Weighted Robust Principal Component Analysis with Gammatone Auditory Filterbank for Singing Voice Separation

Feng Li(✉) and Masato Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
{lifeng, akagi}@jaist.ac.jp

Abstract. This paper presents a proposed extension of robust principal component analysis (RPCA) with weighting (WRPCA) based on gammatone auditory filterbank for singing voice separation. Although the conventional RPCA is an effective method to separate singing voice and music accompaniment, it makes some strong assumptions. For example, drums may lie in the sparse subspace instead of being low-rank, which decreases the separation performance in many real-world applications, especially for drums existing in the mixture music signal. Accordingly, the proposed WRPCA method utilizes different weighted values between sparse (singing voice) and low-rank matrices (music accompaniment). In addition, we developed an extended RPCA on cochleagram using an alternative time-frequency (T-F) representation based on gammatone auditory filterbank. We also applied IBM/IRM estimation to improve the separation results. Evaluation results show that WRPCA achieves better separation performance than the conventional RPCA, especially for the IBM estimation method.

Keywords: Singing voice separation · Robust principal component analysis (RPCA) · Weighted · Gammatone auditory filterbank · Cochleagram · IBM/IRM estimation

1 Introduction

In recent years, singing voice separation has attracted considerable interest and attention in many real-world applications. It attempts to separate singing voice and music accompaniment parts of a music recording, which is a very significant technology for music information retrieval (MIR) [1] and chord recognition [2]. However, current state-of-the-art results are still far behind human hearing capability. The existing problems of singing voice separation are challenging [3].

Many previous studied methods have been proposed with the goal of overcoming the difficulty in separation tasks. Most of them have attempted to use the distinctive characteristic of each source and rarely studied the human auditory system. Huang *et al.* [4] proposed a robust principal component analysis

(RPCA) method for singing voice separation, which decomposed an input matrix into a sparse matrix plus a low-rank matrix. Inspired by a sparse and low-rank model, Yang [5] proposed new sparse and low-rank matrices that were based on the incorporation of harmonicity priors and a back-end drum removal procedure. He [6] also proposed a multiple low-rank representation (MLRR) to decompose a magnitude spectrogram into two low-rank matrices.

As stated above, RPCA is an effective method to separate singing voice from the mixture signal. It decomposes a given amplitude spectrogram (matrix) of a mixture signal into the sum of a low-rank matrix (music accompaniment) and a sparse matrix (singing voice). Since music instruments can reproduce the same sounds each time in the same music, so its magnitude spectrogram can be considered as a low-rank structure part. Singing voice, on the contrary, varies significantly and has a sparse distribution in the spectrogram domain owning to its harmonic structure part, resulting in a spectrogram with a sparse structure part. Although RPCA has been successfully applied to singing voice separation, it makes some strong assumptions. For example, drums may lie in the sparse subspace instead of being low-rank, which decreases the separation performance, especially for drums existing in the mixture music signal.

Even if all of the existing methods (e.g., RPCA and MLRR) can obtain acceptable separation results from mixture music signals, they ignore the features of the human auditory system, which plays a vital role in improving the quality of separation results. Recently a study was published hinting that cochleagram, as an alternative time-frequency (T-F) analysis based on gammatone filterbank, is more suitable than spectrogram for source separation [7]. This is because, cochleagram is derived from non-uniform T-F transform whereas T-F units in low-frequency regions have higher resolutions than in the high frequency regions, which closely resembles the functions of the human ear. Similarly, singing voice performances are quite different from music accompaniment on cochleagram. The spectral energy centralizes in a few T-F units for singing voice and thus can be assumed to be sparse. On the other hand, music accompaniment on the cochleagram has similar spectral patterns and structures that can be captured by a few basis vectors, so it can be hypothesized as a low-rank subspace. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of spectrogram.

To overcome the above-mentioned problems, we propose a weighted method to make sure different scale values are obtained to describe sparse and low-rank matrices. The method, which we call Weighted Robust Principal Component Analysis (WRPCA), chooses different weighted values between sparse and low-rank matrices. In addition, with the purpose of imitating the human auditory system, we adopt gammatone auditory filterbank as the first stage of WRPCA in cochleagram processing. Finally, in order to obtain better separation results, we further apply ideal binary mask (IBM) or ideal ratio mask (IRM) [8] to enforce the constraints between an input mixture signal and the output results.

The rest of this paper is organized as follows. In Section 2, we review the conventional RPCA and RPCA for singing voice separation. In Section 3, we

describe the proposed WRPCA on cochleagram and its application to mask estimation. In Section 4, we evaluate WRPCA on the ccMixture and DSD100 datasets. Finally, we draw conclusions and describe future work in Section 5.

2 Background

In this section, we briefly review the conventional RPCA and RPCA method for singing voice separation.

2.1 Principle of RPCA

Candés *et al*. [9] presented a convex program RPCA, which decomposed an input matrix $M \in \mathbb{R}_{m\times n}$ into the sum of a low-rank matrix $L \in \mathbb{R}_{m\times n}$ plus a sparse matrix $S \in \mathbb{R}_{m\times n}$. The problem can be formulated as follows:

$$\begin{aligned} &min\ |L|_* + \lambda|S|_1, \\ &s.t.\ M = L + S. \end{aligned} \tag{1}$$

where $|\cdot|_*$ denotes the nuclear norm (sum of singular values), $|\cdot|_1$ is the L_1-norm (sum of absolute values of matrix entries), and λ is a positive constant parameter between the low-rank matrix L and the sparsity matrix S. Candés *et al*. suggested $\lambda = 1/\sqrt{max(m,n)}$ [9]. Furthermore, this convex program can be solved by accelerated proximal gradient (APG) or augmented Lagrange multipliers (ALM) [10] (we used an inexact version of ALM in a baseline experiment).

2.2 RPCA for singing voice separation

Huang *et al*. assumed that the RPCA method can be applied to the task of separating singing voice and music accompaniment from the mixture music signal [4]. On account of the music accompaniment part, music instruments can reproduce the same sounds each time in the same music, so its magnitude spectrogram can be considered as a low-rank matrix structure. Singing voice part, in contrast, varies significantly and has a sparse distribution in the spectrogram domain due to its harmonic structure part, resulting in a spectrogram with a sparse matrix structure. Therefore, we can use the RPCA method to decompose an input matrix into a sparse matrix (singing voice) and a low-rank matrix (music accompaniment). However, it makes some strong assumptions. For instance, drums may lie in the sparse subspace instead of being low-rank, which decreases the separation performance, especially for drums existing in the mixture signal.

3 Proposed Method

In this section, we first explain the proposed WRPCA method and then describe its application to IBM/IRM estimation. Finally, we give a block diagram of a singing voice separation system.

Algorithm 1 WRPCA for Singing Voice Separation

Input: Mixture signal $M \in \mathbb{R}_{m\times n}$, weight w.
1: **Initialization**: $\rho, \mu_0, L_0 = M, J_0 = 0, k = 0$.
2: While not convergence **do** :
3: **repeat**
4: $S_{k+1} = \arg\min_S |S|_1 + \frac{\mu_k}{2}|M + \mu_k^{-1}J_k - L_k - S|_F^2$.
5: $L_{k+1} = \arg\min_L |L|_{w,*} + \frac{\mu_k}{2}|M + \mu_k^{-1}J_k - S_{k+1} - L|_F^2$.
6: $J_{k+1} = J_k + \mu_k(M - L_{k+1} - S_{k+1})$.
7: $\mu_{k+1} = \rho * \mu_k$.
8: $k \leftarrow k + 1$.
9: **end while.**
Output: $S_{m\times n},\ L_{m\times n}$.

3.1 Principle of WRPCA

WRPCA is an extension of RPCA, which has different scale values between sparse and low-rank matrices. The corresponding model can be defined as follows:

$$\begin{aligned} &min\ |L|_{w,*} + \lambda|S|_1, \\ &s.t.\ M = L + S. \end{aligned} \tag{2}$$

where $|L|_{w,*}$ is the low-rank matrix with different weighted values, while S is the sparse matrix. $M \in \mathbb{R}_{m\times n}$ is an input matrix, which consists of $L \in \mathbb{R}_{m\times n}$ and $S \in \mathbb{R}_{m\times n}$, and $\lambda > 0$ is a trade-off constant parameter between the sparse matrix S and the low-rank matrix L. We used $\lambda = 1/\sqrt{max(m,n)}$ as suggested by Candés *et al.* [9]. We also adopted an efficient inexact version of the augmented Lagrange multiplier (ALM) [10] to solve this convex model. The corresponding augmented Lagrange function is defined as follows:

$$\begin{aligned} J(M, L, S, \mu) = \ & |L|_w, * + \lambda|S|_1 + < J, M - L - S > \\ & + \frac{\mu}{2}|M - L - S|_F^2. \end{aligned} \tag{3}$$

where J is the Lagrange multiplier and μ is a positive scaler. The process corresponding to mixture music signal separation can be seen in **Algorithm 1** WRPCA for singing voice separation. The value of M is a mixture music signal from the observed data. After the separation by using WRPCA, we can obtain a sparse matrix S (singing voice) and a low-rank matrix L (music accompaniment).

3.2 Weighted values

In this paper, we mainly focus on the fact that nuclear norm minimization and L_1-norm affect not only the sparsity and low-rankness of the two decomposed matrices, but also their relative scales. However, the RPCA method simply ignores the differences between the scales of the sparse and low-rank matrices.

In order to solve this problem, and inspired by the success of weighted nuclear norm minimization [11], we adopted different weighted value strategies to trim the low-rank matrix during the singing voice separation processing. This enables the features of the separated matrices to be better represented.

Lemma 1. Set $M = U \sum V^T$ as the singular value decomposition (SVD) of $M \in \mathbb{R}_{m \times n}$, where

$$\sum = \begin{pmatrix} diag(\delta_1(M), \delta_2(M), ..., \delta_n(M)) \\ 0 \end{pmatrix}, \tag{4}$$

and $\delta_i(M)$ denotes the i-th singular value of M. If the positive regularization parameter C exists and the positive value $\varepsilon < min(\sqrt{C}, \frac{C}{\delta_1(M)})$ holds, by using the reweighting formula $W_i^l = \frac{C}{\delta_i(L_l)+\varepsilon}$ [12] with initial estimation $L_0 = M$, the reweighted problem has the closed-from solution:
$L^* = U \sum' V^T$, where

$$\sum' = \begin{pmatrix} diag(\delta_1(L^*), \delta_2(L^*), ..., \delta_n(L^*)) \\ 0 \end{pmatrix}, \tag{5}$$

and

$$\delta_i(L^*) = \begin{cases} 0 \\ \dfrac{c_1 + \sqrt{c_2}}{2} \end{cases} \tag{6}$$

where $c_1 = \delta_i(M) - \varepsilon$ and $c_2 = (\delta_i(M) + \varepsilon)^2 - 4C$. Gu *et al.* [11] described a more specific proof of ***Lemma 1***. In our experiments, we empirically set the regularization parameter C as the maximum matrix size, which enabled us to obtain the best separation performance results, e.g., $C = max(m, n)$.

3.3 Application to mask estimation

After obtaining the separation results of sparse S and low-rank matrices L by using WRPCA, we applied IBM/IRM estimation to further improve the separation performance. A block diagram of the singing voice separation system is illustrated in Fig. 1. It consists of two stages: WRPCA on cochleagram and singing voice separation based on IBM/IRM estimation. The first stage performs the cochlear analysis with gammatone filter, calculates the cochleagram of the mixture music signal, and then decomposes matrixes into sparse and low-rank matrices by using WRPCA. The second stage applies IBM/IRM estimation to improve the separation results. The IBM and IRM are defined as [8]

$$M_{ibm} = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \tag{7}$$

and

$$M_{irm} = \frac{S_{ij}}{S_{ij} + L_{ij}} \tag{8}$$

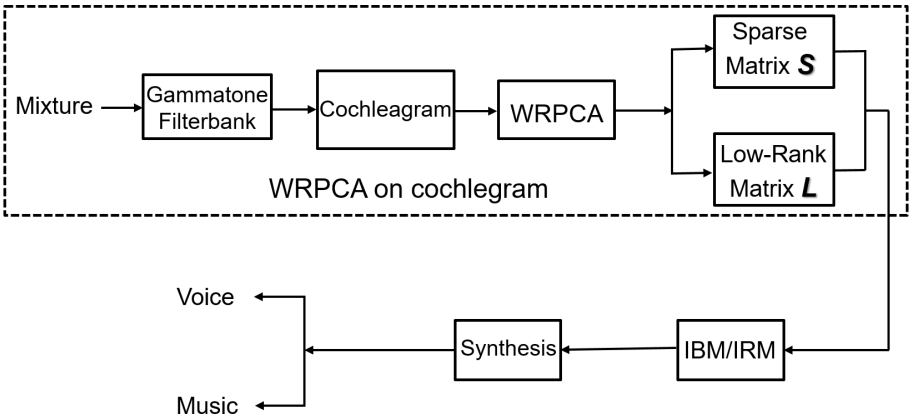

Fig. 1. Block diagram of singing voice separation system.

where M_{ibm} and M_{irm} are the values of IBM estimation and IRM estimation, respectively. S_{ij} and L_{ij} are the values of the sparse and low-rank matrices. The separated matrices can be synthesized as described by Wang *et al.* [13].

4 Experimental Evaluation

In this section, we show how evaluated WRPCA by using two different datasets: ccMixture[1] and DSD100[2], and how we compared it with the conventional RPCA.

4.1 Experimental datasets

In our experiments, we used two different datasets to evaluate WRPCA. The first was the ccMixter dataset, for which we chose 43 full stereo songs with only 30-second fragments (from 0'30" to 1'00") at the same time of each song, which is the maximum period of all songs containing singing voice. Each audio contains three parts: singing voice, music accompaniment, and a mixture of them.

The second was the Demixing Secrets Dataset 100 (DSD100), which was also used in the 2016 Signal Separation Evaluation Campaign (SiSEC) [3]. To reduce computations, we adopted only 30-second fragments (from 1'45" to 2'15") at the same time for all songs, which comprised 36 development songs and 46 test songs. Because there are four sources (bass, drums, vocals and others) for each track, we considered the sum of bass, drums and others was the music accompaniment.

4.2 Experiment conditions

We mainly focused on monaural source separation in our experiments. This is even more difficult than multichannel source separation since only a single chan-

[1] http://www.ccmixter.org/
[2] http://liutkus.net/DSD100.zip

nel is available. We downmixted the two-channel stereo mixtures into a single mono channel and obtained an average value for each channel. All experiment data were sampled at 44.1 kHz. We set parameters for cochleagram analysis: 128 channels, 40∼11025 Hz frequency range, and 256 frequency length. To compare the results with those obtained with WRPCA, we calculated the input feature by using short-time Fourier transform (STFT) and inverse STFT (ISTFT), which is a part of contrast experiments that have been performed on spectrogram for conventional RPCA and WRPCA. We used a window size of 1024 samples, a hop size of 256 samples for the STFT and an FFT size of 1024.

To confirm the effectiveness of WRPCA, we assessed its quality of separation in terms of the source-to-distortion ratio (SDR) and the source-to-artifact ratio (SAR) by using the BSS-EVAL 3.0 metrics [14] and the normalized SDR (NSDR). The estimated signal $\hat{S}$(t) is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t). \tag{9}$$

where $S_{target}(t)$ is the allowable deformation of the target sound, $S_{interf}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and $S_{artif}(t)$ is an artifact term that may correspond to the artifact of the separation method. The SDR, SAR and NSDR are defined as

$$SDR = 10log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \{e_{interf}(t) + e_{artif}(t)\}^2}. \tag{10}$$

$$SAR = 10log_{10}\frac{\sum_t \{S_{target}(t) + e_{interf}(t)\}^2}{\sum_t e_{artif}(t)^2}. \tag{11}$$

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v). \tag{12}$$

where $\hat{v}$ is the separated voice part, v is the original clean signal, and x is the original mixture. The NSDR is used to estimate the overall improvement in the SDR between x and $\hat{v}$.

The higher values of the SDR, SAR and NSDR represent that the method exhibits better separation performance of source separation. The SDR represents the quality of the separated target sound signals. The SAR represents the absence of artificial distortion. All the metrics are expressed in dB.

4.3 Experiment results

To examine WRPCA, we first evaluated it on the ccMixture dataset. Fig. 2 shows the comparison results of conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. The first four methods (RPCA, RPCA with IRM, RPCA with IBM and WRPCA) are calculated on spectrogram (without gammatone filterbank), while WRPCA with IRM and WRPCA with IBM are calculated on cochleagram (with gammatone filterbank). From the experiment results obtained with the SDR and SAR, we can see that WRPCA gets better results on the ccMixture dataset, especially

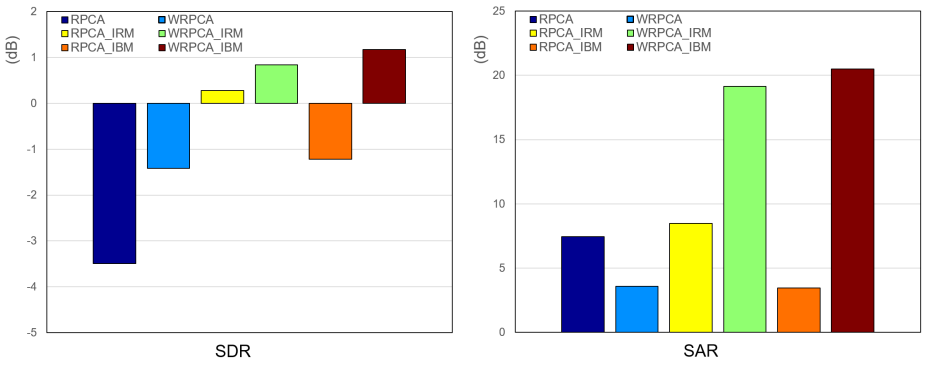

Fig. 2. Comparison of singing voice separation results on **ccMixture** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively.

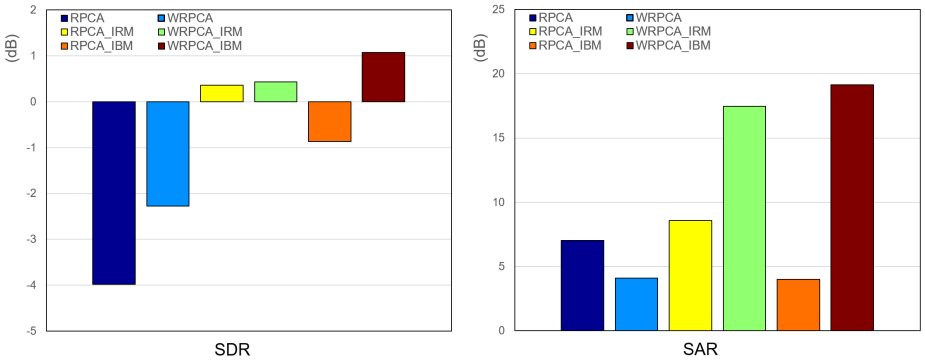

Fig. 3. Comparison of singing voice separation results on **DSD100** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively.

for the IBM estimation (with gammatone filterbank). In contrast, the conventional RPCA got worse results than the others. We also evaluated WRPCA on the DSD100 dataset. Fig. 3 shows the comparison results obtained with the conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. The results clearly show that WRPCA obtains better separation results on the DSD100 dataset, especially for the IBM estimation (with gammatone filterbank). However, the opposite results were obtained with the conventional RPCA. In terms of the SAR in Fig. 2 and Fig. 3, WRPCA with IBM on cochleagram (with gammatone filterbank) attained higher values than others, while the RPCA with IBM (without gammatone filterbank) had the worst values among them.

The NSDR provides overall improvement in the SDR; in other words, it provides better separation performance in singing voice separation. Fig. 4 shows

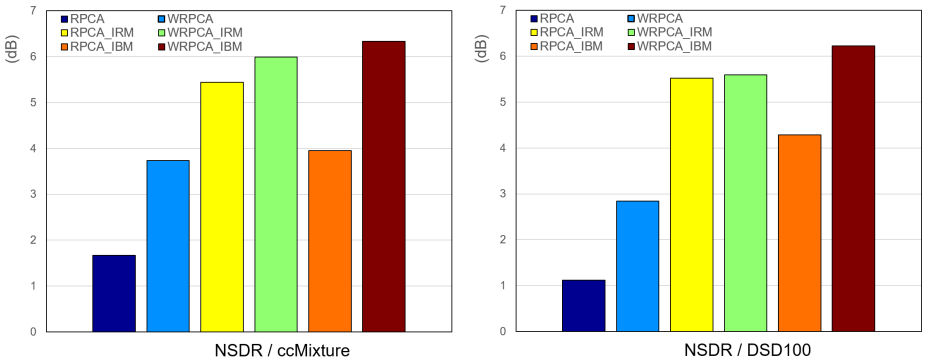

Fig. 4. Comparison of singing voice separation results between **ccMixture** and **DSD100** datasets among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. Note that SDRs for the original datasets, ccMixture and DSD100, are -5.16dB and -5.11dB, respectively.

the NSDR results we obtained with WRPCA on the ccMixture and DSD100 datasets. The results show that the best performance was achieved by WRPCA with IBM (with gammatone filterbank).

Therefore, from the results of Fig. 2, Fig. 3 and Fig. 4, we can confirm that WRPCA on cochleagram provides better separation performance than RPCA on spectrogram under the same conditions with or without IBM or IRM. Moreover, WRPCA provided better results than RPCA without gammatone filterbank and IBM/IRM. We also can see that WRPCA on cochleagram with IBM (with gammatone filterbank) provides better separation results in all evaluation standard methods. However, RPCA with IBM does not provide values as good as those provided by RPCA with IRM.

5 Conclusions and Future Work

In this paper, we proposed an extension of RPCA with weighting on cochleagram (WRPCA). It is based on gammatone auditory filterbank and application to IBM/IRM estimation for singing voice separation. The cochleagram of the mixture signal was decomposed into sparse (singing voice) and low-rank matrices (music accompaniment) by using WRPCA, then IBM/IRM estimation was utilized to improve the separation results. Experimental results obtained on the ccMixture and DSD100 datasets confirmed that WRPCA outperforms the conventional RPCA method in singing voice separation tasks, especially for WRPCA on cochleagram based on gammatone auditory filterbank with IBM estimation. In future work, since prior information (e.g., melody annotations) and spatial information (e.g., localization and isolation) are very significant for separating singing voice from mixture music signals, we will attempt to fuse all of them to improve the separation performance.

Acknowledgments. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan Scholarship and the China Scholarship Council (CSC) Scholarship.

References

1. Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.: Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96(4), 668-696 (2008)
2. Fujishima, T.: Real-time chord recognition of musical sound: A system using common lisp music. In: Proceedings of the International Computer Music Conference (ICMC), 464-467 (1999)
3. Liutkus, A., Stöter, F., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontecave, J.: The 2016 signal separation evaluation campaign. In: Proceedings of 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017), Grenoble, France, 323-332 (2017)
4. Huang, P.S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M.: Singing-voice separation from monaural recordings using robust principal component analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP), 57-60 (2012)
5. Yang, Y.H.: On sparse and low-rank matrix decomposition for singing voice separation. In: Proceedings of the 20th ACM international conference on Multimedia. ACM, 757-760 (2012)
6. Yang, Y.H.: Low-rank representation of both singing voice and music accompaniment via learned dictionaries. In: Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), 427-432 (2013)
7. Gao, B., Woo, W.L., Dlay, S.S.: Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and itakura-saito nonnegative matrix two-dimensional factorizations. IEEE Transactions on Circuits and Systems I: Regular Papers, 60(3), 662-675 (2013)
8. Li, Y.P., Wang, D.L.: On the optimality of ideal binary time-frequency masks. Speech Communication, 51(3), 230-239 (2009)
9. Candés, E., Li, X.D., Ma, Y., Wright, J.: Robust principal component analysis?. Journal of the ACM, 58(3), 11:1-11:37 (2011)
10. Lin, Z.C., Chen, M.M., Ma, Y.: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report, UILU-ENG-09-2215 (2009)
11. Gu, S.H., Xie, Q., Meng, D.Y., Zuo, W.M., Feng, X.C., Zhang, L.: Weighted nuclear norm minimization and its applications to low level vision. International Journal of Computer Vision, 121(2), 183-208 (2017)
12. Candés, E., Wakin, M., Boyd, S.: Enhancing sparsity by reweighted l_1 minimization. Journal of Fourier Analysis and Applications, 14(5-6), 877-905 (2008)
13. Wang, D.L., Brown, G.J.: Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press (2006)
14. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 14(4), 1462-1469 (2006)