공학박사 학위논문

고유 특성을 활용한 음악에서의 보컬 분리

Separation of Singing Voice from Music Exploiting Its Distinct Characteristics

2018 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

정 일 영

공학박사 학위논문

고유 특성을 활용한 음악에서의 보컬 분리

Separation of Singing Voice from Music Exploiting Its Distinct Characteristics

2018 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

정 일 영

고유 특성을 활용한 음악에서의 보컬 분리

Separation of Singing Voice from Music Exploiting Its Distinct Characteristics

지도교수 이 교 구

이 논문을 공학박사 학위논문으로 제출함

2018 년 1 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

정 일 영

정일영의 공학박사 학위논문을 인준함

2018 년 1 월

위 원 장	이 원 종	(인)
부위원장	이 교 구	(인)
위 원	곽 노 준	(인)
위 원	남 주 한	(인)
위 원	이 석 진	(인)

Abstract

Singing voice separation (SVS) refers to the task or the method of decomposing music signal into singing voice and its accompanying instruments. It has various uses, from the preprocessing step, to extract the musical features implied in the target source, to applications for itself such as vocal training.

This thesis aims to discover the common properties of singing voice and accompaniment, and apply it to advance the state-of-the-art SVS algorithms. In particular, the separation approach as follows, which is named 'characteristics-based,' is concentrated in this thesis. First, the music signal is assumed to be provided in monaural, or as a single-channel recording. It is more difficult condition compared to multiple-channel recording since spatial information cannot be applied in the separation procedure. This thesis also focuses on unsupervised approach, that does not use machine learning technique to estimate the source model from the training data. The models are instead derived based on the low-level characteristics and applied to the objective function. Finally, no external information such as lyrics, score, or user guide is provided. Unlike blind source separation problems, however, the classes of the target sources, singing voice and accompaniment, are known in SVS problem, and it allows to estimate those respective properties.

Three different characteristics are primarily discussed in this thesis. Continuity, in the spectral or temporal dimension, refers the smoothness of the source in the particular aspect. The spectral continuity is related with the tim-

bre, while the temporal continuity represents the stability of sounds. On the other hand, the low-rankness refers how the signal is well-structured and can be represented as a low-rank data, and the sparsity represents how rarely the sounds in signals occur in time and frequency.

This thesis discusses two SVS approaches using above characteristics. First one is based on the continuity and sparsity, which extends the harmonic-percussive sound separation (HPSS). While the conventional algorithm separates singing voice by using a two-stage HPSS, the proposed one has a single stage procedure but with an additional sparse residual term in the objective function. Another SVS approach is based on the low-rankness and sparsity. Assuming that accompaniment can be represented as a low-rank model, whereas singing voice has a sparse distribution, conventional algorithm decomposes the sources by using robust principal component analysis (RPCA). In this thesis, generalization or extension of RPCA especially for SVS is discussed, including the use of Schatten p-/l_p-norm, scale compression, and spectral distribution. The presented algorithms are evaluated using various datasets and challenges and achieved the better comparable results compared to the state-of-the-art algorithms.

Keywords: Singing voice separation, Optimization, Music signal processing
Student Number: 2013-30733

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A signal obtained from a real-world is generally a mixture, in other words, it consists of various co-occurring sources. When analyze or extract information from the captured signal, mainly focusing on a specific target source, then the others are considered as noise that disturbs the analysis procedure. In this case, therefore, the appropriate algorithm to extract the target source and remove the others, or source separation algorithm is required. Fig. 1.1 shows how source separation algorithm works for an environmental audio signal as an example.

Source separation algorithms can be applied to various domains, and they have been developed in domain- or task-specific in usual. When monitoring electrical activities of neurons in the brain using electroencephalography (EEG) or magnetoencephalogram (MEG), the captured signal is often corrupted by undesired noise, such as eye blinking or muscle movement. Therefore source separation is performed as a preprocessing step to extract the clean desired signal from the observed mixture [1] . On the other hand, when a speech signal

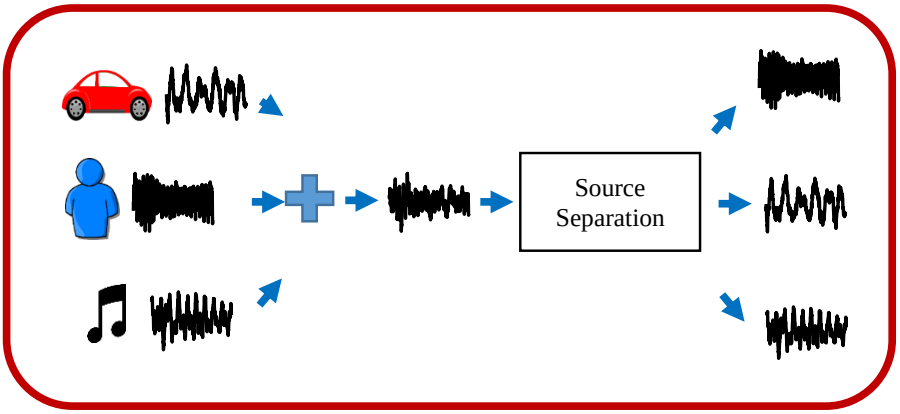

Fig. 1.1 An example of source separation framework

is captured using a microphone for speech recognition, various environmental noise may coincide. In this situation, source separation in terms of speech enhancement or noise reduction can be executed to prevent the degeneration of recognition quality [2].

Music signals, which is the main aim of this thesis, is also mixtures since it contains various instrumental tracks such as a drum, piano, guitar and so on, and also has a singing voice or sound effects. However, the definition of 'source' in the music signal can be varied depending on the application. Some may aim to extract a specific instrument such and considered a sum of all other ones as noise [3], while some others separated all the instruments individually [4]. On the other hand, a single source can be considered as not only an individual instrument but also a group of them. For example, harmonic-percussive sound separation (HPSS) categorizes musical instruments into two groups, which are

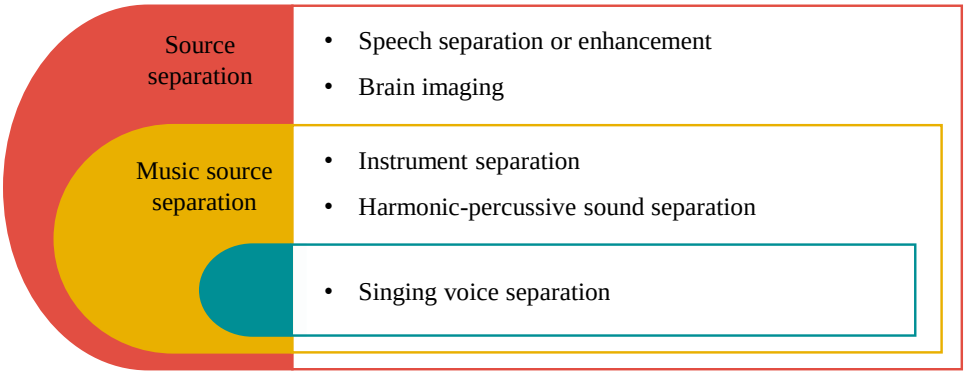

Fig. 1.2 Category of source separation tasks, focusing on the singing voice separation.

harmonic instruments and percussive instruments [5, 6].

This thesis focuses on the separation of singing voice and its accompanying instruments or singing voice separation (SVS), which is one of the music source separation problems as shown in Fig. 1.2. Although the primary target source of SVS is singing voice, this thesis also consider accompaniment as another source rather than noise. Therefore the primary goal of this thesis is to develop SVS algorithms which achieve high separation quality for both singing voice and accompaniment.

The rest of this chapter is organized as follows. In 1.1 the motivations that why SVS is important and is a challenging task are introduced. Several applications are listed in 1.2. Basic definitions of the task in both mathematical and conceptual are derived in 1.3. The main goals of the thesis and the subtasks to

achieve it is described in 1.5, and the outline of the thesis is briefly introduced in 1.6.

1.1 Motivation

Developing SVS algorithm is an important task for the following reasons. First, most of the music has both singing voice and accompaniment. Although there are some exceptions such as a cappella (singing voice only) or instrumental music (accompaniment only), it is relatively rare, especially in the modern popular music. Considering the usability of the algorithm, therefore SVS is one of the music source separation algorithms that can be applied most widely.

In addition, singing voice and accompaniment have distinct roles and provide unique information for the music. In case of singing voice, the information about singer, lyrics, lead melody, and even emotion of music can be acquired. From the accompaniment, information about instruments played in music, chords, and rhythm can be obtained. Since when analyzing a source the other one is not just useless but disturbing analysis as noise, thus the separation of sources is important preprocessing step for the understanding of music.

However, developing SVS algorithm is challenging due to the following difficulties. First, it is difficult to represent singing voice using a simple model because of its irregular patterns compared to other musical instruments. For example, all the singers have different timbre based on their gender, age, nationality or personal character. In addition, there are various singing styles that one singer can do, including falsetto, shouting, screaming and so on. Since singing voice is mostly based on the lyrics, it also has a variation depending

on the pronunciation and note. Therefore conducting a model which represent the shared characteristics in these varied singing voice and which distinguish it from other accompaniment may be the first step for the development of SVS.

1.2 Applications

SVS can be applied to the numerous applications. Below are the examples of them, which are categorized into three groups, singing voice-related, accompaniment-related, and other applications.

Singing voice-related applications Various MIR tasks use singing voice in music, and SVS is required as a preprocessing step when only the mixture with accompaniment is provided for the tasks. Singer identification [7], singing voice activity detection [8], singing voice melody estimation [9], lyric recognition [10] and singing voice-to-lyric alignment [11] are the examples of singing voice-related MIR applications.

Accompaniment-related applications First, SVS can be applied as the one part of cascade instrument separation framework. For example, when separating singing voice, harmonic instruments, and percussive instruments, the accompaniment signal separated using SVS can be considered as a sum of a harmonic and percussive instrument. It also leads to various instrument-specific MIR applications, such as chord estimation [12], or tempo and beat estimation [13]. In addition, many other applications besides information retrieval use accompaniment, including karaoke and vocal training.

Other applications Obtaining the individual sources from music allows the diverse reproduction of the music or the rich listening experience of users. For example, it can extend the conventional equalizer which scales for each frequency in general to scales for each source [14]. In case of the music reproducing such as remixing or upmixing, source separation allows the individual source-wise processing including voice conversion or source localization [15]. In addition, the source-wise processing is also useful for the music visualization for the information retrieval [16] or artistic representation.

1.3 Definitions and keywords

When a mixture m is obtained, it can be represented as a sum of sound units as follows:

$$m(n) = \sum_{k} u_k(n), \tag{1.1}$$

where n denotes the time index, and u_k denotes k-th sound unit. Under the assumption that all the sound unit u_k is corresponds to singing voice or accompaniment, then (1.1) can be alternately represented as follows:

$$m(n) = v(n) + a(n), \tag{1.2}$$

where v and a denotes the singing voice and accompaniment signal occurred in m, respectively, which can be represented as

$$v(n) = \sum_{k \in \mathbf{V}} u_k(n), \tag{1.3}$$

$$a(n) = \sum_{k \in \mathbf{A}} u_k(n), \tag{1.4}$$

where $\mathbf{V}$ is a set of sound unit indices which is correspond to singing voice, and $\mathbf{A}$ is a complement set of $\mathbf{V}$, which correspond to accompaniment. The main goal of SVS is to find v and a in (1.2) from m.

For all the steps in studies for SVS, including model estimation and evaluation, defining which sound unit u_k is correspond to singing voice and what is not is mandatory. However, it is not required to be precise, and previous studies tend to group them roughly to be intuitively agreeable. The following definition may be an example, and it is used for this thesis. It is noted that these are rough definitions and not considered precisely in the development of the SVS algorithms presented in this thesis.

Singing voice is roughly defined as all the musical sounds played by using the human voice. It includes singing, rapping, and chorus, and even scat, whistling, screaming and growling. However, non-vocal sounds occurred from human such as clapping is not considered as singing voice.

Accompaniment is defined as all the musical sound which is not considered as singing voice. It includes all the typical instruments including piano, guitar, and drum, and also synthesized sounds or sound effects.

1.4 Evaluation criteria

The evaluation criteria for SVS can be varied depending on the purpose of its applications. One of the simplest method is to calculate the difference between the original sources and the separated ones, by using mean square error, for example. If the separated sources will be directly provided to user, then the

separation quality should be evaluated by them. On the other hand, the performance difference of the specific application between with/without SVS can be measured when SVS is used as a preprocessing step of it.

Belows are the detailed explanation for the evaluation approaches.

Numerical measurement

Numeric criteria simply measure the error in low-level between the estimated and the target signals. Decomposition-based measurements presented by Vincent et al. is one of the most widely used for the evaluation of blind source separation and even for music source separation [17] . It decomposes the separated output signal $\hat{s}$ as follows:

$$\hat{s} = s_{target} + e_{interf} + e_{noise} + e_{artif}, \tag{1.5}$$

where s denotes the original target source, and e_{interf}, e_{noise}, and e_{artif} denotes the errors, which are the interferences, additional noise e.g. sensor noise, and the artifacts occur in the separation procedure, respectively. In music source separation tasks, e_{noise} is often ignored or considered as zero. Focusing on the specific noise, separation quality can be measured as follows:

$$\text{SIR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2}, \tag{1.6}$$

$$\text{SAR} = 10\log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{artif}||^2}, \tag{1.7}$$

$$\text{SDR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{artif}||^2}, \tag{1.8}$$

where SIR, SAR, and SDR denotes the source-to-interferences ratio, source-to-artifacts ratio, and source-to-distortion ratio, respectively. Here, SIR and SAR

can be interpreted as 'how much the non-target sources are eliminated', and 'how less the target source is damaged by artifacts'. SDR is considered as the overall quality of the separation results.

In a signal with long time duration, the intensity of the signal is varied over time. In this case, the above measurements are biased to the separation results at the high-intensity time region by definitions, while human perception is also sensitive to the errors in the low-intensity region. To narrow the gap between human subjectivity and numerical evaluation results, segmental SDR, which calculates the SDR for each segment of the source and takes an average of them, can be used.

Several methods for summarizing the measurements of multiple data have been used. Because each mixture can have different input signal-to-noise ratio, normalized SDR calculates the gain of SDR comparing before and after separation procedure. It is defined as follows:

$$\text{NSDR} = \text{SDR}(\hat{s}, s_{target}) - \text{SDR}(m, s_{target}), \tag{1.9}$$

where m is a input mixture. In addition, global NSDR (GNSDR) calculates the mean of the multiple evaluation data with weights by its respective time duration. It is defined as follows:

$$\text{GNSDR} = \frac{\sum_i T_i \text{NSDR}_i}{\sum_i T_i}, \tag{1.10}$$

where T_i and NSDR_i denote the time duration and NSDR of i-th data, respectively.

Subjective evaluation

Above numeric evaluation criteria does not exactly represents how human evaluate it subjectively. It is because of the difference between the energy of error and how the listeners perceived it. For example, there are various revealed psychoacoustic characteristics or effects including perceptual scale of frequency [18] , frequency dependent absolute threshold of hearing (ATH) or loudness [19, 20], auditory masking [19], and missing fundamental [21]. In particular, when the application provides the separated sources to users for being played, the subjective quality of SVS can be more important than the numerical one.

Two different approach for evaluating the subjective quality is possible, and the first one is to ask human directly. For example, Emiya et al. conducted an evaluation protocol that asks users to address the following four tasks respectively [22]:

1) rate the global quality compared to the reference for each test signal;

2) rate the quality in terms of preservation of the target source in each test signal;

3) rate the quality in terms of suppression of other sources in each test signal;

4) rate the quality in terms of absence of additional artificial noise in each test signal.

Another approach is to use the evaluation algorithm that predicts the subjective scores. Various algorithms are presented for the specific domains, for instance, Perceptual evaluation of speech quality (PESQ) for speech signal. In case of source separation, Perceptual evaluation methods for audio source separation (PEASS), which consists of the overall, target-related, interference-

related, and artifacts-related perceptual score is the most popular algorithm [22].

Application-dependent evaluation

SVS can be used in variety of applications as discussed in Section 1.3. In this case, the performance of applications can be used to evaluate the separation quality since it is the main purpose of SVS. For example, the accuracy of melody extraction, or singing voice detection was used to measure the separation quality [23, 24, 25].

1.5 Topics of interest

The primary goal of this thesis is to develop a novel algorithm for SVS. To this end, the following sub-task is stated as shown in Fig. 1.3.

Characteristics Studying and finding the relevant characteristics for singing voice and accompaniment. It should be able to represent and distinguish the classes, as well as be applied easily to the objective function for SVS. In particular, this thesis mainly focuses on the three characteristics, which are continuity, low-rankness, and sparsity.

Objective function Developing the objective function that represents SVS task. The function will be considered relevant when its optimal solutions correspond to the target singing voice and accompaniment. Two different approaches are tried in this thesis, that one is based on the continuity and sparsity, while another one is based on the low-rankness and sparsity.

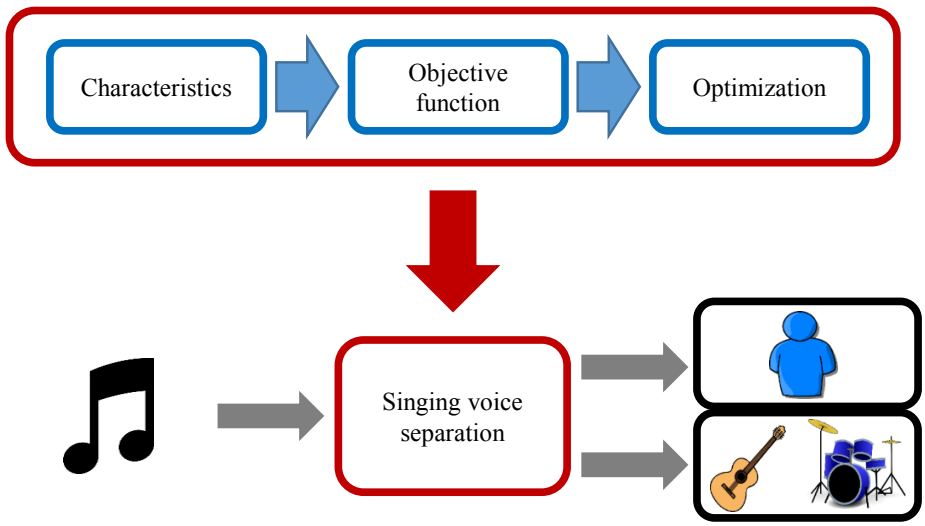

Fig. 1.3 Topics of interest in this paper. Developing SVS algorithm, which is the primary goal of this thesis, can be decomposed into three subtasks: Finding characteristics, deriving objective function using the characteristics, and solving the objective function using relevant optimization method.

Optimization Deriving the optimization method for the presented objective function. The method should minimize the objective function efficiently, which means fast computational speed and low memory usage. In this thesis, convex optimization methods are applied to the algorithms, including the augmented Lagrangian multiplier (ALM) and auxiliary function method.

1.6 Outline of the thesis

Chapter 2 describes the background of SVS. First, previous studies and algorithms for SVS are briefly described. These are categorized into four groups, which are characteristics-based, spatial, machine learning-based, and informed approach. In addition, useful information for studying SVS is introduced including public datasets and challenges. Finally, several evaluation criteria to measure the separation quality is explained.

Chapter 3 consists of the discussions about the characteristics of singing voice and accompaniment. Three different characteristics – continuity, low-rankness, and sparsity – are discussed, including what those are and how singing voice and accompaniment are different in terms of those characteristics.

Chapter 4 describes the SVS approach which is based on the continuity and sparsity. The conventional algorithm, which separates singing voice by using harmonic-percussive sound separation twice, is introduced. After that, the proposed algorithms using harmonic-percussive-residual sound separation is presented.

Chapter 5 describes another approach which is based on the low-rankness and sparsity. The conventional algorithm which is based on the robust principal component analysis (RPCA) is introduced, and the proposed algorithm tries to generalize or extend the conventional one. At first, RPCA which uses the nuclear and the l_1-norm is generalized to Schatten p-norm and l_p-norm, and even adding a proper scale compression step. In addition, another characteristics of singing

voice and accompaniment, which is called spectral distribution, is introduced, and it is applied to RPCA which is called weighted RPCA.

Chapter 2

Background

2.1 Spectrogram-domain separation framework

Various audio source separation algorithms share the similar framework, which we called 'spectrogram-domain separation framework'. Fig.2.1 shows the overall flows of it and below is a brief explanation about it.

Time-frequency representation is a relevant alternate domain to analyze a time signal. Short-time Fourier transform (STFT), which takes discrete Fourier transform with sliding window is one of the most popular approach. A STFT of x, $\boldsymbol{X}$, is as follows:

$$\boldsymbol{X}(f,t) = \sum_{n} x_t(n) e^{\frac{-j2\pi fn}{N}}, \tag{2.1}$$

where $x_t(n) = w(n)x(Wt+n)$. N and W denote the size and hop size of window, respectively. w is a windowing function, such as hamming or hanning function. Since STFT is a linear operation, Additivity in (1.2) is hold as follows:

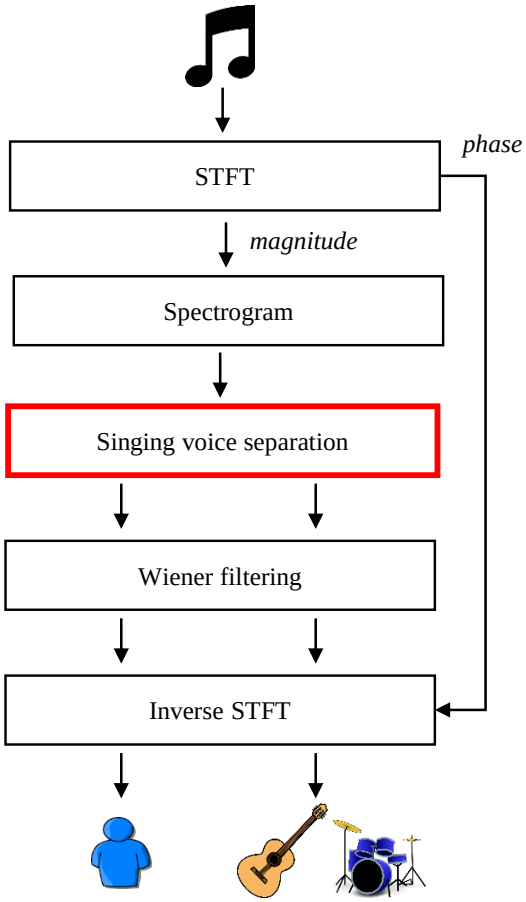

Fig. 2.1 Framework of spectrogram-domain singing voice separation

$$\boldsymbol{M}(f,t) = \boldsymbol{V}(f,t) + \boldsymbol{A}(f,t). \tag{2.2}$$

However, finding the characteristics of singing voice and accompaniment in STFT domain is difficult because of phase which occurs irregularly. Therefore, various audio source separation is done by discarding phase and remaining mag-

nitude only as $|\boldsymbol{M}|$, which is called spectrogram. In the spectrogram-domain approach, it is often assumed that the additivity of (1.2) and (2.2) is approximately hold in spectrogram domain as follows:

$$|\boldsymbol{M}(f,t)|^2 = |\boldsymbol{V}(f,t)|^2 + |\boldsymbol{A}(f,t)|^2. \tag{2.3}$$

More generalized assumption is used in some studies as follows:

$$|\boldsymbol{M}(f,t)|^{2\gamma} = |\boldsymbol{V}(f,t)|^{2\gamma} + |\boldsymbol{A}(f,t)|^{2\gamma}, \tag{2.4}$$

where γ is in the range of $(0,1]$ and denotes the scale compression parameter. For convenience, in the rest of the thesis $|\boldsymbol{X}|^{2\gamma}$ with proper γ is simplified as X. Fig. 2.2 shows an example of spectrogram of audio signal which is singing voice.

The outputs using spectrogram-based separation framework are also spectrograms. However, this separated spectrogram has no information about phase, thus it cannot be reconstruct the separated time signal. In addition, the approximation of (2.3) and (2.4) is not precise, so there may be errors after the separation. To compensate these two problems, soft masking or Wiener-like filtering is usually applied as a postprocessing step for source separation. It is performed as follows:

$$\boldsymbol{V}(f,t) = \frac{V(f,t)}{V(f,t)+A(f,t)}\boldsymbol{M}(f,t), \tag{2.5}$$

$$\boldsymbol{A}(f,t) = \frac{A(f,t)}{V(f,t)+A(f,t)}\boldsymbol{M}(f,t), \tag{2.6}$$

The time-domain signals of the sources are then reconstructed from V and A by using inverse STFT.

There are a number of modified version of above framework as follows.

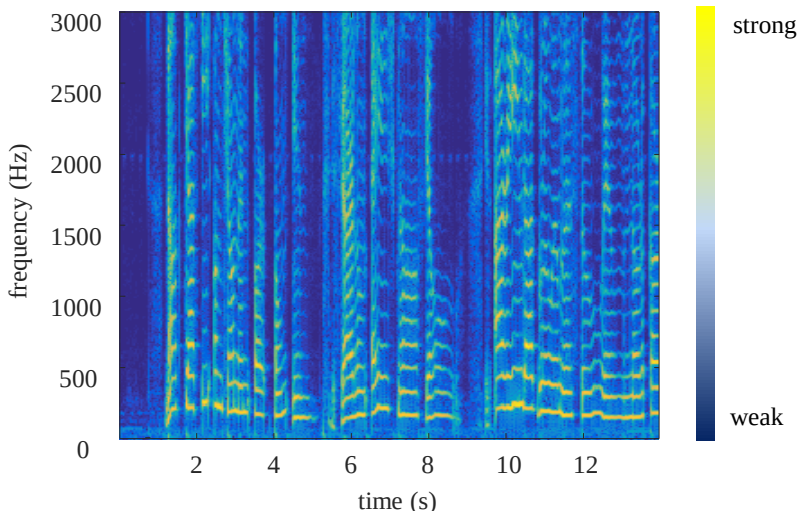

Fig. 2.2 An example of spectrogram of singing voice signal. The spectrogram is zoomed to specific time-frequency range, and represented in log-scale for visual convenience.

Spectrogram is replaced to the other time-frequency representation methods such as constant-Q transform [26]. On the other hand, normalization technique such as spectral standardization or principal component analysis can be used when SVS algorithm is based on the machine learning-based approach [27].

Wiener filtering can be omitted or replaced to other masking techniques such as binary masking [28].

Single stage framework , that separates singing voice and accompaniment from mixture directly, can be modified to have multiple separation stage. In this case, the sources are gradually enhanced via each separation stage, and the

latter separation algorithm uses the enhanced sources from the output of the former algorithm. Belows are the example of multiple stage separation frameworks

– Singing voice activity detection (VAD) → RPCA [29]
– RPCA → f_0 detection [24]
– Weighted RPCA (wRPCA) → VAD → wRPCA with updated weight [30]
– Deep neural network (DNN) → spatial estimation → DNN [27]

2.2 Approaches for singing voice separation

Conventional approaches and algorithms for SVS are introduced in this section. In particular, the algorithm is grouped into four approaches, which are named to characteristics-based, spatial, machine learning-based, and informed approach. Characteristics-based SVS is an approach that estimates the sources based on the characteristics nature of them. It can be considered as a fundamental approach even for the other extended ones, and it is the main aim of this thesis. Spatial approach, in addition to the source characteristics, exploits the mixing characteristics of sources which would represents the locations or the room condition. Simple algorithms such as spatial filtering which assumed that singing voice is always located in center also shows a remarkable separation results, and more complex ones tries to estimates the mixing or unmixing matrix. In case of machine learning-based approach, it also estimates the source characteristics also but by using the training data.

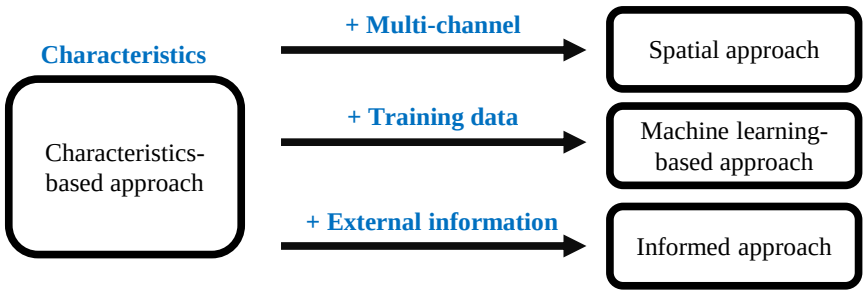

Fig. 2.3 General approaches of singing voice separation.

2.2.1 Characteristics-based approach

Fundamental frequency-based approach

SVS based on fundamental frequency (f_0) is one of the most traditional approach [31]. It is based on the observation that singing voice is the most predominant source in the most of music signal. In addition, the spectrum of singing voice has a strong harmonic structure, with a few exceptions of unvoiced sound. From these insights, f_0-based SVS basically detects the sequence of predominant f_0 from the music and extract the time-frequency coefficients which correspond to the harmonics of the sequence. Further studies have tried to improve f_0 detection algorithm, including preprocessing steps, to remove f_0 and its harmonics which is not corresponds to singing voice, or to handle the unvoiced singing voice [32]. However, f_0-based approach has several drawbacks: 1) Separation quality highly depends on the accuracy of f_0 detection algorithm. 2) It is not guaranteed that the detected f_0 is corresponding to singing voice. 3) It cannot separate the singing voice which is not corresponds to f_0, such as plosive sounds.

Repetition-based approach

Another remarkable approach for SVS is based on the repetitive nature of accompaniment [33, 34]. This is based on the observation that various musical sources tend to repeat over time, depending on its tempo or speed. It happened more especially in case of 'background' music sources, for example a drum loop or guitar riff, while it is relatively weak in 'foreground' sources, which is singing voice in general. REPET, which is based on these insights, separate singing voice as follows. First, the tempo of music is estimated and the time length of repetition is calculated based on it. Spectrogram is sliced to have the calculated time length, and the repeating accompaniment is estimated by taking median operation over the slices. The residuals in the spectrogram that cannot be represented by using the median of slice is considered as singing voice and separated. There are several algorithms that extend original REPET, including adaptive accompaniment estimation using moving-median [35], or based on similarity matrix [36].

2.2.2 Spatial approach

Spatial filtering

Most of popular music is produced in multi-channel format i.e. stereo. Although those music 'recordings' are in general not recorded in real-world but mixed in studio, spatial characteristics are often applied like real-world recording to provide the spatial impression in music. In the studio mixing procedure, singing voice is often located in the center, while other instruments are located widely by using panning.

Although this approach is cannot be verified because the direction of source image indeed depends on the favor of music producer, but it is practically useful and applied to many real-world music player or editor, such as audacity.

(Un)mixing matrix estimation

When capturing a sound by a microphone, it is considered that the captured signal consists the information about original sound image and the spatial characteristics, which is related with the path between the source and the microphone. On the other hand, even if the music is mixed by the producer, he can apply various spatial effects to the sources so that they can be perceived as if they are in a specific location. While a monaural recording is difficult to be decomposed into the original source and spatial effects, it can be tried in multi-channel situation since all the microphones record the same source images, but at the different relative path. Independent component analysis is one of the most popular algorithm in this approach for the blind source separation problem [37].

In case of music source separation or even SVS, prior information for the source images is often combined with spatial information. Ozerov et al. extended the conventional nonnegative matrix factorization (NMF) to deal with the multichannel signal [38], and Nugraha et al. used DNN to estimate the spectral distribution of the source images [27].

2.2.3 Machine learning-based approach

Machine learning-based source separation also can be considered as an approach that uses the source characteristics. However, it estimates the characteristics

model by learning from the training data, while the characteristics-based approach in general derives the objective function from the observation.

Nonnegative matrix factorization

NMF is an algorithm to represent a nonnegative matrix as a multiplication of two (or more) nonnegative matrix. It is widely applied to music source separation algorithms, because the nonnegative assumption is well-suited for music spectrogram which is nonnegative and also can be approximated as a combination of nonnegative source spectrograms [4, 39, 40].

Deep learning

In the recent years, 'deep learning' approaches have achieved remarkable performance in most of machine learning tasks, including recognition of image, speech, or video, generating artistic images or music, and even mastering the game of Go. It also has been applied to the music source separation and its separation quality outperforms that of the other conventional approaches. Remarkably, algorithms using deep learning approaches ranked on top in SiSEC 2016 and SiSEC 2015 [41, 42], while none of the submitted algorithms used deep learning in SiSEC 2013 [43], which was held just before 2015.

2.2.4 informed approach

One may expect that the separation quality would be improved when the additional information is provided in the separation procedure, compared to when using the waveform only. Informed source separation is an approach to reveal which information can be applied and how to.

Score of an instrument provides the information about when when is played in which notes and how long it is continued. Since it is highly related with f_0 trajectory, the audio signal corresponds to the score can be separated basically by extracting the harmonic components of f_0. Various algorithms for music source separation have been presented [44, 45, 46], and even for SVS [47].

Lyrics of music represent the phonemes of singing voice. Since only singing voice can have lyrics in music, it is expected to be a key information for separation. Most studies on SVS or speech separation use an example-based approach that derives the source closer to the signal synthesized in the text or lyrics [48, 49].

User-guided information can be various depending on the applications. For example, users can guide the melody of singing voice by humming it. In that case the SVS algorithm detects the melody sequence or f_0 of humming, then separates them and those harmonic components [50]. On the other hand, user can guide directly by annotating where the sources exist or not on the spectrogram. Although it requires to users the knowledge about spectrogram, but ideally they can provides the perfectly aligned information in ideal. In addition, separation results can be improved by interacting with the user [51, 52, 53, 54, 55, 56].

2.3 Datasets and challenges

2.3.1 Datasets

MIR-1K dataset consists of 110 Chinese pop songs [32]. Singing voice and accompaniment are recorded in the right and left channel, respectively, with 16,000Hz sample rate. Each music is divided into segments with 4 to 13 seconds time duration and the number of segment is 1000. The people who sang the singing voice track are 8 females and 11 males, and not professional singers. In case of accompaniment track, it was played by using karaoke-style virtual instruments. MIR-1K dataset is one of the first dataset for SVS task that released in public, and has reasonable size for evaluate the algorithms. However, it is not enough for the machine learning approaches, especially deep learning approaches which requires training data. In addition, the genre or style in dataset is slightly biased, and the quality of music such as sample rate, singing skills, or karaoke-style accompaniments is far from real music.

iKala dataset is similar to the MIR-1K dataset but with better quality [29]. In consists of 252 30-second music clips which are excerpted from 206 musics. 100 additional excerpts are not released but reserved for MIREX. Each source is recorded separately as MIR-1K dataset, but with 44,100Hz sample rate. For the singing voice, six professional singers were hired to sing the songs. The dataset is not publicly disclosed but is provided after the license agreement with the exception of the clips for MIREX.

Beach boys dataset denotes a set of music recordings collected from the album *Good Vibrations: Thirty Years of the Beach Boys* and *The Pet Sounds Ses-*

sions by *Beach Boys* which are released in 1993 and 1997, respectively [57, 58]. For each recordings, singing voice and accompaniment are provided separately as split stereo recording (stereo format where one channel is singing voice and the another one is accompaniment), or as two different recordings (a cappella and instrumental). To use the dataset for SVS experiments, two channel or recording is mixed and SVS algorithms tried to recover the original sources. Although Beach boys dataset is valuable since it contains actual popular music recordings which achieved huge success, the number of recording is small (5 [57] and 10 [58], and 1 recording is duplicated) and all is from the same artist.

MSD100 and DSD100 dataset consists of 100 music and these are split to development and test set. Each music consists of 4 recording, which are vocal (singing voice), drum, bass, and others. When using it for the experiments of SVS, the sum of drum, bass and others are considered as accompaniment. The difference between MSD100 and DSD 100 is that the recordings in DSD100 is scaled by professional music producer to be similar as real-world popular music [41, 42] .

2.3.2 Challenges

A number of challenges about SVS were held to encourage researchers to develop and share there algorithm.

MIREX or music information retrieval evaluation exchange is an annual challenge that consists of various tasks related with music information retrieval problem. SVS was included as a subtask of MIREX since 2014. The partici-

pants are asked to submit the source code of their SVS algorithm, then it is evaluated by organizer by using iKala dataset.

SiSEC or signal separation evaluation campaign was held every one and half year and consists of various source separation problems. Music source separation is a subtask of campaign, which aims to separate the mixture to four sources (vocals, drums, bass, and others). Because it is not mandatory to separated all the individual sources, the task is useful even for SVS that separates singing voice (vocals) and accompaniment (sum of all other sources).

Chapter 3

Characteristics of music sources

3.1 Introduction

The characteristics of singing voice and accompaniment are discussed in this chapter. Because characteristics-based SVS algorithms do not use any machine learning approach to characterize the sources, it is required to find the characteristics which represent each source well and even distinguish each based on assumption and/or observation. In addition, these should be able to be represented in a mathematical format to so can be derived into objective function for SVS. Therefore, appropriate characteristics should be able to lead the objective function whose optimal solution corresponds to the separated sources.

3.2 Spectral/temporal continuity

3.2.1 Continuity of a spectrogram

Continuity of a spectrogram denotes the similarity of its coefficients with its neighbor ones, in other words, how smooth the spectrogram is. Since a spectrogram has a form with two dimension of frequency and time, continuity can be individually considered for each dimension. If coefficients in a spectrogram is highly similar with its neighbor to the frequency axis, then it can be said that it has a high spectral continuity. On the other hand, it has a high temporal continuity if its neighbor coefficients to the time axis are similar.

An audio signal which can be represented as a spectrogram with a high a spectral/temporal continuity may be expected to be sounds as follows. If a sound has a salient f_0 with a strong harmonic structure, then its spectral continuity may be relatively low because a coefficient which belong to f_0 harmonic may have large value, while its neighbor which is not belong to f0 harmonic is small. Therefore, it is expected that a spectrogram with high spectral continuity has broadband spectra. In case of temporal continuity, a spectrogram with high temporal continuity is expected to consist of 'stable' sounds, which is sustained for long time and rarely changed.

Continuity of spectrogram can be measured by calculating the overall difference between neighbor coefficients. Here, a difference can be defined in various form but sum of square error is widely used thanks to its simplicity. Spectral continuity of a spectrogram X, $C_f(X)$ is defined as follows:

$$C_f(X) = -\sum_{f,t}(X_{f,t} - X_{f-1,t})^2. \tag{3.1}$$

Similarly, its temporal continuity $C_t(X)$ is defined as follows:

$$C_t(X) = -\sum_{f,t}(X_{f,t} - X_{f,t-1})^2. \tag{3.2}$$

In (3.1) and (3.2), the higher C_f and C_t means the higher continuity.

3.2.2 Continuity of musical sources

Each musical source has different degree of continuity. In case of percussive instruments, such as drums, these have mostly unpitched broadband sound.[1] On the other hand, since a percussive sound is occurred by a single hit of instrument, it instantly attenuate and therefore it has a short sustain time. Consequently, percussive sound has relatively high spectral continuity but low temporal continuity.

Harmonic instruments has opposite characteristics compared to percussive ones in terms of continuity. Most of harmonic instruments have pitched sound that have strong harmonic structure, which is expected to have a low spectral continuity. On the other hand, once a harmonic sound is played it continued during the respective note length, so it is expected to have relatively stable temporal characteristics, at least compared to the percussive one. Therefore, harmonic sound has relatively low spectral continuity but high temporal continuity.

Singing voice also has different continuity characteristic compared to harmonic or percussive instruments. Moreover, due to complex characteristics of

[1]There are also various pitched percussive instruments like glockenspiel. In this thesis, however, percussive and harmonic instruments are distinguished based on those musical roles: rhythmic or harmonic. Therefore pitched percussive instruments are also classified as harmonic instruments. In addition, it is noted that the main purpose of defining percussive and harmonic instrument in this thesis is to verify that the singing voice is hardly grouped into any of them.

singing voice, its continuity is needed to be discussed with respective spectral/temporal resolution of spectrogram. First, although it basically depends on the syllables in lyrics, singing voice in general has strong harmonic structure and f_0 which is belongs to the note frequency. However, when large window size is used for spectrogram, this spectral continuity goes stronger because the neighbors of harmonic coefficients in spectrum also activated in a analysis window because of vibrato of singing voice. With a fixed analysis window size, it is expected that the spectral continuity of singing voice is weaker than percussive instruments but stronger than harmonic instruments.

In case of temporal continuity, although singing voice is sounded based on the note length as harmonic instruments, it is relatively less 'stable' over time due to the fast tremolo or vibrato of singing voice. If the analysis window size goes smaller, enough to be faster than this unstable activity, then the spectrogram may be more stable locally. With a fixed analysis window size, it is expected that the temporal continuity of singing voice is stronger that percussive instruments but weaker than harmonic instruments.

Fig. 3.1 shows the example spectrograms of sources, and its simple representation emphasizing its continuity.

3.3 Low-rankness

3.3.1 Low-rankness of a spectrogram

Given a matrix X, its rank is the maximum number of linearly independent columns of X. Representing or approximating data as a low-rank matrix is widely applied in many applications, including data encoding or denoising in image signal processing [59]. Since a spectrogram of audio signal is also a two-

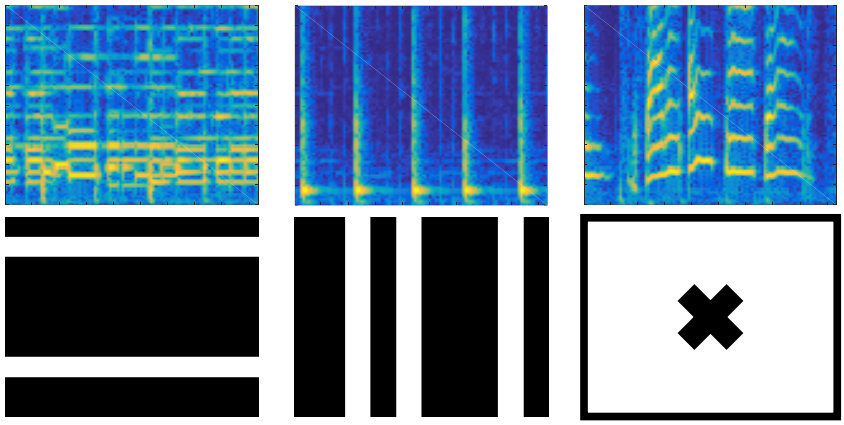

Fig. 3.1 Comparison of continuity in (a) harmonic instruments, (b) percussive instruments, and (c) singing voice. Top row is the excerpts of spectrogram, and bottom row is their simplified representation as ridges. It is noted that singing voice cannot be represented neither horizontal nor vertical ridges.

dimensional matrix, its rank is useful to analyze its characteristics.

The low-rankness of a matrix X, $L(X)$, can be simply represented based on its rank as follows:

$$L(x) = -\text{rank}(X), \tag{3.3}$$

where the higher $L(X)$ represents the more low-rankness. However, it is difficult to use (3.3) directly for real-world applications. First, most of data captured in real-world, even audio spectrograms, are full-rank due to its randomness and noise. In addition, the rank minimization problem is known to be NP-hard [60]. To overcome these problems, the nuclear norm is often used as a approximation of rank. the nuclear norm of X, $||X||_*$, is defined by

$$||X||_* = \sum_i \sigma_i, \tag{3.4}$$

where σ_i is the i-th singular value. it is noted that the nuclear norm is equivalent

to the l_1-norm of the singular values, while the rank, which is the number of non-zero singular values, is equivalent to the l_0-norm of them. Therefore, the low-rankness of (3.3) can be approximated as

$$L(X) = -||X||_*. \tag{3.5}$$

Practically, the rank or nuclear norm of a spectrogram of the audio signal is related with the diversity of sounds in the signal. If it consists of many unique sounds then it cannot be represented as a combination of a few number of the spectra, and even of the orthogonal vectors, thus it leads to have higher rank. On the contrary, if a few number of sounds are occurs repetitively in the signal, its spectrogram may have low rank.

However, above discussion is not always hold, especially when excessively many sounds occur in the signal. Because the sounds in the signal is not always orthogonal, it is possible to approximately represent many spectra using a few number of vectors. Moreover, the spectrogram of a mixture which consists of excessively many sounds tends to be 'blurred', and can be represented as a low-rank matrix.

3.3.2 Low-rankness of musical sources

To discuss about low-rankness of singing voice and accompaniment, the following characteristics of each have to be considered. At first, if there are many sound elements in a source, then in general its spectrogram could be expected to be a high-lank matrix. However, if those sound elements are similar and can be approximated using a few spectra, then there still a possibility of low-rankness. In addition, if many sound elements are simultaneously occurred then the ob-

tained mixture spectrum can be 'blurred' and it also can be approximated by using those smoothed spectrum model. Singing voice and accompaniment both has a characteristics that leads to be low-rank or high-rank. Below is a detailed discussion about those characteristics.

In case of accompaniment, every note produced by instruments can be considered as a unique sound elements. However, the number of these elements is quite limited since most of musical accompaniment is composed using limited number of instruments and notes. Instead those elements are frequently reproduced over the whole track, thus its spectrogram can be easily represented as a combination of a few number of spectra, or unique vectors.

On the contrary, there are plenty of variation in singing voice, including the singer characteristics (gender, age, singing style, etc.) and the pronunciation of lyrics. In addition, the unit source of singing voice is rarely mixed since there are in general one or a few number of singers in a music track. Therefore, it is a reasonable conclusion that a spectrogram of singing voice may have high rank.

Fig. 3.2 shows the example spectrogram of singing voice and accompaniment, and Fig. 3.3 shows the singular values of the example singing voice and accompaniment. Comparing singing voice and accompaniment in Fig. 3.3, the most of energy is concentrated in a few number of singular values in case of accompaniment.

3.4 Sparsity

3.4.1 Sparsity of a spectrogram

Sparsity of a matrix is the contrast concept with density of it. If the most of elements in a matrix is zero, than it is called as a sparse matrix. When the

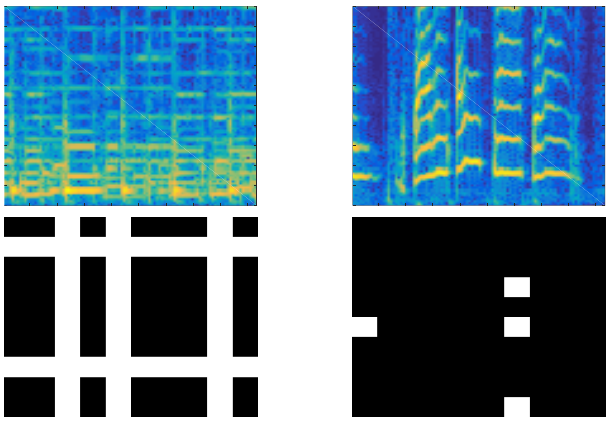

Fig. 3.2 Comparison of low-rankness and sparsity in (a) accompaniment and (b) singing voice. Top row is the excerpts of spectrogram, and bottom row is their simplified binary representation.

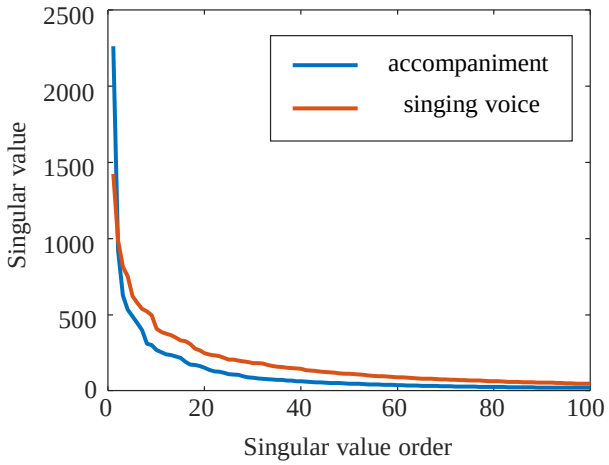

Fig. 3.3 Singular value distribution of accompaniment and singing voice. These are computed from the magnitude spectrogram after normalization. First 100 singular values are represented for visual convenience.

matrix has a unit variance, then it also can be considered that most of energy of a matrix is concentrated in a few elements.

In case of audio signal, the value of a coefficient in a spectrogram represents the energy of in the respective time-frequency region, thus if a spectrogram is sparse then it means that sounds mainly occur in specific time frame and/or frequency bin, while there are almost silence in the other time or frequency.

Assuming there are many unique sounds with independent time-frequency distribution which an audio signal consist of, then this audio signal tends to have a Gaussian distribution due to the central limit theorem. In other words, if a signal consist of a few unique sounds, it is expected that it may have a distribution which is far from the Gaussian. Since a unique sound in general have very sparse distribution because it rarely occurs over time, here 'far from the Gaussian' means far to be sparse rather than far to be dense.

Ideally the sparsity of a matrix X, $S(X)$, is represented by calculating the number of nonzero coefficients in a matrix as follows.

$$S(x) = -\text{nonzero}(X) \tag{3.6}$$

where nonzero denotes the number of nonzero coefficients. As the low-rankness, it is also often relaxed using l_1-norm as follows:

$$S(X) = -\sum_{f,t} |X_{f,t}| \tag{3.7}$$

It is noted that l_1-norm can be considered as a convex relaxation of l_0-norm, which is equivalent to nonzero.

3.4.2 Sparsity of musical sources

A sparsity of musical source is related with the following factors. First, if the source has strong harmonic structure, it is sparser than the source with broad-

band spectra because most of energy is concentrated in a few harmonic coefficients. On the other hand, if the source is rarely played in music or has short sustain time, it is expected to be sparse since there are many silence regions in spectrogram which are closed to zero. Finally, the source which consist of many instruments or is played in polyphonic, it is in general less sparse because the coefficients in spectrogram tends to have a Gaussian distribution due to the central limit theorem.

From the above discussion, it is expected that singing voice has sparser distribution compared to accompaniment: it has strong harmonic structure while percussive instruments in accompaniment do not have, there are part without singing voice in music especially between verses or lines, when accompaniment is still played, and there are in usual one or a few number of singers in a music recording and they sing in monophonic, while accompaniment consists of various polyphonic instruments.

If we compare singing voice with a single instrument instead of accompaniment, there can be other instruments which has sparser distribution than singing voice. When there is an instrument or a sound effect which is played only once in a music, obviously it is much sparser than singing voice. However, this kind of instruments has not only has high sparsity, but also has low-rankness those silence does not increase its rank. Therefore, singing voice can be considered as one of the instrument with the highest sparsity compared to its low-rankness.

3.5 Experiments

This this section, the above discussions about the characteristics of musical sources are empirically verified from the experiments using the actual dataset. We used the development data in DSD100 dataset, which consists of 50 music tracks with 44.1kHz sample rate. Each track contains four stereo sources, which are vocals, drum, bass, and others.

As a preprocessing step, we first remixed or redefined the provided sources to be harmonic instruments (bass+others), percussive instruments (drum), accompaniment (drum+bass+others), and singing voice (vocals). The sources are then down-mixed to be mono by averaging two channels. Each tracks was split to 10 seconds segments, and the total number of segments is 1276. Magnitude spectrograms are obtained using STFT with 4096 window size (93ms for 44.1kHz) and 1024 shift.

When measuring (3.1), (3.2), (3.5), or (3.7), its actual value might be meaningless since it depends on not only its characteristics but also its scale, the number of unit sounds, or the proportion of silence legion. For the measurement that is invariant to these unintended conditions, we instead focused on the relations between those values. Two different experiments are discussed in this section, where the one compare the spectral and temporal continuity, while the another one compare the sparsity and low-rankness.

Fig. 3.4 shows the spectral/temporal continuity of harmonic instruments, singing voice, and percussive instruments, as well as the linear regression of respective sources. As expected, harmonic instruments tends to have the higher temporal continuity C_t compared to percussive ones, when fixing the spectral continuity C_f. Singing voice, on the other hand, showed the intermediate char-

acteristics between those two instruments. On the other hand, Fig. 3.5 shows the comparison of low-rankness and sparsity of singing voice and accompaniment and those linear regression. Although there exists some overlap between two sources, accompaniment tends to have higher low-rankness in the fixed sparsity.

3.6 Summary

In this chapter, the characteristics of musical sources are discovered in the spectrogram domain. Three characteristics were in particular focused, which are spectral/temporal continuity, low-rankness, and sparsity. These were first defined for a matrix or a general audio spectrogram, and the equations for the measurements were also introduced. After that, the differences between singing voice and accompaniment, in aspects of these characteristics, are discussed.

Table 3.1 shows the summary of discussion in this chapter. Because of the unstable and unrepeated patterns of singing voice, it does not have continuous or low-rank characteristics in general, but has sparse distribution. In case of accompaniment spectrogram, in consists of many instruments which frequently reproduce the same sound and it leads to have low-rankness but not sparsity. The instruments in accompaniment can be categorized into to groups, which are harmonic instruments with high temporal continuity and percussive instruments with high spectral continuity.

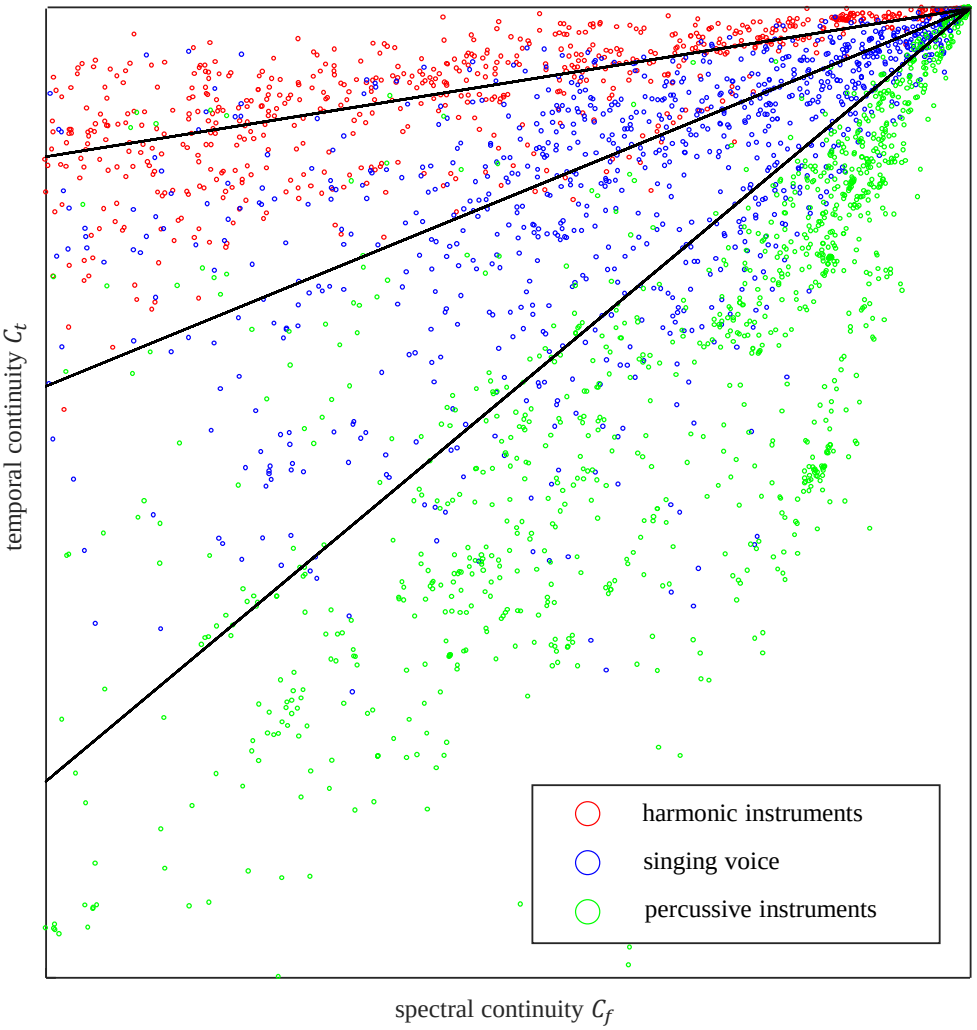

Fig. 3.4 Visualization of spectral/temporal continuity of the sources. Black lines represent the linear regression with zero offset of harmonic instruments, singing voice, and percussive instruments, from top to bottom. Each line has the regression coefficient of 0.153, 0.389, and 0.797, respectively.

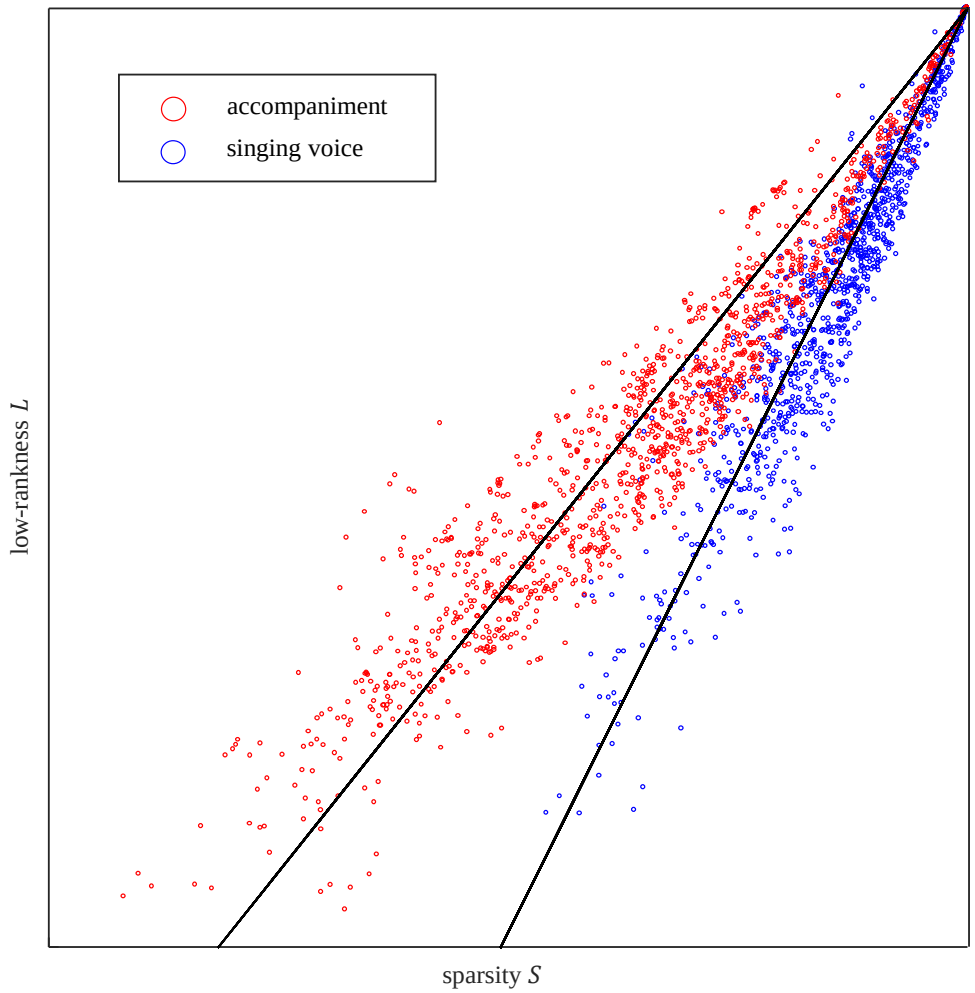

Fig. 3.5 Visualization of sparsity and low-rankness of the sources. Black lines represent the linear regression with zero offset of accompaniment and singing voice, from top to bottom. Each line has the regression coefficient of 0.025 and 0.039, respectively.

Table 3.1 Summary of the characteristics comparison between singing voice and accompaniment. Blue H and red L denote that the source has high or low characteristics respectively.

		Singing voice	Accompaniment	
			Harmonic	Percussive
Continuity	Spectral	L	L	H
	Temporal	L	H	L
Low-rankness		L	H	
Sparsity		H	L	

Chapter 4

Singing voice separation using continuity and sparsity

4.1 Introduction

In this chapter, based on the original work by Jeong et al. [61], an approach for SVS using spectral/temporal continuity and sparsity is explained. As discussed in Chapter 3, harmonic and percussive sounds can be distinguished in terms of spectral/temporal continuity. Harmonic sound, which has strong harmonic structure and long sustain time, has low spectral continuity and high temporal continuity. On the other hand, percussive sounds have relatively broadband spectra and short sustain time, thus it leads to high spectral continuity and low temporal continuity. Singing voice, which has strong harmonic structure but also has unstable temporal dynamics, is closer to harmonic sounds than to percussive ones, although it is also quite percussive compared to the other harmonic instruments. From these observations, separation of singing voice has

been tried to define the another source between harmonic and percussive instruments, and extract it by extending HPSS algorithms.

Two-stage HPSS framework is the one of the most widely applied approach for SVS. At the first HPSS stage, the music signal is decomposed into two tracks, which contain harmonic and percussive sound, and singing voice is included in harmonic sound. The separated harmonic sound, which is in fact also contains singing voice, is decomposed again in the second HPSS stage with different time-frequency resolutions. In this stage the singing voice is separated by considering it as percussive sound. Various HPSS algorithms have been tried to be applied to SVS. Tachibana et al. used temporal/spectral continuity-based HPSS [5, 14, 23], FitzGerald et al. used a median filtering and matrix factorization approach [6, 26], and Zhu et al. used NMF and basis selection [62].

Our proposed algorithm is also based on the observation that singing voice is neither exactly harmonic nor percussive. Instead of using the HPSS twice in a cascaded way, we formulate the vocal separation problem in a single optimization framework using additional constraints that allow the residual in the HPSS process but forces it to be sparse and nonnegative.

The rest of this chapter is organized as follows. In Section 2, an algorithm of SVS using two-stage HPSS, which gave a motivation for the proposed algorithm, is described. In Section 3, the proposed algorithm is described, where we define the objective function for optimization and present the derivation of the update rule, including the pseudocode. In Section 4, we present the experimental results and discussion, followed by conclusions and directions for future work in Section 5.

4.2 SVS using two-stage HPSS

This section introduced the conventional SVS algorithm using two-stage HPSS. From various algorithms based on a similar approach, The work of Tachibana et al. that the proposed algorithm directly aims to extend is introduced [23].

4.2.1 Harmonic-percussive sound separation

In a music signal, the harmonic and percussive components usually have distinctive characteristics. The harmonic sounds generally have a very strong harmonic structure, and the sustain time is relatively long, resulting in parallel, horizontal ridges in the spectrogram. On the other hand, the percussive sounds are very short and broadband, and therefore shown as vertical ridges in the spectrogram. Based on this observation, Ono et al. proposed an algorithm to separate the harmonic and percussive components from the spectrogram by minimizing the temporal/spectral gradients of the separated spectrograms to enhance the horizontal/vertical ridges [5, 14]. By approximating that the spectrogram of music is same as a sum of spectrogram of harmonic and percussive sounds, it can be represented as follows:

$$M = |\boldsymbol{M}|^{2\gamma} = H + P, \tag{4.1}$$

where $\boldsymbol{M}$ denotes the STFT of music signal, and $M \in \mathbb{R}^{F \times T}$, $H \in \mathbb{R}^{F \times T}$, and $P \in \mathbb{R}^{F \times T}$ denote the spectrogram of mixture, harmonic instruments, and percussive instruments, respectively. F and T denote the number of frequency bins and time frame, respectively.γ is a parameter to compress the original magnitude spectrogram, to emphasize the difference of two sources in terms of

continuity. Applying the continuity model in Chapter 3, Ono et al. proposed the following objective function to separate each source by maximizing the spectral/temporal continuity of percussive/harmonic sounds, respectively [5, 14]:

$$\begin{aligned} J(H,P) &= \frac{1}{2}\sum_{f,t}(H_{f,t}-H_{f,t-1})^2 + \frac{\alpha}{2}\sum_{f,t}(P_{f,t}-P_{f-1,t})^2, \\ s.t. \quad & H+P=M, \quad H\geq 0, \quad P\geq 0, \end{aligned} \tag{4.2}$$

where $X_{f,t}$ denotes the (f,t)-th coefficient of X, and α denotes a weight parameter between the spectral and temporal continuity. The non-negativity constraint is to make the separated H and P to be a spectrogram. On the other hand, they also presented the variation of (4.2) whose equality between a mixture M and a sum of H and P is relaxed by using Kullback–Leibler (KL) divergence as follows:

$$\begin{aligned} J(H,P) = &\frac{1}{2}\sum_{f,t}(H_{f,t}-H_{f,t-1})^2 + \frac{\alpha}{2}\sum_{f,t}(P_{f,t}-P_{f-1,t})^2 \\ &+\sum_{f,t}M_{f,t}\ln\frac{M_{f,t}}{H_{f,t}+P_{f,t}} - M_{f,t} + (H_{f,t}+P_{f,t}), \\ & H\geq 0, \quad P\geq 0. \end{aligned} \tag{4.3}$$

Besides above ones, various algorithms for HPSS have been presented. For example, median filtering or NMF was applied for HPSS [6, 63].

4.2.2 SVS using two-stage HPSS

Meanwhile, the characteristics of a singing voice signal are very unique; thus, it is difficult to classify it exclusively into harmonic or percussive components. Even though singing voice signals contain a strong harmonic structure unlike

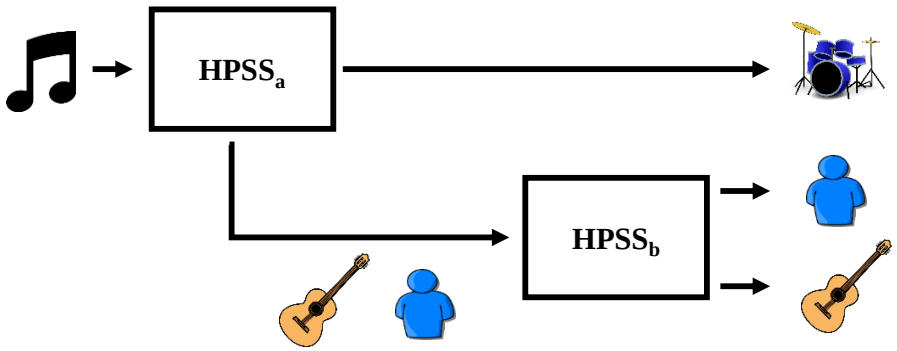

Fig. 4.1 Framework for SVS using two-stage HPSS. HPSSa and HPSSb denote HPSS with respective parameter settings.

percussive instruments, it reveals, at the same time, temporally unstable properties that are distinct from the harmonic ones. Practically, singing voice components are usually shown as horizontal but rapidly changing harmonic ridges in the spectrogram, thus it can be grouped into harmonic or percussive sounds depending on the time-frequency resolution.

Based on these discussion, this resolution can be parameterized for the HPSS algorithm to induce singing voice to be separated harmonics or percussive sounds. In addition, when the music signal is decomposed by using HPSS into harmonic instruments with singing voice and percussive instruments, then the former one can be separated again into harmonic instruments and singing voice by using HPSS again with different parameters. From these discussions, SVS algorithm using two-stage HPSS was presented by Tachibana et al., and Fig. 4.1 shows the framework of the algorithm [23].

Since the core insight for this SVS framework is to use HPSS twice, it is possible the other HPSS methods can be implied. Fitzgarald et al. presented a similar approach but using median filtering-based HPSS [26].

4.3 Proposed algorithm

Improving the abovementioned SVS algorithm using two-stage SVS, the single-stage SVS using harmonic-percussive-sparse separation algorithm is introduced in this section.

Although singing voice has a unique characteristics that is different from other instruments in terms of continuity, it is difficult to represent the numeric threshold to distinguish it. Moreover, the sources cannot be classified precisely due to the overlap between them as in Fig. 3.4.

Instead of representing the singing voice as moderate continuities, the proposed SVS algorithm uses another characteristics, sparsity, to distinguish it from the other instruments. In addition, instead of two-stage framework it separates all the sources by solving a single objective function. Fig. 4.2 shows the framework of proposed algorithm. Assuming there are percussive and harmonic instruments as well as singing voice in a music signal, its spectrogram can be represented as a combination of the vertical and horizontal ridges, and sparse components which are not continuous. Therefore, it is expected that the sources can be separated by obtaining the ridges and components from it.

The proposed method is also similar with the other conventional SVS algorithms, which described a singing voice signal as a residual that cannot be represented using an accompaniment model [28, 33, 34, 35, 64, 65]. Furthermore, a singing voice signal has a certain structure, which means that the energy of the singing voice is concentrated in a few time/frequency bins, and thus is often modeled using l_1-norm minimization in the spectrogram domain [28, 64].

Taking into account the unique properties of a singing voice signal that belongs to neither harmonic nor percussive sounds, we first assume that the

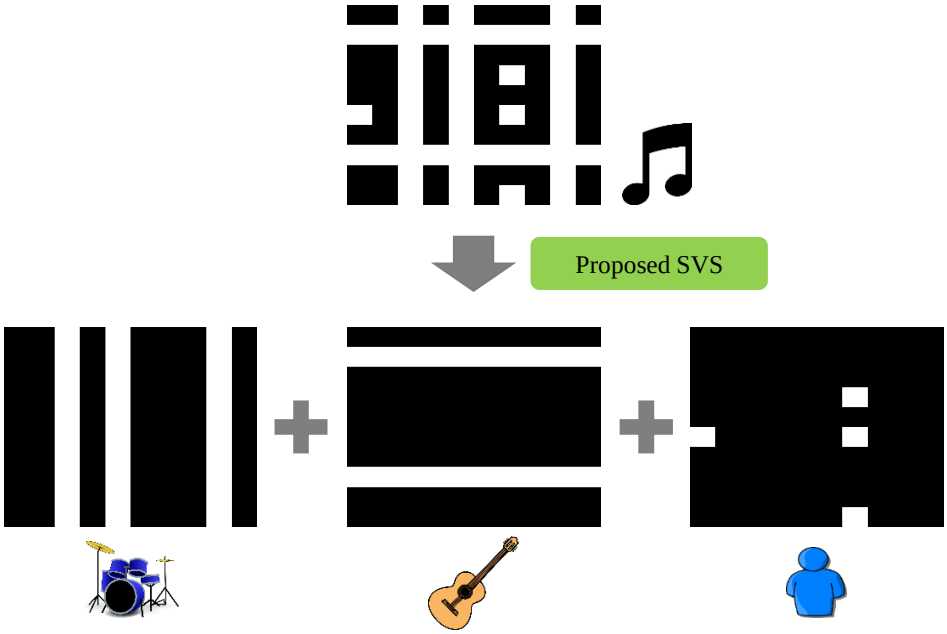

Fig. 4.2 Framework for the proposed algorithm.

spectrogram of the music signal can be approximately represented as a sum of the harmonic, percussive, and singing voice components as

$$M = |\boldsymbol{M}|^{2\gamma} = H + P + V, \tag{4.4}$$

where $\boldsymbol{M}$ is the STFT of an input music signal, and M, H, P, and V denote the scale-compressed spectrograms of the input, harmonic, percussive, and vocal components, respectively. $|\cdot|^{2\gamma}$ denotes the element-wise power operation, and the scale parameter γ, in the interval of (0, 1], denotes the compression rate, as presented by Ono et al. [5]. In this paper, we empirically set to be 0.25.

Based on the abovementioned characteristics of each component, we then derive objective function J to separate the singing voice and the accompaniment

as follows:

$$J(H,P,V) = \frac{1}{2}\sum_{f,t}(H_{f,t} - H_{f,t-1})^2 + \frac{\alpha}{2}\sum_{f,t}(P_{f,t} - P_{f-1,t})^2 + \phi\sum_{f,t}|V_{f,t}|, \tag{4.5}$$

$$s.t. \quad H + P + V = M, \quad H \geq 0, \quad P \geq 0, \quad V \geq 0,$$

where f and t are the frequency and the time indices, respectively. Two parameters, α and ϕ, are used to control the relative weights among the objective terms; $\alpha > 0$ denotes the relative smoothness of P; and $\phi > 0$ determines the weight for the l_1-norm minimization of the singing voice. In particular, in order to guarantee the scale invariance of the objective function J, the value of ϕ should be decided in the form $\phi = kE$, where k is a constant, and $E = \frac{1}{N}\sum_{f,t}|W_{f,t}|$, where N denotes the number of coefficients in the spectrogram W. With this ϕ, the objective function $J(\beta H, \beta P, \beta V)$ with the scaled input βW, where $\beta > 0$ is a scale parameter, can be expressed as $\beta^2 J(H,P,V)$, which leads to the same separation results.

Basically, this objective function is similar to the conventional harmonic/percussive separation algorithm; the first and second terms are the same as the objective function in Ono's algorithm [5]. However, by adding the third term, which we want to be the vocal, to the objective function and by imposing the sparsity and nonnegativity constraints to this extra term, we formulate the vocal separation problem into a single optimization framework.

Here, we derive the iterative update rule to minimize the objective function. Assuming that the present H, P, and V satisfy the nonnegativity constraints, the absolute sign in the l_1-norm in the singing voice term can be ignored and

thus is modified as follows:

$$\phi \sum_{f,t} |V_{f,t}| = \phi \sum_{f,t} (M_{f,t} - H_{f,t} - P_{f,t}), \tag{4.6}$$

and the objective function $J(H, P, V)$ can be represented as $J(H, P)$ by using only H and P.

The differentiations of the objective function by H and P are given by

$$\begin{aligned} \frac{\partial J}{\partial H_{f,t}} &= (2H_{f,t} - H_{f,t+1} - H_{f,t-1}) - \phi, \\ \frac{\partial J}{\partial P_{f,t}} &= \alpha(2P_{f,t} - P_{f+1,t} - H_{f-1,t}) - \phi, \end{aligned} \tag{4.7}$$

respectively. With the other terms fixed, the optimal values that make the differentiations to be zero can easily be found as follows:

$$\begin{aligned} H_{f,t} &\leftarrow \frac{H_{f,t+1} + H_{f,t-1}}{2} + \frac{\phi}{2}, \\ P_{f,t} &\leftarrow \frac{P_{f+1,t} + P_{f-1,t}}{2} + \frac{\phi}{2\alpha}. \end{aligned} \tag{4.8}$$

It can be interpreted that the optimal values are the sum of 1) averages of their temporal/spectral neighbor components to minimize the gradient terms and 2) extra values to minimize the residual components. Because these two terms are obviously nonnegative, the nonnegativity constraints on H and P hold after the update. However, since it does not hold for V, which is $M - H - P$, the minimum boundary must be set to ensure the nonnegativity of V.

Algorithm 1 shows the overall procedure of the proposed SVS algorithm. Steps 1 through 7 explain the abovementioned procedure. First, an input music signal is transformed into the spectrogram domain using STFT and is scale-compressed. Then, H and P are initialized to zero. For each iteration, H and P are updated based on Steps 5 and 6, with a minimum filter to satisfy the

nonnegativity condition of the singing voice component. After the iteration is done, the scale-compressed spectrograms of singing voice and accompaniment, which are the main aims of the algorithm, are estimated as Steps 8 and 9.

Because the separation is performed in the scale-compressed spectrogram domain and the constraint of $M = V + A$ does not ensure that $\boldsymbol{M} = \boldsymbol{V} + \boldsymbol{A}$, the perfect reconstruction of the input signal cannot be guaranteed using the separated spectrograms. To overcome this problem, we use a generalized Wiener filter as shown in Steps 10 and 11, where and are the separated singing voice and accompaniment components in the original STFT domain. By using the generalized Wiener filter, it is guaranteed that $\boldsymbol{M} = \boldsymbol{V} + \boldsymbol{A}$; thus, they can be directly converted into time domain signals using the inverse STFT. Finally, a high-pass filter is applied as a postprocessing step to remove the components at low frequencies from the vocal signal because vocal signal is rarely present at low frequencies. The removed signal is considered to be part of the accompaniment signal and is added to it.

4.4 Experimental evaluation

4.4.1 MIR-1k Dataset

To quantitatively evaluate the proposed vocal separation algorithm, we used the MIR-1 K database, which consists of 1000 music clips sung by amateur singers [32]. Singing voice and accompaniment tracks are recorded separately, and we mixed the signals in -5 dB, 0 dB, and 5 dB singing voice-to-accompaniment ratio (VAR) conditions.

We used the sampling rate of 16 kHz and the analysis window size of 1024 samples with a 3/4 overlapping ratio. The parameters α and ϕ were set to be

Algorithm 1 Pseudocode for the optimization of singing voice separation algorithm using harmonic-percussive-sparse separation.

$\boldsymbol{M} \leftarrow \mathrm{STFT}(m)$ ▷ m: music signal
$M \leftarrow |\boldsymbol{M}|^{2\gamma}$
$H \leftarrow 0$
$P \leftarrow 0$
while $iter \leq maxiter$ **do**
    $H_{f,t} \leftarrow \min(\frac{H_{f,t+1}+H_{f,t-1}}{2} + \frac{\phi}{2}, M_{f,t} - P_{f,t})$
    $P_{f,t} \leftarrow \min(\frac{P_{f+1,t}+H_{f-1,t}}{2} + \frac{\phi}{2\alpha}, M_{f,t} - H_{f,t})$
end while
$V \leftarrow M - H - P$
$A \leftarrow H + P$
$\boldsymbol{V} \leftarrow \boldsymbol{M}\frac{V^{\frac{1}{2\gamma}}}{V^{\frac{1}{2\gamma}}+A^{\frac{1}{2\gamma}}}$
$\boldsymbol{A} \leftarrow \boldsymbol{M}\frac{A^{\frac{1}{2\gamma}}}{V^{\frac{1}{2\gamma}}+A^{\frac{1}{2\gamma}}}$
$v \leftarrow \mathrm{ISTFT}(\boldsymbol{V})$ ▷ ISTFT: inverse STFT
$a \leftarrow \mathrm{ISTFT}(\boldsymbol{A})$
$\hat{a} \leftarrow \mathrm{LPF}(v)$ ▷ LPF: low-pass filtering with predefined cutoff frequency
$v \leftarrow v - \hat{a}$ ▷ v: separated singing voice
$a \leftarrow v + \hat{a}$ ▷ a: separated accompaniment

0.25 and $0.025E$, respectively. The number of the iterations was 200. For fair comparison, the final results were obtained using a high-pass filter with a 110 Hz cut-off frequency, which is the same as that used in Tachibana's algorithm [23], while the highest performance was obtained when a 120 Hz cut-off frequency was used. As a performance metric, we used GNSDR, which is widely used for the evaluation of SVS algorithms. Detailed explanation for GNSDR is in Section 1.

We evaluate the proposed algorithm with several conventional ones [23, 31, 32, 33]. To briefly describe each algorithms, **Li** first detects the f_0 of singing voice and separate its harmonic components [31], **Hsu** is similar with **Li** but

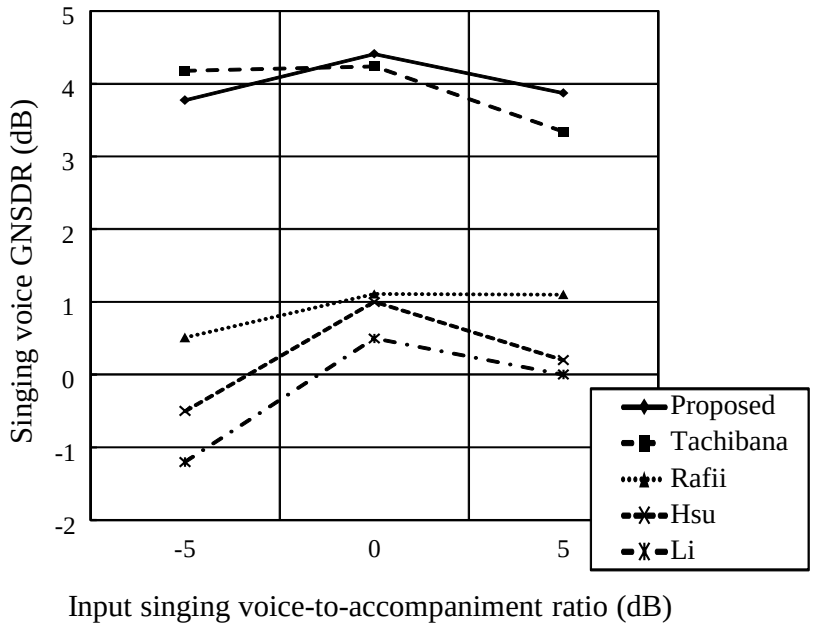

Fig. 4.3 Comparison of GNSDR in different singing voice-to-accompaniment conditions.

uses additional process for the 'unvoiced' singing voice using Gaussian mixture model (GMM), and **Rafii** removes the repetitive component from a music spectrogram which is considered as accompaniment [33]. **Tachibana** uses the two-stage HPSS explained in Section 4.2.2 [23].

As shown in Fig. 4.3, the proposed method shows the highest GNSDR compared to other conventional algorithms with VAR values of 0 dB and 5 dB. It shows relatively low GNSDR with a VAR of -5 dB compared to Tachibana's algorithm. One possible explanation is that vocal components with small values tend to converge to zero because of the sparsity constraint. Finding an additional compensation mechanism is required as a future work.

4.4.2 Beach boys Dataset

For the next experiment, Real-world music with longer time length was used. Although it is difficult to obtain original multitrack recordings for evaluation purposes, the album *Good Vibrations*: *Thirty Years of the Beach Boys* by the Beach Boys, which was released in 1993, contains several tracks where the vocal is recorded in one channel and all the accompaniment in the other one [57]. Despite the limitations that these recordings are by the same artist and in the same genre, they have been considered a useful dataset for the evaluation of vocal separation algorithms in many papers [26, 35].

We set the parameters to be the same as those used in the MIR-1K experiments, but the cut-off frequency of the high-pass filter was set differently to 100 Hz for a fair comparison [26, 35]. For the same reason, we computed the mean SDR instead of GNSDR, and it was calculated using BSS-EVAL metrics [17]. Table 4.1 shows the overall separation performance achieved using the Beach Boys dataset. Considering that the reported SDRs of the separated vocal for FitzGerald's method [26] were -1.48 dB, 1.54 dB, and 1.89 dB in -6 dB, 0 dB, and 6 dB input VAR conditions (no pretraining), respectively, the results show that the proposed method achieves comparable or higher separation performance, even though a direct comparison is not appropriate because the detailed experimental conditions such as the exact tracks used and the size of the segmented input signal, as well as the main criteria for evaluation, were not the same.

Table 4.1 Evaluation results of proposed SVS algorithm using Beach boys dataset. All the result is in dB.

Input VAR	Singing voice			Accompaniment		
	SDR	SIR	SAR	SDR	SIR	SAR
-6 dB	-2.74	-1.08	6.14	5.85	8.09	11.12
0 dB	2.00	4.82	6.51	2.23	3.38	10.82
6 dB	5.80	11.87	7.35	-1.76	-0.89	9.76

4.4.3 iKala dataset in MIREX 2014

The presented algorithm was submitted to Music Information Retrieval Evaluation eXchange (MIREX) 2014. iKala datasets was used to evaluate the algorithms, and three evaluation criteria, GNSDR for singing voice, GNSDR for accompaniment, and running time, were used.

Parameters used in the submission is as follows. We used the analysis window size of 2048 samples with a 3/4 overlapping ratio. The parameters α and ϕ were set to be 0.25 and $0.01E$.

11 algorithms was submitted from 8 teams. Belows are the brief description of the approach that each teams used.

IIY first separates singing voice and accompaniment using RPCA. Melody contour of singing voice is then detected from the separated singing voice, by using f_0 detection and singing voice activity detection. The separated singing

voice is enhanced again by extracting the f_0-related harmonic components. This study was later developed by Ikemiya et al. [24].

RNA detects singing voice using a support vector machine, and also detects its predominant f_0. Singing voice is isolated using detected f_0 and harmonic sinusoidal modeling, then it is reconstructed from the estimated sinusoidal model parameters. The accompaniment signal is obtained by subtracting the estimated singing voice from music [66].

GW uses a Bayesian NMF approach, and the basis estimated by NMF are clustered into two groups using K-means clustering or NMF [67].

RP uses REPET-SIM, which is a extended version of REPET [34]. To handle the non-periodically repeating accompaniments, the similarity matrix is used to find the indices of repeating components, while the conventional REPET assumed that the components are repeated periodically based on its tempo. Repeating spectrogram, which enhances the repeating components in the spectrogram is considered as accompaniment, and the residual is extracted as singing voice [36]. This study is later extended by Rafii et al. [68].

LFR uses kernel additive modeling, which is based on local regression of a specific time-frequency coefficient in spectrogram using its neighbor ones. Various types of kernel such as vertical (frequency axis) for percussive sounds, horizontal (time axis) for harmonic sounds, periodic for repeating sounds, and cross-shape for the detailed local characteristics. It is also combined with a compression algorithm to reduce the computational cost [69].

YC uses the spectral-temporal modulation features extracted from the auditory spectrogram. Each time-frequency coefficients are clustered into three groups, singing voice, harmonic, percussive sounds, by using a two-stage clustering process [70].

HKHS used recurrent neural network. a magnitude spectrum of music mixture is used as a model input, and the model is trained to predict the magnitude spectra of singing voice and accompaniment which the mixture consists of. The submission also presented several ideas, including mask estimation and the discriminative objective function [71].

As shown in Table 4.2, the proposed algorithm which is denoted to JL1 achieved remarkable results in GNSDR for both singing voice and accompaniment. However, compared to the other algorithms with similar results, GNSDR for accompaniment is slightly lower than that of singing voice, and it is needed to be developed. In terms of computation efficiency, the proposed algorithm was executed in the lowest runtime, and it makes the proposed algorithm to be used in real-world applications.

4.5 Conclusion

Focusing on the unique characteristics of the vocal distinct from the accompaniment in a music signal, we proposed an algorithm for separating the vocal and accompaniment signal from monaural music using a single optimization framework.

We assumed that an accompaniment signal can be represented as the sum

Table 4.2 Results of singing voice separation algorithms submitted to MIREX 2014. JL1 denotes the implementation of the proposed algorithm.

Name	Algorithm	GNSDR (dB)		Runtime (hh)
		Voice	Music	
IIY2	RPCA+pitch	4.48	7.87	02
IIY1	RPCA+pitch	4.22	7.79	02
JL1	**Proposed**	**4.16**	**5.63**	**01**
RNA1	Pitch	3.69	7.32	06
GW1	NMF	2.89	5.25	24
RP1	Repetition	2.86	5.03	01
LFR1	KAM	0.65	3.09	03
YC1	EM	-0.82	-3.12	13
HKHS1	RNN	-1.40	0.35	06
HKHS2	RNN	-1.94	0.52	06
HKHS3	RNN	-2.48	0.14	06

of the sustained harmonic and percussive sounds, and that the sparse residual components that cannot be regarded as exclusively either harmonic or percussive may be identified as the vocal signal. Although the proposed algorithm is an extended version of the previously proposed HPSS algorithm, which must be used twice in succession for SVS, the derivation of the proposed algorithm is simpler, and the quantitative evaluation demonstrates that it achieves improved or comparable performance in various singing voice-to-accompaniment conditions.

Because the proposed algorithm is based on the harmonic but nonstable characteristics of a singing voice signal, which makes it distinguishable from

both sustained harmonic and percussive accompaniment signals, it is obvious that the performance degrades when the input music signal contains a nonstable harmonic accompaniment or a sustained vocal. For example, a guitar sound with strong vibrato could be incorrectly separated as a vocal, while a sustained vocal with weak fluctuation could be separated as an accompaniment. To overcome these limitations of the proposed algorithm, we will exploit more characteristics of singing voice and accompaniment signals and use machine-learning approaches.

Chapter 5

Singing voice separation using low-rankness and sparsity

5.1 Introduction

Continuing Chapter 4, another approach for SVS which uses low-rankness and sparsity is explained in this chapter. The contents of the chapter are based on the original works by Jeong et al. [30, 72].

As discussed in Section 3.3, singing voice and accompaniment have salient differences in terms of low-rankness, as well as sparsity. Because of repetitive nature of accompaniment, it often can be represented as a combination of a few numbers of vectors, whereas singing voice has strong sparsity because the sounds occurred by voice rarely coincide. Assuming that a music spectrogram can be represented as a sum of the singing voice and accompaniment, it is a reasonable approach to separate them by decomposing the low-rank and sparse components from those mixture.

If the low-rank component is the primary target in the low-rank/sparse decomposition, then the residual components can be considered as errors or noise with a sparse distribution. In this case, the decomposition procedure is often explained as a low-rank approximation that is robust against noise. For example, NMF, which approximates a nonnegative matrix as a multiplication of two low-rank nonnegative matrices [73, 74], is extended to robust NMF by using the sparse error function such as $l_{2,1}$-norm [75], the difference between l_1- and l_2-norm [76, 77], Cauchy function [78], correntropy induced metric [79], and Huber function [79]. On the other hand, adding an additional outlier matrix instead of changing error function is also widely used for the robustness of NMF [64, 80]. In case of PCA, RPCA, which is similar to the PCA-based dimensionality reduction but uses l_1-norm error function, is one of the most popular approach [81]. Various SVS algorithm have been proposed by using above approaches. Sprechmann et al. proposed an SVS algorithm using robust NMF [64], and Huang et al. used RPCA [28]. In this section, approaches for SVS based on RPCA mainly focused.

The rest of this section is organized as follows. In Section 5.2, the algorithms of RPCA and its application to SVS are introduced, and their limitations and improvement methods are discussed. In Section 5.3, generalization of conventional RPCA using Schatten p- and l_p-norm is described. In Section 5.4, another extended RPCA which uses weighted l_1-norm is introduced, as well as its application for SVS using the spectral distribution and singing voice activity. Finally, we make a summary in Section 5.5.

5.2 SVS using robust principal component analysis

5.2.1 Robust principal component analysis

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

$$\begin{aligned} J(L,S) &= \mathrm{rank}(L) + \lambda\,\mathrm{nonzero}(S), \\ s.t. \quad & L + S = M, \end{aligned} \tag{5.1}$$

where $M \in \mathbb{R}^{F\times T}$, $L \in \mathbb{R}^{F\times T}$, and $S \in \mathbb{R}^{F\times T}$ are the mixture, low-rank, and sparse matrix, respectively. $\mathrm{rank}(\cdot)$ and $\mathrm{nonzero}(\cdot)$ denote the rank and the number of nonzero components in a matrix, respectively. λ denotes the relative weight between two terms. Since above objective function is difficult to solve, Candès et al. presented its convex relaxation, or RPCA, as follows [81]:

$$\begin{aligned} J(L,S) &= |L|_* + \lambda|S|_1, \\ s.t. \quad & L + S = M, \end{aligned} \tag{5.2}$$

where $|\cdot|_*$ and $|\cdot|_1$ denote the nuclear norm (sum of singular values) and l_1-norm (sum of the absolute values of matrix elements), respectively. These properly approximate $\mathrm{rank}(\cdot)$ and $\mathrm{nonzero}(\cdot)$ in (5.1) and allow to solve it in a convex formulation. As in (5.1), λ decides the relative importance between two norms. Candès et al. suggested $\lambda = 1/\sqrt{\max(F,T)}$ [81], and Huang et al. generalized it as $\lambda = k/\sqrt{\max(F,T)}$ with a parameter k [28].

5.2.2 Optimization for RPCA using augmented Lagrangian multiplier method

ALM method is one efficient method for optimization of (5.2) which uses the following objective function [81]:

$$J(L,S) = |L|_* + \lambda|S|_1 + < Y, L+S-M > + \frac{\mu}{2}||L+S-M||_F^2, \tag{5.3}$$

which is also often alternately written as follows [82]:

$$J(L,S) = |L|_* + \lambda|S|_1 + \frac{\mu}{2}||L+S-M+\frac{1}{\mu}Y||_F^2. \tag{5.4}$$

When S is fixed, (5.4) can be simplified as follows:

$$J(L) = \frac{1}{2}||L-F||_F^2 + \frac{1}{\mu}|L|_*, \tag{5.5}$$

where $F = M - S - \frac{1}{\mu}Y$. Likewise, it can be simplified when L is fixed as

$$J(S) = \frac{1}{2}||S-G||_F^2 + \frac{\lambda}{\mu}|S|_1, \tag{5.6}$$

where $G = M - L - \frac{1}{\mu}Y$. Optimization of (5.2) using a generic Lagrange multiplier algorithm is done by minimizing L and S in (5.3) or (5.4), and updating the Lagrange multiplier matrix Y via $Y_k \leftarrow Y + \mu(L+S-M)$.

More practically, one can take a strategy that iteratively minimize L and S of (5.5) and (5.6), respectively, which is easier to find the optimal solution for each iteration. The optimal L in (5.5) can be directly obtained as

$$L \leftarrow U_F \delta_{1/\mu}(\Lambda_F) V_F, \tag{5.7}$$

where $U_X \Lambda_X V_X = X$ is the singular value decomposition of X. δ is the element-wise shrinkage operator that is $\delta_\tau(x) = \text{sgn}(x)\max(|x| - \tau, 0)$. Similarly, the optimal S in (5.6) can be obtained as

$$S \leftarrow \delta_{\lambda/\mu}(G). \tag{5.8}$$

Algorithm 2 describe a pseudocode for the optimization of RPCA.

Algorithm 2 Pseudocode for the optimization of robust principal component analysis

set $0 < \rho < 1$, $\mu > 0$
$L, S, Y \leftarrow 0$
while $iter \leq maxiter$ **do**
    update L as (5.7)
    update S as (5.8)
    update Y by $Y \leftarrow Y + \mu(L + S - M)$
    (optional) update μ by $\mu > \rho\mu$
end while

5.2.3 SVS using RPCA

Huang et al. suggested that RPCA can be applied to separate the singing voice and the accompaniment from music signal [28]. Fig. 5.1 shows the concept of RPCA-based SVS algorithm. In the case of accompaniment, instruments often reproduce the same sounds in the same music, therefore its magnitude spectrogram can be represented as a low-rank matrix. On the contrary, singing voice has a sparse distribution in the spectrogram domain due to its strong harmonic structure. Therefore, M, L, and S in (5.2) can be considered as a spectrogram of the input music, accompaniment, and singing voice, respectively. After the separation is done in the spectrogram domain, the waveforms of sources are obtained by performing inverse STFT with the same phase of original mixture.

Although RPCA has been successfully applied to SVS, there is still plenty of room for improvement. One of the main factor of its limits is the simplicity of RPCA. Since RPCA has only one parameter, λ, it is difficult to adapt SVS by using the parameter tuning strategy. In addition, the nuclear norm and l_1-norm in RPCA, which is used for the simplicity in convex optimization, is not exactly fitted for the practical SVS task.

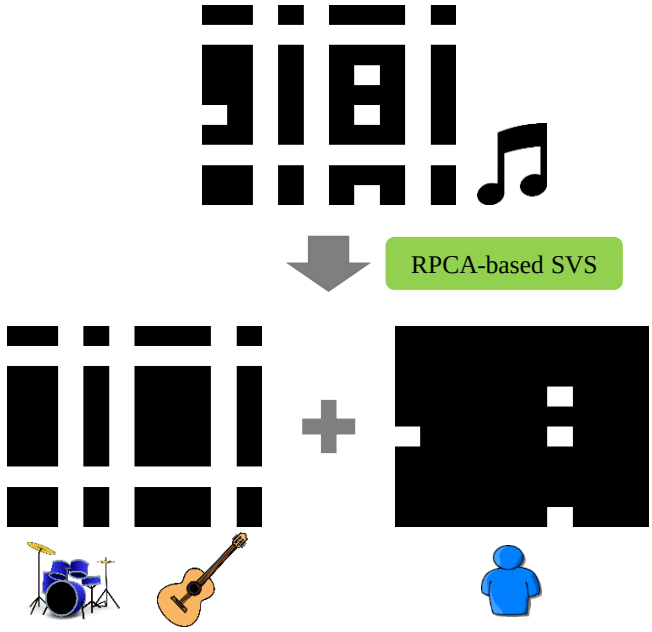

Fig. 5.1 Framework for the RPCA-based SVS.

Several improvement methods can be applied to overcome the above limits. First, two norms in RPCA can be generalized by using Schatten p- and l_p-norm. The advantage of the norm generalization is not only modifying RPCA to be closer to the ideal low-rank/sparse decomposition, but also adding more parameters that can be tuned to maximize the separation quality. However, additional consideration for optimization is required since it is not convex when $0 \leq p < 1$. In addition, the generalization can be also tried in terms of a preprocessing of input matrix.

Another improvement method is applying the task-specific characteristics into the algorithm. Since the input and output matrix of RPCA for SVS is a time-frequency representation of music signal, it would be appropriate to

apply the temporal-spectral characteristics of musical sources. Considering the spectral characteristics, a spectral distribution is one of the most important factor to distinguish sources. On the other hand, temporal activation can be also useful to represents the temporal characteristics.

5.3 SVS using generalized RPCA

5.3.1 Generalized RPCA using Schatten p- and l_p-norm

The nuclear norm and the l_1-norm in RPCA are specific cases of the Schatten p-norm and l_p-norm when $p = 1$. From the definition of the l_p-norm of a vector, the l_p-norm of a matrix S can be defined as follows:

$$||S||_p = (\sum_{f,t}(|S_{f,t}|^p))^{\frac{1}{p}}. \tag{5.9}$$

Similarly, the Schatten p-norm of matrix L is defined as

$$||L||_{Sp} = (\sum_{i}\sigma_i^p)^{\frac{1}{p}}. \tag{5.10}$$

Although (5.9) and (5.10) strictly define the norms only when $p \geq 1$, in this thesis, we do not differentiate them from a quasi-norm, which means $0 < p < 1$ as in [82]. Thus, the l_p-norm and the Schatten p-norm to the p power are

$$||S||_p^p = \sum_{f,t}(|S_{f,t}|^p), \tag{5.11}$$

$$||L||_{Sp}^p = \sum_{i}\sigma_i^p. \tag{5.12}$$

When p is closed to zero, $||S||_p^p$ and $||L||_{Sp}^p$ approximately represent the number of non-zero components and that of the non-zero singular values of S and L, which denote the sparsity and rank, respectively. Therefore, we can reasonably

infer that the smaller p is, the closer to (5.1) our approximation becomes. Practically, a gap exists between the ideal model and the real situation, and the optimal p may lie in $0 < p < 1$. In this paper, we use the same p for the two norms for convenience and to reduce the number of parameters. The aim of the extended RPCA using the Schatten p- and l_p-norms (pRPCA) is to minimize the following objective function:

$$\begin{aligned} J(L,S) &= ||L||_{Sp}^{p} + \lambda||S||_{p}^{p}. \\ s.t. \quad & L + S = M, \end{aligned} \tag{5.13}$$

It is noted that (5.13) is non-convex, and there is a trade-off between (5.2) (convex but further from (5.1)) and (5.13) (closer to (5.1) but non-convex). A more suitable function for a particular application may be determined through experimental comparison.

5.3.2 Comparison of pRPCA with robust matrix completion

To solve (5.13), we refer to an existing similar algorithm for the matrix completion [82]. The presented objective function is as expressed as follows:

$$J(L) = ||L_{\Omega} - M_{\Omega}||_{p}^{p} + \gamma||L||_{Sp}^{p}, \tag{5.14}$$

where $M_{\Omega} = \{M_{f,t}|(f,t) \in \Omega\}$ denotes the given (observed) values in matrix M. The aim of (5.14) is to estimate original matrix L from the incomplete and noisy observation using the low-rank and the sparse noise models. If all values in M are given and $L - M$ is introduced to matrix S, the objective function can be expressed with equality constraint as follows:

$$\begin{aligned} J(L,S) &= ||L||_{Sp}^{p} + \lambda||S||_{p}^{p}, \\ s.t. \quad & L - S = M, \end{aligned} \tag{5.15}$$

where $\lambda \propto 1/\gamma$. It is noted that (5.15) is the same as (5.13) except for the sign of S in the equality constraint.

5.3.3 Optimization method of pRPCA

In order to solve (5.13), we apply the method of ALM which is also used in RPCA [81] and Schatten p-/l_p- norm robust matrix completion [82]. It uses the following unconstrained objective function:

$$J(L,S) = ||L||_{Sp}^{p} + \lambda||S||_{p}^{p} + \frac{\mu}{2}||S + L - M + \frac{1}{\mu}\Lambda||_{F}^{2}, \tag{5.16}$$

where Λ is an ALM. It aims to solve (5.16) by incrementing it iteratively. When L is fixed, (5.16) can be simplified as follows:

$$J(S) = \frac{1}{2}||S - H||_{f}^{2} + \frac{\lambda}{\mu}|||S|||_{p}^{p}, \tag{5.17}$$

where $H = M - L - \frac{1}{\mu}\Lambda$. Likewise, when S is fixed, (5.16) can be simplified as follows:

$$J(L) = \frac{1}{2}||L - G||_{F}^{2} + \frac{1}{\mu}|||L|||_{Sp}^{p}, \tag{5.18}$$

where $G = M - S - \frac{1}{\mu}\Lambda$.

Algorithm 3 describes the iterative procedure to solve (5.13). For details in solving (5.17) and (5.18), please refer to [82].

5.3.4 Discussion of the normalization factor for λ

In this section, we discuss the normalization factor for λ under various p values. Let us assume we have matrix M, which is separated into $L + S = M$ using

Algorithm 3 Pseudocode for the optimization of pRPCA

set $0 < \rho < 1$, $\mu > 0$
$L, S, Y \leftarrow 0$
while $iter \leq maxiter$ **do**
    update S by solving (5.17)
    update L by solving (5.18)
    update Y by $Y \leftarrow Y + \mu(L + S - M)$
    update μ by $\mu > \rho\mu$
end while

pRPCA with $J(L,S) = ||L||_{Sp}^p + \lambda||S||_p^p$. If we have another matrix M', which is an n-times repetition of M as $M' = [MM \cdots MM]$, we expect that M' should be separated using pRPCA into $L' + S' + M'$, where $L' = [LL \cdots LL]$ and $S' = [SS \cdots SS]$. On the basis of this objective, we propose normalization factor λ' , where $\lambda = k\lambda'$, as follows: if the singular value decomposition (SVD) of L is $U_L \Lambda_L V_L$, then the SVD of L' is $L' = U(\sqrt{n}\Lambda)(\sqrt{n^{-1}}[VV \cdots VV])$. Therefore, the Schatten p-norm of L' to the p power is given by

$$||L'||_{Sp}^p = n^{\frac{p}{2}}||L||_{Sp}^p. \tag{5.19}$$

On the other hand, the p-norm of S' to the p power is given by

$$||S'||_p^p = n||S||_p^p, \tag{5.20}$$

From this property, the proper normalization factor λ' can be determined to make $|L'||_{Sp}^p + \lambda||S'||_p^p = n^{\frac{p}{2}}||L||_{Sp}^p + nk\lambda'||S||_p^p$ be equal to $|L||_{Sp}^p + \lambda||S||_p^p = ||L||_{Sp}^p + k\lambda'||S||_p^p$ as

$$\lambda' = n^{\frac{p}{2}-1}. \tag{5.21}$$

Thus, λ' should be changed depending on the relative size of the matrix, i.e., to the power $\frac{p}{2} - 1$ to be exact. In general, we finally determine λ' by considering

both the dimensions of the matrix as follows:

$$\lambda' = \max(f,t)^{\frac{p}{2}-1}. \tag{5.22}$$

We note that $\lambda' = \sqrt{\max(f,t)^{-1}}$ when $p = 1$, which is the same as that suggested for RPCA [81].

5.3.5 Generalized RPCA using scale compression

In practical experiments, the magnitude spectrogram is not a suitable domain for audio source separation. Instead, a proper scale compression can increase the separation performance [14, 61]. We present the extended RPCA by applying a scale compression and generalized Wiener filtering step (SC-RPCA). The objective function of the SC-RPCA is expressed as follows:

$$\begin{aligned} J(L,S) &= ||\hat{L}||_* + \lambda||\hat{S}||_1, \\ s.t. \quad \hat{L} + \hat{S} &= M^\alpha, \end{aligned} \tag{5.23}$$

where $0 < \alpha < 1$. It is noted that the optimization process of the SC-RPCA is same as that of the RPCA except by taking the input matrix as M^α instead of M. To obtain L and S from $\hat{L}$ and $\hat{S}$, we used the generalized Wiener filters as follows:

$$S_{f,t} = \frac{\hat{S}_{f,t}}{\hat{L}_{f,t} + \hat{S}_{f,t}} M_{f,t}, \tag{5.24}$$

$$L_{f,t} = \frac{\hat{L}_{f,t}}{\hat{L}_{f,t} + \hat{S}_{f,t}} M_{f,t}. \tag{5.25}$$

These filters allow perfect reconstruction of the input M from the separated components L and S.

5.3.6 Experimental results

The MIR-1K dataset is used to evaluate the proposed algorithms [32]. It consists of 1000 wav files of Chinese karaoke pop songs from amateur singers. Each file contains vocal and accompaniment tracks recorded separately with a 4–13-s duration and 16-kHz sampling rate. We also used the BSS-EVAL 3.0 as the evaluation criteria, which includes SIR, SAR, and SDR [17].

In all experiments, we first mix the signals in -5, 0, and 5-dB VAR. The spectrograms of the mixtures are generated using a 1024-size Hamming window with a hop size of 256. First, we discuss the value of λ for the SC-RPCA. Table 5.1 lists the GNSDR result of the separated singing voice and shows that smaller k values (where $\lambda = k\lambda'$) are needed for a smaller α value because when α becomes smaller, the spectrogram of the input music becomes smoother and can be easily approximated using the low-rank model. Therefore, to maintain the overall amount of residual matrix S, the weight of the l_1-norm minimization should be smaller. From Table 5.1, we determine the values of α and k to be 0.4 and 0.6, respectively. On the other hand, the other parameters in pRPCA ($k = 1.5, p = 0.4$) and the conventional RPCA ($k = 1.5$) are also chosen empirically to maximize the separation performance.

The overall separation performance using RPCA, pRPCA, and SC-RPCA are listed in Table 5.2 and Table 5.3. The interesting point shown in the tables is that pRPCA tends to show better performance in higher VARs, whereas SC-RPCA shows better performance in lower VARs. However, both algorithms show better results than the conventional RPCA in most of the mixing conditions.

Table 5.1 GNSDR of the separated singing voice using SC-RPCA over various values of α and k. The input VAR is 0 dB

		α			
		0.3	0.4	0.5	0.6
k	0.4	4.51	3.86	3.02	2.39
	0.6	4.47	4.74	4.11	3.43
	0.8	3.51	4.65	4.64	4.18
	1	2.41	3.92	4.52	4.47

Table 5.2 Performance comparison of the separated singing voice.

	Input VAR								
	-5 dB			0 dB			5 dB		
	GNS DR	SIR	SAR	GNS DR	SIR	SAR	GNS DR	SIR	SAR
RPCA	3.53	0.83	6.15	3.91	6.99	8.41	3.08	12.89	10.60
*p*RPCA	3.52	1.15	5.63	4.06	7.12	8.63	**4.04**	12.84	12.26
SC-RPCA	**4.25**	1.76	6.54	**4.74**	8.31	8.81	4.01	14.34	11.29

5.4 SVS using RPCA and spectral distribution

5.4.1 RPCA with weighted $l_1$norm

Since λ in (5.2) is a global parameter for all the element of M, or $M_{f,t}$, once its value is decided then all $M_{f,t}$ have the same importance for the low-rankness of $L_{f,t}$ and the sparsity of $S_{f,t}$. However, it is not always proper in actual situation, and might be too simple. For example, if we know that $L_{f,t} = 0$ for some (f, t), we may able to choose the value of λ to be $\lambda = 0$ for those element. If $S_{f,t} = 0$,

Table 5.3 Performance comparison of the separated accompaniment.

	Input VAR								
	-5 dB			0 dB			5 dB		
	GNS DR	SIR	SAR	GNS DR	SIR	SAR	GNS DR	SIR	SAR
RPCA	1.36	7.96	12.92	2.97	3.98	12.24	4.11	-0.02	11.14
*p*RPCA	1.27	9.76	9.90	**4.09**	6.90	9.17	**6.97**	4.60	8.20
SC-RPCA	**1.59**	8.03	14.07	3.48	4.59	12.58	5.29	1.45	10.63

on the contrary, we may set $\lambda \to \infty$. To apply the different weight for each element, we present RPCA with weighted l_1-norm, or wRPCA, which replace λ to the weighting matrix Λ as:

$$\begin{aligned} \text{minimize} \quad & |L|_* + |\Lambda \otimes S|_1, \\ s.t. \quad & L + S = M, \end{aligned} \tag{5.26}$$

where $\otimes$ denotes the element-wise multiplication operator. Note that $|\Lambda \otimes S|_1$ is a weighted l_1-norm of S, which has been presented in a number of previous studies [83, 84]. To solve (5.26), optimization method for RPCA such as ALM method can be directly used, just by replacing λ to Λ.

5.4.2 Proposed method: SVS using wRPCA

We extended previous RPCA-based SVS framework, by using wRPCA instead of RPCA in particular. We refer several previous studies to design the separation framework [29, 85, 86].

Nonnegativity constraint

At first, we added a nonnegativity constraint in (5.26) as follows:

$$\begin{aligned} &\text{minimize} \quad |L|_* + |\Lambda \otimes S|_1, \\ &s.t. \quad L + S = M, \quad L \geq 0, \quad S \geq 0. \end{aligned} \tag{5.27}$$

This constraint prevent that large value of $\Lambda_{f,t}$ makes large negative value for S. The optimization of (5.27) is similar as of (5.2) or (5.26) but L and S are rectified as $x \leftarrow \max(x, 0)$ in every iteration.

Two-stage framework using VAD

There were two opposite studies on SVS and VAD. Chan et al. suggested that additional vocal activity information can improve SVS [29]. On the other hand, Lehner and Widmer suggested that SVS can improve the accuracy of VAD algorithm [25]. To apply both of these suggestions, we conducted the two-stage framework as shown in Fig. 5.2. At the first stage, the sources are separated without vocal activity information. Next, vocal activity is detected using the separated singing voice. In the second separation stage, the sources are separated again with detected vocal activity information. We basically used VAD algorithm presented by Lehner et al. which uses well-designed mel-frequency cepstral coefficients (MFCC) as features [86]. In addition, we also used the vocal variance features which were also proposed in their other studies [85]. For the classification, we used random forest with 500 trees, and used threshold of 0.55. As a post-processing step, median filtering was applied to the frame-wise classification results with 7 frames filter length (1.4s). Note that above framework is also based on the previous study [86]. Because the temporal resolution of spectrogram and VAD might be different, we aligned them by considering

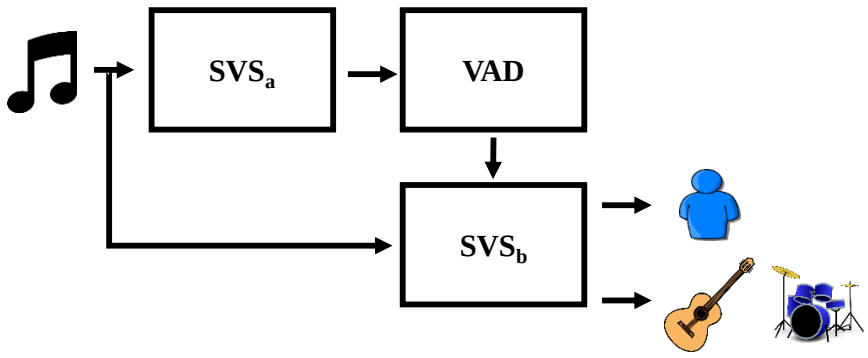

Fig. 5.2 Framework of singing voice separation using two-stage wRPCA and VAD.

those absolute time indices so that we can obtain the frame-wise VAD results.

Choosing the value for Λ

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

$$\Lambda = k\lambda\Delta, \tag{5.28}$$

where λ is $1/\sqrt{max(F,T)}$ suggested by Candès et al. [81], and k is a global parameter used by Huang et al. [28]. In this work, we empirically set it to be $k = 0.6$. Δ is a element-wise weighting matrix which is our main interest.

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

$$\Delta_{f,t} = \frac{b_A(f)}{b_V(f)}, \tag{5.29}$$

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

$$\begin{aligned} b_A(f) &= \sum_t |A_{f,t}|, \\ b_V(f) &= \sum_t |V_{f,t}|, \end{aligned} \tag{5.30}$$

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

This variance ratio might be different when only vocal-activated frames are estimated. At least it will be smaller than (5.29) in overall, since all the non-vocal frames where singing voice is absent are excluded. In addition, since we know that there is no singing voice in the non-vocal frames, we can set the weight for those frames to infinite so the singing voice can be successfully eliminated. Consequently, we set $\hat{\Delta}$ for the second separation stage as follows:

$$\hat{\Delta}_{f,t} = \begin{cases} \frac{\hat{b}_A(f)}{\hat{b}_V(f)}, & \text{if } p(t) = 1, \\ \infty, & \text{otherwise}, \end{cases} \tag{5.31}$$

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

$$\hat{b}_A(f) = \sum_t |\hat{A}_{f,t}|,$$
$$\hat{b}_V(f) = \sum_t |\hat{V}_{f,t}|, \tag{5.32}$$

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

Handling multi-channel signals

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

5.4.3 Experimental results using DSD100 dataset

We applied our SVS algorithm to the dataset and the evaluation criteria from sixth community-based signal separation evaluation campaign (SiSEC 2016): professionally-produced music recordings (MUS) [87]. This campaign provided Demixing Secrets Dataset 100 (DSD100), which consist 50 tracks for training ('dev') and other 50 for testing ('test'). All the tracks are sampled at 44.1kHz

and have stereo channels. Because there are 4 sources (vocals, bass, drums, and others) for each track, we considered the sum of bass, drums, and others as accompaniment. We used the dev set only to set Λ and $\hat{\Lambda}$, and even to train the VAD algorithm. In our experiments, VAD scores 0.87 F-score and 84% accuracy from the test set. As the evaluation criteria, it measures SDR, image-to-spatial distortion ratio (ISR), SIR, and SAR based on BSS-Eval [17]. To generate the spectrogram of music, we took the magnitude of short-time Fourier transform with Hanning window of 4096 samples and half overlap.

Fig. 5.3 shows the comparison of conventional RPCA, wRPCA, and two-stage wRPCA with VAD, and Tabel 5.4 shows the numerical values of the median of SDR. From this result, we can find that the proposed wRPCA improve SDR score for both singing voice and accompaniment, and even VAD does. However, the improvement from VAD is considerably degraded in the test set compared to the dev set. Considering that VAD for dev data makes almost perfect accuracy since it is trained by itself, we can expect that the better VAD algorithm is required to maximize its effectiveness. Example results are shown in Fig. 5.4 and Fig. 5.5. Compared to the conventional RPCA, it is observed that wRPCA successfully improve the separation quality, especially in the low-frequency region, and even VAD does in the non-vocal frames in particular.

5.4.4 Comparison with state-of-the-arts in SiSEC 2016

The presented SVS algorithms using wRPCA was submitted to SiSEC 2016:MUS. This task aims to separates music into 4 sources, which are labeled as vocal, drum, bass, and others, but it is not mandatory to obtain all

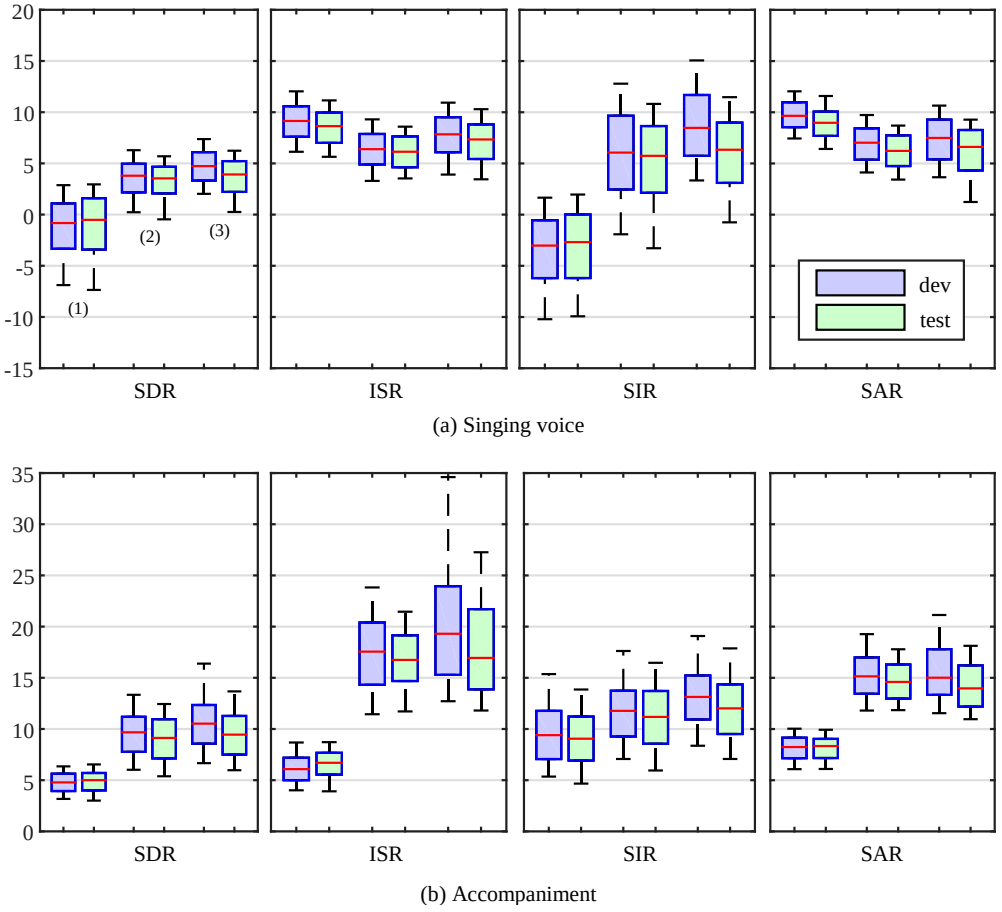

Fig. 5.3 Comparison of singing voice separation results using (1) conventional RPCA, (2) proposed wRPCA, and (3) wRPCA with VAD.

the sources. In addition, a sum of drum, bass, and other is considered as accompaniment, thus separating singing voice and accompaniment only is also possible.

Fig. 5.6 Fig. 5.7 show the results of the submissions in SiSEC 2016. Belows are the brief explanation for the submissions.

Table 5.4 Numerical values of median SDR in Fig. 5.3.

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

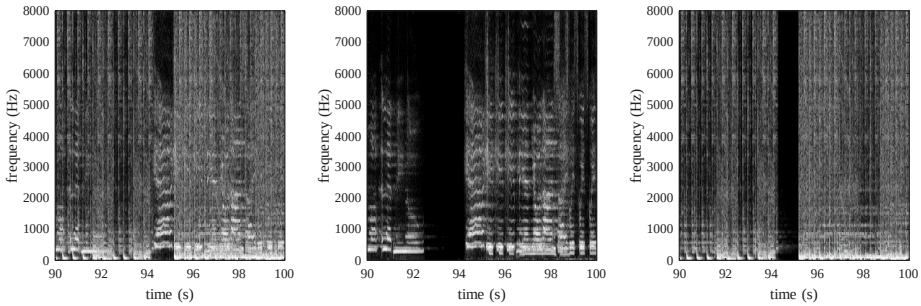

Fig. 5.4 Log-spectrograms of example mixture, singing voice, and accompaniment. Audio clips are excerpted from 'AM Contra - Heart Peripheral' in the dev set of DSD100.

GRA used ensemble methods with multiple DNNs. The DNNs are trained with those respective setting, such as target (source or mask), masking type (binary or soft), or discriminate objective function [88].

HUA used a conventional RPCA-based SVS algorithm [28].

KON used RNN that jointly optimize the mask [89].

RAF used REPET-based separation algorithms as **RP** in Section 4.4.3 [34, 35, 36].

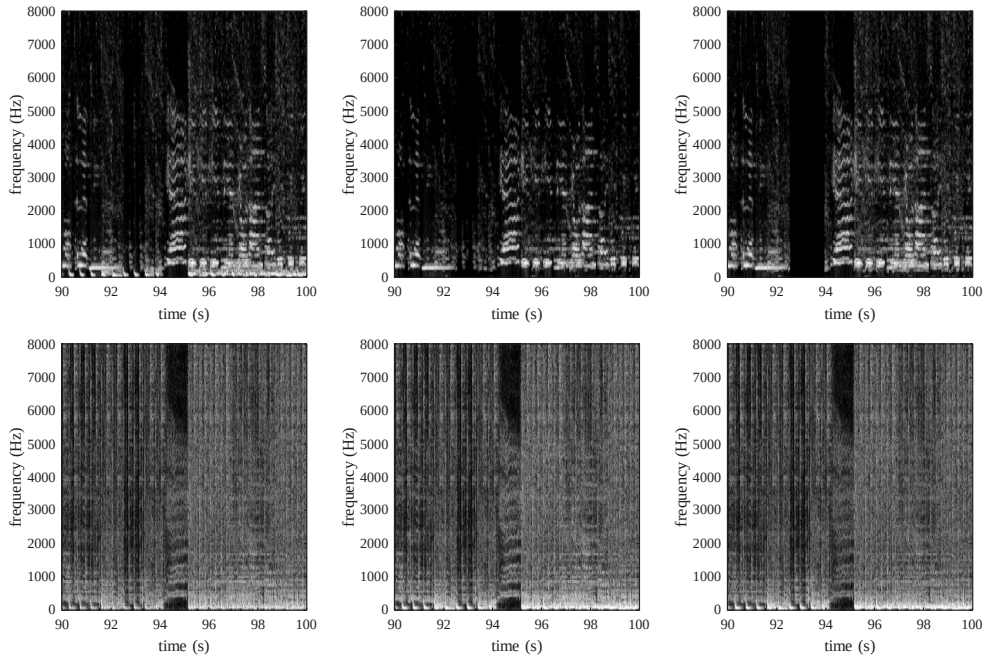

Fig. 5.5 Log-spectrograms of separated singing voice (top) and accompaniment (bottom). Input mixture is same as in Fig. 5.4.

KAM used KAM as **LFR** in Section 4.4.3 [69].

CHA used CNN which uses spectrograms as an input and output. In particular, it used the different size for the convolution filters in each layer, to have vertical or horizontal shape [90].

DUR used a mid-level representation, including pitch and timbre, which is provided by a source/filter model [91].

OZE used the flexible audio source separation toolbox (FASST), which conducts source separation frameworks using generalized expectation-

maximization [92].

STO used DNN-based algorithm, which uses STFT of common fate model (CFM) as an input and an output of separation model [93].

NUG used Two-stage multichannel DNN which has a similar framework to **UHL** [27, 94].

UHL used DNN, bi-directional LSTM, or those linear combination to learn a model that obtains the musical instruments from the music spectrogram. It used two-stage network framework, that the spectral densities of instruments in mono are estimated in the first stage, then it is recursively updated in second stage with spatial parameter updates. In addition, it applied several data augmented technique that randomizing the channel order, source amplitudes, or source combination to generate a mixture [27, 95].

IBM is not a submitted algorithm, but it is displayed as an expected maximum performance using the ideal binary mask.

The submitted algorithms can be categorized into two groups. The first one is conventional approaches, which is based on the characteristics modeling or statistic estimation. On the other hand, another approaches are based on machine learning, including NMF, DNN, convolutional neural network (CNN), or recurrent neural network (RNN). Fig. 5.6 and Fig. 5.7 show that the proposed methods outperform the conventional approaches. Moreover, they also shows the comparable results with the deep learning-based approaches.

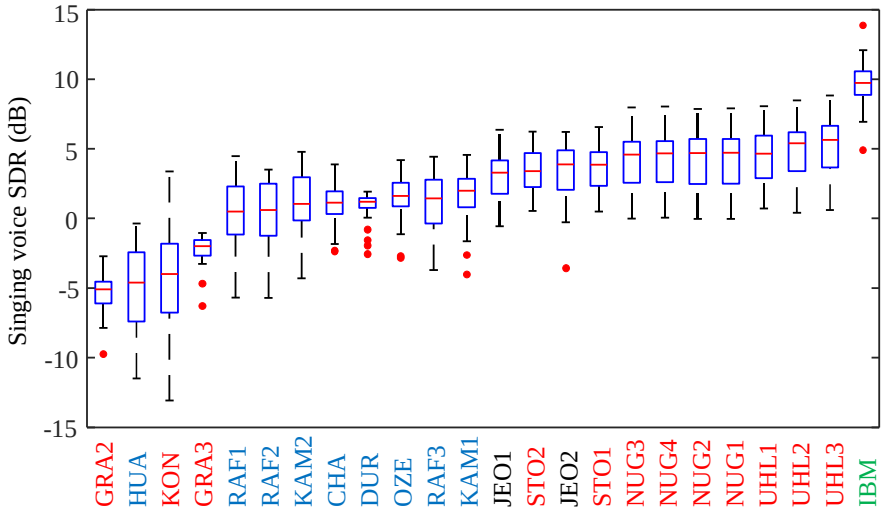

Fig. 5.6 Boxplot of singing voice SDR of the submissions in SiSEC 2016. Submissions are ordered in median on singing voice SDR. Colors of submission names represents its approach, which is conventional (red), machine learning-based (blue), proposed method (black), and ideal results (green).

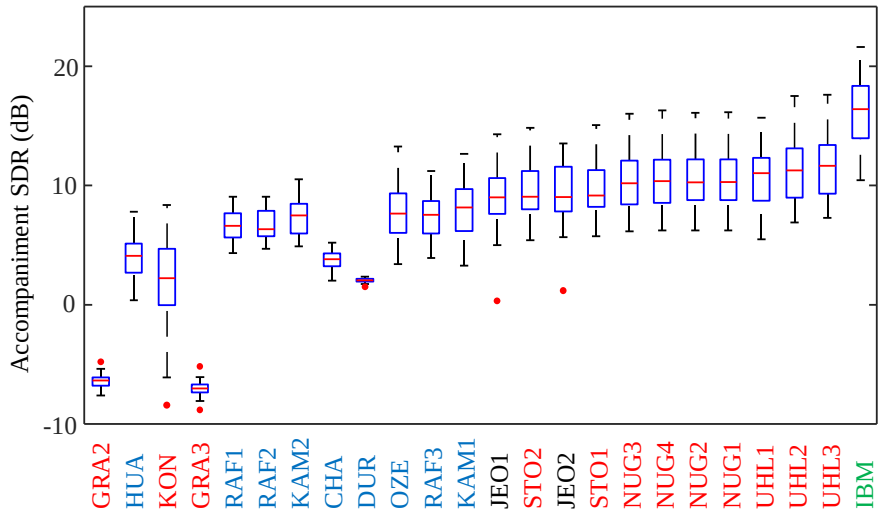

Fig. 5.7 Boxplot of accompaniment SDR of the submissions in SiSEC 2016. The order and color of submission is same as Fig. 5.6.

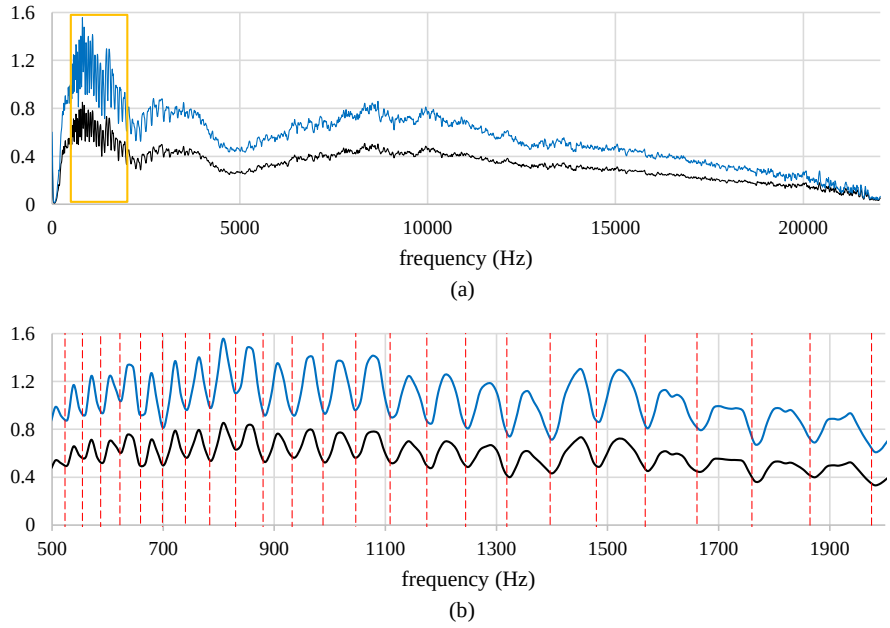

Fig. 5.8 (a) $(\frac{b_A(f)}{b_V(f)})^{-1}$ (black) and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ (blue) where $(\cdot)^{-1}$ is for visibility, and (b) the enlarged plot in the range of (500, 2000), which is marked as a yellow square. Red dotted line denotes the frequencies that correspond to musical note (C#5 to B6).

5.4.5 Discussion

Since the main contribution of our work is the use of Λ and $\hat{\Lambda}$, more accurately, Δ and $\hat{\Delta}$, we discuss in depth about the characteristics of them. Fig. 5.8 shows the plots of $(\frac{b_A(f)}{b_V(f)})^{-1}$ and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ where $(\cdot)^{-1}$ is for visibility. Higher value means that the singing voice is stronger than the accompaniment in that frequency bin. What follows are several interesting insights we found from these plots.

- $(\frac{b_A(f)}{b_V(f)})^{-1}$ and $(\frac{\hat{b}_A(f)}{\hat{b}_V(f)})^{-1}$ both show similar trends but only the scales are different, and we expect it means that the spectral characteristics of accompa-

niment are similar between in vocal and non-vocal frames.

- Singing voice is extremely weaker than accompaniment in very low frequency range (lower than 100Hz). It is reasonable because singing voice is mostly distributed in f_0 and its harmonics, which is rarely occur in those range, while some instruments such as bass and drums can be. Some previous studies for SVS have applied this characteristics by using high-pass filtering [61, 23].

- Some peaks can be found from the envelope, that are located around 0.7, 1.5, 3, and 8kHz. we expect it is related with the formants of singing voice.

- From Fig. 5.8 (b), we found an interesting phenomena that the singing voice is relatively weak in the frequency bins which correspond to the musical notes compared to those neighbor frequency bins. Although it needs more experiments to clarify the reason, we made some possible hypotheses as follows: 1) the mainlobe of singing voice may wider than that of accompaniment, 2) singing voice has stronger vibrato in general, and it may cause the 'blurred peak' in a long window length, or 3) singers frequently fail to sound exact note frequency, and make more errors than the instrumental players.

5.5 Summary

In this section, SVS algorithms based on low-rankness of accompaniment and sparsity of singing voice were discussed. The conventional RPCA-based SVS algorithm was briefly discussed, including its motivation and algorithm as well as its optimization method.

Although the RPCA concept is appropriate for SVS problem, it still needs to be extended or generalized for this specific usage. Two generalized RPCA-based approaches have been presented in this section. First, we have proposed

the application of the Schatten p- and l_p-norms instead of the nuclear norm and l_1-norm, respectively. We have also presented a simple scale compression process to make a spectrogram more proper representation for decomposition. Experimental results show that both methods yield performance better than or comparable to the conventional RPCA. Our next step will be to minimize the computational intensity of pRPCA. Because most of the operation time is spent by SVD, we would be able to significantly reduce it using the inexact SVD used in RPCA [81]. Furthermore, we plan to combine pRPCA and SC-RPCA in a single framework. Finally, we will investigate the use of other acoustic characteristics, such as harmonicity or timbre, to help separate vocal from the rest.

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

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The goal of the thesis was to develop SVS system by applying the common properties of singing voice and accompaniment in music signals. To this end, we first discussed for the three distinct characteristics, that is, continuity, low-rankness, and sparsity (Chapter 3). This discussion includes the definitions, those mathematical representations, as well as those meaning in audio spectrograms. In addition, musical sources such as harmonic instruments, percussive instruments, accompaniment, and singing voice are discussed in terms of those characteristics. It was even empirically shown by using the actual music data.

Algorithms for SVS was presented based on these characteristics. First, we presented an algorithm based on the continuity and sparsity (Chapter 4). In particular, we extended the conventional SVS algorithm which uses two-stage HPSS, by applying additional sparse residual which is considered as singing

voice. An objective function for the single-stage separation framework is derived, and its separation quality evaluated by using various datasets and challenges shows better or comparable results compared to other conventional algorithms. Another approach uses the low-rankness and sparsity, and we have presented several generalization or extension method for the original RPCA-based SVS algorithm (Chapter 5). We first proposed to generalize the nuclear norm and l_1-norm to Schatten p-norm and l_p-norm, which are closer to ideal low-rankness and sparsity, and even allow SVS algorithm to use one more parameter p. On the other hand, wRPCA, which extends RPCA by using weighted l_1-norm instead of l_1-norm, was introduced. We introduced another useful characteristics, spectral distribution, and presented wRPCA-based SVS algorithm whose weight is set based on this characteristics.

6.2 Contributions

The main contributions of this thesis can be summarized in the following points:

- **A broad range of review on SVS**: A comprehensive review of SVS was provided. The importance and applications of SVS have been introduced, as well as the important keywords and evaluation criteria. We have categorized the algorithms for SVS into four groups–characteristics-based, spatial, machine learning-based and inform approach–and introduced those representative methods. Finally, we have listed the popular datasets and challenges.

- **In-depth discussion of singing voice and accompaniment characteristics**: We have discussed how singing voice and accompaniment

can be distinguished, especially in terms of the low-level features. Three characteristics, continuity, low-rankness, and sparsity have mainly been focused. We also have shown the numerical estimation results for those characteristics using the actual music data.

- **A novel algorithm for SVS using continuity and sparsity**: We presented a novel SVS algorithm, which extends HPSS algorithm but with an additional sparse residual. Simple optimization strategy have also been presented. The separation quality have been evaluated by using various datasets, and verified that the presented algorithm shows the better or comparable quality with efficient computation, compared to the state-of-the-art algorithms.

- **Generalized RPCA-based SVS using Schatten p- and l_p-norm**: We presented pRPCA as an generalized version of RPCA with Schatten p- and l_p-norm, which is more accurate to approximate the rank and sparsity. We have also discussed the optimization method of pRPCA as well as the normalization factor for the weighting parameter. In addition, we have applied pRPCA for the SVS task. The optimal parameters for the SVS task have been found empirically, and the separation quality has been evaluated in various mixing condition.

- **Scale compression for RPCA-based SVS**: We have presented a simple method to improve the RPCA-based SVS by using scale compression. From the experimental results, we have empirically found the compression rate which shows the highest separation quality.

- **RPCA with weighted l_1-norm**: We have presented wRPCA, that replace the l_1-norm of RPCA into weighted l_1-norm. It allows to choose the different importance between low-rankness and sparsity, for each coefficients in the matrix.

- **wRPCA-based SVS applying the spectral distribution and VAD**: We have presented a novel SVS algorithm which is based on wRPCA. We have used the spectral distribution of sources to decide the values of weights. We have also presented the two-stage SVS framework which uses vocal activity detection. Compared to the conventional RPCA-based SVS algorithm, the proposed algorithm shows the meaningful improvements in numerical evaluation.

6.3 Future work

6.3.1 Discovering various characteristics for SVS

In this thesis, we have discussed various characteristics of singing voice and accompaniment, including continuity, low-rankness, sparsity as well as spectral distribution. Unfortunately, because all the proposed SVS algorithms use only a subset of them, integrating all the characteristics into a single separation algorithm remains as a future task. In addition, there are also other important characteristics which have been widely used in other SVS algorithms but did not mainly focused in this thesis, including the predominant f_0 of singing voice or repetition of accompaniment, and it is even needed to be integrated in the future.

In addition, it is required to discover novel characteristics for SVS. Considering other music source separation or MIR-related tasks, there are various

features which imply the musical concepts. For example, harmonicity has been widely used for music transcription, and vibrato or tremolo have been applied for instrument recognition, as well as singing voice detection. Group sparsity, which is similar to sparsity but for each group rather than each element, is useful for NMF to let a source to occur in the specific time region but totally eliminated in the other region. As a future work, in-depth discussion for this characteristics, especially for singing voice and accompaniment, is required.

6.3.2 Expanding to other SVS approaches

As introduced in 2.2, the characteristics-based SVS can be considered as the lowest-level approach. To maximize the separation quality, it is mandatory to expand the methods to the higher-level approach, such as spatial or machine learning-based one.

One simple method is to use the separation results of characteristics to estimate the power spectral density (PSD) for spatial approach. Assuming that all the unit sounds in a source occurs in the same location, the multi-channel observation can be modeled by using PSD of source and the filters for each microphone. Therefore, knowing PSD can be helpful to improve the separation quality in spatial approaches. Similarly, the separation results can be used as a prior in case of machine learning-based approach such as NMF-based SVS.

6.3.3 Applying the characteristics for deep learning models

Deep learning is a part of machine learning methods, whose features are not designed by human engineers but learned from data using a general-purpose learning procedure. It outperforms the conventional approaches in almost all

the machine learning tasks regardless of the field, including image (object recognition [96]), natural language (translation [97]), audio (speech recognition [98]), and even games (the game of Go [99]). As discussed in 5.4.4, deep learning-based method also top-ranked in SVS with huge performance gap compared to conventional ones.

Can knowledge of source characteristics be helpful for deep learning-based SVS algorithms, even though it may be expected to learn those characteristics by themselves? It may be debatable. Since deep learning model learns the characteristics by itself, domain knowledge can be considered not only information but also kind of bias. Moreover, since deep learning studies are trying to develop an end-to-end framework, which uses a raw signal as an input data without any preprocessing or feature extraction, one may expect that characteristics-based approaches will become increasingly meaningless.

However, we still believe that it is still useful, especially for architecture design. Because deep learning architectures consist of multiple layers, which represents from the low-level features to its abstracted high-level ones, it is helpful to guide the model what to learn in the low layer and how to abstract those features in the high layer. For example, we have shown that proper settings for receptive field in neural network can improve genre recognition accuracy [100]. As the future work, it is required to design the relevant architecture by implying the characteristics of data.

Bibliography

[1] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] D. Barry, D. Fitzgerald, E. Coyle, and B. Lawlor, "Drum source separation using percussive feature detection and spectral modulation," 2005.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components

by complementary diffusion on spectrogram," in *Proceedings of European Signal Processing Conference (EUSIPCO)*. IEEE, 2008, pp. 1–4.

[6] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2010.

[7] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, vol. 13, 2002, p. 17.

[8] M. E. Markaki, A. Holzapfel, and Y. Stylianou, "Singing voice detection using modulation frequency feature." in *SAPA@ INTERSPEECH*, 2008, pp. 7–10.

[9] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3. IEEE, 2005, pp. iii–17.

[10] M. McVicar, D. P. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3117–3121.

[11] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 16, no. 2, pp. 338–349, 2008.

[12] H. Papadopoulos and G. Peeters, "Large-scale study of chord estimation algorithms based on chroma representation and hmm," in *Proceedings of International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2007, pp. 53–60.

[13] M. A. Alonso, G. Richard, and B. David, "Tempo and beat estimation of musical signals." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2004.

[14] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 139–144.

[15] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–7.

[16] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of ACM Multimedia Conference (ACMMM)*. ACM, 1999, pp. 77–80.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[19] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. CRC Press, 2004.

[20] S. S. Stevens, "The measurement of loudness," *The Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 815–829, 1955.

[21] J. Schnupp, I. Nelken, and A. King, *Auditory neuroscience: Making sense of sound*. MIT press, 2011.

[22] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[23] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 1, pp. 228–237, 2014.

[24] 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 Acoustics, Speech, and Signal Processing*, vol. 24, no. 11, pp. 2084–2095, 2016.

[25] B. Lehner and G. Widmer, "Monaural blind source separation in the context of vocal detection." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 309–315.

[26] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62–73, 2010.

[27] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proceedings of European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1748–1752.

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

[29] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2015, pp. 718–722.

[30] I.-Y. Jeong and K. Lee, "Singing voice separation using rpca with weighted l1-norm," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Springer, 2017, pp. 553–562.

[31] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.

[32] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 18, no. 2, pp. 310–319, 2010.

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

[34] ——, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 21, no. 1, pp. 73–84, 2013.

[35] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012, pp. 53–56.

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

[37] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[38] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans-*

actions on Acoustics, Speech, and Signal Processing, vol. 18, no. 3, pp. 550–563, 2010.

[39] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2003, pp. 177–180.

[40] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[41] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Springer, 2015, pp. 387–395.

[42] 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, 2017, pp. 323–332.

[43] N. Ono, Z. Koldovský, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2013, pp. 1–6.

[44] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[45] J. Driedger, H. Grohganz, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proceedings of ACM Multimedia Conference (ACMMM)*. ACM, 2013, pp. 541–544.

[46] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 888–891.

[47] C. Joder and B. W. Schuller, "Score-informed leading voice separation from monaural audio," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 277–282.

[48] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[49] Z. Chen, P.-S. Huang, and Y.-H. Yang, "Spoken lyrics informed singing voice separation," in *Proceedings of Hacking Audio and Music Research (HAMR)*, 2013. [Online]. Available: http://labrosa.ee.columbia.edu/hamr2013/proceedings/doku.php/singingseparation

[50] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2009, pp. 69–72.

[51] I.-Y. Jeong and K. Lee, "Informed source separation from monaural music with limited binary time-frequency annotation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2015, pp. 489–493.

[52] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 883–887.

[53] N. J. Bryan, G. J. Mysore, and G. Wang, "Isse: an interactive source separation editor," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2014, pp. 257–266.

[54] N. Q. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1567–1571.

[55] A. Lefevre, F. Bach, and C. Févotte, "Semi-supervised {NMF} with time-frequency annotations for single-channel source separation," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[56] A. Lefèvre, F. Glineur, and P.-A. Absil, "A convex formulation for informed source separation in the single channel setting," *Neurocomputing*, vol. 141, pp. 26–36, 2014.

[57] The beach boys, "Good vibrations: Thirty years of the beach boys," Capitol Records, 1993.

[58] ——, "The pet sounds sessions," Capitol Records, 1997.

[59] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Computing Surveys*, vol. 47, no. 2, p. 36, 2015.

[60] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.

[61] 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, pp. 1197–1200, 2014.

[62] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 21, no. 10, pp. 2096–2107, 2013.

[63] J. Park and K. Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 148–154.

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

[65] 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 (ATC)*. IEEE, 2008, pp. 243–246.

[66] P. Rao, N. Nayak, and S. Adavanne, "Singing voice separation using adaptive window harmonic sinusoidal modeling," *The Music Information Retrieval Exchange MIREX 2014*, 2014.

[67] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 507–512.

[68] Z. Rafii and B. Pardo, "Online repet-sim for real-time speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 848–852.

[69] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2015, pp. 76–80.

[70] F. Z. Yen, Y.-J. Luo, and T.-S. Chi, "Singing voice separation using spectro-temporal modulation features." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 617–622.

[71] 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 International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 477–482.

[72] I.-Y. Jeong and K. Lee, "Vocal separation using extended robust principal component analysis with schatten p/lp-norm and scale compression," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2014, pp. 1–6.

[73] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[74] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 556–562.

[75] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2011, pp. 673–682.

[76] A. B. Hamza and D. J. Brady, "Reconstruction of reflectance spectra using robust nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3637–3642, 2006.

[77] C. Ding and D. Kong, "Nonnegative matrix factorization using a robust error function," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2033–2036.

[78] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proceedings of IEEE Workshop on Applications of Sig-*

nal Processing to Audio and Acoustics (WASPAA). IEEE, 2015, pp. 1–5.

[79] L. Du, X. Li, and Y.-D. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization," in *Proceedings of IEEE International Conference on Data Mining (ICDM)*. IEEE, 2012, pp. 201–210.

[80] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, no. 2, pp. 192–200, 2011.

[81] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011.

[82] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p-norm and\ell _p-norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, 2015.

[83] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ 1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[84] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, "Analyzing weighted l1 minimization for sparse recovery with nonuniform sparse models," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1985–2001, 2011.

[85] B. Lehner, G. Widmer, and R. Sonnleitner, "On the reduction of false positives in singing voice detection," in *Proceedings of IEEE Interna-*

tional Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2014, pp. 7480–7484.

[86] B. Lehner, R. Sonnleitner, and G. Widmer, "Towards light-weight, real-time-capable singing voice detection." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 53–58.

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

[88] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Single-channel audio source separation using deep neural network ensembles," in *Proceedings of Audio Engineering Society Convention (AES)*. Audio Engineering Society, 2016.

[89] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[90] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Springer, 2017, pp. 258–266.

[91] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[92] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[93] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 126–130.

[94] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[95] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2017.

[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision*

and pattern recognition (CVPR), 2016, pp. 770–778.

[97] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[98] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning(ICML)*, 2016, pp. 173–182.

[99] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[100] I.-Y. Jeong and K. Lee, "Learning temporal features using a deep neural network and its application to music genre classification." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 434–440.

초 록

보컬 분리란 음악 신호를 보컬 성분과 반주 성분으로 분리하는 일 또는 그 방법을 의미한다. 이러한 기술은 음악의 특정한 성분에 담겨 있는 정보를 추출하기 위한 전처리 과정에서부터, 보컬 연습과 같이 분리 음원 자체를 활용하는 등의 다양한 목적으로 사용될 수 있다.

본 논문의 목적은 보컬과 반주가 가지고 있는 고유한 특성에 대해 논의하고 그것을 활용하여 보컬 분리 알고리즘들을 개발하는 것이며, 특히 '특징 기반' 이라고 불리는 다음과 같은 상황에 대해 중점적으로 논의한다. 우선 분리 대상이 되는 음악 신호는 단채널로 제공된다고 가정하며, 이 경우 신호의 공간적 정보를 활용할 수 있는 다채널 환경에 비해 더욱 어려운 환경이라고 볼 수 있다. 또한 기계 학습 방법으로 데이터로부터 각 음원의 모델을 추정하는 방법을 배제하며, 대신 저차원의 특성들로부터 모델을 유도하여 이를 목표 함수에 반영하는 방법을 시도한다. 마지막으로, 가사, 악보, 사용자의 안내 등과 같은 외부의 정보 역시 제공되지 않는다고 가정한다. 그러나 보컬 분리의 경우 암묵 음원 분리 문제와는 달리 분리하고자 하는 음원이 각각 보컬과 반주에 해당한다는 최소한의 정보는 제공되므로 각각의 성질들에 대한 분석은 가능하다.

크게 세 종류의 특성이 본 논문에서 중점적으로 논의된다. 우선 연속성의 경우 주파수 또는 시간 측면으로 각각 논의될 수 있는데, 주파수축 연속성의 경우 소리의 음색적 특성을, 시간축 연속성은 소리가 안정적으로 지속되는 정도를 각각 나타낸다고 볼 수 있다. 또한, 저행렬계수 특성은 신호의 구조적 성질을 반영하며 해당 신호가 낮은 행렬계수를 가지는 형태로 표현될 수 있는지를 나타내며, 성김 특성은 신호의 분포 형태가 얼마나 성기거나 조밀한지를 나타낸다.

본 논문에서는 크게 두 가지의 보컬 분리 방법에 대해 논의한다. 첫 번째 방법은 연속성과 성김 특성에 기반을 두고 화성 악기-타악기 분리 방법 (harmonic-percussive sound separation, HPSS) 을 확장하는 방법이다. 기존의 방법이 두 번의 HPSS 과정을 통해 보컬을 분리하는 것에 비해 제안하는 방법은 성긴 잔여 성분을 추가해 한 번의 보컬 분리 과정만을 사용한다. 논의되는 다른 방법은 저행렬계수 특성과 성김 특성을 활용하는 것으로, 반주가 저행렬계수 모델로 표현될 수 있는 반면 보컬은 성긴 분포를 가진다는 가정에 기반을 둔다. 이러한 성분들을 분리하기 위해 강인한 주성분 분석 (robust principal component analysis, RPCA) 을 이용하는 방법이 대표적이다. 본 논문에서는 보컬 분리 성능에 초점을 두고 RPCA 알고리즘을 일반화하거나 확장하는 방식에 대해 논의하며, 트레이스 노름과 l_1 노름을 각각 샤텐 p 노름과 l_p 노름으로 대체하는 방법, 스케일 압축 방법, 주파수 분포 특성을 반영하는 방법 등을 포함한다. 제안하는 알고리즘들은 다양한 데이터셋과 대회에서 평가되었으며 최신의 보컬 분리 알고리즘들보다 더 우수하거나 비슷한 결과를 보였다.

주요어: 보컬 분리, 최적화, 음악 신호 처리

학 번: 2013-30733