Combining F0 and non-negative constraint robust principal component analysis for singing voice separation

Feng Li[a,b,*], Masato Akagi[b]

[a] *Department of Computer Science and Technology, Anhui University of Finance and Economics, Bengbu 233030, China*
[b] *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

Abstract

Separating singing voice from a musical mixture remains an important task in the field of music information retrieval. Recent studies on singing voice separation have shown that robust principal component analysis (RPCA) with rank-1 constraint approach can improve separation quality. However, the performance of separation is limited because the vocal part can not be described well by the separated matrix. Therefore, prior information such as fundamental frequency (F0) should be considered. F0 can significantly improve separation performance by removing the spectral components of non-repeating instruments (e.g., bass and guitar). In this paper, we propose a novel singing voice separation algorithm by combining prior information and non-negative constraint RPCA, which incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm. In addition, we use the original phase recovery in estimating the spectral components of the separated singing voice. Experimental results on the iKala and MIR-1K datasets show higher efficiency of the proposed algorithm compared with state-of-the-art methods in terms of separation accuracy.

Keywords: Singing voice separation; Robust principal component analysis;

*Corresponding author
Email addresses: lifeng@jaist.ac.jp (Feng Li), akagi@jaist.ac.jp (Masato Akagi)

Non-negative rank-1 constraint; F0

1. Introduction

Singing voice separation is a process of separating the singing voice from a musical mixture and is widely used in chord recognition [1], singer identification [2], music auto-tagging [3], singing lyric recognizer [4], and music information retrieval [5] [6]. However, state-of-the-art methods of singing voice separation are still far behind human hearing capability, so this task remains extremely challenging [7] [8] due to the musical instruments involved and the time-varying spectral overlap between the singing voice and accompaniment.

Until recently, deep learning-based methods [9] [10] [11] [12] [13] [14] [15] [16] are perhaps the most widely used supervised learning methods for singing voice separation. In particular, Convolutional Neural Networks (CNN) seem to be especially adapted for this separation task. Lin et al. [11] proposed CNN-based model with ideal binary masking and cross entropy for singing voice separation. He [12] also proposed a sound level invariant singing voice separation by CNN with two types of data augmentation, frame normalization and zero-mean convolution. U-Net architecture described in [15] is modified the convolutional layers on the down-sampling and up-sampling sides. Although they have proven to be effective for separating the singing voice, a large amount of training data is needed in advance, which makes these models difficult to apply in the case of small audio data. Additionally, when there is a mismatch between the training and the testing samples [17], separation quality decreases due to overfitting. For this reason, unsupervised methods are often preferable for singing voice separation, particularly when only a limited amount of audio data is available or when there is no additional prior information [18]. Many unsupervised methods are inspired by, or loosely based on, non-negative matrix factorization (NMF) [19] [20] [21] [22], which is a type of dimensionality reduction that decomposes a non-negative matrix into a non-negative basis matrix and a non-negative activation matrix by using an iterative cost-minimization algorithm with multiplicative

update rules. Although NMF has shown impressive results in singing voice separation, it is difficult to determine the appropriate number of non-negative basis vectors. An algorithm based on robust principal component analysis (RPCA) [23] is effective for singing voice separation because singing voice can be well modeled as a sparse matrix, while the accompaniment is modeled as a low-rank matrix.

Inspired by the sparse and low-rank model for singing voice separation, Yang [24] proposed the multiple low-rank representations to decompose a magnitude spectrogram into two low-rank matrices. In a similar vein, a new RPCA-based method that incorporates harmonicity priors and a back-end drum removal procedure was proposed by [25]. Sprechmann et al. [26] proposed real-time online singing voice separation by robust low-rank modeling. Yu et al. [27] proposed sparse and low-rank representation with pre-learned dictionaries under the alternating direction method of multipliers framework. Rafii et al. [28] proposed a repeated accompaniment concept for background music and used the repeating pattern extraction technique for separating the repeating music part from the non-repeating singing voice in a mixed signal. Jeong et al. [29] proposed an extension of RPCA by generalizing the nuclear norm and the l_1-norm to Schatten-p norm and l_p-norm, respectively.

As mentioned above, RPCA is an effective strategy to separate singing voice from a musical mixture. It decomposes a given amplitude spectrogram of a mixture signal into a sum of a low-rank matrix and a sparse matrix. Because musical instruments reproduce almost the same sounds every time, a given note is played in a given song, the magnitude spectrogram of these sounds can be considered as a low-rank structure. Singing voice, in contrast, varies significantly but has a sparse distribution in the spectrogram domain to its harmonic structure. Although RPCA has been successfully applied to singing voice separation, it fails when there are significant differences in dynamic range among the different background instruments. Some instruments such as drums correspond to singular values with tremendous dynamic range. Because RPCA uses a nuclear norm to estimate the rank of the low-rank matrix, it over-estimates the

rank of the matrix that includes drum sounds. The accuracy of such separation thus decreases, as drums may be placed in a sparse subspace instead of being low-rank.

To solve these problems, Mikami et al. [30] proposed a residual drum sound estimation method for singing voice separation. Jeong et al. [31] proposed an extension of RPCA with weighted l_1-norm minimization for singing voice separation but only studied the different weighted values on a sparse matrix without including the low-rank matrix. In another work, Li et al. [32] proposed an extension of the RPCA algorithm called weighted robust principal component analysis (WRPCA), which utilizes different weighted values to describe the low-rank matrix for singing voice separation. However, it suffers from high computational cost due to computing the singular value decomposition at each iteration. Therefore, Li et al. [33] proposed an extension of RPCA with rank-1 constraint (CRPCA) that can improve both the separation performance and running time. But the quality of singing voice separation is limited because the vocal part can not be described well by the separated matrix. Separation algorithm with additional prior information such as fundamental frequency (F0) can enhance the effectiveness of separation results [34]. Because F0 varies over time and is a property of the parts played by various singing voice and accompaniment, it can greatly improve separation quality by removing the spectral components of non-repeating instruments (e.g., bass and guitar). Li et al. [35] proposed a method on singing voice separation by predominant F0 estimation with singing voice detection. Hsu et al. [36] proposed a tandem algorithm that estimated F0 information and separated the singing voice jointly and iteratively. Virtanen et al. [37] proposed a separation algorithm from polyphonic music accompaniment by combining F0 estimation and non-negative spectrogram factorization. Chan et al. [38] proposed an informed group-sparse representation for singing voice separation with the idea of informed separation incorporating F0 estimation.

Motivated by the above considerations, in this paper, we propose a novel singing voice separation algorithm by combining prior information and non-

negative rank-1 constraint RPCA (NCRPCA) called informed non-negative rank-1 constraint RPCA (NCRPCAi), which incorporates human-labeled F0 and non-negative rank-1 constraint minimization of singular values in RPCA for separating the singing voice. Furthermore, to minimize the reconstruction error when synthesizing the singing voice, we use the original phase recovery in estimating the spectral components of the separated singing voice.

In summary, we briefly summarize the main contributions of this paper as follows.

- We propose a novel singing voice separation NCRPCAi algorithm, which incorporates human-labeled F0 and non-negative rank-1 constraint minimization of singular values in RPCA to separate singing voice from a musical mixture.
- In addition, to minimize the reconstruction error when synthesizing the singing voice, we use the original phase recovery in estimating the spectral components of the separated singing voice.
- The proposed method yields state-of-the-art separation results on the iKala and MIR-1K datasets.

The remainder of this paper is organized as follows. Section 2 elaborates on our proposed NCRPCAi algorithm. Section 3 and Section 4 describe the framework of the reconstructed voice spectrogram and phase recovery. Experimental results are presented in Section 5. Finally, we conclude this paper in Section 6.

2. Informed NCRPCA

Informed NCRPCA is an extension of RPCA, which incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA. The NCRPCAi model can be defined as

$$\begin{aligned} &\text{minimize} \quad \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda|S|_1 + \frac{\gamma}{2}|S - E_0|, \\ &\text{subject to } X = L + S, L \geq 0, S \geq 0. \end{aligned} \tag{1}$$

where E_0 denotes the reconstructed voice spectrogram from F0. In section 3, we describe the value of E_0 in detail. The L is a low-rank matrix, $X \in \mathbb{R}_{m\times n}$ is an input matrix, and $\lambda > 0$ is a trade-off constant parameter between the sparse matrix S and the low-rank matrix L. The $\delta_i(L)$ is the i-th singular value of L. $\gamma > 0$ is a parameter. The same value $\lambda = \gamma = 1/\sqrt{\max(m,n)}$ as suggested by [38] [39]. We adopt an inexact augmented Lagrange multiplier (iALM) [40] to solve this convex model. The corresponding augmented Lagrange function is defined as

$$\begin{aligned} J(X,L,S,\mu) = \ & \min \sum_{i=2}^{min(m,n)} \delta_i(L) + \lambda|S|_1 \\ + < J, X-L-S > + & \frac{\mu}{2}|X-L-S|_F^2 + \frac{\gamma}{2}|S-E_0|, \end{aligned} \tag{2}$$

where J is the Lagrange multiplier, μ is a positive value, and $< J, X-L-S >$ denotes $J_{k+1} = J_k + \mu_k(X - L_{k+1} - S_{k+1})$.

From the above Lagrangian function, we can obtain the non-negative values of L and S,

$$\begin{aligned} L_{k+1} = & \min_L \sum_{i=2}^{\min(m,n)} \delta_i(L) + \langle J_k, X-L-S_k\rangle \\ & + \frac{\mu_k}{2}|X-L-S_k|_F^2 + \frac{\gamma}{2}|S_k - E_0|, \end{aligned} \tag{3}$$

$$\begin{aligned} S_{k+1} = & \min_S \ \lambda|S|_1 + \langle J_k, X-L_k-S\rangle \\ & + \frac{\mu_k}{2}|X-L_k-S|_F^2 + \frac{\gamma}{2}|S-E_0|, \end{aligned} \tag{4}$$

2.1. Update rules based on rank-1 constraint

As suggested by Oh et al. [41], the update rules of L and S are obtained as

$$L_{k+1} = P_{1,\mu_k^{-1}}(X - S_k + \mu_k^{-1}J_k), \tag{5}$$

$$S_{k+1} = Q_{\lambda\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1}J_k + \gamma E0), \tag{6}$$

and $P_{1,\mu_k^{-1}}(\cdot)$ can be defined as

$$P_{1,\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V_Y^T, \tag{7}$$

Algorithm 1 NCRPCAi for singing voice separation.

Input: Mixture signal X ($X \in \mathbb{R}_{m\times n}$), F0

1: **Initialize**: $\rho > 1, \mu_0 > 0, \lambda = \gamma > 0, k = 0, J_0 = L_0 = S_0 = 0$.

2: While not converge, **do** :

3: $L_{k+1} = P_{1,\mu_k^{-1}}(X - S_k + \mu_k^{-1} J_k)$.

4: $L_{k+1} = max\ (L_{k+1},\ 0)$.

5: $S_{k+1} = Q_{\lambda\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1} J_k + \gamma E0)$.

6: $S_{k+1} = max\ (S_{k+1},\ 0)$.

7: $J_{k+1} = J_k + \mu_k(X - L_{k+1} - S_{k+1})$.

8: $\mu_{k+1} = \rho * \mu_k$.

9: $k = k + 1$.

10: **end while**.

Output: $L_{m\times n} \geq 0,\ S_{m\times n} \geq 0$.

where the soft-thresholding operator [42] can be defined as

$$Q_{\mu_k^{-1}}(D_{Y_2}) = sign(D_{Y_2}) \cdot max(|D_{Y_2}| - \mu_k^{-1}, 0), \tag{8}$$

where $Y = Y_1 + Y_2$ ($Y \in \mathbb{R}_{m\times n}$), $D_{Y_1} = diag(\delta_1, 0, ..., 0)$, $D_{Y_2} = diag(0, \delta_2, ..., \delta_{min(m,n)})$, and δ_1 and δ_2 are the first and second singular values.

The specific process for separating singing voice from a mixed music signal is outlined in Algorithm 1. The input value of X is a musical mixture signal and F0 is the human-labeled from the observed audio data. E0 can be obtained from the values of F0. After the separation using the NCRPCAi algorithm, we can obtain a low-rank matrix L (accompaniment) and a sparse matrix S (singing voice).

3. Reconstructed voice spectrogram

To obtain the aforementioned reconstructed voice spectrogram E_0 from F0, we define harmonic masking M_h by the human-labeled F0 as the following equation:

$$M_h(t,f) = \begin{cases} 1 & nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & others, \end{cases} \tag{9}$$

where F_t is F0 estimated at frame t, n is the index of a harmonic part, and w is a frequency width for extracting the energy around each harmonic part, which we set to $w = 80\ Hz$ as suggested by [34]. Therefore, we define the reconstructed vocal spectrogram from the vocal annotations as

$$E_0 = X \odot M_h(t,f), \tag{10}$$

where $\odot$ denotes the element-wise multiplication operator (Hadamard product).

4. Phase recovery

We calculate the magnitude spectrogram (X) by short-time Fourier transform (STFT) in a musical mixture. Additionally, we estimate the magnitude and the phase of each source to resynthesize the singing voice in the time domain. The original phase P [43] can be defined as

$$P = angle(X); \tag{11}$$

Therefore, the recovered spectrogram $\widetilde{X}$ with the original phase in the complex coordinate can be obtained as

$$\widetilde{X} = S \odot cos(P) + i(S \odot sin(P)), \tag{12}$$

where S is the value of the sparse matrix separated by NCRPCAi algorithm.

Figure 1 shows an example of the waveform and spectrogram comparison of the clean and separated results using the proposed NCRPCAi and NCRPCA

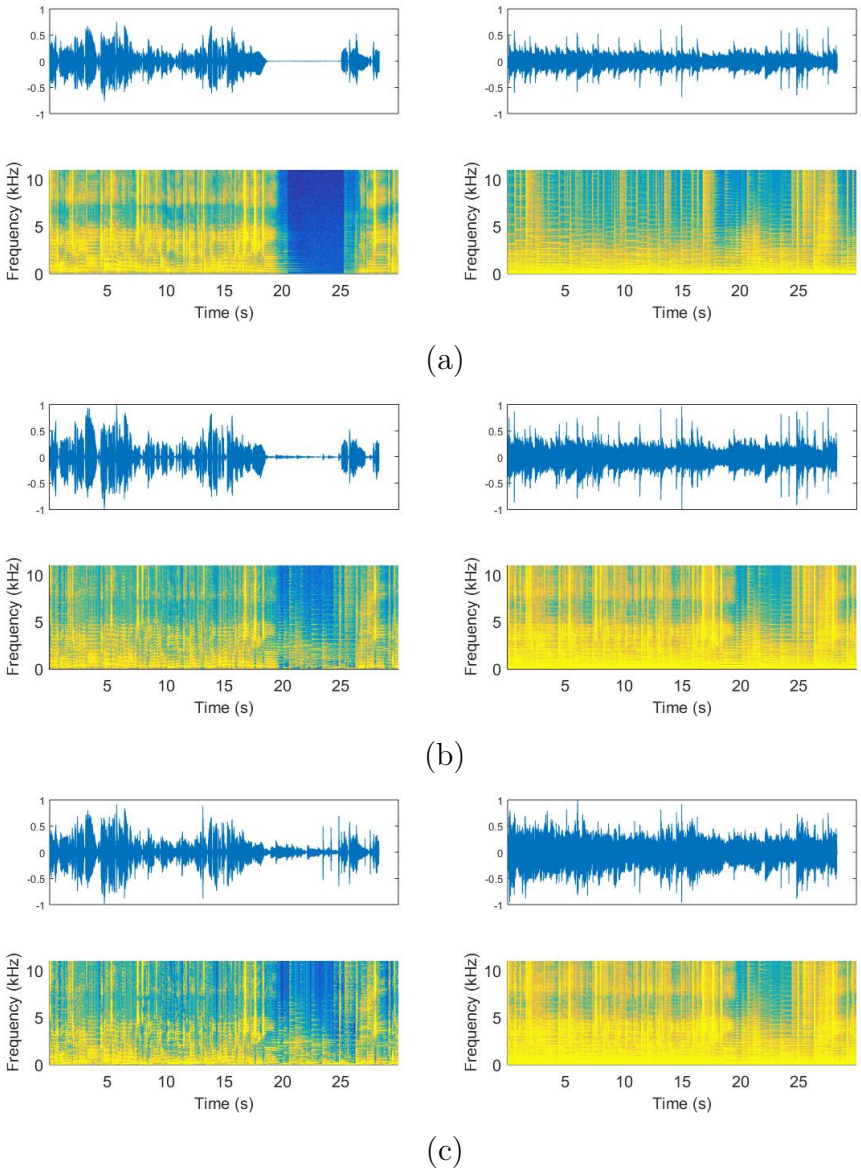

Figure 1: Example of waveform and spectrogram comparison of the clean and separated audio using NCRPCAi and NCRPCA methods on the iKala dataset (*71716_chorus*). The left parts are for singing voice and the right parts are for accompaniment. (a) is clean audio **(Top)**, (b) and (c) are the audio separated by NCRPCAi **(Middle: SDR is 12.30 dB)** and NCRPCA **(Bottom: SDR is 6.82 dB)**, respectively.

algorithms on the iKala dataset (*71716_chorus*). The left parts are for singing voice and the right parts are for the accompaniment. (a) is clean audio, (b) and (c) are the singing voice and accompaniment separated by NCRPCAi and NCRPCA, respectively. As shown in the figure, (b) contains the least amount of interference from the background music (accompaniment), in other words, NCRPCAi performs much better than NCRPCA.

5. Experimental evaluation

This section will focus on evaluating the proposed method and comparing it with the previous ones at the different evaluation metrics.

5.1. Experiment settings

To confirm the effectiveness of the proposed algorithm, our evaluation is carried out on two datasets on singing voice separation. One is iKala dataset [44][1]. This dataset contains 252 clips, each 30 sec long. Each song in the database is recorded in a wave file, sampled with 44.1 kHz, and has two channels. One channel is a ground truth singing voice, and the other is a ground truth music accompaniment. To reduce memory usage, we downsampled all the audio from 44.1 kHz to 22.05 kHz and computed its STFT by sliding a hamming window of 1411 samples with a 75% overlap to obtain the spectrogram. The mixture was of the singing voice and accompaniment at 0 dB signal-to-noise ratio ($SNR = 0$).

The other experiment dataset is MIR-1K dataset[45][2]. It contains 1000 Chinese pop songs recorded at 16 kHz sampling rate. The duration of each song slip ranges from 4 to 13 seconds. The right channel is singing voice and the left channel is background music. The mixture was of the singing voice and accompaniment at 0 dB signal-to-noise ratio.

[1] http://mac.citi.sinica.edu.tw/ikala/

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

5.2. Evaluation metrics

To evaluate the performance of the proposed method, we assessed its separation performance in terms of source-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifact ratio (SAR), and normalized SDR (NSDR) by using the BSS-EVAL evaluation toolbox 3.0 [46] [47][3]. The principle of the separation performance measures is to be decompose estimate signal $\hat{S}$(t) of a source $S_i(t)$ is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t), \tag{13}$$

where $S_{target}(t)$ is the allowable deformation of the target sound $S_i(t)$, $S_{interf}(t)$ 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 such as musical noise or the deformations included by the separation method that are not allowed. The formulas for SDR, SIR, SAR, and NSDR are respectively defined as

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \left(S_{interf}(t) + S_{artif}(t)\right)^2}, \tag{14}$$

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \tag{15}$$

$$SAR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \tag{16}$$

and

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \tag{17}$$

where $\hat{v}$ is the separated voice part, v is the clean singing voice signal, and x is the clean mixture value.

Higher values of SDR, SIR, SAR, and NSDR mean that the method exhibits better separation performance in singing voice separation task. More specifically, the value of SDR indicates the overall quality of the separated target sound signals, the value of SIR reflects the suppression of the interfering

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

source, and the value of SAR represents the absence of artificial distortion. In this work, we report the global values of SDR, SIR, SAR, and NSDR, respectively. In other words, the separation results are described with GSDR, GSIR, GSAR, and GNSDR, respectively. In a similar vein, higher values of GSDR, GSIR, GSAR, and GNSDR represent better quality of separation, especially the value of GNSDR, which is the most important metric in the aspect of overall performance evaluation. All the metrics are expressed in decibels.

5.3. Experimental results

In this section, we evaluate the proposed algorithm on the iKala and MIR-1K datasets, and compare it with unsupervised and supervised methods. To compare with the informed RPCA, the model can be defined as

$$\begin{aligned} &\text{minimize } |L|_* + \lambda|S|_1 + \frac{\gamma}{2}|S - E_0|, \\ &\text{subject to } X = L + S. \end{aligned} \tag{18}$$

where E_0 denotes the reconstructed voice spectrogram from F0.

5.3.1. Comparison with RPCA method

Table 1 shows the experimental results of the proposed algorithm and RPCA method on the iKala dataset. The results in this table confirm that NCRPCA shows better separation performance than RPCA. Meanwhile, with the corresponding approach of using F0, NCRPCAi also shows much better results than RPCAi in all evaluation metrics on the iKala dataset (252).

- RPCA: [23]
- RPCAi: Informed RPCA [48]
- NCRPCA: Non-negative constraint RPCA (Proposed 1)
- NCRPCAi: Informed NCRPCAi (Proposed 2)

Table 1: Singing voice separation results on the iKala dataset in dB **(252)**.

Method	GSDR	GSIR	GSAR	GNSDR
RPCA	6.41	8.37	12.65	2.46
RPCAi	11.91	18.09	13.46	7.96
NCRPCA	6.75	9.73	11.19	2.80
NCRPCAi	**12.03**	**18.31**	**13.54**	**8.08**

5.3.2. Comparison with state-of-the-art methods

In order to compare our proposed method with state-of-the-art supervised methods, we used 208 clips for testing in the experiment and another 44 clips for obtaining codebooks in the training process. The supervised methods mainly utilize online dictionary learning [49]. We used the SPAMS toolbox[4] to learn codebooks on 44 clips with the dictionary size of 100 atoms and the remaining 208 clips for testing.

- LRR: Low-Rank Representation [50]
- LRRi: Informed LRR [38]
- GSR: Group-Sparse Representation [38]
- GSRi: Informed GSR [38]

Table 2 shows the experimental results of the proposed NCRPCAi and state-of-the-art methods on the iKala dataset (208). These results were obtained with the supervised (LRR, LRRi, GSR, and GSRi) and unsupervised (RPCA, RPCAi, NCRPCA, and NCRPCAi) methods, respectively.

[4]http://spams-devel.gforge.inria.fr/

Table 2: Singing voice separation results on the iKala dataset in dB **(208)**.

Method	GSDR	GSIR	GSAR	GNSDR
LRR	7.73	11.41	11.17	3.93
LRRi	11.55	16.92	13.38	7.75
GSR	6.30	7.63	**14.80**	2.50
GSRi	11.51	16.34	13.63	7.71
RPCA	6.21	8.14	12.53	2.41
RPCAi	11.74	17.82	13.31	7.93
NCRPCA	6.55	9.49	11.05	2.74
NCRPCAi	**11.85**	**18.04**	13.39	**8.05**

The results in this table indicate that all methods performed better when using F0 than without it. The proposed NCRPCAi showed better results than supervised methods which use online dictionary learning (LRR, LRRi, GSR, and GSRi). As for the most important separation performance metric, the GNSDR, the proposed NCRPCAi method shows the best results among all methods with the value of 8.05 dB.

5.3.3. Comparison with RPCAi method

We also comparison with the proposed method with RPCAi on the MIR-1K dataset. Table 3 shows the experiment results of singing voice and accompaniment, respectively. From the experimental results obtained with GSDR, GSIR, GSAR, and GNSDR in the above table, again, it clearly shows that the proposed NCRPCAi delivered the best separation results between the separated singing voice and accompaniment parts.

Table 3: Singing voice separation results on the MIR-1K dataset in dB.

Method	GSDR	GSIR	GSAR	GNSDR
Singing Vocie				
RPCAi	5.40	8.89	8.91	5.25
NCRPCAi	**6.97**	**10.42**	**10.41**	**6.83**
Accompaniment				
RPCAi	4.79	10.21	7.00	4.68
NCRPCAi	**6.66**	**12.81**	**8.41**	**6.54**

5.4. Discussion

The experiment results are described in Tables 1, 2 and 3. From the values on the three tables, we can clearly see that NCRPCA with F0 achieves better separation results than without it on the iKala and MIR-1K datasets. Therefore, the prior information F0 is useful for the separation results. As demonstrated in [34] [36] [37], there are dependencies between singing voice separation and F0 estimation, which allows for improving the performance on one task by integrating information obtained via a method designed for the other. On the contrary, the inaccuracy of F0 estimation will lead to the worse values of recovered singing voice (e.g., E0), which brings the separated sparse matrix is not exact from the mixture matrix in Algorithm 1. It means that the results of separation performance are not better. The more accuracy of F0 estimation from singing voice is expected to be further research.

6. Conclusion

In this paper, we proposed a singing voice separation method by combining prior information and non-negative constraint RPCA that incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA. In addition, we used the original phase recovery in estimating the spectral components of the separated singing voice. Experimental results on the iKala and MIR-1K datasets demonstrate that the proposed NCRPCAi algorithm outperforms the conventional RPCA and state-of-the-art methods. For future work, since the dependencies between singing voice separation and F0 estimation, the accuracy of F0 estimation will bring the different separation results. This is a topic worth further studying. Future research will pursue this possibility.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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. The authors would like to thank the anonymous reviewers for their invaluable comments in improving the quality of this paper.

References

[1] T. Fujishima, Realtime chord recognition of musical sound: a system using common lisp music, in: Processing of International Computer Music Association (ICMC), 1999, pp. 464-467.

[2] M. N. Chinthaka, C.S. Xu, Y. Wang, Singer identification based on vocal and instrumental models, in: Proceedings of the 17th International Conference on Pattern Recognition (ICPR), 2004, pp. 375-378.

[3] J. Lee, J. Nam, Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging, IEEE signal processing letters, vol. 24, no. 8, 2017, pp. 1208-1212.

[4] C. K. Wang, R. Y. Lyu, Y. C. Chiang, An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker, in: Eighth European Conference on Speech Communication and Technology, 2003, pp. 1197-1200.

[5] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, Content-based music information retrieval: current directions and future challenges, Proceedings of the IEEE, vol. 96, no. 4, 2008, pp. 668-696.

[6] M. Goto, R. B. Dannenberg, Music interfaces based on automatic music signal analysis: new ways to create and listen to music, IEEE Signal Processing Magazine, vol. 36, no. 1, 2019, pp. 74-81.

[7] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, J. Fontecave, In the 2016 signal separation evaluation campaign, in: Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Springer, Cham, 2017, pp. 323-332.

[8] E.J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Krupse, L. Yang, An introduction to signal processing for singing-voice analysis: high notes in the effort to automate the understanding of vocals in music, IEEE Signal Processing Magazine, vol. 36, no. 1, 2019, pp. 82-94.

[9] J.R. Hershey, Z. Chen, J.L. Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in Proceedings

of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 31-35.

[10] W. Yuan, B. He, S. Wamg, J. Wang, M. Unoki, Enhanced feature network for monaural singing voice separation, Speech Communication, 2019, pp. 1-6.

[11] K. W. E. Lin, B. T. Balamurali, E. Koh, S. Liu, D. Herremans, Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy, Neural Computing and Applications, 2018, pp. 1-14.

[12] K. W. E. Lin, M. Goto, Zero-mean convolutional network with data augmentation for sound level invariant singing voice separation, in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 251-255.

[13] A. J. R. Simpson, G. Roma, M. D. Plumbley, Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network, in: Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Springer, Cham, 2015, pp. 429-436.

[14] Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 4, 2018, pp. 787-796.

[15] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, Singing voice separation with deep U-Net convolutional networks, in: Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 745-751.

[16] Z. Rafii, A. Liutkus, F.R. Stöter, S.I. Mimilakis, D. FitzGerald, B. Pardo, An overview of lead and accompaniment separation in music, IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26, no. 8, 2018, pp. 1307-1335.

[17] D. L. Wang, J.T. Chen, Supervised speech separation based on deep learning: an overview, IEEE/ACM Transactions on Audio, Speech and Language Processing, 2018, pp. 1702-1726.

[18] N. Tengtrairat, W.L. Woo, Single-channel separation using underdetermined blind autoregressive model and least absolute deviation, vol. 147, Neurocomputing, 2015, pp. 412-425.

[19] T. O. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no .3, 2007, pp. 1066-1074.

[20] M.N. Schmidt, M. Mϕrup, Nonnegative matrix factor 2-D deconvolution for blind single channel source separation, in: Proceedings of Independent Component Analysis and Blind Signal Separation (ICA), 2006, pp. 700-707.

[21] A. Chanrungutai, C. A. Ratanamahatana, Singing voice separation for mono-channel music using non-negative matrix factorization, in: Proceedings of International Conference on Advanced Technologies for Communications, 2008, pp. 243-246.

[22] S. Mirzaei, H. V. hamme, Y. Norouzi, Blind audio source counting and separation of anechoic mixtures using the multichannel complex NMF framework, Signal Processing, vol. 115, 2015, pp. 27-37.

[23] P. S. Huang, S. D. Chen, P. Smaragdis, M. H. Johnson, Singing-voice separation from monaural recordings using robust principal component analysis, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 57-60.

[24] Y. H Yang, Low-rank representation of both singing voice and music accompaniment via learned dictionaries, in: Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR), 2013, pp. 427-432.

[25] Y. H. Yang, On sparse and low-rank matrix decomposition for singing voice separation, in: Proceedings of the 20th ACM international conference on Multimedia (MM), 2012, pp. 757-760.

[26] P. Sprechmann, A. Bronstein, G. Sapiro, Real-time online singing voice separation from monaural recordings using robust low-rank modeling, in: Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR), 2012, pp. 67-72.

[27] S.W. Yu, H.J., Z.Y. Duan, Singing voice separation by low-rank and sparse spectrogram decomposition with prelearned dictionaries, Journal of the Audio Engineering Society, vol. 65, no. 5, 2017, pp. 377-388.

[28] Z. Rafii, B. Pardo, Repeating pattern extraction technique (REPET): a simple method for music/voice separation, IEEE transactions on audio, speech, and language processing, vol. 21, no. 1, 2013, pp. 73-84.

[29] I.Y. Jeong, K. Lee, Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints, IEEE Signal Processing Letters, vol. 21, no. 10, 2014, pp. 1197-1200.

[30] S. Mikami, A. Kawamura, Y. Iiguni, Residual drum sound estimation for RPCA singing voice extraction, in: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 442-446.

[31] I.Y. Jeong, K. Lee, Singing voice separation using RPCA with weighted l_1-norm, in: Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Springer, Cham, 2017, pp. 553-562.

[32] F. Li, M. Akagi, Weighted robust principal component analysis with gammatone auditory filterbank for singing voice separation, in: Proceedings of the International Conference on Neural Information Processing (ICONIP), 2017, pp. 849-858.

[33] F. Li, M. Akagi, Unsupervised singing voice separation based on robust principal component analysis exploiting rank-1 constraint, in: Proceedings of 26th European Signal Processing Conference (EUSIPCO), 2018, pp.1920-1924.

[34] Y. Ikemiya, K. Itoyama, K. Yoshii, Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 11, 2016, pp. 2084-2095.

[35] Y.P. Li, D.L. Wang, Separation of singing voice from music accompaniment for monaural recordings, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, 2007, pp. 1475-1487.

[36] C.L. Hsu, D.L. Wang, A tandem algorithm for singing pitch extraction and voice separation from music accompaniment, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, 2012, pp. 1482-1491.

[37] T. Virtanen, A. Mesaros, M. Ryynänen, Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music, in: Proceedings of INTERSPEECH, 2008, pp. 17-22.

[38] T.-S. T. Chan, Y.-H. Yang, Informed group-sparse representation for singing voice separation, IEEE Signal Processing Letters, vol. 24, no. 2, 2017, pp. 156-160.

[39] E. J. Candés, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM, vol. 58, no. 3, 2011.

[40] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv :1009.5055, 2010.

[41] T. Oh, Y. Tai, J. Bazin, H. Kim, I.S. Kweon, Partial sum minimization of singular values in robust PCA: Algorithm and applications, IEEE transac-

tions on pattern analysis and machine intelligence, vol. 38, no. 4, 2016, pp. 744-758.

[42] E. Hale, W. Yin, Y. Zhang, Fixed-point continuation for ℓ_1-minimization: Methodology and convergence, SIAM Journal on Optimization, vol. 19, no. 3, 2008, pp. 1107-1130.

[43] E. Vincent, N. Bertin, R. Gribonval, F. Bimbot, From blind to guided audio source separation: How models and side information can improve the separation of sound, IEEE Signal Processing Magazine, vol. 31, no. 3, 2014, pp. 107-115.

[44] T.S. Chan, T.C. Yeh, Z.C. Fan, H.W. Chen, L. Su, Y.H. Yang, R. Jang, Vocal activity informed singing voice separation with the iKala dataset, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 718-722.

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

[46] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE transactions on audio, speech, and language processing, vol. 14, no. 4, 2006, pp. 1462-1469.

[47] E. Vincent, S. Araki, F. Theis, R. Gribonval, G. Nolted, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunkerf, D. Lutter, N.Q.K. Duong, The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges, Signal Processing, vol. 92.8, 2012, pp. 1928-1936.

[48] Z. Chen, P.-S. Huang, Y.-H. Yang, Spoken lyrics informed singing voice separation, in: Proceedings of HAMR, 2013.

[49] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009.

[50] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, 2013, pp. 171-184.