Singing Voice Separation and Vocal F0 Estimation Based on Mutual Combination of Robust Principal Component Analysis and Subharmonic Summation

Yukara Ikemiya, *Student Member, IEEE*, Katsutoshi Itoyama, *Member, IEEE*, and Kazuyoshi Yoshii, *Member, IEEE*

***Abstract*—This paper presents a new method of singing voice analysis that performs mutually-dependent singing voice separation and vocal fundamental frequency (F0) estimation. Vocal F0 estimation is considered to become easier if singing voices can be separated from a music audio signal, and vocal F0 contours are useful for singing voice separation. This calls for an approach that improves the performance of each of these tasks by using the results of the other. The proposed method first performs robust principal component analysis (RPCA) for roughly extracting singing voices from a target music audio signal. The F0 contour of the main melody is then estimated from the separated singing voices by finding the optimal temporal path over an F0 saliency spectrogram. Finally, the singing voices are separated again more accurately by combining a conventional time-frequency mask given by RPCA with another mask that passes only the harmonic structures of the estimated F0s. Experimental results showed that the proposed method significantly improved the performances of both singing voice separation and vocal F0 estimation. The proposed method also outperformed all the other methods of singing voice separation submitted to an international music analysis competition called MIREX 2014.**

***Index Terms*—Robust principal component analysis (RPCA), subharmonic summation (SHS), singing voice separation, vocal F0 estimation.**

I. Introduction

SINGING voice analysis is important for active music listening interfaces [1] that enable a user to customize the contents of existing music recordings in ways not limited to frequency equalization and tempo adjustment. Since singing voices tend to form main melodies and strongly affect the moods of musical pieces, several methods have been proposed for editing the three major kinds of acoustic characteristics of singing voices: fundamental frequencies (F0s), timbres, and volumes. A system of speech analysis and synthesis called TANDEM-STRAIGHT [2], for example, decomposes human voices into F0s, spectral envelopes (timbres), and non-periodic components. High-quality F0- and/or timbre-changed singing voices can then be resynthesized by manipulating F0s and spectral envelopes. Ohishi *et al.* [3] represents F0 or volume dynamics of singing voices by using a probabilistic model and transfers those dynamics to other singing voices. Note that these methods deal only with isolated singing voices. Fujihara and Goto [4] model the spectral envelopes of singing voices in polyphonic audio signals to directly modify the vocal timbres without affecting accompaniment parts.

To develop a system that enables a user to edit the acoustic characteristics of singing voices included in a polyphonic audio signal, we need to accurately perform *both* singing voice separation and vocal F0 estimation. The performance of each task could be improved by using the results of the other because there is a complementary relationship between them. If singing voices were extracted from a polyphonic audio signal, it would be easy to estimate a vocal F0 contour from them. Vocal F0 contours are useful for improving singing voice separation. In most studies, however, only the *one-way* dependency between the two tasks has been considered. Singing voice separation has often been used as preprocessing for vocal F0 estimation, and vice versa.

In this paper we propose a novel singing voice analysis method that performs singing voice separation and vocal F0 estimation in an interdependent manner. The core component of the proposed method is preliminary singing voice separation based on robust principal component analysis (RPCA) [5]. Given the amplitude spectrogram (matrix) of a music signal, RPCA decomposes it into the sum of a low-rank matrix and a sparse matrix. Since accompaniments such as drums and rhythm guitars tend to play similar phrases repeatedly, the resulting spectrogram generally has a low-rank structure. Since singing voices vary significantly and continuously over time and the power of singing voices concentrates on harmonic partials, on the other hand, the resulting spectrogram has a not low-rank but sparse structure. Although RPCA is considered to be one of the most prominent ways of singing voice separation, non-repetitive instrument sounds are inevitably assigned to a sparse spectrogram. To filter out such non-vocal sounds, we estimate the F0 contour of singing voices from the sparse spectrogram based on a saliency-based F0 estimation method called subharmonic summation (SHS) [6] and extract only a series of harmonic structures corresponding to the estimated F0s. Here we propose a novel F0 saliency spectrogram in the time-frequency (TF) domain by leveraging the results of RPCA. This can avoid the negative effect of accompaniment sounds in vocal F0 estimation.

Manuscript received December 3, 2015; revised March 28, 2016 and May 25, 2016; accepted May 25, 2016. Date of publication June 7, 2016; date of current version September 2, 2016. The study was supported by JST OngaCREST Project, JSPS KAKENHI 24220006, 26700020, and 26280089, and Kayamori Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roberto Togneri.

The authors are with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: ikemiya@kuis.kyoto-u.ac.jp; itoyama@kuis.kyoto-u.ac.jp; yoshii@kuis.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2577879

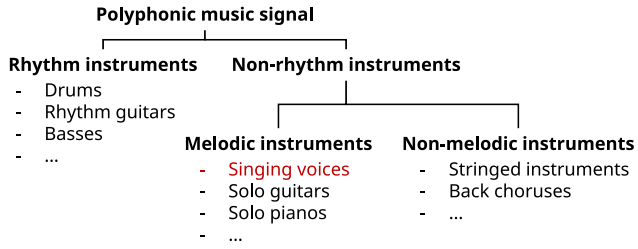


Fig. 1. Typical instrumental composition of popular music.

Our method is similar in spirit to a recent method of singing voice separation that combines rhythm-based and pitch-based methods of singing voice separation [7]. It first estimates two types of *soft* TF masks passing only singing voices by using a singing voice separation method called REPET-SIM [8] and a vocal F0 estimation method (originally proposed for multiple-F0 estimation [9]). Those soft masks are then integrated into a unified mask in a weighted manner. On the other hand, our method is deeply linked to human perception of a main melody in polyphonic music [10], [11]. Fig. 1 shows an instrumental composition of popular music. It is thought that humans easily recognize the sounds of rhythm instruments such as drums and rhythm guitars [10] and that in the residual sounds of non-rhythm instruments, spectral components that have predominant harmonic structures are identified as main melodies [11]. The proposed method first separates the sounds of rhythm instruments by using a TF mask estimated by RPCA. Main melodies are extracted as singing voices from the residual sounds by using another mask that passes only predominant harmonic structures. Although the main melodies do not always correspond to singing voices, we do not deal with vocal activity detection (VAD) in this paper because many promising VAD methods [12]–[14] can be applied as pre- or post-processing of our method.

The rest of this paper is organized as follows. Section II introduces related works. Section III explains the proposed method. Section IV describes the evaluation experiments and the MIREX 2014 singing-voice-separation task results. Section V describes the experiments determining robust parameters for the proposed method. Section VI concludes this paper.

II. Related Work

This section introduces related works on vocal F0 estimation and singing voice separation. It also reviews some studies on the combination of those two tasks.

A. Vocal F0 Estimation

A typical approach to vocal F0 estimation is to identify F0s that have predominant harmonic structures by using an F0 saliency spectrogram that represents how likely the F0 is to exist in each TF bin. A core of this approach is how to estimate a saliency spectrogram [15]–[19]. Goto [15] proposed a statistical multiple-F0 analyzer called *PreFEst* that approximates an observed spectrum as a superimposition of harmonic structures. Each harmonic structure is represented as a Gaussian mixture model (GMM) and the mixing weights of GMMs corresponding to different F0s can be regarded as a saliency spectrum. Rao *et al.* [16] tracked multiple candidates of vocal F0s including the F0s of locally predominant non-vocal sounds and then identified vocal F0s by focusing on the temporal instability of vocal components. Dressler [17] attempted to reduce the number of possible overtones by identifying which overtones are derived from a vocal harmonic structure. Salamon *et al.* [19] proposed a heuristics-based method called *MELODIA* that focuses on the characteristics of vocal F0 contours. The contours of F0 candidates are obtained by using a saliency spectrogram based on SHS. This method achieved the state-of-the-art results in vocal F0 estimation.

B. Singing Voice Separation

A typical approach to singing voice separation is to make a TF mask that separates a target music spectrogram into a vocal spectrogram and an accompaniment spectrogram. There are two types of TF masks: soft masks and binary masks. An ideal binary mask assigns 1 to a TF unit if the power of singing voices in the unit is larger than that of the other concurrent sounds, and 0 otherwise. Although vocal and accompaniment sounds overlap with various ratios at many TF units, excellent separation can be achieved using binary masking. This is related to a phenomenon called auditory masking: a louder sound tends to mask a weaker sound within a particular frequency band [20].

Nonnegative matrix factorization (NMF) has often been used for separating a polyphonic spectrogram into nonnegative components and clustering those components into vocal components and accompaniment components [21]–[23]. Another approach is to exploit the temporal and spectral continuity of accompaniment sounds and the sparsity of singing voices in the TF domain [24]–[26]. Tachibana *et al.* [24], for example, proposed harmonic/percussive source separation (HPSS) based on the isotropic natures of harmonic and percussive sounds. Both components were estimated jointly via maximum a posteriori estimation. Fitzgerald *et al.* [25] proposed an HPSS method applying different median filters to polyphonic spectra along the time and frequency directions. Jeong *et al.* [26] statistically modeled the continuities of accompaniment sounds and the sparsity of singing voices. Yen *et al.* [27] separated vocal, harmonic, and percussive components by clustering frequency modulation features in an unsupervised manner. Huang *et al.* [28] have recently used a deep recurrent neural network for supervised singing voice separation.

Some state-of-the-art methods of singing voice separation focus on the repeating characteristics of accompaniment sounds [5], [8], [29]. Accompaniment sounds are often played by musical instruments that repeat similar phrases throughout the music, such as drums and rhythm guitars. To identify repetitive patterns in a polyphonic audio signal, Rafii *et al.* [29] took the median of repeated spectral segments detected by an autocorrelation method, and improved the separation by using a similarity matrix [8]. Huang *et al.* [5] used RPCA to identify repetitive structures of accompaniment sounds. Liutkus *et al.* [30] proposed kernel additive modeling that combines many

conventional methods and accounts for various features like continuity, smoothness, and stability over time or frequency. These methods tend to work robustly in several situations or genres because they make few assumptions about the target signal. Driedger *et al.* [31] proposed a cascading method that first decomposes a music spectrogram into harmonic, percussive, and residual spectrograms, each of which is further decomposed into partial components of singing voices and those of accompaniment sounds by using conventional methods [28], [32]. Finally, the estimated components are reassembled to form singing voices and accompaniment sounds.

C. One-Way or Mutual Combination

Since singing voice separation and vocal F0 estimation have complementary relationships, the performance of each task can be improved by using the results of the other. Some vocal F0 estimation methods use singing voice separation techniques as preprocessing for reducing the negative effect of accompaniment sounds in polyphonic music [24], [29], [33], [34]. This approach results in comparatively better performance when the volume of singing voices is relatively low [35]. Some methods of singing voice separation use vocal F0 estimation techniques because the energy of a singing voice is concentrated on an F0 and its harmonic partials [32], [36], [37]. Virtanen *et al.* [32] proposed a method that first separates harmonic components using a predominant F0 contour. The residual components are then modeled by NMF and accompaniment sounds are extracted. Singing voices and accompaniment sounds are separated by using the learned parameters again.

Some methods perform both vocal F0 estimation and singing voice separation. Hsu *et al.* [38] proposed a tandem algorithm that iterates these two tasks. Durrieu *et al.* [39] used source-filter NMF for directly modeling the F0s and timbres of singing voices and accompaniment sounds. Rafii *et al.* [7] proposed a framework that combines repetition-based source separation with F0-based source separation. A unified TF mask for singing voice separation is obtained by combining the TF masks estimated by the two types of source separation in a weighted manner. Cabañas-Molero *et al.* [40] proposed a method that roughly separates singing voices from stereo recordings by focusing on the spatial diversity (called *center extraction*) and then estimates a vocal F0 contour for the separated voices. The separation of singing voices is further improved by using the F0 contour.

III. Proposed Method

The proposed method jointly executes singing voice separation and vocal F0 estimation (Fig. 2). Our method uses RPCA to estimate a mask (called an RPCA mask) that separates a target music spectrogram into low-rank components and sparse components. The vocal F0 contour is then estimated from the separated sparse components via Viterbi search on an F0 saliency spectrogram, resulting in another mask (called a harmonic mask) that separates harmonic components of the estimated F0 contour. These masks are integrated via element-wise multiplication, and finally singing voices and accompaniment sounds are obtained by separating the music spectrogram according to the integrated mask. The proposed method can work well for complicated music audio signals. Even if the volume of singing voices is relatively low and music audio signals contain various kinds of musical instruments, the harmonic structures (F0s) of singing voices can be discovered by calculating an F0 saliency spectrogram from an RPCA mask.

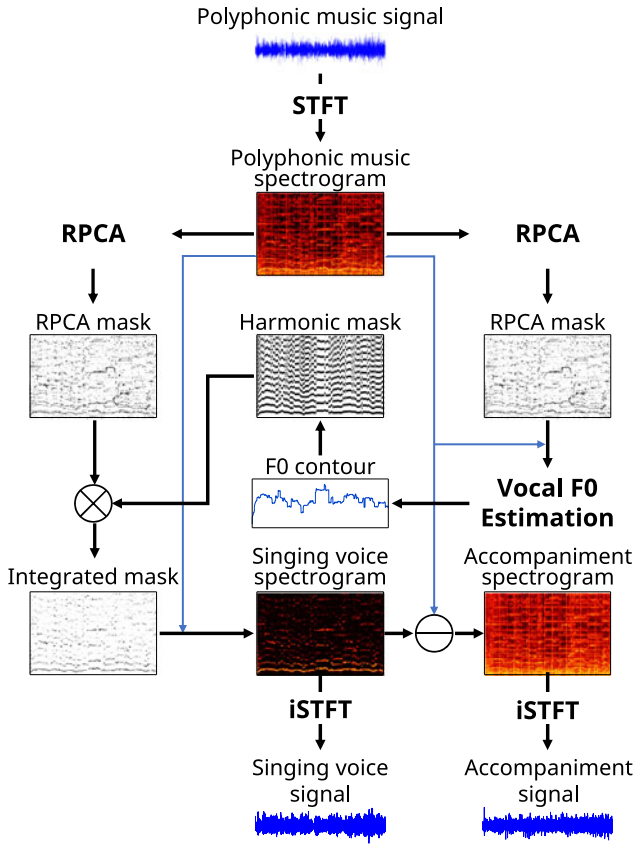


Fig. 2. Overview of the proposed method. First an RPCA mask that separates low-rank components in a polyphonic spectrogram is computed. From this mask and the original spectrogram, a vocal F0 contour is estimated. The RPCA mask and the harmonic mask calculated from the F0 contour are combined by multiplication, and finally the singing voice and the accompaniment sounds are separated using the integrated mask.

A. Singing Voice Separation

Vocal and accompaniment sounds are separated by combining TF masks based on RPCA and vocal F0s.

1) Calculating an RPCA Mask: A singing voice separation method based on RPCA [5] assumes that accompaniment and vocal components tend to have low-rank and sparse structures, respectively, in the TF domain. Since spectra of harmonic instruments (e.g., pianos and guitars) are consistent for each F0 and the F0s are basically discretized at a semitone level, harmonic spectra having the same shape appear repeatedly in the same musical piece. Spectra of non-harmonic instruments (e.g., drums) also tend to appear repeatedly. Vocal spectra, in contrast,

rarely have the same shape because the vocal timbres and F0s vary continuously and significantly over time.

RPCA decomposes an input matrix $\boldsymbol{X}$ into the sum of a low-rank matrix $\boldsymbol{X}_L$ and a sparse matrix $\boldsymbol{X}_S$ by solving the following convex optimization problem:

$$\text{minimize } \|\boldsymbol{X}_L\|_* + \hat{\lambda}\|\boldsymbol{X}_S\|_1 \text{ (subject to } \boldsymbol{X}_L + \boldsymbol{X}_S = \boldsymbol{X}\text{)}, \quad \hat{\lambda} = \frac{\lambda}{\sqrt{\max(T,F)}}, \tag{1}$$

where $\boldsymbol{X}$, $\boldsymbol{X}_L$, and $\boldsymbol{X}_S \in \mathbb{R}^{T\times F}$, $\|\cdot\|_*$ and $\|\cdot\|_1$ represent the nuclear norm (also known as the trace norm) and the L1-norm, respectively. λ is a positive parameter that controls the balance between the low-rankness of $\boldsymbol{X}_L$ and the sparsity of $\boldsymbol{X}_S$. To find optimal $\boldsymbol{X}_L$ and $\boldsymbol{X}_S$, we use an efficient inexact version of the augmented Lagrange multiplier (ALM) algorithm [41].

When $\boldsymbol{X}$ is the amplitude spectrogram given by the short-time Fourier transform (STFT) of a target music audio signal (T is the number of frames and F is the number of frequency bins), the spectral components having repetitive structures are assigned to $\boldsymbol{X}_L$ and the other varying components are assigned to $\boldsymbol{X}_S$. Let t and f be a time frame and a frequency bin, respectively ($1 \le t \le T$ and $1 \le f \le F$). We obtain a TF soft mask $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA}} \in \mathbb{R}^{T\times F}$ by using Wiener filtering:

$$M^{(\mathrm{s})}_{\mathrm{RPCA}}(t,f) = \frac{|X_S(t,f)|}{|X_S(t,f)| + |X_L(t,f)|}. \tag{2}$$

A TF binary mask $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA}} \in \mathbb{R}^{T\times F}$ is also obtained by comparing $\boldsymbol{X}_L$ with $\boldsymbol{X}_S$ in an element-wise manner as follows:

$$M^{(\mathrm{b})}_{\mathrm{RPCA}}(t,f) = \begin{cases} 1 & \text{if } |X_S(t,f)| > \gamma |X_L(t,f)| \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

The gain γ adjusts the energy between the low-rank and sparse matrices. In this paper the gain parameter is set to 1.0, which was reported to achieve good separation performance [5]. Note that $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA}}$ is used only for estimating a vocal F0 contour in Section III-B.

Using $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA}}$ or $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA}}$, the vocal spectrogram $\boldsymbol{X}^{(*)}_{\mathrm{VOCAL}} \in \mathbb{R}^{T\times F}$ is roughly estimated as follows:

$$\boldsymbol{X}^{(*)}_{\mathrm{VOCAL}} = \boldsymbol{M}^{(*)}_{\mathrm{RPCA}} \odot \boldsymbol{X}, \tag{4}$$

where $\odot$ indicates the element-wise product. If the value of λ for singing voice separation is different from that for F0 estimation, we execute two versions of RPCA with different values of λ (Fig. 2). If we were to use the same value of λ for both processes, RPCA would be executed only once. In section V we discuss the optimal values of λ in detail.

2) Calculating a Harmonic Mask: Using a vocal F0 contour $\hat{\boldsymbol{Y}} = \{\hat{y_1}, \hat{y_2}, \ldots, \hat{y_T}\}$ (see details in Section III-B), we make a harmonic mask $\boldsymbol{M}_{\mathrm{H}} \in \mathbb{R}^{T\times F}$. Assuming that the energy of vocal spectra is localized on the harmonic partials of vocal F0s, we defined $\boldsymbol{M}_{\mathrm{H}} \in \mathbb{R}^{T\times F}$ as:

$$\boldsymbol{M}_{\mathrm{H}}(t,f) = \begin{cases} & 0 < f - w^n_{\mathrm{u}} \le W, \\ & w^n_{\mathrm{l}} = \mathrm{f}\left(nh_{\hat{y}_t} - \frac{w}{2}\right), \\ \mathrm{w}(n;W) \ \text{ if } & w^n_{\mathrm{u}} = \mathrm{f}\left(nh_{\hat{y}_t} + \frac{w}{2}\right), \\ & W = w^n_{\mathrm{l}} - w^n_{\mathrm{u}} + 1, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathrm{w}(n;W)$ denotes the nth value of a window function of length W, $\mathrm{f}(h)$ denotes the index of the nearest time frame corresponding to a frequency h [Hz], n is the index of a harmonic partial, w is a frequency width [Hz] for extracting the energy around the partial, $h_{\hat{y}_t}$ is the estimated vocal F0 [Hz] of frame t. We chose the Tukey window whose a shape parameter is set to 0.5 as a window function.

3) Integrating the Two Masks for Singing Voice Separation: Given the RPCA mask (soft) $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA}}$ and the harmonic mask $\boldsymbol{M}_{\mathrm{H}}$, we define an integrated soft mask $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA+H}}$ as follows:

$$\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA+H}} = \boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA}} \odot \boldsymbol{M}_{\mathrm{H}}. \tag{6}$$

Furthermore, an integrated binary mask $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA+H}}$ is also defined as:

$$\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA+H}}(t,f) = \begin{cases} 1 & \text{if } \boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA+H}}(t,f) > 0.5 \\ 0 & \text{otherwise.} \end{cases}. \tag{7}$$

Although the integrated masks have fewer spectral units assigned to singing voices than the RPCA mask and the harmonic mask do, they provide better separation quality (see the comparative results reported in Section V).

Using the integrated masks $\boldsymbol{M}^{(*)}_{\mathrm{RPCA+H}}$, the vocal and accompaniment spectrograms $\hat{\boldsymbol{X}}^{(*)}_{\mathrm{VOCAL}}$ and $\hat{\boldsymbol{X}}^{(*)}_{\mathrm{ACCOM}}$ are given by

$$\hat{\boldsymbol{X}}^{(*)}_{\mathrm{VOCAL}} = \boldsymbol{M}^{(*)}_{\mathrm{RPCA+H}} \odot \boldsymbol{X}, \quad \hat{\boldsymbol{X}}^{(*)}_{\mathrm{ACCOM}} = \boldsymbol{X} - \hat{\boldsymbol{X}}^{(*)}_{\mathrm{VOCAL}}. \tag{8}$$

Finally, time signals (waveforms) of singing voices and accompaniment sounds are resynthesized by computing the inverse STFT with the phases of the original music spectrogram.

B. Vocal F0 Estimation

We propose a new method that estimates a vocal F0 contour $\hat{\boldsymbol{Y}} = \{\hat{y_1}, \ldots, \hat{y_T}\}$ from the vocal spectrogram $\boldsymbol{X}^{(\mathrm{b})}_{\mathrm{VOCAL}}$ by using the binary mask $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA}}$. A robust F0-saliency spectrogram is obtained by using both $\boldsymbol{X}^{(\mathrm{b})}_{\mathrm{VOCAL}}$ and $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA}}$ and a vocal F0 contour is estimated by finding an optimal path in the saliency spectrogram with the Viterbi search algorithm.

1) Calculating a Log-Frequency Spectrogram: We convert the vocal spectrogram $\boldsymbol{X}^{(\mathrm{b})}_{\mathrm{VOCAL}} \in \mathbb{R}^{T\times F}$ to the log-frequency spectrogram $\boldsymbol{X}'_{\mathrm{VOCAL}} \in \mathbb{R}^{T\times C}$ by using spline interpolation on the dB scale. A frequency h_f [Hz] is translated to the index

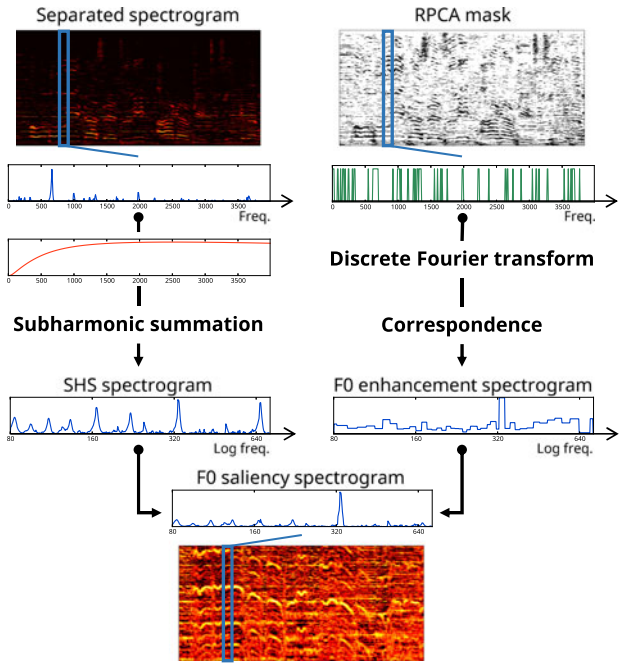


Fig. 3. An F0-saliency spectrogram is obtained by integrating an SHS spectrogram derived from a separated vocal spectrogram with an F0 enhancement spectrogram derived from an RPCA mask.

of a log-frequency bin c $(1 \leq c \leq C)$ as follows:

$$c = \left\lfloor \frac{1200 \log_2 \frac{h_f}{h_{\mathrm{low}}}}{p} + 1 \right\rfloor, \tag{9}$$

where h_{low} is a predefined lowest frequency [Hz] and p a frequency resolution [cents] per bin. The frequency h_{low} must be sufficiently low to include the low end of a singing voice spectrum (i.e., 30 Hz).

To take into account the non-linearity of human auditory perception, we multiply the A-weighting function $R_A(f)$ to the vocal spectrogram $\boldsymbol{X}^{(\mathrm{b})}_{\mathrm{VOCAL}}$ in advance. $R_A(f)$ is given by

$$R_A(f) = \frac{12200^2 h_f^4}{(h_f^2 + 20.6^2)(h_f^2 + 12200^2)} \times \frac{1}{\sqrt{(h_f^2 + 107.7^2)(h_f^2 + 737.9^2)}}. \tag{10}$$

This function is a rough approximation of the inverse of the 40-phon equal-loudness curve[1] and is used for amplifying the frequency bands that we are perceptually sensitive to, and attenuating the frequency bands that we are less sensitive to [19].

2) Calculating an F0-Saliency Spectrogram: Fig. 3 shows the procedure of calculating an F0-Saliency spectrogram. We calculate a SHS spectrogram $\boldsymbol{S}_{\mathrm{SHS}} \in \mathbb{R}^{T \times C}$ from the tentative vocal spectrogram $\boldsymbol{X}'_{\mathrm{VOCAL}} \in \mathbb{R}^{T \times C}$ in the log-frequency domain. SHS [6] is the most basic and light-weight algorithm that underlies many vocal F0 estimation methods [19], [42]. $\boldsymbol{S}_{\mathrm{SHS}}$ is given by

$$S_{\mathrm{SHS}}(t, c) = \sum_{n=1}^{N} \beta_n \boldsymbol{X}'_{\mathrm{VOCAL}} \left(t, c + \left\lfloor \frac{1200 \log_2 n}{p} \right\rfloor \right), \tag{11}$$

where c is the index of a log-frequency bin $(1 \leq c \leq C)$, N is the number of harmonic partials considered, and β_n is a decay factor (0.86^{n-1} in this paper).

We then calculate an F0 enhancement spectrogram $\boldsymbol{S}_{\mathrm{RPCA}} \in \mathbb{R}^{T \times C}$ from the RPCA mask $\boldsymbol{M}_{\mathrm{RPCA}}$. To improve the performance of vocal F0 estimation, we propose to focus on the regularity (periodicity) of harmonic partials over the linear frequency axis. The RPCA binary mask $\boldsymbol{M}_{\mathrm{RPCA}}$ can be used for reducing half or double pitch errors because the harmonic structure of the singing voice strongly appears in it.

We first take the discrete Fourier transform of each time frame of the binary mask as follows:

$$F(t, k) = \left| \sum_{f=0}^{F-1} M^{(\mathrm{b})}_{\mathrm{RPCA}}(t, f) e^{-i \frac{2\pi k f}{F}} \right|. \tag{12}$$

This idea is similar to the cepstral analysis that extracts the periodicity of harmonic partials from log-power spectra. We do not need to compute the log of the RPCA binary mask because $\boldsymbol{M}_{\mathrm{RPCA}} \in \{0, 1\}^{T \times F}$. The F0 enhancement spectrogram $\boldsymbol{S}_{\mathrm{RPCA}}$ is obtained by picking the value corresponding to a frequency index c:

$$S_{\mathrm{RPCA}}(t, c) = F\left(t, \left\lfloor \frac{h_{\mathrm{top}}}{h_c} \right\rfloor \right), \tag{13}$$

where h_c is the frequency [Hz] corresponding to log-frequency bin c and h_{top} is the highest frequency [Hz] considered (Nyquist frequency).

Finally, the reliable F0-saliency spectrogram $\boldsymbol{S} \in \mathbb{R}^{T \times C}$ is given by integrating $\boldsymbol{S}_{\mathrm{SHS}}$ and $\boldsymbol{S}_{\mathrm{RPCA}}$ as follows:

$$S(t, c) = S_{\mathrm{SHS}}(t, c) S_{\mathrm{RPCA}}(t, c)^{\alpha}, \tag{14}$$

where α is a weighting factor for adjusting the balance between $\boldsymbol{S}_{\mathrm{SHS}}$ and $\boldsymbol{S}_{\mathrm{RPCA}}$. When α is 0, $\boldsymbol{S}_{\mathrm{RPCA}}$ is ignored, resulting in the standard SHS method. While each bin of $\boldsymbol{S}_{\mathrm{SHS}}$ reflects the total volume of harmonic partials, each bin of $\boldsymbol{S}_{\mathrm{RPCA}}$ reflects the number of harmonic partials.

3) Executing Viterbi Search: Given the F0-saliency spectrogram $\boldsymbol{S}$, we estimate the optimal F0 contour $\hat{\boldsymbol{Y}} = \{\hat{y_1}, \cdots, \hat{y_T}\}$ by solving the following problem:

$$\hat{\boldsymbol{Y}} = \underset{y_1, \ldots, y_T}{\operatorname{argmax}} \sum_{t=1}^{T-1} \left\{ \log \frac{\boldsymbol{S}(t, y_t)}{\sum_{c=c_l}^{c_h} \boldsymbol{S}(t, c)} + \log G(y_t, y_{t+1}) \right\}, \tag{15}$$

where c_l and c_h are the lowest and highest log-frequency bins of an F0 search range. $G(y_t, y_{t+1})$ is the transition cost function from the current F0 y_t to the next F0 y_{t+1}. $G(y_t, y_{t+1})$ is defined as

$$G(y_t, y_{t+1}) = \frac{1}{2b} \exp\left(-\frac{|c_{y_t} - c_{y_{t+1}}|}{b} \right) \tag{16}$$

[1] http://replaygain.hydrogenaud.ioproposalequal_loudness.html

TABLE I
DATASETS AND PARAMETERS

	Number of clips	Length of clips	Sampling rate	Window size	Hopsize	N	λ	w	α
MIR-1K	110	20–110 sec	16 kHz	2048	160	10	0.8	50	0.6
MedleyDB	45	17–514 sec	44.1 kHz	4096	441	20	0.8	70	0.6
RWC-MDB-2001	100	125–365 sec	44.1 kHz	4096	441	20	0.8	70	0.6

where $b = \sqrt{\frac{150^2}{2}}$ and c_y indicates the log-frequency [cents] corresponding to log-frequency bin c. This function is equivalent to the Laplace distribution whose standard deviation is 150 [cents]. Note that the shifting interval of time frames is 10 [ms]. This optimization problem can be efficiently solved using the Viterbi search algorithm.

IV. EXPERIMENTAL EVALUATION

This section reports experiments conducted for evaluating singing voice separation and vocal F0 estimation. The results of the *Singing Voice Separation* task of MIREX 2014, which is a world-wide competition between algorithms for music analysis, are also shown.

A. Singing Voice Separation

Singing voice separation using different binary masks was evaluated to verify the effectiveness of the proposed method.

1) Datasets and Parameters: The MIR-1K dataset[2] (*MIR-1K*) and the MedleyDB dataset (*MedleyDB*) [43] were used for evaluating singing voice separation. Note that we used the 110 "Undivided" song clips of MIR-1K and the 45 clips of MedleyDB listed in Table II. The clips in MIR-1K were recorded at a 16 kHz sampling rate with 16 bit resolution and the clips in MedleyDB were recorded at a 44.1 kHz sampling rate with 16 bit resolution. For each clip in both datasets, singing voices and accompaniment sounds were mixed at three signal-to-noise ratios (SNR) conditions: -5, 0, and 5 dB.

The datasets and the parameters used for evaluation are summarized in Table I, where the parameters for computing the STFT (window size and hopsize), SHS (the number N of harmonic partials), RPCA (a sparsity factor λ), a harmonic mask (frequency width w), and a saliency spectrogram (a weighting factor α) are listed. We empirically determined the parameters w and λ according to the results of grid search (see details in Section V). The same value of λ (0.8) was used for both RPCA computations in Fig. 2. The frequency range for the vocal F0 search was restricted to 80–720 Hz.

2) Compared Methods: The following TF masks were compared.

1) *RPCA*: Using only an RPCA soft mask $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA}}$
2) *H*: Using only a harmonic mask $\boldsymbol{M}_{\mathrm{H}}$
3) *RPCA-H-S*: Using an integrated soft mask $\boldsymbol{M}^{(\mathrm{s})}_{\mathrm{RPCA+H}}$
4) *RPCA-H-B*: Using an integrated binary mask $\boldsymbol{M}^{(\mathrm{b})}_{\mathrm{RPCA+H}}$
5) *RPCA-H-GT*: Using an integrated soft mask made by using a ground-truth F0 contour

TABLE II
SONG CLIPS IN *MedleyDB* USED FOR EVALUATION

Artists	Songs
A Classic Education	Night Owl
Aimee Norwich	Child
Alexander Ross	Velvet Curtain
Auctioneer	Our Future Faces
Ava Luna	Waterduct
Big Troubles	Phantom
Brandon Webster	Dont Hear A Thing, Yes Sir I Can Fly
Clara Berry And Wooldog	Air Traffic, Boys, Stella, Waltz For My Victims
Creepoid	Old Tree
Dreamers Of The Ghetto	Heavy Love
Faces On Film	Waiting For Ga
Family Band	Again
Helado Negro	Mitad Del Mundo
Hezekiah Jones	Borrowed Heart
Hop Along	Sister Cities
Invisible Familiars	Disturbing Wildlife
Liz Nelson	Coldwar, Rainfall
Matthew Entwistle	Dont You Ever
Meaxic	Take A Step, You Listen
Music Delta	80s Rock, Beatles, Britpop, Country1, Country2, Disco, Gospel, Grunge, Hendrix, Punk, Reggae, Rock, Rockabilly
Night Panther	Fire
Port St Willow	Stay Even
Secret Mountains	High Horse
Steven Clark	Bounty
Strand Of Oaks	Spacestation
Sweet Lights	You Let Me Down
The Scarlet Brand	Les Fleurs Du Mal

6) *ISM*: Using an ideal soft mask

"RPCA" is a conventional RPCA-based method [5]. "H" used only a harmonic mask created from an estimated F0 contour. "RPCA-H-S" and "RPCA-H-B" represent the proposed methods using soft masks and binary masks, respectively, and "RPCA-H-GT" means a condition that the ground-truth vocal F0s were given (the upper bound of separation quality for the proposed framework). "ISM" represents a condition that oracle TF masks were estimated such that the ground-truth vocal and accompaniment spectrograms were obtained (the upper bound of separation quality of TF masking methods). two Note that even ISM is far from perfect separation because it is based on naive TF masking, which causes nonlinear distortion (e.g., musical noise). For H, RPCA-H-S and RPCA-H-B, the accuracies of vocal F0 estimation are described in Section IV-B.

3) Evaluation Measures: The *BSS_EVAL* toolbox[3] [44] was used for measuring the separation performance. The principle of *BSS_EVAL* is to decompose an estimate $\hat{s}$ of a true source

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

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

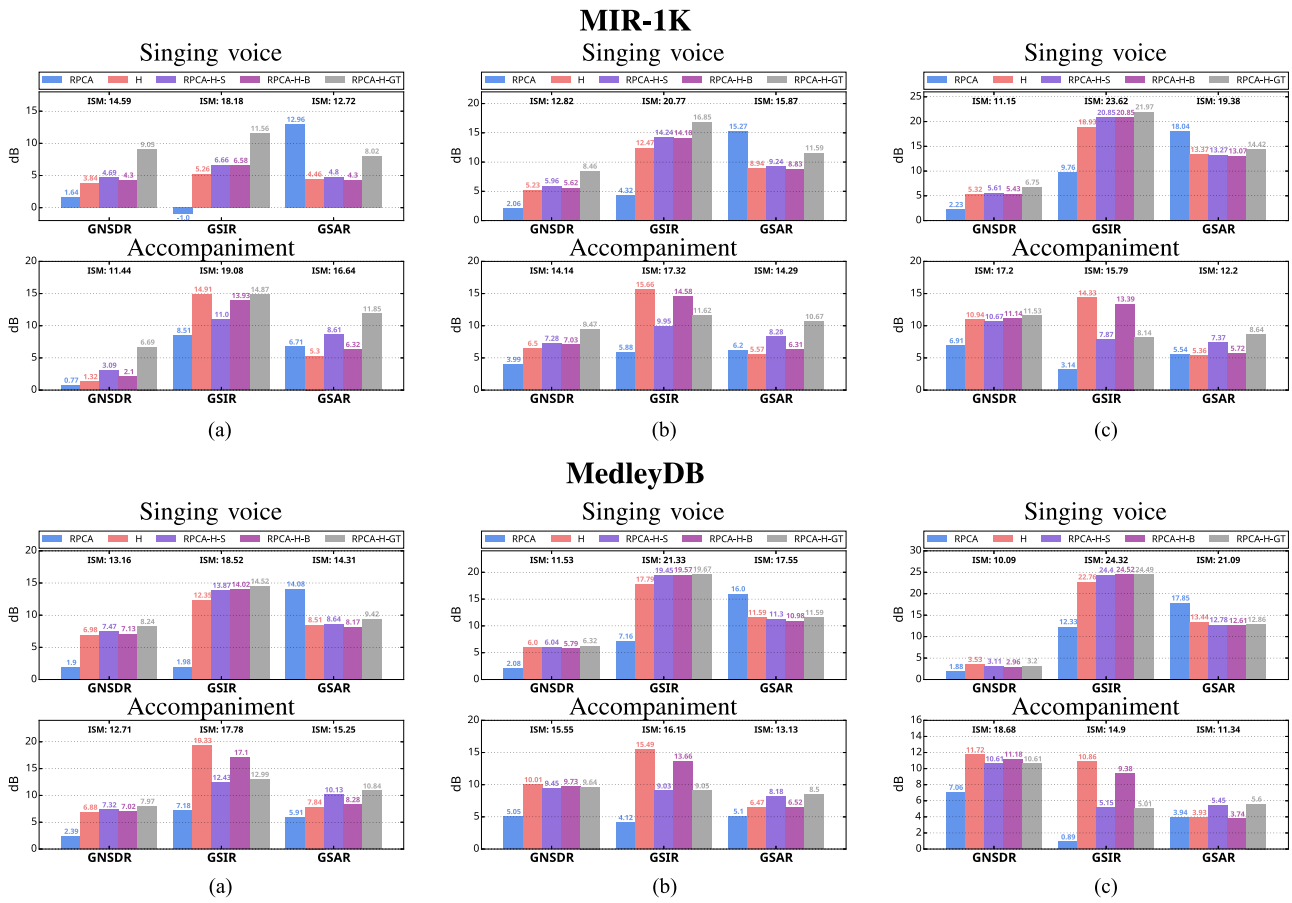


Fig. 4. Comparative results of singing voice separation using different binary masks. The upper section shows the results for MIR-1K and the lower section for MedleyDB. From left to right, the results for mixing conditions at SNRs of −5, 0, and 5 dB are shown. The evaluation values of "ISM" are expressed with letters in order to make the graphs more readable. (a) −5 dB SNR, (b) 0 dB SNR, (c) 5 dB SNR.

signal s as follows:

$$\hat{s}(t) = s_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t), \quad (17)$$

where s_{target} is an allowed distortion of the target source s and e_{interf}, e_{noise} and e_{artif} are respectively the interference of the unwanted sources, perturbing noise, and artifacts in the separated signals (such as musical noise). Since we assume that an original signal consists of only vocal and accompaniment sounds, the perturbing noise e_{noise} was ignored. Given the decomposition, three performance measures are defined: the Source-to-Distortion Ratio (SDR), the Source-to-Interference Ratio (SIR) and the Source-to-Artifacts Ratio (SAR):

$$\text{SDR}(\hat{s}, s) := 10\log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2}\right), \quad (18)$$

$$\text{SIR}(\hat{s}, s) := 10\log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}\right), \quad (19)$$

$$\text{SAR}(\hat{s}, s) := 10\log_{10}\left(\frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}\right), \quad (20)$$

where $\|\cdot\|$ denotes a Euclidean norm. two In general, there is a trade-off between SIR and SAR. When only reliable frequency components are extracted, for example, the interference of unwanted sources is reduced (SIR is improved) and the nonlinear distortion is increased (SAR is degraded).

We then calculated the Normalized SDR (NSDR) that measures the improvement of the SDR between the estimate $\hat{s}$ of a target source signal s and the original mixture x. To measure the overall separation performance we calculated the Global NSDR (GNSDR), which is a weighted mean of the NSDRs over all the mixtures x_k (weighted by their length l_k):

$$\text{NSDR}(\hat{s}, s, x) = \text{SDR}(\hat{s}, s) - \text{SDR}(x, s), \quad (21)$$

$$\text{GNSDR} = \frac{\sum_k l_k \text{NSDR}(\hat{s}_k, s_k, x_k)}{\sum_k l_k}. \quad (22)$$

In the same way, the Global SIR (GSIR) and the Global SAR (GSAR) were calculated from the SIRs and the SARs. For all these ratios, higher values represent better separation quality.

Since this paper does not deal with the VAD and we intended to examine the effect of the harmonic mask for vocal separation, we used only the voiced sections for evaluation; that is to say, the amplitude of the signals in unvoiced sections was set to 0 when calculating the evaluation scores.

4) Experimental Results: two As shown in Fig. 4, the proposed method using soft masks (RPCA-H-S) and the proposed method using binary masks (RPCA-H-B) outperformed RPCA

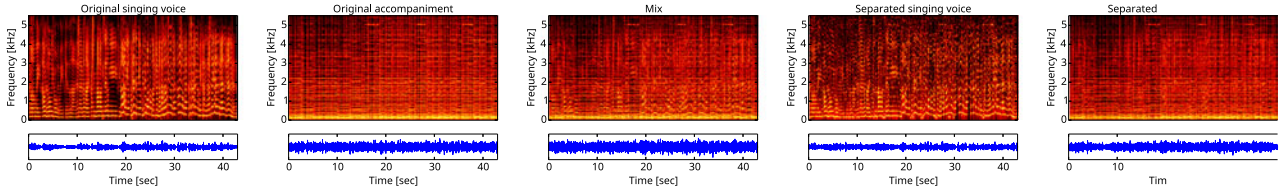


Fig. 5. An example of singing voice separation by the proposed method. The results of "Coldwar / LizNelson" in MedleyDB mixed at a −5 dB SNR are shown. From left to right, an original singing voice, an original accompaniment sound, a mixed sound, a separated singing voice, and a separated accompaniment sound are shown. The upper figures are spectrograms obtained by taking the STFT and the lower figures are resynthesized time signals.

TABLE III
EXPERIMENTAL RESULTS FOR VOCAL F0 ESTIMATION (AVERAGE ACCURACY [%] OVER ALL CLIPS IN EACH DATASET)

		PreFEst-V		MELODIA-V		MELODIA		
Database	SNR [dB]	w/o RPCA	w/ RPCA	w/o RPCA	w/ RPCA	w/o RPCA	w/ RPCA	Proposed
MIR-1K	−5	36.45	42.99	53.48	**60.69**	54.37	59.50	57.78
	0	50.70	56.15	76.88	**80.90**	78.09	79.91	75.48
	5	63.77	66.32	88.87	**90.26**	88.89	89.33	85.42
MedleyDB	original mix	70.83	72.25	70.69	74.93	71.24	73.40	**81.90**
	−5	71.82	72.72	72.05	76.75	74.56	75.32	**82.68**
	0	80.91	81.02	86.59	89.20	87.34	87.54	**90.31**
	5	86.39	85.41	92.63	**93.93**	93.08	92.50	93.15
RWC-MDB-P-2001		69.81	71.71	67.79	71.64	69.89	70.30	**80.84**
Average of all datasets		66.24	68.57	76.12	79.79	77.18	78.48	**80.95**

and H in terms of GNSDR in most settings. This indicates that extraction of harmonic structures is useful for singing voice separation in spite of F0 estimation errors and that combining an RPCA mask and a harmonic mask is effective for improving the separation quality of singing voices and accompaniment sounds. The removal of the spectra of non-repeating instruments (e.g., bass guitar) significantly improved the separation quality. two When vocal sounds are much louder than accompaniment sounds (MedleyDB, 5 dB SNR), H outperformed RPCA-H-B and RPCA-H-S in GNSDR. This indicates that RPCA masks tend to excessively remove the frequency components of vocal sounds in such a condition. RPCA-H-S outperformed RPCA-H-B in GNSDR, GSAR, and GSIR of the singing voice. On the other hand, RPCA-H-B outperformed RPCA-H-S in GSIR of the accompaniment and H outperformed both RPCA-H-B and RPCA-H-S. This indicates that a harmonic mask is useful for singing voice suppression.

Fig. 5 shows an example of an output of singing voice separation by the proposed method. We can see that vocal and accompaniment sounds were sufficiently separated from a mixed signal even though the volume level of vocal sounds was lower than that of accompaniment sounds.

B. Vocal F0 Estimation

We compared the vocal F0 estimation of the proposed method with conventional methods.

1) Datasets: MIR-1K, MedleyDB, and the RWC Music Database (*RWC-MDB-P-2001*) [45] were used for evaluating vocal F0 estimation. RWC-MDB-P-2001 contains 100 song clips of popular music which were recorded at a 44.1 kHz sampling rate with 16 bit resolution. The dataset contains 20 songs with English lyrics performed in the style of American popular music in the 1980s and 80 songs with Japanese lyrics performed in the style of Japanese popular music in the 1990s.

2) Compared Methods: The following four methods were compared.

1) *PreFEst-V*: PreFEst (saliency spectrogram) + Viterbi search
2) *MELODIA-V*: MELODIA (saliency spectrogram) + Viterbi search
3) *MELODIA*: The original MELODIA algorithm
4) *Proposed*: F0-saliency spectrogram + Viterbi (*proposed method*)

PreFEst [15] is a statistical multi-F0 analyzer that is still considered to be competitive for vocal F0 estimation. Although PreFEst contains three processes—the *PreFEst-front-end* for frequency analysis, the *PreFEst-core* computing a saliency spectrogram, and the *PreFEst-back-end* that tracks F0 contours using multiple agents—we used only the *PreFEst-core* and estimated F0 contours by using the Viterbi search described in Section III-B3 ("PreFEst-V"). *MELODIA* is a state-of-the-art algorithm for vocal F0 estimation that focuses on the characteristics of vocal F0 contours. We applied the Viterbi search to a saliency spectrogram derived from MELODIA ("MELODIA-V") and also tested the original MELODIA algorithm ("MELODIA"). In this experiment we used the MELODIA implementation provided as a vamp plug-in.[4]

Singing voice separation based on RPCA [5] was applied before computing conventional methods as preprocessing ("w/ RPCA" in Table III). We investigated the effectiveness of the proposed method in conjunction with preprocessing of singing voice separation.

[4]http://mtg.upf.edu/technologies/melodia

3) Evaluation Measures: We measured the raw pitch accuracy (RPA) defined as the ratio of the number of frames in which correct vocal F0s were detected to the total number of voiced frames. An estimated value was considered correct if the difference between it and the ground-truth F0 was 50 cents (half a semitone) or less.

4) Experimental Results: Table III shows the experimental results of vocal F0 estimation, where each value is an average accuracy over all clips. The results show that the proposed method achieved the best performance in terms of average accuracy. With MedleyDB and RWC-MDB-P-2001 the proposed method significantly outperformed the other methods, while the performance of MELODIA-V and MELODIA were better than that of the proposed method with MIR-1K. This might be due to the different instrumentation of songs included in each dataset. Most clips in MedleyDB and RWC-MDB-P-2001 contain the sounds of many kinds of musical instruments, whereas most clips in MIR-1K contain the sounds of only a small number of musical instruments.

These results are originated from the characteristics of the proposed method. In vocal F0 estimation, the spectral periodicity of an RPCA binary mask is used to enhance vocal spectra. The harmonic structures of singing voices appear clearly in the RPCA mask when music audio signals contain various kinds of repetitive musical instrument sounds. The proposed method therefore works well especially for songs of particular genres such as *rock* and *pops*.

C. MIREX2014

We submitted our algorithm to the *Singing Voice Separation* task of the Music Information Retrieval Evaluation eXchange (MIREX) 2014, which is a community-based framework for the formal evaluation of analysis algorithms. Since the datasets are not freely distributed to the participants, MIREX provides meaningful and fair scientific evaluations.

There is some difference between our submission for MIREX and the algorithm described in this paper. The major difference is that only an SHS spectrogram (with the exception of an F0 enhancement spectrogram in Section III-B2) was used as a saliency spectrogram in the submission. Instead a simple VAD method based on an energy threshold was used after singing voice separation.

1) Dataset: 100 monaural clips of pop music recorded at 44.1-kHz sampling rate with 16-bit resolution were used for evaluation. The duration of each clip was 30 seconds.

2) Compared Methods: 11 submissions participated in the task.[5] The submissions **HKHS1**, **HKHS2** and **HKHS3** are algorithms using deep recurrent neural networks [28]. **YC1** separates singing voices by clustering modulation features [27]. **RP1** is the REPET-SIM algorithm that identifies repetitive structures in polyphonic music by using a similarity matrix [8]. **GW1** uses Bayesian NMF to model a polyphonic spectrogram, and clusters the learned bases based on acoustic features [23]. **JL1** uses the temporal and spectral discontinuity of singing voices [26], and **LFR1** uses light kernel additive modeling based on the algorithm in [30]. **RNA1** first estimates predominant F0s and then reconstructs an isolated vocal signal based on harmonic sinusoidal modeling using estimated F0s. **IIY1** and **IIY2** are our submissions. The only difference between IIY1 and IIY2 is their parameters. The parameters for both submissions are listed in Table IV.

[5] www.music-ir.org/mirex/wiki/2014:Singing_Voice_Separation_Results

TABLE IV
PARAMETER SETTINGS FOR MIREX2014

	Window size	Hopsize	N	λ	w
IIY1	4096	441	15	1.0	100
IIY2	4096	441	15	0.8	100

3) Evaluation Results: Fig. 6 shows the evaluation results for all submissions. Our submissions (IIY1 and IIY2) provided the best mean NSDR for both vocal and accompaniment sounds. Even though the submissions using the proposed method outperformed the state-of-the-art methods in MIREX 2014, there is still room for improving their performances. As described in Section V-A, the robust range for the parameter w is from 40 to 60. We set the parameter to 100 in the submissions, however, and that must have considerably reduced the sound quality of both separated vocal and accompaniment sounds.

V. PARAMETER TUNING

In this section we discuss the effects of parameters that determine the performances of singing voice separation and vocal F0 estimation.

A. Singing Voice Separation

The parameters λ and w affect the quality of singing voice separation. λ is the sparsity factor of RPCA described in Section III-A1 and w is the frequency width of the harmonic mask described in Section III-A2. The parameter λ can be used to trade off the rank of a low-rank matrix with the sparsity of a sparse matrix. The sparse matrix is sparser when λ is larger and is less sparse when λ is smaller. When w is smaller, fewer spectral bins around an F0 and its harmonic partials are assigned as singing voices. This is the recall-precision trade-off of singing voice separation. To examine the relationship between λ and w, we evaluated the performance of singing voice separation for combinations of λ from 0.6 to 1.2 in steps of 0.1 and w from 20 to 90 in steps of 10.

1) Experimental Conditions: MIR-1K was used for evaluation at three mixing conditions with SNRs of -5, 0, and 5 dB. In this experiment, a harmonic mask was created using a ground-truth F0 contour to examine only the effects of λ and w. GNSDRs were calculated for each parameter combination.

2) Experimental Results: Fig. 7 shows the overall performance for all parameter combinations. Each unit on a grid represents the GNSDR value. It was shown that λ from 0.6 to 1.0 and w from 40 to 60 provided robust performance in all mixing conditions. In the -5 dB mixing condition, an integrated mask performed better for both of the singing voice and the

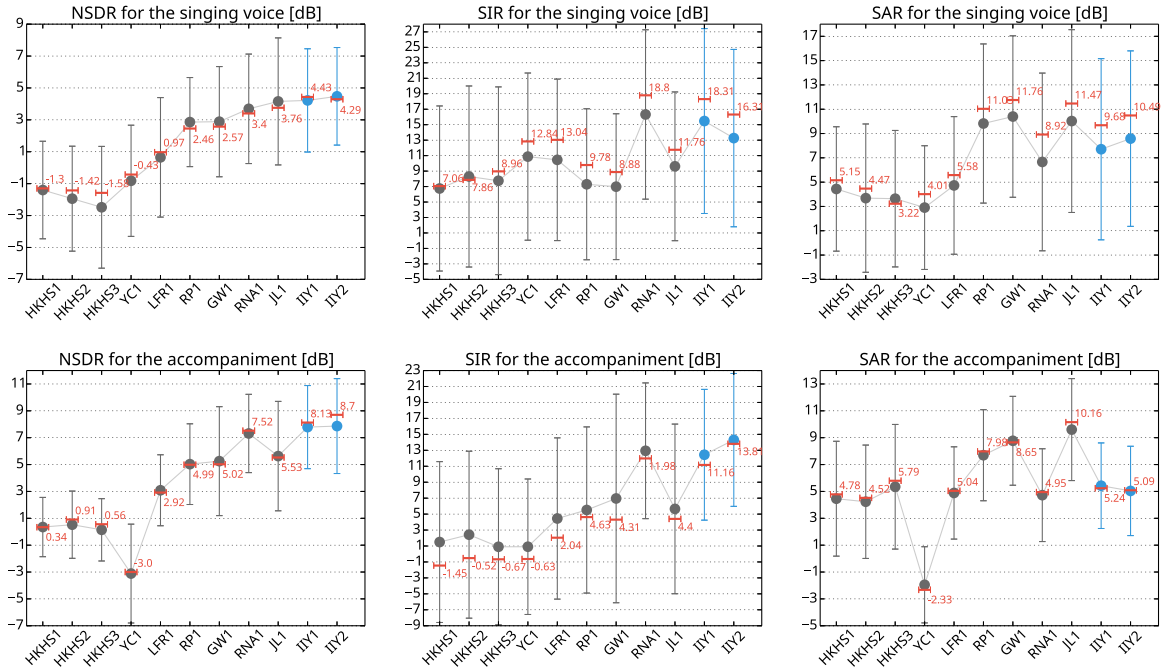


Fig. 6. Results of the *Singing Voice Separation* task in MIREX2014. The circles, error bars, and red values represent means, standard deviations, and medians for all song clips, respectively.

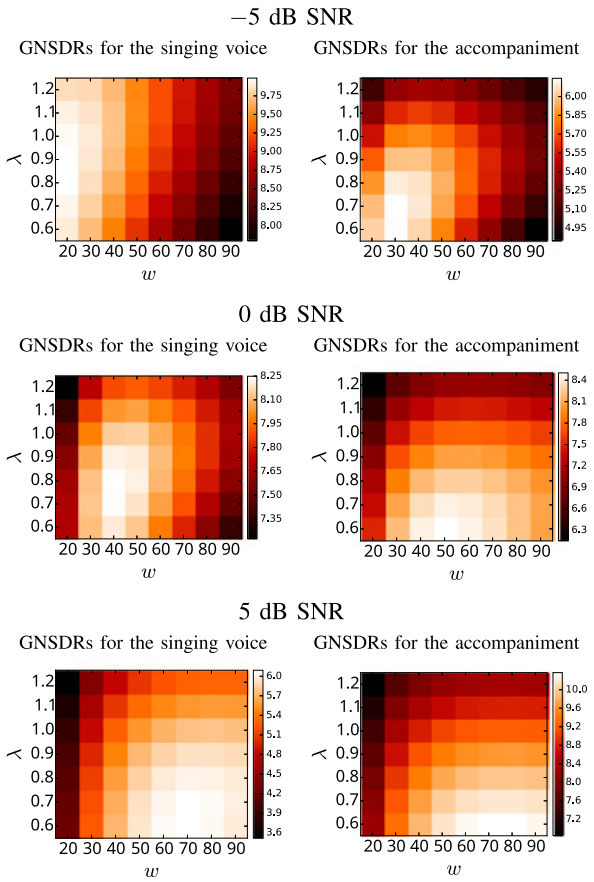


Fig. 7. Experimental results of grid search for singing voice separation. GNSDR for MIR-1K is shown in each unit. From top to bottom, the results of −5, 0, and 5 dB SNR conditions are shown. The left figures show results for the singing voice and the right figures for the music accompaniment. In all parts of this figure, lighter values represent better results.

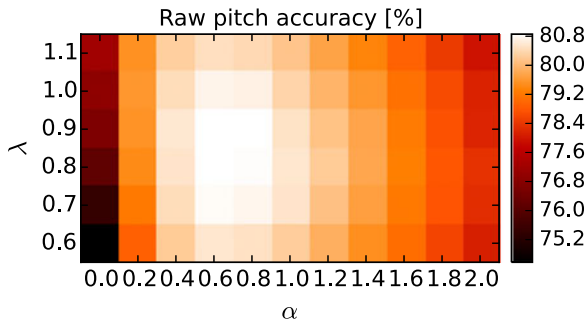


Fig. 8. Experimental results of grid search for vocal F0 estimation. The mean raw pitch accuracy for RWC-MDB-P-2001 is shown in each unit. Lighter values represent better accuracy.

accompaniment when w was smaller. This was because most singing voice spectra were covered by accompaniment spectra and only few singing voice spectra were dominant around an F0 and harmonic partials in the condition.

B. Vocal F0 Estimation

The parameters λ and α affect the accuracy of vocal F0 estimation. λ is the sparsity factor of RPCA and α is the weight parameter for computing the F0-saliency spectrogram described in Section III-B2. α determines the balance between an SHS spectrogram and an F0 enhancement spectrogram in a F0-saliency spectrogram, and there must be range of its value that provides robust performance. We evaluated the accuracy of singing voice separation for combinations of λ from 0.6 to 1.1 in steps of 0.1 and α from 0 to 2.0 in steps of 0.2. RWC-MDB-P-2001 was used for evaluation, and RPA was measured for each parameter combination.

Fig. 8 shows the overall performance for all parameter combinations of grid search. Each unit on a grid represents RPA for each parameter combination. It was shown that λ from 0.7 to 0.9 and α from 0.6 to 0.8 provided comparatively better performance than any other parameter combinations. RPCA with λ within the range separates vocal sounds to a moderate degree for vocal F0 estimation. The value of α was also crucial to estimation accuracy. The combinations with $\alpha = 0.0$ yielded especially low RPAs. This indicates that an F0 enhancement spectrogram was effective for vocal F0 estimation.

VI. Conclusion

This paper described a method that performs singing voice separation and vocal F0 estimation in a mutually-dependent manner. The experimental results showed that the proposed method achieves better singing voice separation and vocal F0 estimation than conventional methods do. The singing voice separation of the proposed method was also better than that of several state-of-the-art methods in MIREX 2014, which is an international competition in music analysis. In the experiments on vocal F0 estimation, the proposed method outperformed two conventional methods that are considered to achieve the state-of-the-art performance. Some parameters of the proposed method significantly affect the performances of singing voice separation and vocal F0 estimation, and we found that a particular range of those parameters results in relatively good performance in various situations.

We plan to integrate singing voice separation and vocal F0 estimation in a unified framework. Since the proposed method performs these tasks in a cascading manner, separation and estimation errors are accumulated. One promising way to solve this problem is to formulate a unified likelihood function to be maximized by interpreting the proposed method from a viewpoint of probabilistic modeling. To discriminate singing voices from musical instrument sounds that have sparse and non-repetitive structures in the TF domain like singing voices, we attempt to focus on both the structural and timbral characteristics of singing voices as in [35]. It is also important to conduct subjective evaluation to investigate the relationships between the conventional measures (SDR, SIR, and SAR) and the perceptual quality.

References

[1] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 1441–1444.

[2] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 3933–3936.

[3] Y. Ohishi, D. Mochihashi, H. Kameoka, and K. Kashino, "Mixture of Gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3714–3718.

[4] H. Fujihara and M. Goto, "Concurrent estimation of singing voice F0 and phonemes by using spectral envelopes estimated from polyphonic music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 365–368.

[5] P. S. Huang, S. D. Chen, P. Smaragdis, and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 57–60.

[6] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, 1988.

[7] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 22, no. 12, pp. 1884–1893,Dec. 2014.

[8] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Oct. 2012, pp. 583–588.

[9] Z. Duan and B. Pardo, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[10] C. Palmer and C. L. Krumhansl, "Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity," *Perception Psychophysics*, vol. 41, no. 6, pp. 505–518, 1987.

[11] A. Friberg and S. Ahlbäck, "Recognition of the main melody in a polyphonic symbolic score using perceptual knowledge," *J. New Music Res.*, vol. 38, no. 2, pp. 155–169, 2009.

[12] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 1885–1888.

[13] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1252–1261, Oct. 2011.

[14] B. Lehner, G. Widmer, and S. Böck, "A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks," in *Proc. Eur. Signal Process. Conf.*, 2015, pp. 21–25.

[15] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, 2004.

[16] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. Audio, Speech, Language Process*, vol. 18, no. 8, pp. 2145–2154, Nov. 2010.

[17] K. Dressler, "An auditory streaming approach for melody extraction from polyphonic music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 19–24.

[18] V. Arora and L. Behera, "On-line melody extraction from polyphonic audio using harmonic cluster tracking," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 21, no. 3, pp. 520–530, Mar. 2013.

[19] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 6, pp. 1759–1770, Aug. 2012.

[20] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Norwell, MA, USA: Kluwer, 2005, pp. 181–197.

[21] A. Chanrungutai and C. A. Ratanamahatan, "Singing voice separation in mono-channel music using non-negative matrix factorization," in *Proc. Int. Conf. Adv. Technol. Commun.*, 2008, pp. 243–246.

[22] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 21, no. 10, pp. 2096–2107, Oct. 2013.

[23] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 507–512.

[24] 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 Trans. Audio, Speech, Lang. Process*, vol. 22, no. 1, pp. 228–237, Jan. 2014.

[25] D. Fitzgerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Trans. Electron. Signal Process.*, vol. 4, no. 1, pp. 62–73, 2010.

[26] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *Signal Process. Lett.*, vol. 21, no. 10, pp. 1197–1200, 2014.

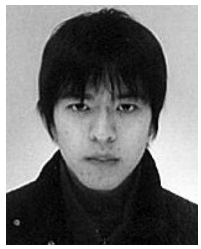
[27] F. Yen, Y.-J. Luo, and T.-S. Chi, "Singing voice separation using spectro-temporal modulation features," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 617–622.

[28] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 477–482.

[29] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 21, no. 1, pp. 71–82, Jan. 2013.

[30] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[31] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 126–130.
[32] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. ISCA Tutorial Res. Workshop Statistical Perceptual Audition*, 2008, pp. 17–20.
[33] C. L. Hsu and J. R. Jang, "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 525–530.
[34] T.-C. Yeh, M.-J. Wu, J.-S. Jang, W.-L. Chang, and I.-B. Liao, "A hybrid approach to singing pitch extraction based on trend estimation and hidden Markov models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 457–460.
[35] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.
[36] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, no. 4, pp. 1475–1487, May 2007.
[37] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 3, pp. 638–648, Mar. 2010.
[38] C. L. Hsu, D. Wang, J. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 5, pp. 1482–1491, Jul. 2012.
[39] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.
[40] P. Cabañas-Molero, D. M. Muñoz, M. Cobos, and J. J. López, "Singing voice separation from stereo recordings using spatial clues and robust F0 estimation," in *Proc. AEC Conf.*, 2011, pp. 239–246.
[41] Y. M. Z. Lin and M. Chen, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Math. Program.*, 2009.
[42] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2007, pp. 373–374.
[43] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 155–160.
[44] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[45] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2002, pp. 287–288.



Yukara Ikemiya received the B.S. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2013 and 2015, respectively. He is currently working for an electronics manufacturer in Japan. His research interests include music information processing and speech signal processing. He has attained the best result in the Singing Voice Separation task of MIREX 2014. He is a Member of the Information Processing Society of Japan.

Katsutoshi Itoyama (M'13) received the B.E. degree, the M.S. degree in informatics, and the Ph.D. degree in informatics, all from Kyoto University, Kyoto, Japan, in 2006, 2008, 2011, respectively. He is currently an Assistant Professor at the Graduate School of Informatics, Kyoto University, Japan. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He received the 24th TAF Telecom Student Technology Award and the IPSJ Digital Courier Funai Young Researcher Encouragement Award. He is a Member of the IPSJ and ASJ.

Kazuyoshi Yoshii received the Ph.D. degree in informatics from Kyoto University, Japan, in 2008. He is currently a Senior Lecturer at Kyoto University. His research interests include music signal processing and machine learning. He has received several awards including the IPSJ Yamashita SIG Research Award and the Best-in-Class Award of MIREX 2005. He is a Member of the Information Processing Society of Japan and Institute of Electronics, Information, and Communication Engineers.