Separation of Singing Voice from Music using Robust Principle Component Analysis and its Extensions

Feng Li

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Separation of Singing Voice from Music using Robust Principle Component Analysis and its Extensions

Feng Li

Supervisor: Professor Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]

September 2019

Abstract

The development in multimedia technologies has promoted dramatically the rapid growth of music data in recent years. There are various different applications for people's demands in music such as information retrieval, identification and handing. However, singing voice and background music are related to each other in the mixed music, the mutual interference has brought huge obstacles to music information processing. The problem of how to extract the audio information from music has become an important research topic. As the part of music information retrieval, the technologies of singing voice separation are facing unprecedented challenge.

The objective of this research is to deal with the problem of singing voice separation from monaural recordings. It is even more difficult than multichannel since the spatial information cannot be applied in the separation procedure. Singing voice separation is a technique for separating or extracting singing voice from a musical mixture, which has found many applications in the wide areas such as singer identification, singing evaluation and query by humming. This is a relatively easy separation task of the human auditory system, but it becomes more difficult when we attempt to simulate this problem in a computational method. To achieve the task of singing voice separation, this study mainly focuses on robust principal component analysis (RPCA) and its extensions.

RPCA has been recently proposed of popularization and effectiveness way of separation approach that separates singing voice and accompaniment from a mixture music. 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 is an effective approach to separate singing voice from the mixed audio signal, 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 algorithm over-estimates 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. Thus, it motivates us to describe exactly the separated low-rank matrix.

To overcome the disadvantage of RPCA for singing voice separation, two extensions of RPCA algorithm are proposed in this dissertation. One is called weighted robust principal component analysis (WRPCA). It uses different weighted values to describe the low-rank matrix for singing voice separation. Additionally, incorporating the proposed WRPCA with gammatone auditory filterbank for singing voice separation. The significance of WRPCA can describe different low-rank matrix under the conditions of human's auditory perceptual properties. Because the cochleagram is derived from non-uniform time-frequency transform whereas time-frequency units in low-frequency regions have higher resolutions than in the high-frequency regions, which closely resembles the functions of the human ear. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of the spectrogram.

Another extension of RPCA with rank-1 constraint called constraint RPCA (CRPCA). It 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. Thus, it not only provides a robust solution to large dynamic range differences among instruments but also reduces the computation complexity. Then, incorporating the proposed CRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. In addition, constructing coalescent masking and vocal activity detection on CRPCA method to constrain the temporal segments that allowed to constrain singing voice from the mixed music datum. Finally, combining F0 and non-negative rank-1 constraint

RPCA, which incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm.

In conclusion, this dissertation proposes two extensions of the effective optimization algorithms concentrating on RPCA for singing voice separation. One is using different weighted value for describing the separated low-rank matrix. The other is exploring rank-1 constraint minimization of singular value in RPCA. In terms of source-to-artifact ratio, the previous is better than the later. However, CRPCA obtains better separation quality than WRPCA in singing voice separation. The outcomes of this research contribute to further improving the technologies related to music information retrieval. Additionally, the potential contribution of this research is to deal with the problems of noise reduction and speech enhancement by using the separated low-rank and sparse model. Since the background noise is assumed as the part of low-rank component and the human speech is regarded as the part of sparse component.

Keywords: Singing voice separation, robust principal component analysis, weighted, rank-1 constraint, F0.

Acknowledgments

The past three academic years for my PhD study has been full of fun and fruitful period of my life in Japan. As I am finishing up my study by now, I can easily admit that all my achievements own to my advisors, teachers, labmates, friends, and family.

First and foremost, I would like to express my deepest appreciation to Prof. Masato Akagi, my Ph.D supervisor, who gives his extremely support, continuous teaching, and patient guidance during the period of my doctoral study at Japan Advanced Institute of Science and Technology (JAIST). This dissertation could not be accomplished without his precious suggestions and well-directed advice. I am grateful for his insights and encouragement, which accelerate my academic research in cracking complex problems.

I would also like to express my gratitude towards co-advisor Prof. Masashi Unoki of JAIST in Japan. Professor Unoki checked carefully and gave me many precious comments for my presentation and research report.

I am also grateful to minor research advisor Prof. Jianwu Dang of JAIST and Tianjin University for his suggestion and comments, as well as being a member of the dissertation committee. When I talked with Prof. Dang, I could feel his energy for research which inspire me a lot.

I would like to express my thanks to Prof. Mark Hasegawa-Johnson, University of Illinois at Urbanna-Champaign (UIUC), who give me an opportunity to study in the USA. I also would like to thank Dr. Kaizhi Qian from UIUC, who gives me lots of suggestions and comments when I stay at UIUC.

I would like to thank Dr. Rieko Kubo, Dr. Yongwei Li and Dr. Yawen Xue, who gave me much information during my doctoral study and gave me many helps between the life and research at JAIST.

I want to appreciate Dr. Zhichao Peng and all members in Acoustic Information Science (AIS) Laboratory for their valuable comments, helps and encouragements at JAIST. The discussion among us always provided me many new ideas and gave me a lot of happiness during the boring research period.

I am also thankful to the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan Scholarship and the China Scholarship Council (CSC) of China Scholarship to support me during my doctoral studies in Japan.

Finally, I would like to delicate this dissertation to my parents, my father and my mother, for their forever encouragement and love over many years. Without their support and understand, I can not achieve all things I have now.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

RPCA	Robust Principal Component Analysis
WRPCA	Weighted Robust Principal Component Analysis
CRPCA	Constraint Robust Principal Component Analysis
STFT	short-time Fourier transform
ISTFT	Inverse short-time Fourier transform
MLRR	Multiple Low-Rank Representation
VAD	Vocal Activity Detection
IBM	Ideal Binary Masking
IRM	Ideal Ratio Masking
MIR	Music Information Retrieval
ReLu	Rectified Linear Units
DNN	Deep Neural Network
KAM	Kernel Additive Modeling
CNN	Convolutional Network Network
LSTM	Long Short-Term Memory
ALM	Augmented Lagrange Multipliers
iALM	inexact Augmented Lagrange Multiplier
REPET	REpeating Pattern Extraction Technique
ADMM	Alternating Direction Method of Multipliers
APG	Accelerated Proximal Gradient
T-F	Time-Frequency
F0	Fundamental Frequency
SDR	Source-to-Distortion Ratio
SIR	Source-to-Interference Ratio
SAR	Source-to-Drtifact Ratio
NSDR	Normalized SDR

NCRPCA	Non-negative Constraint Robust Principal Component Analysis
NCRPCAi	Informed Non-negative Constraint Robust Principal Component Analysis
SVD	Singular Value Decomposition
NMF	Non-negative Matrix Factorization
KL	Kullback-Leibler
IS	Itakura-Saito
EUC	Euclidean
CI	Cochlear Implant

Chapter 1

Introduction

1.1 Motivation

In recent years, the development in multimedia technologies has promoted dramatically the rapid growth of music data. There are various different applications for people's demands in music such as information retrieval, identification and handing. However, singing voice and background music are related to each other in the mixed music, the mutual interference has brought huge obstacles to music information processing. The problem of how to extract the audio information from music signal has become an important topic. As the part of music information retrieval, the technologies of singing voice separation are facing unprecedented challenge.

Singing voice separation is a technique for separating or extracting singing voice from a musical mixture, which has found many applications in the wide areas like music information retrieval [1], singer identification [2], music emotion recognition [3], chord recognition [4], melody extraction [5], drum extraction [6], Karaoke applications [7], and education for musical instruments [8].

This is a relatively easy separation task of the human auditory system, but it becomes more difficult when we attempt to simulate this problem in a computational method. Although there are many methods for singing voice separation, the separation quality is not well because the many instruments are coexisting in the background music. The separation results of state-of-the-art methods are still far behind human hearing capability. The existing problems of singing voice separation are still facing severe challenging [9] [10]. Therefore, it is an important task

for solving the problem of singing voice separation.

Many academic challenges about singing voice separation were also hold in the previous years. For example, the organizations of Music Information Retrieval Evaluation eXchange (MIREX) and Signal Separation Evaluation Campaign (SiSEC) are also evaluated for singing voice separation task. MIREX is an annual challenges, which contains of various tasks related to the problems of music information retrieval. Since 2014, singing voice separation is included as a sub-task of MIREX. SiSEC is held one and a half year. It consists of the different problems about audio source separation task. Music source separation is also included as a sub-task of SiSEC, which separates singing voice (vocals) from the musical mixture (vocals, drums, bass, and others).

Motivated by the above considerations, an effective optimization algorithm plays an important role in singing voice separation. In particular, the audio information of singing voice can be described exactly and improve the separation quality from the music. This study mainly focuses on solving the problem of singing voice separation in monaural recording. It is even more difficult than multichannel since the spatial information cannot be applied in the separation procedure. Therefore, research in the field of monaural singing voice separation become very hot topic. Many methods are focused on unsupervised and supervised learning. As for supervised method, deep learning is the most popular method in monaural singing voice separation. However, a large number of training data are needed in advanced. So, the unsupervised learning has made great progress in singing voice separation.

Therefore, to obtain better separation performance from the observed mixed music, the effective optimization algorithm is need to be solved by unsupervised learning method in the singing voice separation task, the main works of the dissertation are focused on different optimization approaches for singing voice separation.

1.2 Methodology

Currently, robust principal component analysis (RPCA) [11] has been recently proposed of popularization and effectiveness way of separation approach that separates singing voice and accompaniment from a mixture music in monaural recording. 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).

As for the mixture music, 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. Therefore, RPCA method can be well-described as the part of singing voice from the mixed music signal by the separated sparse matrix. The mixture music signal can be described as the low-rank and sparse model. And the process of RPCA decomposition is very suited to the singing voice separation task.

Although the model of RPCA has been successfully applied to singing voice separation task, it fails when there are significant differences in dynamic range between 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 over-estimates the rank of a matrix that includes drum sounds. The accuracy of such singing voice separation results thus decreases, as drums may be placed in the sparse subspace instead of being low-rank.

Therefore, to obtain the better separation performance in singing voice separation, this dissertation mainly focuses on RPCA and its extension for singing voice separation. Two extensions of RPCA were proposed in this dissertation. Figure 1.1 shows the proposed methods for singing voice separation in this dissertation. The methods are mainly focus on RPCA for singing voice separation.

The first extension of RPCA called WRPCA method. On the one hand, evaluate the proposed WRPCA for singing voice separation. It utilizes different weighted values to constraint the separated low-rank matrix. The experimental evaluation is carried out on the ccMixter dataset and on the DSD100 dataset. On the other hand, combining the proposed WRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. And the experiments are conducted on the ccMixter and DSD100 datasets. However, WRPCA suffers from high computational cost due to computing the singular value decomposition at each iteration during the separation processing. Hence, the running time of WRPCA is slower than RPCA. Therefore, we propose another deformation instead of WRPCA method for singing voice separation.

The another extension of RPCA called CRPCA method, which utilizes the rank-1 constraint

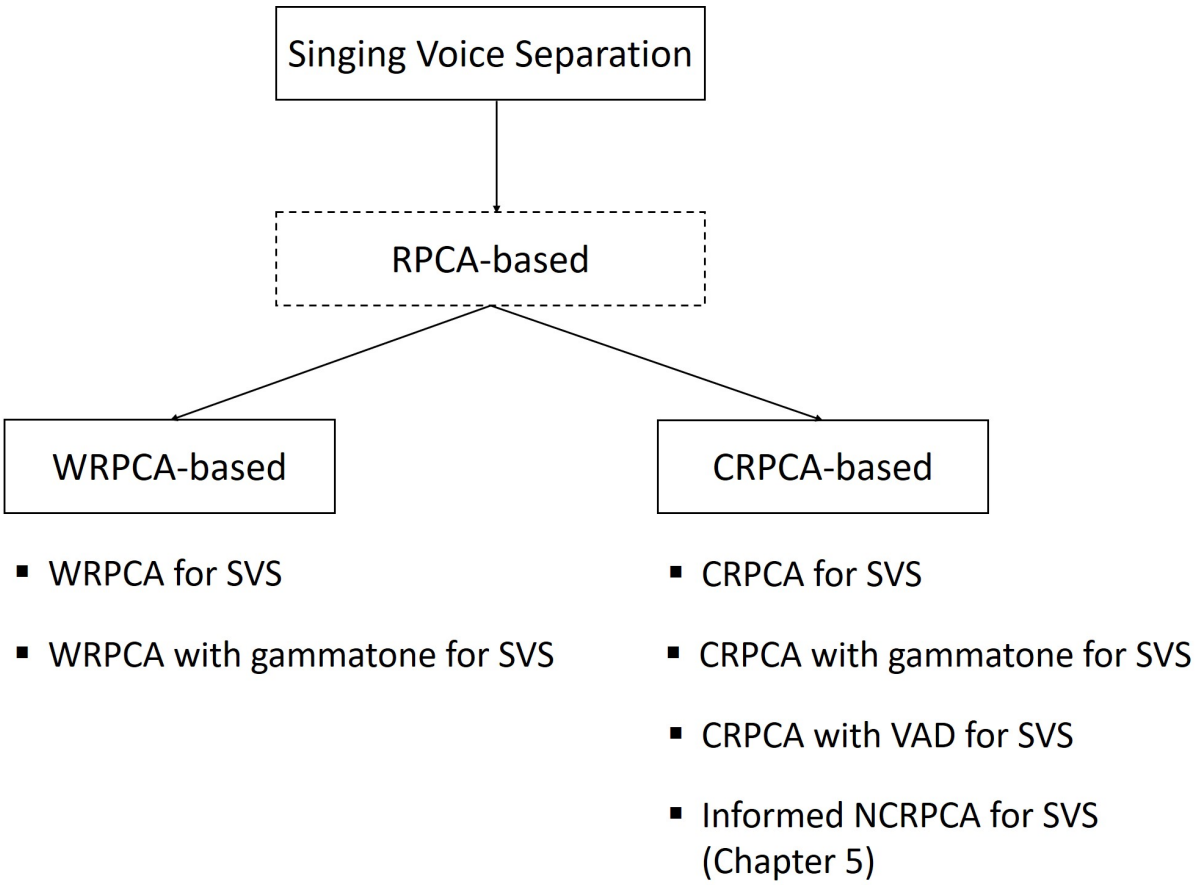


Figure 1.1: The proposed methods for singing voice separation (SVS) in the dissertation.

minimization of singular values in RPCA instead of minimizing the nuclear norm for separating singing voice from the mixture music. Thus, it not only provides a robust solution to large dynamic range differences among instruments but also reduces the computation complexity. The experiment are conducted by combining other feature to evaluate the proposed methods on the different databases.

Firstly, we evaluate CRPCA model on the ccMixter and DSD100 datasets. Secondly, we combine the proposed CRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. Thirdly, we construct the coalescent masking and vocal activity detection (VAD) to constrain the temporal segments that allowed to constrain singing voice. The results on the ccMixter and DSD100 datasets reveal that the proposed method are very effective than the previous in singing voice separation task. Finally, we introduce a singing voice separation method by combining the human-labeled F0 and non-negative CRPCA to separated the singing voice from the mixture music. Experiment evaluation is compared with the previous methods on the iKala dataset.

As for the above discussed two extensions of RPCA method, WRPCA and CRPCA, respectively. In terms of source-to-artifact ratio, WRPCA obtains the better results than CRPCA in singing voice separation. However, CRPCA can get the better separation performance in source-to-distortion ratio and source-to-interference ratio.

1.3 Research goal

The goal of this research is to deal with the problem of singing voice separation from monaural recordings. It is even more difficult than multichannel since the spatial information cannot be applied in the separation procedure.

To achieve the task of singing voice separation, this study mainly focuses on RPCA and its extensions. Because RPCA is one of the popularization of such separation algorithm. It decomposes a given amplitude spectrogram of a mixture signal into the sum of a low-rank matrix (accompaniment) and a sparse matrix (singing voice).

Since the instruments reproduce nearly the same sounds every time, 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 algorithm has been successfully applied to singing voice separation, it fails when one singular value (e.g., drums) is much larger than all others (e.g., bass, guitar or other accompanying instruments). The accuracy of such separation results thus decreases, as drums may be placed in the sparse subspace instead of being the low-rank from mixture original matrix.

With regards to RPCA-based approach, the main method in this dissertation mainly focuses on solving the disadvantage of RPCA algorithm for singing voice separation in monaural recordings.

1.4 Organization of the dissertation

Figure 1.2 shows organization of the dissertation and the remainder of the dissertation is structured as follows:

- **Chapter 2** provides the background about related work of singing voice separation. First, introduces the previous studies and methods in the task of singing voice separation. Then, gives some related databases in singing voice separation tasks in this dissertation. Finally, explains several evaluation metrics to measure the separation performance of the proposed methods.

- **Chapter 3** proposes an extension of RPCA called WRPCA, which describes the different weighted values to constraint the separated low-rank matrix. The experimental evaluation is carried out on the ccMixter dataset and on the DSD100 dataset. In addition, combines the proposed WRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. All the experiments are conducted on the ccMixter and DSD100 datasets.

- **Chapter 4** describes another 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. Then, combines the proposed CRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. Finally, constructs coalescent masking and incorporates vocal activity detection to constrain the temporal segments that allowed to constrain singing voice. The experiments are evaluated on the ccMixter and DSD100 datasets.

- **Chapter 5** proposes a singing voice separation method by combining F0 and non-negative CRPCA, which incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm. Experimental evaluation are conducted on the iKala dataset.

- **Chapter 6** first summarizes all of this work in this dissertation. Then, draws the conclusions focuses on the proposed methods for singing voice separation. And the future works are discussed in the end.

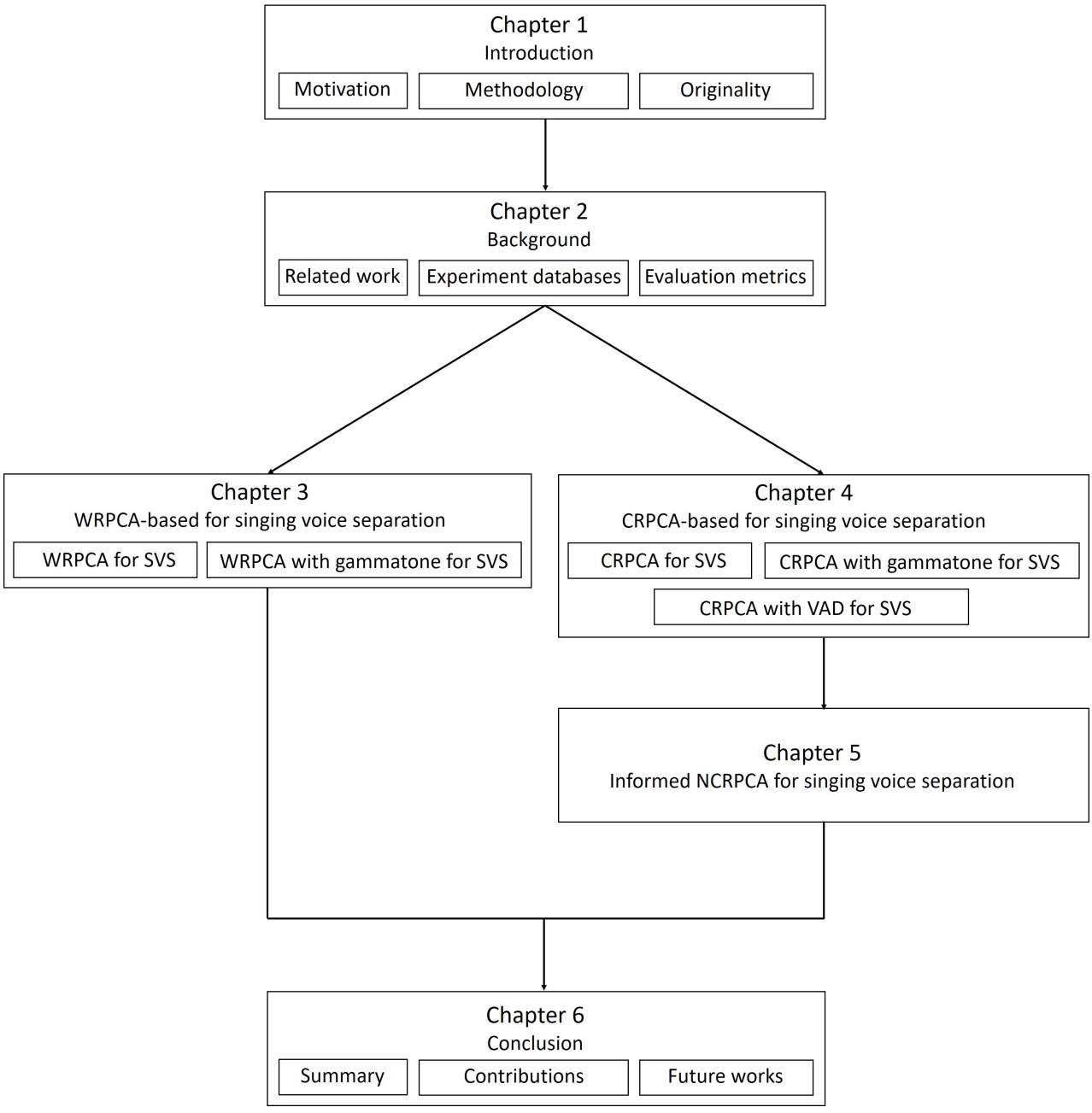


Figure 1.2: Organization of the dissertation.

Chapter 2

Background

Mixture music signal is very popular in our daily life, which is the main research target in this dissertation. It contains the singing voice and various instruments (e.g., piano, drums, guitar and others). Singing voice separation is a technique for separating singing voice from a musical mixture and has been intensively studied in recent years. This technique can be used for many applications including music information retrieval [1], Karaoke applications [7], chord recognition [4], music auto-tagging [12], singing lyric recognizer [13] [14], melody extraction [5], and fundamental frequency (F0) estimation [15].

However, the results on state-of-the-art methods are still far behind human hearing capability. The existing problems of singing voice separation are still faced with serious challenges [9] [10] [16] due to the musical instruments involved and time-varying spectral overlap between singing voice and background music.

Figure 2.1 illustrates the system of singing voice separation. This figure shows that after separating the singing voice from mixture system by the separation algorithm, the separated singing voice and accompaniment can be obtained from the musical mixture. Therefore, it is obvious that effective optimization algorithms play a significant role in the process of the separation task.

Until recently, there have been many approaches proposed to solve the difficult in singing voice separation tasks. It can be divided into two categories: unsupervised and supervised learning methods.

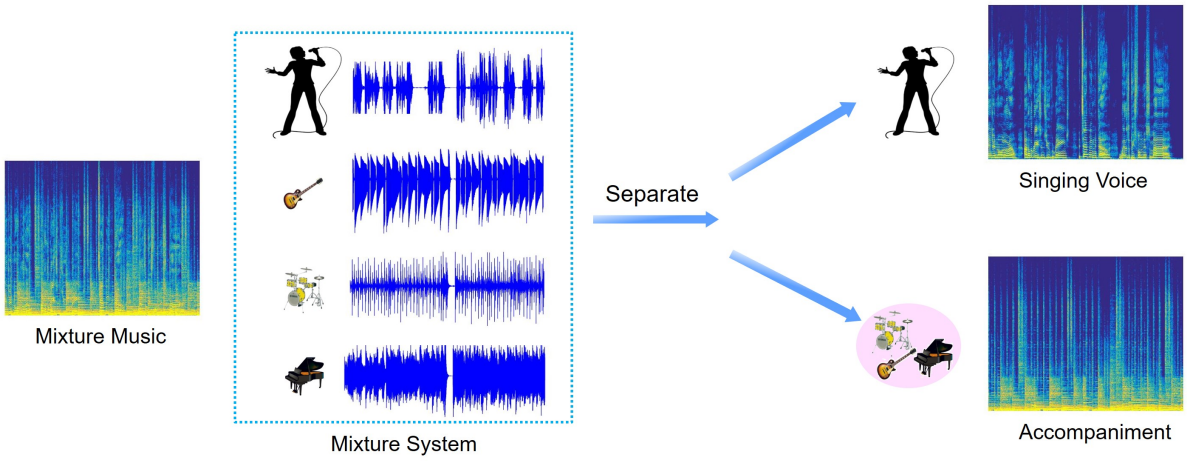


Figure 2.1: Illustrate the system of singing voice separation.

2.1 Related work

According to the previous studies, the separation approaches are mainly divided into two categories: unsupervised and supervised methods, respectively. In terms of unsupervised methods for singing voice separation, sparse or low-rank approximation assumption is typical methods for singing voice separation, for example, Non-negative Matrix Factorization (NMF) and RPCA methods. Additionally, another popular approach for singing voice separation is based on the repetitive nature of background music (REPET). As for the supervised method, deep learning-based methods are very poplar for singing voice separation.

2.1.1 Non-negative matrix factorization

NMF [17] [18] [19] [20] [21] [22] is a specially sparse representation algorithm model for singing voice separation, 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. The matrix decomposition model can be defined as follows:

$$V \approx WH, \tag{2.1}$$

where $V(V \in R_{m \times n})$ is an observed non-negative matrix that represents an amplitude spectrogram of sound source signals, $W(W \in R_{m \times k})$ is a non-negative basis matrix of a sound signal as column vectors, $H(H \in R_{k \times n})$ is a non-negative activation matrix that corresponds to the activation of each

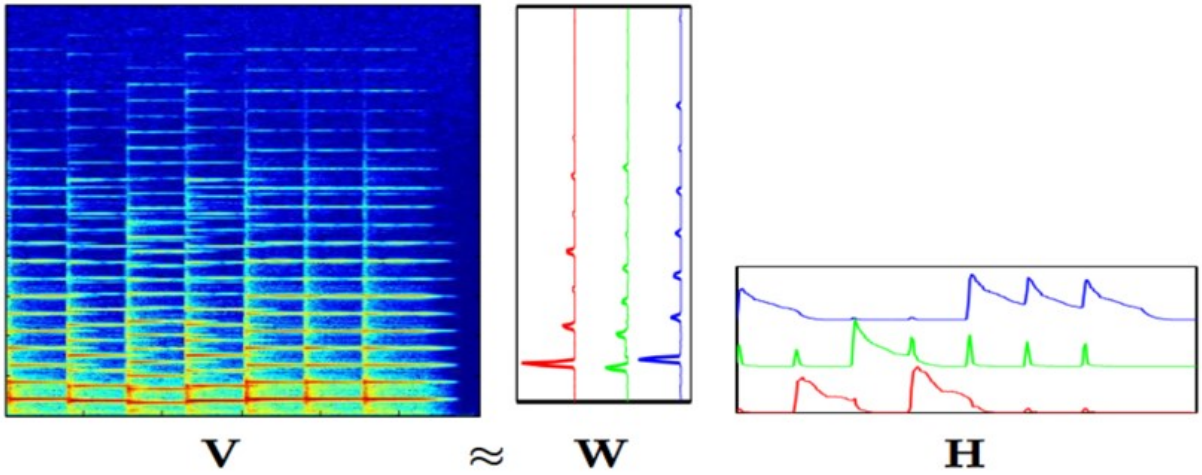


Figure 2.2: The decomposition model of NMF, which uses KL divergence and $K = 3$ on Mery Had a Little Lamb. **V** is the mixture matrix, **W** is the basic matrix and **H** is the activation matrix which describes the time-varying gains for each basis vector.

basis vector of W, m and n are the rows and columns of observed sound signals, respectively. And k is the number of supervised signal basis vectors. Usually, we choose m × k + k × n << m × n; hence reducing the dimensions of the input data. Figure 2.2 gives an example of NMF decomposition. And the basic vector is set as $k = 3$ with KL divigence on Mery Had a Little Lamb. The basic matrix shows representative spectral patterns, while the activation matrix illustrates time-varying gains for each basis vector.

The β-divergence [23] [24] is a family of cost functions parameterized by a signal shape parameter β and can be defined as

$$D_\beta(y|x) = \begin{cases} \dfrac{y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}}{\beta(\beta-1)}, & \beta \in \mathbb{R} \setminus \{0,1\} \\ \dfrac{y}{x} - log\dfrac{y}{x} - 1, & (\beta = 0) \\ ylog\dfrac{y}{x} + x - y. & (\beta = 1) \end{cases} \tag{2.2}$$

Generally, the cost functions in NMF can be calculated by the following three distances: Itakura-Saito divergence ($\beta = 0$), Kullback-Leibler divergence ($\beta = 1$), and Euclidean distance

($\beta = 2$). The corresponding formulas are given as

$$D_\beta(y|x) = \begin{cases} \frac{y}{x} - log\frac{y}{x} - 1, & (\beta = 0) \\ ylog\frac{y}{x} + x - y, & (\beta = 1) \\ \frac{1}{2}(y - x)^2. & (\beta = 2) \end{cases} \tag{2.3}$$

In NMF, the multiplicative update rules for W and H have been derived to minimize each of the three divergences and without the need for constraints to enforce non-negativity. In order to reduce dimension, commonly, set to a small number, which results in NMF being a low-rank matrix approximation method. Therefore, the multiplicative update rules are derived as follows for the Euclidean distance (EUC),

$$W \leftarrow W \otimes \frac{VH^T}{WHH^T}. \tag{2.4}$$

$$H \leftarrow H \otimes \frac{W^T V}{W^T WH}. \tag{2.5}$$

Kullback-Leibler divergence (KL),

$$W \leftarrow W \otimes \frac{\frac{V}{WH}H^T}{1H^T}. \tag{2.6}$$

$$H \leftarrow H \otimes \frac{W^T\frac{V}{WH}}{W^T 1}. \tag{2.7}$$

and Itakura-Saito divergence (IS)

$$W \leftarrow W \otimes \frac{\frac{V}{(WH)^2}H^T}{\frac{1}{WH}H^T}. \tag{2.8}$$

$$H \leftarrow H \otimes \frac{W^T\frac{V}{(WH)^2}}{W^T\frac{1}{WH}}. \tag{2.9}$$

where the operator $\otimes$ denotes element-wise multiplication of two matrices (Hadamard product), $\frac{V}{WH}$ denotes element-wise division, $(WH)^2$ denotes element-wise exponentiation, and 1 denotes a matrix of ones of appropriate dimension.

Owing to the non-negative assumption can be suited for the non-negative values of music spectrogram and also be approximated as the combination of the non-negative audio source

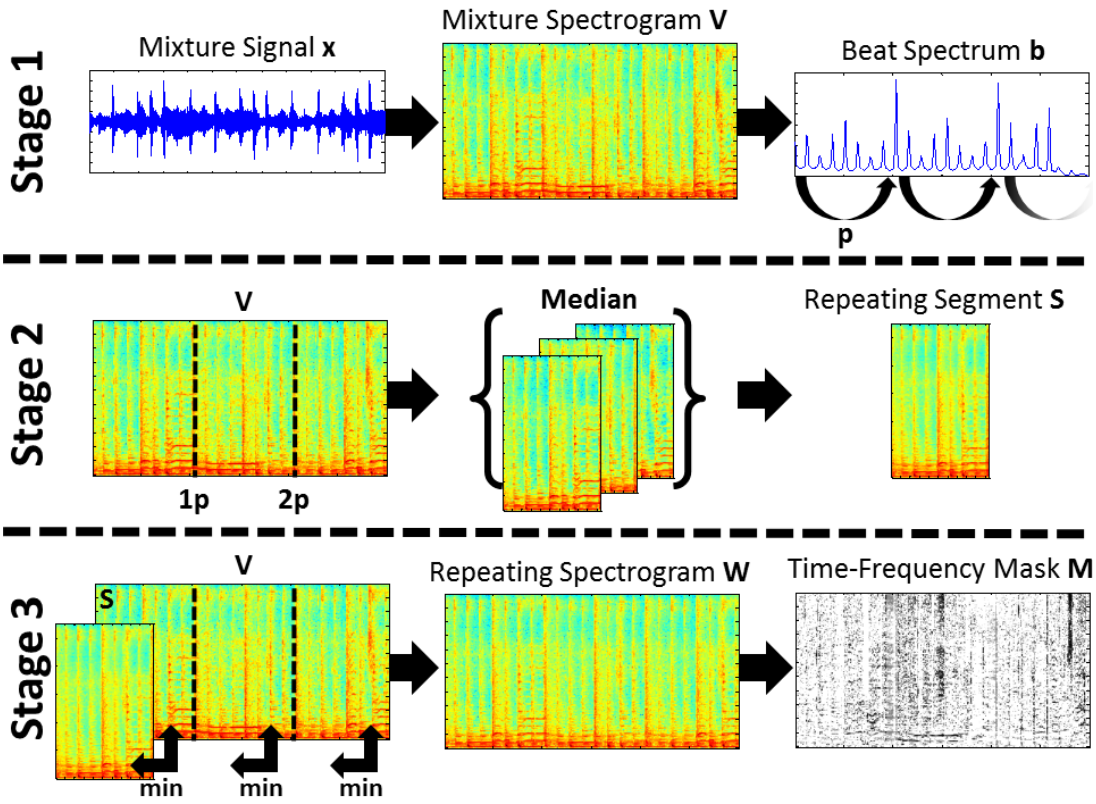


Figure 2.3: Overview of the REPET method for singing voice separation. Stage 1: calculate the beat of mixed music and then estimate the time length of repetition according to the calculation values on the beat. Stage 2: slice the mixture spectrogram and calculate the repeating accompaniment by taking median operation. Stage 3: extract the residual parts in the spectrogram that cannot be represented and separate it as singing voice part [26].

spectrogram, therefore, NMF can be also applied to singing voice separation. Although NMF has shown impressive results in monaural audio source separation, it is difficult to determine the appropriate number of nonnegative basis vectors.

2.1.2 REPET-based approach

REPET-based approaches are also popular for singing voice separation, which is according to the feature of background repetition characteristics [25] [26]. This methodology is based on the observation that the different individual sources tend to repeat over time, depending on its beat or speech.

Rafii et al. [25] [26] used the REPET algorithm for separating the repeating music part of the non-repeating singing voice in a musical mixture signal. The basic idea was to identify the periodically repeating segments in the mixture audio, then compared them to a repeating seg-

ment model derived from them, and finally extracted the repeating patterns via time-frequency masking.

Figure 2.3 illustrates the overview of REPET method for singing voice separation. In the first stage, calculate the beat of mixed music and then estimate the time length of repetition according to the calculation values on the beat. In the second stage, slice the mixture spectrogram and calculate the repeating accompaniment by taking median operation. In the third stage, extract the residual parts in the spectrogram that cannot be represented and separate it as singing voice part. Additionally, there are some methods that extend the original REPET method, including adaptive REPET by using the moving-median [27], or by using the similarity matrix [28].

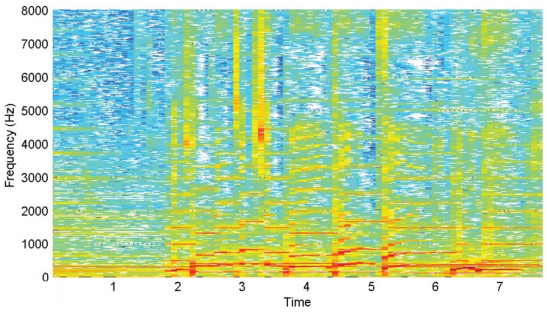
2.1.3 Robust principle component analysis

Candés et al. [29] proposed a convex RPCA model, 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 $S \in \mathbb{R}_{m\times n}$. The model can be defined as follows:

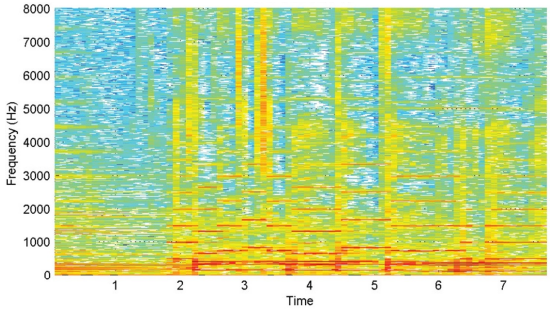
$$\begin{aligned} &\text{minimize } |L|_* + \lambda|S|_1, \\ &\text{subject to } X = L + S. \end{aligned} \tag{2.10}$$

where $|\cdot|_1$ is the L_1-norm, which is the sum of absolute values of matrix entries, $|\cdot|_*$ denotes the nuclear norm (sum of singular values), and $\lambda > 0$ is a positive constant parameter between the parts of sparsity matrix S and low-rank matrix L. Moreover, this convex model can be solved by accelerated proximal gradient (APG) or augmented Lagrange multiplier (ALM) [30]. According to the previous study [11], an inexact version of ALM (iALM) was used as a baseline for comparison in the dissertation.

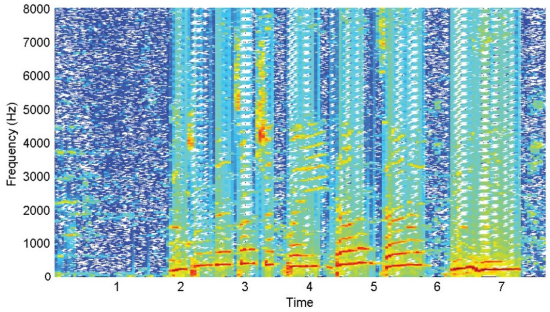
Huang et al. [11] proposed a method on RPCA for singing voice separation, which is an effective approach because the singing voice can be well modeled as a sparse matrix, while the accompaniment as well modeled as a low-rank matrix. RPCA has been extensively and successively applied in other signal processing applications like speech enhancement [31] [32] [33], SAR imaging [34] [35], direction of arrivals tracking [36] and also in computer vision applications [37] [38] [39]. 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 [40]



(a) Original Matrix



(b) Low-Rank Matrix



(c) Sparse Matrix

Figure 2.4: Example of system of singing voice separation by using RPCA [11]. (a) is the original matrix $\boldsymbol{X}$ (musical mixture), (b) is the separated low-rank matrix $\boldsymbol{L}$ (accompaniment), and (c) is the separated sparse matrix $\boldsymbol{S}$ (singing voice).

for singing voice separation. Figure 2.4 shows an example of the singing voice separation by using RPCA, the top spectrogram is the original matrix (musical mixture), the left bottom is separated low-rank matrix (accompaniment) and the right bottom is the separated sparse matrix (singing voice), respectively.

In a similar vein, Yang [41] proposed multiple low-rank representations (MLRR) to decompose a magnitude spectrogram into two low-rank matrices. Sprechmann et al. [42] proposed a real-time online singing voice separation by robust low-rank modeling. Fourer et al. [43] proposed a novel unsupervised singing voice detection method which uses single-channel blind source separation algorithm as a preliminary step. Chan et al. [44] proposed using informed group-sparse representation with the idea of pitch annotations separation. Pu et al. [45] proposed an approach in audio separation with the assistance of visual information. In addition, he [46] also proposed a non-linear generalization of RPCA, which uses two autoencoder network to realize the low-rank and sparse matrix decomposition. One autoencoder for low-rank part and the other one for the sparse part.

As stated above, RPCA is an effective algorithm to separate singing voice from the mixture

music signal, which can be well-described as the part of singing voice from the mixed music signal by the separated sparse matrix. And the mixture music signal can be described as the low-rank and sparse model. So, the process of RPCA decomposition is suited to the singing voice separation task. Additionally, many previous studies have shown that such decomposition is very effective in singing voice separation applications. 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.

2.1.4 Deep learning

Recently, deep learning has been received much attention for singing voice separation. Convolutional Network Network (CNN) architecture has been successful in audio source separation, especially in singing voice separation [47] [48] [49].

Chandna et al. [47] utilized the convolutional filters specifically designed for audio database and allowed a significant gain in processing time over a simple multi-layer perception, in the fully connected layer, dimensional reduction allows the model to learn a more compact representation of the input data from which the source can be separated. Takahashi et al. [48] extended DenseNet to tackle the music source separation with the proposed MDenseNet architecture.

In addition, he [49] proposed MMDenseLSTM framework for audio source separation, which is a variant of CNN architecture. It integrates long short-term memory (LSTM) in multiple scales with skip connection to efficiently model long-term structures within an audio context. There are also many new neural network based on extension of CNN architecture [50] [51] [52] [53] [54] [55] [56] [57], including the U-Net architecture and its variant [58] [59] [60] [61] [62] [63].

Deep neural network (DNN)-based models [7] [58] [64] [65] [66] [67] are perhaps the most widely used supervised learning models for singing voice separation. Figure 2.5 gives an example of musical source separation by using DNN architecture. Although they have proven effective for separating singing voice, a large number of training data are needed in advance,

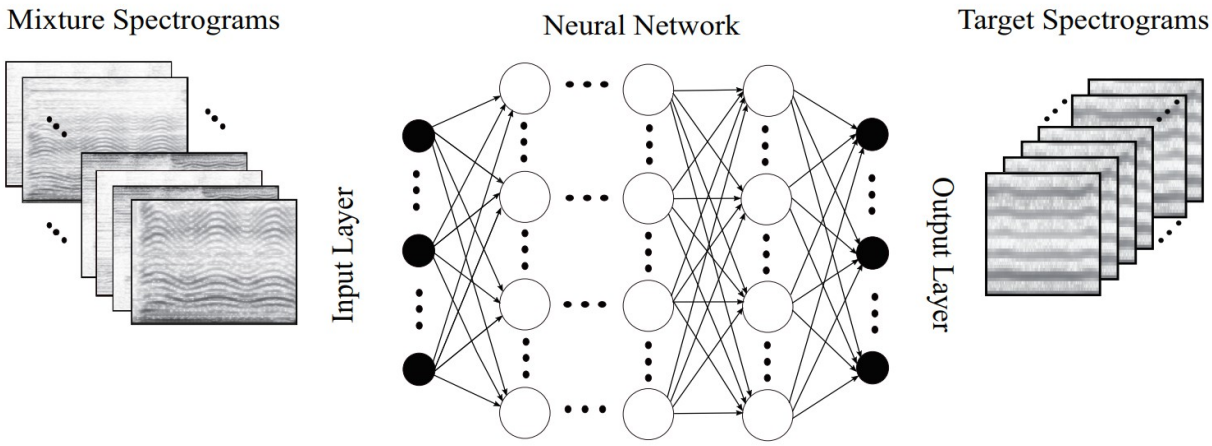


Figure 2.5: DNN architecture for musical source separation. The mixture magnitude spectrograms are set as inputs, and source magnitude spectrograms of the desired source S_j are set as the targets [68].

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 [69], separation quality decreases due to overfitting.

2.2 Experiment databases

In this dissertation, to evaluate the proposed optimization algorithm on the task of singing voice separation, a reasonable and feasible dataset is contributed to confirm the effectiveness of the proposed algorithm. In this chapter, several public databases are introduced that are commonly used in singing voice separation task. The general description of all the experiment databases are showed in Table 2.1.

2.2.1 MIR-1K dataset

MIR-1K dataset [70][1] contains 1000 Chinese pop songs recorded at 16 kHz sampling rate with 16 bit resolution. The duration of each song slip ranges from 4 to 13 seconds. The length of all datasets are 133 minutes. These song clips were extracted from 110 Chinese karaoke pop songs. The singing voice and the clean music accompaniment were recorded at the right and left channels, respectively. And the singing voice sung by amateur singers (8 females and 11

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

Table 2.1: All the experiment databases

Name	**Clips**	**Duration** (s)	**Sampling rate** (kHz)
MIR-1K	1000	4 ~ 13	16
ccMixter	50	77 ~ 456	44.1
DSD100	100	141 ~ 435	44.1
iKala	252	30	44.1

males), and the music accompaniment retrieved from the popular Chinese songs. This is the first dataset for singing voice separation task that released in public.

2.2.2 ccMixter dataset

ccMixter dataset [71][2] contains 50 full songs with duration rang from 1 minute 17 seconds to 7 minutes 36 seconds. Each audio music data contains the following three parts: singing voice, background music, and a mixture music, respectively. These audio dataset is extracted from 110 karaoke songs which contain a mixture track and a music accompaniment track. And the songs are freely selected from the 5000 Chinese popular music songs and sung by their lab-mates of 8 females and 11 males. Most of the singers are amateur and do not have professional music training.

2.2.3 DSD100 dataset

DSD100 dataset contains 100 full stereo songs of different audio data was recorded as the Demixing Secrets Dataset (DSD100). It ranges from 2 minutes 21 seconds to 7 minutes 15 seconds, as also used for the 2016 Signal Separation Evaluation Campaign (SiSEC) [9][3], which is split into 50 train (*Dev*) and 50 test ($Test$) songs. Each datum consists of bass, drums, other, and singing voice, respectively.

[2]https://members.loria.fr/ALiutkus/kam/
[3]http://liutkus.net/DSD100.zip

2.2.4 iKala dataset

iKala dataset [72][4] contains 252 song clips with duration 30 seconds. Each song clip in this database is recorded sampled with 44.1 kHz. And there is two channels in a wave file. The right channel is the ground truth singing voice, and the left channel is the ground truth background music. This dataset also contain the human-labeled fundamental frequency estimation of each audio data.

2.3 Evaluation metrics

In this dissertation, to evaluate the separation performance of the proposed method on audio source separation (e.g., singing voice separation), assess 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 metrics [73] [74][5]. The estimated signal $\hat{S}$(t) is defined as

$$\hat{S}(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t), \tag{2.11}$$

where $S_{target}(t)$ denotes the allowable deformation of the target sound, $S_{interf}(t)$ denotes the allowable deformation of the sources that account for the interferences of the undesired sources, and $S_{artif}(t)$ denotes the artifact term that may correspond to the artifact of the separation method.

The formulas for SDR, SIR, SAR, and NSDR are respectively defined as

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \left(S_{interf}(t) + S_{artif}(t)\right)^2}, \tag{2.12}$$

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t S_{interf}(t)^2}, \tag{2.13}$$

$$SAR = 10\log_{10}\frac{\sum_t (S_{target}(t) + e_{interf}(t))^2}{\sum_t e_{artif}(t)^2}, \tag{2.14}$$

[4] http://mac.citi.sinica.edu.tw/ikala/
[5] http://bass-db.gforge.inria.fr/bss_eval/

and

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \tag{2.15}$$

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, SAR, and NSDR mean that the corresponding separation algorithm exhibits better separation performance in terms of the 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 the metrics are expressed in dB.

In addition, report the global of SDR, SIR, SAR, and NSDR in the experiment. In other words, the separation results in GSDR, GSIR, GSAR, and GNSDR are performed, respectively. The equations are defined as follows

$$GSDR = \frac{\sum_{i=1}^{n} SDR_i}{n}, \tag{2.16}$$

$$GSIR = \frac{\sum_{i=1}^{n} GSIR_i}{n}, \tag{2.17}$$

$$GSAR = \frac{\sum_{i=1}^{n} NSAR_i}{n}, \tag{2.18}$$

and

$$GNSDR = \frac{\sum_{i=1}^{n} NSDR_i}{n}. \tag{2.19}$$

In a similar vein, the higher values of GSDR, GSIR, GSAR, and GNSDR represent better quality of the separation approach. Noticeable, especially in the GNSDR, which is the most important measure metric for the overall improve the separation performance in the singing voice separation task.

The subjective evaluation is also important for evaluating the separation quality. According the previous research on subjective evaluation, asking many different people to listen the separated singing voice and accompaniment from the mixture music signal, and than giving a reasonable scores according the evaluation standards. So, the related work will be done in the future work.

In a conclusion, in terms of human use of the separated results (separated singing voice), the subjective evaluation quality of singing voice separation becomes very important than the objective evaluation.

Chapter 3

WRPCA-based singing voice separation

In this chapter, an effective strategy is to deal with the problem of singing voice separation by using weighted robust principal component analysis (WRPCA). It constraints the value of separated low-rank matrix by utilizing the different weighted values. According to previous study [11], RPCA-based method is an effective strategy for singing voice separation because singing voice can be well modeled as a low-rank matrix. However, 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 over-estimates 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. Therefore, WRPCA can solve this problem by using the different weighted values to describe the separated low-rank matrix. So the separation quality can be improved due to the drums are described as low-rank matrix.

Therefore, in this chapter, there are two experiments to evaluate the proposed WRPCA method for singing voice separation. On the first experiment, we evaluate the proposed WRPCA for singing voice separation on the spectrogram with the ccMixter and DSD100 datasets. On the second experiment, we combine the proposed WRPCA method with gammatone auditory filterbank on the cochleagram for singing voice separation. To confirm the effectiveness of this proposed method, the experimental evaluation is also carried out on the ccMixter and DSD100 datasets.

3.1 WRPCA for singing voice separation

According to the previous studies, RPCA is an effective method to separate singing voice from the mixed music signal, which decomposes a given amplitude spectrogram (matrix) of a music signal into the sum of a low-rank matrix (music accompaniment) and a sparse matrix (singing voice). Since music accompaniment tends to have a similar phase, resulting in a spectrogram with the low-rank structure part. While singing voice varies significantly and continuously over time, resulting that a spectrogram has a sparse structure part. Although RPCA has been successfully applied to singing voice separation, it has a strong assumption. For example, drums may lie in the sparse subspace instead of being low-rank, which lead that the separation performance is decreased in many real world applications, especially for the drums existing in music signal. To copy with this problem, in this section, a weighted value method to make sure different scale values to describe sparse and low-rank matrices called WRPCA was proposed, which is choosing different weighted values between low-rank and sparse matrices.

3.1.1 Principal of WRPCA

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

$$\begin{aligned} &\text{minimize } |L|_{w,*} + \lambda |S|_1, \\ &\text{subject to } X = L + S, \end{aligned} \tag{3.1}$$

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. $\lambda = 1/\sqrt{\max(m,n)}$ was used as suggested by Candés et al. [29]. I also adopted an efficient inexact ALM [30] to solve this convex model. The corresponding augmented Lagrange function is defined as

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

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

$$\Sigma = \begin{pmatrix} diag(\delta_1(X), \delta_2(X), ..., \delta_n(X)) \\ \\ 0 \end{pmatrix}, \tag{3.2}$$

and $\delta_i(X)$ denotes the i-th singular value of X. If the positive regularization parameter C exists and the positive value $\varepsilon < min(\sqrt{C}, \frac{C}{\delta_1(X)})$, by using the reweighting formula $w_i^l = \frac{C}{\delta_i(L_l)+\varepsilon}$ [75], the weighted values will converge to

$$L^* = U\Sigma' V^T, \tag{3.3}$$

where

$$\Sigma' = \begin{pmatrix} diag(\delta_1(L^*), \delta_2(L^*), ..., \delta_n(L^*)) \\ \\ 0 \end{pmatrix}, \tag{3.4}$$

and

$$\delta_i(L^*) = \begin{cases} 0 \\ \dfrac{c_1 + \sqrt{c_2}}{2} \end{cases} \tag{3.5}$$

where $c_1 = (\delta_i(X) - \varepsilon)$ and $c_2 = ((\delta_i(X) + \varepsilon)^2 - 4C)$ [76]. In this study, set the regularization parameter C as the maximum matrix size, which enables us to obtain the best separation performance results on the audio data, e.g., $C = max(m, n)$ [77].

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, finally, obtain a low-rank matrix L (accompaniment) and a sparse

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

matrix S (singing voice). Therefore, WRPCA method decomposed 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 a singular value decomposition (SVD) at each iteration, which in turns leads to slow running time.

3.1.2 Experimental evaluation

In this section, the proposed WRPCA is evaluated on two different databases.

Experiment settings

One is the ccMixter dataset, to reduce the computations in the experiment, 30 seconds clip (from 0'30" to 1'00") was used at the same time of each song, which is the maximum period of all songs containing singing voice, but there are still exist 2 songs with no singing voice during this period, adopt to another period (from 1'30" to 2'00") in this 2 songs.

The other is DSD100 dataset. I also use only 30 seconds clip (from 1'45" to 2'15"), which

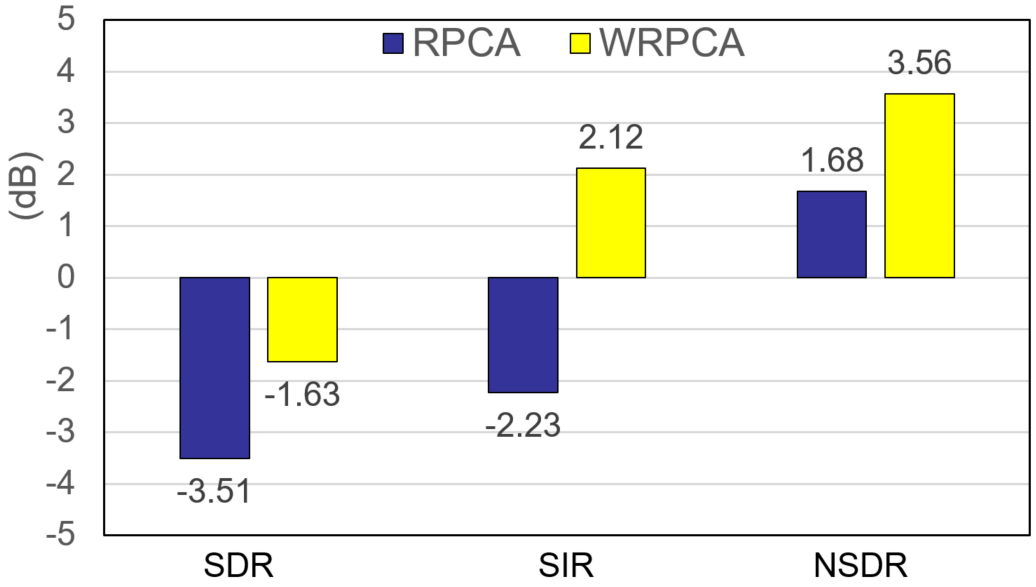


Figure 3.1: Comparison of singing voice separation results using RPCA and the proposed WRPCA on the ccMixter dataset. Note that SDR for the original dataset is -5.19 dB.

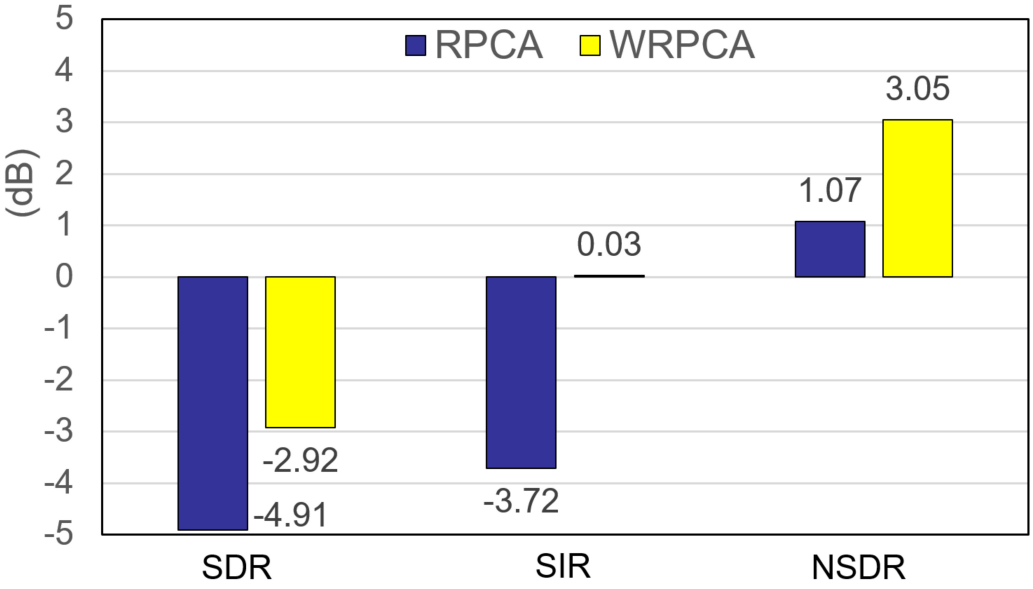
is the only period where all 100 full stereo songs contain singing voice. Because there are 4 sources (bass, drums, vocals and others) for each track, considering the sum of bass, drums and others as music accompaniment part.

In this chapter, the experiment mainly focuses on monaural source separation. It is even more difficult than multichannel source separation since only one single channel information was available. The two-channel stereo mixtures were downmixted into a single channel and obtained an average value of each channel. All experimental data are sampled at 44.1 kHz. The input feature is calculated using short-time Fourier transform (STFT) and inverse short-time Fourier transform (ISTFT). A window size of 1024 samples and a hop size of 256 samples for the STFT. And FFT size is 1024.

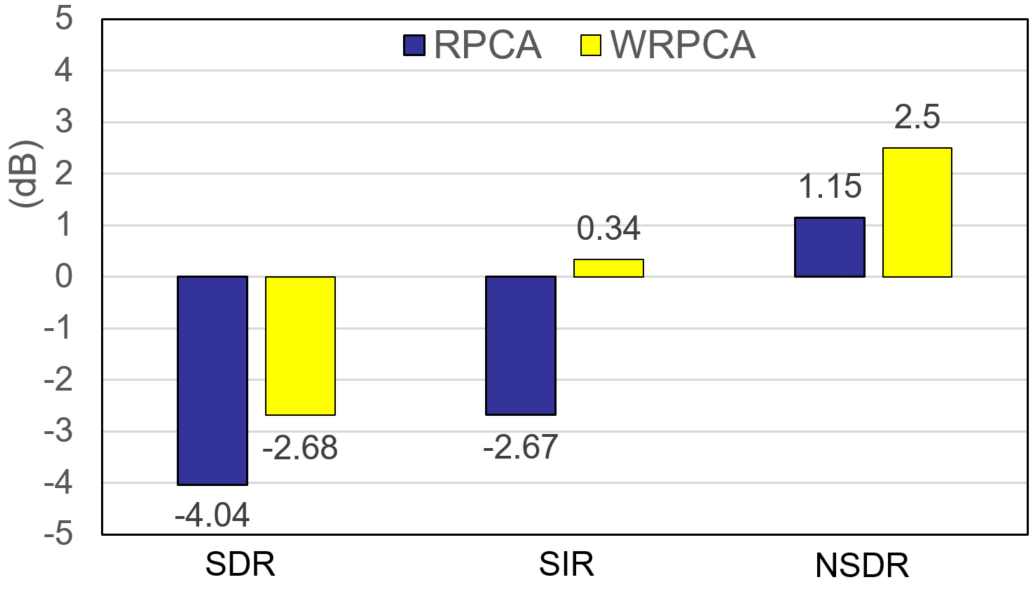
To confirm the effectiveness of our proposed method, the quality of separation is assessed in terms of SDR, SIR, and NSDR using the BSS-EVAL evaluation toolbox 3.0 metrics. All the metrics are expressed in dB.

3.1.3 Result and conclusion

Figure 3.1 shows the experiment results of SDR, SIR and NSDR between WRPCA and RPCA on the ccMixter dataset. The experiment results show that the proposed method gets better results on the ccMixter dataset.



(a)



(b)

Figure 3.2: Comparison of singing voice separation results using conventional RPCA and the proposed WRPCA on the DSD100 dataset. (a) is the set of DSD100/***dev*** data; (b) is the set of DSD100/***test*** data. Note that SDRs for the original datasets, ***dev*** and ***test***, are -5.98 dB and -5.18 dB, respectively.

In addition, the experiment results are compared with the conventional RPCA on the DSD100 dataset. Figure 3.2(a) is the separation results of SDR, SIR and NSDR on ***dev*** data (top); Figure 3.2(b) is the separation results of SDR, SIR and NSDR on ***test*** data (bottom). The above two figures show that the proposed WRPCA method also yields promising experimental results than the conventional RPCA method on the DSD100 dataset.

In this chapter, an extension of RPCA with different weighted values for singing voice separation was proposed. The experimental results on the ccMixter and DSD100 datasets show clearly that the proposed method outperforms the conventional RPCA for the singing voice separation on the two databases.

3.2 WRPCA with gammatone auditory filterbank for singing voice separation

Even if the previous proposed WRPCA method can obtain acceptable separation results from mixture music signals, they ignore the features of the human auditory system, which plays a vital role in improving the quality of separation results. Recently a study was published hinting that cochleagram, as an alternative time-frequency (T-F) analysis based on gammatone filterbank, is more suitable than spectrogram for source separation [78]. This is because, cochleagram is derived from non-uniform T-F transform whereas T-F units in low-frequency regions have higher resolutions than in the high frequency regions, which closely resembles the functions of the human ear. Similarly, singing voice performances are quite different from music accompaniment on cochleagram. The spectral energy centralizes in a few T-F units for singing voice and thus can be assumed to be sparse. On the other hand, music accompaniment on the cochleagram has similar spectral patterns and structures that can be captured by a few basis vectors, so it can be hypothesized as a low-rank subspace. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of spectrogram.

To overcome the above-mentioned problems and imitate the human auditory system, adopt gammatone auditory filterbank as the first stage of WRPCA in cochleagram processing. Finally, apply ideal binary mask (IBM) or ideal ratio mask (IRM) [79] to enforce the constraints between an input mixture signal and the output results.

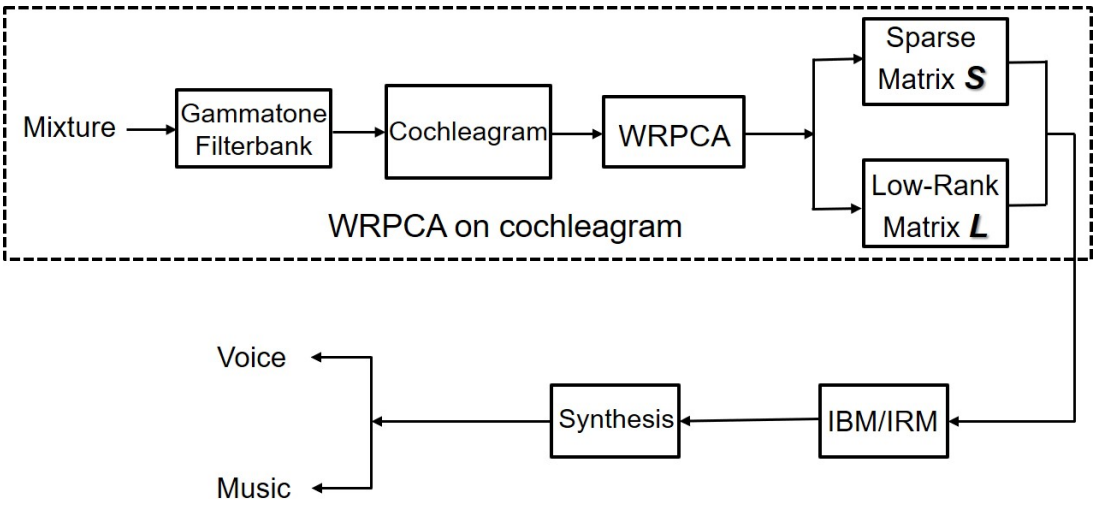


Figure 3.3: Block diagram of the proposed singing voice separation

3.2.1 Application to mask estimation

After obtaining the separation results of sparse S and low-rank matrices L by using WRPCA, applied IBM and IRM estimations to further improve the separation performance. A block diagram of the singing voice separation system is illustrated in Figure 3.3. It consists of two stages: WRPCA on cochleagram and singing voice separation based on IBM and IRM estimations. The first stage performs the cochlear analysis with gammatone filter, calculates the cochleagram of the mixture music signal, and then decomposes matrixes into sparse and low-rank matrices by using WRPCA. The second stage applies IBM/IRM estimation to improve the separation results. The IBM and IRM are defined as [79]

$$M_{ibm} = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \tag{3.6}$$

and

$$M_{irm} = \frac{S_{ij}}{S_{ij} + L_{ij}} \tag{3.7}$$

where M_{ibm} and M_{irm} are the values of IBM estimation and IRM estimation, respectively. S_{ij} and L_{ij} are the values of the sparse and low-rank matrices. The separated matrices can be synthesized as described by Wang *et al*. [69].

3.2.2 Experimental evaluation

In this section, we introduce how evaluated WRPCA by using two different databases: ccMixter and DSD100 datasets, and how compared it with the conventional RPCA.

Experiment settings

In this chapter, to evaluate WRPCA, two different databases are used in this experiment. The first was the ccMixter dataset, for which I chose 43 full stereo songs with only 30 seconds clip (from 0'30" to 1'00") at the same time of each song, which is the maximum period of all songs containing singing voice. Each audio contains three parts: singing voice, music accompaniment, and a mixture of them.

The second was the DSD100 dataset. To reduce computations, 30 seconds clip were adopted (from 1'45" to 2'15") at the same time for all audio data, which comprised 36 development songs and 46 test songs. Each track consists of four sources, for example, bass, drums, vocals and others. In the experiment, two-channel stereo mixtures were downmixted into a single mono channel and obtained an average value for each channel. All experiment data were sampled at 44.1 kHz. Setting parameters for cochleagram analysis: 128 channels, 40~11025 Hz frequency range, and 256 frequency length. To compare the results with those obtained with WRPCA, calculated the input feature by using STFT and ISTFT, which is a part of contrast experiments that have been performed on spectrogram for conventional RPCA and WRPCA. The window size of 1024 samples was used, a hop size of 256 samples for the STFT and an FFT size of 1024.

3.2.3 Result and conclusion

To evaluate WRPCA algorithm, the first experiment was evaluated on the ccMixter dataset. Figure 3.4 and 3.5 indicate the comparison results of conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. The methods of RPCA, RPCA with IRM, RPCA with IBM and WRPCA) are calculated on spectrogram (without gammatone filterbank), while WRPCA with IRM and WRPCA with IBM are calculated on cochleagram (with gammatone filterbank). The experiment results obtained with the SDR and SAR show that WRPCA gets better results on the ccMixter dataset, especially for the IBM estimation (with gammatone filterbank). In contrast, the conventional RPCA got

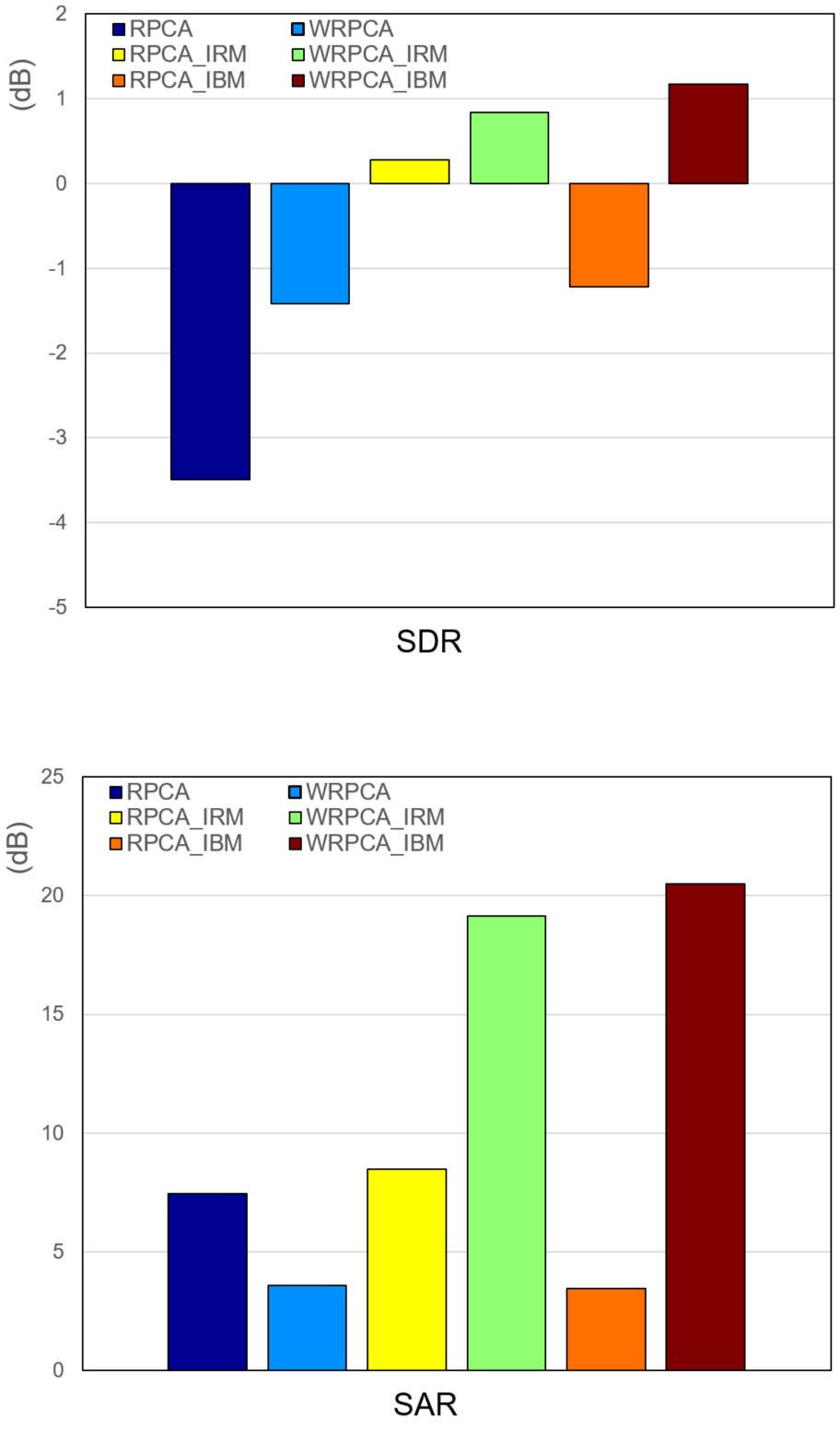


Figure 3.4: Comparison of singing voice separation results on the **ccMixter** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively.

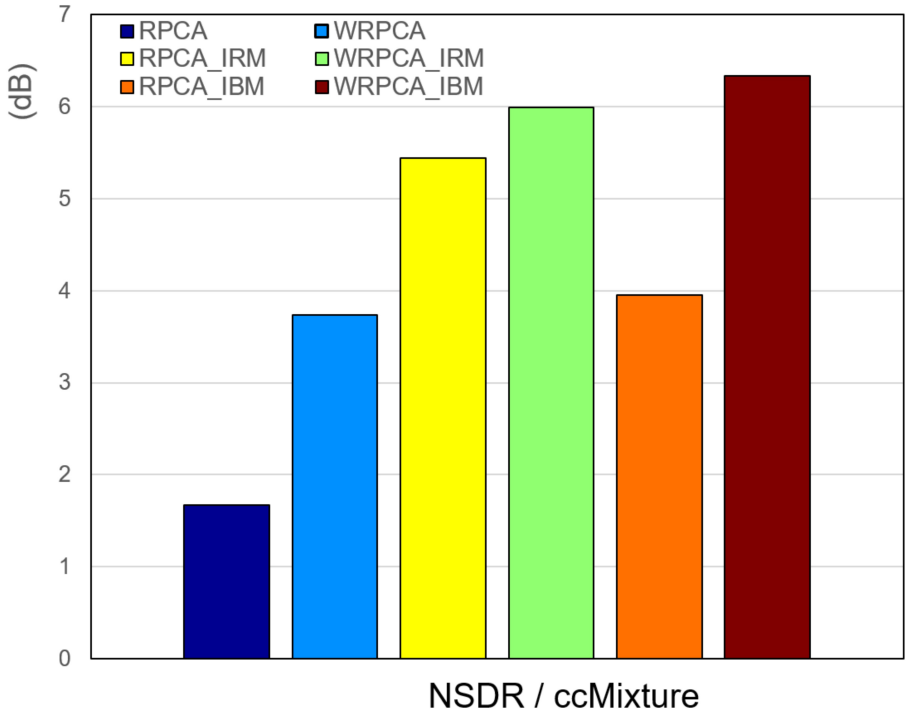


Figure 3.5: Comparison of singing voice separation results on the **ccMixter** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. Note that SDR for the original datasets, ccMixter is -5.16 dB.

worse results than the others. The second experiment was evaluated on the DSD100 dataset. Figure 3.6 and 3.7 show the comparison results obtained with the conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. The results clearly show that WRPCA obtains better separation results on the DSD100 dataset, especially for the IBM estimation (with gammatone filterbank). However, the opposite results were obtained with the conventional RPCA. In terms of the SAR in the experiment, WRPCA with IBM on cochleagram (with gammatone filterbank) attained higher values than others, while the RPCA with IBM (without gammatone filterbank) had the worst values among them.

The NSDR provides overall improvement in the SDR; in other words, it provides better separation performance in singing voice separation. Figure 3.5 and 3.7 show the NSDR results with WRPCA on the ccMixter and DSD100 datasets. The results indicate that the best performance was achieved by WRPCA with IBM (with gammatone filterbank).

According to the results of Figures 3.4, 3.5 and 3.6, the proposed WRPCA on cochleagram provides better separation performance than RPCA on spectrogram with or without IBM

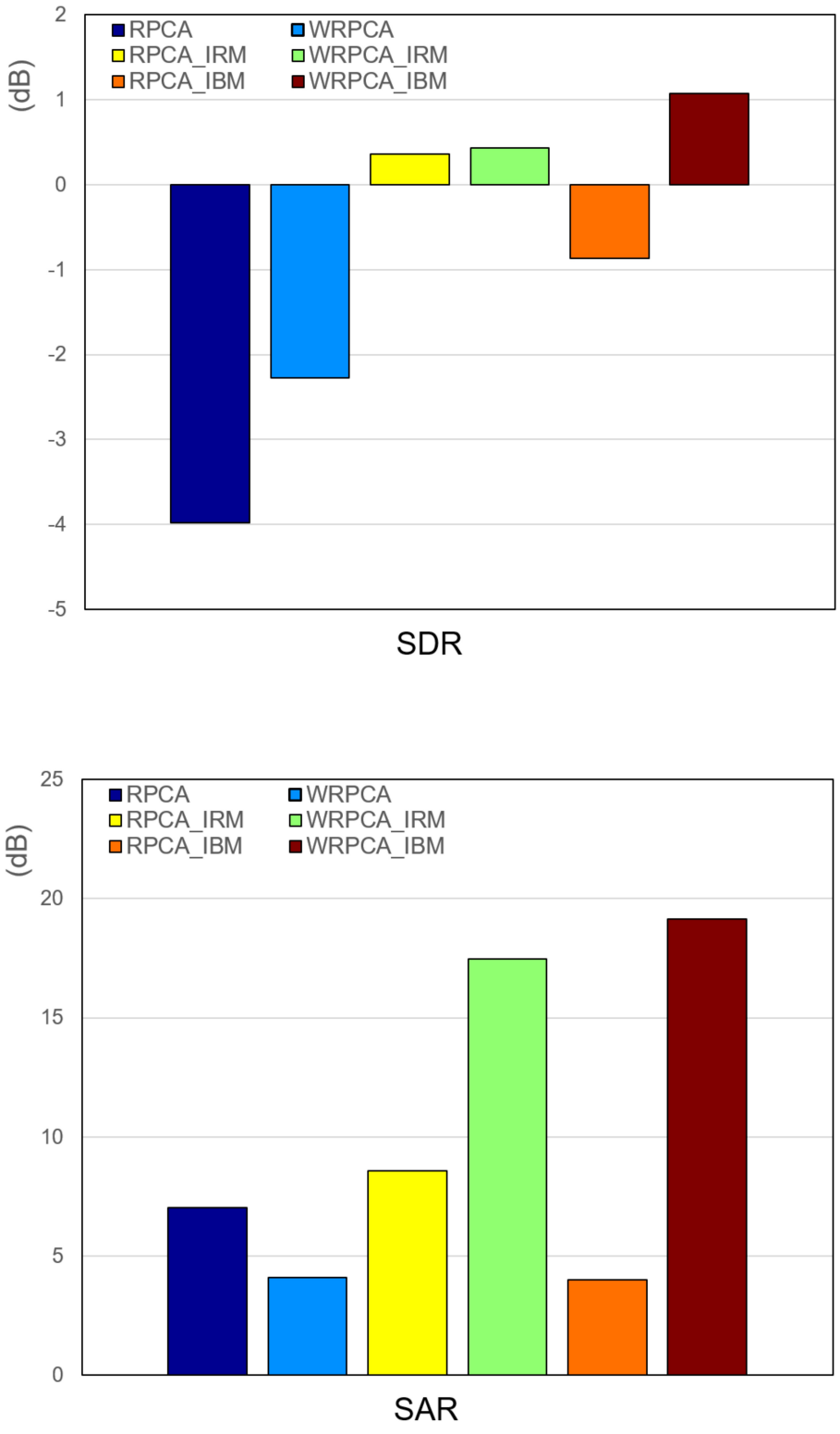


Figure 3.6: Comparison of singing voice separation results on the **DSD100** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively.

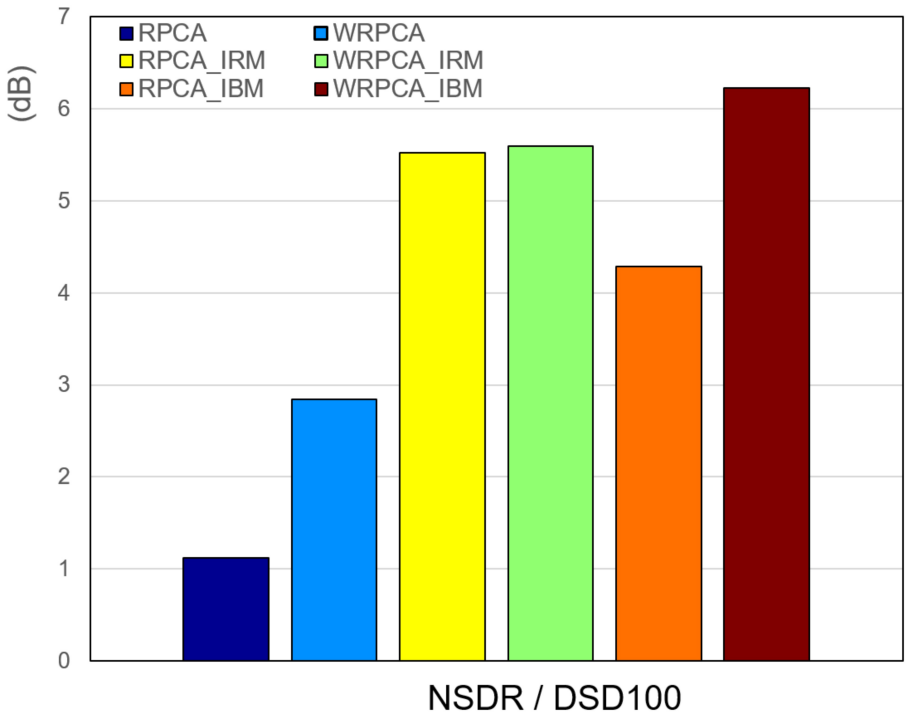


Figure 3.7: Comparison of singing voice separation results on the **DSD100** dataset among conventional RPCA, RPCA with IRM, RPCA with IBM, WRPCA, WRPCA with IRM, and WRPCA with IBM, respectively. Note that SDR for the original datasets, DSD100 is -5.11 dB.

or IRM. Moreover, WRPCA provided better results than RPCA without gammatone filterbank and IBM or IRM estimations. Additionally, WRPCA on cochleagram with IBM (with gammatone filterbank) provides better separation results in all evaluation standard methods. However, RPCA with IBM does not provide values as good as those provided by RPCA with IRM.

In this work, an extension of RPCA with weighting on cochleagram (WRPCA). It is based on gammatone auditory filterbank and application to IBM/IRM estimation for singing voice separation. The cochleagram of the mixture signal was decomposed into sparse (singing voice) and low-rank matrices (music accompaniment) by using WRPCA, then IBM/IRM estimation was utilized to improve the separation results. Experimental results obtained on the ccMixter and DSD100 datasets confirmed that WRPCA outperforms the conventional RPCA method in singing voice separation tasks, especially for WRPCA on cochleagram based on gammatone auditory filterbank with IBM estimation.

3.3 Discussion and summary

This chapter proposes a different weighted value in the low-rank and sparse decomposition model for singing voice separation. In addition, utilize the proposed WRPCA algorithm with gammatone filterbank on cochleagram instead of spectrogram for singing voice separation. The experiment results show that the cochleagram is better than spectrogram, Owing to the cochleagram is derived from non-uniform T-F transform whereas T-F units in low-frequency regions have higher resolutions than the high frequency regions, which closely resembles the functions of the human ear. In addition, by utilizing different weighted values to describe the separated low-rank matrix is better to the conventional RPCA. This is because the drums can be described by the separated sparse matrix.

Chapter 4

CRPCA-based singing voice separation

In this chapter, the main task is studied on another extension of RPCA, which uses a rank-1 constraint robust principal component analysis called Constraint RPCA (CRPCA) and its application to singing voice separation.

Firstly, describing CRPCA for singing voice separation, which uses rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm, which not only provides a robust solution to large dynamic range differences among instruments but also reduces the computation complexity.

Secondly, utilizing the proposed CRPCA on cochleagram based on gammatone auditory filterbank for singing voice separation.

Thirdly, combining F0 and non-negative constraint RPCA for singing voice separation. In addition, to minimize the reconstruction error when synthesizing the singing voice, using the original phase recovery in estimating the spectral components of the separated singing voice of the musical mixture.

Fourthly, incorporating CRPCA with vocal activity detection for singing voice separation. The proposed CRPCA method utilizes rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm. 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.

4.1 CRPCA for singing voice separation

As mentioned the above chapter, RPCA can be used as an efficient strategy to separate singing voice in a mixture music signal, which decomposes the given amplitude spectrogram (matrix) of a mixture music signal into the sum of a sparse matrix (singing voice) and a low-rank matrix (music accompaniment). Owing to the part of background music can reproduce the same signal in the same song, the magnitude spectrogram therefore can be considered as a part of low-rank matrix. Singing voice, on the other hand, varies significantly and has a sparse distribution since its harmonic structure part in the spectrogram domain, resulting in a spectrogram with a sparse matrix part. Although RPCA has been successfully applied to singing voice separation, it ignores the different characteristic values of SVD and computational complexity to minimize the nuclear norm for separating singing voice. Thus the separation performance decreases due to drums may lie in the sparse subspace instead of being low-rank. In another work, WRPCA [77] [80], which chose the different weighted values to describe the low-rank matrix for singing voice separation. However, it suffers from high computational cost, as it requires an SVD at each iteration. Hence the running time of WRPCA method is slower than the conventional RPCA method. Recently a partial sum minimization of singular values instead of minimizing the nuclear norm in RPCA [81] was published, which used the minimized rank way to solve the different values of SVD in image processing, especially for background subtraction under the condition of rank-1 constraint.

To solve these problems, the partial sum minimization of SVD and computation complexity are significant for separating singing voice from the mixture music signal. In this chapter, combining the idea in [81] and propose a novel extension of RPCA exploiting rank-1 constraint, which utilizes the rank-1 constraint minimization singular values in RPCA instead of minimizing the nuclear norm for singing voice separation. Owing to rank-1 constraint in the background music, which is very similar to background subtraction, as the background music has a larger variation in richness than singing voice among different songs. In addition, rank-1 constraint can utilize a prior target rank to separate background music and singing voice from the mixture music signal, which leads to a reduction in computation complexity. Therefore, the proposed CRPCA can not only describe the different values of SVD under the rank-1 constraint information, but also the computation complexity is reduced. Finally, apply time-frequency masking to further improve the separation results.

In previous studies [77], [80], WRPCA was used to separate singing voice from mixture music signal. On account of the part of background music can reproduce the same signal in the same song, although it has different values, the magnitude spectrogram can still be considered as a part of low-rank matrix. Singing voice signal, on the other hand, varies significantly and has a sparse distribution since its harmonic structure part in the spectrogram domain, resulting in a spectrogram with a sparse matrix part. So utilizing WRPCA method to decompose an input matrix structure part into a low-rank matrix part and a sparse matrix part. The separated results outperform RPCA in the different audio data. However, it suffers from high computational cost during computing an SVD at each iteration, which leads to the running time to be slow. To overcome the disadvantages of both RPCA and WRPCA, a partial sum minimization of singular value based on rank-1 constraint called CRPCA was proposed. The aim is to fully utilize a prior rank-1 constraint to minimize the partial sum of singular values in RPCA.

4.1.1 Principal of CRPCA

CRPCA is a novel extension of RPCA, which exploiting rank-1 constraint for singing voice separation. The model is defined as follows:

$$\begin{aligned} &\text{minimize} \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda|S|_1, \\ &\text{subject to } X = L + S. \end{aligned} \tag{4.1}$$

where L is the value of low-rank matrix, S is the value of sparse matrix. $X \in \mathbb{R}_{m\times n}$ is the value of an input matrix, which consists of $L \in \mathbb{R}_{m\times n}$ and $S \in \mathbb{R}_{m\times n}$, and $\lambda > 0$ is a positive constant parameter between the sparse matrix S and the low-rank matrix L. And $\delta_i(L)$ is the i-th singular value of L. $\lambda = 1/\sqrt{\max(m,n)}$ was used as suggested in [29]. I also used an efficient iALM [30] method to solve this convex model in this work. The augmented Lagrangian function can be defined as follows:

$$J(X,L,S,\mu) = \min \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda|S|_1 + \langle J, X-L-S\rangle + \frac{\mu}{2}|X-L-S|_F^2.$$

where J is the Lagrange multiplier and μ is a positive scalar. The process of separating singing voice from the mixture music signal can be seen in **Algorithm 2** CRPCA for singing voice

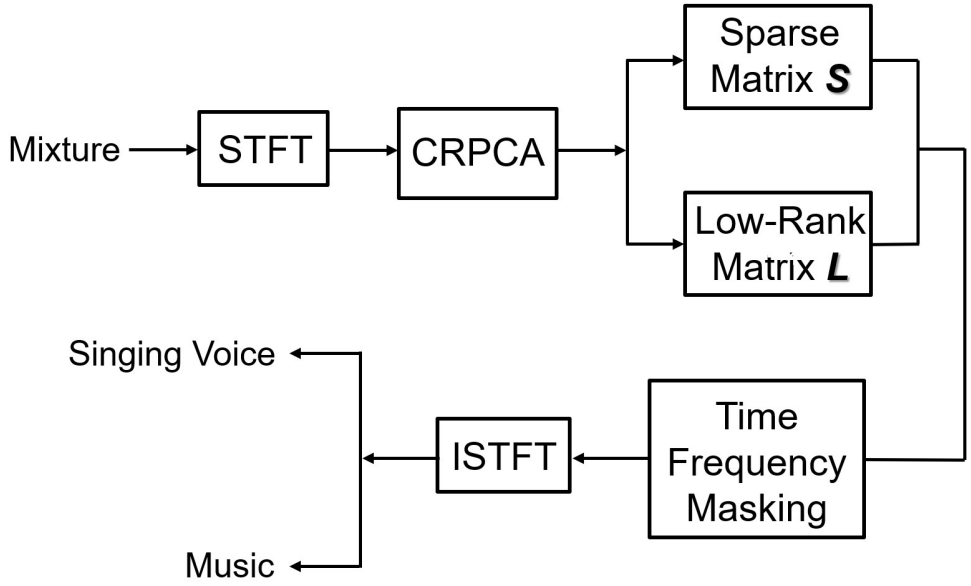


Figure 4.1: Block diagram of the proposed singing voice separation system.

separation. The value of X is a mixture music signal from the observed audio data. After separated by using CRPCA, finally, obtain a low-rank matrix L (music accompaniment) and a sparse matrix S (singing voice).

From the augmented Lagrangian function, the following two sub-problems about L and S are solved as follows

$$L_{k+1} = \min_{L} \sum_{i=2}^{\min(m,n)} \delta_i(L) + \langle J_k, X - L - S_k \rangle + \frac{\mu_k}{2}|X - L - S_k|_F^2. \tag{4.2}$$

$$S_{k+1} = \min_{S} \lambda|S|_1 + \langle J_k, X - L_k - S \rangle + \frac{\mu_k}{2}|X - L_k - S|_F^2. \tag{4.3}$$

As suggested by Oh et al. [81], the update rules of L and S are equivalent to solve the above two sub-problems as the following two equations:

$$L_{k+1} = P_{1,\mu_k^{-1}}(X - S_k + \mu_k^{-1} J_k) \tag{4.4}$$

$$S_{k+1} = Q_{\lambda\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1} J_k) \tag{4.5}$$

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

$$P_{1,\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V_Y^T \tag{4.6}$$

where $Y = Y_1 + Y_2$ $(Y \in \mathbb{R}_{m\times n})$, $D_{Y_1} = diag(\delta_1, 0, ..., 0)$, $Q_{\mu_k^{-1}}(D_{Y_2}) = sign(D_{Y_2}) \cdot max(|D_{Y_2}| - \mu_k^{-1}, 0)$

Algorithm 2 CRPCA for singing voice separation

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

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

2: While not converge,

3: **do** :

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

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

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

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

8: $k = k + 1$.

9: **end while**.

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

is the soft-thresholding operator [82], $D_{Y_2} = diag(0, \delta_2, ..., \delta_{\min(m,n)})$, δ_1 and δ_2 are the first and second singular values. In order to improve the separation performance, after separated by using CRPCA, adopt ideal binary time frequency masking (IBM) estimation to further improve the separation results. The masking M_{ibm} is defined as follows:

$$M_{ibm} = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \tag{4.7}$$

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

A block diagram of the proposed singing voice separation system can be seen in Figure 4.1. For each mixture audio in the test dataset, apply a STFT and ISTFT based on being separated by using CRPCA. Furthermore, utilize IBM method to further improve the separation results. And finally, obtain low-rank matrix L (music accompaniment) and sparse matrix S (singing voice), respectively.

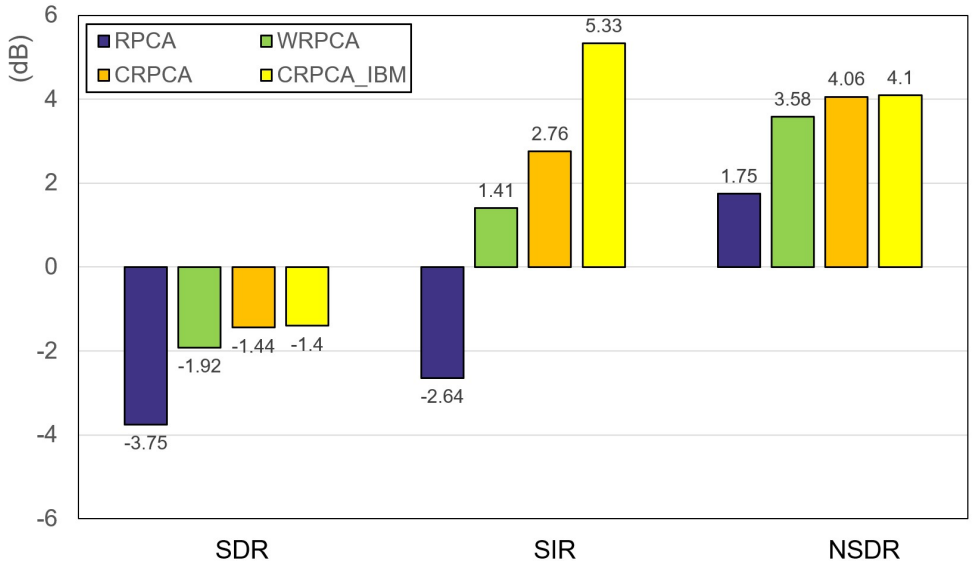


Figure 4.2: Comparison of singing voice separation results on the **ccMixter** dataset among RPCA, WRPCA, CRPCA and CRPCA with IBM on SDR, SIR, and NSDR, respectively.

4.1.2 Experimental evaluation

In this chapter, the proposed CRPCA and CRPCA with IBM are evaluated by using two different datasets: ccMixter and DSD100 datasets, respectively. And compare it with RPCA and WRPCA methods.

Experiment settings

In the experiments, to evaluate the performance of the proposed CRPCA method, two different datasets are used to compare with RPCA, WRPCA, CRPCA and CRPCA with IBM methods. The first one was the ccMixter dataset, which contains 50 full songs with durations ranging from 1'17" to 7'36". Each audio data contains three parts: music accompaniment, singing voice, and a mixture of them, respectively. The other one was the DSD100 dataset. I considered the sum of drums, bass and other as music accompaniment part. The target was to separate singing voice from the music accompaniment in the mixture music signal.

In the experiment, evaluated the proposed CRPCA and CRPCA with IBM methods mainly concentrate on monaural source separation tasks. It was even more difficult than multichannel source separation due to only one single channel was available. The two-channel stereo mixture datasets were downmixed into a single channel and obtained an average value of each channel.

All data were sampled at 44.1 kHz. The input feature was calculated using STFT and ISTFT.

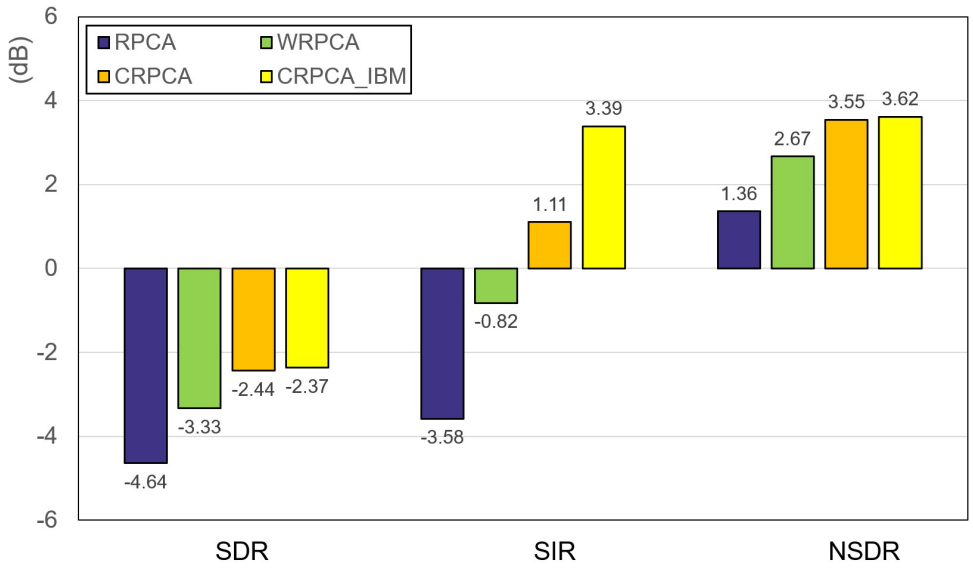


Figure 4.3: Comparison of singing voice separation results on the **DSD100** dataset among RPCA, WRPCA, CRPCA and CRPCA with IBM on SDR, SIR, and NSDR, respectively.

A window size of 1024 samples and a hop size of 256 samples. All the experiments were run by using MATLAB R2015a, on a PC win10, X64-based processor, RAM32GB with i7-6700K CPU@4.00 GHz.

4.1.3 Result and conclusion

To confirm CRPCA and CRPCA with IBM methods, the experiment is evaluated the proposed method on the ccMixter dataset. Figure 4.2 shows the comparison of singing voice separation results on RPCA, WRPCA, CRPCA and CRPCA with IBM (CRPCA_IBM). The experiment results are obtained with SDR, SIR and NSDR show that CRPCA obtains better separation results than RPCA and WRPCA, especially for using IBM estimation on the ccMixter dataset. With regard to SIR, CRPCA with IBM has a significant improved result among them. The degree of separation result values of SIR, RPCA and CRPCA with IBM, are -2.64 dB and 5.33 dB, respectively.

CRPCA and CRPCA with IBM methods were evaluated on the DSD100 dataset. Figure 4.3 shows the comparison results with RPCA, WRPCA, CRPCA and CRPCA with IBM. The experiment results clearly reveal that CRPCA with IBM also obtains better separation performance on the DSD100 dataset. These two result figures indicate that rank-1 constraint minimization can improve the separation performance than minimizing the nuclear norm in RPCA,

Table 4.1: 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**

even exceed the previous proposed WRPCA method. In addition, after using IBM estimation by CRPCA, especially with regard to SIR values among them. RPCA and CRPCA with IBM, are -3.58 dB and 3.39 dB, respectively.

As the above results shown, although WRPCA can get better separation results, the running time is longer than RPCA on two datasets. Owing to CRPCA can utilize a prior target rank to separate singing voice from the mixture music signal, no matter separation performance or running time, the rank-1 constraint minimization singular values in RPCA is better than the nuclear norm for separating singing voice. Furthermore, applied IBM method to improve the separation performance. In terms of running time, CRPCA is preferable under the same conditions on two datasets.

Compared the running time of the proposed method with those of the previous methods of the above-mentioned two datasets. Table 4.1 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.

In this work, a novel unsupervised approach that extends RPCA exploiting rank-1 constraint for singing voice separation task was proposed. The experiment evaluation results on the ccMixter and DSD100 datasets indicated that CRPCA outperforms the conventional RPCA and WRPCA, especially for using time frequency masking. In addition, with regard to the running time, CRPCA is shorter than others under the same conditions while WRPCA is the worst among them.

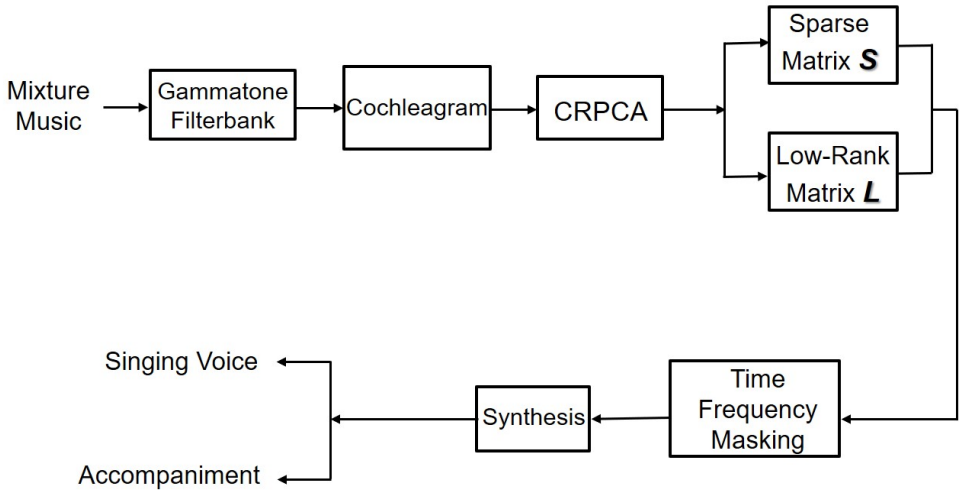


Figure 4.4: Block diagram of the proposed singing voice separation system.

4.2 CRPCA with gammatone auditory filterbank for singing voice separation

According to the previous studies, inspired by a sparse and low-rank model, the proposed an effective extension of RPCA with rank-1 constraint (CRPCA) [83]. Although it can get better separation results than RPCA in singing voice separation task, there is still exists a lot of room for improvement. Recently a study was published hinting that cochleagram, as an alternative time-frequency analysis based on gammatone filterbank, is more suitable than spectrogram for source separation [78] [80]. This is because, cochleagram is derived from non-uniform time-frequency transform whereas time-frequency units in low-frequency regions have higher resolutions than the high-frequency regions, which closely resembles the functions of the human ear. Similarly, singing voice performances are quite different from music accompaniment on cochleagram. The spectral energy centralizes in a few time-frequency units for singing voice and thus can be assumed to be sparse. On the other hand, music accompaniment on the cochleagram has similar spectral patterns and structures that can be captured by a few basis vectors, so it can be hypothesized as a low-rank subspace. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of the spectrogram.

To improve the separation performance, combine gammatone auditory filterbank with cochleagram by using CRPCA algorithm. In addition, further apply time-frequency masking estimation [79] to enforce the constraints between an input mixture music signal and the output results [84].

4.2.1 Gammatone filterbank and cochleagram

The Gammatone filterbank [85] is a cochlear filtering representation which decomposes an input signal into the time-frequency domain using a lot of gammatone filters. The impulse response of a gammatone filter centered at frequency w is obtained as follows:

$$g(w,t) = \begin{cases} t^{h-1}e^{-2\pi vt}cos(2\pi wt),\ t > 0 \\ 0,\ others \end{cases} \tag{4.8}$$

where h represents the order of filter, v stands for the rectangular bandwidth which increases as the center frequency w increases. The filter output response $r(c,t)$ can be expressed as follows:

$$r(c,t) = x(t) * g(w_c,t) \tag{4.9}$$

where '*' indicates the convolution in time domain, c is a particular filter channel and the center frequency is w_c. So this function can be shifted backwards by using (h-1)/($2\pi v$) to compensate for the filter delay. The output of each filter channel is cut into time-frequency with half of overlap between the consecutive frames. And finally, the time-frequency spectra of all the filter outputs are constructed to form the cochleagram.

4.2.2 CRPCA using time-frequency masking

After separated by using CRPCA, in order to improve the separation performance, apply binary time-frequency masking estimation to further improve the separation results. The masking b_m is defined as follows:

$$b_m = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \tag{4.10}$$

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

A block diagram of the proposed unsupervised singing voice separation system can be illustrated in Figure 4.4. For each mixture music audio in the test dataset, calculate the cochleagram of the mixture music audio under the condition of gammatone filterbank, after that decompose the matrix into low-rank matrix L (music accompaniment) and sparse matrix S (singing voice)

by using CRPCA method, and then, deal with the separated sparse and low-rank matrices by using time-frequency masking. Finally, the separated matrices can be synthesized as described in [69].

4.2.3 Experimental evaluation

In the experiments, firstly, evaluated the proposed method on the MIR-1K dataset, which contains 1000 song clips with durations ranging from 4 to 13 seconds. The data were extracted from 110 Chinese karaoke pop songs.

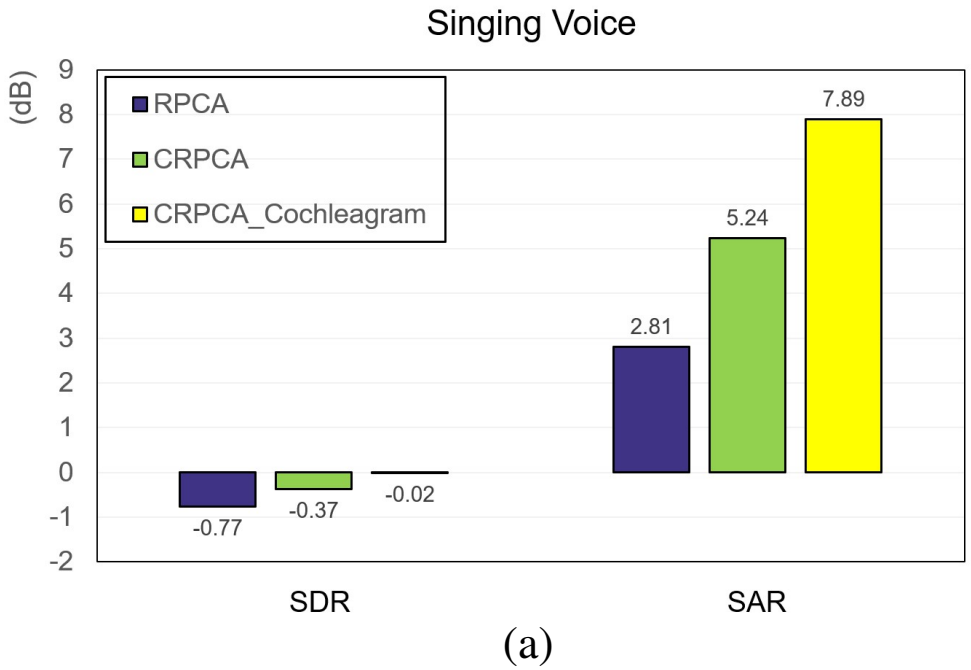
Experiment settings

All experiment data were sampled at 16 kHz. The parameters for cochleagram analysis: 128 channels, 40~8000 Hz frequency range, and 256 frequency length. To compare the results with those obtained with CRPCA, and calculated the input feature by using STFT and ISTFT, which is a part of baseline experiments that have been performed on spectrogram for the conventional RPCA method. The window size of 1024 samples was used, a hop size of 256 samples for the STFT and an FFT size of 1024.

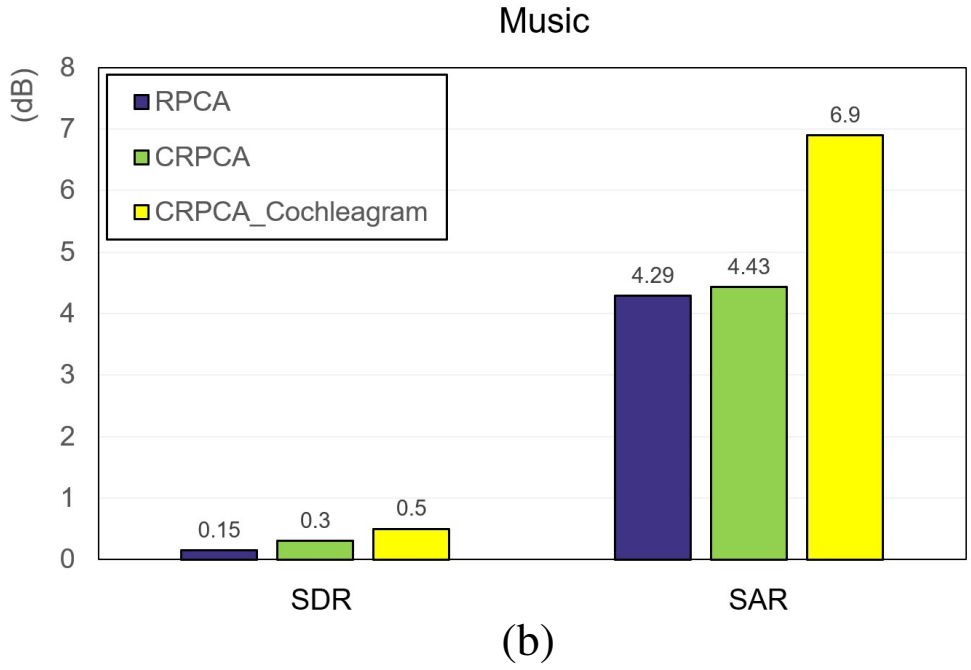
4.2.4 Result and conclusion

To examine the proposed method, the experiment was evaluated on the MIR-1K dataset. Figure 4.5 shows the comparison results of conventional RPCA, CRPCA and CRPCA on cochleagram, respectively. All methods were run by using binary time-frequency masking estimation. The experiment results show that the proposed method can improve the separation performance between singing voice and music. In terms of singing voice, the separation performance is worse than the part of music in SDR. On the contrary, the SAR of the proposed method has the highest value among them. In addition, the SAR obtains a significant improvement on cochleagram between the parts of singing voice and music.

In this work, a novel unsupervised method to deal with the singing voice separation task has been proposed. The experimental results on the MIR-1K dataset show that the proposed method outperforms the conventional RPCA method.



(a)



(b)

Figure 4.5: Comparison of unsupervised singing voice separation results on the MIR-1K dataset among of the conventional RPCA, CRPCA and CRPCA on cochleagram, respectively.

4.3 CRPCA with vocal activity detection for singing voice separation

As stated above, inspired by this melody extraction, which plays a vital role in separating singing voice [86] [87] [88] [89], convert the CRPCA output to an ideal binary masking, in-

corporate it with a harmonic masking to create a coalescent masking, and apply the coalescent masking to extract the singing voice. In addition, adopt a vocal activity detection (VAD) algorithm to constrain the temporal segments in which singing voice may occur.

4.3.1 Proposed method

In this chapter, combining the proposed CRPCA with IBM estimation to further improve the singing voice separation results from mixed music signal, we define the function M_{ibm} is defined as

$$M_{ibm}(i,j) = \begin{cases} 1 & S_{ij} \geq L_{ij} \\ 0 & S_{ij} < L_{ij} \end{cases} \tag{4.11}$$

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

Owing to the vocal F0 estimation can significantly improve the separation performance of singing voice [86], so the F0 contour properly plays a crucial role in the process of separation. Subharmonic summation is an efficient technique for this calculation [88] [90]. In the experiment, adopt the salience function $H(t, s)$, which is formulated as

$$H(t,s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2(n)), \tag{4.12}$$

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 chapter. 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

$$C = argmax \sum_{t=1}^{T-1} \left(\log a_t H(t, s_t) + \log T(s_t, s_{t+1}) \right), \tag{4.13}$$

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. The Viterbi search algorithm [91] is used

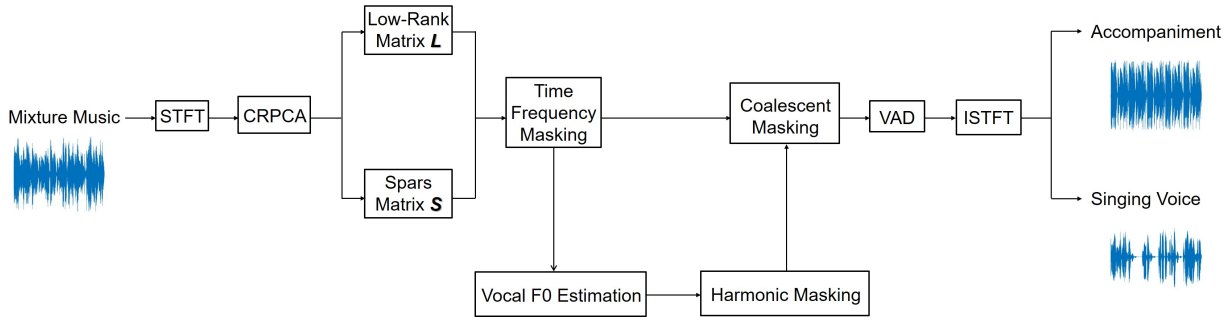


Figure 4.6: Block diagram of the proposed singing voice separation system.

to optimize the melody contour C value.

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

$$M_h(t,f) = \begin{cases} 1 & nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & others \end{cases} \tag{4.14}$$

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.

Combining the harmonic masking M_h with ideal binary time frequency masking M_{ibm} to construct the coalescent masking. The corresponding formulation M_c can be described as

$$M_c = M_{ibm} \odot M_h \tag{4.15}$$

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

To obtain better separation performance and optimize the value of coalescent masking, 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

$$\Omega(t) = \sum_f \left(\frac{1}{H_f} \sum_{n=1}^{H_f} P(t, s + 1200 \log_2(n)) \right)^{1.8}, \tag{4.16}$$

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 the proposed singing voice separation system is given in Figure 4.6. For each mixture music in the test dataset, first, apply a magnitude STFT [92] to obtain X, therefore the separate X into the corresponding low-rank matrix L and sparse matrix S by using the CRPCA method. Second, utilize coalescent masking to constrain the time-frequency masking to only those times and frequencies that constrain harmonics. The value of VAD is adopted to improve the separation performance by discriminating the vocal and non-vocal frames. Finally, utilize an ISTFT [93] to obtain the music accompaniment and singing voice parts.

In this work, randomly excerpted example 30 seconds clip from the ccMixter dataset (*Alex Beroza_-_To_Be_Sensitive_(with_mindmapthat*). Figures 4.7 and 4.8 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.

As shown in the figures, the spectrogram of Figure 4.7 contains the greatest amount of interference from background music signal (accompaniment) in the recovered singing, while in Figure 4.7(f) contains the least. As for the comparison with accompaniment in Figure 4.8, the proposed CRPCA using coalescent masking and VAD has the best value of separation performance among them.

4.3.2 Experimental evaluation

In this work, two databases are evaluated for evaluating the proposed algorithm, the first is the ccMixter dataset, which contains 50 full songs. Each audio datum contains three parts: singing voice, accompaniment, and a mixture of the two, respectively. The second dataset is DSD100 dataset. Each datum consists of bass, drums, other, and singing voice, respectively. In the experiments, all the data are used as test data, 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.

Experiment settings

All experiments main focus on the monaural source separation task. Therefore, the two-channel stereo mixture databases were downmixed into a single channel. I evaluated the whole audio

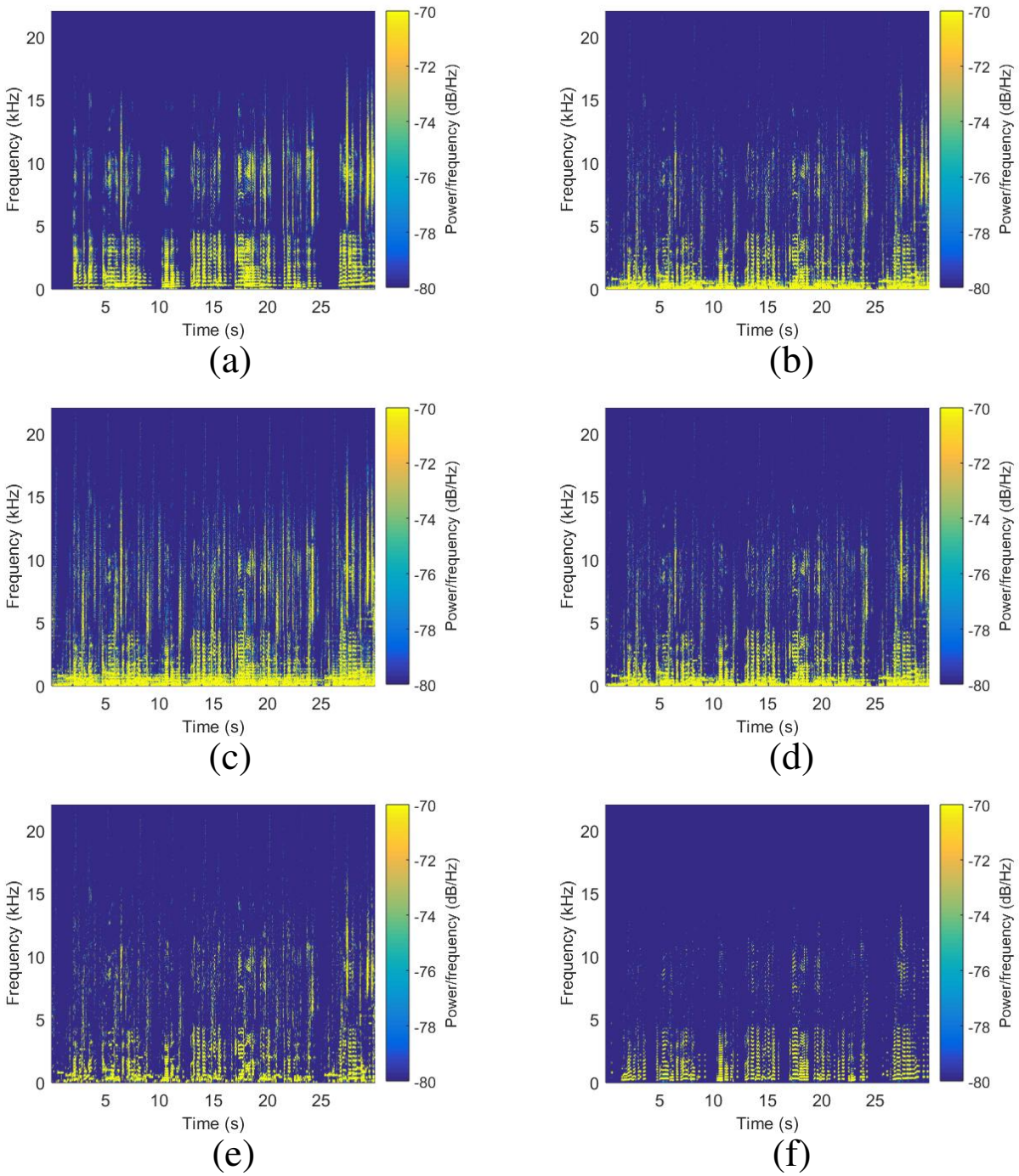


Figure 4.7: Example of spectrograms are excerpted from the ccMixter dataset: **(a)** spectrogram of original singing voice, **(b)** spectrogram of separated singing voice by RPCA, **(c)** spectrogram of separated singing voice by WRPCA, **(d)** spectrogram of separated singing voice by CR-PCA (Proposed 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.

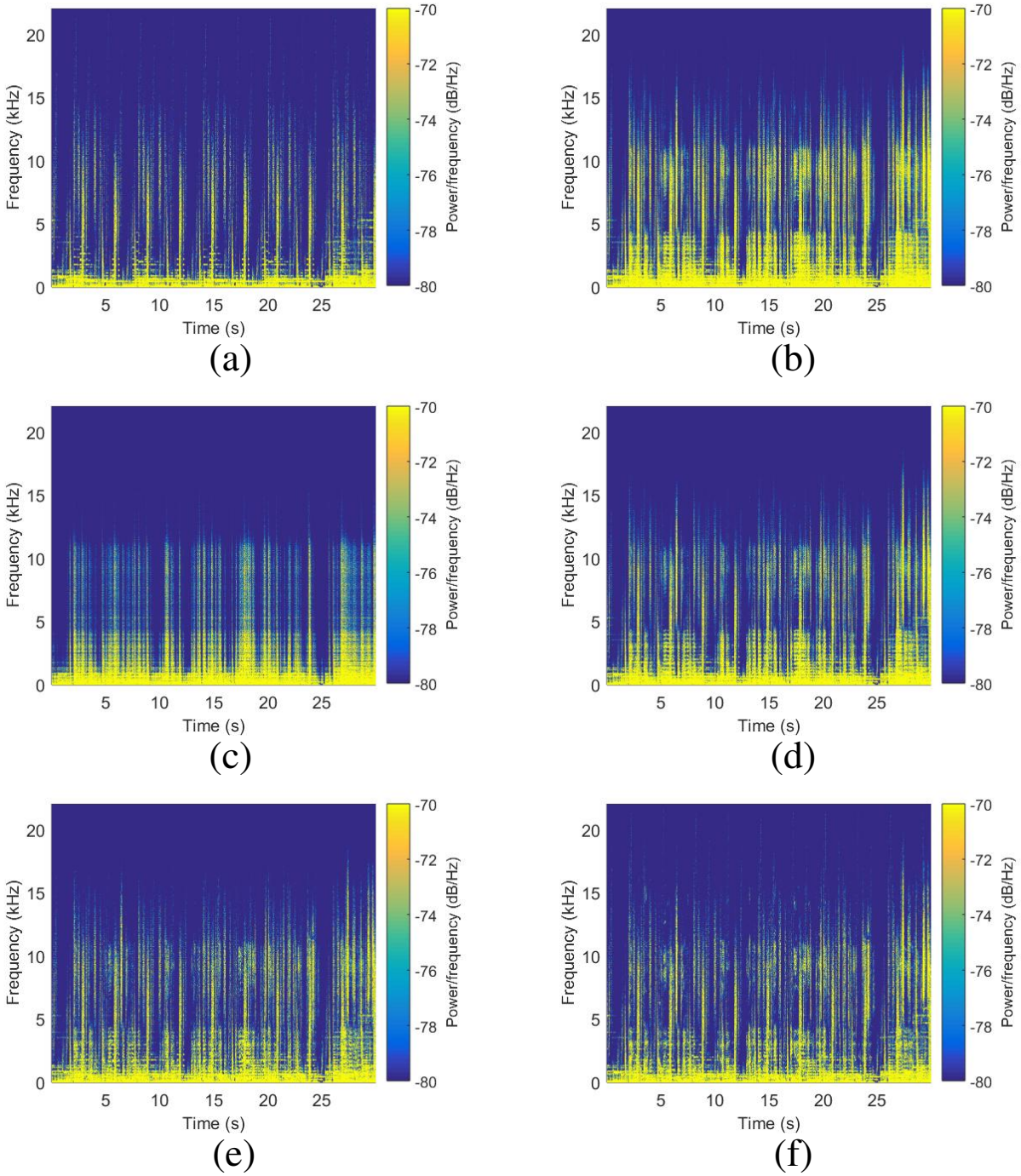


Figure 4.8: Example of spectrograms are excerpted from the ccMixter dataset: **(a)** spectrogram of original accompaniment, **(b)** spectrogram of separated accompaniment by RPCA, **(c)** spectrogram of separated accompaniment by WRPCA, **(d)** spectrogram of separated accompaniment by CRPCA (Proposed 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.

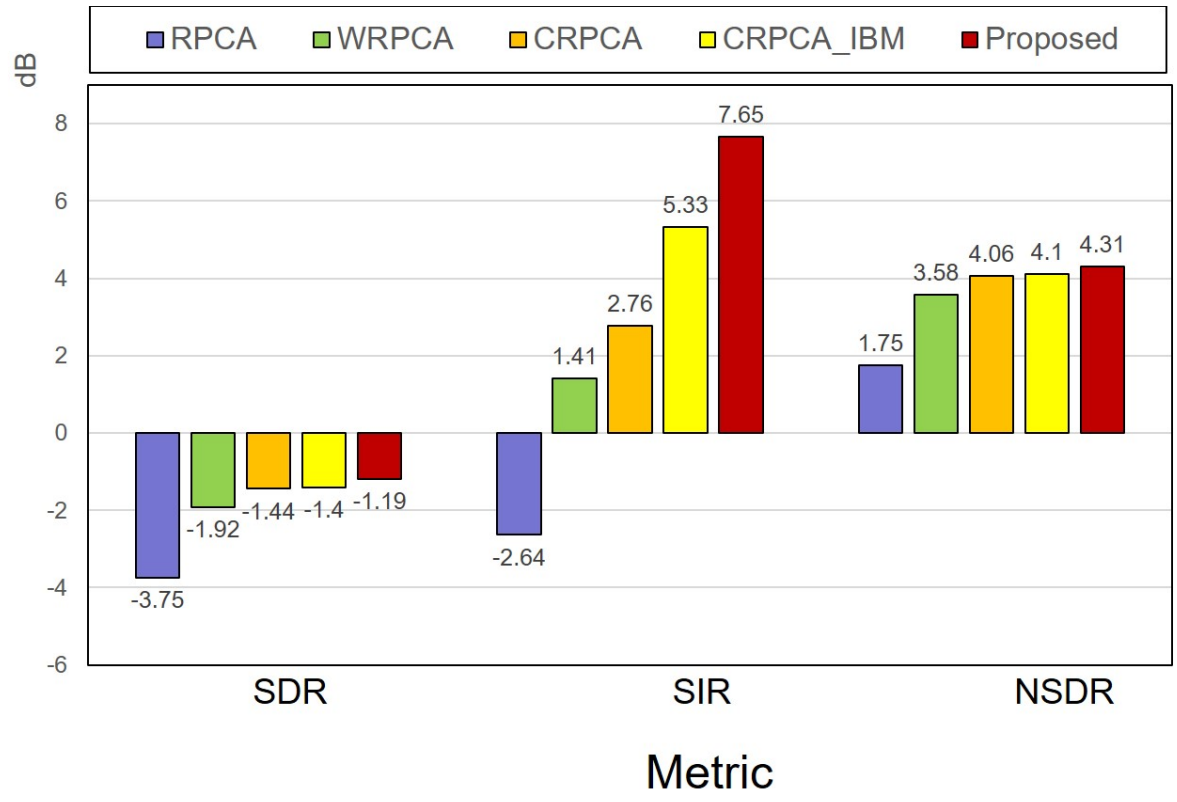


Figure 4.9: Comparison of the separation results on the **ccMixter** dataset for conventional RPCA, WRPCA, CRPCA, CRPCA with IBM, and CRPCA using coalescent masking and VAD in terms of SDR, SIR, and NSDR, respectively.

datum rather than just partial lengths on both databases. 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.

4.3.3 Result and conclusion

For the ccMixter dataset, all comparisons of singing voice separation results with the conventional RPCA, WRPCA, and the proposed methods (e.g., CRPCA only, CRPCA with IBM and CRPCA using coalescent masking and VAD) are shown in Figure 4.9. From the experimental results obtained with the SDR, SIR, and NSDR indicate that CRPCA using coalescent masking and VAD gets better separation results than others. On the contrary, the conventional RPCA was the worse in the separation task on the ccMixter dataset.

Figure 4.10 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

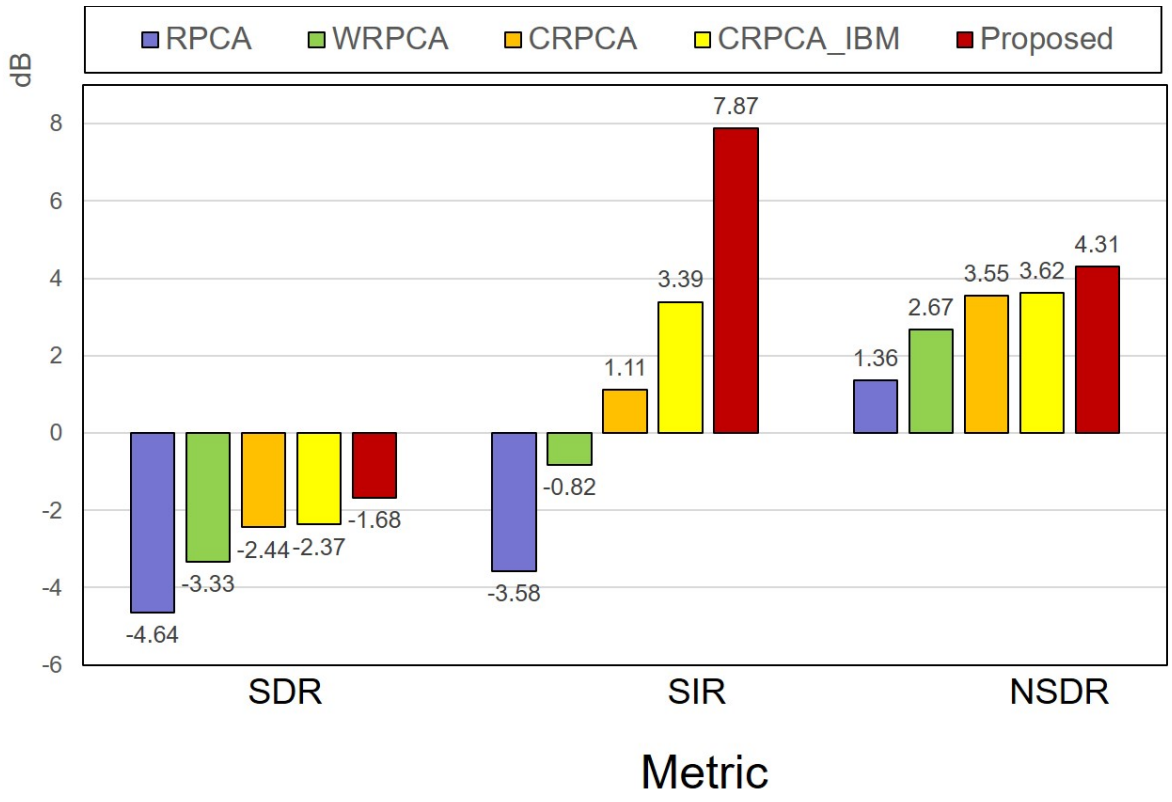


Figure 4.10: Comparison of the separation results on the **DSD100** dataset for conventional RPCA, WRPCA, CRPCA, CRPCA with IBM, and CRPCA using coalescent masking and VAD in terms of SDR, SIR, and NSDR, respectively.

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.

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 databases. 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 separation performance in terms of NSDR, the proposed method delivered improvements by +2.56 dB and +2.95 dB on the ccMixter and DSD100 datasets, respectively. Indeed, as for the value of SIR, the proposed method yielded estimates with significantly less interference, +10.29 dB and +11.45 dB, respectively.

In this chapter, a blind monaural singing voice separation based on an extension of RPCA

was proposed, which 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 among them.

4.4 Discussion and summary

This chapter proposes a novel extension of RPCA by exploring rank-1 constraint called CRPCA for singing voice separation. CRPCA utilizes rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm, which not only describes the different values of singular values decomposition, but also reduces the computation complexity, especially better than the previous proposed WRPCA algorithm in the task of singing voice separation on the different datasets.

In addition, utilizing the proposed CRPCA algorithm with the different feature to improve the separation performance from the mixed music. One is using gammatone filterbank on cochleagram instead of spectrogram for singing voice separation. Another is evaluating the proposed CRPCA with VAD method for singing voice separation. All the experiment are evaluated on the same database, the results indicate that the proposed CRPCA is better than conventional RPCA and WRPCA methods. Noticeable, the proposed CRPCA with VAD is better than without it.

Finally, we give an example by comparison with the different separation methods. Figure 4.11 is the spectrogram of the mixed music by combining singing voice with the background music (drums). The red box reflects the different of drums in the mixture music.

Figure 4.12 gives the corresponding parts of singing voice and drums by using different separation methods. The separation result is described as the following spectrograms. The mixture music is mixed by singing voice and drums. The left are the singing voice and the right are the background music (drums). (a) and (b) are the clean singing voice and drums. (c) and (d) are the separated results by using CRPCA, (e) and (f) are the separated results by using

WRPCA, (g) and (h) are the separated results by using RPCA. As shown in the left figures, the spectrograms of singing voice by separated with CRPCA and WRPCA contains the less amount of interference from the background music, especially for the part of the red box in the left spectrograms. In other words, WRPCA and CRPCA can separated drums well than RPCA in the mixed music.

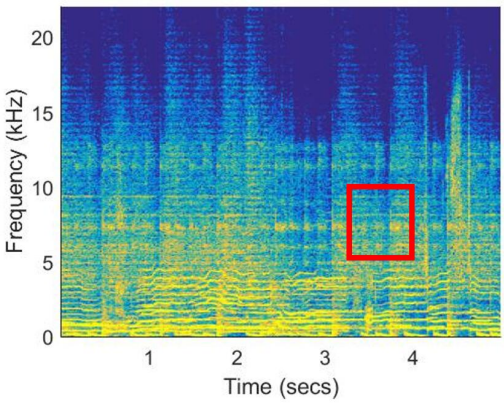


Figure 4.11: Spectrogram of the mixed music by combining singing voice with drums.

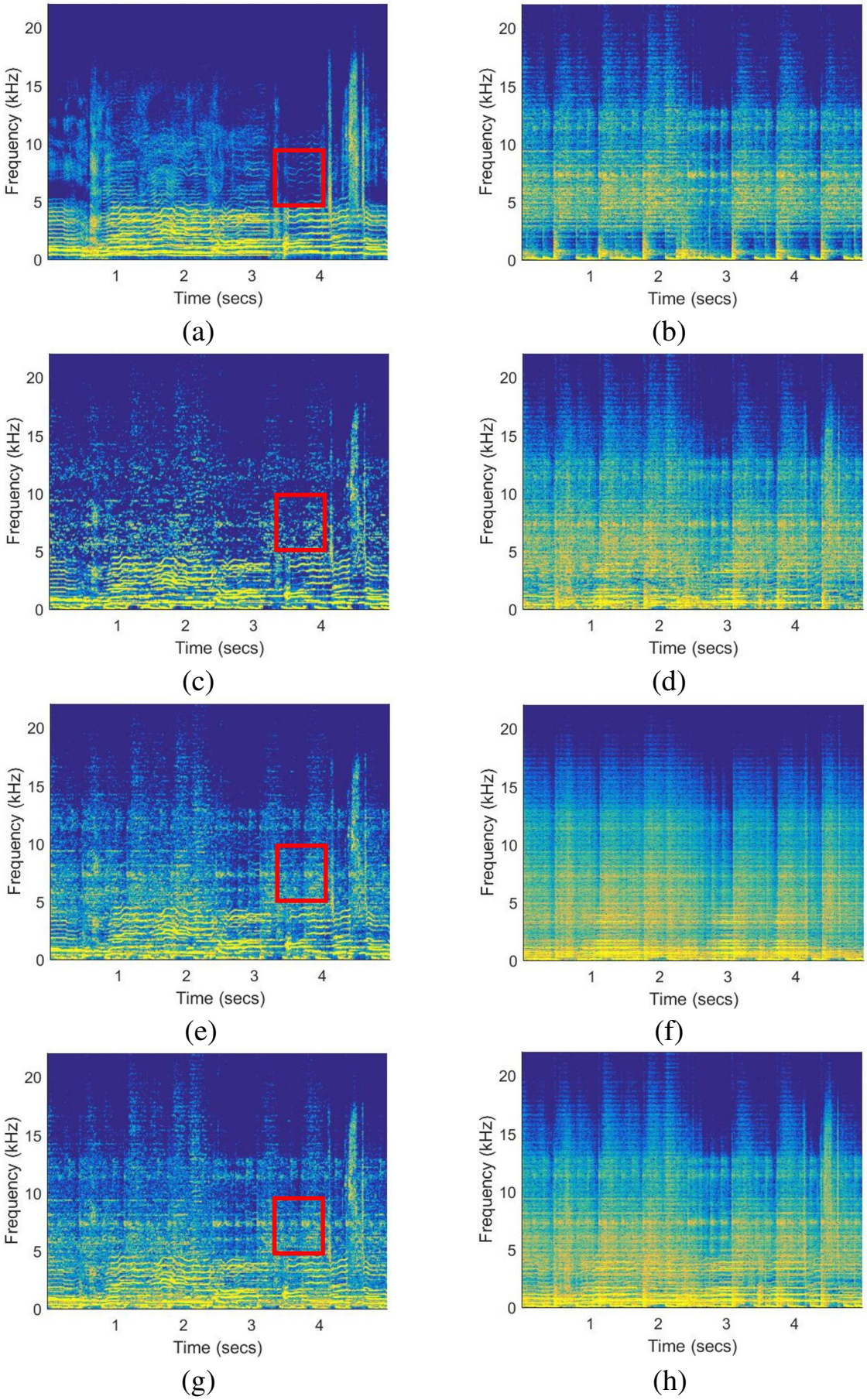


Figure 4.12: Separation results by using different separation methods.

Chapter 5

Informed NCRPCA for singing voice separation

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 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 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 chapter, a novel singing voice separation algorithm was proposed 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, use the original phase recovery in estimating the spectral components of separated singing voice.

Inspired by this sparse and low-rank model for singing voice separation, Yang [41] 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 [40]. Sprechmann et al. [42] proposed a real-time online singing voice separation by the robust low-rank modeling. Yu et al. [94] proposed sparse and low-rank representation with pre-learned dictionaries under the alternating direction method of multiplier framework for singing voice separation. Jeong et

al. [95] proposed an extension of RPCA by generalizing the nuclear norm and the l_1-norm to Schatten-p norm and l_p-norm, respectively.

To deal with these problems, Mikami et al. [96] proposed a residual drum sound estimation method for singing voice separation. Jeong et al. [97] proposed an extension of RPCA with weighted l_1-norm minimization for singing voice separation but only studied the different weighted values of a sparse matrix without including the low-rank matrix. In another work, Li et al. [80] proposed an extension of the RPCA algorithm called weighted robust principal component analysis, 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. [83] proposed an extension of RPCA with rank-1 constraint 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. A separation algorithm with additional prior information such as fundamental frequency (F0) can enhance the effectiveness of separation results [86]. 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).

Motivated by the above considerations, in this chapter, a novel singing voice separation algorithm was proposed by combining the prior information and non-negative rank-1 constraint RPCA (NCRPCA) called informed non-negative rank-1 constraint RPCA (NCRPCAi), which incorporates the human-labeled F0 and non-negative rank-1 constraint minimization of singular values in RPCA for separating the singing voice from the mixture music. Furthermore, to minimize the reconstruction error when synthesizing the singing voice, the original phase recovery is used in estimating the spectral components of the separated singing voice of the musical mixture.

5.1 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

$$\begin{aligned} &\text{minimize} \sum_{i=2}^{\min(m,n)} \delta_i(L) + \lambda|S|_1 + \frac{\gamma}{2}|S - E_0|, \\ &\text{subject to } X = L + S, L \geq 0, S \geq 0. \end{aligned} \tag{5.1}$$

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 [44] [29]. We adopt an inexact augmented Lagrange multiplier (iALM) [30] to solve this convex model. The corresponding augmented Lagrange function is defined as

$$\begin{aligned} J(X, L, S, \mu) = \min \sum_{i=2}^{min(m,n)} \delta_i(L) + \lambda|S|_1 \\ + < J, X - L - S > + \frac{\mu}{2}|X - L - S|_F^2 + \frac{\gamma}{2}|S - E_0|, \end{aligned} \tag{5.2}$$

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,

$$\begin{aligned} L_{k+1} = \min_L \sum_{i=2}^{\min(m,n)} \delta_i(L) + \langle J_k, X - L - S_k\rangle \\ + \frac{\mu_k}{2}|X - L - S_k|_F^2 + \frac{\gamma}{2}|S_k - E_0|, \end{aligned} \tag{5.3}$$

$$\begin{aligned} S_{k+1} = \min_S \lambda|S|_1 + \langle J_k, X - L_k - S\rangle \\ + \frac{\mu_k}{2}|X - L_k - S|_F^2 + \frac{\gamma}{2}|S - E_0|, \end{aligned} \tag{5.4}$$

5.1.1 Update rules based on rank-1 constraint

As suggested by Oh et al. [81], the update rules of L and S are obtained as

$$L_{k+1} = P_{1,\mu_k^{-1}}(X - S_k + \mu_k^{-1} J_k), \tag{5.5}$$

$$S_{k+1} = Q_{\lambda\mu_k^{-1}}(X - L_{k+1} + \mu_k^{-1} J_k + \gamma E0), \tag{5.6}$$

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

$$P_{1,\mu_k^{-1}}(Y) = U_Y(D_{Y_1} + Q_{\mu_k^{-1}}(D_{Y_2}))V_Y^T, \tag{5.7}$$

where the soft-thresholding operator [82] can be defined as

$$Q_{\mu_k^{-1}}(D_{Y_2}) = sign(D_{Y_2}) \cdot max(|D_{Y_2}| - \mu_k^{-1}, 0), \tag{5.8}$$

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).

5.2 Reconstructed voice spectrogram

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

$$M_h(t, f) = \begin{cases} 1 & nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & others, \end{cases} \tag{5.9}$$

Algorithm 3 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 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 set to $w = 80\ Hz$ as suggested by [86] [98]. Therefore, define the reconstructed vocal spectrogram from the vocal annotations as

$$E_0 = X \odot M_h(t, f), \tag{5.10}$$

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

5.3 Phase recovery

We calculate the magnitude spectrogram (X) by 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 [100] as can be defined

$$P = angle(X); \tag{5.11}$$

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

$$\widetilde{X} = S \odot cos(P) + i(S \odot sin(P)), \tag{5.12}$$

where S is the value of the sparse matrix separated by NCRPCAi algorithm, $\odot$ denotes the element-wise multiplication operator (Hadamard product).

Figures 5.1, 5.2, and 5.3 show an example of the waveform and spectrogram comparison of the clean and separated results using the proposed NCRPCAi and NCRPCA 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 this figure, Figure 5.1(b) contains the least amount of interference from the background music (accompaniment), in other words, NCRPCAi performs much better than NCRPCA. And the value of SDR in Figure 5.1(b) is 12.30 dB.

5.4 Experimental evaluation

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

5.4.1 Experiment settings

To confirm the effectiveness of the proposed algorithm for singing voice separation, our evaluation is carried out on the iKala dataset. 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, I 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 ($SNR = 0$). In order to evaluate the proposed method with the previous ones, 208 clips was used for the testing in the supervised method. The rest 44 clips for the training the codebooks in the comparison methods.

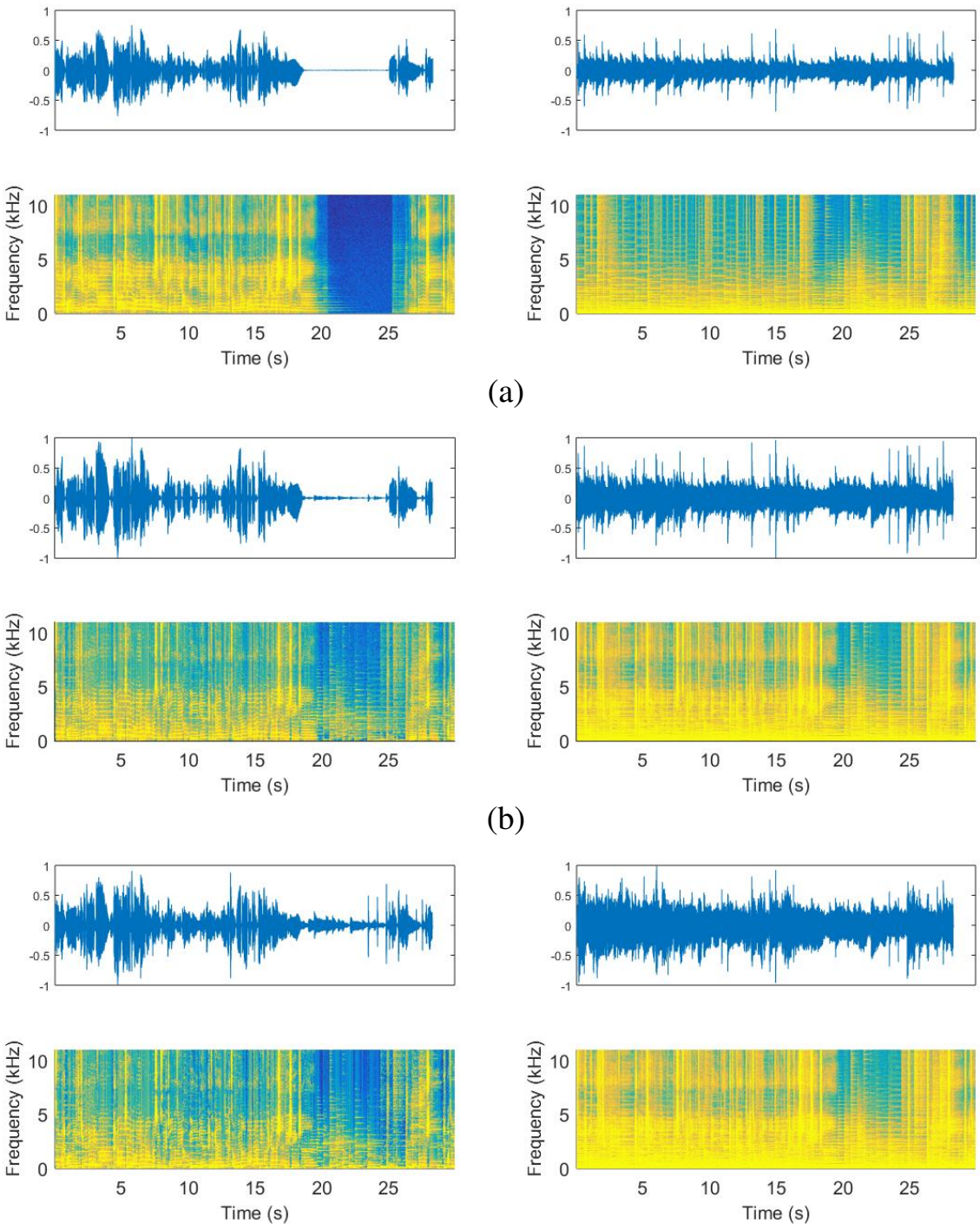


Figure 5.1: Example of waveform and spectrogram comparison of the clean and separated audio using NCRPCAi and NCRPCA methods on the iKala dataset (*71716_chorus*). Left are singing voice and the right are accompaniment. (a) is the clean audio **(Top)**, (b) and (c) are the separated audio by NCRPCAi **(Middle: SDR is 12.30 dB)** and NCRPCA **(Bottom: SDR is 6.82 dB)**, respectively.

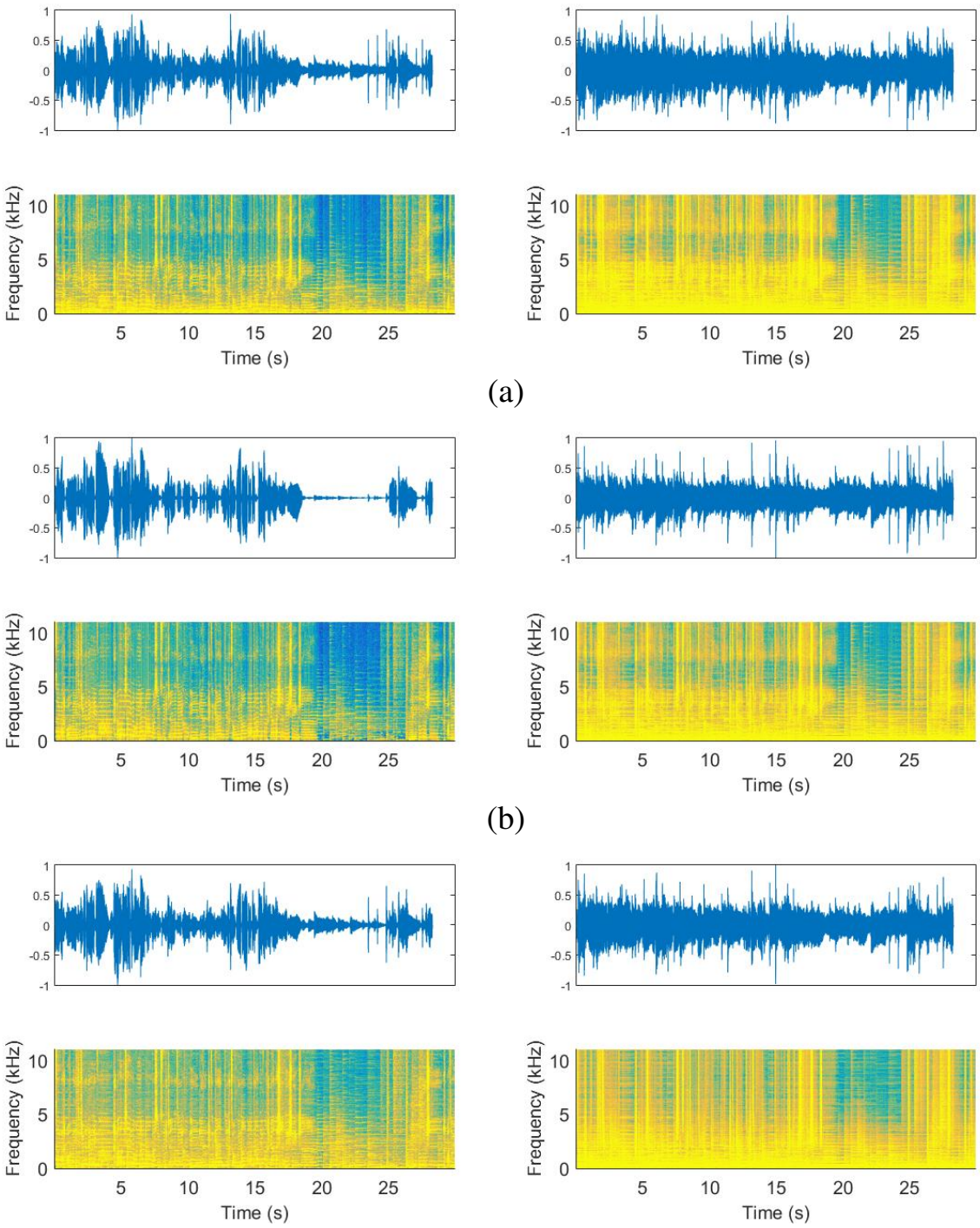


Figure 5.2: Example of waveform and spectrogram comparison of the separation results by using RPCA, RPCAi, and LRR methods on the iKala dataset (*71716_chorus*). Left are singing voice and the right are accompaniment. (a) is the separated audio by RPCA **(Top: SDR is 5.62 dB)**, (b) and (c) are the separated audio by RPCAi **(Middle: SDR is 12.28 dB)** and LRR **(Bottom: SDR is 8.05 dB)**, respectively.

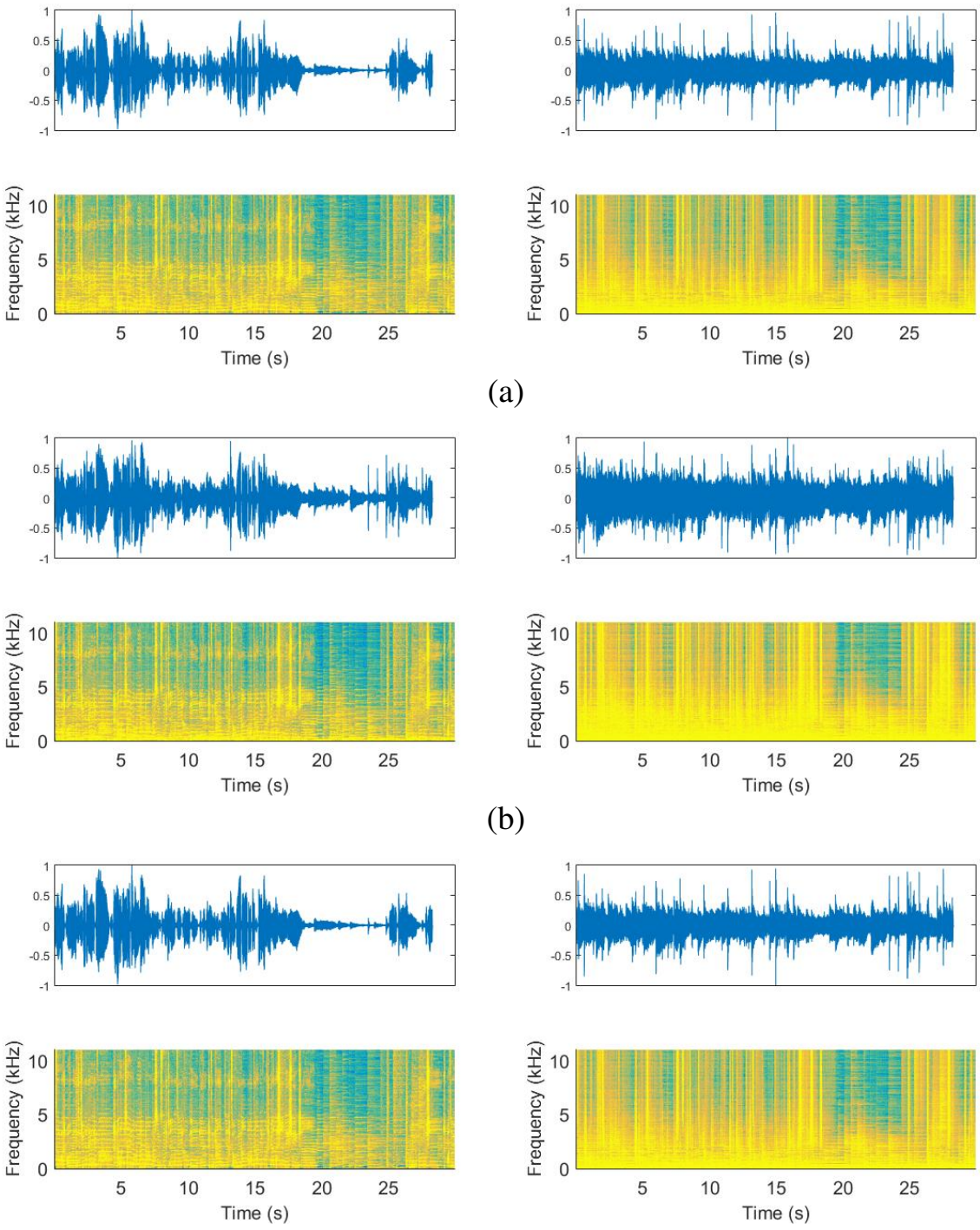


Figure 5.3: Example of waveform and spectrogram comparison of the separation results by using LRRi, GSR, and GSRi methods on the iKala dataset (*71716_chorus*). Left are singing voice and the right are accompaniment. (a) is the separated audio by LRRi **(Top: SDR is 12.18 dB)**, (b) and (c) are the separated audio by GSR **(Middle: SDR is 5.89 dB)** and GSRi **(Bottom: SDR is 12.18 dB)** methods, respectively.

Table 5.1: 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	2.8
NCRPCAi	**12.03**	**18.31**	**13.54**	**8.08**

5.4.2 Evaluation metrics

To evaluate the performance of the proposed method, assessed its separation performance in terms of SDR, SIR, SAR, and NSDR. Higher values of SDR, SIR, SAR, 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, the value of SIR reflects the suppression of the interfering source, and the value of SAR represents the absence of artificial distortion. This work reports the metrics of 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.

5.4.3 Result and conclusion

This section evaluates the proposed algorithm on the iKala dataset and compares it with unsupervised and supervised methods.

Comparison with RPCA method

Table 5.1 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 algorithms of using F0, NCRPCAi also shows much better results than RPCAi in all evaluation metrics on the iKala dataset (252).

Table 5.2: 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**

- RPCA: [11]
- RPCAi: Informed RPCA [14]
- NCRPCA: Non-negative Constraint RPCA (Proposed 1)
- NCRPCAi: Informed NCRPCAi (Proposed 2)

Comparison with state-of-the-art methods

In order to properly comparison with state-of-the-art supervised methods, in this experiment, 208 clips was used for testing in the experiment. The other 44 clips for obtaining codebooks in the training process. The supervised methods mainly utilized the online dictionary learning [101]. The SPAMS toolbox[1] is used to learn codebooks on the 44 clips, the dictionary size is 100 atoms, and the remaining 208 clips for testing.

- LRR: Low-Rank Representation [102]
- LRRi: Informed LRR [44]
- GSR: Group-Sparse Representation [44]
- GSRi: Informed GSR [44]

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

Table 5.2 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.

The results in this table indicate that all methods (LRR, GSR, and NCRPCA) performed better when using F0 than without it. The proposed NCRPCAi showed even better results than the 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.

Chapter 6

Conclusion

In this chapter, first, overall of this research is summarized. Then, discuss the contributions. Finally, future work is introduced.

6.1 Summary

This research mainly focuses on solving the singing voice separation problem. To achieve the better separation performance, the main approaches are proposed by extending RPCA method. Since RPCA has been a recently proposed of the popularization on singing voice separation algorithm that separates singing voice and musical accompaniment from the monaural recordings. Although RPCA is an effective approach to the 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).

Therefore, to overcome this disadvantage, the original topics in this dissertation are mainly research on two different deformations of RPCA and the related effective optimization algorithms with the auditory feature.

The first deformation is called WRPCA, which uses the different weighted values to describe the separated matrix. WRPCA can accurately estimate the rank of a observed matrix that include drums sounds by separated low-rank matrix. Then, combining the proposed WRPCA with gammatone auditory filterbank for singing voice separation. The significance of WRPCA can describe different low-rank matrix under the conditions of human's auditory perceptual properties. The experimental results show that the proposed method are effective for the

separated singing voice than the previous method on the ccMixter and DSD100 datasets.

The second deformation is called CRPCA, which utilizes rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm. CRPCA not only provides a robust solution to large dynamic range differences among instruments but also reduces the computation complexity. Evaluation results show that the proposed CRPCA method can achieve better separation performance than the previous methods on the ccMixter and DSD100 datasets. The running time on CRPCA is shorter than the previous proposed methods under the same conditions. Second, combining the proposed CRPCA with gammatone auditory filterbank on cochleagram, which uses an alternative time-frequency representation method to imitate human auditory system for singing voice separation. And applying the time-frequency masking estimation to improve the separation results. Evaluation results show that CRPCA with gammatone auditory filterbank achieves better separation performance than the conventional RPCA, especially for the time-frequency masking estimation method on the MIR-1K dataset. Third, further quality improvement is achieved by converting CRPCA to an ideal binary masking, incorporating it with harmonic masking to create a coalescent masking, and combining with a vocal activity detection. Evaluation results on the ccMixter and DSD100 datasets show that the proposed method achieves better separation performance than the previous methods. Finally, proposing a singing voice separation method by combining F0 and non-negative CRPCA. Experimental evaluation show that the proposed method obtain better results than others on the iKala dataset.

In a conclusion, the goal of this research is to solve the problem of singing voice separation. More specifically, two extensions of RPCA-based were proposed for separating the singing voice from the mixed music in the dissertation. One is called WRPCA, which uses the different weighted values to describe the separated low-rank matrix. The other is called CRPCA, which adopts the rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm. This is because the separated matrix by RPCA has the same weighted singular values which cannot suitable for complex audio mixtures, especially for the drums. In addition, utilizing the cochleagram on gammatone filterbank instead of spectrogram for singing voice separation. Because the cochleagram is derived from non-uniform time-frequency transform whereas time-frequency units in low-frequency regions have higher resolutions than the high-frequency regions, which closely resembles the functions of the human ear. Therefore, it is

promising to separate singing voice via the proposed extension methods on cochleagram instead of the spectrogram.

With respect to WRPCA and CRPCA method on the same datasets. WRPCA obtains the better results than CRPCA in SAR for singing voice separation. However, CRPCA can get the better separation performance than WRPCA in SIR and SDR.

6.2 Contributions

The contributions of this dissertation are to present a set of optimization algorithms focus on extensions of RPCA for singing voice separation that have the capability to separate the singing voice from the mixed music signal in monaural recordings. The other applications can be used by this sparse and low-rank model. More specifically, they contributions of this research can be summarized as follows:

- The main contribution of this study solved the problem of singing voice separation by the extensions of RPCA algorithm, WRPCA and CRPCA, respectively. After obtains singing voice from the mixed music, as the pre-processing, the outcomes can be used to improve the performance like singer identification, music emotion recognition, singing voice synthesis, etc.

- The potential contribution of this research is to deal with the problems of noise reduction and speech enhancement by using the separated low-rank and sparse model. Since the background noise is assumed as the part of low-rank component and the human speech is regarded as the part of sparse component.

6.3 Future works

This study concentrates on solving the problem of singing voice separation using RPCA and its extensions. For the future works, the remain work in this dissertation could be further studied to improve the quality of singing voice from the mixed audio signal. The specific information can be roughly described as follows:

- As for the proposed evaluation methods, sound distortions of the separated singing voice in WRPCA and CRPCA need to be further improved in singing voice separation task.

- In order to reduce the influence of noise in the process of separation, investigating robust graph embedding/learning approaches [103] [104] [105] to optimize the separation performance of the mixed audio signal. In addition, combining the time-frequency masking and F0 to improve the separation performance.

- Although the proposed methods are effective in the objective evaluation, this objective evaluation criteria cannot meet the whole singing problem and using BSS evaluation (e.g., SDR, SIR, SAR, and NSDR) is also limited to actual application. It is because of the difference between the energy of error and how the listeners perceived it. More specifically, when the application provides the separated sources to users for being played (e.g., Kalaoke), the subjective quality of singing voice can be more important than the numerical one. For future work, we will do some subjective evaluation experiments on the singing voice separation problem. According to the different listeners, evaluate the quality in terms of preservation of the target source in each test signal. In addition, evaluate the global quality compared to the reference for each test signal.

Bibliography

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

[2] M. N. Chinthaka, C.S. Xu, and 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] Y. H. Yang and H. H. Chen, "Music emotion recognition," *CRC Press*, 2011.

[4] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," in *Proceedings of the International Computer Music Conference (ICMC)*, 1999, pp. 464-467.

[5] 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.

[6] C. Laroche, H. Papadopoulost, M. Kowalski, and G. Richard, "Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 46-50.

[7] A. J. R. Simpson, G. Roma, and 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.

[8] A. Hille and S. Jürgen, "How learning a musical instrument affects the development of skills," *Economics of Education Review*, vol. 44, 2015, pp. 56-82.

[9] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and 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.

[10] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, 2018, pp. 293-305.

[11] P. S. Huang, S. D. Chen, P. Smaragdis, and 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.

[12] J. Lee and 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.

[13] C. K. Wang, R. Y. Lyu, and 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.

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

[15] R. M. Bittner, B. McFee, and J. P. Bello, "Multitask learning for fundamental frequency estimation in music," in *arXiv preprint arXiv:1809.00381*, 2018.

[16] D. L. Wang and 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.

[17] 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.

[18] M.N. Schmidt and 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.

[19] A. Chanrungutai and 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.

[20] S. Mirzaei, H. V. hamme, and 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.

[21] V. Tuomas, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, 2017, pp. 1066-1074.

[22] C. Févotte, E. Vincent, and A. Ozerov, "Audio source separation," *chapter single-channel audio source separation with NMF: divergences, constraints and algorithms,Springer*, 2018.

[23] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Charleston SC, USA*, pp. 32-39, 2006.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computating*, vol. 23, no. 9, pp. 2421-2456, 2011.

[25] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 221–224.

[26] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): a simple method for music/voice separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, 2013, pp. 73-84.

[27] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 53–56.

[28] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 583–588.

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

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

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

[32] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *Journal of Computational Information Systems*, vol. 10, no. 10, 2014, pp. 4403-4410.

[33] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, T. Kawahara, and 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*, vol. 26, no. 2, 2018, pp. 215-230.

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

[35] F. Biondi, "A polarimetric extension of low-rank plus sparse decomposition and radon transform for ship wake detection in synthetic aperture sadar images," *IEEE Geoscience and Remote Sensing Letters*, 2018.

[36] A. Das, "A Bayesian sparse-plus-low-rank matrix decomposition method for direction-of-arrival tracking," *IEEE Sensors Journal*, vol. 17, no. 15, 2017, pp. 4894-4902.

[37] T. Bouwmans, S. Javed, H. Zhang, and Z. Lin, "On the applications of robust PCA in image and video processing," *Proceedings of the IEEE*, 2018, pp. 1427-1457.

[38] T. Bouwmans, A. Sobral, S. Javed, S. Jung, and 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*, vol. 23, 2017, pp. 1-71,

[39] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: robust PCA, robust subspace tracking and robust subspace recovery," *IEEE Signal Processing Magazine*, vol. 35, no. 4, 2018, pp. 32-55.

[40] Y. H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *Proc. ACM Multimedia*, 2012, pp. 757-760.

[41] 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.

[42] P. Sprechmann, A. Bronstein, and 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.

[43] D. Fourer and G. Peeters, "Single-channel blind source separation for singing voice detection: a comparative study," in *arXiv preprint arXiv*:1805.01201, 2018.

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

[45] J. Pu, Y. Panagakis, S. Petridis, J. Shen, and M. Pantic, "Blind audio-visual localization and separation via low-rank and sparsity," *IEEE Transactions on Cybernetics*, 2018, pp. 2168-2267.

[46] J. Pu, Y. Panagakis, and M. Pantic, “Learning low rank and sparse models via robust autoencoders,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3192-3196.

[47] P. Chandna, M. Miron, J. Janer, and E. Gomez, “Monoaural audio source separation using deep convolutional neural networks,” in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2017.

[48] N. Takahashi and Y. Mitsufuji, “Multi-scale multi-band DenseNets for audio source separation,” in *Proc. IEEE Workshop Appl. Signal Process.Audio Acoust.*, Oct. 2017, pp. 21–25.

[49] N. Takahashi, N. Goswami, and Y. Mitsufuji, “MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation,” in *arXiv preprint arXiv*:1805.02410, 2018.

[50] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Singing-voice separation from monaural recordings using deep recurrent neural networks,” in *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 477-482.

[51] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, “Deep neural network based instrument extraction from music,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2135-2139.

[52] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, “Deep clustering and conventional networks for music separation: stronger together,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 61-65.

[53] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, “Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 721-725.

[54] M. Michelashvili, S. Benaim, and L. Wolf, "Semi-supervised Monaural singing voice separation with a masking network trained on synthetic mixtures," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 291-295.

[55] W. Yuan, S. Wamg, X. Li, M. Unoki, and W. Wang, "Proximal deep recurrent neural network for monaural singing voice separation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 286-290.

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

[57] K. W. E. Lin and 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.

[58] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proceedings of 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 745-751.

[59] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[60] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proceedings of 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[61] J. Oh, D. Kim and Se-Young Yun, "Spectrogram-channels u-net: a source separation model viewing each channel as the spectrogram of each source," in *arXiv preprint arXiv:1810.11520*, 2018.

[62] V. S. Narayanaswamy, S. Katoch, J. J. Thiagarajan, H. Song, and A. Spanias, "Audio source separation via multi-scale learning with dilated dense U-Nets," in *arXiv preprint arXiv:1904.04161.*, 2019.

[63] C. S. J. Doire, “Online singing voice separation using a recurrent one-dimensional U-NET trained with deep feature losses,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3752-3756.

[64] Z. Rafii, A. Liutkus, and F.R. Stöter, S.I. Mimilakis, D. FitzGerald, and 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.

[65] E.M. Grais, M.U. Sen, and 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.

[66] J.R. Hershey, Z. Chen, J.L. Roux, and 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.

[67] Y. Luo, Z. Chen, and 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.

[68] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.R. Stöter, “Musical source separation: An introduction,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, 2019, pp.31-40.

[69] D. L. Wang and G. J. Brown, “Computational auditory scene analysis: principles, algorithms, and applications,” *Wiley-IEEE Press, Hoboken*, 2006.

[70] C. L. Hsu and 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.

[71] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, “Kernel additive models for source separation,” *IEEE Transactions on Signal Processing*, vol. 62, no. 16, 2014, pp. 4298-4310.

[72] T.S. Chan, T.C. Yeh, Z.C. Fan, H.W. Chen, L. Su, Y.H. Yang, and 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.

[73] E. Vincent, R. Gribonval, and 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.

[74] E. Vincent, S. Araki, F. Theis, R. Gribonval, G. Nolted, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunkerf, D. Lutter, and N.Q.K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges,"*Signal Processing*, vol. 92, no. 8, 2012, pp. 1928-1936.

[75] E.J. Candés, M.B. Wakin, and S.P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, 2008, pp. 877-905.

[76] S.H. Gu, Q. Xie, D.Y. Meng, W.M. Zuo, X.C. Feng, and L. Zhang, "Weighted nuclear norm minimization and its applications to low level vision," *International Journal of Computer Vision*, vol. 121, no. 2, 2017, pp. 183-208.

[77] F. Li and M. Akagi, "Singing voice separation using weighted robust principal component analysis," in *Proc. ASJ Autumn Meeting*, 2017, pp. 559-562.

[78] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and itakura–saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 3, 2013, pp. 662-675.

[79] Y. P. Li and D. L. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol 51, no. 3, 2009, pp. 230-239.

[80] F. Li and 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.

[81] T. H. Oh, Y. W. Tai, J. C. Bazin, H. Kim, and I. S. Kweon, "Partial sum minimization of singular values in robust PCA: algorithm and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, 2016, pp. 744-758.

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

[83] F. Li and 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. 1929-1933.

[84] F. Li and M. Akagi, "'Unsupervised singing voice separation using gammatone auditory filterbank and constraint robust principal component analysis," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2018, pp.1924-1928.

[85] G. N. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on neural networks*, vol. 15, no. 5, 2004, pp. 1135-1150.

[86] Y. Ikemiya, K. Itoyama, and 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.

[87] J. Salamon, E. Gómez, D.P.W. Ellis. and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, 2014, pp. 118-134.

[88] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, 2012, pp. 1759-1770.

[89] M. Chen, B. Li, and T.-S. Chi, "CNN based two-stage multi-resolution end-to-end model for singing melody extraction," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1005-1009.

[90] D.J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the acoustical society of America*, vol. 83, no. 1, 1998, pp. 257-264.

[91] G.D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, 1973, 268-278.

[92] S.H. Nawab, T.F. Quatieri, and J.S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, no. 4, 1983, pp. 986-998.

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

[94] S. Yu, H. Zhang, and Z. 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.

[95] I.Y. Jeong and 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.

[96] S. Mikami, A. Kawamura, and 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.

[97] I.Y. Jeong and 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.

[98] F. Li and M. Akagi, "Blind monaural singing voice separation using rank-1 constraint robust principal component analysis and vocal activity detection," *Neurocomputing*, vol. 350, 2019, pp. 44-52.

[99] F. Li and M. Akagi, "Combining F0 and Non-Negative Constraint Robust Principal Component Analysis for Singing Voice Separation," *Signal Processing. (Under Review)*

[100] E. Vincent, N. Bertin, R. Gribonval, and 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.

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

[102] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and 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.

[103] N. Shahid, V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Robust principal component analysis on graphs," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2812-2820.

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

[105] Z. Kang, H. Pan, S. Hoi, and Z. Xu, "Robust graph learning from noisy data," *IEEE transactions on cybernetics*, 2019.

Publications

Journal Paper

[1] **Feng Li** and Masato Akagi, "Blind Monaural Singing Voice Separation Using Rank-1 Constraint Robust Principal Component Analysis and Vocal Activity Detection," *Neurocomputing*, vol. 350, pp. 44-52, 2019.

[2] **Feng Li** and Masato Akagi, "Combining F0 and Non-Negative Constraint Robust Principal Component Analysis for Singing Voice Separation," *Signal Processing*. (Revise)

Book Chapter

[3] **Feng Li** and Masato Akagi, "Weighted Robust Principal Component Analysis with Gammatone Auditory Filterbank for Singing Voice Separation," *Neural Information Processing, Lecture Notes in Computer Science, LNCS 10639, Springer*, pp. 849-858, 2017.

International Conference

[4] **Feng Li**, Kaizhi Qian, Mark Hasegawa-Johnson, and Masato Akagi, "Monaural Singing Voice Separation Using Fusion-Net with Time-Frequency Masking," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2019.

[5] **Feng Li** and Masato Akagi, "Unsupervised Singing Voice Separation Using Gammatone Auditory Filterbank and Constraint Robust Principal Component Analysis," in *Proceed-*

ings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.1924-1928, 2018.

[6] **Feng Li** and Masato Akagi, "Unsupervised Singing Voice Separation Based on Robust Principal Component Analysis Exploiting Rank-1 Constraint," in *Proceedings of 26th European Signal Processing Conference (EUSIPCO)*, pp.1920-1924, 2018.

[7] **Feng Li** and Masato Akagi, "Weighted Robust Principal Component Analysis with Gammatone Auditory Filterbank for Singing Voice Separation," in *Proceedings of the 24th International Conference on Neural Information Processing (ICONIP), Springer*, pp. 849-858, 2017.

Domestic Conference

[8] **Feng Li**, Masato Akagi, and Mark Hasegawa-Johnson, "Audio Source Separation Using Deep Fully Residual Convolutional Neural Network," in *Proceedings of ASJ Spring Meeting*, pp. 1309-1312, 2019.

[9] **Feng Li** and Masato Akagi, "Auditory-Inspired Audio Source Separation Using Constraint Robust Principal Component Analysis based on Cochleagram," in *Proceedings of ASJ Spring Meeting*, pp. 395-398, 2019.

[10] **Feng Li** and Masato Akagi, "Audio Source Separation via Constraint Robust Principal Component Analysis," in *Proceedings of ASJ Spring Meeting*, pp. 555-558, 2018.

[11] **Feng Li** and Masato Akagi, "Singing Voice Separation using Weighted Robust Principal Component Analysis," in *Proceedings of ASJ Autumn Meeting*, pp. 559-562, 2017.