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

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)

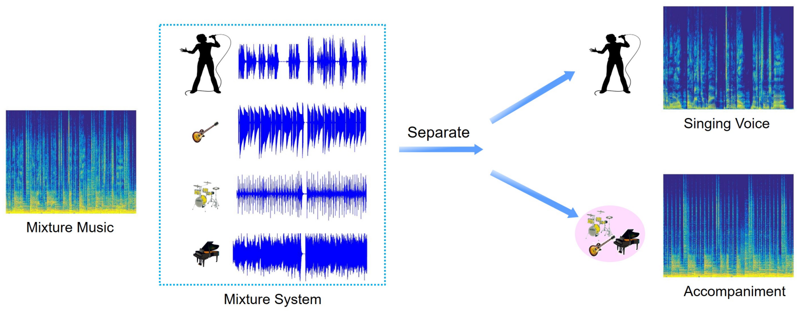

Figure 1: Illustration of blind monaural singing voice separation system.

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 back-end 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 [43] [44] [45] [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 [42] by melody extraction, which plays a vital role in separating singing voice [47] [48] [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 [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

Candés 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

$$\begin{aligned} &\text{minimize } |L|_* + \lambda |S|_1, \\ &\text{subject to } X = L + S, \end{aligned} \tag{1}$$

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 Candés 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

$$\begin{aligned} &\text{minimize } |L|_{w,*} + \lambda |S|_1, \\ &\text{subject to } X = L + S, \end{aligned} \tag{2}$$

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 Candés et al. [37]. We also adopted an efficient inexact ALM [50] to solve this convex model. The corresponding augmented Lagrange function is defined as

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

where J is the Lagrange multiplier and μ 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

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

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 Candés et al [51] proposed reconstruct sparse signals, the reweighing formula can be defined as

$$
w_i^l = \frac{C}{\delta_i(L_l) + \varepsilon}, \tag{5}
$$

so the weighted values will converge to

$$
L^* = U\Sigma' V^T, \tag{6}
$$

where

$$
\Sigma' = \begin{pmatrix} diag(\delta_1(L^*), \delta_2(L^*), ..., \delta_n(L^*)) \\ 0 \end{pmatrix}, \tag{7}
$$

and

$$
\delta_i(L^*) = \begin{cases} 0 \\ \dfrac{c_1 + \sqrt{c_2}}{2} \end{cases} \tag{8}
$$

Algorithm 1 WRPCA for singing voice separation

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

Initialize: $\rho, \mu_0, L_0 = X, J_0 = 0, k = 0$.

While not converge,

do :

$S_{k+1} = \arg\min |S|_1 + \frac{\mu_k}{2}|X + \mu_k^{-1} J_k - L_k - S|_F^2$.

$L_{k+1} = \arg\min |L|_{w,*} + \frac{\mu_k}{2}|X + \mu_k^{-1} J_k - S_{k+1} - L|_F^2$.

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

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

$k = k + 1$.

end while.

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

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

$$\begin{aligned} &\text{minimize} \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda |S|_1, \\ &\text{subject to } X = L + S, \end{aligned} \tag{9}$$

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 Candés 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

$$\begin{aligned} J(X, L, S, \mu) = \ &\min \sum_{i=2}^{min(m,n)} \delta_i(L) + \lambda |S|_1 \\ &+ < J, X - L - S > + \frac{\mu}{2} |X - L - S|_F^2, \end{aligned} \tag{10}$$

where J is the Lagrange multiplier and μ is a positive value.

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

$$\begin{aligned} L_{k+1} = \ &\min_L \sum_{i=2}^{\min(m,n)} \delta_i(L) + \langle J_k, X - L - S_k \rangle \\ &+ \frac{\mu_k}{2} |X - L - S_k|_F^2, \end{aligned} \tag{11}$$

$$\begin{aligned} S_{k+1} = \ &\min_S \ \lambda |S|_1 + \langle J_k, X - L_k - S \rangle \\ &+ \frac{\mu_k}{2} |X - L_k - S|_F^2, \end{aligned} \tag{12}$$

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

$$L_{k+1} = P_{1,\mu_k^{-1}}(X - S_k + \mu_k^{-1} J_k), \tag{13}$$

Algorithm 2 CRPCA for singing voice separation

Input: Mixture signal $X \in \mathbb{R}_{m\times n}$.

Initialize: $\rho > 1, \mu_0 > 0, k = 0, L_0 = S_0 = 0$.

While not converge,

do :

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

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

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

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

$\quad k = k + 1$.

end while.

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

$$S_{k+1} = Q_{\lambda\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1} J_k), \tag{14}$$

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

$$P_{1,\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V_Y^T, \tag{15}$$

where the soft-thresholding operator [53] can be defined as

$$Q_{\mu_k^{-1}}(D_{Y_2}) = sign(D_{Y_2}) \cdot max(|D_{Y_2}| - \mu_k^{-1}, 0), \tag{16}$$

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 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

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

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

$$H(t, s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2(n)), \tag{18}$$

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

$$C = argmax \sum_{t=1}^{T-1} \left(\log a_t H(t, s_t) + \log T(s_t, s_{t+1})\right), \tag{19}$$

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

$$M_h(t, f) = \begin{cases} 1 & nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & others \end{cases} \tag{20}$$

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

$$M_c = M_{ibm} \otimes M_h \tag{21}$$

where M_{ibm} and M_h are the time frequency masking and harmonic masking, respectively, and $\otimes$ 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

$$S_{vocal} = M_c \otimes X \tag{22}$$

where $\otimes$ 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

$$\Omega(t) = \sum_f \left(\frac{1}{H_f} \sum_{n=1}^{H_f} P(t, s + 1200 \log_2(n)) \right)^{1.8}, \tag{23}$$

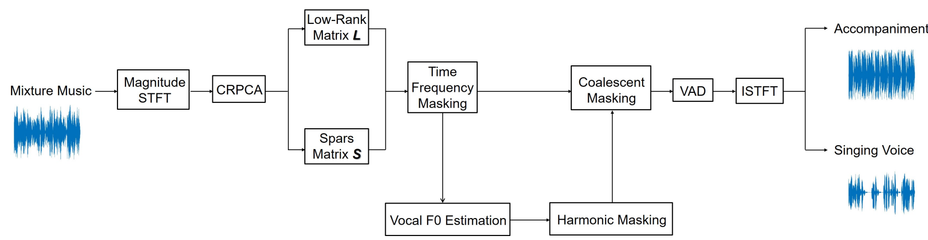

Figure 2: Block diagram of proposed blind monaural singing voice separation system.

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.

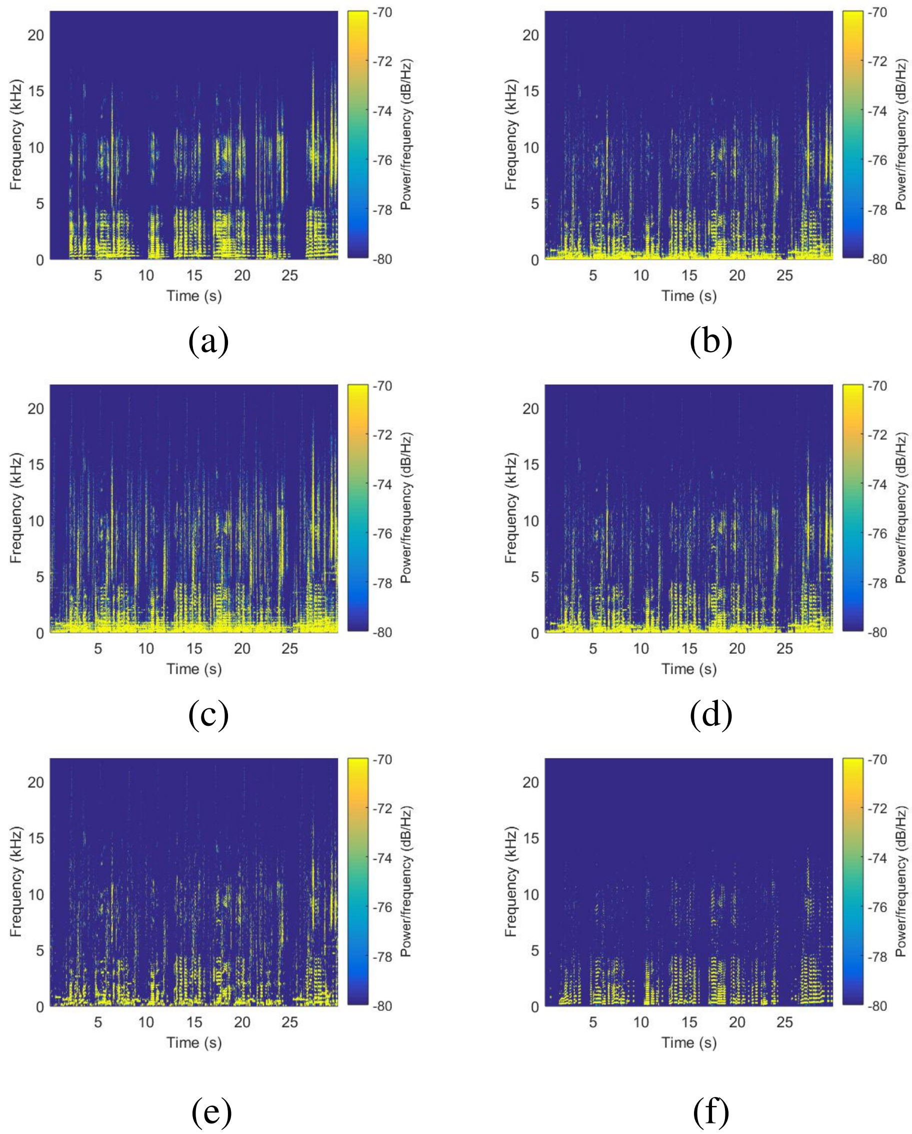

Figure 3: Spectrograms are excerpted from AlexBeroza_-_To_Be_Sensitive_(with_mindmapthat) in the ccMixter dataset: **(a)** spectrogram of original singing voice, **(b)** spectrogram of separated singing voice by RPCA, **(c)** spectrogram of separated singing voice by WRPCA, **(d)** spectrogram of separated 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.

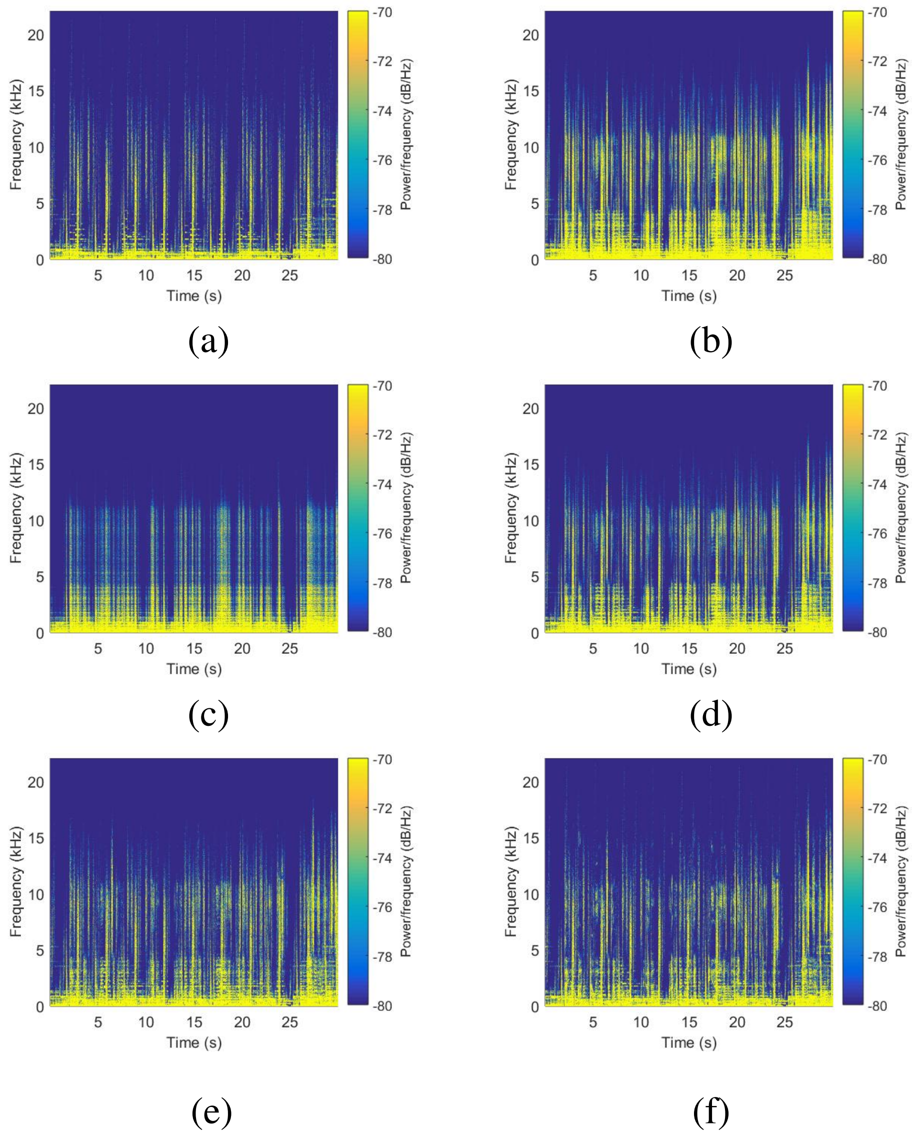

Figure 4: Spectrograms are excerpted from AlexBeroza_-_To_Be_Sensitive_(with_mindmapthat) in ccMixter dataset: **(a)** spectrogram of original accompaniment, **(b)** spectrogram of separated accompaniment by RPCA, **(c)** spectrogram of separated accompaniment by WRPCA, **(d)** spectrogram of separated 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.

[1] https://members.loria.fr/ALiutkus/kam/
[2] http://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][3]. The estimated signal $\hat{S}$(t) is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t), \tag{24}$$

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

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \left(S_{interf}(t) + S_{artif}(t)\right)^2}, \tag{25}$$

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \tag{26}$$

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \tag{27}$$

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

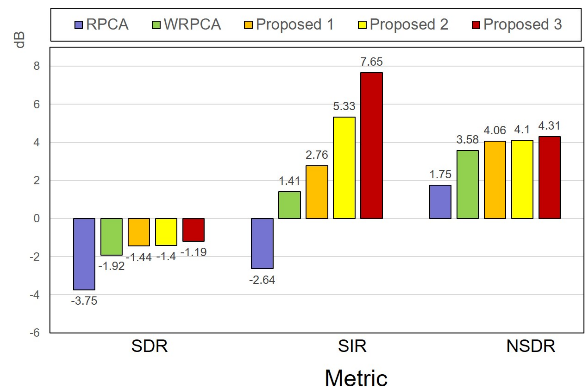

Figure 5: Comparison of monaural singing voice separation results on **ccMixter** dataset for conventional RPCA, WRPCA, CRPCA, CRPCA with IBM, and CRPCA using coalescent masking and VAD in terms of SDR, SIR, and NSDR, respectively.

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-

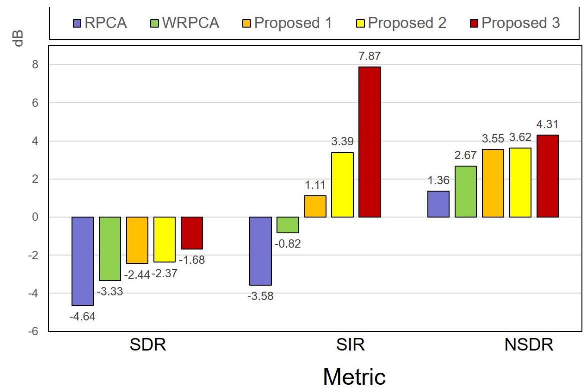

Figure 6: Comparison of monaural singing voice separation results on **DSD100** dataset for conventional RPCA, WRPCA, CRPCA, CRPCA with IBM, and CRPCA using coalescent masking and VAD in terms of SDR, SIR, and NSDR, respectively.

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 1: Running time (hh:mm:ss)

Dataset	RPCA	WRPCA	CRPCA
ccMixter	02:04:40	03:03:31	**00:52:10**
DSD100	04:34:30	06:49:28	**01:54:17**

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

[1] 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.

[2] S. Jo and C. D. Yoo, Melody extraction from polyphonic audio based on particle filter, in: Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, Netherlands, 2010, pp. 357-362.

[3] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, Content-based music information retrieval: current directions and future challenges, in: proceedings of the IEEE, 96.4, 2008, pp. 668-696.

[4] 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.

[5] K. Z. Qian, Y. Zhang, S. Y. Chang, X. S. Yang, D. Florêncio, M. H. Johnson, Speech enhancement using bayesian wavenet, in: Processsing of Interspeech, 2017, pp. 2013-2017.

[6] T. Higuchi, H. Kameoka, Unified approach for audio source separation with multichannel factorial HMM and DOA mixture model, in: Proceedings of European Signal Processing Conference (EUSIPCO), 2015, pp. 2043-2047.

[7] 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.

[8] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, J. Fontecave, 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.

[9] E.M. Grais, M.U. Sen, H. Erdogan, Deep neural networks for single channel source separation, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3734-3738.

[10] 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.

[11] 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.

[12] Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26.4, 2018, pp. 787-796.

[13] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde, Singing voice separation with deep U-Net convolutional networks, in: Proceed-

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

[14] 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, 26.8, 2018, pp. 1307-1335.

[15] 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.

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

[17] M. Oh, H.M. Park, Blind source separation based on independent vector analysis using feed-forward network, Neurocomputing, 74.17, 2011, pp. 3713-3715.

[18] 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, 15.3, 2007, pp. 1066-1074.

[19] 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.

[20] 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.

[21] Z. Chen, D. Ellis, Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA, October 2013.

[22] C. Sun, Q. Zhang, J. Wang, J. Xie, Noise reduction based on robust principal component analysis, Journal of Computational Information Systems, 10.10, 2014, pp. 4403-4410.

[23] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, T. Kawahara, H. Okuno, Speech Enhancement Based on Bayesian Low-Rank and Sparse Decomposition of Multichannel Magnitude Spectrograms, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26.2, 2018, pp. 215-230.

[24] F. Biondi, Low rank plus sparse decomposition of synthetic aperture radar data for maritime surveillance, International Workshop on Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing, CoSeRa 2016, pp. 75-79.

[25] F. Biondi, A Polarimetric Extension of Low-Rank Plus Sparse Decomposition and Radon Transform for Ship Wake Detection in Synthetic Aperture Radar Images, IEEE Geoscience and Remote Sensing Letters, 2018.

[26] A. Das, A Bayesian Sparse-Plus-Low-Rank Matrix Decomposition Method for Direction-of-Arrival Tracking, IEEE Sensors Journal, 17.15, 2017, pp. 4894-4902.

[27] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, On the Applications of Robust PCA in Image and Video Processing, Proceedings of the IEEE, 2018, pp. 1427-1457

[28] T. Bouwmans, A. Sobral, S. Javed, S. Jung, E. Zahzah, Decomposition into Low-rank plus Additive Matrices for Background/Foreground Separation: A Review for a Comparative Evaluation with a Large-Scale Dataset, Computer Science Review, Volume 23, 2017, pp. 1-71,

[29] N. Vaswani, T. Bouwmans, S. Javed, P. Narayanamurthy, Robust Subspace Learning: Robust PCA, Robust Subspace Tracking and Robust Subspace Recovery, IEEE Signal Processing Magazine, 35.4, 2018, pp. 32-55.

[30] 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.

[31] Y. H Yang, Low-rank representation of both singing voice and music accompaniment via learned dictionaries, in: Proceedings of 18th International Society for Music Information Retrieval Conference (ISMIR), 2013, pp. 427-432.

[32] Z. Rafii, B. Pardo, Repeating pattern extraction technique (REPET): a simple method for music/voice separation, IEEE transactions on audio, speech, and language processing, 21.1, 2013, pp. 73-84.

[33] P. Sprechmann, A. Bronstein, G. Sapiro, Real-time online singing voice separation from monaural recordings using robust low-rank modeling, in: Proceedings of 17th International Society for Music Information Retrieval Conference (ISMIR), 2012, pp. 67-72.

[34] D. Fourer, G. Peeters, Single-Channel Blind Source Separation for Singing Voice Detection: A Comparative Study, Preprint, 2018.

[35] T.S.T. Chan, Y.H. Yang, Informed Group-Sparse Representation for Singing Voice Separation, IEEE Signal Processing Letters, 24.2, 2017, 156-160.

[36] J. Pu, Y. Panagakis, S. Petridis, J. Shen, M. Pantic, Blind Audio-Visual Localization and Separation via Low-Rank and Sparsity, IEEE Transactions on Cybernetics, 2018, pp. 2168-2267.

[37] E. J. Candés, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM (JACM), 58.3, 2011.

[38] 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.

[39] 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.

[40] 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.

[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 transactions on pattern analysis and machine intelligence, 38.4, 2016, pp. 744-758.

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

[43] T. Oh, A Novel Low-Rank Constraint Method with the Sparsity Model for Moving Object Analysis, Master Thesis, KAIST, 2012.

[44] W. Leow, Y. Cheng, L. Zhang, T. Sim, L. Foo, Background Recovery by Fixed-rank Robust Principal Component Analysis, in Proceedings of the International Conference on Computer Analysis of Images and Patterns (CAIP), 2013. pp. 54-61.

[45] J. Xue, Y. Zhao, W. Liao, J. Cha, Total Variation and Rank-1 Constraint RPCA for Background Subtraction, IEEE Access, September 2018.

[46] I. Kajo, N. Kamel, Y. Ruichek, A. Mali, SVD-based Tensor Completion Technique for Background Initialization, IEEE Transaction on Image Processing, 2018.

[47] J. Salamon, E. Gómez, D. P.W. Ellis, G. Richard, Melody extraction from polyphonic music signals: approaches, applications, and challenges, IEEE Signal Processing Magazine, 31.2, 2014, pp. 118-134.

[48] J. Salamon, E. Gómez, Melody extraction from polyphonic music signals using pitch contour characteristics, IEEE Transactions on Audio, Speech, and Language Processing, 20.6, 2012, pp. 1759-1770.

[49] 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, 24.11, 2016, pp. 2084-2095.

[50] 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.

[51] E. J. Candés, M. B. Wakin, S. Boyd, Enhancing sparsity by reweighted l_1 minimization, Journal of Fourier analysis and applications, 14.5, 2008, pp. 877-905.

[52] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, L. Zhang, Weighted nuclear norm minimization and its applications to low level vision, International journal of computer vision, 121.2, 2017, pp. 183-208.

[53] E. Hale, W. Yin, Y. Zhang, Fixed-point continuation for ℓ_1-minimization: Methodology and convergence, SIAM Journal on Optimization, 19.3, 2008, pp. 1107-1130.

[54] D.J. Hermes, Measurement of pitch by subharmonic summation, The journal of the acoustical society of America, 83.1, 1998, pp. 257-264.

[55] Forney, G. David, The viterbi algorithm, Proceedings of the IEEE, 61.3, 1973, 268-278.

[56] S.H. Nawab, T.F. Quatieri, J.S. Lim, Signal reconstruction from short-time Fourier transform magnitude, IEEE Transactions on Acoustics, Speech, and Signal Processing, 31.4, 1983, pp. 986-998.

[57] N. Sturmel,L. Daudet, Signal reconstruction from STFT magnitude: A state of the art. In International conference on digital audio effects (DAFx), 2011, pp. 375-386.

[58] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, L. Daudet, Kernel additive models for source separation, IEEE transactions on audio, speech, and language processing, 62.16, 2014, pp. 4298-4310.

[59] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE transactions on audio, speech, and language processing, 14.4, 2006, pp. 1462-1469.

[60] N. Han, J. Wu, Y. Liang, X. Fang; W. Wong, S. Teng, Low-rank and sparse embedding for dimensionality reduction, Neural Networks, 2018, pp. 202-216.

[61] Z. Kang, H. Pan, S. Hoi, Z. Xu, Robust Graph Learning from Noisy Data, Preprint, December 2018.

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.