrics, have been mostly used in the source separation community [30].

4.3. Results and Discussion

The results are averaged over all 15 testing mixtures for six different algorithms and indicated in Table 1. Figure 2 shows the convergence of the proposed algorithm as a function of the number of MU iterations. Performance of the algorithm

Algorithms	SDR (dB)	SIR (dB)	SAR (dB)
Baseline NMF - without training	-0.5	2.4	6.4
Baseline NMF - speech training	5.8	9.6	10.4
NMF non-sparsity	3.5	4.5	10.1
NMF - Block sparsity [16] ($\lambda = 1, \alpha = 1$)	7.9	12.7	11.4
NMF - Component sparsity [18] ($\lambda = 1, \alpha = 0$)	8.0	13.1	11.4
Proposed NMF - Mixed sparsity ($\lambda = 1, \alpha = 0.2$)	**8.3**	**13.3**	**11.5**

Table 1. Average performance of speech enhancement obtained on the test set.

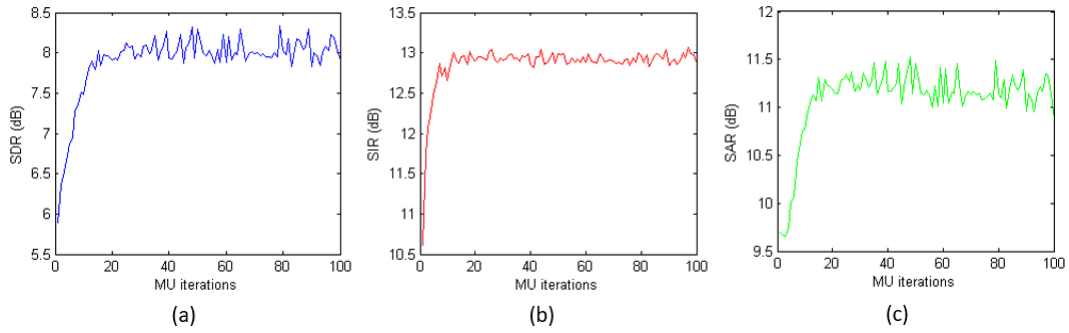

Figure 2. Speech enhancement performance of the proposed method as a function of MU iterations.

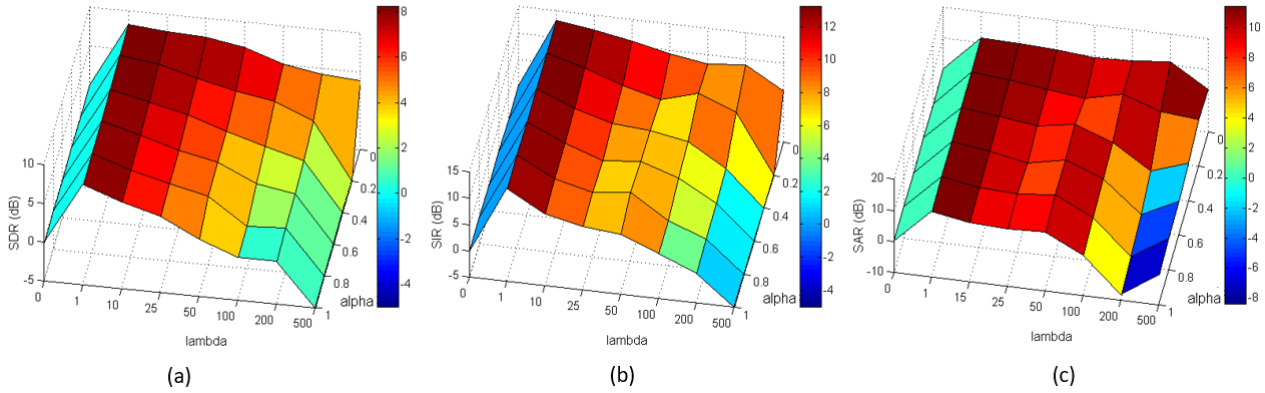

Figure 3. Average speech enhancement performance of the proposed method as a function of λ and α.

as a function of the parameters λ and α is shown in Figure 3.

It is interesting to see in Table 1 that the results obtained by the "NMF non-sparsity" method even were lower than the results of "Baseline NMF - speech training" method. It reveals that the generic noise spectral model itself is redundant and contains some irrelevant spectral patterns with the actual noise in the mixture. Thus the importance of such sparsity penalty is explicitly confirmed by the fact that the results obtained by three algorithms based on the NMF with group sparsity-inducing penalties were far more better than the remaining three algorithms. It is also not surprising to see that the baseline NMF method yielded quite good results when using training data for speech signal (*i.e.* "Baseline NMF - speech training" method gained 5.8 dB SDR), but without training data, the result is very low (*i.e.* "Baseline NMF - without training" method gained -0.5 dB SDR). Finally, the "Proposed NMF - Mixed sparsity" algorithm offers the best speech enhancement performance in terms of all SDR, SIR, and SAR compared to the five existing ones. More specifically, compared to two algorithms based on the NMF with group sparsity-inducing penalties, the proposed NMF - Mixed sparsity method gained 0.4 dB and 0.3 dB SDR higher than those of the "NMF - Block sparsity" method and the "NMF - Component sparsity" method, respectively. The proposed method's results were also far better than results of three first methods. This proves the effectiveness of the successful combination of two state-of-the-art group

sparsity-inducing penalties we have proposed.

Investigating the convergence of proposed method, Figure 2 shows that all measure SDR, SIR, and SAR increases with more number of MU iterations. This confirms that the derived algorithm converges correctly and saturates after about 20 MU iterations.

The average speech separation performance over all mixtures in the test set, as a function of λ and α, is shown in Figure 3. As can be seen, the proposed algorithm is less sensitive to the choice of α and more sensitive to the choice of λ. It is quite stable with the small value of λ, and the result is best with $1 \leq \lambda \leq 25$ and $0 \leq \alpha \leq 0.4$. Overall the proposed algorithm is not very sensitive to the choice of such hyper-parameters and thus in the practical implementation one can set them quite easily.

5. Conclusions

In this paper, we have presented a speaker-dependent single-channel speech separation method based on the matrix factorization framework. Our method employed some different noise signal files to build the general spectral model for noise. For the estimation of the speech and noise signal from their mixture, we proposed the combination of NMF with two types of sparsity constraints. Experimental results showed the effectiveness of the proposed algorithm. Our further investigation showed the algorithm's convergence and its robustness to the choice of hyper-parameters λ and α. These properties are very useful for setting parameter in practical installation of the algorithm.
Future work could be devoted to extend the work to multi-channel case where the spatial model, such as the one considered in [31], for audio sources is incorporated. Additionally, validating the effectiveness of the proposed denoising approach for automatic speech recognition (ASR) would be a particular interest.

References

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer, 2005.

[2] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 126–130.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. NY, USA: IEEE, Oct. 2013, pp. 1–4.

[4] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 Signal Separation Evaluation Campaign," in *Latent Variable Analysis and Signal Separation*. Cham: Springer International Publishing, 2017, vol. 10169, pp. 323–332.

[5] B. V. Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.

[6] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 2578–2581.

[7] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574 – 584, 2015.

[8] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, article ID 24717, 2007.

[9] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[10] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[11] B. King, C. Fevotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation." IEEE, Sep. 2012, pp. 1–6.

[12] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[13] Q. Wang, W. Woo, and S. Dlay, "Informed single-channel speech separation using hmm gmm user-generated exemplar source," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 2087–2100, Dec 2014.

[14] L. Chen, X. Ma, and S. Ding, "Single Channel Speech Separation Using Deep Neural Network," in *Advances in Neural Networks - ISNN 2017*, F. Cong, A. Leung, and Q. Wei, Eds. Cham: Springer International Publishing, 2017, vol. 10261, pp. 285–292.

[15] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 9, pp. 1652–1664, 2016.

[16] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*. MIT Press, 2001, pp. 556–562.

[18] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[19] D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application

to on-the-fly audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 256–260.

[20] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation - a novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[21] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, T. H. Tran, and N. Q. K. Duong, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in *Proc. ACM Int. Sym. on Information and Communication Technology (SoICT)*, 2015, pp. 247–251.

[22] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, and N. Q. K. Duong, "Single-channel speaker-dependent speech enhancement exploiting generic noise model learned by non-negative matrix factorization," in *Proc. IEEE Int. Conf. on Electronics, Information, and Communications (ICEIC)*, 2016, pp. 1–4.

[23] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and Semi-supervised Separation of Sounds from Single-Channel Mixtures," in *Independent Component Analysis and Signal Separation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4666, pp. 414–421.

[24] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[25] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.

[26] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

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

[28] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 21–24.

[29] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Interspeech*, 2012, pp. 17–20.

[30] 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, Jul. 2006.

[31] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–11, 2013.