

Singing Voice Separation Using RPCA with Weighted l_1 -norm

Il-Young Jeong and Kyogu Lee^(✉)

Music and Audio Research Group, Seoul National University,
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
kglee@snu.ac.kr

Abstract. In this paper, we present an extension of robust principal component analysis (RPCA) with weighted l_1 -norm minimization for singing voice separation. While the conventional RPCA applies a uniform weight between the low-rank and sparse matrices, we use different weighting parameters for each frequency bin in a spectrogram by estimating the variance ratio between the singing voice and accompaniment. In addition, we incorporate the results of vocal activation detection into the formation of the weighting matrix, and use it in the final decomposition framework. From the experimental results using the DSD100 dataset, we found that proposed algorithm yields a meaningful improvement in the separation performance compared to the conventional RPCA.

Keywords: Singing voice separation · Robust principal component analysis · Weighted l_1 -norm minimization

1 Introduction

Singing voice separation (SVS), or separating singing voice and accompaniment from a musical mixture is a challenging task. Many of the previous studies have attempted to use the distinctive characteristics of each source: fundamental frequency (f_0) and its harmonic structure of singing voice [11], repeatability [12], spectral/temporal continuity [5, 14], and so on.

Huang *et al.*, on the other hand, proposed to use a low-rank/sparse model for singing voice separation [4]. Approaches based on the low-rank/sparse model assume that accompaniment in music is usually repetitive because the number of instruments and notes in the accompaniment is limited. It is therefore presumed that the spectrogram of the accompaniment can be represented as a low-rank matrix. On the other hand, singing voice can be expressed as a sparse matrix because most of energy is concentrated on the f_0 trajectory and its harmonics. Based on these observations, robust principal component analysis (RPCA) [2] that decomposes a matrix into low-rank and sparse parts, was applied to separate singing voice and accompaniment in a mixture [4].

Although RPCA has been successfully applied to SVS, there is still plenty of room for improvement. Numerous studies have tried to extend the basic

RPCA-based approach. Sprechmann *et al.* presented a robust nonnegative matrix factorization, where an accompaniment spectrogram is represented by a combination of a few nonnegative spectra [13]. Jeong and Lee tried to extend RPCA by generalizing the nuclear norm and l_1 -norm to Schatten- p norm and l_p -norm, respectively, and suggested the appropriate value of p , for SVS in particular [6]. Chan *et al.* imposed additional vocal activation information to RPCA to remove the singing voice in the non-vocal frames [3].

In this paper we focus on the fact that minimization of the nuclear norm and l_1 -norm affects not only the low-rankness and sparsity of two decomposed matrices, but also their relative scale. Therefore, if prior information of their relative scale is known, it can be utilized in matrix decomposition by controlling the relative importance between the nuclear and l_1 -norm minimization terms. Furthermore, each time-frequency component of the spectrogram might have different prior, so we have to apply different weights to each element.

In our work, we construct a weighting matrix using two distinctive features: (1) frequency-dependent variance ratio between accompaniment and singing voice, and (2) the presence of singing voice, which is obtained by conducting a simple vocal activity detection (VAD) algorithm. In doing so, we go through a two-stage process that VAD is performed on the pre-separated singing voice, followed by the re-separation stage using updated the weighting matrix.

2 Algorithm

2.1 Robust Principal Component Analysis

Ideally, the low-rank and the sparse components can be decomposed from their mixture by solving the following optimization problem:

$$\begin{aligned} & \text{minimize} \quad \text{rank}(L) + \lambda \text{nonzero}(S), \\ & \text{s.t.} \quad L + S = M, \end{aligned} \tag{1}$$

where $M \in \mathbb{R}^{F \times T}$, $L \in \mathbb{R}^{F \times T}$, and $S \in \mathbb{R}^{F \times T}$ are the mixture, low-rank, and sparse matrix, respectively. $\text{rank}(\cdot)$ and $\text{nonzero}(\cdot)$ denote the rank and the number of nonzero components in a matrix, respectively. λ denotes the relative weight between two terms. Since above objective function is difficult to solve, Candès *et al.* presented its convex relaxation, or RPCA, as follows [2]:

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

where $|\cdot|_*$ and $|\cdot|_1$ denote the nuclear norm (sum of singular values) and l_1 -norm (sum of the absolute values of matrix elements), respectively. These properly approximate $\text{rank}(\cdot)$ and $\text{nonzero}(\cdot)$ in Eq. (1) and allow to solve it in a convex formulation. As in Eq. (1), λ decides the relative importance between two norms. Candès *et al.* suggested $\lambda = 1/\sqrt{\max(F, T)}$ [2], and Huang *et al.* generalized it as $\lambda = k/\sqrt{\max(F, T)}$ with a parameter k [4].

2.2 RPCA with Weighted l_1 -norm

Since λ in Eq. (2) is a global parameter for all the element of M , or $M_{f,t}$, once its value is decided then all $M_{f,t}$ have the same importance for the low-rankness of $L_{f,t}$ and the sparsity of $S_{f,t}$. However, it is not always proper in actual situation, and might be too simple. For example, if we know that $L_{f,t} = 0$ for some (f, t) , we may able to choose the value of λ to be $\lambda = 0$ for those element. If $S_{f,t} = 0$, on the contrary, we may set $\lambda \rightarrow \infty$. To apply the different weight for each element, we present RPCA with weighted l_1 -norm, or weighted RPCA (wRPCA), which replace λ to the weighting matrix A as:

$$\begin{aligned} & \text{minimize} \quad |L|_* + |A \otimes S|_1, \\ & \text{s.t.} \quad L + S = M, \end{aligned} \tag{3}$$

where \otimes denotes the element-wise multiplication operator. Note that $|A \otimes S|_1$ is a weighted l_1 -norm of S , which has been presented in a number of previous studies [1, 7]. To solve Eq. (3), optimization method for RPCA such as augmented Lagrangian multiplier (ALM) method can be directly used, just by replacing λ to A .

3 Singing Voice Separation

3.1 SVS Using RPCA

Huang *et al.* suggested that RPCA can be applied to separate the singing voice and the accompaniment from music signal [4]. In the case of music accompaniment, instruments often reproduce the same sounds in the same music, therefore its magnitude spectrogram can be represented as a low-rank matrix. On the contrary, singing voice has a sparse distribution in the spectrogram domain due to its strong harmonic structure. Therefore, M , L , and S in Eq. (2) can be considered as a spectrogram of the input music, accompaniment, and singing voice, respectively. After the separation is done in the spectrogram domain, the waveform for each source is obtained by directly applying the phase of the original mixture.

3.2 Proposed Method: SVS Using wRPCA

We extended previous RPCA-based SVS framework, by using wRPCA instead of RPCA in particular. We refer several previous studies to design the separation framework [3, 8, 9].

Nonnegativity Constraint. At first, we added a nonnegativity constraint in Eq. (3) as follows:

$$\begin{aligned} & \text{minimize} \quad |L|_* + |A \otimes S|_1, \\ & \text{s.t.} \quad L + S = M, \quad L \geq 0, \quad S \geq 0. \end{aligned} \tag{4}$$

This constraint prevent that large value of $A_{f,t}$ makes large negative value for S . The optimization of Eq. (4) is similar as of Eq. (2) or Eq. (3) but L and S are rectified as $x \leftarrow \max(x, 0)$ in every iteration.

Two-Stage Framework Using VAD. There were two opposite studies on SVS and VAD. Chan *et al.* suggested that additional vocal activity information can improve SVS [3]. On the other hand, Lehner and Widmer suggested that SVS can improve the accuracy of VAD algorithm [10]. To apply both of these suggestions, we conducted the two-stage framework as follows. At the first stage, the sources are separated without vocal activity information. Next, vocal activity is detected using the separated singing voice. In the second separation stage, the sources are separated again with detected vocal activity information. We basically used VAD algorithm presented by Lehner *et al.* which uses well-designed mel-frequency cepstral coefficients (MFCC) as features [8]. In addition, we also used the vocal variance features which were also proposed in their other studies [9]. For the classification, we used random forest with 500 trees, and used threshold of 0.55. As a post-processing step, median filtering was applied to the frame-wise classification results with 7 frames filter length (1.4s). Note that above framework is also based on the previous study [8]. Because the temporal resolution of spectrogram and VAD might be different, we aligned them by considering those absolute time indices so that we can obtain the frame-wise VAD results.

Choosing the Value for Λ . We choose the value of Λ as follows. At first, we decompose Λ as

$$\Lambda = k\lambda\Delta, \quad (5)$$

where λ is $1/\sqrt{\max(F,T)}$ suggested by Candès *et al.* [2], and k is a global parameter used by Huang *et al.* [4]. In this work, we empirically set it to be $k = 0.6$. Δ is a element-wise weighting matrix which is our main interest.

To select the appropriate value for Δ , we basically focused on the fact that Δ should be smaller when singing voice is relatively stronger than accompaniment, and be larger in the opposite case. If we try to set the frequency-wise weight, therefore it might be reasonable to use the ratio of their variance as

$$\Delta_{f,t} = \frac{b_A(f)}{b_V(f)}, \quad (6)$$

where $b_A(f)$ and $b_V(f)$ are the variances of the accompaniment and singing voice, respectively, in f -th frequency bin. Assuming both singing voice and accompaniment have the Laplacian distribution, they can be estimated by calculating the l_1 -norm for each frequency bin in the training data as follows:

$$\begin{aligned} b_A(f) &= \sum_t |A_{f,t}|, \\ b_V(f) &= \sum_t |V_{f,t}|, \end{aligned} \quad (7)$$

where A and V are the training data of the accompaniment and singing voice, respectively, that all the spectrograms of tracks in the training set are concatenated over time. Note that we assume that both accompaniment and singing voice for training are from the same music, those therefore have the same time length.

This variance ratio might be different when only vocal-activated frames are estimated. At least it will be smaller than Eq. (6) in overall, since all the non-vocal frames where singing voice is absent are excluded. In addition, since we know that there is no singing voice in the non-vocal frames, we can set the weight for those frames to infinite so the singing voice can be successfully eliminated. Consequently, we set $\hat{\Delta}$ for the second separation stage as follows:

$$\hat{\Delta}_{f,t} = \begin{cases} \frac{\hat{b}_A(f)}{\hat{b}_V(f)}, & \text{if } p(t) = 1, \\ \infty, & \text{otherwise,} \end{cases} \quad (8)$$

where $p(t)$ is the vocal activity information for the t -th frame: $p(t) = 1$ for the vocal-activated frames and 0 for the non-vocal ones. $\hat{b}_A(f)$ and $\hat{b}_V(f)$ are similar as $b_A(f)$ and $b_V(f)$, respectively, but estimated from the vocal-activated frames only as

$$\begin{aligned} \hat{b}_A(f) &= \sum_t |\hat{A}_{f,t}|, \\ \hat{b}_V(f) &= \sum_t |\hat{V}_{f,t}|, \end{aligned} \quad (9)$$

where \hat{A} and \hat{V} are the excerpts of A and V , respectively, which include the vocal-activated frames ($p(t) = 1$) only.

Handling Multi-channel Signals. Real-world music data are mostly provided in a multi-channel format *e.g.* stereo. Although the spatial information is helpful for better separation results, it is beyond the scope of this work. Therefore, the tracks are mixed down to a single-channel format. We simply took an average of spectrograms over channel and perform RPCA (or wRPCA) to this averaged spectrogram. We were concerned that the data is spatially biased if we take an average of waveform (center enhanced) or perform the algorithms to each channel separately (left/right enhanced). After the separation of $M = L + S$ is done, the separated singing voice and accompaniment of original multi-channel signal is obtained by using the Wiener-like filter (or soft mask) as $L/(L + S)$ for the accompaniment or $S/(L + S)$ for the singing voice for each channel.

4 Experimental Results

We applied our SVS algorithm to the dataset and the evaluation criteria from sixth community-based signal separation evaluation campaign (SiSEC 2016): professionally-produced music recordings (MUS) [16]. This campaign provided

Demixing Secrets Dataset 100 (DSD100), which consist 50 tracks for training ('dev') and other 50 for testing ('test'). All the tracks are sampled at 44.1 kHz and have stereo channels. Because there are 4 sources (vocals, bass, drums, and others) for each track, we considered the sum of bass, drums, and others as accompaniment. We used the dev set only to set Λ and $\hat{\Lambda}$, and even to train the VAD algorithm. In our experiments, VAD scores 0.87 F-score and 84% accuracy from the test set. As the evaluation criteria, it measures signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) based on BSS-Eval [15]. To generate the spectrogram of music, we took the magnitude of short-time Fourier transform with Hanning window of 4096 samples and half overlap.

Figure 1 shows the comparison of conventional RPCA, wRPCA, and two-stage wRPCA with VAD, and Table 1 shows the numerical values of the median of SDR. From this result, we can find that the proposed wRPCA improve SDR score

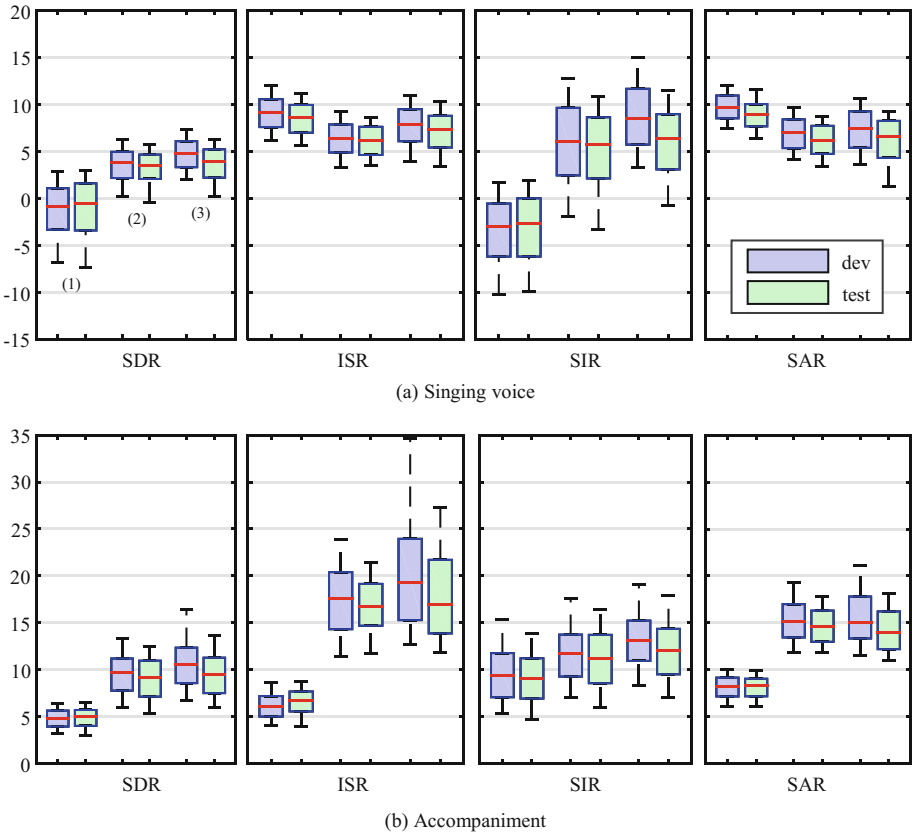
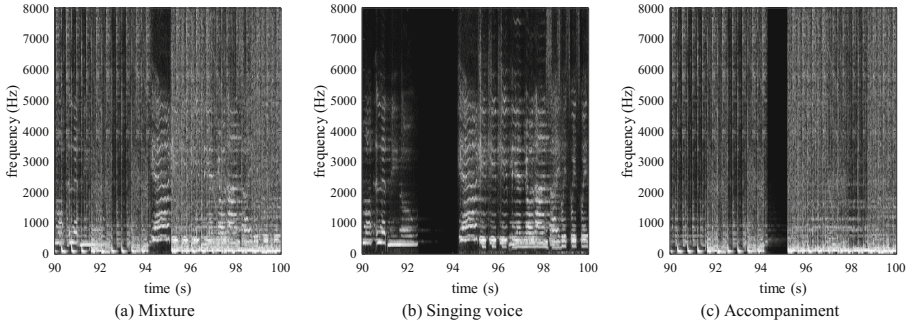
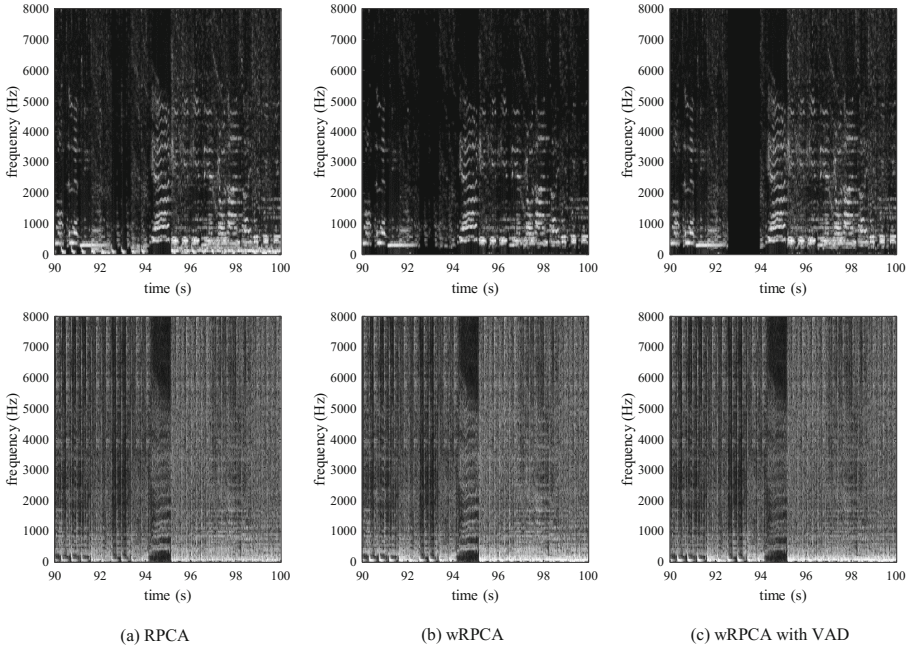


Fig. 1. Comparison of singing voice separation results using (1) conventional RPCA [4], (2) proposed wRPCA, and (3) wRPCA with VAD.

Table 1. Numerical values of median SDR in Fig. 1.

SDR(dB)	dev			test		
	RPCA	wRPCA	wRPCA w/VAD	RPCA	wRPCA	wRPCA w/VAD
Singing voice	-0.83	3.80	4.74	-0.51	3.54	3.92
Accompaniment	4.78	9.68	10.52	5.00	9.13	9.45

**Fig. 2.** Log-spectrograms of example mixture, singing voice, and accompaniment. Audio clips are excerpted from ‘AM Contra - Heart Peripheral’ in the dev set of DSD100.**Fig. 3.** Log-spectrograms of separated singing voice (top) and accompaniment (bottom). Input mixture is same as in Fig. 2.

for both singing voice and accompaniment, and even VAD does. However, the improvement from VAD is considerably degraded in the test set compared to the dev set. Considering that VAD for dev data makes almost perfect accuracy since it is trained by itself, we can expect that the better VAD algorithm is required to maximize its effectiveness. Example results are shown in Figs. 2 and 3. Compared to the conventional RPCA, it is observed that wRPCA successfully improve the separation quality, especially in the low-frequency region, and even VAD does in the non-vocal frames in particular. Audio files are demonstrated at http://margin.snu.ac.kr/svs_wrpca.

5 Discussion

Since the main contribution of our work is the use of Λ and $\hat{\Lambda}$, more accurately, Δ and $\hat{\Delta}$, we discuss in depth about the characteristics of them. Figure 4 shows the plots of $(\frac{b_A(f)}{b_V(f)})^{-1}$ and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ where $(\cdot)^{-1}$ is for visibility. Higher value means that the singing voice is stronger than the accompaniment in that frequency bin. What follows are several interesting insights we found from these plots.

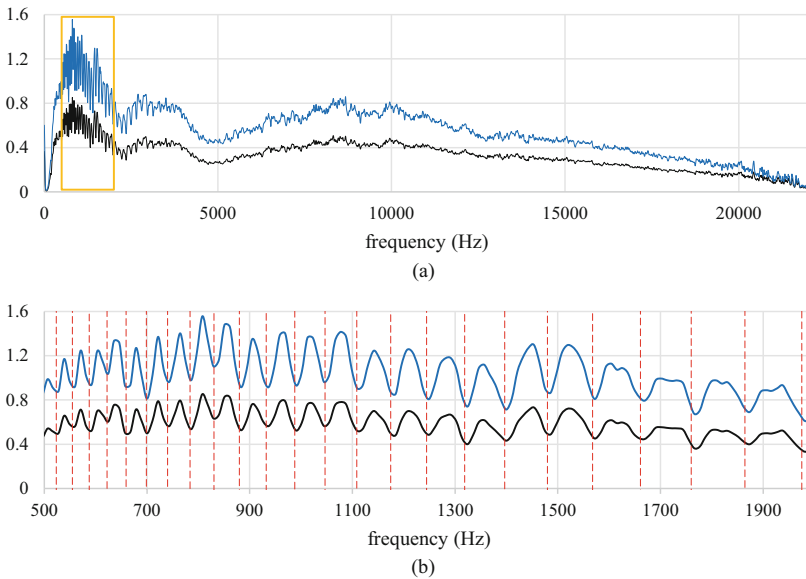


Fig. 4. (a) $(\frac{b_A(f)}{b_V(f)})^{-1}$ (black) and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ (blue) where $(\cdot)^{-1}$ is for visibility, and (b) the enlarged plot in the range of (500, 2000), which is marked as a yellow square. Red dotted line denotes the frequencies that correspond to musical note (C#5 to B6). (Color figure online)

- $(\frac{b_A(f)}{b_V(f)})^{-1}$ and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ both show similar trends but only the scales are different, and we expect it means that the spectral characteristics of accompaniment are similar between in vocal and non-vocal frames.
- Singing voice is extremely weaker than accompaniment in very low frequency range (lower than 100 Hz). It is reasonable because singing voice is mostly distributed in f_0 and its harmonics, which is rarely occur in those range, while some instruments such as bass and drums can be. Some previous studies for SVS have applied this characteristics by using high-pass filtering [5, 14].
- Some peaks can be found from the envelope, that are located around 0.7, 1.5, 3, and 8 kHz. we expect it is related with the formants of singing voice.
- From Fig. 4(b), we found an interesting phenomena that the singing voice is relatively weak in the frequency bins which correspond to the musical notes compared to those neighbor frequency bins. Although it needs more experiments to clarify the reason, we made some possible hypotheses as follows: (1) the mainlobe of singing voice may wider than that of accompaniment, (2) singing voice has stronger vibrato in general, and it may cause the ‘blurred peak’ in a long window length, or (3) singers frequently fail to sound exact note frequency, and make more errors than the instrumental players.

6 Conclusion

A novel framework for RPCA-based SVS was presented. In particular, we replaced the l_1 -norm term to the weighted l_1 -norm, and proposed to use the frequency-dependent variance ratio between singing voice and accompaniment to make the weighting matrix. In addition, we apply VAD for SVS by conducting a two-stage separation framework. In future works, we will investigate a method for finding a better weighting matrix Λ . The spatial information that is discarded in the current study also will be tried to be applied in the separation procedure.

Acknowledgments. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8501-16-1016) supervised by the IITP (Institute for Information & communications Technology Promotion)

References

1. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14**(5–6), 877–905 (2008)
2. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *J. ACM* **58**(3), 11 (2011)
3. Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Su, L., Yang, Y.H., Jang, R.: Vocal activity informed singing voice separation with the iKala dataset. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 718–722 (2015)

4. Huang, P.S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M.: Singing-voice separation from monaural recordings using robust principal component analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 57–60 (2012)
5. Jeong, I.Y., Lee, K.: Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints. *IEEE Signal Process. Lett.* **21**(10), 1197–1200 (2014)
6. Jeong, I.Y., Lee, K.: Vocal separation using extended robust principal component analysis with Schatten p/l_p -norm and scale compression. In: IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6 (2014)
7. Khajehnejad, M.A., Xu, W., Avestimehr, A.S., Hassibi, B.: Analyzing weighted minimization for sparse recovery with nonuniform sparse models. *IEEE Trans. Sig. Process.* **59**(5), 1985–2001 (2011)
8. Lehner, B., Sonnleitner, R., Widmer, G.: Towards light-weight, real-time-capable singing voice detection. In: International Society for Music Information Retrieval Conference, pp. 53–58 (2013)
9. Lehner, B., Widmer, G., Sonnleitner, R.: On the reduction of false positives in singing voice detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7480–7484 (2014)
10. Lehner, B., Widmer, G.: Monaural blind source separation in the context of vocal detection. In: International Society for Music Information Retrieval Conference, pp. 309–316 (2015)
11. Li, Y., Wang, D.: Separation of singing voice from music accompaniment for monaural recordings. *IEEE Trans. Audio Speech Lang. Process.* **15**(4), 1475–1487 (2007)
12. Rafii, Z., Pardo, B.: Repeating pattern extraction technique (REPET): a simple method for music/voice separation. *IEEE Trans. Audio Speech Lang. Process.* **21**(1), 73–84 (2013)
13. Sprechmann, P., Bronstein, A.M., Sapiro, G.: Real-time online singing voice separation from monaural recordings using robust low-rank modeling. In: International Society for Music Information Retrieval Conference, pp. 67–72 (2012)
14. Tachibana, H., Ono, N., Sagayama, S.: Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(1), 228–237 (2014)
15. Vincent, E., Gribonval, R., Fvotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
16. Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., Duong, N.Q.K.: The signal separation evaluation campaign (2007–2010): achievements and remaining challenges. *Sig. Process.* **92**(8), 1928–1936 (2012)