Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

A Novel Singing Voice Separation Method Based on Sparse Non-Negative Matrix Factorization and Low-Rank Modeling

S. Mavaddati*(C.A.)

Abstract: A new single channel singing voice separation algorithm is presented in this paper. This field of signal processing provides important capability in various areas dealing with singer identification, voice recognition, data retrieval. This separation procedure is done using a decomposition model based on the spectrogram of singing voice signals. The novelty of the proposed separation algorithm is related to different issues listed in the following: 1) The decomposition scheme employs the vocal and music models learned using sparse non-negative matrix factorization algorithm. The vocal signal and music accompaniment can be considered as sparse and low-rank components of a singing voice segment, respectively. 2) An alternating factorization algorithm is used to decompose input data based on the modeled structures of the vocal and musical components. 3) A voice activity detection algorithm is introduced based on the energy of coding coefficients matrix in the training step to learn the basis vectors that are related to instrumental parts. 4) In the separation phase, these non-vocal atoms are updated to the new test conditions using the domain transfer approach to result in a proper separation procedure with low reconstruction error. The performance evaluation of the proposed algorithm is done using different measures and leads to significantly better results in comparison with the earlier methods in this context and the traditional procedures. The average improvement values of the proposed separation algorithm for PESQ, fwSegSNR, SDI, and GNSDR measures in comparison with previous separation methods in two defined test scenario and three mentioned SMR levels are 0.53, 0.84, 0.39, and 2.19, respectively.

Keywords: Singing Voice Separation, Dictionary Learning, Incoherence, Sparse Coding, Voice Activity Detector.

1 Introduction

A singing voice separation algorithm with proper separation results can be employed in different signal processing fields such as data retrieval, singer identification, voice or lyric recognition [1-3]. The song recorded using a single microphone is investigated in the separation method presented in this paper. The separation process will be more difficult if the vocal (speech) signal is recorded in the presence of music accompaniment (non-vocal) signal that has a high energy level. The goal of this signal processing field is the separation of vocal and non-vocal parts of the recorded singing voice signal to remove the music accompaniment components. So, a separation approach is required to increase either intelligibility or quality of the captured vocal signal with the least distortion. Several singing voice separation algorithms have been presented in different areas such as autocorrelation-based [4], filter-based [5], pitch-based [6-7], low rank representation [8-10], non-negative matrix factorization (NMF) [11-15], robust principal component analysis (RPCA-based) [16-18] and dictionary-based methods [20-22, 24-25]. In [11], a recurrent neural network-based decomposition algorithm is proposed to capture long-term temporal dependencies between data structures using NMF and model sound mixtures. The efficiency of these approaches often depends on the

Iranian Journal of Electrical and Electronic Engineering, 2019.
Paper first received 28 February 2018 and accepted 24 August 2018.
* The author is with the Electronic Department, Faculty of Technology and Engineering, University of Mazandaran, Babolsar, Iran.
E-mail: s.mavaddati@umz.ac.ir.
Corresponding Author: S. Mavaddati.

captured music data during the input frames with vocal absence. In [12], a NMF algorithm is used to find individual components in the mixed data. Also, the proper value for the number of bases is selected by using manually classification of the recovered bases. An algorithm based on NMF is introduced in [13] to decompose the spectrum of music signal and provide criteria for the automatic component selection. Then, a pitch frequency is detected with the separated singing voice components and the calculated error between the separated singing voice and the pure singing voice is used to determine the ability of the separation method. A monaural singing voice separation algorithm is proposed in [14] using NMF to factorize the long-term and short-term mixture spectrograms. Then, a thresholding method is applied to select NMF components. This selection is in such a way that most pitched and percussive elements are separately filtered out from the mixed signal. In [15], a separation method is presented to capture the vocal parts from the mixed signal using a NMF-based framework. This algorithm separates the harmonic, percussive, and vocal structures from the input mixed signal using applying the constraint to each component to enforce its feature to have harmonic or temporal continuity.

In [16], an unsupervised separation algorithm using RPCA has been introduced to model the instrumental components of the singing voice signal. A binary mask using spectrogram is regarded to adjust the separation matrix coefficients. Also, in [18], the separation process is carried out based on a supervised RPCA procedure and deep recurrent neural network to ignore background music from singing voice frames. In dictionary learning-based methods, an input frame is approximately coded using the weighted linear combination of a small number of overcomplete dictionary atoms [19-20]. The separated vocal components are yielded from spares coding of the singing voice signal over learned atoms using an alternating optimization algorithm [19-20].

In this paper, it is shown that how to integrate SNMF into low-rank modeling in order to represent the vocal and non-vocal components of sound mixtures. The factorization parameters are set in such a way that the approximation error is reduced as much as possible and separation process is done with more accuracy. The vocal part and musical accompaniment in singing voice signal can be assumed sparse and low-rank in the time-frequency domain. Since the spectrum of frames involved instrumental content has a repetitive structure and highly correlated with each other, these segments are investigated as low-rank parts in singing voice signal.

The remainder of this paper is structured as follows. In Section 2, the voice singing separation problem is described. Section 3 has a detailed overview of the proposed separation method based on SNMF and low-rank-modeling. In Section 4, the experimental results are expressed. Finally, the experimental results are provided in Section 5 and conclude the paper in Section 6.

2 Problem description

The vocal and non-vocal frames of a singing voice signal can be investigated as sparse and low-rank components in the time-frequency domain. Therefore, the input signal should be transformed into this domain to earn more details. As mentioned, monaural recording in a real environment can be linearly defined as

$$\mathbf{Y}(f,m) = \mathbf{S}(f,m) + \mathbf{L}(f,m) \tag{1}$$

where $\mathbf{Y}(f,m)$, $\mathbf{S}(f,m)$ and $\mathbf{L}(f,m)$ are the spectrograms of the singing voice, vocal and non-vocal signals at frequency bin f and frame number m, respectively. The singing voice signal $\mathbf{Y}\in\mathbb{R}^N$ can be represented linearly by a SNMF algorithm as $\mathbf{Y}=\mathbf{WH}$, where $\mathbf{W}\in\mathbb{R}^{N\times P}$, $P > N$ is a dictionary matrix with P atoms shown by $\{w_p\}_{p=1}^{P}$ with unit norm $\left\|w_{(:,p)}\right\|_2 \forall p = 1,\ldots,P$ and the activity matrix $\mathbf{H}\in\mathbb{R}^P$, $P \gg K$ involves the coefficients of factorization process over $\mathbf{W}$ [27-28]. Non-negativity constraints $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$ apply on both matrices. This factorization problem according to approximation error and sparsity constraint is solved by iterating the following non-negativity constrained least square problems formulated as [24-25]:

$$F(W,H) = \|Y - WH\|^2 + \alpha \sum_p h_p \tag{2}$$

where positive constant α is the weighted coefficient for the second term that calculates sparsity in each row of $\mathbf{H}$. The generalized Kullback-Leibler divergence is used to measure the approximation error. Then, SNMF is formulated using the following optimization problem [24]:

$$F(W,H) = \sum_{n,p} \left(Y_{n,p}.\log\left(Y_{n,p} / (WH)_{n,p}\right) - Y_{n,p} + (WH)_{n,p} \right) + \alpha h_{p0} \tag{3}$$

where $\|H\|_0$ is the number of non-zero coefficients in each row of $\mathbf{H}$ or sparsity constraint K.

3 Overview of the Proposed Method

Fig. 1 shows the block diagram of the proposed method. At first, the singing voice signals in the train and test steps are transformed into STFT domain. The detail of each block has been explained in the following.

3.1 Dictionary Learning Using SNMF

In this paper, a new solution for singing voice separation problem using incoherent dictionary learning is presented. The idea of using dictionary learning

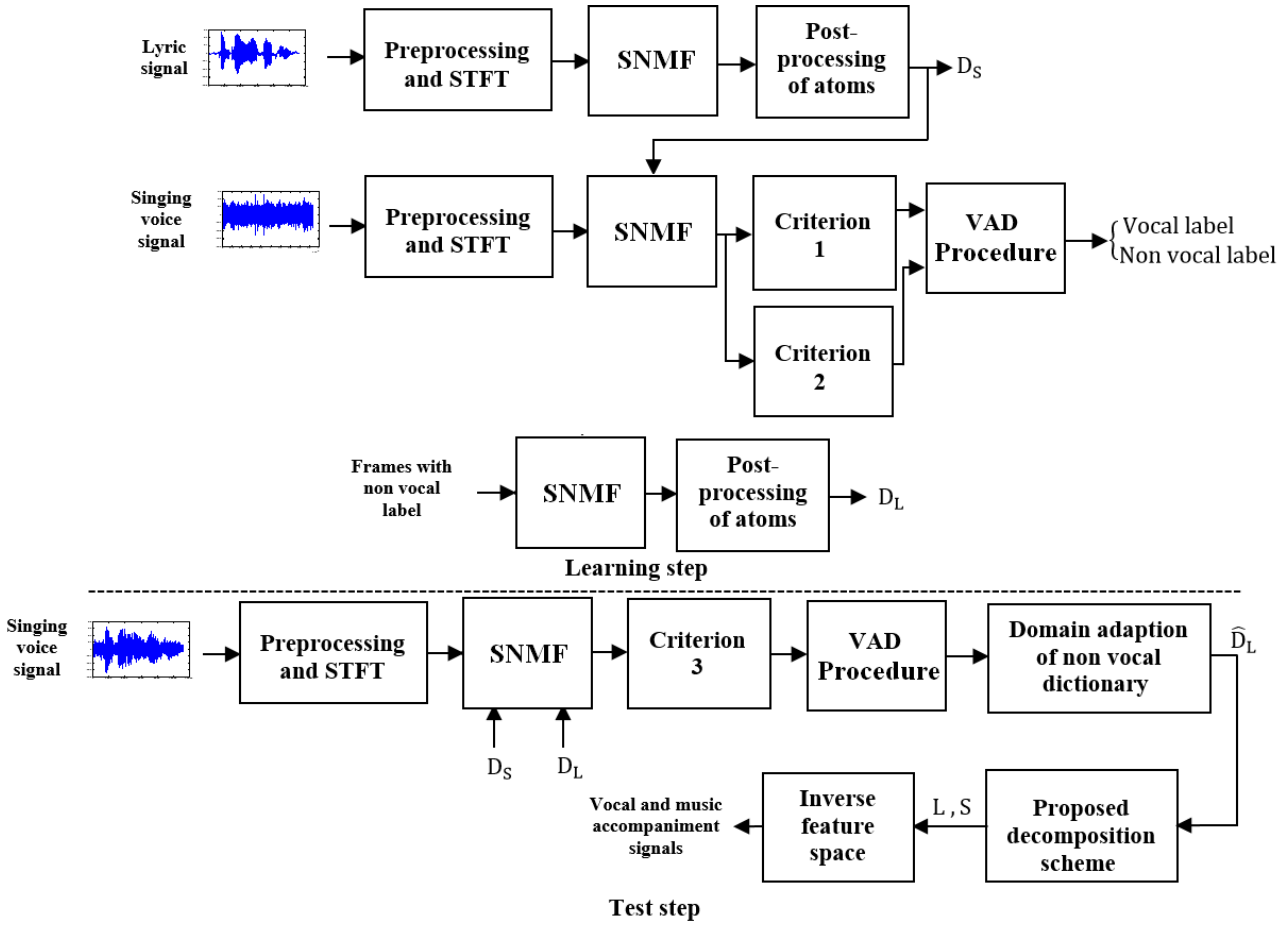


Fig. 2 Block diagram of the proposed separation procedure of singing voice signal with learning and test steps.

concepts is one of the most interest fields in signal processing research. The input frame is approximately modeled as a weighted linear combination of a small number of basis vectors to compress data or reduce dimensionality. The sparseness constraint is imposed on activity coefficients to control the speed of learning process. The SNMF algorithm in the first step of dictionary learning algorithm is utilized to learn a generative model for vocal and non-vocal segments of the singing voice signal.

The atoms related to the vocal signal S as a sparse component have been trained over the original clean voice signal as

$$W_S^*, H_S^* = \arg\min_{W_S, H_S} \|S - W_S H_S\|^2 \quad \text{s.t. } \|H_S\|_0 \le K \qquad (4)$$

This pre-trained dictionary is employed to capture enough data in learning procedure of non-vocal or low-rank data L. The non-vocal dictionary atoms are learned based on the captured music components using a voice activity detection (VAD) procedure in the next step. The non-vocal segments based on the energy of sparse coefficients in coding over vocal dictionary are detected using this energy-based VAD algorithm. If an input frame has low sparse coefficients energy, it means that this frame will not include speech nature and cannot be coded properly over vocal atoms. Therefore a right decision will be made against the input non-vocal frame. These captured frames L are employed for learning non-vocal atoms as

$$W_L^*, H_L^* = \arg\min_{W_L, H_L} \|L - W_L H_L\|^2 \quad \text{s.t. } \|H_L\|_0 \le K \qquad (5)$$

Then, an alternating optimization method is used to solve the factorization sub-problems on the sparsity and low-rank characteristic of speech and music frames in the test step.

A main characteristic in the learning process is mutual coherence between atoms to determine the dependency between atoms. Lower mutual coherence value results in a dictionary with independent basis vectors as much as possible. In this paper, a post-processing method is employed to earn incoherent dictionaries. This parameter is adjusted in such a way that the reconstruction error is reduced to the separation process is done with more accuracy. This goal will be earned when atoms have low coherence value as equiangular tight frame (ETF). Atom coherence is defined as the maximum absolute value of the cross-correlations between atoms or maximum absolute value of the off-diagonal elements of $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ with the normalized atoms [26]. The iterative projection and rotation (IPR)

method proposed in [27] is used in order to attain incoherent dictionaries in a post-processing step. IPR sets the structural and spectral constraints on the number of non-zero eigenvalues of **G**. Then, the residual norm of approximation is minimized by atom rotation with an orthogonal matrix [27].

3.2. Domain Transfer Technique

After dictionary learning based on SNMF, the trained non-vocal dictionary atoms in the test step are adapted to the new ones according to the initial and final frames of the input segments. This adaptation process is carried out using the domain transfer technique which was previously used for speech enhancement in the presence of background piano signal [20, 23, 28]. This adaptation technique is employed to transfer non-vocal dictionary atoms to an adapted dictionary based on the background music of the test environment captured based on the VAD algorithm defined in Section 3.3. This is a useful approach when faced with a music component that has different content in learning and test steps. In fact, this mismatch can be alleviated as much as possible using this adaptation scheme [28].

This domain adaptation method that is based on an analytical solution to update dictionary atoms is utilized in this paper. In a standard dictionary learning algorithm, the whole or most parts of data such as image or speech signal is required that is time-consuming. Using domain transfer technique just a few patches of data in the test environment are employed in the learning process [28]. Therefore, the source domain dictionary is transferred to a target domain for image denoising with a dictionary-regularization term designed based on the energy function. In the proposed supervised algorithm for speech enhancement in [20] the domain adaptation technique is used to transfer a learned noise dictionary to a dictionary adapted to the noise circumstances captured based on the test environment. Using this technique, the observed sparse speech data can be coded over the adapted noise dictionary with lower sparse reconstruction error. This technique has a prominent role to obtain better enhancement results particularly when the noise is non-stationary. It is clear that encountering with non-stationary noises can severely reduce the performance of speech enhancement algorithms. In [23], this transfer technique is used for speech enhancement in wavelet packet transform domain. Therefore, an adapted separation scheme based on the noisy space is carried out and the main drawbacks seen in the earlier dictionary-learning-based speech enhancement methods are solved.

3.3. Proposed Voice Activity Detector

Our proposed VAD scheme in time-frequency domain based on dictionary learning technique involves different sections illustrated as follows:

Criterion 1: As be mentioned, the sparse and low-rank components of input frames are detected using an energy-based VAD algorithm over vocal and non-vocal learned dictionaries. In training step, vocal dictionary W_s trained over the input frames involves only lyric wave or vocal components. The singing voice frames Y is sparsely factorized over W_s to yield activity matrix H_s. Then the similarity between the original frame Y and the reconstructed frame $\hat{Y}$ achieved based on the sparse factorization over W_s, is calculated. If the approximation error $\|Y-W_sH_s\|_2$ is high or $\|Y-\hat{Y}\|_2 < \varepsilon$, the input frame is labeled as vocal. On the other hand, if $\|Y-\hat{Y}\|_2 > \varepsilon$, the input label will have non-vocal structure. This similarity measure with high value expresses that the input frame has speech structure since it has been properly coded over sparse dictionaryW_s.

Criterion 2: In order to ensure the correctness of our decision about the input frame label, factorization process of the observed data is done over the initial and final frames of the input singing voice signal defined by W_{L0}. Since these frames have non-vocal nature, if the energy of the activity matrix H_s is high, it means that this frame has non-vocal or low-rank content and will be properly factorized over W_{L0}. If this factorization is done with high activity coefficient energy, the vocal label is assigned.

Using this condition, a label will be devoted to each input frame as vocal or non-vocal part. As can be seen in Fig. 1, these criteria are employed in the learning step. Also, the input frame that does not satisfy the mentioned conditions is not used in the test step.

Criterion 3: In the test step that sparse and low-rank dictionaries are trained, a non-negative factorization process based on the energy of each row of the activity matrix over the composite dictionary is carried out. The singing voice signal Y in the test step is factorized over the composite dictionary using SNMF procedure as

$$H_S \ , H_L = SNMF\left(Y \ , \left[W_S W_L\right]\right) \tag{6}$$

Then, the energy of the activity coefficients related to each dictionary is computed. If the energy in this representation over is more than the defined threshold, the non-vocal label is assigned to input frame and use in the next step included atom adaptation according to the test circumstances.

3.4 Decomposition Procedure

In this paper, a new separation method for singing voice is introduced without setting new parameters to adjust non-negative coefficients in the separation step and eliminate having a complicated training scheme. So, an alternating optimization method is introduced to solve the factorization problems on the sparsity and low-rank components of voice and instrumental parts in the test step. In some presented separation methods, previous knowledge about the statistics of input data is neglected while using the learned models results in a

better performance especially about singing voice separation problem. RPCA is one of the basic procedures for sparse low-rank decomposition that uses an alternating algorithm to analyze data content by imposing constraints on rank and sparsity of each observed frame [17]. The sparse and low-rank components are obtained using hard thresholding and singular value decomposition (SVD) of the input frame, respectively [17].

$$\min \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \text{ subject to } \mathbf{Y} = \mathbf{L} + \mathbf{S} \tag{7}$$

where $\|.\|_*$ denote the nuclear norm and $\mathbf{L}$ shows the rank parameter. The following sub-problems have been formulated in RPCA separation techniques [17, 25]

$$\begin{aligned} \mathbf{L}^t &= \arg\min_{rank(L)\le r} \left\|\mathbf{Y}-\mathbf{L}-\mathbf{S}^{t-1}\right\|_F^2 \\ \mathbf{S}^t &= \arg\min_{Card(S)\le k \;\&S\ge 0} \left\|\mathbf{Y}-\mathbf{L}^t-\mathbf{S}\right\|_F^2 \end{aligned} \tag{8}$$

The rank value of low-rank component $\mathbf{L}\in\mathbb{R}^{N\times B}$ is shown using rank(.) bounded to $r \le \min(N,B)$. Also, Card(.) is the cardinality index of the sparse part and $\|.\|_F$ is the Frobenius norm.

In the proposed factorization method, the pre-trained models of vocal and non-vocal components learned using SNMF are imposed to (7) as

$$\begin{aligned} &\min \mathbf{W}_L.\mathbf{H}_{L^*} + \lambda \mathbf{W}_S.\mathbf{H}_{S1} \\ &\text{s.t. } \mathbf{Y} = \mathbf{W}_L.\mathbf{H}_L + \mathbf{W}_S.\mathbf{H}_S \end{aligned} \tag{9}$$

Using alternately solving this separation problem, Eq. (9) can be formulated in a low-rank modeling as

$$\begin{aligned} &\min_{\mathbf{H}_L} \frac{1}{2}\left\|\mathbf{Y}-\mathbf{W}_L.\mathbf{H}_L-\mathbf{S}\right\|_2^2 + \lambda_{\mathbf{H}_L}\left\|\mathbf{H}_L\right\|_1 + \lambda_{\mathbf{L}}\left\|\mathbf{D}_L\mathbf{H}_L\right\|_* \\ &\min_{\mathbf{H}_S} \frac{1}{2}\left\|\mathbf{Y}-\mathbf{L}-\mathbf{W}_S.\mathbf{H}_S\right\|_2^2 + \lambda_{\mathbf{H}_S}\left\|\mathbf{H}_S\right\|_1 \end{aligned} \tag{10}$$

where λ_{H_L} and λ_{H_S} denote the sparsity index for activity matrices. Also, λ_{L} denotes the rank value of the instrumental component. The low-rank and sparse components of each input data are separated using analytical solutions proposed in [23] based on SVD technique and hard thresholding. Our proposed algorithm for solving the separation problem using learned dictionaries for vocal and non-vocal parts is detailed in Algorithm 1.

Algorithm 1 Proposed singing voice separation problem based on SNMF.

Input:$\mathbf{W}_S$, $\mathbf{W}_L$, Y, t_{itr} (Number of iteration), r (Rank value), T (Threshold value)
Output:S , L
Initialization: $\mathbf{H}_S$ = [0], $\mathbf{H}_L$ = [0]
For: t = 1**:** t_{itr}
%Update activity matrix $\mathbf{H}_L$ and low-rank component L
H_L^* = SNMF(Y,$\mathbf{W}_L$)
$U \Lambda V = \text{SVD}(\mathbf{W}_L.\mathbf{H}_L^*)$
$L^{t+1} = \sum_{i=1}^{r} \lambda_i U_i V_i$
$H_L^{t+1} = \text{SNMF}\left(L^{t+1}, \mathbf{W}_L\right)$
%Update activity matrix $\mathbf{H}_S$ and low-rank component S
$H_S^* = \text{SNMF}\left(Y, \mathbf{W}_S\right)$
$S^{t+1} = \text{Threshold}\left(\mathbf{W}_S.H_S^*, T\right)$
$H_S^{t+1} = \text{SNMF}\left(S^{t+1}, \mathbf{W}_S\right)$
$t = t+1$
End for

4 Implementation Details

The proposed separation method is assessed using different evaluation measures using MIR-1K corpus [29]. This large multi-talker database has been used for research about singing voice separation problem that includes 1000 song recording of 19 singers of both genders with sentences about 4s to 13s. These implementations are carried out in two scenarios: supervised test (ST) and unsupervised test (UT). In ST scenario, the singers in train and test steps are similar but not in UT situation. In fact, the supervised situation involves similar singers and different lyrics in train and test steps This train and test sets include 400 and 100 song of 4 male and 5 female singers in train step and 3 male and 4 female singers in the test step of the unsupervised scenario. The input data is resampled from 16 kHz to 8 kHz and feature coefficients are yielded using a 256-point STFT. The signals are framed with 37.5 ms segment length and 40% overlap with a Hamming window. All framing, pre-processing steps are the same in train and test steps. It should be mentioned that the structure of noise in noisy signal does not possess an exact sparse representation in a dictionary trained over pure noise signal.

It should be mentioned that the structure of noise in noisy signal does not possess an exact sparse representation in a dictionary trained over pure noise signal.

It should be mentioned that the structure of vocal and music segments are very different so the non-vocal frames in the singing voice signal in both ST and UT scenarios do not possess an adequate representation over the learned vocal basis and will be coded over low-rank atoms in the composite dictionary. The threshold values in criteria 1-3 in VAD procedure are set based on the experimental simulation results. The ε parameter is set to 0.25 for all speech signals in all SMR levels. Since the music signal has low-rank components, the rank value is selected based on the experimental results and high correlation between the consecutive frames of the music signal. The rank value in the decomposition scheme based on SNMF technique is set to 2. The magnitude of signal spectrums have been used in

training procedure and the phase is kept unchanged during the synthesis. The challenge in the singing voice separation problem occurs when the level of background music is high. Therefore, the lower level of SMR (signal to music ratio) is considered in the simulations. It is clear that the results will be much better in higher SMR values such. Each singing voice in the test step is combined with background music at SMRs of -5dB, 0dB, and 5dB. The BSS-EVAL toolbox is employed to assess the proposed algorithm [30]. All dictionary atoms are initialized with random training data chosen from input frames. After dictionary learning using SNMF, the atoms are decorrelated using IPR algorithm as mentioned in Section 3.

5 Evaluation

This section provides experimental results to compare our separation approach with some baseline methods and the earlier algorithm using different evaluation measures. These implementations are done to show the main role of dictionary learning in singing voice separation process with imposing apriori information about data structures. The resulted measurement values are averaged over all test signals. For better evaluation with more details, the proposed approach is compared with the previous method introduced in [8-9, 16, 18]. The selected assessment measures are frequency-weighted segmental SNR (fwSegSNR), PESQ scores [31], speech distortion index (SDI) [32] and global normalized source to distortion ratio (GNSDR) [16, 18]. The fwSegSNR as a speech evaluation measure is chosen to determine the quality of the estimated vocal over in different frequency bands [33]. PESQ measure estimates mean opinion score as a subjective criterion to determine intelligibility of capture vocal signal [31]. Another measure is SDI index that determines instrumental part included in the separated vocal signal and defined as [32]

$$\mathrm{SDI}\left(S,\hat{S}\right)=\frac{E\left\{\left[S\left(n\right)-\hat{S}\left(n\right)\right]^{2}\right\}}{E\left\{S^{2}\left(n\right)\right\}} \tag{11}$$

where $s(n)$ and $\hat{s}(n)$ are the initial vocal signal and its estimated form at the sampling time index n, respectively. The GNSDR measure is formulated by averaging over all captured signals as [16]:

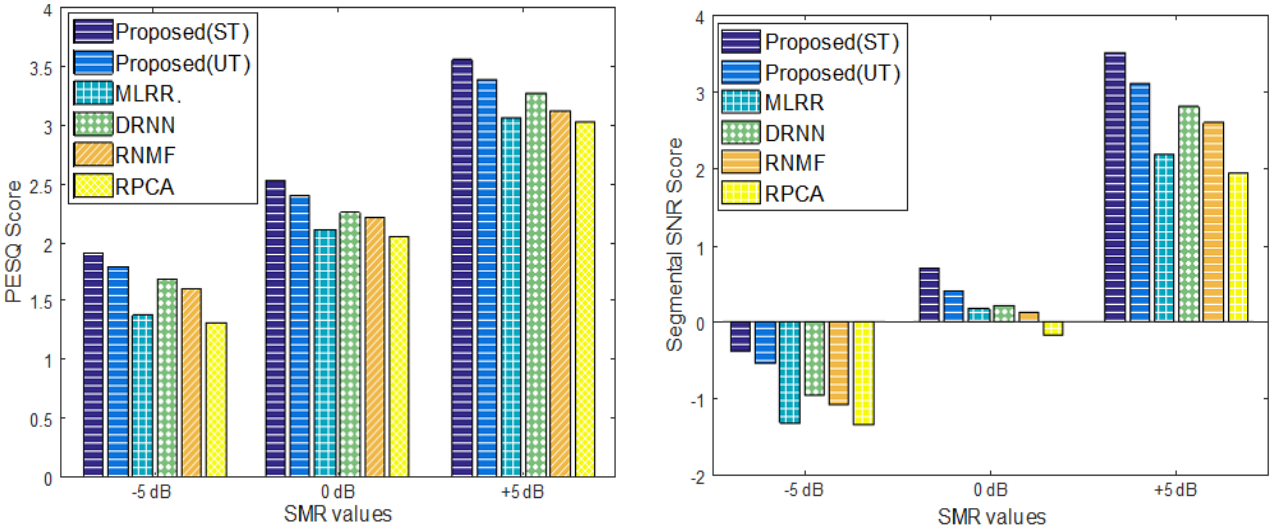
$$\mathrm{GNSDR}\left(S,\hat{S},Y\right)=\frac{\sum_{n=1}^{N}\alpha_{n}\left[\mathrm{NSDR}\left(S,\hat{S}\right)-\mathrm{NSDR}\left(S,Y\right)\right]}{E\left\{s^{2}\left(n\right)\right\}} \tag{12}$$

where NSDR is the normalized source to distortion ratio. N is the total number of tests song and α_n is the weighted factor related to the test signal length. A vocal signal with more quality and intelligibility is estimated when higher values for fwSegSNR, PESQ and GNSDR have been earned. In terms of SDI measure, lower values show better separation result and less distortion in the captured vocal signal. The MATLAB implementations of fwSegSNR and PESQ provided in [34-35] are used. Our implementation is done on a Windows 64-bit based computer with Core i5 3.2GHz CPU in all train and test scenarios. A comprehensive model trained with an acceptable reconstruction error for vocal and music data using SNMF in combination with low-rank modeling is used in our separation process. Therefore, the frequency contents of a test signal can be exactly represented using the corresponding atoms, for example, the music spectrum frames do not have activity coefficients over vocal atoms and should be sparsely modeled by the non-vocal atoms using the presented factorization scheme. It is worth mentioning that imposing sparse constraint to the learning process and using domain adaptation technique to reduce the mismatch effect between train and test circumstances result in a lower reconstruction error and a precise factorization scheme. The separation results of ST and UT scenarios measured by the PESQ and fwSegSNR scores at different SMR values for the proposed method in comparison with other mentioned algorithms are shown in Fig. 2. Also, the SDI and GNSDR measure values are expressed in Fig. 3. All results in this Section obtain from averaging over all test data. In order to have a proper comparison with more details between different algorithms, these results are reported in Tables 2 and 3.

In Figs. 2-3, it is shown that the proposed method in the ST situation earns considerably better results than other mentioned methods that are based on other processing such as low-rank modeling (RNMF, MLRR), neural network (DRNN) and RPCA. This superiority achieves at all SMR conditions and for different assessment measures so our separation algorithm outperforms other methods in this field [8-9, 16, 18].

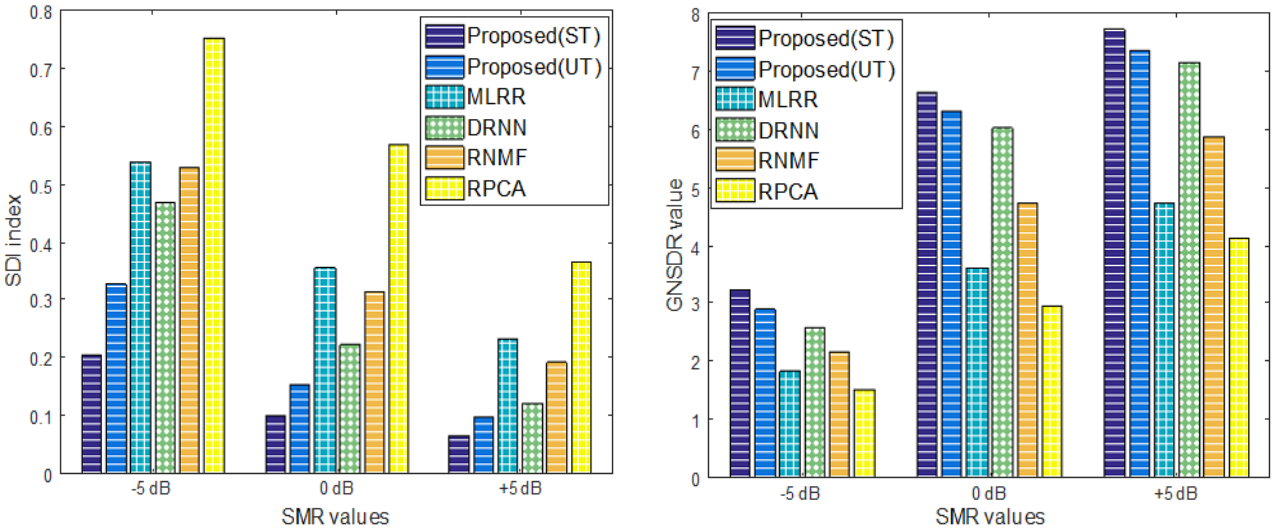
Table 1 The abbreviations used for different compared methods.

Proposed (ST)	Proposed algorithm in the supervised test
Proposed (UT)	Proposed separation algorithm in the unsupervised test
MLRR	Separation algorithm using multiple low-rank representation [8]
DRNN	Separation algorithm using deep recurrent neural networks [18]
RNMF	Separation algorithm using robust low-rank non-negative matrix factorization (RNMF) [9]
RPCA	Separation algorithm using robust principal component analysis (RPCA) [16]



(a) (b)

Fig. 2 Performance comparison of different methods in terms of PESQ and fwSegSNR scores at different SMR values.



(a) (b)

Fig. 3 Performance comparison of different methods in terms of SDI and GNSDR at different SMRs.

Table 2 The results of PESQ score and fwSegSNR measure in SMRs of -5, 0 and 5 for different methods.

	PESQ			fwSegSNR		
	-5dB	0dB	+5dB	-5dB	0dB	+5dB
Proposed (ST)	1.9206	2.5317	3.5696	-0.3880	0.7081	3.5211
Proposed (UT)	1.8042	2.4056	3.3950	-0.5432	0.4074	3.1137
DRNN [18]	1.6954	2.2638	3.2830	-0.9604	0.2156	2.8126
MLRR [8]	1.3916	2.1168	3.0772	-1.3230	0.1764	2.1854
RNMF [9]	1.6170	2.2246	3.1360	-1.0780	0.1176	2.6068
RPCA [16]	1.3230	2.0580	3.0380	-1.3524	-0.1764	1.9404

Table 3 The results of SDI index and GNSDR measure in SMRs of -5, 0 and 5 for different methods.

	SDI			GNSDR		
	-5dB	0dB	+5dB	-5dB	0dB	+5dB
Proposed (ST)	0.2060	0.1009	0.0669	3.2495	6.6542	7.7503
Proposed (UT)	0.3296	0.1545	0.0989	2.9197	6.3438	7.3914
DRNN [18]	0.4692	0.2244	0.1224	2.5899	6.0431	7.1780
MLRR [8]	0.5406	0.3570	0.2346	1.8522	3.6358	4.7530
RNMF [9]	0.5304	0.3162	0.1938	2.1854	4.7530	5.8898
RPCA [16]	0.7548	0.5712	0.3672	1.5288	2.9792	4.1454

These results emphasize the significant role of employing SNMF in combination with low-rank modeling and separation process based on the proposed optimization algorithm. As can be seen, our proposed method properly estimates vocal signal from background music component that has periodic content for any musical instruments. This repetitive content is well-structured to provide a generative incoherent dictionary, so a non-vocal segment can be sparsely modeled over related atoms with high accuracy that results in a precise separation procedure.

The proposed separation algorithm can obtain the improvement values 0.53, 0.84, 0.39, and 2.19 for the PESQ, fwSegSNR, SDI, and GNSDR measures, respectively. These values are achieved in comparison with previous separation methods in two defined test scenario and three mentioned SMR levels are.

The UT situation gains slightly lower results than ST scenario since it is similar to speaker independent case in speaker recognition methods with different singers and lyrics in train and test steps. The vocal atoms in ST have more coherent to input frames observed in the test step than UT situation that results in a factorization process with lower reconstruction error or separation with more quality.

Another reason for the superiority of the proposed method is due to two cases, updating non-vocal atoms according to the background music of test step to alleviate the mismatch between instrumental components with train phase and also training incoherence atoms using IPR for vocal and non-vocal framed based on the proposed VAD. Learning incoherent atoms causes that a well-structured music

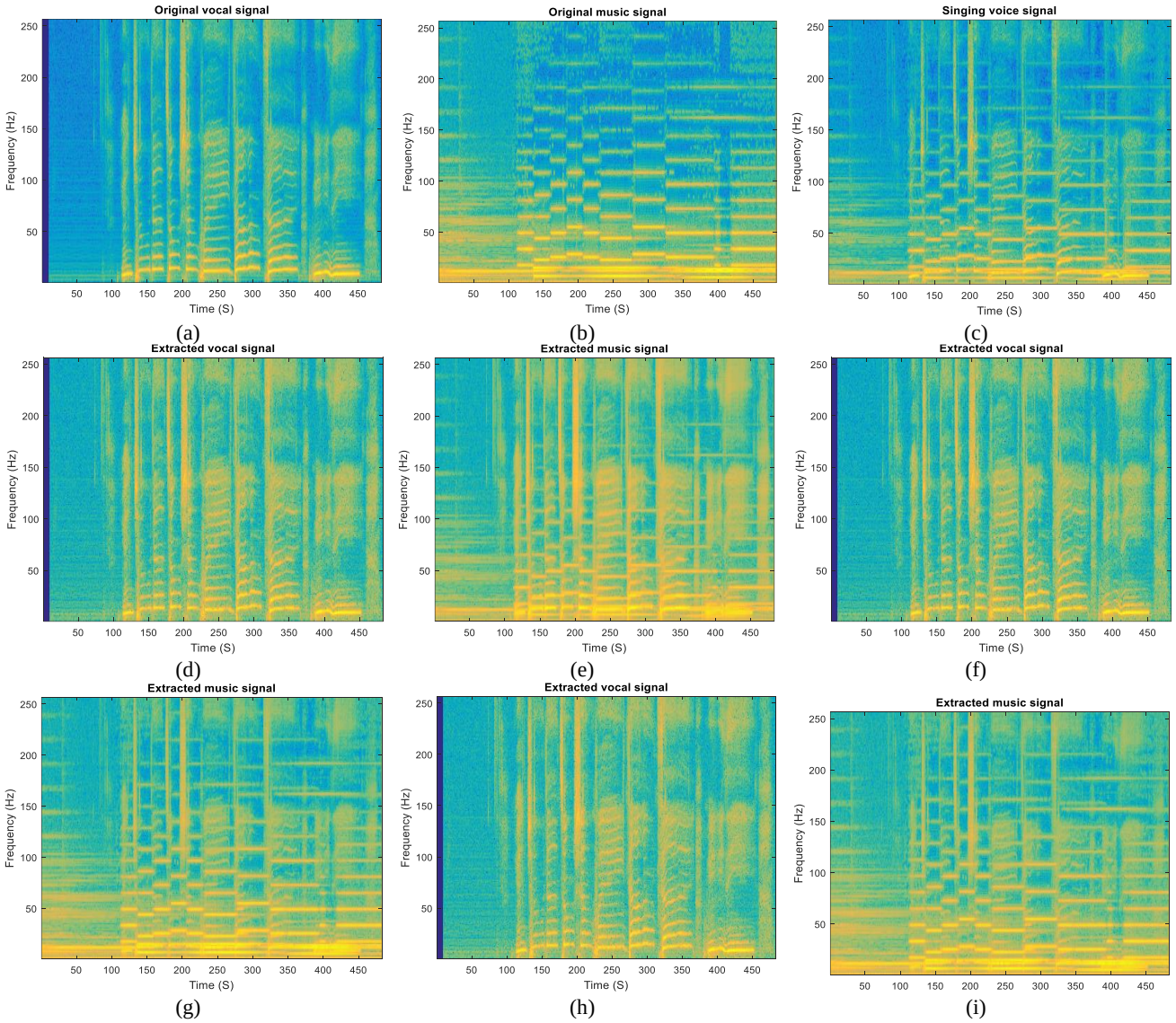


Fig. 4 The spectrograms of a) clean vocal signal, b) original music instrumental, c) singing voice signal at SMR = 0dB. The estimated signals using the proposed method at SMR = 5dB: d) Vocal and e) Non-vocal parts. The estimated signals using the proposed method at SMR = 0dB: f) Vocal and g) Non-vocal parts. The estimated signals using the proposed method at SMR = -5dB: h) Vocal and i) Non-vocal parts.

frames are ignored in sparse representation using SNMF over vocal atoms.

Moreover, using an alternating optimization method over learned incoherent dictionaries for solving constrained decomposition problem of sparsity and rank level components have a prominent role in improving separation quality. Also, the lowest performance is yielded using RPCA method, since it works using in an unsupervised way without any prior information about the content of the input frames.

In order to have more performance evaluation about the proposed method, the spectrogram of the mentioned singing voice separation algorithms is investigated. The lyric "abjones_1_01" in MIR-1K database was combined with the background music at SMRs of -5dB, 0dB, and 5dB. The spectrograms of the clean vocal, non-vocal and mixed signals are shown in Fig. 4. Moreover, the separated sparse and low-rank components of singing voice signal by the presented algorithm are presented in this Figure.

Also, the decomposition results of the observation data to sparse part as vocal segment and low-rank part as background music frame for these methods at SMR=0 have been plotted in Fig. 5. These results are obtained in the unsupervised situation. Consideration of these spectrograms shows that our learning-based separation algorithm has been able to eliminate successfully the background music from singing voice signal at different SMR values and the estimated vocal spectrogram is separated with more accuracy than other approaches. Also, it is shown that other methods have more visible residual music components that result in more vocal distortion than our proposed algorithm.

6 Conclusions

In this paper, a new decomposition procedure is introduced for the singing voice separation problem based on the combination of SNMF and low-rank modeling. An incoherence generative model is learned for vocal and non-vocal parts of the singing voice signal considered as sparse and low-rank components in the time-frequency domain. A dictionary with incoherent atoms can represent input frames with low approximation error that results in a precise factorization process. This alternating factorization method includes vocal and non-vocal models to decompose sparse and low-rank parts of singing voice signal obtained from monaural recordings. Also, an energy-based VAD scheme is presented to capture enough music frames for atom training related to non-vocal data. These learned atoms are updated corresponding to the estimated background music using domain transfer technique in the test step. Using apriori information about components of the input frame using SNMF increases the quality and intelligibility of the estimated vocal signal especially in low SMR values. The experimental results using different assessment measures show that the presented algorithm obtained significantly better results than the earlier methods in this context and the traditional methods.

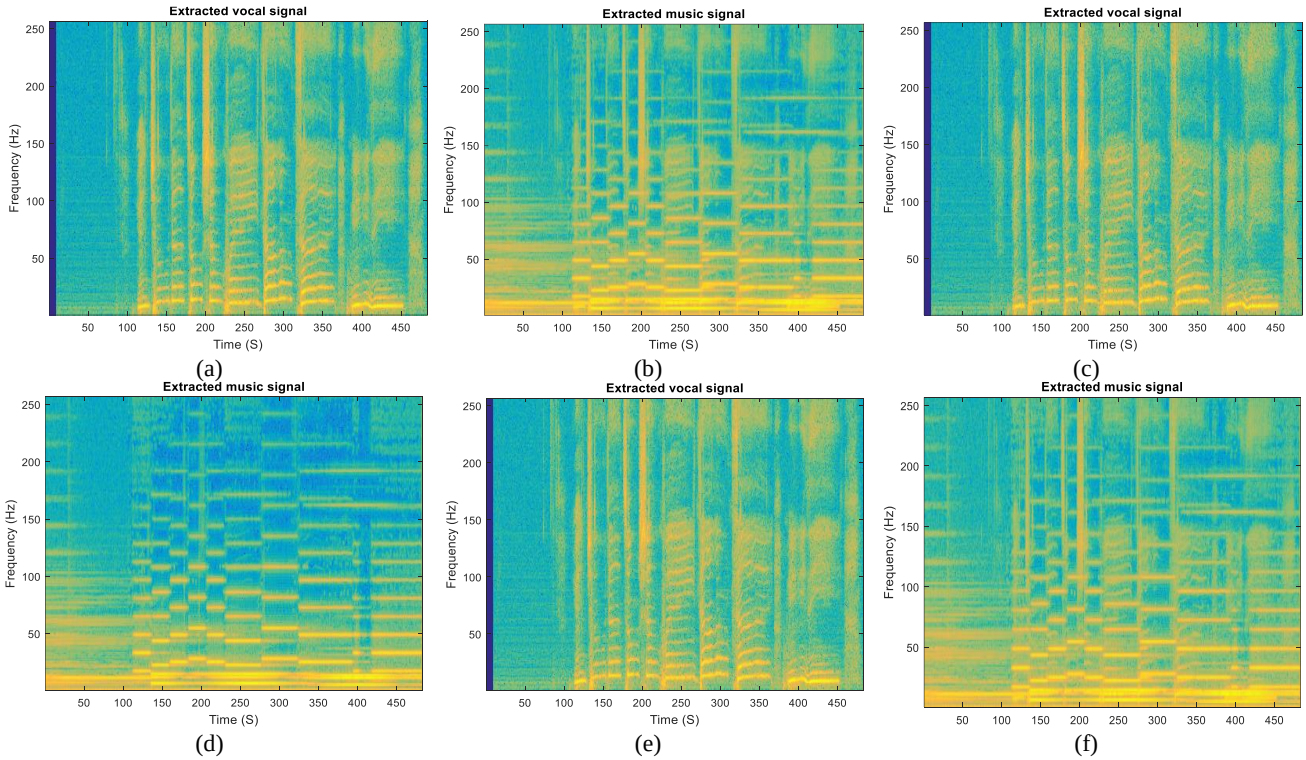


Fig. 5 The spectrograms of captured sparse and low-rank components at SMR=0dB using: The proposed method, a) Vocal and b) Non-vocal signals. DRNN method, c) Vocal and d) Non-vocal signals. MLRR algorithm, e) Vocal and f) Non-vocal signals.

References

[1] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyric synchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 6, pp. 1252–1261, 2011.

[2] M. Lagrange, A. Ozerov, and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," in *13th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 595–560, 2012.

[3] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *8th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 467–470, 2007.

[4] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pp. 221–224, 2011.

[5] J. L. Durrieu, G. Richard, B. David and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 3, pp. 564–575, 2010.

[6] L. Yipeng and W. DeLiang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 4, pp. 1475-1487, 2007.

[7] C. L. Hsu, J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 2, pp. 310-319, 2010.

[8] Y. H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *14th International Society for Music Information Retrieval Conference*, pp. 427–432, 2013.

[9] P. Sprechmann, A. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *13th International Society for Music Information Retrieval Conference*, pp. 67–72, 2012.

[10] Y. H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *ACM Multimedia*, pp. 757–760, 2012.

[11] N. Boulanger, G. Mysore, and M. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7019–7023, 2014.

[12] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *DMRN Summer Conference*, Glasgow, pp. 23–24, 2005.

[13] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in *International Conference on Advanced Technologies for Communications*, 2008.

[14] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 10, pp. 2096–2107, 2013.

[15] E. Ochiai, T. Fujisawa, and M. Ikehara, "Vocal separation by constrained non-negative matrix factorization," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 480–483, Dec.2015.

[16] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa, "Singing voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60, 2012.

[17] E. J. Candes, L. Xiaodong, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, Vol. 58, No. 3, pp. 1-39, 2011.

[18] P. S. Huang, M. Kim, M. Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[19] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 20, pp. 1698–1712, 2012.

[20] S. Mavaddati, S. M. Ahadi, and S. Seyedin, "Speech enhancement using sparse dictionary learning in wavelet packet transform domain," *Computer, Speech & Language*, Vol. 44, pp. 22-47, 2017.

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing over-complete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, Vol. 54, pp. 4311–4322, 2006.

[22] S. Mavaddati, S. M. Ahadi, and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement," *IET Signal Processing*, Vol. 9, No. 7, pp. 1–9, 2015.

[23] S. Mavaddati, S. M. Ahadi, and S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Communication*, Vol. 76, pp. 42–60, 2016.

[24] W. Liu, N. Zheng, and X. Lu, "Non-negative matrix factorization for visual coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 293–296, 2003.

[25] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis. Bioinformatics," Vol. 23, No. 12, pp. 1495–1502, 2007.

[26] D. T. You, J. Q. Han, G. B. Zheng, T. R. Zheng, and J. Li, "Sparse representation with optimized learned dictionary for robust voice activity detection," *Circuits, Systems, and Signal Processing*, Vol. 33, pp. 2267–2291, 2014.

[27] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Processing Letters*, Vol. 20, No. 5, pp. 475–478, 2013.

[28] G. Chen, C. Xiong, and J. J. Corso, "Dictionary transfer for image denoising via domain adaptation," in *Proceedings of IEEE International Conference on Image Processing*, pp.1189–1192, 2012.

[29] MIR-1K Dataset, https://sites.google.com/site/unvoicedsoundseparation/mir-1k.

[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, 2006.

[31] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band importance functions," *he Journal of the Acoustical Society of America*, Vol. 125, No. 5, pp. 3387–3405, 2009.

[32] J. Benesty, *Springer handbook of speech processing*. Springer's publication, pp. 843–871, 2008.

[33] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 24, No. 5, pp. 380–391, 1976.

[34] P. C. Loizou, *Speech enhancement: Theory and practice*. Taylor & Francis, New York, 2007.

[35] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of International Conference on Acoustics, Speech, Signal Processing*, pp. 749–752, 2001.

S. Mavaddati received the Ph.D. degree in Electrical Engineering from Amirkabir University of Technology, in 2016 and the M.Sc. degree in Electrical Engineering from the University of Mazandaran, Babolsar, Iran, in 2010. In 2017, she joined the Faculty of Technology and Engineering, University of Mazandaran. Her major fields of interest include voice and speech processing, image processing, evolutionary computation and artificial intelligence.

© 2019 by the authors. Licensee IUST, Tehran, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license (https://creativecommons.org/licenses/by-nc/4.0/).