Circuits, Systems, and Signal Processing (2020) 39:3652–3681
https://doi.org/10.1007/s00034-019-01338-0

A Novel Singing Voice Separation Method Based on a Learnable Decomposition Technique

Samira Mavaddati[1]

Received: 27 January 2019 / Revised: 25 December 2019 / Accepted: 27 December 2019 /
Published online: 8 January 2020

Abstract

In this paper, a new monaural singing voice separation algorithm is presented. This field of signal processing provides important information in many areas dealing with voice recognition, data retrieval, and singer identification. The proposed approach includes a sparse and low-rank decomposition model using spectrogram of the singing voice signals. The vocal and non-vocal parts of a singing voice signal are investigated as sparse and low-rank components, respectively. An alternating optimization algorithm is applied to decompose the singing voice frames using the sparse representation technique over the vocal and non-vocal dictionaries. Also, a novel voice activity detector is presented based upon the energy of the sparse coefficients to learn atoms related to the non-vocal data in the training step. In the test phase, the learned non-vocal atoms of the music instrumental part are updated according to the non-vocal components captured from the test signal using domain adaptation technique. The proposed dictionary learning process includes two coherence measures: atom–data coherence and mutual coherence to provide a learning procedure with low reconstruction error along with a proper separation in the test step. The simulation results using different measures show that the proposed method leads to significantly better results in comparison with the earlier methods in this context and the traditional procedures.

Keywords Singing voice separation · Dictionary learning · Incoherence · Sparse coding · Voice activity detector

✉ Samira Mavaddati
s.mavaddati@umz.ac.ir

[1] Electronic Department, Faculty of Technology and Engineering, University of Mazandaran, Babolsar, Iran

1 Introduction

A singing voice separation approach with high separation ability is applicable in many areas dealing with singer identification, voice and lyric recognition, and data retrieval [12, 13, 20]. This paper focuses on the singing voice separation process when a single microphone records the songs. The difficulties in this area are more prominent when the vocal signal is recorded in the presence of non-vocal or music accompaniment signal with high energy level. The purpose of this processing is to separate the vocal and non-vocal components of the recorded signals and remove the non-vocal parts. So, the proposed separation algorithm should be designed to increase either intelligibility or quality of the vocal signal without causing any distortion [22].

The singing voice signal includes two components: the components containing the vocal content (voice-only components) and the ones involving the non-vocal content (music accompaniments).

The voice or vocal signal is approximately sparse in the time–frequency domain. This domain provides a detailed analysis with high resolution for the vocal signals. The sparse representation algorithm models a voice frame with a linear combination of a limited number of atoms based on a dictionary learning procedure. On the other hand, the non-vocal signal can be assumed as a low-rank component, since its spectrum (time–frequency representation) in different time frames is highly correlated with each other.

Various singing voice separation algorithms are proposed using different basic approaches over the years such as autocorrelation-based [34], filter-based [10], pitch-based [15, 44], low-rank-based representation [39, 42, 43], cluster-based [24, 25], neural network-based [6, 19], and the approaches including nonnegative matrix factorization (NMF) [4].

The autocorrelation-based algorithm is only based on self-similarity measure. This algorithm works by extraction of the repeating musical components without need to prior training procedure. This method has the advantage of being simple, fast, and blind [34]. The filter-based method proposes a source/filter model to extract the musical accompaniment from the polyphonic audio signals. This unsupervised algorithm discriminates between the components of the singing voice signal by assuming that the energy of these parts is different [10]. The pitch-based algorithm introduced in [44] classifies an input signal into the vocal and non-vocal components. A pitch detection method detects the pitch frequency of the singing voice signal, and the detected pitch in the separation stage categorizes the time–frequency segments into the singing voice signal. In [15], a spectral subtraction method is applied to track the pitch frequency using the segmentation and the grouping stages. At first, the input singing voice signal is decomposed into small parts in different time–frequency resolutions. Then, the unvoiced parts are identified by Gaussian mixture models (GMM) as a classifier.

The low-rank-based separation methods assume that the magnitude spectrogram of a song can be considered as a superposition of low-rank and sparse components corresponding to the instrumental part and the vocal part, respectively

[39, 42, 43]. In [42], at first, the online dictionary learning is presented to model the vocal and instrumental parts of the clean signals. Then, the low-rank representations of these components are computed and the magnitude spectrogram is decomposed into two low-rank matrices to show the musical components. In [39], an online single-channel separation method is presented based on robust principal component analysis (RPCA). This method decomposes the signal into a low-rank component with its repetitive structure and a sparse component with its quasi-harmonic structure. A sparse and low-rank decomposition scheme using the RPCA technique is presented in [43] that works based on the harmonic similarity between the sinusoids components. The effectiveness of deep clustering on the task of singing voice separation is considered in the cluster-based methods [24, 25]. In [24], time–frequency representation of input sources is utilized to a stack of several recurrent layers, followed by a feed-forward layer to yield a time–frequency mask. Then, this mask is applied to the Mel frequency filter bank and the signal recovers using the inverse Fourier transform. In [25], it was shown that deep clustering on a singing voice separation task can outperform the conventional networks in the supervised and unsupervised conditions. Moreover, an optimized deep clustering and the conventional mask-inference networks are combined to classify the instrument and the vocal components.

Source separation procedures in [6, 19] are performed using deep convolutional neural networks (CNNs). In [19], U-Net architecture is employed to train two separate models for the extraction of the instrumental and vocal components of a signal in the time–frequency domain. A soft mask is obtained from the output of the final decoder layer of the U-Net and this output is multiplied element-wise with the spectrogram of the mixed signal to result in the final estimation of the components. Also in [6], a CNN is applied to estimate the time–frequency soft masks for source separation. The dimensional reduction process in the connected layer leads to a more compact representation of the input singing voice signal. This method only models the vocal components, while the instruments are used primarily to increase the dissimilarity with these components. In [4], the extracted features based on the temporal information are used to learn models with NMF. Then, a recurrent neural network is applied to capture long-term temporal dependencies in the singing voice signal without need to have temporal constraints.

In addition to the mentioned categories, there are other methods to solve the singing voice separation problem [11, 33, 35]. In [33], a priori probabilistic approach based on Bayesian modeling and statistical estimation is used to separate the mixed sources. Also, an expectation maximization algorithm optimizes the separation procedure. In [35], the periodically repeating frames in the singing voice signal are detected, and then, a comparison process with a repeating model is done via time–frequency masking. This algorithm can trace the pitch contour and works as a preprocessor in different signal processing fields. The U-net style network is applied in [11] to train the vocal and musical components. The vocal source can be captured from the mixture signal when the likelihood of the source from a given mixture is maximized. So, a decomposition scheme is proposed to maximize the likelihood by using implicit density [11].

In [16], the RPCA technique is employed to model the repetitive structure of the non-vocal components in the input signals [5]. In this unsupervised separation method, a binary time–frequency mask is estimated to earn the separation matrix with more accuracy. The algorithm presented in [17], similar to the RPCA-based method [5], can enhance the vocal signals corrupted by the background non-vocal signals. In this approach, separation of the monaural singing voice signal is performed in a supervised manner by using a deep recurrent neural network (DRNN). The different temporal connections of this network are learned using limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm. In [18], a singing voice separation approach in an interdependent manner based on the RPCA technique is presented. Using this technique, the contours of vocal fundamental frequency are estimated to make a time–frequency mask. Then, the decomposition process including RPCA procedure is performed using this mask.

In [31], a singing voice separation procedure using the empirical wavelet transform is presented to capture the repetitive structure of the non-vocal components and design a wavelet filter. The components related to the repetitive content of the non-vocal components are selected with the maximum frequency detected in the spectrum of singing voice signals.

In recent years, there is an increasing interest in using the sparse representation and dictionary learning techniques in the voice processing systems [30, 38]. The purpose of sparse coding is to approximately model the data frames as a weighted linear combination of a small number of the dictionary atoms [1]. A redundant dictionary is learned using the input signals to model each observed frame with the sparse linear combination of a fixed number of atoms [1, 9]. A trained dictionary that involves different bases with the unit norms in its column makes a generative model to represent precisely the content of input frame. In [38], least angle regression with the coherence criterion algorithm (LARC) followed by K-singular value decomposition (K-SVD) technique is utilized for dictionary learning. In [30], orthogonal matching pursuit (OMP) and limited-memory BFGS algorithms are used for sparse coding and dictionary training of speech and different noise signals.

In [27], a separation problem based on combination of sparse nonnegative matrix factorization (SNMF) and low-rank modeling is introduced. An incoherence generative model is learned for different components of the singing voice signal. Also, a factorization method using the model learned based on SNMF algorithm is utilized to decompose the sparse and low-rank parts of the singing voice signal.

In this paper, a new method for incoherent dictionary learning of the singing voice frames is introduced. For this purpose, the atom–data coherence and mutual coherence parameters should be considered. These parameters are adjusted in such a way that the approximation error is reduced as much as possible and the separation process is carried out with more accuracy [29]. In order to solve the singing voice separation problem using a dictionary-based approach, a new optimization method is presented based on the proposed learnable sparse and low-rank decomposition scheme and domain adaptation procedure. The vocal and non-vocal parts of the singing voice signal in the time–frequency domain are considered as sparse and low-rank components. In the training process, the vocal dictionary is learned on the clean singing voice signals. This dictionary is used in the next step to provide

enough data for the non-vocal dictionary learning process. The non-vocal atoms are learned based on the captured non-vocal frames using the proposed voice activity detector (VAD) algorithm. The presented energy-based VAD detects the non-vocal segments using the energy of coefficient matrix in the sparse coding of the input signal over a vocal dictionary. The RPCA technique is used in combination with the sparse representation over the vocal and non-vocal dictionaries to alternately solve the proposed optimization problem [28]. In the test step, the atoms in the learned non-vocal dictionary are adapted to the new ones according to the initial and final sections of the observed signal. This process is performed by using the domain adaptation technique which was previously used for speech enhancement in the presence of piano noise [7, 28, 30].

The remainder of this paper is organized as follows. In Sect. 2, the voice singing separation problem is illustrated. Then, in Sect. 3, a brief overview of the incoherent dictionary learning, domain adaptation technique, proposed voice activity detector, and sparse low-rank decomposition procedure is explained. The details of the presented decomposition model are addressed in Sect. 3. In Sect. 4, the experimental results are expressed. Finally, the evaluation results are provided in Sect. 5 and the paper is concluded in Sect. 6.

2 Problem Description

This paper introduces a novel singing voice separation algorithm based on the low-rank sparse decomposition scheme. The sparse and low-rank components of the singing voice signal are considered as vocal and non-vocal parts in the time–frequency domain, respectively. It should be noted that the vocal signal has no data redundancy in the time domain. Hence, this signal is usually transferred into another feature space to result in a better sparse representation. A suitable domain is the short-time Fourier transform (STFT). The vocal signal with speech content can be considered as the sparse component in the time–frequency domain. Also, the background music is regarded as a low-rank component since its frames are correlated in the time–frequency domain. Therefore, the singing voice signal is transferred into the STFT feature domain to display the low-rank sparse time–frequency representation. Monaural singing voice signal in the STFT domain can be linearly modeled as [21, 32]:

$$\boldsymbol{Y}(n,m) = \boldsymbol{S}(n,m) + \boldsymbol{L}(n,m) \tag{1}$$

where $\boldsymbol{Y}(n, m)$, $\boldsymbol{S}(n, m)$, and $\boldsymbol{L}(n, m)$ are the spectrograms of the singing voice, vocal and non-vocal signals at frequency bin n and the frame number m, respectively. The singing voice signal $\boldsymbol{Y} \in \mathbb{R}^{N\times M}$ can be represented linearly by the sparse coding of atoms as $\boldsymbol{Y}=\boldsymbol{DX}$, where $\boldsymbol{D} \in \mathbb{R}^{N\times P}, P > N$ is an overcomplete or redundant dictionary with P atoms shown by $\{\boldsymbol{d}_p\}_{p=1}^{P}$ with unit norm. The sparse coefficient matrix $\boldsymbol{X} \in \mathbb{R}^{P\times M}, P \gg K$ contains the sparse coefficients of $\boldsymbol{Y}$ with the cardinality parameter K [21, 32]. N indicates the number of frequency bins, and P denotes the number of dictionary atoms. Also, the value of cardinality parameter K determines that how

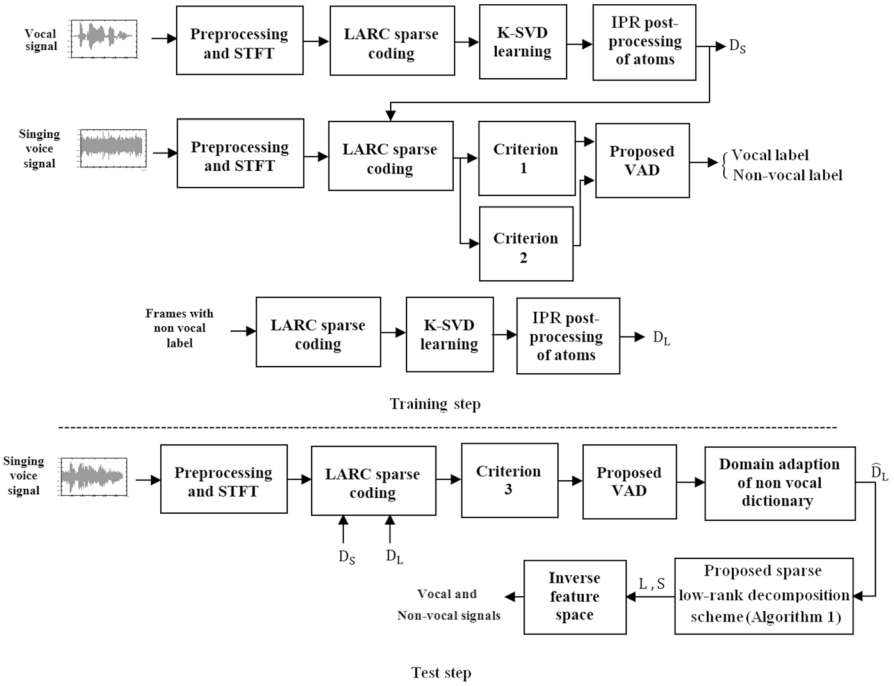

Fig. 1 Block diagram of the proposed separation procedure of the singing voice signal with the training and test steps

many atoms can participate in the representation of each input frame (each column of Y matrix). In fact, each column of X includes only K nonzero elements.

An overcomplete or redundant dictionary is a dictionary with more columns than its rows due to better representation of the data frames over the space bases. For example, a dictionary with twice the number of rows in its columns is named the overcomplete dictionary with a redundancy rate of 2. The sparse representation problem according to the approximation error and the sparsity constraint is formulated as [21, 32]:

$$\boldsymbol{X}^* = \arg\min_{X} ||\boldsymbol{Y} - \boldsymbol{DX}||_F^2 \quad \text{s.t.} \quad ||\boldsymbol{X}||_0 \leq K \tag{2}$$

where $||X||_0$ is the number of nonzero coefficients in each row of $\boldsymbol{X}$ bounded to the sparsity constraint K.

3 Overview of Proposed Method

The block diagram of the proposed method is shown in Fig. 1. The input signals in the training and test steps are first transferred into the STFT domain. The role of each block in Fig. 1 is illustrated with details in the following subsections.

3.1 Preprocessing of Input Data

In the first step of the preprocessing phase, the input signals are resampled from 16 kHz to 8 kHz. In the simulations, the input signal is divided into several frames to obtain stationarity on the data frame. The input data are segmented with 37.5 ms frame length and 40% overlap. Then, the pre-emphasis procedure is performed by applying the first-order difference equation [23]. Then, a Hamming window with 300 samples is applied to each frame [23]. Then, the feature space is achieved using a 300-point STFT.

3.2 LARC Sparse Coding and K-SVD Dictionary Learning

K-SVD is the first method to learn a redundant dictionary from a set of training data [1]. This learning algorithm is proposed for image denoising by learning an overcomplete dictionary. This flexible algorithm works easily with any sparse representation method. Using the K-SVD algorithm, each input data is modeled with the linear combination of the sparse coefficients of K atoms in a singular value decomposition procedure to provide a better fit with the training data. The idea of using sparse representation is one of the most interesting areas in different signal processing fields.

Two main parameters should be considered in order to learn a dictionary, the atom–data coherence, and the mutual coherence between the atoms. The first one determines the dependency between the dictionary atoms and the training data. If the value of the atom–data coherence is high, it will result in a better fitting between the dictionary atoms and the training data and then a lower reconstruction error. The second one expresses the dependency between the dictionary atoms and is obtained using the maximum value calculated from the correlations between the different atoms. The lower value for the mutual coherence results in a dictionary with independent basis vectors as much as possible. A coherence criterion is employed in the proposed dictionary learning process to yield the overcomplete dictionaries with the incoherent atoms.

In the first step of the dictionary learning procedure, the LARC sparse coding algorithm is utilized for the sparse representation based on the atom–data coherence constraint [38]. This algorithm is a generalization of the least angle regression method with the stopping condition including the residual coherence. In this approach, a variable cardinality parameter is set. The noisy signal in the enhancement step is coded sparsely over the composite dictionary as [38]:

$$\begin{aligned} &\boldsymbol{X}_S^*, \boldsymbol{X}_L^* = \mathrm{LARC}\left(\boldsymbol{Y}, \left[\boldsymbol{D}_S \boldsymbol{D}_L\right], \text{coherence value}\right) \rightarrow \\ &\boldsymbol{X}_S^*, \boldsymbol{X}_L^* = \arg\min_{\boldsymbol{X}_S, \boldsymbol{X}_L} ||\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}||_F^2 \rightarrow \arg\min_{\boldsymbol{X}_S, \boldsymbol{X}_L} \left\| \boldsymbol{Y} - \left[\boldsymbol{D}_S \boldsymbol{D}_L\right] \begin{bmatrix} \boldsymbol{X}_S \\ \boldsymbol{X}_L \end{bmatrix} \right\|_F^2 \end{aligned} \tag{3}$$

where $\boldsymbol{X}_S^*$ and $\boldsymbol{X}_L^*$ show the sparse and low-rank coefficients for representation of the input frame over the dictionaries $\boldsymbol{D}_S$ and $\boldsymbol{D}_L$. These dictionaries are related to the

vocal and non-vocal components. The obtained sparse coefficients of Eq. 3 are used in the reconstruction of the vocal and non-vocal frames as:

$$\hat{\boldsymbol{S}} = \boldsymbol{D}_S \boldsymbol{X}_S^*, \quad \hat{\boldsymbol{L}} = \boldsymbol{D}_L \boldsymbol{X}_L^* \tag{4}$$

The ability to maintain a balance between the confusion and distortion in the source signal is obtained by using a stop condition in the sparse coding process. This condition is based upon the atom–data coherence measure. A high value for the cardinality parameter or too dense sparse coding results in the source confusion since the number of dictionary atoms is not enough for proper representation. Source distortion occurs when the sparse coding is performed with a low cardinality parameter or too sparse coding. So, the number of required atoms will not be enough for sparse representation and input data cannot be coded exactly over these atoms. Therefore, the cardinality parameter must be set precisely [28].

3.3 IPR Post-processing of Atoms

A better matching between the input frames and dictionary atoms will be obtained when each dictionary has low coherence value similar to an equiangular tight frame (ETF). The ETF is a matrix in a Euclidean space with a set of unit vectors in its columns and the coherence value as small as possible [2]. The problem of finding a dictionary with low mutual coherence between its normalized atoms can be solved by analyzing Gram matrix $\boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}$. The coherence criterion is described by the maximum absolute value of the off-diagonal elements of the Gram matrix with the normalized atoms [2]. If all off-diagonal elements are the same, a dictionary with the minimum self-coherence value has been found. In the training phase of the presented algorithm, an iterative projection and rotation (IPR) method proposed in [2] is used as a post-processing step in order to yield the incoherent dictionaries. The IPR method was first introduced for the incoherent dictionary learning to represent the music signals [2]. In the proposed algorithm, the incoherent dictionary learning based on the K-SVD/LARC followed by IPR is applied to result in a lower coherence value than other dictionary learning algorithms such as K-SVD or K-SVD followed by LARC coding [28–30]. Due to the desirable results reported in [2], this dictionary learning method is used in this paper.

3.4 The Proposed Voice Activity Detector

In recent years, different VAD algorithms in various feature spaces were presented based on the dictionary learning technique [45, 46]. In [45], a dictionary-based VAD algorithm is introduced in the time domain including a detection process based on the average energies of different short and long segments of the input signal. In [46], a VAD algorithm is presented using an optimized dictionary learning procedure to yield the incoherent dictionaries. The features in this scheme are the modified version of the ones designed in [45]. Also, in [40], a VAD algorithm using nonnegative

sparse coding is presented in the STFT domain using the energy of sparse coefficients with a conditional random field (CRF) method.

In the proposed separation scheme, an energy-based VAD algorithm is introduced in the time–frequency domain using the incoherent K-SVD technique and LARC sparse coding algorithm.

In [30], our previous work for speech enhancement, a VAD algorithm was introduced based on the trained dictionaries with L-BFGS optimization algorithm in the wavelet packet transform domain. The modified version of this VAD algorithm in the time–frequency domain is employed in this paper. The proposed VAD scheme is detailed as follows [30].

First criterion In the first step of the proposed VAD algorithm, the input signals are transferred into the STFT domain. Then, a vocal dictionary $\boldsymbol{D}_S$ is learned by the K-SVD/LARC algorithm over the input frames involved only the vocal components. The presented VAD algorithm uses the energy of coefficient matrices calculated from the sparse representation of the singing voice signal over the learned vocal dictionary $\boldsymbol{D}_S$. Then, the singing voice frames $\mathbf{Y}$ are sparsely represented by $\boldsymbol{D}_S$ using LARC coding, and then, the similarity between the original frame $\mathbf{Y}$ and the reconstructed data frame $\hat{\mathbf{Y}}$ is calculated based on this representation:

$$\boldsymbol{X}_S^* = \mathrm{LARC}\big(\mathbf{Y}, \boldsymbol{D}_S, \text{coherence value}\big) \rightarrow \hat{\mathbf{Y}} = \boldsymbol{D}_S \boldsymbol{X}_S^* \tag{5}$$

where $\boldsymbol{X}_S$ is the sparse coefficient matrix coded by the LARC algorithm. If $||\mathbf{Y} - \hat{\mathbf{Y}}||_F^2 < \varepsilon_1$ or the approximation error $||\mathbf{Y} - \mathbf{D}_S \boldsymbol{X}_S^*||_F^2$ is low, the input frame will have a vocal label.

If $||\mathbf{Y} - \hat{\mathbf{Y}}||_F^2 > \varepsilon_1$, the input label will be non-vocal. A high value for this similarity measure means that the input frame has non-vocal structure since it has not been properly coded over the vocal dictionary.

Second criterion In order to ensure that our decision about the label of the input frame is true, the sparse representation is carried out over the initial and final frames of the input singing voice signal. These frames make a primary dictionary defined by $\mathbf{D}_{L0}$ that only contain the musical content due to the initial and final silence of the signal. The label of the captured frames, vocal or non-vocal tags, can be found due to the approximation error of the input frame over this dictionary. If the energy of the sparse coefficient matrix over $\mathbf{D}_{L0}$ is low, it means that the input frame does not have a proper representation on the non-vocal atoms captured from the initial and final frames of the singing voice signal. So, it indicates that the input frame has vocal content and will not be coded precisely on the atoms involved non-vocal structure. Hence, the correct label is assigned to the input frame:

$$\boldsymbol{X}_S^* = \mathrm{LARC}\big(\mathbf{Y}, \mathbf{D}_{L0}, \text{coherence value}\big) \rightarrow \mathbf{E}_0 = {}^1\!/_P \sum_{p=1}^{P} \boldsymbol{x}_{S,p}^{*2} \tag{6}$$

where $\boldsymbol{x}_S^*$ indicates the rows of the coefficient matrix $\boldsymbol{X}_S^*$ and P is the number of these rows. If the sparse coding matrix $\mathbf{E}_0$ has low coefficient energy, $|\mathbf{E}_0| < \varepsilon_2$, the vocal label assigned to the input frame in the previous step is true. Otherwise, the detected label will change to the new one.

If this representation is performed with high sparse coefficient energy, $|\mathbf{E}_0| > \epsilon_2$, the non-vocal label is assigned to the frame. If the detected label according to the first criterion was vocal, the estimated label will be changed to the new one. Therefore, a label will be assigned to each input frame using the results of the first and second criteria in the training step.

As mentioned, the second criterion emphasizes that the decision about the label of the input frame according to the first condition is true. If this label is specified as the vocal and the sparse representation of this frame over dictionary D_{L0} has low energy, this vocal label will be approved. Also, if the label of the input frame by applying the first condition is detected as the non-vocal and the sparse coding of this frame over dictionary D_{L0} has high energy, then the non-vocal label will be confirmed. Otherwise, if the estimated labels for the input frame in the first and second conditions are not the same, the label will be changed. In this paper, it is assumed that the eight initial and final frames of the input singing voice signal only include musical content.

Third criterion The last two criteria are utilized in the training step. In the test step, the dictionaries related to the sparse and low-rank components are available. Then, a voice activity detector scheme is proposed based on the energy of the coefficient matrices in the sparse representation of the observed data over a composite dictionary $[\mathbf{D}_S \mathbf{D}_L]$. The singing voice signal in the test step is coded sparsely over the composite dictionary. The detected frames captured by the proposed VAD scheme are learned over the composite dictionary as [38]:

$$\boldsymbol{X}_S^*, \boldsymbol{X}_L^* = \mathrm{LARC}\big(\boldsymbol{Y},\ [\boldsymbol{D}_S \boldsymbol{D}_L], \text{coherence value}\big) \rightarrow \arg\min_{\boldsymbol{X}_S, \boldsymbol{X}_L} \left\| \boldsymbol{Y} - [\boldsymbol{D}_S \boldsymbol{D}_L] \begin{bmatrix} \boldsymbol{X}_S \\ \boldsymbol{X}_L \end{bmatrix} \right\|_F^2 \tag{7}$$

where $\boldsymbol{X}_S^*$ and $\boldsymbol{X}_L^*$ show the sparse coefficients of the input data over $\boldsymbol{D}_S$ and $\boldsymbol{D}_L$ that indicate the vocal and non-vocal dictionaries. The energy of sparse coefficients is computed as:

$$\boldsymbol{E}_S = {}^1\!/_{P_1} \sum_{p_1=1}^{P_1} \boldsymbol{x}_{S,p_1}^{*2}, \quad \boldsymbol{E}_L = {}^1\!/_{P_2} \sum_{p_2=1}^{P_2} \boldsymbol{x}_{L,p_2}^{*2} \tag{8}$$

where P_1 and P_2 denote the rows of the coefficient matrices $\boldsymbol{X}_S^*$ and $\boldsymbol{X}_L^*$. If the energy of input frame in this representation over each dictionary is high, the related label according to the dictionary class is assigned to this frame.

The block diagram of the proposed VAD algorithm based on the mentioned criterions in the training and test step is illustrated in Fig. 2.

3.5 Domain Adaptation of Non-vocal Atoms

In the proposed test step, the domain adaptation technique is employed to adapt the non-vocal dictionary to the new one according to the characteristics of the test signal. This adaptation process was first applied as an analytical solution for image denoising to update the learned atoms [7]. Then, a low-rank sparse

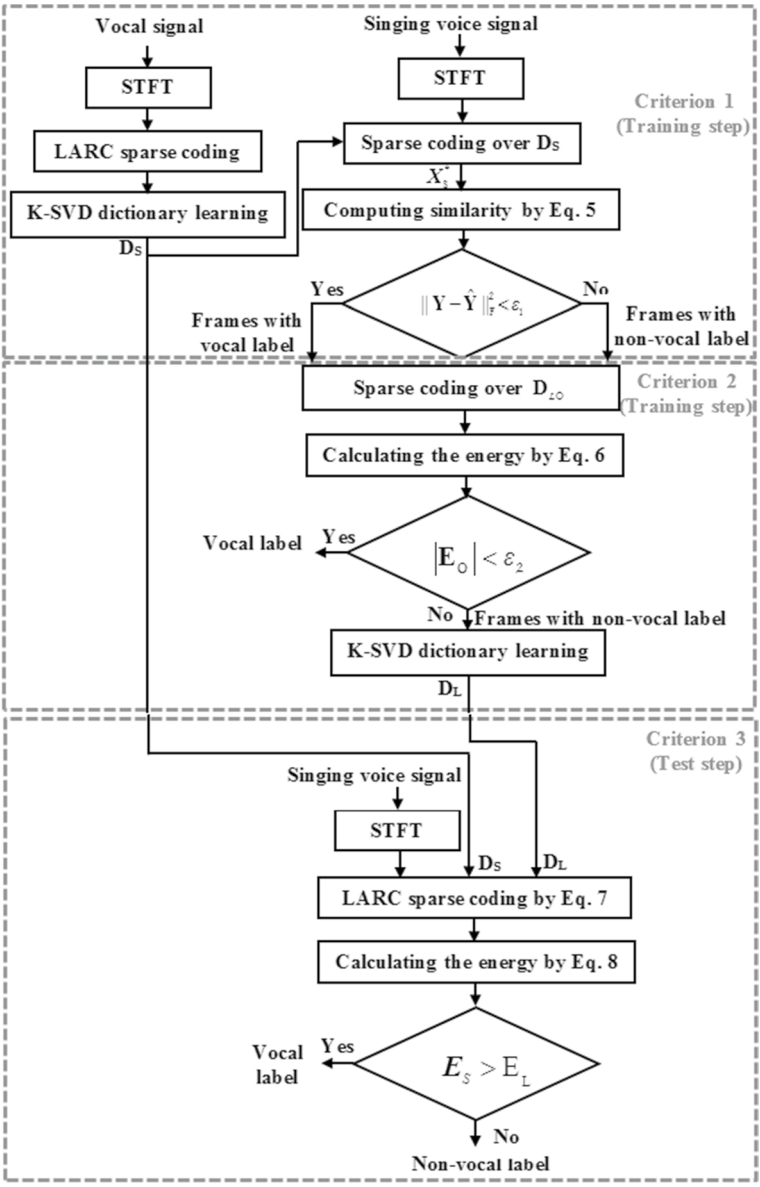

Fig. 2 Block diagram of the proposed VAD procedure based on three criterions in training and test steps

decomposition scheme was introduced in [28] to reduce the mismatch between the characteristics of the training and test signals and enhance the speech frames. In [30], the adaptive supervised and semi-supervised speech enhancement algorithms were presented using the sparse coding in the wavelet packet transform (WPT) domain. Also, the dictionary adaptation technique is applied to update the

atoms of the primary noise dictionary according to the estimated noise in the test environment. It has been shown in [28, 30] that this technique is very effective to achieve better enhancement results especially in the presence of stationary or non-stationary noise signals.

In this paper, the domain adaptation procedure is used to update the atoms of the non-vocal dictionary based on the non-vocal frames captured by the test signal. These frames are estimated using the proposed VAD algorithm defined in Sect. 3.4. Hence, the adapted atoms can sparsely code the observed data based on the characteristics of the non-vocal signal in the test space with low approximation error. This adaptation procedure leads to an effective factorization process especially when the non-vocal components in the training and test steps have the different structure. In fact, the mismatch between the training and test signals has been alleviated based on the atom adaptation technique as much as possible.

3.6 Proposed Decomposition Scheme

The sparse and low-rank components of the singing voice signal in the time–frequency domain are considered as the vocal and non-vocal components, respectively. Some methods ignore the effect of using knowledge about the statistics of the input data, while using the learned models with these statistics leads to a better performance, especially in the singing voice separation problem. Therefore, an alternating decomposition algorithm involved the sparse and low-rank models of the input components is used to solve this separation problem. It is noteworthy that employing incoherent vocal and non-vocal dictionaries as expressed in Sect. 3.2 results in a proper separation process based on the content of the input data.

The basic algorithm for sparse low-rank decomposition is RPCA that uses an alternating projection algorithm to decompose the data structure by setting the constraints on the rank and sparsity parameters for each observed signal [5]. In the RPCA algorithm, the sparse and low-rank components are obtained from hard thresholding based on a shrinkage function and singular value decomposition (SVD) of the observed signals, respectively [5]. The RPCA as a convex optimization algorithm recovers a low-dimensional matrix from high-dimensional observations and decomposes the input signals into the sparse and low-rank components [5]. In recent years, RPCA is used in the different fields of voice processing such as singing voice separation problem [16]. This technique solves the following convex optimization problem since the vocal and non-vocal components of the singing voice signal can be represented by sparse and low-rank components in the time–frequency domain [5, 28]:

$$\arg\min\left(||\boldsymbol{L}||_{*}+\lambda||\boldsymbol{S}||_{1}\right),\quad \text{subject to}\quad \boldsymbol{Y}=\boldsymbol{L}+\boldsymbol{S} \tag{9}$$

where $||.||_{*}$ defines the nuclear norm as the sum of the singular values and shows that $\boldsymbol{L}$ should be low rank. Different Lagrangian-based or projection-based techniques have been proposed to solve these subproblems [5, 28]:

$$\begin{aligned} \boldsymbol{L}^t &= \underset{\text{rank}(L)\le r}{\arg\min} \, ||\boldsymbol{Y} - \boldsymbol{L} - \boldsymbol{S}^{t-1}||_F^2 \\ \boldsymbol{S}^t &= \underset{\text{Card}(S)\le k \& S\ge 0}{\arg\min} \, ||\boldsymbol{Y} - \boldsymbol{L}^t - \boldsymbol{S}||_F^2 \end{aligned} \tag{10}$$

where $||.||_{\text{F}}$ is the Frobenius norm. Also, rank(.) denotes the rank of the low-rank components bounded to $r \le \min(N, B)$ for $\boldsymbol{L} \in \mathbb{R}^{\text{N}\times\text{B}}$. Card(.) shows the cardinality value of each sparse component.

In this paper, the RPCA method is combined with the sparse coding of the vocal and non-vocal dictionaries (trained using K-SVD algorithm according to Sect. 3.2) to alternately solve the optimization problem [28]. The nonnegative coefficients in this optimization procedure are adjusted during the sparse representation step without adding the new parameters to prevent from a complicated learning approach. The separation problem can be formulated as [28]:

$$\arg\min \left(||\boldsymbol{D}_L \cdot \boldsymbol{X}_L||_* + \lambda||\boldsymbol{D}_S \cdot \boldsymbol{X}_S||_1\right), \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{D}_L \cdot \boldsymbol{X}_L + \boldsymbol{D}_S \cdot \boldsymbol{X}_S \tag{11}$$

The following subproblems are obtained to alternately solve the separation problem and update the sparse matrices by applying the vocal and non-vocal dictionaries in Eq. 9 [28]:

$$\begin{aligned} &\underset{\boldsymbol{X}_L}{\arg\min} \left(\frac{1}{2}||Y - \boldsymbol{D}_L\boldsymbol{X}_L - S||_F^2 + \lambda_{\boldsymbol{X}_L}||\boldsymbol{X}_L||_1 + \lambda_{\boldsymbol{L}}||\boldsymbol{D}_L\boldsymbol{X}_L||_*\right) \\ &\underset{\boldsymbol{X}_s}{\arg\min} \left(\frac{1}{2}||Y - \boldsymbol{L} - \boldsymbol{D}_S\boldsymbol{X}_S||_F^2 + \lambda_{\boldsymbol{X}_S}||\boldsymbol{X}_S||_1\right) \end{aligned} \tag{12}$$

where $\lambda_{\text{x}_\text{L}}$ and $\lambda_{\text{x}_\text{S}}$ are the weighting factors for the sparsity constraints of the coefficient matrices. Also, λ_{L} indicates the rank value of the non-vocal components. The analytical solution of these subproblems using the SVD algorithm and hard thresholding of the low-rank and sparse components is proposed in [28]. The proposed sparse low-rank decomposition algorithm based on the learned dictionaries for the vocal and non-vocal signals is shown in Algorithm 1 [28].

4 Experimental Results

The redundancy rates of the overcomplete dictionaries for the vocal and non-vocal data have been set to 4 and 2, respectively. The dictionary learning for the non-vocal frames with the periodic content is performed usually with low approximation error. This means that the input data with the harmonic structure can be represented exactly by the trained atoms. For this reason, the lower redundancy rate is sufficient for this data type.

These experiments are carried out in two situations: singer-dependent (SD) and singer-independent tests (SI). In the SD situation, the singers in the training and test steps are the same. Since the vocal and non-vocal frames have very distinct structure, the non-vocal frames in both SD and SI scenarios do not have an adequate sparse coding over the vocal dictionary. Therefore, these frames are coded by the low-rank atoms in the composite dictionary. The threshold

values in the criteria 1–2 of the proposed VAD algorithm, ε_1 and ε_2, are set to 0.15 and 0.25, respectively. Also, the number of iteration (itr), coherence value (C), threshold value (T), and rank value (r) defined in Algorithm 1 are adjusted according to the experimental simulation results to 20, 0.025, 0.1, and 2, respectively. In the training phase, only the magnitude spectrums of input signals have been used in the learning process. The phase spectrum is kept unchanged during the synthesis.

In the test phase, each singing voice signal is mixed with the background music at SNRs of -5 dB, 0 dB, and 5 dB. The BSS-EVAL toolbox is used to assess the performance of the proposed algorithm [41].

Algorithm 1: Proposed sparse low-rank decomposition algorithm based on the learned dictionaries

Input: D_S, D_L, Y, K (cardinality for LARC), *itr* (number of iteration), C (coherence value), T (threshold value), r (rank value)

Output: X_S, X_L, S, L

Initialization: $X_S^* = [0]$, $X_L^* = [0]$

for t=1 → *itr*

% Update low-rank component L and related coefficient matrix X_L

$$X_L^* = (D_L^T . D_L)^{-1} \left(D_L^T \left(Y - D_S . X_S^t \right) \right)$$

$$UAV = SVD\left(D_L . X_L^* \right)$$

$$\mathrm{L}^t = \sum_{i=1}^{r} \lambda_i \mathrm{U}_i \mathrm{V}_i$$

$$X_L^{t+1} = LARC\left(D_L, L^t, C, K \right)$$

$$\mathrm{L}^{t+1} = \mathrm{D}_L . \mathrm{X}_L^{t+1}$$

% % Update sparse component S and related coefficient matrix X_S

$$X_S^* = (D_S^T . D_S)^{-1} \left(D_S^T \left(Y - D_L . X_L^{t+1} \right) \right)$$

$$S^t = HardThreshold\left(D_S . X_S^*, T \right)$$

$$X_S^{t+1} = LARC \quad D_S, S^t, C, K$$

$$S^{t+1} = D_S . X_S^{t+1}$$

t=t+1;

end for

All dictionaries are initialized with the training data chosen randomly from the input frames. Then, the learned dictionaries with the decorrelated atoms are obtained by applying the IPR algorithm as mentioned in Sect. 3.3. All framing, preprocessing, and redundancy rates are the same in the training and test steps.

5 Evaluation

The several experiments have been performed to evaluate the performance of the proposed algorithm using the MIR-1K corpus.[1] This large multi-talker database is provided with recordings of 1000 songs for research on the singing voice separation problem. This corpus involves the songs recorded by 19 singers of both genders with sentences about 4–13 s. In the implementations, four male and five female singers in the training step and three male and four female singers for the singer-independent test have been selected. The training and test sets include 400 and 200 songs.

In this paper, the effect of the dictionary learning technique in the singing voice separation problem is considered. The performance of the proposed method is compared with some baseline methods and the earlier algorithms in this context. The reported results are obtained from averaging over all test signals. In order to assess the simulation results with more details, the proposed approach is compared with the methods introduced in [16, 17, 39, 42] and my previous research presented in [27].

The MLRR-based (multiple low-rank representation) method proposed in [42] uses 1024-point STFT by sliding a Hamming window with 25% overlap to obtain the spectrogram. Then, the magnitude and the phase part of spectrogram are used in the separation procedure. The spectrogram of each frame in [39] is computed using a window size of 1024 with 40% overlap. A gradient descent method is used to train dictionaries used in the robust low-rank nonnegative matrix factorization (RNMF) procedure. In the RPCA-based separation algorithm proposed in [16], the spectrogram is calculated using a window size of 1024 with 40% overlap. Then, the inexact augmented Lagrange multiplier (ALM) method is applied to solve the RPCA algorithm. In [17], a DRNN is used with three hidden layers of 1000 hidden units with the mean squared error criterion and joint masking training. The spectral representation is extracted using a 1024-point STFT with 50% overlap, and the magnitude of spectra is considered as input features to the neural network. This neural network is optimized by back-propagating the gradients with respect to train the objectives. The limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm is used to train the models from random initialization, and the maximum epoch is set to 400. In [27], a monaural voice separation method is presented based on SNMF and low-rank modeling in order to represent the vocal and non-vocal components of sound mixtures. In this paper, a factorization procedure is designed to reduce the approximation error and result in a separation process with more accuracy. The magnitude of signal spectrums has been employed in the learning process, and the phase of signals is kept unchanged during the synthesis procedure. The rank value in SNMF algorithm is selected based on the experimental results and high correlation between the consecutive frames of the background

[1] https://sites.google.com/site/unvoicedsoundseparation/mir-1k.

music. The unsupervised version of this algorithm is utilized in the simulation results. This method in the simulation results of this paper is indicated with ((SNMF)) algorithm.

All of these methods used to compare the performance of the proposed algorithm are evaluated on MIR-1K dataset [16, 17, 27, 39, 42].

The assessment measures employed in the simulations are frequency-weighted segmental SNR (fwSegSNR), perceptual evaluation of speech quality (PESQ score), speech distortion index (SDI) [26], and global normalized source-to-distortion index (GNSDI) in the following of the evaluation frameworks performed in [7, 29].

The fwSegSNR as a speech quality assessment measure is similar to the segmental SNR with an additional averaging over the frequency bands corresponding to the ear's critical bands [14]. PESQ score as an objective measure uses a perceptual model to estimate the mean opinion score that is a subjective criterion and determines the speech intelligibility [26]. Another instrumental measure to evaluate the performance of a speech signal is SDI [3]. This measure quantifies the non-vocal part included in the separated vocal signal and is defined as:

$$\mathrm{SDI}\left(S(t),\hat{S}(t)\right)=\frac{E\left\{\left[S(t)-\hat{S}(t)\right]^2\right\}}{E\left\{S^2(t)\right\}} \tag{13}$$

where $s(t)$ and $\hat{s}(t)$ are the initial vocal signal and its estimation captured from the separation algorithm at the sampling time index t, respectively. The GNSDI is computed by averaging over all test signals as [16]:

$$\mathrm{GNSDI}\left(S(t),\hat{S}(t),Y(t)\right)=\frac{\sum_{i=1}^{I}a_i\left[\mathrm{NSDI}\left(S_i(t),\hat{S}_i(t)\right)-\mathrm{NSDI}\left(S_i(t),Y_i(t)\right)\right]}{E\left\{S_i^2(t)\right\}} \tag{14}$$

where I is the total number of the test songs and a_i is the weighted coefficient corresponding to the length of the related test signals. The NSDI denotes the normalized source-to-distortion ratio and is defined as

$$\mathrm{NSDI}\left(S(t),\hat{S}(t),Y(t)\right)=\mathrm{SDI}\left(\hat{S}(t),S(t)\right)-\mathrm{SDI}(S(t),Y(t)) \tag{15}$$

The NSDI value determines the improvement of the SDI parameter between the observed signal $Y(t)$ and the estimated singing voice signal $\hat{s}(t)$ [16]. The SDI parameter can be calculated from Eq. 13.

A separation process with more quality is yielded when the fwSegSNR, PESQ, and GNSDI measures have higher values. Also, the lower values for SDI parameter achieve better separation results. In this paper, the MATLAB implementations of the fwSegSNR and PESQ measures provided by [23, 36] are used. The results are the average of all test signals in the mentioned conditions. The MATLAB software on a Windows 64-bit-based computer with Core i5 3.2 GHz CPU is employed for the training and test steps.

Table 1 Results of the atom coherence, atom–data coherence, and SNR values of the data approximation for the vocal and non-vocal signals

	Vocal signal	Non-vocal signal
K-SVD/OMP [1]		
Atom coherence	0.89	0.90
Atom–data coherence	0.48	0.54
SNR (dB)	10.2	10.5
K-SVD/LARC [38]		
Atom coherence	0.87	0.89
Atom–data coherence	0.69	0.78
SNR (dB)	10.8	11.1
K-SVD/LARC/IPR (proposed)		
Atom coherence	**0.49**	**0.54**
Atom–data coherence	**0.74**	**0.79**
SNR (dB)	**11.4**	**11.7**

The best results obtained from different algorithms are highlighted in bold

5.1 Results

As described in Sects. 3.2 and 3.3, the K-SVD/LARC dictionary learning process, employed in this paper, increases the data–atom coherence. This procedure considers the residual coherence parameter between each data frame and the learned atoms. Also, an overcomplete dictionary with minimum coherence value between its learned atoms is obtained using the IPR technique followed by the atom correction step. The obtained coherence values are reported in Table 1 for different dictionary learning algorithms. The best results obtained from different algorithms are highlighted in bold. These algorithms involve the dictionary learning using K-SVD with OMP sparse coding [1], K-SVD with LARC coding [38], and the proposed incoherent dictionary scheme based on the K-SVD with LARC sparse coding followed by IPR post-processing. The reported SNR values are the ratio of the input matrix $\mathbf{Y}$ to the reconstruction error of sparse coding over the dictionary $\mathbf{D}$:

$$\mathrm{SNR}(\mathbf{Y}, \mathbf{D}X) = 20\log_{10}\left(||\mathbf{Y}||_F^2 / ||\mathbf{Y} - \mathbf{D}X||_F^2\right) \tag{16}$$

As shown in Table 1, the calculated atom–data coherence and SNR values for the K-SVD/LARC training procedure are higher than the K-SVD/OMP method, since the matching between the dictionary atoms and the training data increases based on the LARC coding and its variable cardinality parameter. Also, the IPR technique has been used in the training step to decorrelate the learned atoms and yields a dictionary closer to the ETF along with decreasing the approximation error. Therefore, the atom coherence values obtained in the training step are lower than the first two mentioned training algorithms. The separation results measured by the PESQ and segmental SNR scores at different SNR values are given in Figs. 3 and 4. These results are obtained for the SD and SI scenarios of the proposed method in comparison with other mentioned algorithms.

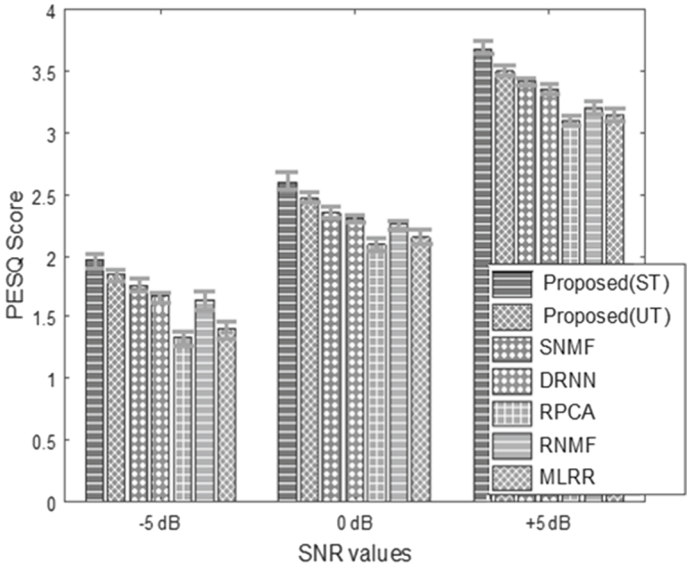

Fig. 3 Performance comparison of different methods in terms of PESQ scores at different SNR values for separation of the vocal component

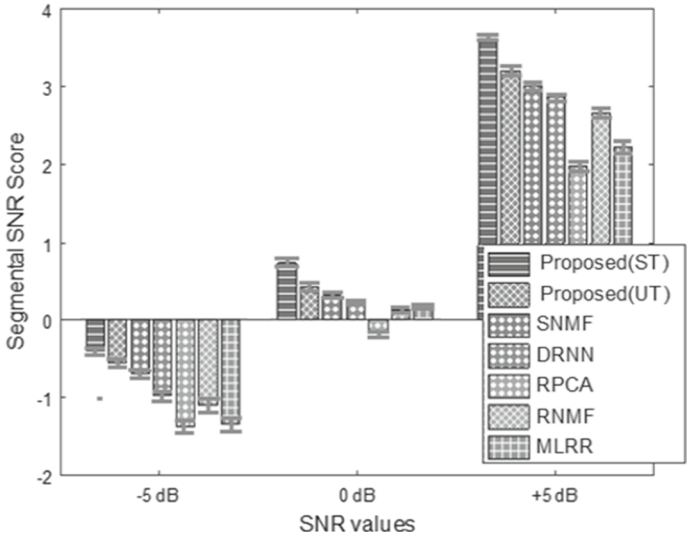

Fig. 4 Performance comparison of different methods in terms of frequency-weighted segmental SNR values at different SNRs for separation of the vocal components

In order to better show the calculated measures for the mentioned approaches, the results of SDI and GNSDI are reported in Figs. 5 and 6, respectively. These results are obtained from the averaging over all test signals. Also, the results shown

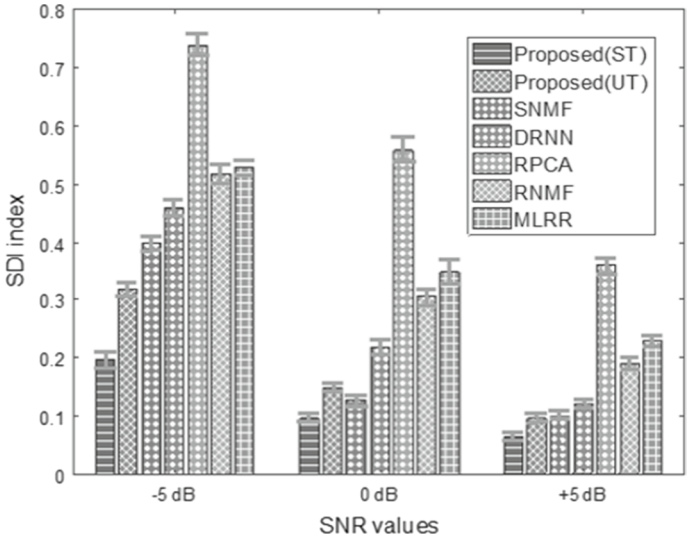

Fig. 5 Performance comparison of different methods in terms of SDI measure at different SNR values for separation of the vocal components

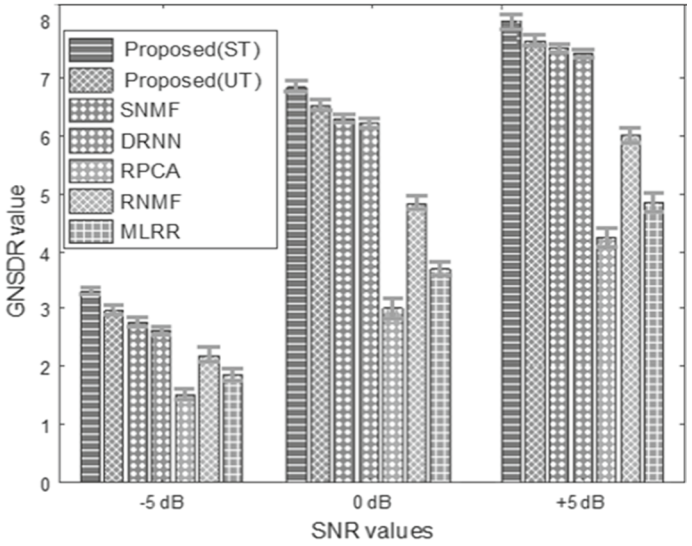

Fig. 6 Performance comparison of different methods in terms of GNSDI value at different SNR values for separation of the vocal components

in Figs. 3 and 6 are reported in Tables 2 and 3 in order to have better representation with more details for a proper comparison between the algorithms.

As explained, more quality in the separation process is obtained when the fwSeg-SNR, PESQ, and GNSDI scores have high values. In terms of SDI measure, lower

Table 2 Results of separation for the vocal components using PESQ score and fwSegSNR measure at SNRs of −5, 0, and 5 for different methods

	PESQ			fwSegSNR		
	−5 dB	0 dB	+5 dB	−5 dB	0 dB	+5 dB
Proposed (SD)	**1.98**	**2.61**	**3.68**	**−0.41**	**0.73**	**3.63**
Proposed (SI)	1.86	2.48	3.50	−0.56	0.42	3.21
SNMF [27]	1.77	2.36	3.42	−0.70	0.31	3.01
DRNN [17]	1.73	2.31	3.35	−0.98	0.22	2.87
MLRR [42]	1.65	2.27	3.20	−1.11	0.12	2.66
RNMF [39]	1.42	2.16	3.14	−1.35	0.18	2.23
RPCA [16]	1.35	2.11	3.10	−1.38	−0.18	1.98

The best results obtained from different algorithms are highlighted in bold

Table 3 Results of separation for the vocal components using SDI and GNSDI measure at SNRs of −5, 0, and 5 for different methods

	SDI			GNSDI		
	−5 dB	0 dB	+5 dB	−5 dB	0 dB	+5 dB
Proposed (SD)	**0.24**	**0.098**	**0.065**	**3.35**	**6.86**	**7.99**
Proposed (SI)	0.32	0.15	0.096	3.01	6.54	7.62
SNMF [27]	0.40	0.13	0.10	2.82	6.31	7.51
DRNN [17]	0.46	0.22	0.12	2.67	6.23	7.40
MLRR [42]	0.52	0.31	0.19	2.23	4.85	6.01
RNMF [39]	0.53	0.35	0.23	1.89	3.71	4.85
RPCA [16]	0.74	0.56	0.36	1.56	3.04	4.23

The best results obtained from different algorithms are highlighted in bold

values describe better separation results and less distortion in the captured vocal signal. As can be seen, the proposed method in the singer-dependent situation (with similar singers and different lyrics in the training and test steps) achieves better measure values than other approaches that are based on the low-rank modeling (SNMF, RNMF, MLRR), neural network (DRNN), and RPCA techniques at all SNR conditions [15, 19, 24, 42]. Therefore, it can be concluded that the proposed method outperforms the mentioned algorithms in the singing voice separation problem.

The purpose of singing voice separation algorithm is voice recognition, voice retrieval, singer identification, or lyric recognition. In these entire processing fields, the quality of the separated vocal signal is important. In fact, the performance of the separation algorithm presented is evaluated based on the captured vocal signal, and not on the extracted non-vocal component. A precise separation of the vocal signal spectrum leads to an appropriate estimation for the non-vocal components. Therefore, in order to have more investigation about the separation results, the mentioned evaluation measures have been applied to the extracted non-vocal signals and the performance of the proposed algorithm for estimation of these components has been assessed. The results of this simulation for different methods using SDI and fwSegSNR measures at different SNR values are reported in Table 4. The obtained results are consistent with those reported in Figs. 3 and 6 and Tables 2 and 3 and show that

Table 4 Results of separation for the non-vocal components using SDI and fwSegSNR measures at SNRs of −5, 0, and 5 for different methods

	SDI			fwSegSNR		
	−5 dB	0 dB	+5 dB	−5 dB	0 dB	+5 dB
Proposed (SD)	**0.28**	**0.10**	**0.08**	**−0.46**	**0.68**	**3.54**
Proposed (SI)	0.36	0.19	0.15	−0.58	0.60	3.19
SNMF [27]	0.41	0.23	0.18	−0.82	0.43	2.91
DRNN [17]	0.47	0.28	0.21	−1.12	0.31	2.85
MLRR [42]	0.53	0.40	0.29	−1.25	0.17	2.34
RNMF [39]	0.58	0.48	0.36	−1.36	0.02	2.18
RPCA [16]	0.79	0.63	0.41	−1.45	−0.33	1.84

The best results obtained from different algorithms are highlighted in bold

the proposed decomposition scheme can outperform other presented algorithms in this context at different SNR values. This superiority is more prominent at lower SNR values.

The effectiveness of the dictionary learning algorithm to remove the background music from the speech signal has been proven in [28–30]. The reported results in these references emphasize on the prominent role of the learning-based technique that can be employed in the singing voice separation problem. In these noise separation procedures, the best results were achieved in the presence of piano noise that is a periodic signal [28–30].

The proposed method can separate properly the vocal signals from the background non-vocal components since the non-vocal signals made from any musical instruments are well structured and can train the generative incoherent dictionaries. Therefore, the non-vocal parts of the singing voice signals are sparsely modeled by the trained atoms with more accuracy.

It is obvious that the quality of the separated vocal signals in the SD test is better than the SI situation. This is due to this fact that the dictionary learning in the SD test allows us to have the vocal atoms with more coherent in the training and test scenarios. Therefore, a sparse coding with lower approximation error and the separation procedure with more quality attains.

In Figs. 3 and 6, the reason of better results in the SI situation is that although the vocal content in the test step is not precisely found in the vocal dictionary atoms, the adapted non-vocal dictionary using a variable sparsity value and the incoherent atoms can exactly model any background non-vocal component. Hence, the non-vocal segments of the observed frames are coded in the corresponding transferred non-vocal dictionaries. Also, coding the vocal components by the related dictionary with variable cardinality value has a main effect in the proposed optimization process.

In the dictionary learning process, achieving incoherent atoms is very important. In the incoherent dictionary learning, the vocal and non-vocal atoms are incoherent to each other and the variable cardinality parameter in the LARC coding effectively prevents from sparse coding of the structured non-vocal components over the vocal dictionary. So, the cardinality value in this algorithm is limited to the number of columns in the learned dictionary.

Table 5 Results of separation for the vocal component using PESQ score and fwSegSNR measure at SNRs of −5, 0, and 5 for the proposed SD and SI scenarios with and without domain adaptation step

	PESQ			fwSegSNR		
	−5 dB	0 dB	+5 dB	−5 dB	0 dB	+5 dB
SD scenario	**1.98**	**2.61**	**3.68**	**−0.41**	**0.73**	**3.63**
SI scenario	1.86	2.48	3.50	−0.56	0.42	3.21
SD without domain adaptation step	1.83	2.49	3.53	−0.54	0.62	3.55
SI without domain adaptation step	1.77	2.36	3.41	−0.64	0.38	3.10

The best results obtained from different algorithms are highlighted in bold

Table 6 Results of separation for the vocal component using SDI and GNSDI measures at SNRs of −5, 0, and 5 for the proposed SD and SI scenarios with and without domain adaptation step

	SDI			GNSDI		
	−5 dB	0 dB	+5 dB	−5 dB	0 dB	+5 dB
SD scenario	**0.24**	**0.098**	**0.065**	**3.35**	**6.86**	**7.99**
SI scenario	0.32	0.15	0.096	3.01	6.54	7.62
SD without domain adaptation step	0.34	0.13	0.092	3.25	6.71	7.73
SI without domain adaptation step	0.39	0.23	0.16	2.74	6.24	7.46

The best results obtained from different algorithms are highlighted in bold

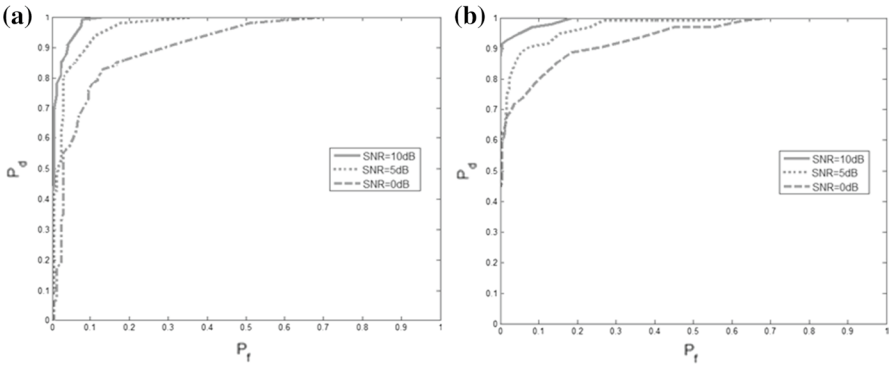

Fig. 7 ROC curves of the proposed VAD at different SNR values for: **a** detection of vocal activity, **b** detection of non-vocal activity

To evaluate the role of domain adaptation step, the performance of the proposed vocal separation method in the two defined scenarios, SD and SI, is considered with different assessment measures. These results are shown in Tables 5 and 6 based on the PESQ, fwSegSNR, SDI, and GNSDI measures. The reported results show that the proposed singing voice separation approaches obtain better results than these scenarios without the domain adaptation step. Therefore, it can result that the atom adaptation step has a prominent effect in the presented separation process.

In order to assess the performance of the proposed VAD algorithm, receiver operating characteristic (ROC) curves are employed. The P_d and P_f in this evaluation measure indicate the probability of the vocal/non-vocal label detections and the probability of false alarm, respectively. In this experiment, the performance of the presented VAD at SNRs of 0 dB, 5 dB, and 10 dB is investigated. These results are shown in Fig. 7. As expected, the proposed VAD yields a better accuracy at high SNR values. The ROC curves show that the proposed method can detect precisely the labels related to the input vocal and non-vocal frames.

In order to have more evaluations, the statistical test is performed to assess the performance of the presented algorithm in different conditions. This test should be performed when the implementations consist of different conditions and various compared methods. This test shows that whether there is any significant difference between the compared methods or not. These conditions contain different methods, measurements, and SNR values in the two mentioned scenarios. For this purpose, the nonparametric Friedman test with the Holms post hoc test without any initial assumption is applied to perform the statistical significance test for the estimated accuracy values attained from more than two algorithms [8, 37].

In this test, the results with more accuracy will be obtained when the number of conditions is sufficiently bigger than the number of methods. So to increase the number of conditions, this test was performed in two situations. Firstly, for the results of PESQ and SDI scores and then for fwSegSNR and GNSDI measures, these two classes of criteria contain the same value range. $R_j = \frac{1}{I}\sum_{i=1}^{I} r_{ij}$ denotes the average performance rank of each measure for the jth approach out of J methods evaluated on I conditions. In this test, the approaches with better performance will have lower rank values [8]. In our simulations, the number of methods and different conditions in each situation is $J=5$ and $I=24$, respectively. This statistical test begins with a null hypothesis test that means all the methods have the same quality. It is possible that this hypothesis is accepted or rejected during this test. The original Friedman test is explained as:

$$\chi_F^2 = \frac{12I}{J(J+1)}\left[\sum_{j=1}^{J} R_j^2 - \frac{J(J+1)^2}{4}\right] \tag{17}$$

Also, a modified statistic of the Friedman test F_F based on F-distribution with $(J-1)$ and $(J-1)\times(I-1)$ degrees of freedom is defined as $(I-1)\chi_F^2/\left(I(J-1)-\chi_F^2\right)$ [8].

The null hypothesis will be rejected if F_F is greater than the critical value of χ_F^2. A post hoc test can be performed in order to determine which algorithm has better performance [8]. In this test, $\mathcal{Z}_j = \left(R_0 - R_j\right)/\sqrt{J(J+1)/6I}$ is calculated for each assessment method. R_0 is related to the method with the best average rank or the worst performance. The ρ-value at the statistical significance level $\alpha = 0.05$ is calculated by using the area under the standard normal distribution and outside of the range $(-\mathcal{Z}, \mathcal{Z})$. The simulation results of Friedman test for different singing voice separation methods, SD and SI, three SNR values, PESQ, and SDI scores are reported in Table 7. Also, these results for fwSegSNR and GNSDI measurements are expressed in Table 8.

Table 7 PESQ and SDI scores in the statistical Friedman test with the Holms post hoc test for different methods and various conditions

Methods	Average rank (R_j)	$\mathcal{Z}$	ρ-value	Holm($\alpha/(J-i)$)
Proposed (SI)	**1.09**	**7.4321**	**0**	**0.0071**
SNMF [27]	**1.83**	**6.3121**	**0**	**0.0083**
DRNN [17]	2.23	5.6922	0	0.0100
MLRR [42]	2.97	3.7122	0.0002	0.0125
RNMF [39]	3.21	2.8142	0.0012	0.0167
RPCA [16]	3.87	2.4356	0.0149	0.0250

The best results obtained from different algorithms are highlighted in bold

Table 8 fwSegSNR and GNSDI measures in the statistical Friedman test with the Holms post hoc test for different methods and various conditions

Methods	Average rank (R_j)	$\mathcal{Z}$	ρ-value	Holm($\alpha/(J-i)$)
Proposed (SI)	**1.02**	**7.7125**	**0**	**0.0071**
SNMF [27]	**1.92**	**6.8210**	**0**	**0.0083**
DRNN [17]	2.56	6.0351	0	0.0100
MLRR [42]	2.80	3.6587	0.0003	0.0125
RNMF [39]	3.54	2.7101	0.0067	0.0167
RPCA [16]	3.89	2.4011	0.0163	0.0250

The best results obtained from different algorithms are highlighted in bold

The values of χ_F^2 for the obtained results in Tables 7 and 8 are 56.5, 63.1, respectively. Also, the values of F_F in these tables are 24.35 and 33.64. The critical F-value with (5-1) and (5-1)$\times$(24-1) degrees of freedom is equal to 2.7955. Therefore, the initial null hypothesis is rejected and a post hoc test can be used to compare different separation methods with each other since the obtained value F_F is greater than the critical F-value. If the Holm's critical value calculated for each method is greater than the corresponding ρ-value, it can be concluded that this approach has better performance than other algorithms as listed in the statistical test table [37].

The algorithm with the highest $\mathcal{Z}$ value has the best performance since the difference between R_0 and R_j in the numerator of its formula will be bigger.

The average rank close to 1 reported in Tables 7 and 8 illustrates the best average rank among other algorithms. These results show that the proposed method performs statistically better than the other mentioned algorithms in all conditions. The RPCA method related to R_0 has the lowest performance with the average rank of 3.88 in Tables 7 and 8.

In order to have more evaluations about the performance of the proposed method, the spectrogram plots of the mentioned singing voice separation algorithms are considered. The lyric "abjones_1_01" of MIR-1 K database was mixed by the background music at SNRs of -5 dB, 0 dB, and 5 dB. The spectrograms of the clean vocal, non-vocal, and singing voice signals are shown in Fig. 8.

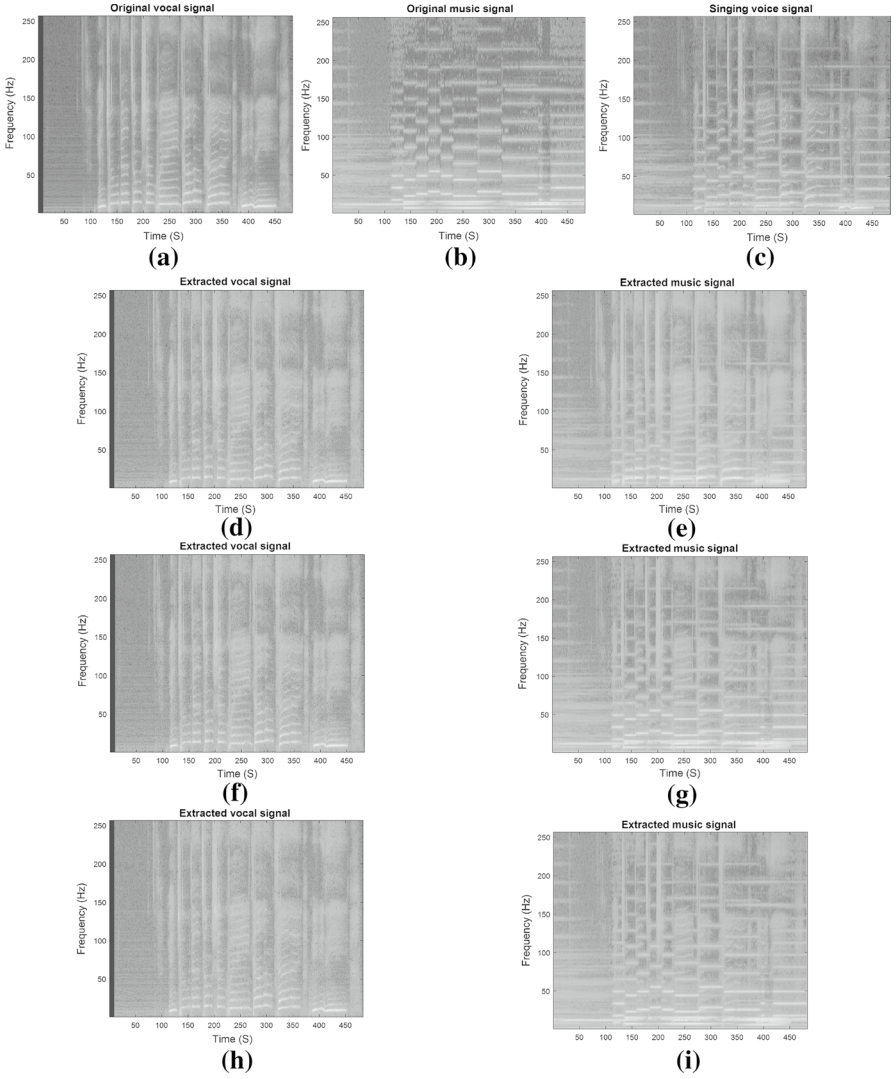

Fig. 8 Spectrograms of **a** clean vocal signal, **b** original non-vocal signal, **c** singing voice signal at SNR = 0 dB. The decomposed signals using the proposed method at SNR = 5 dB: **d** vocal signal and **e** non-vocal signal. The decomposed signals using the proposed method at SNR = 0 dB: **f** vocal signal and **g** non-vocal signal. The decomposed signals using the proposed method at SNR = −5 dB: **h** vocal signal and **i** non-vocal signal

Moreover, the separated sparse and low-rank components of the singing voice signal estimated by the proposed method can be seen in this figure.

Also, the decomposed parts of the observed singing voice signal, the sparse component (vocal part), and the low-rank component (non-vocal part) for different mentioned methods at SNR = 0 are shown in Fig. 9. The results of the proposed method in Figs. 8 and 9 are obtained in the SI situation.

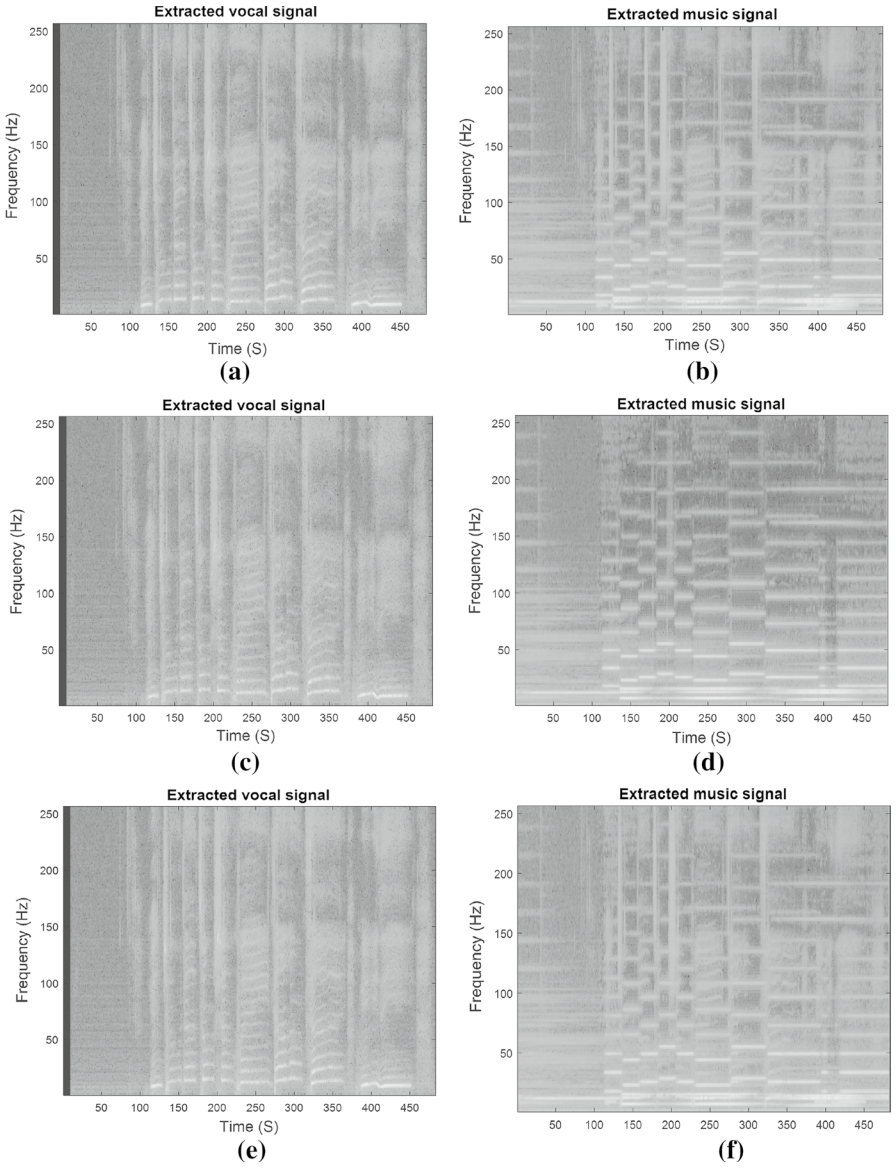

Fig. 9 Spectrograms of the extracted sparse and low-rank components at SNR = 0 dB using: The proposed method, **a** vocal signal and **b** non-vocal signal. DRNN method [19], **c** vocal signal and **d** non-vocal signal. MLRR algorithm [15], **e** vocal signal and **f** non-vocal part

The spectrograms of the sparse and low-rank components decomposed using the proposed learning-based separation scheme show that this approach can eliminate successfully the background non-vocal signal from the singing voice signal at different SNR values.

The captured spectrograms for the vocal component show that the proposed separation method separates the singing voice signal with more accuracy than other comparison methods. Also, it can be seen that other methods have more non-vocal residual or vocal distortion than the proposed algorithm.

5.2 Discussion

In this paper, the vocal and non-vocal dictionaries are learned by the comprehensive data with an acceptable approximation error. Therefore, the frequency content of each input frame is exactly represented using the related atoms. If a mismatch exists between the components of the training and test data, the spectrum of the non-vocal components cannot be coded by the vocal dictionary and it should be sparsely represented over the non-vocal dictionary by LARC sparse coding.

In fact, applying a sparse coding method with variable cardinality parameter to obtain lower reconstruction error, as well as the domain adaptation technique, has important effects in the simulation results.

The superiority of the proposed method results from two issues, using domain adaptation technique to alleviate the mismatch between the non-vocal components in the training and test steps and also learning incoherence dictionaries for the vocal and non-vocal signals by the proposed VAD algorithm. On the other hand, the non-vocal signals are well structured and are used to learn an incoherent dictionary by the IPR method. So, the non-vocal components of the observed frames are disregarded in the sparse representation using LARC coding over the vocal atoms. The incoherent dictionary learning process helps us to attenuate this problem and achieve better results than other algorithms. Also, using an alternating learned-based optimization method to solve the constrained decomposition problem has a prominent role in the improvement of the separation quality. A lower performance is obtained with the RPCA method since it works using an unsupervised scheme without any prior information about the structure of observed data.

6 Conclusions

In this paper, a new decomposition algorithm is presented for the singing voice separation problem. A low-rank sparse decomposition model in the time–frequency domain is applied to estimate the vocal and non-vocal parts of the singing voice signal. An incoherent dictionary learning scheme is presented based on the LARC representation including the atom–data coherence parameter and IPR post-processing method to train the incoherent atoms. This modified training process increases the mutual coherence between the training data and the learned atoms and also reduces the coherence between the dictionary atoms. An alternating optimization method based on the sparse coding over the vocal and non-vocal models is used to decompose the sparse and low-rank parts of the singing voice signal obtained from the monaural recordings. Also, a novel voice activity detector scheme is introduced

based on the energy of the sparse coefficient matrix to learn the atoms related to the non-vocal data. These atoms are adapted to the captured non-vocal components in the test signal by applying the domain transfer technique. In fact, all information about the vocal and non-vocal components is considered in order to increase the quality of the separated vocal signal, especially at the low SNR values. The experimental results using different measures and statistical significance test show that the proposed scheme leads to significantly better results than the earlier methods in this context and the basic and traditional approaches.

Acknowledgements The author wishes to thank Professor P. Loizou for making the source codes of the fwSegSNR and PESQ for the objective quality evaluations publicly available. The author also thanks Christian D. Sigg for publishing the MATLAB implementations of the LARC algorithm.

References

1. M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing over-complete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**, 4311–4322 (2006)
2. D. Barchiesi, M.D. Plumbley, Learning incoherent dictionaries for sparse approximation using iterative projections and rotations. IEEE Trans. Signal Process. **61**, 2055–2065 (2013)
3. J. Benesty, *Springer Handbook of Speech Processing* (Springer, Berlin, 2008)
4. N. Boulanger, G. Mysore, 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) (2014), pp. 7019–7023
5. E.J. Candes, L. Xiaodong, Y. Ma, J. Wright, Robust principal component analysis? J. ACM **58**, 1–39 (2011)
6. P. Chandna, M. Miron, J. Janer, E. Gomez, Monoaural audio source separation using deep convolutional neural networks, in International Conference on Latent Variable Analysis and Signal Separation (2017), pp. 258–266
7. G. Chen, C. Xiong, J.J. Corso, Dictionary transfer for image denoising via domain adaptation, in Proceedings of IEEE International Conference on Image Processing (2012), pp. 1189–1192
8. J. Demsar, Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**, 2845–2862 (2001)
10. J.L. Durrieu, G. Richard, B. David, C. Fevotte, Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Trans. Audio Speech Lang. Process. **18**, 564–575 (2010)
11. Z.C. Fan, Y.L. Lai, J.S.R. Jang, SVSGAN: singing voice separation via generative adversarial network, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
12. H. Fujihara, M. Goto, A music information retrieval system based on singing voice timbre, in ISMIR (2007), pp. 467–470
13. H. Fujihara, M. Goto, J. Ogata, H.G. Okuno, Lyric synchronizer: automatic synchronization system between musical audio signals and lyrics. J. Sel. Top. Signal Process. **5**, 1252–1261 (2011)
14. A. Gray, J. Markel, Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **24**, 380–391 (1976)
15. C.L. Hsu, J.S.R. Jang, On the improvement of singing voice separation for monaural recordings using the MIR-1 K dataset. IEEE Trans. Audio Speech Lang. Process. **18**, 310–319 (2010)
16. P.S. Huang, S.D. Chen, P. Smaragdis, M. Hasegawa, Singing voice separation from monaural recordings using robust principal component analysis, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012), pp. 57–60

17. P.S. Huang, M. Kim, M. Johnson, P. Smaragdis, Singing-voice separation from monaural recordings using deep recurrent neural networks, in International Society for Music Information Retrieval Conference (2014)
18. Y. Ikemiya, K. Itoyama, K. Yoshii, Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation. J. IEEE/ACM TASLP **24**, 2084–2095 (2016)
19. A. Jansson, E.J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, Singing voice separation with deep U-Net convolutional networks, in Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR) (2017), pp. 745–751
20. M. Lagrange, A. Ozerov, E. Vincent, Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning, in Proceedings of the ISMIR (2012), pp. 595–560
21. H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, advances in neural information processing systems. Adv. Neural. Inf. Process. Syst. **19**, 801–808 (2007)
22. Y. Li, D.L. Wang, Singing voice separation from monaural recordings, in Proceedings of the International Conference of Music Information Retrieval (2006), pp. 176–179
23. P.C. Loizou, *Speech Enhancement: Theory and Practice* (Taylor and Francis, London, 2007)
24. Y. Luo, Z. Chen, D.P.W. Ellis, Deep clustering for singing voice separation, in MIREX, task of Singing Voice Separation (2016), pp. 1–2
25. Y. Luo, Z. Chen, J.R. Hershey, J.L. Roux, N. Mesgarani, Deep clustering and conventional networks for music separation: Stronger together, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017), pp. 61–65
26. J. Ma, Y. Hu, P.C. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band importance functions. J. Acoust. Soc. Am. **125**, 3387–3405 (2009)
27. S. Mavaddati, A novel singing voice separation method based on sparse non-negative matrix factorization and low-rank modeling. Iran. J. Electr. Electron. Eng. **15**, 1–17 (2019)
28. S. Mavaddaty, S.M. Ahadi, S. Seyedin, A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation. Speech Commun. **76**, 42–60 (2016)
29. S. Mavaddaty, S.M. Ahadi, S. Seyedin, Modified coherence-based dictionary learning method for speech enhancement. Signal Process. IET **9**, 537–545 (2015)
30. S. Mavaddaty, S.M. Ahadi, S. Seyedin, Speech enhancement using sparse dictionary learning in wavelet packet transform domain. Comput. Speech Lang. **44**, 22–47 (2017)
31. A.R. Nerkar, M.A. Joshi, Singing-voice separation from monaural recordings using empirical wavelet transform, in International Conference on Advanced Communication Control and Computing Technologies (2016), pp. 795–800
32. B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1. Vis. Res. **37**, 3311–3325 (1997)
33. A. Ozerov, P. Philippe, F. Bimbot, R. Gribonval, Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Trans. Audio Speech Lang. Process. **15**, 1564–1578 (2007)
34. Z. Rafii, 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) (2011), pp. 221–224
35. Z. Rafii, B. Pardo, Repeating pattern extraction technique (REPET): a simple method for music/voice separation. IEEE Trans. Audio Speech Lang. Process. **21**, 73–84 (2013)
36. A. Rix, J. Beerends, M. Hollier, 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 (2001), pp. 749–752
37. D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th edn. (Chapman & Hall/CRC, Boca Raton, 2000)
38. C.D. Sigg, T. Dikk, J.M. Buhmann, Speech enhancement using generative dictionary learning. IEEE Trans. Acoust. Speech Signal Process. **20**, 1698–1712 (2012)
39. P. Sprechmann, A. Bronstein, G. Sapiro, Real-time online singing voice separation from monaural recordings using robust low-rank modeling, in Proceedings of the 13th International Society for Music Information Retrieval Conference (2012), pp. 67–72
40. P. Teng, Y. Jia, Voice activity detection via noise reducing using non-negative sparse coding. IEEE Signal Process. Lett. **20**, 475–478 (2013)

41. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469 (2006)
42. Y.H. Yang, Low-rank representation of both singing voice and music accompaniment via learned dictionaries, in Proceedings of the 14th International Society for Music Information Retrieval Conference (2013), pp. 427–432
43. Y.H. Yang, On sparse and low-rank matrix decomposition for singing voice separation, in ACM Multimedia (2012), pp. 757–760
44. L. Yipeng, W. DeLiang, Separation of singing voice from music accompaniment for monaural recordings. IEEE Trans. Audio Speech Lang. Process. **15**, 1475–1487 (2007)
45. D.T. You, J.Q. Han, G.B. Zheng, T.R. Zheng, Sparse power spectrum based robust voice activity detector, in IEEE International Conference on Acoustics, Speech, and Signal Processing (2012), pp. 289–292
46. D.T. You, J.Q. Han, G.B. Zheng, T.R. Zheng, J. Li, Sparse representation with optimized learned dictionary for robust voice activity detection. Circuits Syst. Signal Process. **33**, 2267–2291 (2014)