<table>
<thead>
<tr>
<th>Title</th>
<th>Blind Monaural Singing Voice Separation Using Rank-1 Constraint Robust Principal Component Analysis and Vocal Activity Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Li, Feng; Akagi, Masato</td>
</tr>
<tr>
<td>Citation</td>
<td>Neurocomputing, 350: 44-52</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2019-04-17</td>
</tr>
<tr>
<td>Type</td>
<td>Journal Article</td>
</tr>
<tr>
<td>Text version</td>
<td>author</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10119/17247">http://hdl.handle.net/10119/17247</a></td>
</tr>
<tr>
<td>Rights</td>
<td>Copyright (C)2019, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (CC BY-NC-ND 4.0). [<a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a>] NOTICE: This is the author's version of a work accepted for publication by Elsevier. Feng Li and Masato Akagi, Neurocomputing, 350, 2019, 44-52, <a href="http://dx.doi.org/10.1016/j.neucom.2019.04.030">http://dx.doi.org/10.1016/j.neucom.2019.04.030</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The methodology and experimental setup were discussed, focusing on the robustness and performance of the proposed method. The effectiveness of the rank-1 constraint in improving separation accuracy was highlighted. The integration of robust principal component analysis with vocal activity detection was demonstrated to yield promising results in blind monaural singing voice separation.</td>
</tr>
</tbody>
</table>

日本先端科学技术研究推進機構

JAIST
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
Blind monaural singing voice separation using rank-1 constraint robust principal component analysis and vocal activity detection

Feng Li*, Masato Akagi
Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Abstract

In this paper, a novel blind separation method for monaural singing voice based on an extension of robust principal component analysis (RPCA) using a rank-1 constraint called Constraint RPCA (CRPCA) is proposed. Although the conventional RPCA is an effective method to separate singing voice from the mixed audio signal, it fails when one singular value (e.g., drum) is much larger than all others (e.g., other accompanying instruments). The proposed CRPCA method utilizes rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm, which not only provides a solution robust to large dynamic range differences among instruments but also reduces the computation complexity. Further quality improvement is achieved by converting CRPCA to an ideal binary masking, combining it with harmonic masking to create a coalescent masking, and finally, combining with a vocal activity detection. Evaluation results on ccMixter and DSD100 datasets show that the proposed method achieves better separation performance than the previous methods.

Keywords: Blind monaural singing voice separation; Robust principal component analysis; Rank-1 constraint; Coalescent masking; Vocal activity detection

*Corresponding author

Email addresses: lifeng@jaist.ac.jp (Feng Li), akagi@jaist.ac.jp (Masato Akagi)
1. Introduction

Monaural singing voice separation has received much attention in recent years for its range of potential applications including singer identification [1], melody extraction [2], music information retrieval (MIR) [3], chord recognition [4], speech enhancement [5], and computational auditory scene analysis (CASA) [6]. This type of separation is even more difficult than multichannel source separation since only one channel is used [7]. Blind monaural singing voice separation is a technique for extracting singing voice from a set of single channel mixed music signals without any additional prior information. The blind monaural singing voice separation system is shown in Fig. 1.

Mixture music consists of singing voice and background music including drums, bass, and other instruments. After separation by the proposed method, we obtain the target singing voice and accompaniment parts from the mixture music.

There have been many methods proposed to overcome the difficulty in separation tasks. However, state-of-the-art methods for singing voice separation are still far behind human hearing capability, especially for single-channel sources, and the task remains extremely challenging [8] due to the musical instruments involved and time-varying spectral overlap between singing voice and background music. Research in the field of monaural singing voice separation can be divided into two categories: supervised and unsupervised learning methods. Supervised learning methods mainly rely on prior knowledge about the mixed audio sources. Deep neural network (DNN)-based models [9] [10] [11] [12] [13] [14] are perhaps the most widely used supervised
learning models for singing voice separation. Although they have proven effective for separating singing voice, a large number of training data are needed in advance, which makes these models difficult to apply in case of small audio data. In addition, when there is a mismatch between training and testing samples [15], separation quality decreases due to overfitting. In light of this, unsupervised methods are often preferable for monaural singing voice separation, particularly when only a limited amount of audio data is available or when there is no additional prior information [16] [17]. Many unsupervised methods are inspired by, or loosely based on, non-negative matrix factorization (NMF) [18] [19] [20], 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 using an iterative cost-minimization algorithm with multiplicative update rules. Although NMF has shown impressive results in monaural audio source separation, it is difficult to determine the appropriate number of nonnegative basis vectors.

Robust principal component analysis (RPCA) [7] is an effective approach for singing voice separation because singing voice can be well modeled as a sparse matrix, while accompaniment as well modeled as a low-rank matrix. RPCA has been extensively and successively applied in other signal processing applications like speech enhancement [21] [22] [23], SAR imaging [24] [25], direction of arrivals tracking [26] and also in computer vision applications [27] [28] [29]. Inspired by this sparse and low-rank model, a new RPCA-based method that incorporates harmonicity priors and a backend drum removal procedure was proposed [30]. In a similar vein, Yang [31] proposed multiple low-rank representations (MLRR) to decompose a magnitude spectrogram into two low-rank matrices. Rafii et al. [32] proposed a repeated accompaniment concept for background music and used the Repeating Pattern Extraction Technique (REPET) for separating the repeating music part from the non-repeating singing voice in a mixture signal. Sprechmann et al. [33] proposed a real-time online singing voice separation by robust low-rank modeling. Fourer et al. [34] proposed a novel unsupervised singing voice detection method which uses single-channel Blind Audio Source Separation (BASS) algorithm as a preliminary step. Chan et al. [35] proposed using informed group-sparse representation with the idea of pitch annotations separation. Pu et al. [36] proposed an approach in audio separation with the assistance of visual
information.

As stated above, RPCA is an effective way 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 (accompaniment) and a sparse matrix (singing voice). Since musical instruments reproduce nearly 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 it uses nuclear norm to estimate the rank of the low-rank matrix, RPCA algorithms similar to those in [37] over-estimate the rank of a matrix that includes drum sounds. The accuracy of such separation results thus decreases, as drums may be placed in the sparse subspace instead of being low-rank.

To overcome these issues, Mikami et al. [38] proposed a residual drums sound estimation method for singing voice separation. Jeong et al. [39] 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 rather than including the low-rank matrix as well. In another approach, Li and Akagi [40] proposed an extension of the RPCA algorithm called weighted robust principal component analysis (WRPCA), which uses 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 (SVD) at each iteration. Hence, the running time of WRPCA is slower than RPCA. Recently, a partial sum minimization of singular values as an alternative to minimizing the nuclear norm in RPCA [41] was proposed, which uses minimized rank to determine the different values of SVD in image processing. In response to the above problems, in this paper, we extend the idea in [41] and propose an extension of RPCA exploiting the rank-1 constraint (CRPCA) [42], which utilizes the rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm for separating singing voice from the mixture music. There are other
works which used rank-1 constraint RPCA in computer vision application \cite{43} \cite{44} \cite{45} \cite{46}. To the best of our knowledge, this is the first work using different singular values for the singing voice separation task. CRPCA not only describes the different values of SVD but also reduces the computation complexity. This present study extends the preliminary work \cite{42} by melody extraction, which plays a vital role in separating singing voice \cite{47} \cite{48} \cite{49}, we convert the CRPCA output to an ideal binary masking, combine it with a harmonic masking to create a coalescent masking, and apply the coalescent masking to extract the singing voice. In addition, we adopt a vocal activity detection (VAD) algorithm to constrain the temporal segments in which singing voice may occur.

To sum up, in this paper, we propose a blind separation method based on rank-1 constraint RPCA for monaural singing voice. The major contributions of this paper are summarized as follows.

- We present an extension of RPCA called CRPCA, which constraints the low-rank matrix in RPCA to have rank greater than or equal to one, thereby describing the sensitively of RPCA to dynamic range variation.

- We construct coalescent masking, which consists of time-frequency masking fused with harmonic masking. In addition, we use VAD to constrain the temporal segments that are allowed to contain singing voice.

- We demonstrate through a detailed experiment on monaural singing voice separation that the proposed method can achieve a significant improvement of separation performance over the conventional RPCA and even exceeds the previously proposed WRPCA \cite{40}.

The remainder of this paper is structured as follows. In Section 2, we briefly review related work on singing voice separation focusing on RPCA-based methods. The proposed CRPCA method is described in Section 3. In Sections 4 and 5, we introduce the coalescent masking and VAD, respectively. Then, the results and analysis of the experiments on benchmark datasets are provided in Section 6. We conclude in Section 7 with a brief summary.
2. Related work

This section briefly reviews the conventional RPCA. Then, we discuss the previously proposed WRPCA and its application to singing voice separation.

2.1. Principle of RPCA

Candès et al. [37] presented a convex program RPCA, which decomposed an input matrix \( X \in \mathbb{R}^{m \times n} \) into the sum of a low-rank matrix \( L \in \mathbb{R}^{m \times n} \) and a sparse matrix \( S \in \mathbb{R}^{m \times n} \). This problem can be formulated as

\[
\begin{align*}
& \text{minimize} \ |L|_* + \lambda |S|_1, \\
& \text{subject to} \ X = L + S,
\end{align*}
\]

(1)

where \(|L|_*\) denotes the nuclear norm (sum of singular values), \(|S|_1\) denotes the \(L_1\)-norm (sum of absolute values of matrix entries), and \(\lambda > 0\) is a positive constant balancing the relative importance of model violations between the low-rank matrix \(L\) and sparse matrix \(S\). As Candès et al. [37] suggested, we set \(\lambda = 1/\sqrt{\max(m,n)}\) in this work. Furthermore, this convex program can be solved by accelerated proximal gradient (APG) or augmented Lagrange multipliers (ALM) [50]. There are two versions of ALM methods: inexact and exact. We use the efficient inexact ALM algorithm for solving the RPCA problem as a baseline method for comparison in our experiments [7].

2.2. Principle of WRPCA

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

\[
\begin{align*}
& \text{minimize} \ |L_{w,*} + \lambda S|_1, \\
& \text{subject to} \ X = L + S,
\end{align*}
\]

(2)

where \(w\) is a vector of weights and \(|L_{w,*}\) is the low-rank matrix computed using weighted singular value minimization, \(S\) is the sparse 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 \(L\). We used \(\lambda = 1/\sqrt{\max(m,n)}\) as suggested by Candès et al. [37]. We also
adopted an efficient inexact ALM [50] to solve this convex model. The corresponding augmented Lagrange function is defined as

\[
J(X, L, S, \mu) = |L_\ast|_* + \lambda|S|_1 + \langle J, X - L - S \rangle + \frac{\mu}{2}|X - L - S|_F^2,
\]

(3)

where \( J \) is the Lagrange multiplier and \( \mu \) is a positive scalar.

In RPCA, nuclear norm minimization and \( L_1 \)-norm affect not only the sparsity and low-rankness of the two decomposed matrices but also their relative scale values. In order to better balance their scale values, WRPCA uses different weighted value strategies to trim the low-rank matrix during each stage of the singing voice separation processing.

Set \( X = U\Sigma V^T, X \in \mathbb{R}^{m \times n} \), where

\[
\Sigma = \begin{pmatrix} \text{diag}(\delta_1(X), \delta_2(X), \ldots, \delta_n(X)) \\ 0 \end{pmatrix},
\]

(4)

and \( \delta_i(X) \) denotes the \( i \)-th singular value of \( X \). If the positive regularization parameter \( C \) exists and the positive value \( \varepsilon < \min(\sqrt{C}, \frac{C}{\delta_1(X)}) \), using Candès et al [51] proposed reconstruct sparse signals, the reweighing formula can be defined as

\[
w_i' = \frac{C}{\delta_i(L_\ast) + \varepsilon},
\]

(5)

so the weighted values will converge to

\[
L_\ast = U\Sigma' V^T,
\]

(6)

where

\[
\Sigma' = \begin{pmatrix} \text{diag}(\delta_1(L_\ast), \delta_2(L_\ast), \ldots, \delta_n(L_\ast)) \\ 0 \end{pmatrix},
\]

(7)

and

\[
\delta_i(L_\ast) = \begin{cases} 0 \\ \frac{c_1 + \sqrt{c_2}}{2} \end{cases}
\]

(8)
Algorithm 1 WRPCA for singing voice separation

**Input:** Mixture signal $X \in \mathbb{R}^{m \times n}$, weight vector $w$.

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

1. **Initialize:** $\rho, \mu_0, L_0 = X, J_0 = 0, k = 0$.
2. **While** not converge, **do**:
   3. $S_{k+1} = \arg \min |S|_1 + \frac{\mu_k}{2} |X + \mu_k^{-1} J_k - L_k - S|_F^2$.
   4. $L_{k+1} = \arg \min |L_{w, \ast, \ast} + \frac{\mu_k}{2} |X + \mu_k^{-1} J_k - S_{k+1} - L_k|_F^2$.
   5. $J_{k+1} = J_k + \mu_k (X - L_{k+1} - S_{k+1})$.
   6. $\mu_{k+1} = \rho \ast \mu_k$.
   7. $k = k + 1$.
8. **end while.**

where $c_1 = (\delta_i (X) - \varepsilon)$ and $c_2 = ((\delta_i (X) + \varepsilon)^2 - 4C)$ [52]. In this work, we empirically set the regularization parameter $C$ as the maximum matrix size, which enables us to obtain the best separation performance results on the audio dataset, e.g., $C = \max(m, n)$ [40].

The specific process for separating singing voice from the mixed music signal is outlined in **Algorithm 1**, where the value of $X$ is a mixed music signal from the observed audio datum. After separation by WRPCA, we obtain a low-rank matrix $L$ (accompaniment) and a sparse matrix $S$ (singing voice). Therefore, we can use the WRPCA method to decompose an input matrix into a low-rank matrix part and a sparse matrix part. The separation results outperform the RPCA method in different audio data. However, it suffers from high computational cost due to computing an SVD at each iteration, which in turns leads to slow running time.

3. **Constraint RPCA (CRPCA)**

CRPCA is an extension of RPCA in which the low-rank matrix is constrained to have rank greater than or equal to one. Because of this constraint, the first singular
value can be removed from the nuclear norm, thereby freeing the first basis vector to represent a component with very high singular value such as the average drumset or average background noise. The corresponding model can be defined as

$$\min_{\sum_i \delta_i(L) + \lambda |S|_1,}$$

subject to

$$X = L + S,$$

where $L$ is the low-rank matrix and $S$ is the sparse 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 $L$. $\delta_i(L)$ is the $i$-th singular value of $L$. We use the same value $\lambda = 1/\sqrt{\max(m,n)}$ as suggested by Candès et al. [37]. We also adopt an efficient inexact version of the ALM [50] to solve this convex model. The corresponding augmented Lagrange function is defined as

$$J(X, L, S, \mu) = \min_{\sum_i \delta_i(L) + \lambda |S|_1} + <J, X - L - S> + \frac{\mu}{2} |X - L - S|_F^2,$$  

where $J$ is the Lagrange multiplier and $\mu$ is a positive value.

From the above Lagrangian function, we can obtain the following two sub-problems related to $L$ and $S$:

$$L_{k+1} = \min_L \sum_i \delta_i(L) + <J_k, X - L - S_k> + \frac{\mu_k}{2} |X - L - S_k|_F^2,$$

$$S_{k+1} = \min_S \lambda |S|_1 + <J_k, X - L_k - S> + \frac{\mu_k}{2} |X - L_k - S|_F^2,$$

3.1. Update rules based on rank-1 constraint

As suggested by Oh et al. [41], the update rules of $L$ and $S$ are equivalent to solving the above two sub-problems, as

$$L_{k+1} = P_{1/\mu_k}^{-1}(X - S_k + \mu_k^{-1} J_k),$$
Algorithm 2 CRPCA for singing voice separation

Input: Mixture signal $X \in \mathbb{R}_{mix}$.  
1: Initialize: $\rho > 1, \mu_0 > 0, k = 0, L_0 = S_0 = 0$.  
2: While not converge,  
3: do :  
4: $L_{k+1} = P_{\mu_k^{-1}}(X - S_k + \mu_k^{-1}J_k)$.  
5: $S_{k+1} = Q_{\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1}J_k)$.  
6: $J_{k+1} = J_k + \mu_k(X - L_{k+1} - S_{k+1})$.  
7: $\mu_{k+1} = \rho \cdot \mu_k$.  
8: $k = k + 1$.  
9: end while.  
Output: $L_{mix}, S_{mix}$.  

\[ S_{k+1} = Q_{\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1}J_k), \]  
\[ \text{and } P_{\mu_k^{-1}}(\cdot) \text{ can be defined as } \]  
\[ P_{\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V^T_Y, \]  
\[ \text{where the soft-thresholding operator } [53] \text{ can be defined as } \]  
\[ Q_{\mu_k^{-1}}(D_{Y_2}) = \text{sign}(D_{Y_2}) \cdot \max(|D_{Y_2}| - \mu_k^{-1}, 0), \]  
\[ \text{where } Y = Y_1 + Y_2 \text{ (} Y \in \mathbb{R}_{mix} \text{), } D_{Y_1} = \text{diag}(\delta_1, 0, ..., 0), \text{ } D_{Y_2} = \text{diag}(0, \delta_2, ..., \delta_{\min(m,n)}), \]  
\[ \text{and } \delta_1 \text{ and } \delta_2 \text{ are the first and second singular values. } \]

The separation process corresponding to the mixed music signal is outlined in Algorithm 2. The input value $X$ is a mixed music signal from the observed audio data. Finally, after the algorithm convergences, we obtain a low-rank matrix $L$ (accompaniment) and a sparse matrix $S$ (singing voice).
4. Coalescent masking

4.1. Time frequency masking

We apply ideal binary time frequency masking (IBM) [7] to further improve the separation results from low-rank and sparse matrices by CRPCA. The function $M_{ibm}$ is defined as

$$M_{ibm}(i, j) = \begin{cases} 
1 & S_{ij} \geq L_{ij} \\
0 & S_{ij} < L_{ij} 
\end{cases} \quad (17)$$

where $S_{ij}$ and $L_{ij}$ are the values of the sparse and low-rank matrices.

4.2. Vocal F0 estimation

Vocal F0 estimation can significantly improve the separation performance of singing voice [49], so extracting the F0 contour properly is crucial. Subharmonic summation is an efficient technique for this calculation [48] [54]. In this work, we adopt the salience function $H(t, s)$, which is formulated as

$$H(t, s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2 (n)),$$  

where $t$ and $s$ indicate frame index and logarithmic frequency, respectively. $P(t, s)$ represents the power at frame $t$ and frequency $s$, $N$ is the number of harmonic parts, and $h_n$ is a decaying factor, $0.84^{n-1}$ in this paper. Log frequency $s$ is measured in cents (1200 cents per octave), and $P(t, s)$ is computed with a frequency resolution of 200 bins per octave (6 cents per bin).

The optimal melody contour $C$ can be solved by using an optimal path problem formulated as

$$C = \arg\max_T \sum_{t=1}^{T-1} (\log a_t H(t, s_t) + \log T(s_t, s_{t+1})),$$  

where $T(s_t, s_{t+1})$ is a transition probability that indicates the likelihood of the current F0 moving to the next F0 in the consecutive frame, and $a_t$ is a normalization factor that makes the salience values sum to one within the range of the F0 search. We use the Viterbi search algorithm [55] to optimize the melody contour $C$ value.
4.3. Harmonic masking

In accordance with the previous research, we define the harmonic masking $M_h$ by the above-mentioned obtained vocal F0 as

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

(20)

where $F_t$ is the vocal 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.

4.4. Coalescent masking

In this section, we propose a coalescent masking, which is combining harmonic masking $M_h$ with ideal binary time frequency masking $M_{ibm}$. The corresponding formulation $M_c$ can be described as

$$M_c = M_{ibm} \odot M_h$$

(21)

where $M_{ibm}$ and $M_h$ are the time frequency masking and harmonic masking, respectively, and $\odot$ denotes the element-wise multiplication operator.

Finally, the temporal segments in which singing voice can be obtained by using the coalescent masking, the following formulas can be defined as

$$S_{vocal} = M_c \odot X$$

(22)

where $\odot$ denotes the element-wise multiplication operator.

5. Vocal activity detection

To obtain better separation performance and optimize the value of coalescent masking, we apply a VAD algorithm to constrain the temporal segments in which singing voice. Singing voice only be detected in frames $t$ such that $\Omega(t) > k$, where $k$ is a threshold. The cost function $\Omega(t)$ can be defined as

$$\Omega(t) = \sum_f \left( \frac{1}{H_f} \sum_{n=1}^{H_f} P(t, s + 1200 \log_2(n)) \right)^{1.8}$$

(23)
where $H_f = (F_s/2f)$ is the number of harmonics of the frequency $f$ that exist at frequencies below the Nyquist rate $F_s/2$. $P(t, s)$ stands for the power at frame $t$ and log frequency $s$.

A block diagram of our proposed blind monaural singing voice separation system is given in Fig. 2. For each mixture music in the test dataset, we first apply a magnitude short-time Fourier transform (STFT) [56] to obtain $X$, then separate $X$ into the corresponding low-rank matrix $L$ and sparse matrix $S$ by using the CRPCA method. We then utilize coalescent masking to constrain the time-frequency masking to only those times and frequencies that constrain harmonics. VAD is adopted to improve the separation performance by discriminating the vocal and non-vocal frames. Finally, we use an inverse short-time Fourier transform (ISTFT) [57] to obtain the accompaniment and singing voice parts from the mixture music.

In this work, we randomly excerpted example 30-second audio data units from the ccMixter dataset. Figures 3 and 4 show the spectrograms of separated singing voice parts and separated accompaniment parts from the mixed music signal. Different separation methods are used to compare the original spectrograms, singing voice, and accompaniment, respectively. As shown in the figures, the spectrogram of Fig. 3(b) contains the greatest amount of interference from background music signal (accompaniment) in the recovered singing, while in Fig. 3(f) contains the least. In other words, the latter is better than the former in singing voice separation task. As for the comparison with accompaniment in Fig. 4, CRPCA using coalescent masking and VAD has the best value of separation performance among them.
Figure 3: Spectrograms are excerpted from AlexBeroza\_To\_Be\_Sensitive\_(with\_mindmaphat) in the ccMixer dataset: (a) spectrogram of original singing voice, (b) spectrogram of separated singing voice by RPCA, (c) spectrogram of separated singing voice by WRPCA, (d) spectrogram of separated singing voice by CRPCA (Proposed 1), (e) spectrogram of separated singing voice by CRPCA with IBM (Proposed 2), (f) spectrogram of separated singing voice by CRPCA using coalescent masking and VAD (Proposed 3), respectively.
Figure 4: Spectrograms are excerpted from AlexBeroza\textunderscore To\textunderscore Be\textunderscore Sensitive\textunderscore (with\textunderscore mindmapthat) in ccMixter dataset: (a) spectrogram of original accompaniment, (b) spectrogram of separated accompaniment by RPCA, (e) spectrogram of separated accompaniment by WRPCA, (d) spectrogram of separated accompaniment by CRPCA (Proposed 1), (e) spectrogram of separated accompaniment by CRPCA with IBM (Proposed 2), (f) spectrogram of separated accompaniment by CRPCA using coalescent masking and VAD (Proposed 3), respectively.
6. Experimental results and analysis

We performed experiments using two different datasets for the singing voice separation task: ccMixter \[58\]^1 and DSD100 \[8\]^2. Conventional RPCA \[7\] and WRPCA \[40\] are included for comparison.

- Proposed 1: CRPCA only
- Proposed 2: CRPCA with IBM
- Proposed 3: CRPCA using coalescent masking and VAD

6.1. Experiment datasets and conditions

The ccMixter dataset contains 50 full songs with durations ranging from 1’17” to 7’36”. Each audio datum contains three parts: singing voice, accompaniment, and a mixture of the two, respectively.

The Demixing Secrets Dataset (DSD100) contains 100 full stereo songs of different styles with durations ranging from 2’21” to 7’15”, as also used for the 2016 Signal Separation Evaluation Campaign (SiSEC) \[8\], which is split into 50 training (dev) and 50 test songs. Each datum consists of bass, drums, other, and singing voice. In our experiments, all data are conducted as the test data. We consider the sum of drums, bass, and other as the accompaniment part. The objective is to separate the singing voice from the accompaniment in a mixed music signal.

Our main focus in these experiments is the monaural source separation task. This task is typically even more difficult than multichannel source separation due to the availability of only one channel. Therefore, the two-channel stereo mixture datasets we used were downmixed into a single channel. We evaluated the whole audio datum rather than just partial lengths on both datasets. All experiment data were sampled at 44.1 kHz. STFT and ISTFT with a window size of 1024 samples and a hop size of 256 samples were used. All experiments were run using MATLAB R2015a on a PC win10, X64-based processor, RAM 32GB with i7-6700K CPU@4.00 GHz.

---

1https://members.loria.fr/ALiutkus/kam/
2http://liutkus.net/DSD100.zip
To evaluate the effectiveness of the proposed method, we assessed its separation performance in terms of source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and normalized SDR (NSDR) by using the BSS-EVAL 3.0 metrics \[59\].

The estimated signal \( \hat{S}(t) \) is defined as

\[
\hat{S}(t) = S_{\text{target}}(t) + S_{\text{interf}}(t) + S_{\text{artif}}(t),
\]

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

\[
SDR = 10 \log_{10} \frac{\sum t S_{\text{target}}(t)^2}{\sum t (S_{\text{interf}}(t) + S_{\text{artif}}(t))^2},
\]

\[
SIR = 10 \log_{10} \frac{\sum t S_{\text{target}}(t)^2}{\sum t S_{\text{interf}}(t)^2},
\]

\[
NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v),
\]

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

Higher values of SDR, SIR, and NSDR mean that the method exhibits better separation performance in terms of the singing voice separation tasks. More specifically, the value of SDR indicates the overall quality of the separated target sound signals, while the value of SIR reflects the suppression of the interfering source. All metrics are expressed in dB.

### 6.2. Results and discussions

For the ccMixter dataset, all comparisons of singing voice separation results with the conventional RPCA, WRPCA, and proposed methods (CRPCA only, CRPCA with

\[3\text{http://bass-db.gforge.inria.fr/bss_eval/}\]
IBM and CRPCA using coalescent masking and VAD) are shown in Fig. 5. From the experimental results obtained with the SDR, SIR, and NSDR, we can clearly see that CRPCA using coalescent masking and VAD gets better separation results than others.

Fig. 6 shows the results with the conventional RPCA, WRPCA, and the proposed methods on the DSD100 dataset. From the experimental results obtained with SDR, SIR, and NSDR values, again, it clearly shows that the proposed CRPCA using coalescent masking and VAD delivered the best separation results. Moreover, the value of SIR was improved by more than 10 dB in comparison with the conventional RPCA.

We also compared the running time of the proposed method with those of the previous methods on the above-mentioned two datasets. Table 1 lists the running time of each method on the ccMixter and DSD100 datasets. The running time on CRPCA was much shorter than on RPCA or WRPCA, while WRPCA had the worst results.

As the above-mentioned experimental results demonstrate, although WRPCA obtained better separation results than the conventional RPCA, the running time was much longer than RPCA on both datasets. CRPCA can utilize a prior target rank to separate audio source from the mixture signals, regardless of separation performance or running time, which leads to the superiority of CRPCA to RPCA and WRPCA. In the case of running time, WRPCA had the worst performance. As for the sepa-
ration performance in terms of NSDR, our proposed method delivered improvements by +2.56 dB and +2.95 dB on the ccMixter and DSD100 datasets, respectively. Indeed, in terms of SIR, the proposed method yielded estimates with significantly less interference, +10.29 dB and +11.45 dB, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RPCA</th>
<th>WRPCA</th>
<th>CRPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ccMixter</td>
<td>02:04:40</td>
<td>03:03:31</td>
<td><strong>00:52:10</strong></td>
</tr>
<tr>
<td>DSD100</td>
<td>04:34:30</td>
<td>06:49:28</td>
<td><strong>01:54:17</strong></td>
</tr>
</tbody>
</table>

7. Conclusion

In this paper, we have proposed blind monaural singing voice separation based on an extension of RPCA exploiting the constraint that the accompaniment spectrogram must have rank greater than or equal to one, and permitting its first singular values to be arbitrarily large without penalty. Time-frequency masking and harmonic masking are combined to construct coalescent masking, and VAD is utilized to constrain the singing voice and accompaniment values. Experimental results on the ccMixter and
DSD100 datasets demonstrate that the proposed method outperforms the conventional RPCA and WRPCA methods. As for running time, CRPCA is faster than RPCA and WRPCA under the same conditions, while WRPCA is the slowest. For future work, we will investigate robust graph embedding/learning approaches [60] [61] to optimize the separation performance from the mixed audio signal.

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. We also would like to thank Prof. Mark Hasegawa-Johnson (UIUC, US) for his valuable suggestion and revision in this work.

References


ings of 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 745-751.


Feng Li is currently pursuing his Ph.D. degree in Acoustic Information Science (AIS) Laboratory at Japan Advanced Institute of Science and Technology (JAIST) in Japan. He is a member of European Association for Signal Processing (EURASIP), the Asia Pacific Neural Network Society (APNNS), and the Acoustical Society of Japan (ASJ). His research interests include audio source separation, speech enhancement and noise reduction.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and
1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in human beings, and the signal processing of speech.