Neural Computing and Applications (2020) 32:3311–3322
https://doi.org/10.1007/s00521-018-3757-x

ORIGINAL ARTICLE

Singing voice separation with pre-learned dictionary and reconstructed voice spectrogram

Chenghong Yang[1] · Hongjuan Zhang[1]

Received: 9 December 2017 / Accepted: 28 September 2018 / Published online: 6 October 2018

Abstract

Recently the mixture spectrogram of a song is usually considered as a superposition of a sparse spectrogram and a low-rank spectrogram, which correspond to the vocal part and the accompaniment part of the song, respectively. Based on this observation, one can separate singing voice from the background music. However, the quality of such separation might be limited, since the vocal part may be not described very well by low rank, and moreover its more prior information, such as annotation, should be considered when designing separation algorithm. Based on these considerations, in this paper, we present two categories, time–frequency-based source separation algorithms. Specifically, one incorporates both the vocal and instrumental spectrograms as sparse matrix and low-rank matrix, meanwhile combines some side information of vocal part, i.e., the reconstructed voice spectrogram from the annotation. The others further consider both the vocal and instrumental spectrograms as sparse matrix and group-sparse matrix, respectively. Evaluations on the iKala dataset show that the proposed methods are effective and efficient for both the separated singing voice and music accompaniment.

Keywords Singing voice separation · Low rank · Group-sparse · Dictionary Learning

1 Introduction

Automatic singing voice separation, which intends to extract the singing voice from the music mixture, has received much attention in the field of audio signal processing in recent years [1–4]. It has broad applications in singer identification [5, 6], automatic singing transcription [7], automatic lyrics alignment [8], music information retrieval [9], and content-based music retrieval [10, 11].

The singing voice separation task solicits competing entries to blindly separate the singer's voice from music recordings. However, musical sound sources are often strongly correlated in time and frequency, a musical recording is often infeasible without additional knowledge about the sources a decomposition. It is extremely difficult for computer systems, although the human auditory system can easily distinguish the vocal and instrumental of music recording. The main challenges come from the variety of simultaneous sound sources as well as the rich pitch and timbre variations of singing voice. As summarized below, recently, many algorithms have been proposed to separate singing voice from music recording, yet the progress is still limited.

Based on Robust Principal Component Analysis (RPCA) [12], which is a matrix factorization algorithm for solving underlying low-rank and sparse matrices. Suppose we are given a large data matrix X, and know that it may be decomposed as $X = A + E$, where A is a low-rank matrix and E is a sparse matrix. Huang et al. [13] have separated singing voice from music accompaniment with the assumption that the repetitive music accompaniment lie in a low-rank subspace, while the singing voices can be regarded as relatively sparse within songs. The main drawback to this approach is that the resulting sparse matrix often contains instrumental solo or percussion [14, 15]. Yang [14] further incorporated harmonicity priors and a back-end drum removal procedure to improve the separation. Su and Yang proposed a novel artist identification method based on sparse features learned from both the magnitude and phase parts of the spectrum [15]. Papadopoulos et al. [16] presented an adaptive formulation

Grants No. 11501351.

✉ Hongjuan Zhang
zhanghongjuan@shu.edu.cn

[1] Department of Mathematics, Shanghai University, Shanghai 200444, People's Republic of China

of RPCA that incorporates music content information to guide the decomposition.

Further, the data samples can be represented as linear combinations of the bases in a given dictionary [17], i.e., $X = DZ + E$, where X is a input matrix, D is a dictionary of sparse matrices and E is a sparse matrix. For music spectrograms, we assume that the instrumentals are repetitive lies in a low-rank matrix and the vocal are sparse lies in a sparse matrix. Because the vocal part of a song can sometimes be low rank as well, the quality of such separation might be limited. Yang [18] considered both the singing voice and music accompaniment as low-rank matrices and employed an online dictionary learning algorithm [19] to learn the structures of singing voice and accompaniment sounds from clean vocal and instrumental signals as prior knowledge introduced in the decomposition process.

Chen and Ellis applied the RPCA framework to speech enhancement assuming that the background noise is low-rank and the speech is sparse [20, 21]. Differently, they incorporated the pre-learned idea to decompose the sparse components into the product of a pre-learned speech dictionary and a sparse activation matrix. They proposed to use the sum of a low-rank matrix and a residual to identify the background noise [20]. This approach, however, cannot be directly used for singing voice separation in music. This is because the background music is often much more non-stationary than background noise and may not be well represented by a low-rank matrix and a residual. To solve this problem, Yu et al. [22] proposed Low-rank and Sparse representation with Pre-learned Dictionaries (LSPD) under the Alternating Direction Method of Multipliers (ADMM) framework for singing voice separation. First, they pre-learned universal voice and music dictionaries from isolated singing voice and background music training data. Then, in addition to using a sparse spectrogram and a low-rank spectrogram to model the singing voice and the background music, respectively, they added a residual term to capture the components that are not well modeled by either the sparse or the low-rank terms.

A partial solution for the sparse matrix contains instrumental solo or percussion is to incorporate reliable annotations for the sparse part using informed RPCA (hereafter RPCAi) [23]. Chan et al. [24] presents a modified RPCA algorithm, the algorithm is then applied to separate the singing voice from the instrumental accompaniment using vocal activity information, this work represents one of the first attempts to incorporate vocal activity information into the RPCA algorithm, while vocal activity detection has been studied extensively [25, 26]. Ikemiya et al. [27] proposed a novel framework for improving both vocal F0 estimation and singing voice separation by making effective use of the mutual dependency of those tasks. Therefore, we present the first model named Low-rank, Sparse representation with Pre-learned Dictionaries and side Information (LSPDi) under the ADMM framework. First, we pre-learn two dictionaries about foreground singing voice and background music and use a sparse spectrogram and a low-rank spectrogram to model them, respectively. Then, a residual term is added to capture the components that are not well modeled by either the sparse or the low-rank term. Finally, we incorporate the reconstructed voice spectrogram from the annotation when separating vocal and music.

However, low-rank optimizations are computationally inefficient due to the use of singular value decompositions. To motivate a new representation, Chan et al. [28] proposed to separate singing voice by informed group-sparse representation with the idea of informed separation incorporating pitch annotations. In jazz and popular music, a few chord symbols are enough to compactly represent the harmonic structure of a piece. One observation is that there are many empty rows in this representation. Therefore, a promising strategy for the inverse problem is to encourage row sparsity given an instrumental dictionary. On that basis, we present the second model named Group-Sparse, Sparse with Pre-learned Dictionaries (GSPD) and the third model named Group-Sparse, Sparse with Pre-learned Dictionaries and side Information (GSPDi). Firstly, we use a sparse spectrogram and group-sparse spectrogram to define the singing voice and the background music, respectively. In addition, a residual term is added to fit the components that are not identified by the low-rank and the sparse part. Specially, we pre-learned voice and music dictionaries from clean singing voice. Evaluations on the iKala dataset [29] show their better performance than existing methods.

The rest of this paper is organized as follows. The overview of the existing model is presented in Sect. 2. Section 3 presents the proposed methods. In Sect. 4, dictionary and vocal activity information are described. The simulation results are presented in Sect. 5. Final section concludes this work.

2 Existing algorithm

Low-rank and sparse representation with pre-learned Dictionaries (LSPD) method [22], shown as,

$$\begin{aligned} &\min_{Z_1,Z_2} \lambda_1 \|Z_1\|_* + \lambda_2 \|Z_2\|_1 + \lambda_3 \|E\|_1 \\ &s.t. X = D_1 Z_1 + D_2 Z_2 + E \end{aligned} \tag{1}$$

where X is the input spectrogram, $D_1 \in R^{m \times k_1}$ is a pre-learned dictionary of the music accompaniment, $D_2 \in R^{m \times k_2}$ is a pre-learned dictionary of the singing voice, $D_1 Z_1$

is the separated instrumentals, D_2Z_2 is the separated voice, E denotes the residual part. λ_1, λ_2, and λ_3 are three weighting parameters for balancing the different regularization terms in this model.

It is noting that LSPD utilizes a sparse spectrogram and a low-rank spectrogram to model the singing voice and the background music, respectively, and adds a residual term to capture the components that are not well modeled by either the sparse or the low-rank terms, which improves the performance of related existing methods to some extent [13, 18, 30]. In the following section, we will present three new algorithms.

3 The proposed method

3.1 Low-rank, sparse representation with pre-learned dictionaries and side information (LSPDi)

In order to improve the separation quality of LSPD further, we considered more prior information and added them, i.e., the reconstructed voice spectrogram from the annotation, to the following model,

$$\begin{aligned}&\min_{Z_1,Z_2} \lambda_1\|Z_1\|_* + \lambda_2\|Z_2\|_1 + \lambda_3\|E\|_1 + \frac{\gamma}{2}\|D_2Z_2 - E_0\|_F^2\\&s.t. X = D_1Z_1 + D_2Z_2 + E\end{aligned} \tag{2}$$

Here, all parameters in model (2) are in accordance with model (1), and E_0 denotes the reconstructed voice spectrogram from the annotation. $\|\cdot\|_F$ denotes the Frobenius norm (square root of the sum of the squares of its elements), so that a subtraction can be calculated between the separated voice and the reconstructed voice spectrogram from the annotation. In the following, we also use the ADMM algorithm [31] to solve the optimization problem, by introducing two auxiliary variables J_1 and J_2 as well as three equality constraints,

$$\begin{aligned}&\min_{Z_1,Z_2,J_1,J_2} \lambda_1\|J_1\|_* + \lambda_2\|J_2\|_1 + \lambda_3\|E\|_1\\&\quad + \frac{\gamma}{2}\|D_2Z_2 - E_0\|_F^2\\&s.t. X = D_1Z_1 + D_2Z_2 + E, Z_1 = J_1, Z_2 = J_2\end{aligned} \tag{3}$$

The unconstrained augmented Lagrangian $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L} = &\lambda_1\|J_1^T\|_* + \lambda_2\|J_2\|_1 + \lambda_3\|E\|_1 + \frac{\gamma}{2}\|D_2Z_2 - E_0\|_F^2\\&+ < Y_1, X - D_1Z_1 - D_2Z_2 - E >\\&+ < Y_2, Z_1 - J_1 > + < Y_3, Z_2 - J_2 >\\&+ \frac{\mu}{2}\Big(\|X - D_1Z_1 - D_2Z_2 - E\|_F^2\\&+ \|Z_1 - J_1\|_F^2 + \|Z_2 - J_2\|_F^2\Big)\end{aligned} \tag{4}$$

where Y_1, Y_2, Y_3 are the Lagrange multipliers. We then iteratively update the solutions for J_1, Z_1, J_2 and Z_2.

Specifically, update J_2 firstly,

$$J_2 = \arg\min_{J_2} \lambda_2\|J_2\|_1 + \frac{\mu}{2}\|J_1 - \left(Z_1 + \mu^{-1}Y_3\right)\|_F^2 \tag{5}$$

that can be solved by the soft-threshold operator

$$J_2 = S_{\frac{\lambda_2}{\mu}}\left(Z_2 + \mu^{-1}Y_3\right) \tag{6}$$

since the spectrogram is non-negative

$$J_2 = \max\left\{S_{\frac{\lambda_2}{\mu}}\left(Z_2 + \mu^{-1}Y_3\right), 0\right\} \tag{7}$$

where 0 is an all zero matrix of the size as J_2.

Then, update Z_2, setting $\frac{\partial\mathcal{L}}{\partial Z_2} = 0$,

$$\begin{aligned}Z_2 = &\left(\left(\frac{\gamma}{\mu} + 1\right)D_2^TD_2 + I\right)^{-1}\\&\left(D_2^T\left(X - D_1Z_1 - E + \frac{\gamma}{\mu}E_0 + \frac{1}{\mu}Y_1\right)\right.\\&\left.- \frac{1}{\mu}Y_3 + J_2\right)\end{aligned} \tag{8}$$

And the other variables can be updated using the similar way. The detailed algorithm is shown as follows.

LSPDi

```
Input: $X$, $D_1$, $D_2$
output: $Z_1$, $Z_2$
initialization: $Z_1 = 0$, $Z_2 = 0$, $J_1 = 0$, $J_2 = 0$, $Y_1 = 0$, $Y_2 = 0$,$Y_3 = 0$
while not converged do
update $Z_1$, $Z_2$:
$Z_1 = (D_1^TD_1 + I)^{-1}(D_1^T(X - D_2Z_2 - E + \mu^{-1}Y_1) - \mu^{-1}Y_2 + J_1)$
$Z_2 = \left(\left(\frac{\gamma}{\mu} + 1\right)D_2^TD_2 + I\right)^{-1}\left(D_2^T\left(X - D_1Z_1 - E + \frac{\gamma}{\mu}E_0\right.\right.$
  $\left.\left. + \frac{1}{\mu}Y_1) - \frac{1}{\mu}Y_3 + J_2\right)$
update $J_1$, $J_2$, $E$:
$U\Sigma V = svd(Z_1 + \frac{1}{\mu}Y_2)$, $J_1 = US_{\frac{\lambda_1}{\mu}}[\Sigma]V^T$
$J_2 = S_{\frac{\lambda_2}{\mu}}(Z_2 + \frac{1}{\mu}Y_3)$
$E = S_{\frac{\lambda_3}{\mu}}(X - D_1Z_1 - D_2Z_2 + \mu^{-1}Y_1)$
```

LSPDi

```
update $Y_1$, $Y_2$, $Y_3$:
$Y_1 = Y_1 + \mu(X - D_1Z_1 - D_2Z_2 - E)$
$Y_2 = Y_1 + \mu(Z_1 - J_1)$
$Y_3 = Y_2 + \mu(Z_2 - J_2)$
end while
```

3.2 Group-sparse, sparse representation with pre-learned dictionaries (GSPD)

Due to the use of singular value decompositions, low-rank optimization is computationally inefficient, which prevents its applications for the processing of big data. Here, group sparse as an replacement is used when designing the optimal model,

$$\begin{aligned}&\min_{Z_1,Z_2} \lambda_1\|Z_1^T\|_{2,1} + \lambda_2\|Z_2\|_1 + \lambda_3\|E\|_1\\ &s.t. X = D_1Z_1 + D_2Z_2 + E\end{aligned} \tag{9}$$

where $\|Z^T\|_{2,1} = \sum_i \sqrt{\sum_j Z_{ij}^2}$ is the row sparsity, which means Z has many empty rows in its representation [32] and corresponds to the sum of the l_2-norms of the rows of Z.

As we all know, the above formulation is not trivial to solve since the $\|\cdot\|_{2,1}$ and $\|\cdot\|_1$ norms are not smooth. Moreover, an additional equality constraint should be considered. Therefore, the alternating direction method of multipliers (ADMM) [33] is applied for this model. ADMM works by first rewriting the constraint(s) into an augmented Lagrange function, then updating each variable in an alternating fashion until convergence. Thus, to solve (9), we first introduce two auxiliary variables J_1 and J_2 for the alternating updates and rewrite the optimization problem as follows:

$$\begin{aligned}&\min_{Z_1,Z_2,J_1,J_2} \lambda_1\|J_1^T\|_{2,1} + \lambda_2\|J_2\|_1 + \lambda_3\|E\|_1\\ &s.t. X = D_1Z_1 + D_2Z_2 + E, Z_1 = J_1, Z_2 = J_2\end{aligned} \tag{10}$$

The unconstrained augmented Lagrangian $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L} = {} & \lambda_1\|J_1^T\|_{2,1} + \lambda_2\|J_2\|_1 + \lambda_3\|E\|_1\\ & + <Y_1, X - D_1Z_1 - D_2Z_2 - E>\\ & + <Y_2, Z_1 - J_1> + <Y_3, Z_2 - J_2>\\ & + \frac{\mu}{2}\Big(\|X - D_1Z_1 - D_2Z_2 - E\|_F^2 + \|Z_1\\ & - J_1\|_F^2 + \|Z_2 - J_2\|_F^2\Big)\end{aligned} \tag{11}$$

where Y_1, Y_2 and Y_3 are the Lagrange multipliers. Model (11) can be minimized with respect to J_1, J_2, Z_1, Z_2 and E, respectively, by fixing the other variables and updating the lagrangian multipliers Y_1, Y_2 and Y_3. For example, the minimization of J_1 reduces to

$$\begin{aligned}J_1 &= \arg\min_{J_1} \lambda_1\|J_1^T\|_{2,1} + \frac{\mu}{2}\|J_1 - \big(Z_1 + \mu^{-1}Y_2\big)\|_F^2\\ &= \left(\left(1 - \frac{\lambda_1}{\|(Z_1 + \mu^{-1}Y_2)_i\|}\right)_+ \big(Z_1 + \mu^{-1}Y_2\big)_i\right)_{i=1}^k\end{aligned} \tag{12}$$

where k is the number of row of J_1, A_i denotes the ith row of A, $(B_i)_{i=1}^k = (B_1^T, \ldots, B_k^T)$ and $C_+ = \max(0, C)$.

Update Z_1, setting $\frac{\partial\mathcal{L}}{\partial Z_1} = 0$,

$$\begin{aligned}Z_1 = {} & \big(D_1^TD_1 + I\big)^{-1}\big(D_1^T\big(X - D_2Z_2 - E + \mu^{-1}Y_1\big)\\ & - \mu^{-1}Y_2 + J_1\big)\end{aligned} \tag{13}$$

The other variables can be updated using the similar way. In the following, the proposed algorithm will be given.

GSPD

```
Input: $X$, $D_1$, $D_2$
output: $Z_1$, $Z_2$
initialization: $Z_1 = 0$, $Z_2 = 0$, $J_1 = 0$, $J_2 = 0$, $Y_1 = 0$, $Y_2 = 0$,$Y_3 = 0$
while not converged do
update $Z_1$, $Z_2$:
$Z_1 = (D_1^TD_1 + I)^{-1}(D_1^T(X - D_2Z_2 - E + \mu^{-1}Y_1) - \mu^{-1}Y_2 + J_1)$
$Z_2 = (D_2^TD_2 + I)^{-1}(D_2^T(X - D_1Z_1 - E + \mu^{-1}Y_1) - \mu^{-1}Y_3 + J_2)$
update $J_1$, $J_2$ $E$:
$J_1 = ((1 - \frac{\lambda_1}{\|(Z_1+\mu^{-1}Y_2)_i\|})_+(Z_1 + \mu^{-1}Y_2)_i)_{i=1}^k$
$J_2 = S_{\frac{\lambda_2}{\mu}}(Z_2 + \frac{1}{\mu}Y_3)$
$E = S_{\frac{\lambda_3}{\mu}}(X - D_1Z_1 - D_2Z_2 + \mu^{-1}Y_1)$
update $Y_1$, $Y_2$, $Y_3$:
$Y_1 = Y_1 + \mu(X - D_1Z_1 - D_2Z_2 - E)$
$Y_2 = Y_1 + \mu(Z_1 - J_1)$
$Y_3 = Y_2 + \mu(Z_2 - J_2)$
end while
```

3.3 Sparse representation with pre-learned dictionaries and side information (GSPDi)

Furthermore, more prior information, i.e., the reconstructed voice spectrogram from the annotation is considered in the following model.

$$\min_{Z_1,Z_2} \lambda_1 \|Z_1^T\|_{2,1} + \lambda_2 \|Z_2\|_1 + \lambda_3 \|E\|_1 + \frac{\gamma}{2} \|D_2 Z_2 - E_0\|_F^2$$
$$s.t. X = D_1 Z_1 + D_2 Z_2 + E, \tag{14}$$

To solve (14), we first introduce two auxiliary variables J_1 and J_2 for the alternating updates and rewrite the optimization problem as follows:

$$\min_{Z_1,Z_2,J_1,J_2} \lambda_1 \|J_1^T\|_{2,1} + \lambda_2 \|J_2\|_1 + \lambda_3 \|E\|_1 + \frac{\gamma}{2} \|D_2 Z_2 - E_0\|_F^2$$
$$s.t. X = D_1 Z_1 + D_2 Z_2 + E, Z_1 = J_1, Z_2 = J_2, \tag{15}$$

The unconstrained augmented Lagrangian $\mathcal{L}$ is given by

$$\begin{aligned} \mathcal{L} = {} & \lambda_1 \|J_1^T\|_{2,1} + \lambda_2 \|J_2\|_1 + \lambda_3 \|E\|_1 + \frac{\gamma}{2} \|D_2 Z_2 - E_0\|_F^2 \\ & + < Y_1, X - D_1 Z_1 - D_2 Z_2 - E > \\ & + < Y_2, Z_1 - J_1 > + < Y_3, Z_2 - J_2 > \\ & + \frac{\mu}{2} \Big(\|X - D_1 Z_1 - D_2 Z_2 - E\|_F^2 \\ & + \|Z_1 - J_1\|_F^2 + \|Z_2 - J_2\|_F^2 \Big) \end{aligned} \tag{16}$$

where Y_1, Y_2, and Y_3 are the Lagrange multipliers. $< \cdot, \cdot >$ denotes the trace inner product, and μ is a positive penalty parameter. We the iteratively update the solutions for J_1, Z_1, J_2 and Z_2. The method of updated variables is similar to the previous two algorithms in above-mentioned Sects. 3.1 and 3.2.

The proposed algorithm is described below.

GSPDi

```
**Input:** $X$, $D_1$, $D_2$
**output:** $Z_1$, $Z_2$
**initialization:** $Z_1 = 0, Z_2 = 0, J_1 = 0, J_2 = 0, Y_1 = 0, Y_2 = 0, Y_3 = 0$
**while not** converged **do**
**update** $Z_1$, $Z_2$:
$Z_1 = (D_1^T D_1 + I)^{-1}(D_1^T(X - D_2 Z_2 - E + \mu^{-1} Y_1) - \mu^{-1} Y_2 + J_1)$
$Z_2 = ((\frac{\gamma}{\mu} + 1) D_2^T D_2 + I)^{-1}(D_2^T(X - D_1 Z_1 - E + \frac{\gamma}{\mu} E_0 + \frac{1}{\mu} Y_1)$
  $- \frac{1}{\mu} Y_3 + J_2)$
**update** $J_1$, $J_2$ $E$:
$J_1 = ((1 - \frac{\lambda_1}{\|(Z_1 + \mu^{-1} Y_2)_i\|})_+ (Z_1 + \mu^{-1} Y_2)_i)_{i=1}^k$
$J_2 = S_{\frac{\lambda_2}{\mu}}(Z_2 + \frac{1}{\mu} Y_3)$
$E = S_{\frac{\lambda_3}{\mu}}(X - D_1 Z_1 - D_2 Z_2 + \mu^{-1} Y_1)$
**update** $Y_1$, $Y_2$, $Y_3$:
$Y_1 = Y_1 + \mu(X - D_1 Z_1 - D_2 Z_2 - E)$
$Y_2 = Y_1 + \mu(Z_1 - J_1)$
$Y_3 = Y_2 + \mu(Z_2 - J_2)$
**end while**
```

4 Dictionary and E_0

4.1 Dictionary

We adopt the idea of **O**nline **D**ictionary **L**earning for Sparse Coding (ODL) [19] to learn the singing voice dictionary from isolated training singing voices.

Given N signals ($x_i \in \mathbb{R}_m$), ODL learns a dictionary D by solving the following joint optimization problem,

$$\min_{D \geq 0, \alpha} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2} \|x_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right) \tag{17}$$
$$s.t. d_j^T d_j \leq 1, \alpha_i \geq 0$$

where $\| \cdot \|_2$ denotes the Euclidean and λ is a regularization parameter. The input frames are extracted from the training set after short-time Fourier transform (STFT). Our implementation of ODL is based on the SPAMS toolbox [19]. Following literature [28], we define the dictionary size to be 100 atoms.

4.2 The reconstructed voice spectrogram from the annotation (E_0)

To get the reconstructed voice spectrogram from the annotation, we first transform the human-labeled vocal pitch contours into a time-frequency binary mask. The authors in literature [27] have proposed a harmonic mask similar to that of the work [34], which only passes integral multiples of the vocal fundamental frequencies [35, 36],

$$M(f,t) = \begin{cases} 1, & if \quad |f - nF_0(t)| < w/2, \exists n \in N^+ \\ \\ 0, & otherwise. \end{cases} \tag{18}$$

Here, $F_0(t)$ is the vocal fundamental frequency at time t, n is the order of the harmonic, and w is the width of the mask. Then, we simply define the vocal annotations as $E_0 = X \circ M$, where $\circ$ denotes the Hadamard product.

5 Evaluation

5.1 Dataset

Our evaluation is based on the iKala dataset [29]. The iKala dataset contains 352 30-s clips of Chinese popular songs in CD quality. The singers and musicians are professionals.

Following literature [28], in our experiments, we randomly select 44 songs for training (i.e., learning the dictionaries D_1 and D_2), leaving 208 songs for testing the

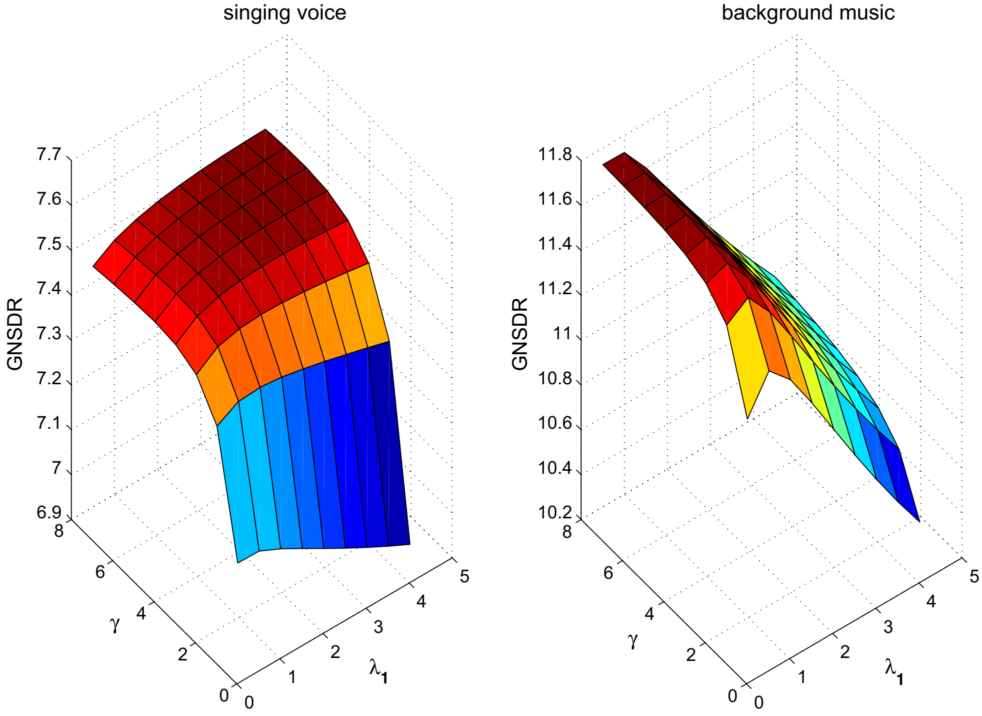

Fig. 1 Separation performance measured by GNSDR for the singing voice (left) and background music (right), using our proposed method LSPDi

performance of separation. The vocals and instrumentals are mixed at signal-to-noise ratio of 0 dB. To reduce the computational cost and the memory footprint of the proposed algorithm, we downsample all the audio recordings from 44,100 to 22,050 Hz. Then, computed its STFT by sliding a Hamming window of 1411 samples with a 75% overlap to obtain the spectrogram [29].

5.2 Evaluation

To measure the quality of the singing voice $\widehat{v}$ with respect to the original clean singing voice v, we use source-to-interference ratio (SIR), source-to-artifacts ratio (SAR) and source-to-distortion ratio (SDR) provided in the commonly used BSS EVAL toolbox version 3.0.[1]

The source-to-distortion ratio (SDR) [37] is computed as follows,

$$\mathrm{SDR}(\widehat{v}, v) = 10\log_{10}\left[\frac{<\widehat{v}, v>^{2}}{\|\widehat{v}^2\|\|v^2\| - <\widehat{v}, v>^{2}}\right]. \tag{19}$$

Normalized SDR (NSDR) is the improvement of SDR from the original mixture x to the separated singing voice $\widehat{v}$ [38, 39], and is commonly used to measure the separation performance for each mixture:

$$\mathrm{NSDR}(\widehat{v}, v, x) = \mathrm{SDR}(\widehat{v}, v) - \mathrm{SDR}(x, v). \tag{20}$$

For overall performance evaluation, the global NSDR (GNSDR) is calculated as,

$$\mathrm{GNSDR} = \frac{\sum_{i=1}^{N} w_i \mathrm{NSDR}(\widehat{v}_i, v_i, x_i)}{\sum_{i=1}^{N} w_i}, \tag{21}$$

where N is the total number of the songs and w_i is the length of the i-th song. We calculate the weighted average of SIR and SAR , which are the Global SIR (GSIR) and Global SAR (GSAR), respectively, over different clips in a similar way. Higher values of SDR, SAR, SIR, GSIR, GSAR GSDR and GNSDR represent better quality of the separation.

5.3 Parameter selection

There are two versions of the proposed method for singing voice separation.

(1) The first one is low-rank representation,

- **LSPD** Supervised method proposed by Yu et al. [22].
- **LSPDi** Proposed LSPDi method with Low-Rank representation and the reconstructed voice spectrogram from the annotation.

(2) The other is group-sparse representation,

- **GSPD** Proposed GSPD method with Group-Sparse representation.

[1] http://bass-db.gforge.inria.fr/.

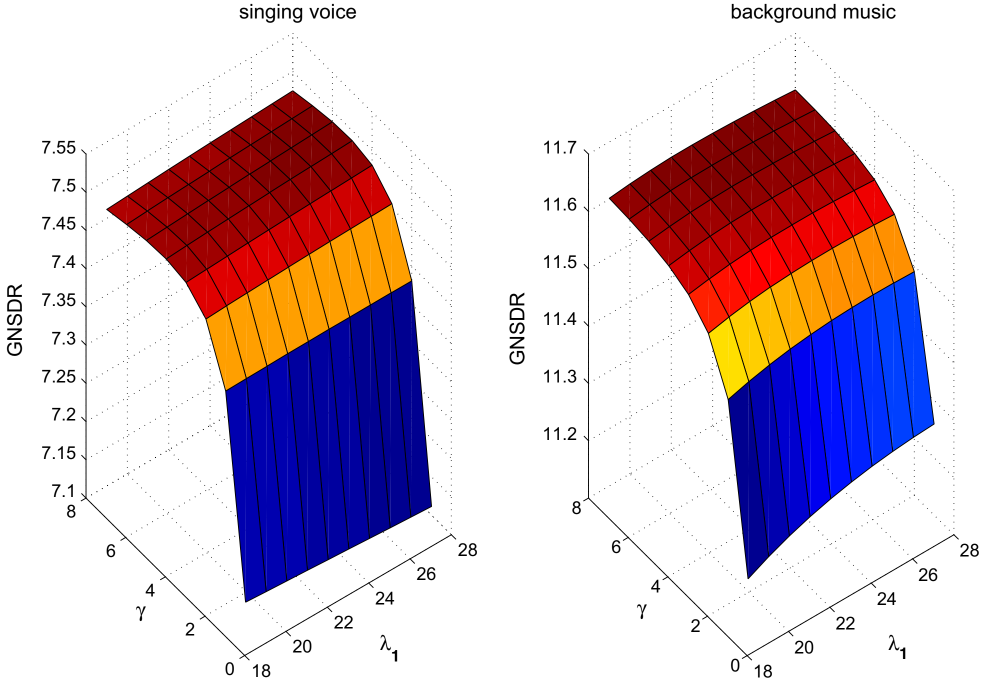

Fig. 2 Separation performance measured by GNSDR for the singing voice (left) and background music (right), using our proposed method GSPDi

- **GSPDi** Proposed GSPDi method with Group-Sparse representation and the reconstructed voice spectrogram from the annotation.

Note that LSPD and GSPD all have three parameters, respectively. Both LSPDi and GSPDi have four parameters.

During parameter selection, we use the indicator of global normalized source-to-distortion ratio (GNSDR) as the evaluation index. The higher the value is, the better the separation quality is. As for all algorithms, i.e., LSPD and LSPDi, GSPD and GSPDi, we set $\lambda_2 = \lambda_3 = 1/\sqrt{\max(m, n)}$ for each $X \in R^{m\times n}$ similar to the work in [29], Here, we only adjust λ_1 and γ in LSPDi and GSPDi.

Figure 1 shows the different GNSDR values of LSPDi for the separated singing voice and background music. First, fixing $\lambda_1 = 1$ (or any other), in the vocal section, the GNSDR monotonically rises at first before the maximum value is achieved and then decreases, when $\gamma = 5$, it reaches the optimal value. Then, we will focus on $\gamma = 5$, in the accompaniment part, fixing $\gamma = 5$, its value first increases and then reaches the maximum after a significant downward trend, reaches it optimal value when $\lambda_1 = 1$. Therefore, in the algorithm LSPDi, we use the parameter $\gamma = 5$ and $\lambda_1 = 1$. Just like LSPDi, we select the parameter $\lambda_1 = 1$ in LSPD.

Based on GSPDi, Fig. 2 presents the GNSDR for the separated singing voice and background music. In the vocal part, we can see that, for any value of λ_1, the GNSDR will always get the maximum value at $\gamma = 5$. So, in the accompaniment part, we fix $\gamma = 5$ and the value of the GNSDR reaches its maximum at $\lambda_1 = 24$, then the values have a significant downward trend. Therefore, in the GSPDi, we set the parameter $\lambda_1 = 24$ and $\gamma = 5$. And we select the parameter $\lambda_1 = 24$ in GSPD, in the same way with GSPDi.

Table 1 Separation quality for the vocal and music for the iKala dataset of LSPD, LSPDi, GSPD and GSPDi

	Vocal			Music		
	GSDR	GSIR	GSAR	GSDR	GSIR	GSAR
LSPD	5.26	11.47	7.19	1.22	2.47	11.73
LSPDi	11.29	19.91	12.16	8.00	16.70	8.88
GSPD	5.29	11.50	7.41	0.98	1.93	11.52
GSPDi	11.29	19.61	12.21	7.88	16.01	8.85

5.4 Comparison of the proposed method

Experimental results show that incorporating the reconstructed voice spectrogram from the annotation ($E0$) can greatly improve the separation performance. As shown in Table 1, Figs. 3 and 4, LSPDi and GSPDi are comparative and both of them achieve a higher performance. Therefore, LPDSi and GSPDi will be used for the following comparisons with the existing methods.

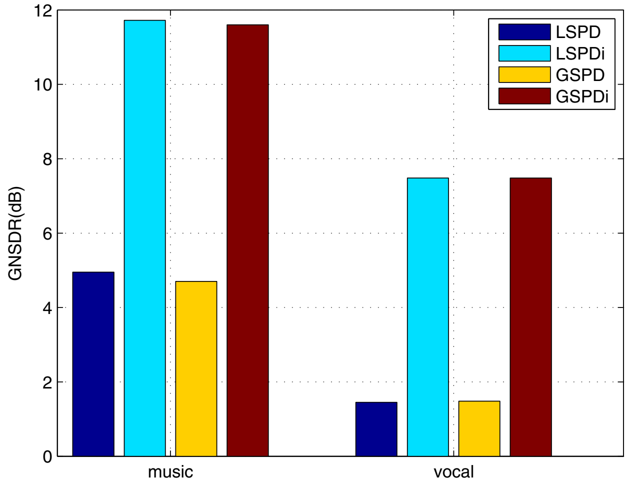

Fig. 3 Separation performance for the music (left) and vocal (right), via the GNSDR, using LSPD, LSPDi, GSPD and GSPDi from left to right

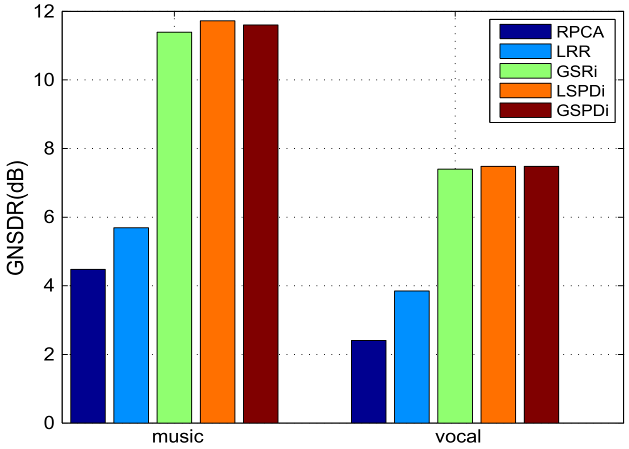

Fig. 5 Separation performance for the singing music (left) and vocal (right), via the GNSDR, using RPCA, LRR, GSRi, LSPDi and GSPDi from left to right

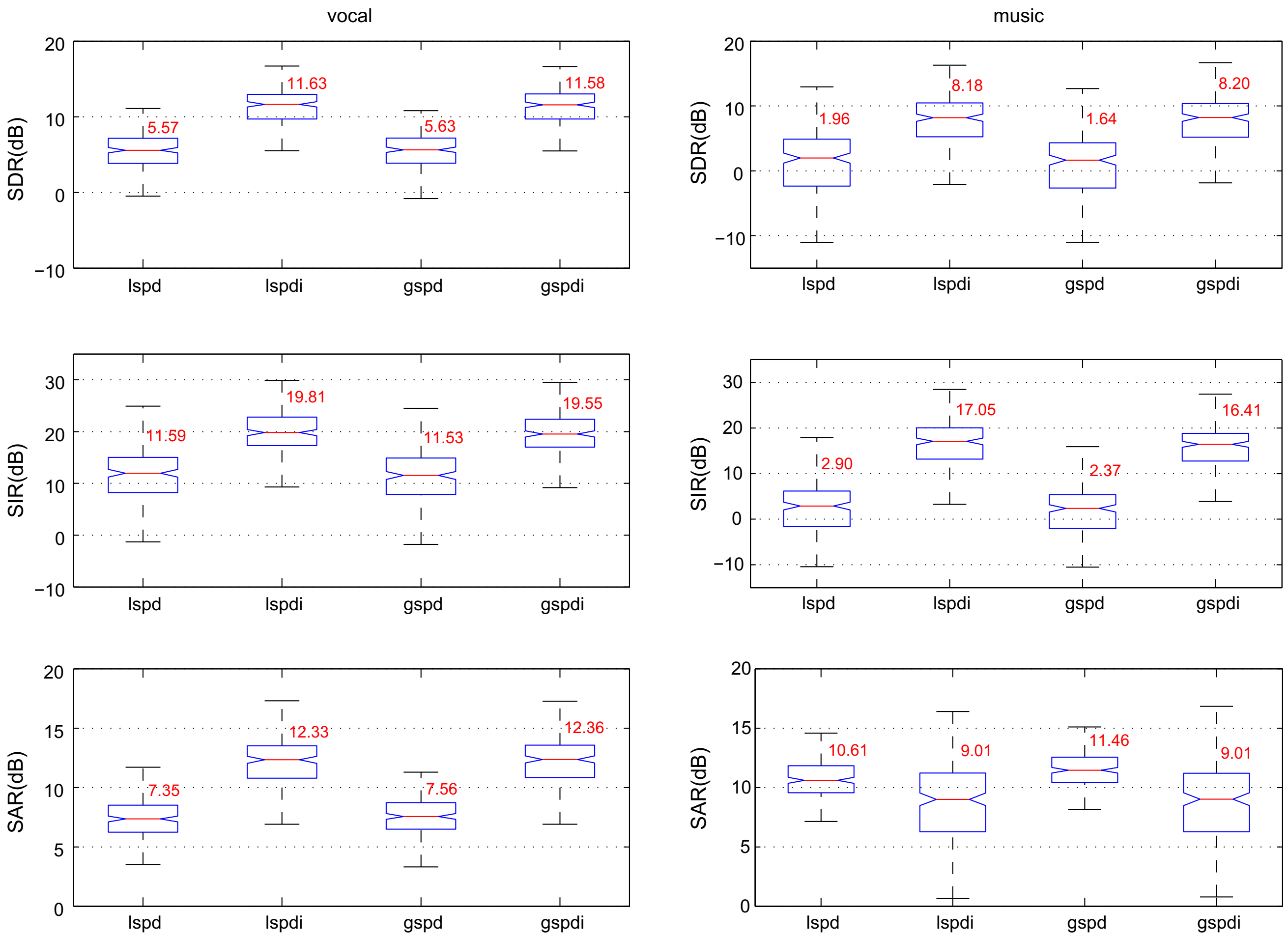

Fig. 4 Separation performance for the vocal (left) and music (right), via the SDR (top row), SIR (middle row) and SAR (bottom row), using LSPD, LSPDi, GSPD, and GSPDi from left to right. The central mark (red horizontal line) in each box is the median of the distribution. Higher values are better

Table 2 Separation quality for the vocal and music for the iKala dataset of RPCA, LRR, GSRi, LSPDi and GSPDi

	Vocal			Music		
	GSDR	GSIR	GSAR	GSDR	GSIR	GSAR
RPCA	6.21	8.14	12.53	0.75	3.23	7.00
LRR	7.66	10.75	11.72	1.96	4.13	8.70
GSRi	11.26	14.93	14.24	7.66	15.23	8.83
LSPDi	11.29	19.91	12.16	8.00	16.70	8.88
GSPDi	11.29	19.61	12.21	7.88	16.01	8.85

5.5 Algorithms for comparison

The proposed above methods are compared with the existing three state-of-the-art singing voice separation algorithms, as RPCA, unsupervised method proposed by Huang et al. [13], LRR, supervised method proposed by Liu et al. [17] and GSRi, supervised method proposed by Chan et al. [28]. In the three experimental algorithms, the original parameters are set as following, λ is set to $\frac{1}{\sqrt{\max(m,n)}}$ and γ is set to $\frac{2}{\sqrt{\max(m,n)}}$.

As shown in Fig. 5, the proposed method has a higher value of global normalized source-to-distortion ratio (GNSDR), which means that the introduction of prior knowledge improve the separation performance.

Several results are shown in Table 2. In the vocal part, proposed algorithm achieves higher GSDR and GSIR than RPCA, LRR and GSRi, which shows that LSPDi, GSPDi have better global separation performance and a better ability to remove the instrumental sounds than RPCA, LRR and GSRi. In the background music part, proposed algorithm achieves higher GSIR and GSAR than RPCA, LRR and GSRi, which suggests that LSPDi, GSPDi has better global separation performance than RPCA, LRR, GSRi and a better ability to remove the singing, a better performs in limiting artifacts during the separation process. The difference on GSDR and GSIR might be significant. So

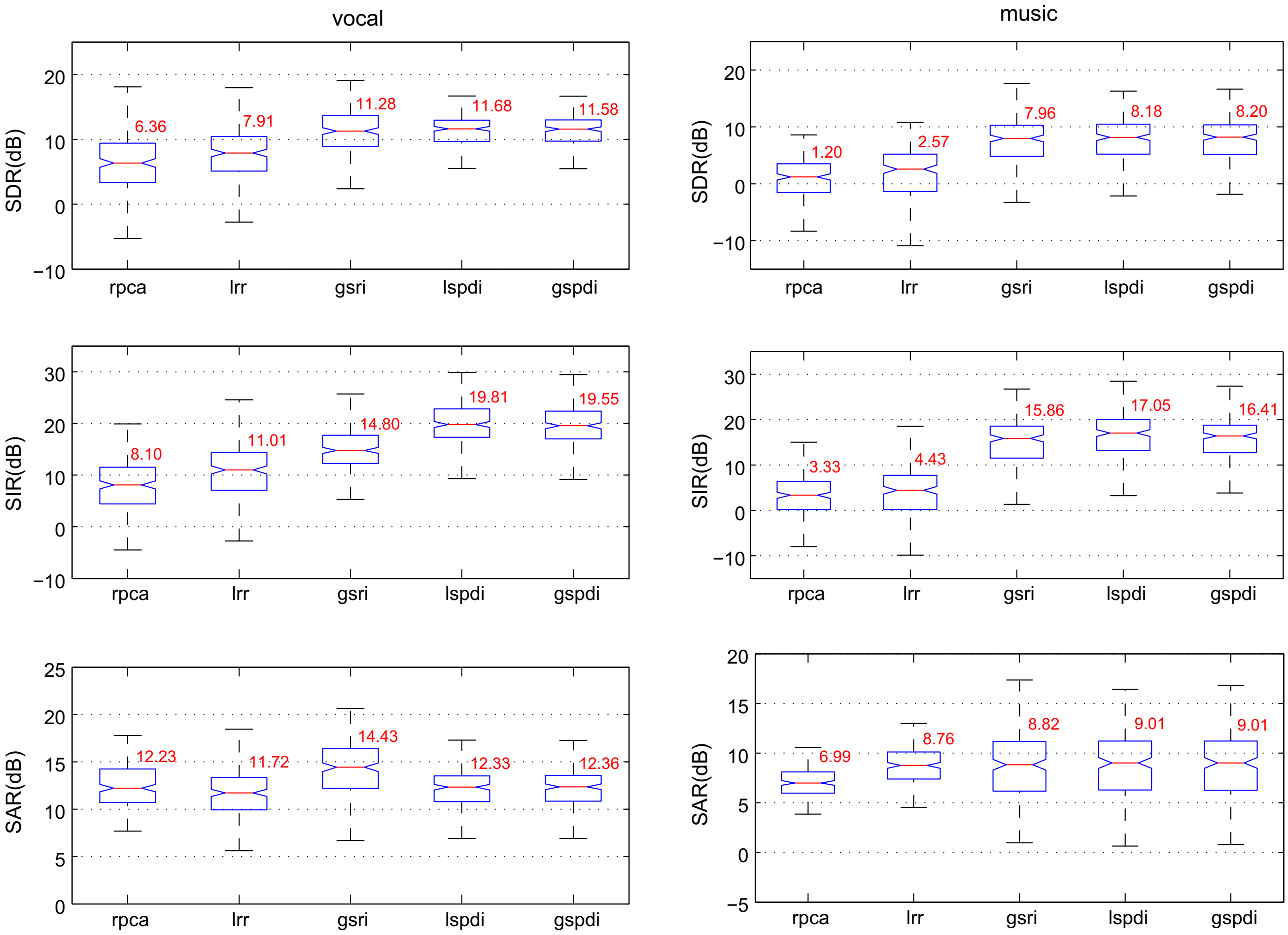

Fig. 6 Separation performance for the vocal (left) and music (right), via the SDR (top row), SIR (middle row) and SAR (bottom row), using RPCA, LRR, GSRi, LSPDi and GSPDi from left to right. The central mark (red horizontal line) in each box is the median of the distribution. Higher value is better

basically it suggests that the proposed method works better on the background music. The above observations show that LSPDi and GSPDi have a better ability to deal with the separation of singing voice and accompaniment.

Figure 6 shows the separation performance for the vocal and background music, respectively, on the iKala dataset, via the SDR (top row), SIR (middle row) and SAR (bottom row). In each column, the boxes from left to right represent Huang's method RPCA, Liu's method LRR, Chan's method GSRi and the proposed LSPDi and GSPDi, respectively.

From Fig. 6, we can see that for the separation of singing voice and music accompaniment, proposed algorithm achieves the highest SDR, SIR and SAR. These results indicate that our proposed LSPDi and GSPDi algorithm have a better overall separation performance for the singing voice and the music components (highest SDR), which exhibits their better capability of limiting artifacts and removing interferences during separation.

6 Conclusions

In this paper, we have presented two categories, time–frequency-based source separation algorithm for music signals. LSPD [22] and LSPDi consider both the vocal and instrumental spectrograms as sparse matrix and low-rank matrix, respectively. GSPD and GSPDi combine both the vocal and instrumental spectrograms as sparse matrix and group-sparse matrix, respectively. Moveover, the dictionaries for the singing voice and background music prelearned from isolated singing voice and background music training data, respectively, have successfully utilized to capture more features of vocal or background music spectrogram, and more prior information are introduced, for example, vocal annotations information. Future developments of the presented method could take advantage of more properties of background music and singing voice with such extensions, for example, some of the recent works [40–64] that employ signal classification maybe further motivate the need of this work.

Acknowledgements The authors acknowledge the insightful comments provided by the anonymous reviewers which have added much to the clarity of the paper. This study was funded by the National Natural Science Foundation of China under the Grant Nos. 11501351 and 11871039.

References

1. Vembu S, Baumann S (2005) Separation of vocals from polyphonic audio recordings. In: Proceedings of the international society for music information retrieval, pp 337–344
2. Ozerov A, Philippe P, Bimbot F, Gribonval R (2007) 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(5):1564–1578
3. Li Y, Wang DL (2007) Separation of singing voice from music accompaniment for monaural recordings. IEEE Trans Audio Speech Lang Process 15(4):1475–1487
4. Durrieu J, David B, Richard G (2011) A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE J Sel Top Signal Process 5(6):1180–1191
5. Berenzweig A, Ellis DPW, Lawrence S (2002) Using voice segments to improve artist classification of music. In: AES 22nd international conference on virtual, synthetic, and entertainment audio, pp 1–8
6. Zwan P, Kostek B (2008) System for automatic singing voice recognition. J Audio Eng Soc 56(9):710–723
7. Wang CK, Lyu RY, Chiang YC (2003) An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. In: Eurospeech, pp 1197–1200
8. Fujihara H, Goto M, Ogata J, Komatani K, Ogata T, Okuno HG (2006) Automatic synchronization between lyrics and music CD recordings based on viterbi alignment of segregated vocal signals. In Proceedings of the IEEE international symposium on multimedia, pp 257–264
9. Tzanetakis G, Martins L, McNally K, Jones R (2010) Stereo panning information for music information retrieval tasks. J Audio Eng Soc 58(5):409–417
10. Fujihara H, Goto M (2007) A music information retrieval system based on singing voice timbre. In Proceedings of the international society for music information retrieval, pp 467–470
11. Zhang Y, Duan Z (2016) Supervised and unsupervised sound retrieval by vocal imitation. J Audio Eng Soc 64(7):533–543
12. Candes EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? J ACM 58(3):1–37
13. Huang PS, Chen SD, Smaragdis P, Johnson MH (2012) Singing voice separation from monaural recordings using robust principal component analysis. In: Proceeding of the IEEE international conference on acoustics, speech and signal processing, pp 57–60
14. Yang YH (2012) On sparse and low-rank matrix decomposition for singing voice separation. In Proceedings of the ACM multimedia, pp 757–760
15. Su L, Yang YH (2013) Sparse modeling for artist identification: exploiting phase information and vocal separation. In: Proceeding of the international society for music information retrieval conference, pp 349–354
16. Papadopoulos H, Ellis DPW (2014) Music-content-adaptive robust principal component analysis for a semantically consistent separation of foreground and background in music audio signals. In: Conference on digital audio effects (DAFx-14), pp 1–8
17. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2013) Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35(1):171–184
18. Yang YH (2013) Low-rank representation of both singing voice and music accompaniment via learned dictionaries. In: Proceedings of the international society for music information retrieval conference, pp 427–432
19. Mairal J, Bach F, Ponce J Sapiro G (2009) Online dictionary learning for sparse coding. In: Proceedings of the international conference machine learning, pp 689–696
20. Chen Z, Ellis DPW (2013) Speech enhancement by sparse, low-rank and dictionary spectrogram decomposition. In: WASPAA, pp 1–4
21. Chen Z, Papadopoulos H, Ellis DPW (2014) Content-adaptive speech enhancement by a sparsely-activated dictionary plus low rank decomposition. In: HSCMA, pp 16–20

22. Yu S, Zhang H, Duan Z (2017) Singing voice separation by Low-rank and sparse spectrogram decomposition with pre-learned dictionaries. J Audio Eng Soc 65(5):377–388
23. Chen Z, Huang PS, Yang YH (2013) Spoken lyrics informed singing voice separation. In: Proceedings of the HAMR. http://labrosa.ee.columbia.edu/hamr2013/proceedings/doku.php/singing_separation
24. Chan TS, Yeh TC, Fan ZC, Chen HW, Su L, Yang YH, Jang R (2015) Vocal activity informed singing voice separation with the iKala dataset. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, pp 718–722
25. Lehner B, Widmer G, Sonnleitner R (2014) On the reduction of false positives in singing voice detection. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 7480–7484
26. Yoshii K, Fujihara H, Nakano T, Goto M (2014) Cultivating vocalactivitydetectionfor music audio signals in a circulation type crowdsourcing ecosystem. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 624–628
27. Ikemiya Y, Yoshii K, Itoyama K (2015) Singing voice analysis and editing based on mutually dependent F0 estimation and source separation. In: Proceedings of the IEEE international conference acoustics, speech signal process, pp 574–578
28. Chan TS, Yang YH (2017) Informed group-sparse representation for singing voice separation. IEEE Signal Process Lett 24(2):156–160
29. Chan TS, Yeh TC, Fan ZC, Chen HW, Sui L, Yang YH, Jang R (2015) Vocal activity informed singing voice separation with the iKala dataset. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 718–722
30. Rafii Z, Pardo B (2013) Repeating pattern extraction technique (REPET): a simple method for music/voice separation. IEEE Trans Audio Speech Lang Process 21(1):73–84
31. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3(1):1–122
32. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc B 68(1):49–67
33. Ma S (2016) Alternating proximal gradient method for convex minimization. J Sci Comput 68(2):546–572
34. Virtanen T, Mesaros A, Ryynanen M (2008) Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In: Proceedings of the ISCA tutorial and research workshop statistic perceptual audition, pp 17–20
35. Durrieu JL, David B, Richard G (2011) A musically motivated midlevel representation for pitch estimation and musical audio source separation. IEEE J Sel Top Signal Process 5(6):1180–1191
36. Ryynanen M, Virtanen T, Paulus J, Klapuri A (2008) Accompaniment separation and karaoke application based on automatic melody transcription. In: Proceedings of the IEEE international conference on multimedia expo, pp 1417–1420
37. Gribonval R, Benaroya L, Vincent E, Fvotte C (2003) Proposals for performance measurement in source separation. In: Proceedings of the international symposium, ICA BSS, pp 763–768
38. Ozerov A, Philippe P, Gribonval R, Bimbot F (2005) One microphone singing voice separation using source-adapted models. In: Proceedings of the IEEE workshop application signal processing to audio acoustics, pp 90–93
39. Ozerov A, Philippe P, Bimbot F, Gribonval R (2007) Adaptation of Bayesian models for single channel source separation and its application to voice/music separation in popular songs. IEEE Trans Audio Speech Lang 15(5):1564–1578
40. Hassan AR, Bhuiyan MIH (2016) Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. Biomed Signal Process Control 24:1–10
41. Hassan AR, Bhuiyan MIH (2016) Automatic sleep scoring using statistical features in the EMD domain and ensemble methods. Biocybernet Biomed Eng 36(1):248–255
42. Hassan AR, Bhuiyan MIH (2016) An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting. Neurocomputing 219(5):76–87
43. Hassan AR, Bhuiyan MIH (2016) A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. J Neurosci Methods 271:107–118
44. Hassan AR, Subasi A (2017) A decision support system for automated identification of sleep stages from single-channel EEG signals. Knowl-Based Syst 128:115–124
45. Hassan AR, Bhuiyan MI (2017) Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. Comput Methods Program Biomed 140:201–210
46. Hassan AR, Bhuiyan MIH (2015) Automatic sleep stage classification. In: 2015 2nd international conference on electrical information and communication technology (EICT), pp 211–216
47. Hassan AR, Bashar SK, Bhuiyan MIH (2015) On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In: International conference on advances in computing, communications and informatics, pp 2238–2243
48. Hassan AR, Bashar SK, Bhuiyan MIH (2015) Automatic classification of sleep stages from single-channel electroencephalogram. In: 2015 annual IEEE India conference (INDICON), pp 1–6
49. Hassan AR, Bhuiyan MIH (2015) Dual tree complex wavelet transform for sleep state identification from single channel electroencephalogram. In: 2015 IEEE international conference on telecommunications and photonics (ICTP), Dhaka, pp 1–5
50. Hassan AR (2016) Computer-aided obstructive sleep apnea detection using normal inverse Gaussian parameters and adaptive boosting. Biomed Signal Process Control 29:22–30
51. Hassan AR, Haque MA (2016) Computer-aided obstructive sleep apnea screening from single-lead electrocardiogram using statistical and spectral features and bootstrap aggregating. Biocybernet Biomed Eng 36:256–266
52. Hassan AR, Haque MA (2016) Computer-aided obstructive sleep apnea identification using statistical features in the EMD domain and extreme learning machine. Biomed Phys Eng Express 2(3):035003
53. Hassan AR (2015) Automatic screening of obstructive sleep apnea from single-lead electrocardiogram. In: 2015 international conference on electrical engineering and information communication technology (ICEEICT)
54. Hassan AR (2015) A comparative study of various classifiers for automated sleep apnea screening based on single-lead electrocardiogram. In: 2015 international conference on electrical and electronic engineering (ICEEE), Rajshahi, pp 45–48
55. Hassan AR, Haque MA (2016) Identification of sleep apnea from single-lead electrocardiogram. In: 2016 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC) and 15th international symposium on distributed computing and applications for business engineering (DCABES), Paris, pp 355–360
56. Hassan AR, Haque MA (2017) An expert system for automated identification of obstructive sleep apnea from single-lead ECG

using random under sampling boosting. Neurocomputing 235:122–130

57. Hassan AR, Haque MA (2016) Computer-aided sleep apnea diagnosis from single-lead electrocardiogram using dual tree complex wavelet transform and spectral features. In: 2015 international conference on electrical and electronic engineering (ICEEE), pp 49–52
58. Hassan AR, Siuly S, Zhang YC (2016) Epileptic seizure detection in EEG signals using tunable-Q factor wavelet transform and bootstrap aggregating. Comput Method Program Biomed 137:247–259
59. Hassan AR, Subasi A (2016) Automatic identification of epileptic seizures from EEG signals using linear programming boosting. Comput Methods Program Biomed 136:67–77
60. Hassan AR, Haque MA (2015) Epilepsy and seizure detection using statistical features in the complete ensemble empirical mode decomposition domain. In: TENCON 2015–2015 IEEE region 10 conference, Macao, pp 1–6
61. Hassan AR, Haque MA (2015) Computer-aided gastrointestinal hemorrhage detection in wireless capsule endoscopy videos. Comput Method Program Biomed 122(3):341–353
62. Bashar SK, Hassan AR, Bhuiyan MIH (2015) Identification of motor imagery movements from EEG signals using dual tree complex wavelet transform. In: 2015 international conference on advances in computing, communications and informatics (ICACCI), Kochi, pp 290–296
63. Bashar SK, Hassan AR, Bhuiyan MIH (2015) Motor imagery movements classification using multivariate EMD and short time Fourier transform. In: 2015 annual IEEE India conference (INDICON), New Delhi, pp 1–6
64. Hassan A, Huda MN, Sarker F, Mamun KA (2016) An overview of brain machine interface research in developing countries: opportunities and challenges. In: ICIEV, pp 396–401