Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms

Hideyuki Tachibana, *Student Member, IEEE,* Nobutaka Ono, *Member, IEEE,* and Shigeki Sagayama, *Member, IEEE*

***Abstract*—We propose a novel singing voice enhancement technique for monaural music audio signals, which is a quite challenging problem. Many singing voice enhancement techniques have been proposed recently. However, our approach is based on a quite different idea from these existing methods. We focused on the fluctuation of a singing voice and considered to detect it by exploiting two differently resolved spectrograms, one has rich temporal resolution and poor frequency resolution, while the other has rich frequency resolution and poor temporal resolution. On such two spectrograms, the shapes of fluctuating components are quite different. Based on this idea, we propose a singing voice enhancement technique that we call two-stage harmonic/percussive sound separation (HPSS). In this paper, we describe the details of two-stage HPSS and evaluate the performance of the method. The experimental results show that SDR, a commonly-used criterion on the task, was improved by around 4 dB, which is a considerably higher level than existing methods. In addition, we also evaluated the performance of the method as a preprocessing for melody estimation in music. The experimental results show that our singing voice enhancement technique considerably improved the performance of a simple pitch estimation technique. These results prove the effectiveness of the proposed method.**

***Index Terms*—singing voice enhancement, multiple resolution, non-stationarity, fluctuation, pitch detection, harmonic and percussive sound separation,**

I. Introduction

THIS paper describes a novel idea to extract singing voices from polyphonic music signals. In many genres of music, especially in the popular musics, the lead vocal is the most impressive and essential part for most listeners, and moreover, it often has much information that is important in music information retrieval (MIR) applications. In fact, many MIR studies, such as automatic lyrics recognition [1], [2], identification of the language of a song [3], automatic singer identification [4], etc., have used the information on singing voices. In addition to the importance as a preprocessing for MIR applications, furthermore, it is also significant in itself in the way that the technique can be applied as a kind of interactive music player, i.e., a vocal/nonvocal equalizer, an automatic karaoke generator [5] and etc.

Manuscript received October 21, 2012; revised March 31, 2013; revised August 17, 2013. This work was supported by the JSPS (Japan Society for the Promotion of Science) Grand-In-Aid No. 22-6961. The associate editors coordinating the review of this manuscript and approving it for publication was Dr. xxxxxx.

H. Tachibana is with the Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo, 113-8656, Japan. (e-mail: tachibana@hil.t.u-tokyo.ac.jp).

N. Ono and S. Sagayama are with the National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan (e-mail: onono@nii.ac.jp, sagayama@nii.ac.jp)

Digital Object Identifier 10.1109/TASLP.2013.2287052

Along with the possibilities of the applications, its technical difficulties also make singing voice enhancement an interesting problem. One of the difficulties comes from the similarity between singing voice and accompaniments, e.g., a piano, a guitar, and percussions. For instance, both the spectra of singing voice and harmonic instruments, such as a piano and a guitar, have harmonic structure. Accordingly, it is difficult for a simple harmonics-extraction technique to detect only the singing voice in polyphonic music signals. Another difficulty is that accompanying instruments do not satisfy some properties of "noise" that have been supposed in conventional signal processing problems, e.g., whiteness and stationarity, and therefore, we cannot expect that a classical noise suppression technique work effectively in singing voice enhancement, because music signals are not white noise nor stationary.

Because of the many potential applications as well as the technical interests described above, many methods on singing voice enhancement in music signals, and other related techniques including singing melody transcription (see section I-A), have been actively studied recently. In most of the existing methods, an input music signal is first transformed from time domain to time-frequency domain, then singing voice is characterized there. Other components such as accompanying instruments are suppressed with time-frequency masking (adaptive Wiener filtering), and finally, the estimated spectrogram of singing voice is transformed back to time domain again. The most important point is how to distinguish the singing voice component from others in a time-frequency representation. Ozerov et al. [6], [7] focused on the difference of spectral distribution (timbre) of singing voice and instruments, and modeled them by Gaussian mixture model. In their method, the GMM was trained in advance in a supervised way, and tuned adaptively for each input. Some studies utilized the pitch information of singing voice. In Li and Wang's method [8], segments including singing voice were first detected based on spectral features. Then, at each of the detected singing-voice segments, the predominant pitch was estimated with using the autocorrelation and thresholding. Hsu and Jang [9] extended this approach to enable to capture unvoiced components of the singing voice with utilizing the spectral envelope information. Another popular stream is based on Non-negative matrix factorization (NMF) of music spectrogram [10], where it is assumed that spectrogram of music can be expressed as an assemblage of a limited number of spectral templates. In Vembu and Baumann's method [11], spectral templates obtained by NMF were classified into singing voice and others with their spectral features such as MFCC, LFPC, and PLP. Virtanen et al. [12] utilized NMF with pitch inference. In their method, the pitch of singing voice was first estimated based on multiple F_0 estimation technique [13], then, the singing voice was roughly removed based on the pitch, and the residual was used for training accompaniment model with NMF. Finally, the singing voice was extracted from the mixture using the derived accompaniment model. In addition to NMF-based approaches, some of other studies also have focused on low-rankness

of music spectrogram. Huand et al. [14], in their PCA-based method, assumed that the spectrogram of accompaniment would lie in lowrank subspace while singing voice would not, as accompaniments are rather repetitive while singing voice are less so. Rafii and Pardo [15] proposed a method "REPET" that suppressed repeating components in spectrogram, i.e., accompaniments. Raj et al. [16] modeled an NMF-like generative signal model and applied probabilistic inferences. Some approaches are based on the harmonicity of singing voice. A method which utilized the harmonicity was proposed by Lagrange et al. [17], in which a technique of computer vision is utilized to pick up the harmonically related spectral peaks of the short-time spectra of singing voice. In summary, majority of the state-of-the-art singing voice extraction techniques considered to extract singing voice on a time-frequency domain utilizing some properties on singing voice such as timbrel features, high-rankness, harmonicity, etc.

In this paper, we propose another approach for singing voice enhancement, focusing on the fluctuation of singing voice, such as vibrato [20]–[22]. In order to capture the fluctuation, we exploit two spectrogram representations with different time-frequency resolutions, which are unlike the existing methods. Our motivation for using two different spectrograms comes from our observation that singing voice has an "intermediate" property between other harmonic instruments and percussive instruments. That is, a singing voice appears similarly to harmonic instruments on an ordinary spectrogram that has 10–30 [ms] temporal resolution, while it should appear rather similar to percussions if the analyzing frame of short-time Fourier transform (STFT) is much longer than the temporal scale of the fluctuation of singing voice. On the basis of the idea, we roughly define three types of musical components, fluctuating, sustained, and transient. Those three types of components can be separated by applying a simple algorithm twice on differently-resolved spectrograms, which separates sinusoidal components and impulsive components, which is called harmonic/percussive sound separation (HPSS) [18], [23]–[25]. In this paper, we describe the details of the discussion above.

According to the experiments we conducted in order to evaluate the performance of the method compared with those of existing methods, it is verified that the method extracts singing voice effectively, indicating around 4 dB Signal to Distortion Ratio (SDR) [6], [26] improvement, which is a considerably higher level than the other methods. In addition, we also describe a straightforward combination of the proposed method and a simple pitch estimation technique, which is one of the possible applications of our singing voice enhancement technique. Experimental results on this task also show the effectiveness of the proposed method.

Note that this paper is the extended version of our previous conference papers [27], [28]. After the conference, Hsu et al. [29] have developed more effective singing voice enhancement technique on the basis of the concept of our previous papers [27], [28] as well as their pitch estimation techniques, but this paper can be placed as the complete description on our proposal method "two-stage HPSS." Another note is that a similar idea was also mentioned by FitzGerald [30] around the same time as our previous works [27], [28].

A. Related Work

In addition to singing voice enhancement, there are many studies that principally focused on the predominant pitch estimation from the mixed music signals, which can be used as preprocessing for designing time-frequency mask, or can be an application of singing voice enhancement. One of the earliest works that addressed the task was PreFEst [31]. Besides PreFEst, several melody tracking algorithms have since been proposed, e.g., the methods by Fujihara et al. [32], Cao et al. [33], Durrieu et al. [34], [35], Salamon and Gómez [36], and Hsu et al. [29], [37] which is a tandem connection of their singing pitch estimation technique based on the trend estimation and a part of our early studies [27], [28]. Some of the other singing pitch transcription methods, e.g., Ryynänen and Klapuri's method [13] and V. Rao and P. Rao's method [38] are based on multiple F_0 estimation, which is also an important topic in the music signal processing area [39], [40]. In addition to the studies above, an exchange on audio melody extraction has been held since 2005 as a part of MIREX (music information retrieval exchange) [41], and many participants have submitted their algorithms to the exchange, including those mentioned above [42].

B. Definition, Notation and Paper Outline

Let $x(t)$ be a real-valued signal, where $t \in \mathbb{Z}, 0 \leq t < f_\mathrm{s}T$ is the discrete time, f_s [Hz] is the sampling rate, and T [s] is the length of the signal. Let us define $x(t) = 0$ when $t < 0, f_\mathrm{s}T \leq t$.

Although there are many ways to represent a signal on time-frequency domain, we simply used a short-time discrete Fourier transform (STFT), which is one of the simplest method, and is invertible with ease. $\tilde{\boldsymbol{X}} = (\tilde{X}_{n,k})_{(n,k)\in\Omega} := \mathrm{STFT}_l[x(t)]$ denotes the complex spectrogram of a signal $x(t)$, where $l \in \mathbb{N}$ denotes the frame length (size of analyzing window) of STFT. For convenience, we assume l is even. Each element $\tilde{X}_{n,k} \in \mathbb{C}$ is defined by discrete Fourier transform (DFT) as follows,

$$\tilde{X}_{n,k} := \sum_{t=0}^{l-1} x(t + ns - l/2) g_1(t) e^{-2\pi j t k / l}, \tag{1}$$

where s is the size of frame shift, which is coordinated with l as $s = l/2$ in this paper, $g_1(t)$ is a window function, and j is the imaginary unit. The subscripts $n, k \in \mathbb{Z}, 0 \leq k < l$ denote the indices of time and frequency, respectively. Note, we can restrict the frequency domain as $0 \leq k < K := l/2 + 1$, because of the redundancy $\tilde{X}_{n,l-k} = \tilde{X}^*_{n,k}$ (complex conjugate). In addition, the values of the spectrogram $\tilde{X}_{n,k}$ outside of the domain $0 \leq n < N := \lceil 2 f_\mathrm{s} T / l \rceil + 1$, is $\tilde{X}_{n,k} = 0$ by definition. Thus we can write the domain Ω as $\Omega = \{(n,k) | n = 0, 1, \ldots, N-1, k = 0, 1, \ldots, K-1\}$, and regard a complex spectrogram $\tilde{\boldsymbol{X}}$ as an element of $\mathbb{C}^{N \times K}$.

The temporal resolution and the frequency resolution of a spectrogram are $s/f_\mathrm{s} (= l/2f_\mathrm{s})$ [s] and f_s/l [Hz], respectively. Thus the product of them is always $1/2$ regardless of the value of l. This fact forms the basis of the proposal method which shall be discussed in section III.

The inverse STFT ($\mathrm{STFT}_l^{-1}[\tilde{\boldsymbol{X}}]$) is defined on the basis of the inverse DFT and the overlap-add (OLA) using a reconstruction window function $g_2(t)$.

Arithmetic operations on spectrograms indicate element-wise operations, e.g., $\boldsymbol{X}/2 = (X_{n,k}/2)$, $\boldsymbol{X} + \boldsymbol{Y} = (X_{n,k} + Y_{n,k})$, $e^{j\angle\tilde{\boldsymbol{X}}} = (e^{j\angle\tilde{X}_{n,k}}) = (\tilde{X}_{n,k}/|\tilde{X}_{n,k}|)$, and $\sqrt{\boldsymbol{X}} = (\sqrt{X_{n,k}})$, etc., where $(X_{n,k})$ is the abbreviation of $(X_{n,k})_{(n,k)\in\Omega}$. For example, $\boldsymbol{X} = |\tilde{\boldsymbol{X}}|^2 \in \mathbb{R}^{N \times K}$ is the squared amplitude of the complex spectrogram $\tilde{\boldsymbol{X}}$, which is called power spectrogram. When we simply write 'spectrogram,' it means power spectrogram in this paper.

The rest of this paper is organized as follows. In section II, we briefly introduce HPSS, a fundamental technique of this paper. In section III, we discuss the fluctuation of singing voice and its effects on the shapes of spectrogram. Moreover, on the basis of the discussion, we describe "two-stage HPSS," which is the main subject of this paper. In section IV, the performance of the proposed method is described. In addition, the effectiveness of our method as a preprocessing for melody estimation is also described. Finally, section V concludes the paper.

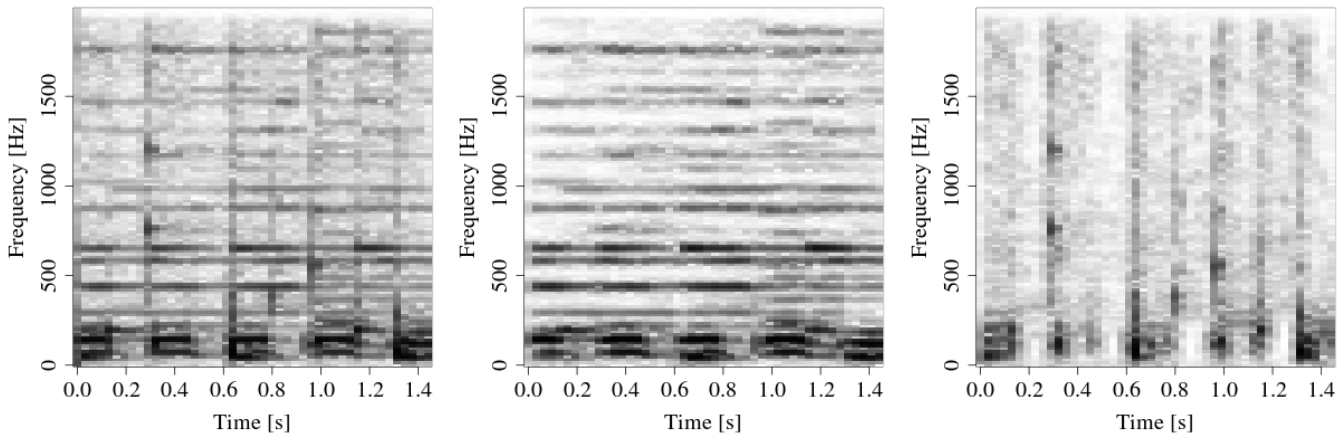


Fig. 1. An example of HPSS separation [18]. (a) Spectrogram of an input music signal, excerpted from RWC-MDB-P-2001 No. 14, RWC Music Database [19]. (b) Spectrogram of the separated H component. (c) Spectrogram of the separated P component. Each spectrogram was obtained under the following condition: frame length was 64 [ms], frame overlap is 32 [ms], and the window function was Hanning window.

Algorithm 1 HPSS updating formulae

procedure HPSS$(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{m}; \boldsymbol{W}, i)$
  $a_1 \leftarrow 2(1+\sigma_{\mathrm{H}}^{-2})$
  $a_2 \leftarrow 2(1+\sigma_{\mathrm{P}}^{-2})$
  for $\forall(n,k), i \le n < i+I, 0 \le k < K$ **do**
    $b_1 \leftarrow \sigma_{\mathrm{H}}^{-2}\left(\sqrt{H_{n-1,k}} + \sqrt{H_{n+1,k}}\right)$
    $b_2 \leftarrow \sigma_{\mathrm{P}}^{-2}\left(\sqrt{P_{n,k-1}} + \sqrt{P_{n,k+1}}\right)$
    $c_1 \leftarrow 2m_{n,k}W_{n,k}$
    $c_2 \leftarrow 2\left(1-m_{n,k}\right)W_{n,k}$
    $H_{n,k} \leftarrow \left(\dfrac{b_1 + \sqrt{b_1^2 + 4a_1c_1}}{2a_1}\right)^2$
    $P_{n,k} \leftarrow \left(\dfrac{b_2 + \sqrt{b_2^2 + 4a_2c_2}}{2a_2}\right)^2$
    $m_{n,k} \leftarrow \dfrac{H_{n,k}}{H_{n,k}+P_{n,k}}$
  end for
end procedure

Algorithm 2 Whole procedure of Sliding HPSS

Preprocessing:
Given an input signal $w(t)$
$\tilde{\boldsymbol{W}} \leftarrow \mathrm{STFT}_l[w(t)]$ ▷ complex spectrogram
$\boldsymbol{W} \leftarrow |\tilde{\boldsymbol{W}}|^2$ ▷ power spectrogram
$\boldsymbol{H} \leftarrow \boldsymbol{W}/2$ ▷ initial value of $\boldsymbol{H}$
$\boldsymbol{P} \leftarrow \boldsymbol{W}/2$ ▷ initial value of $\boldsymbol{P}$
$\forall(n,k), m_{n,k} \leftarrow 0.5$ ▷ initial value of $\boldsymbol{m}$

HPSS Updating based on Sliding Analysis:
for $-I \le i \le N+I$ **do**
  HPSS $(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{m}; \boldsymbol{W}, i)$ ▷ Algorithm 1
end for

Postprocessing:
$\tilde{\boldsymbol{H}} \leftarrow \sqrt{\boldsymbol{m}\boldsymbol{W}}e^{j\angle\tilde{\boldsymbol{W}}}$ ▷ Wiener masking
$\tilde{\boldsymbol{P}} \leftarrow \sqrt{(\boldsymbol{1}-\boldsymbol{m})\boldsymbol{W}}e^{j\angle\tilde{\boldsymbol{W}}}$
$h(t) \leftarrow \mathrm{STFT}_l^{-1}[\tilde{\boldsymbol{H}}]$ ▷ waveform synthesis
$p(t) \leftarrow \mathrm{STFT}_l^{-1}[\tilde{\boldsymbol{P}}]$

II. Harmonic/Percussive Sound Separation

In this section, we make a brief introduction of Harmonic/Percussive Sound Separation (HPSS) [18], [23]–[25], which is a fundamental technique of our singing voice enhancement technique. HPSS is an algorithm that separates a signal $w(t)$ into two classes of components: "harmonic" (sinusoidal) components $h(t)$ and "percussive" (impulsive) components $p(t)$ as follows,

$$w(t) \approx h(t) + p(t). \tag{2}$$

The method separates a signal on power spectrogram domain on the basis of two assumptions. The first assumption is that the power spectrograms of $h(t)$ and $p(t)$, i.e., $\boldsymbol{H} = (H_{n,k})_{(n,k)\in\Omega} \in \mathbb{R}^{N\times K}$ and $\boldsymbol{P} = (P_{n,k})_{(n,k)\in\Omega} \in \mathbb{R}^{N\times K}$, are "smooth" in time and in frequency, respectively. The assumption reflects the nature of harmonic components and percussive components. That is, the harmonic components are rather "smooth" in time, because they are sustained for a while, while percussive components are rather "smooth" in frequency, because they are instantaneous.

Specifically, we defined the smoothness of power spectrogram $\boldsymbol{X} \in \mathbb{R}^{N\times K}$ in time and in frequency as follows,

$$\boldsymbol{X} \text{ is "smooth in time"} \quad \text{when } X_{n,k} \approx X_{n-1,k} \tag{3}$$
$$\boldsymbol{X} \text{ is "smooth in frequency"} \quad \text{when } X_{n,k} \approx X_{n,k-1} \tag{4}$$

On the basis of the definition of "smoothness," we defined criteria to measure how strongly (3) and (4) are satisfied as follows,

$$S_{\mathrm{H}}(\boldsymbol{H}) := \frac{1}{2\sigma_{\mathrm{H}}^2}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}(H_{n,k}^{\gamma} - H_{n-1,k}^{\gamma})^2, \tag{5}$$

$$S_{\mathrm{P}}(\boldsymbol{P}) := \frac{1}{2\sigma_{\mathrm{P}}^2}\sum_{n=0}^{N-1}\sum_{k=1}^{K-1}(P_{n,k}^{\gamma} - P_{n,k-1}^{\gamma})^2, \tag{6}$$

where σ_{H} and σ_{P} are weighting parameters, which were defined empirically as 0.3 in [24], and γ is an exponential factor, which should be 0.5 to make the objective function J that shall be defined later homogeneous (see Appendix.) It is easily confirmed that the value of $S_{\mathrm{H}}(\boldsymbol{H})$ should be small if a spectrogram $\boldsymbol{H}$ is "smooth" in time. Similarly, the value of $S_{\mathrm{P}}(\boldsymbol{P})$ should be small if a spectrogram $\boldsymbol{P}$ is "smooth" in frequency.

The second assumption is that the sum of the separated power spectrograms, i.e., $\boldsymbol{H}+\boldsymbol{P}$, should be almost equal to the original power spectrogram $\boldsymbol{W}=|\mathrm{STFT}_l[w(t)]|^2$. In order to evaluated the proximity of $\boldsymbol{H}+\boldsymbol{P}$ to $\boldsymbol{W}$, we used the generalized Kullbuck-Leibler divergence $D_{\mathrm{KL}}(\boldsymbol{W} \parallel \boldsymbol{H}+\boldsymbol{P})$, defined by

$$D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{Y}) := \sum_{n=0}^{N-1}\sum_{k=0}^{K-1} X_{n,k}\ln\frac{X_{n,k}}{Y_{n,k}} - X_{n,k} + Y_{n,k}. \quad (7)$$

On the basis of these assumptions, HPSS is formulated as an optimization problem to find the optimal spectrograms $\boldsymbol{H}$ and $\boldsymbol{P}$ that minimize the following objective function.

$$\begin{aligned} &\text{minimize } J(\boldsymbol{H},\boldsymbol{P}|\boldsymbol{W}) \\ &\qquad := S_{\mathrm{H}}(\boldsymbol{H}) + S_{\mathrm{P}}(\boldsymbol{P}) + D_{\mathrm{KL}}(\boldsymbol{W}\|\boldsymbol{H}+\boldsymbol{P}) \quad (8)\\ &\text{subject to } \forall(n,k), H_{n,k}\geq 0, P_{n,k}\geq 0. \end{aligned}$$

It is not necessarily easy to optimize $J(\boldsymbol{H},\boldsymbol{P}|\boldsymbol{W})$ directly, since its partial derivatives w.r.t. $H_{n,k}$ and $P_{n,k}$ result in rather complicated forms, but considering an auxiliary parameter $\boldsymbol{m}$ and an auxiliary function $J'(\boldsymbol{H},\boldsymbol{P},\boldsymbol{m}|\boldsymbol{W}) \geq J(\boldsymbol{H},\boldsymbol{P}|\boldsymbol{W})$ which is based on Jensen's inequality, the partial derivatives w.r.t. $H_{n,k}$ and $P_{n,k}$ become simple quadratic equations, and a simple iterative algorithm (Algorithm 1) that is similar to Expectation-Maximization (EM) algorithm is obtained. The detailed derivation is available in [24]. Thus we can separate a spectrogram $\boldsymbol{W}$ into two spectrograms $\boldsymbol{H},\boldsymbol{P}$. This is followed by Wiener filtering and inverse STFT to synthesize audible waveforms $h(t)$ and $p(t)$ ("Postprocessing," Algorithm 2). Hereinafter, let us simplify the notation of the whole procedure as follows,

$$w(t) \xrightarrow{\mathrm{HPSS}(l)} h(t), p(t). \quad (9)$$

The computation cost of each update (Algorithm 1) is not very large, and the solution rapidly converges to near the optimal value within a small number of iterations. Moreover, by the procedure shown in Algorithm 2, the HPSS algorithm is executable in real time. In real time processing, instead of applying updating formula of Algorithm 1 for several times, a sliding block of size I is used as shown in line 8–11, Algorithm 2. Practically, setting I around 10^1 to 10^2, a solution with a sufficient quality is obtained.

Fig. 1 shows an example of HPSS results. In the input spectrogram (a), horizontal and vertical structures are clearly observed. The spectrogram (b) shows the separated H component. It is shown in the spectrogram that most horizontal components in the original spectrogram (a) were separated into this component, while the vertical components were separated into the P components (spectrogram (c).)

III. Singing Voice Enhancement Based on Two-stage HPSS

In this section, we discuss the nature of a singing voice and consider how to extract it from music audio signals. On the basis of the discussion, we show the key idea of our singing voice extraction method, which is the novelty of this paper.

A. Nature of Singing Voice: Intermediate component between 'Harmonic' and 'Percussive'

As described in the introduction, a singing voice differs from instrumental sounds in many ways. For example, its timbrel information is quite different from many other types of instruments. Another difference is the depth of the fluctuation. In this paper, we focus on the fluctuation of singing voice to extract it from mixed music signals.

Let us consider the cases of a piano before a singing voice. Due to its mechanical structure, the pitch of piano tones and its harmonics basically do not change, or only slightly change if any, in a single note. We can also apply the same discussion to some other instruments such as a guitar, though it has more frequent exceptions (e.g., pitch bend) than a piano. As just described, the sound of the pitched instruments such as pianos and guitars basically do not have fluctuation, or have slight fluctuation if any.

A singing voice, unlike those instruments, typically has fluctuation of pitch and amplitude. Since the vocal cord is a human organ which is not as stable as artifacts, it does not generate sounds as flatly as the instruments above do, by and large. Besides the mechanical constraints, many singers fluctuate their singing voice for musical expressions. This fluctuation is called vibrato, which is another reason that a singing voice has much fluctuation than the instruments.

Because of the reasons above, we can assume that a singing voice is "less sinusoidal" than the instruments such as a piano and a guitar in many cases. At the same time, a singing voice is obviously "much more sinusoidal" than percussions. To summarize those two facts, we can regard singing voice as "intermediately sinusoidal," as well as "intermediately non-sinusoidal" sound, between sustained instruments and percussive instruments. In other words, we can consider a third class "intermediate component," which is typified by a singing voice, between "harmonic component" and "percussive component" in HPSS. We denote those three classes as follows,

- $\mathcal{H}$ Stationary, sustained, flatly-played instruments, (*e.g.,* piano, guitar),
- $\mathcal{V}$ Fluctuated quasi-stationary component (*e.g.,* singing voice),
- $\mathcal{P}$ Transient, non-stationary instruments, (*e.g.,* percussion).

B. Intermediate Component Extraction using two-stage HPSS

Let's consider how to extract intermediate component $\mathcal{V}$ with fluctuation from mixed audio signals. Our idea is the utilization of two different spectrograms that have different temporal-frequency resolutions. Note that the spectrogram shape of sound depends on temporal-frequency resolution, which can be controlled by STFT frame length.

First, let us consider a case in which the frame length of STFT is 10 [ms] (i.e., $l = 0.01 \times f_{\mathrm{s}}$). In this case, the frequency resolution of STFT is 50 Hz, because the product of temporal and frequency resolution is $1/2$. Because of its poor frequency resolution, small fluctuation of $\mathcal{V}$ falls in only a few of frequency bins, while the signal occupies many temporal bins because its pitch does not change within such a short duration. Therefore, its appearance on STFT is quite similar to $\mathcal{H}$ (middle row of Fig. 2) in terms of the smoothness. For this reason, when we apply HPSS to the spectrogram whose frame length is l_1 which is short enough, a signal $s(t)$ is roughly separated into $\mathcal{H}+\mathcal{V}$ and $\mathcal{P}$ as follows,

$$s(t) \xrightarrow{\mathrm{HPSS}(l_1)} h_1(t),\ p_1(t), \quad (10)$$

where

$$h_1(t) \approx \mathcal{H}+\mathcal{V}, \quad (11)$$

$$p_1(t) \approx \mathcal{P}. \quad (12)$$

Thus we can remove the $\mathcal{P}$ component from the mixed music signals.

Next, we have to decompose $h_1(t)$ into a harmonic component $\mathcal{H}$ and a singing voice $\mathcal{V}$. In order to achieve that, let us consider a case in which the frame length of STFT is 1 [s]. In this case, the frequency resolution of the spectrogram is 0.5 Hz. In contrast to the previous case, fluctuation of $\mathcal{V}$ is much broader than the frequency resolution, and it occupies many frequency bins in a single frame. Therefore, the appearance of $\mathcal{V}$ component on the spectrogram is not similar to $\mathcal{H}$ but to $\mathcal{P}$, as shown in the bottom row of Fig. 2. Consequently,

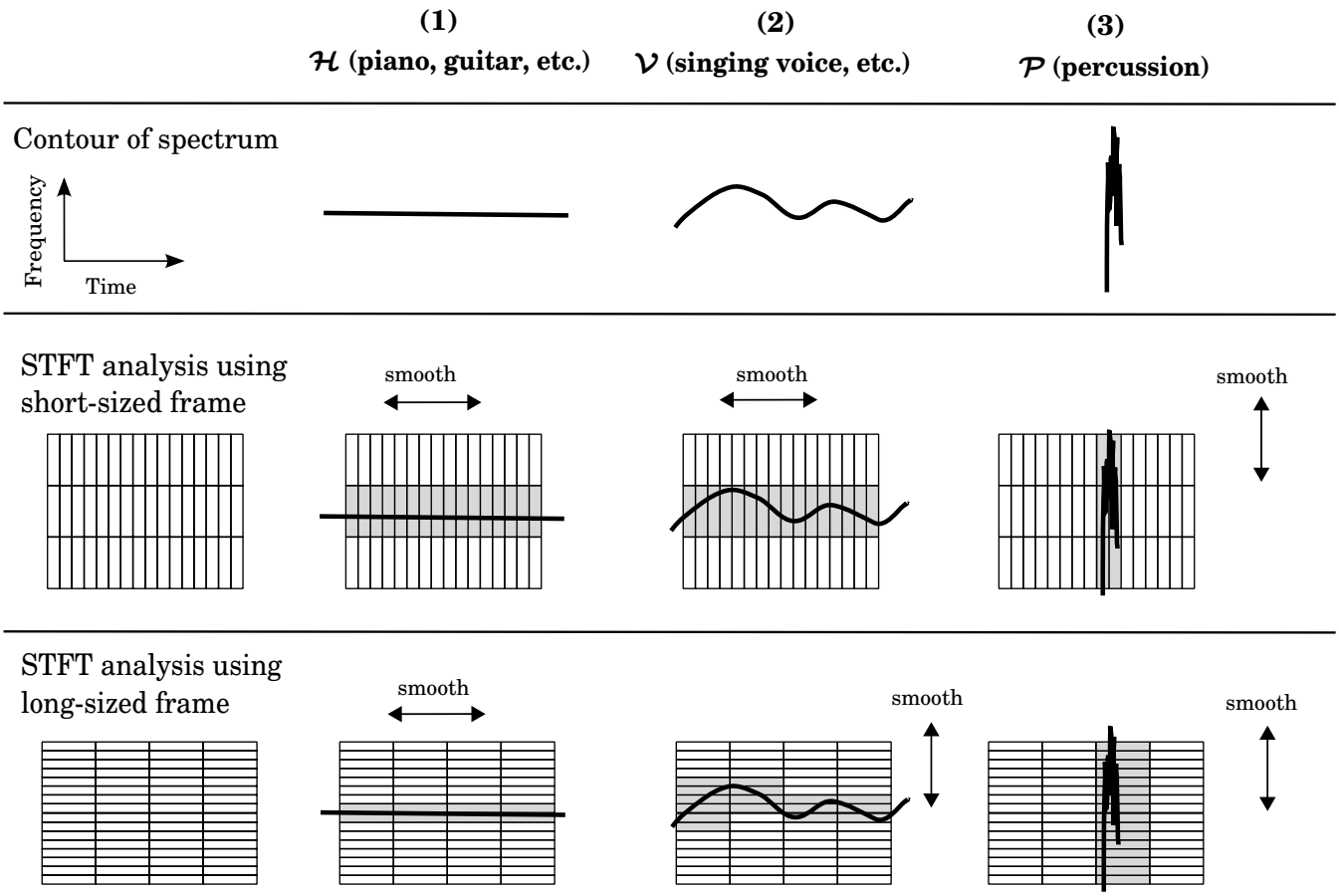


Fig. 2. The dependencies of the appearance of the spectrogram of three types of signals, on each spectrogram. (1) $\mathcal{H}$ component is smooth in time, nonsmooth in frequency, on both short- and long-framed STFT domains. (2) $\mathcal{V}$ component is smooth in time, nonsmooth in frequency, on short-framed STFT domain. However, on long-framed STFT domain, it is nonsmooth in time compared to the width of the time-frequency bins, and smooth in frequency compared to the height of the time-frequency bins. (Note that these figures are not necessarily the exact illustration of the effects of pitch fluctuation, but intuitive ones. Another point to note is that, not only the pitch fluctuation (frequency modulation) but also the amplitude fluctuation (amplitude modulation) expand the bandwidth of the spectrum.) (3) $\mathcal{P}$ component is almost always nonsmooth in time, and smooth in frequency, regardless of the frame length.

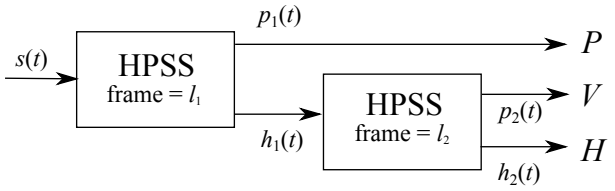


Fig. 3. Diagram of two-stage HPSS.

using a sufficiently long analyzing frame l_2, we can separate $h_1(t)$ into the following two components by HPSS,

$$h_1(t) \xrightarrow{\text{HPSS}(l_2)} h_2(t),\ p_2(t), \tag{13}$$

where

$$h_2(t) \approx \mathcal{H}, \tag{14}$$
$$p_2(t) \approx \mathcal{V}. \tag{15}$$

The obtained $p_2(t)$ roughly corresponds to the $\mathcal{V}$ component, which is the target component.

In summary, applying HPSS twice on differently-resolved two spectrograms separates a music signal into three components, $\mathcal{H}$, $\mathcal{V}$ and $\mathcal{P}$ as shown in Fig. 3, and thus obtained $p_2(t)$ would roughly be the singing voice, which we aimed at in this paper.

C. Experimental Example of two-stage HPSS

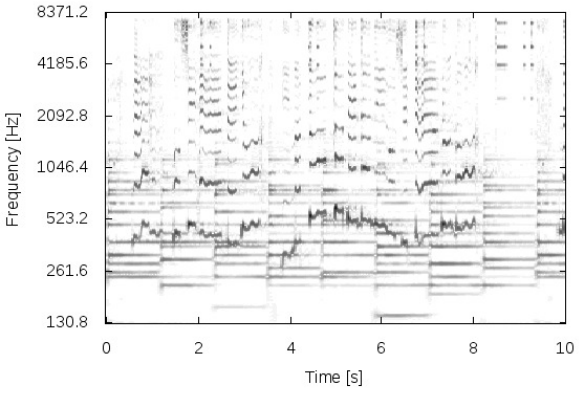
In order to verify the effectiveness of the singing voice enhancement based on two-stage HPSS, we conducted an experiment on singing voice enhancement using a professionally-created music audio signal. The music signals we used for experiments were excerpted from the RWC music database [19]. The data were resampled to 16 kHz and converted into monaural signals by adding both channels of stereo signals.

Fig. 4 shows a result of two-stage HPSS. The figures show the spectrograms of a input signal (Fig. 4 (a)) and the $\mathcal{V}$ component extracted by two-stage HPSS (Fig. 4 (b)). We can see in Fig. 4 (b) that most accompanying sounds are suppressed effectively by the method, and the singing voice is clearer in spectrogram (b) than that of the spectrogram (a). The figures show that the method effectively extract the singing voice from the mixed music audio signal.

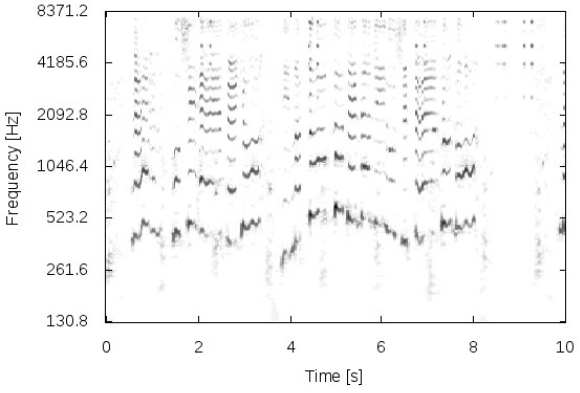
IV. Performance Evaluation of Two-stage HPSS

A. Large Scale Evaluation on Singing Voice Enhancement

1) Experimental Condition: To verify the effectiveness of the two-stage HPSS, we conducted experiments on the singing voice enhancement using music audio signals. The criteria for the performance evaluation of singing voice enhancement were the Normalized SDR (NSDR) and the Global NSDR (GNSDR). NSDR is defined as the



(a)



(b)

Fig. 4. (a) The constant Q spectrogram of an input signal (10 seconds from RWC-MDB-P-2001, No. 25 [19]), (b) the result of two-stage HPSS.

improvement of SDR as follows,

$$\mathrm{NSDR}[\hat{x}(t); s(t), x(t)] = \mathrm{SDR}[\hat{x}(t); x(t)] - \mathrm{SDR}[s(t); x(t)], \quad (16)$$

where $\hat{x}(t), s(t)$ and $x(t)$ denote the estimated signal, the input signal, and the target signal, respectively. SDR [6], [26] (signal to distortion ratio) is defined by

$$\mathrm{SDR}[x(t); y(t)] = 10 \log_{10} \frac{\langle x, y \rangle^2}{\|x\|^2 \|y\|^2 - \langle x, y \rangle^2}, \quad (17)$$

where $\langle x, y \rangle = \sum_t x(t) y(t)$ and $\|x\|^2 = \langle x, x \rangle$. GNSDR is defined as the averaged NSDR of all the pieces, weighted by w_i, the length of i-th piece [6],

$$\mathrm{GNSDR} = \frac{\sum_i w_i \mathrm{NSDR}[\hat{x}_i(t), s_i(t), x_i(t)]}{\sum_i w_i}. \quad (18)$$

Those criteria were also used in some previous works [6], [9], [14], [15].

For evaluation dataset, we exploited the MIR-1K database [9], [43], which is comprised of 1000 Chinese songs sung by amateur singers. The same dataset was used in some of previous works [9], [15]. The length T of each clip was around 4 to 13 [s]. All the data were monaural, and the sample rate of them were f_s =16 kHz. The vocal part and the accompaniment part were recorded separately, and we could mix them in any SNR (signal to noise ratio, i.e., voice to accompaniment ratio) for experiments. In this study, we mixed the singing voice and the accompaniment in $-10, -5, 0, 5$, and 10 dB for the experiment.

The parameters we used were as follows: The frame shift was half the length of the frame length, as described in the definition of STFT. The length of the frames were $l_1/f_\mathrm{s} = 8$ [ms] ($l_1 = 128$ points) and $l_2/f_\mathrm{s} = 512$ [ms] ($l_2 = 8192$ points). Both analyzing window $g_1(t)$ and reconstructing window $g_2(t)$ were sine window, $g_1(t) = g_2(t) = \sin \pi t/l, (0 \leq t < l)$. Under this condition, the following equation is satisfied for any signal $x(t)$, $x(t) \equiv \mathrm{STFT}_l^{-1}[\mathrm{STFT}_l[x(t)]]$. The parameters of HPSS were as follows. The size of the sliding block ($\approx$ number of iteration) I was 30, which is supposed to be sufficient empirically. The values of $\sigma_\mathrm{H}, \sigma_\mathrm{P}$ were 0.3. These values were identical to those which were described in the original HPSS paper [24]. After two-stage HPSS, we applied high pass filter to cut off any components lower than 110 Hz, because vocal components are less likely to appear in such lower frequencies.

2) Results and Discussion: Fig 5 (a) shows the distribution of NSDR for each input SNR condition. For most samples in most SNR conditions, NSDR were larger than 0 dB, i.e., SDR were improved. The method performed well, especially in -5 dB and 0 dB conditions. Fig 5 (b) compares GNSDR between the proposed method and some existing methods. The figure shows that the the proposed method considerably outperformed other methods in any SNR conditions within -5 to 5 dB.

The method was especially effective when the music signal sufficed the assumptions. That is, if the singing voice had sufficient fluctuation, and if it was accompanied by very stationary component and very percussive component, the performance tended to be better. However, the method was not typically effective in the following two cases. One was when the singing voice was not fluctuating sufficiently. When the singer singed flatly, or the singing voice was sustained for a long while with slight fluctuation, the singing voice did not satisfy the assumption that "singing voice is quasi-periodic but a little fluctuating," and two-stage HPSS did not separate the component into $\mathcal{V}$ components, but into $\mathcal{H}$ components. The other was when accompanying sounds were fluctuating. Typical instruments were violin, trumpet, etc., which fluctuate to some extent. Those sounds tended to be separated into $\mathcal{V}$ components, because they satisfy the assumption "quasi-periodic but a little fluctuating." To remove those sounds, we have to use other properties of sounds such as timbre, but this is outside of the scope of this paper.

B. Evaluation as a Preprocessing for Audio Melody Extraction

1) Experimental Condition: As mentioned in the introduction, singing voice enhancement is related to the pitch estimation of melody, and either can be used as the other's preprocessing. In this section, we conducted an experiment to show the effectiveness of the two-stage HPSS as a preprocessing for pitch estimation by comparing the two approaches, one of which is a simple pitch estimation technique, and the other is the tandem connection of two-stage HPSS and the pitch estimator. The pitch estimation algorithm we used in this experiment was a simple probabilistic method based on a spectral likelihood model and a pitch transition model [28].

The data we used for the experiment were excerpted from LabROSA dataset [44], which is referred to as sample data of the audio melody extraction task in MIREX. Nine of 13 pieces were chosen from the dataset under the condition that the melody is performed by a singing voice, and other 4 data were omitted because the melodies are performed by instruments. All the data were monaural, and the sample rate of them were 16 kHz.

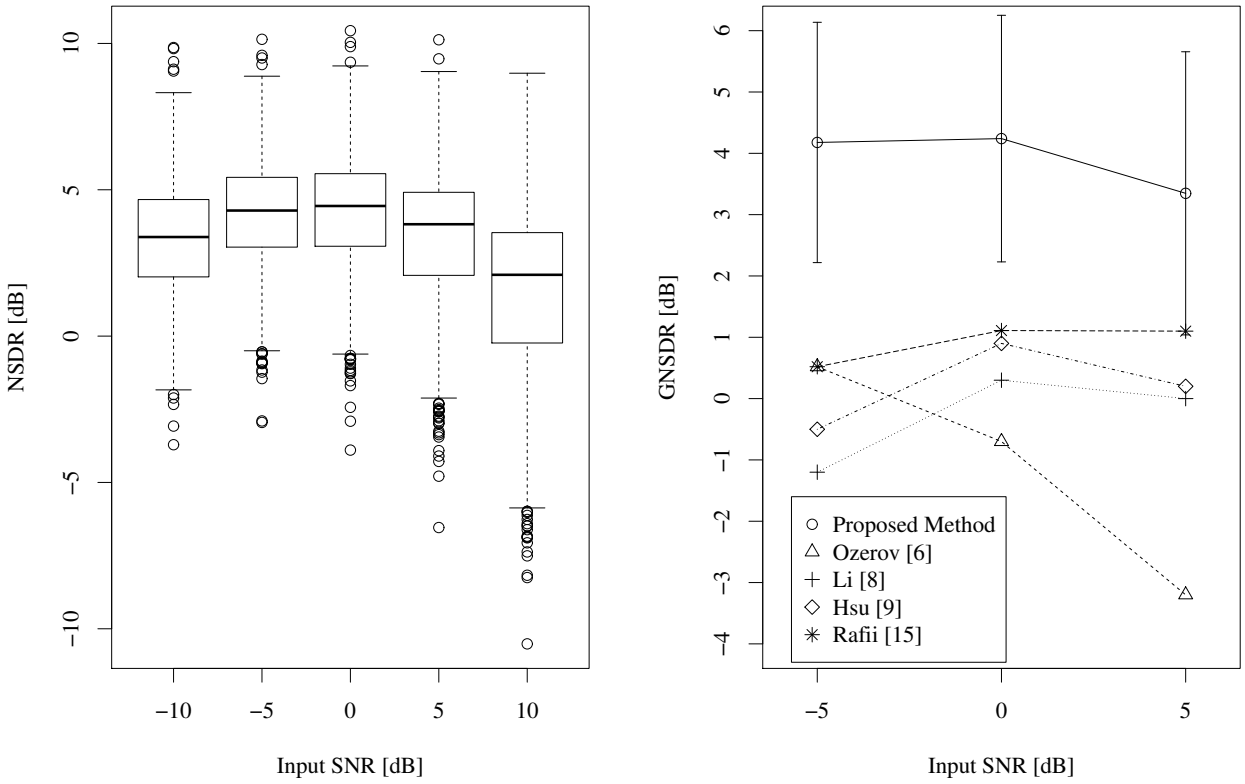


Fig. 5. (a) Boxplot of NSDR of the proposed method for 1000 songs in MIR-1K dataset. (b) GNSDR comparison with some existing singing voice enhancement techniques. The perpendicular bars on the plot of our method indicate the weighted standard deviation of NSDR. GNSDR of existing methods were cited from [9] and [15]. All the GNSDR scores are calculated using 1000 songs in MIR-1K dataset.

The criterion was Raw Pitch Accuracy (RPA), which is defined as the ratio of correctly estimated segments in melody-active segments. The correctness of the estimated pitch for each segment is judged by whether the difference between the estimation and the ground truth is within a quarter tone (half semitone) or not [45].

2) Result and Discussion: Fig. 6 (a) shows an example of the result of the pitch estimation, which is preceded by two-stage HPSS. The pitch sequence was well estimated in melody active segments. Fig. 6 (b) shows a result of the tandem connection of two-stage HPSS and the simple pitch estimator. In this condition, pitch was not correctly estimated sufficiently. These figures show that the result of the pitch estimation preceded by two-stage HPSS was more accurate than the result of the pitch estimation alone. Fig. 7 shows the accuracy ratios for each clip. Compared with the accuracies of the pitch estimation algorithm without the enhancement, it is observed that the singing voice enhancement basically improved the accuracy.

C. Large Scale Evaluation on Audio Melody Extraction in MIREX 2010

We submitted the two-stage HPSS followed by the pitch estimation [28] to the Audio Melody Extraction (AME) evaluations, which was held as a part of Music Information Retrieval EXchange (MIREX) [41] 2009 and 2010. In MIREX evaluations, several datasets and several criteria were used, but in this paper, we focus on the dataset and the criterion that are related to singing voice enhancement. The dataset was MIR-1K dataset, and the criterion concerned was RPA. The conditions were similar to those of the experiments in previous sections.

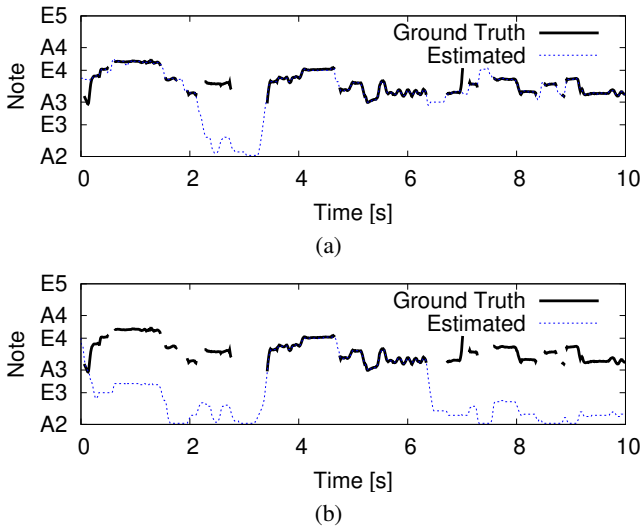


Fig. 6. Estimated melody line and ground truth of train06.wav (excerpted 10 [s]) in LabROSA dataset [44]. (a) The result of pitch estimation preceded by two-stage HPSS. (b) The result of pitch estimation without any preprocessing.

Fig. 8 shows the excerpted results from AME evaluation in MIREX 2010 [46]. It shows that the performance of the proposed method (TOOS1) is comparatively high, especially in a condition in which the volume level of singing voice is low to accompaniments (-5

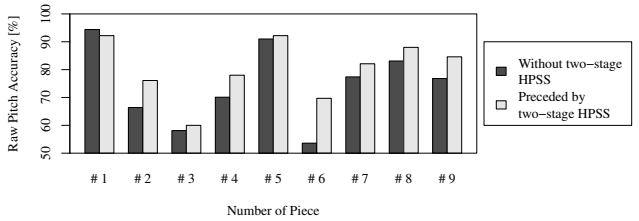


Fig. 7. Raw Pitch Accuracy of melody estimation for each piece in LabROSA dataset. [44]

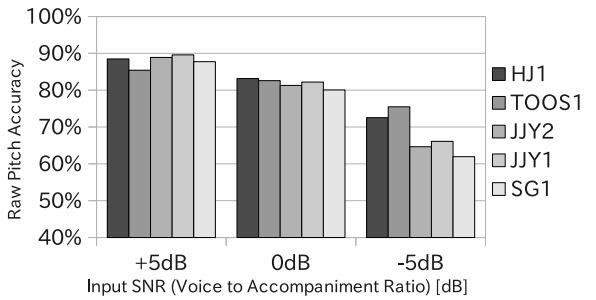


Fig. 8. Excerpted results from AME evaluation, MIREX 2010 [41]: Raw Pitch Accuracy (RPA) in $+5$dB, 0dB, -5dB SNR (voice to accompaniment ratio) conditions. "TOOS1" is our submission, and "HJ1" also uses a part of our singing voice enhancement technique [27], [28] as a preprocessing. This graph shows that the performance of proposed method is high, especially in low SNR (voice to accompaniment ratio) conditions, and comparable to those of other methods in high SNR (voice to accompaniment ratio) conditions. It indirectly shows the effectiveness of our singing voice enhancement method as a preprocessing for audio melody extraction.

dB). The figure also shows that the proposed method also performs comparably to other methods in a condition in which the volume level of melody is relatively high ($+5$ dB). The results, although indirectly, show the effectiveness of our singing voice enhancement method as a preprocessing for melody pitch estimation.

V. Conclusion

In this paper, we described two-stage HPSS, a singing voice enhancement method in monaural music signals. The method extracts the singing voice in music signals focusing on its fluctuation. Such natures of singing voice are exposed on two differently-resolved spectrograms, one is calculated using a short frame (around 10 [ms]), and the other is calculated using a long frame (around 500 [ms]). On the former spectrogram, both $\mathcal{H}$ (sustained, such as a piano) component and $\mathcal{V}$ (quasi-stationary, fluctuating, such as a singing voice) component appear as a "smooth-in-time" components, while $\mathcal{P}$ (transient, such as a percussive instrument) component appears as a "smooth-in-frequency" component. On the latter spectrogram, however, the behavior of $\mathcal{V}$ component is not similar to that of $\mathcal{H}$, but is similar to $\mathcal{P}$ component, i.e., it appears as a "smooth-in-frequency" component. Therefore, two-stage application of HPSS on differently-resolved two spectrograms roughly separates $\mathcal{V}$ component from $\mathcal{H}$ and $\mathcal{P}$ components.

We evaluated the method in the same framework as some previous works. According to experimental evaluations, the performance of the proposed method was considerably higher than those of some previous works and indicated around 4 dB GNSDR for -5, 0, and 5 dB mixtures. We also applied the method as a preprocessing for audio melody extraction and evaluated the performance. Although the subsequent pitch estimation was a very simple algorithm, the accuracy of the estimation was comparatively high compared with other pitch estimation methods in low SNR (voice to accompaniment ratio) conditions, due to the proposed singing voice enhancement method. This result also proves the effectiveness of the proposed method.

There are some works remained in the future, such as an investigation on the robustness of the method for some sound effects such as reverberations and nonlinear distortions that are sometimes observed in the real-world musics. Other future works will include investigations on applications of the method related to music information retrieval tasks, and on an effective utilization of the residual accompanying signals for an automatic karaoke generator.

Appendix A
Scale Invariance (Homogeneity) of J

In this section, the reason the exponential factor γ in (5) and (6) should be 0.5 is described. Let $(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}})$ be the solution of the problem (8), i.e.,

$$(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}}) = \underset{(\boldsymbol{H},\boldsymbol{P})}{\operatorname{argmin}} J(\boldsymbol{H}, \boldsymbol{P}; \boldsymbol{W}). \tag{19}$$

Let us suppose a case that λ times spectrogram $\lambda \boldsymbol{W}$ is given in (8). In that case, not only the input spectrogram, but also the separated spectrograms should be λ times the magnitude of those of the optimal spectrograms $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{P}}$, because it is not preferable that the result is dependent on the volume of the signal. Therefore, the optimal spectrograms $(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}})$ should also be the optimal to the following problem. That is, $(\hat{\boldsymbol{H}}', \hat{\boldsymbol{P}}')$ should be identical to $(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}})$,

$$(\lambda\hat{\boldsymbol{H}}', \lambda\hat{\boldsymbol{P}}') = \underset{\boldsymbol{H},\boldsymbol{P}}{\operatorname{argmin}} J(\boldsymbol{H}, \boldsymbol{P}|\lambda\boldsymbol{W}), \tag{20}$$

$$\Leftrightarrow \quad (\hat{\boldsymbol{H}}', \hat{\boldsymbol{P}}') = \underset{\boldsymbol{H},\boldsymbol{P}}{\operatorname{argmin}} J(\lambda\boldsymbol{H}, \lambda\boldsymbol{P}|\lambda\boldsymbol{W}). \tag{21}$$

Let us consider the dependencies on the scaling λ for each term of the new objective function $J(\lambda\boldsymbol{H}, \lambda\boldsymbol{P}|\lambda\boldsymbol{W})$. It is easily confirmed that

$$S_{\mathrm{H}}(\lambda\boldsymbol{H}) = \lambda^{2\gamma} S_{\mathrm{H}}(\boldsymbol{H}), \tag{22}$$

$$S_{\mathrm{P}}(\lambda\boldsymbol{P}) = \lambda^{2\gamma} S_{\mathrm{P}}(\boldsymbol{P}), \tag{23}$$

$$D_{\mathrm{KL}}(\lambda\boldsymbol{H} + \lambda\boldsymbol{P}|\lambda\boldsymbol{W}) = \lambda D_{\mathrm{KL}}(\boldsymbol{H} + \boldsymbol{P}|\boldsymbol{W}), \tag{24}$$

and the new objective function can be reduced as follows,

$$J(\lambda\boldsymbol{H}, \lambda\boldsymbol{P}|\lambda\boldsymbol{W}) = \lambda J(\boldsymbol{H}, \boldsymbol{P}|\boldsymbol{W}) + (\lambda^{2\gamma} - \lambda)\{S_{\mathrm{H}}(\boldsymbol{H}) + S_{\mathrm{P}}(\boldsymbol{P})\}. \tag{25}$$

The equation indicates that the optimal spectrograms $(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}})$ and $(\hat{\boldsymbol{H}}', \hat{\boldsymbol{P}}')$ that minimize $J(\boldsymbol{H}, \boldsymbol{P}|\boldsymbol{W})$ and $J(\lambda\boldsymbol{H}, \lambda\boldsymbol{P}|\lambda\boldsymbol{W})$ respectively are identical when $\lambda^{2\gamma} - \lambda = 0$, and otherwise, the optimal spectrograms are not generally identical. Hence, $2\gamma = 1$ should be satisfied for scale-invariance.

Acknowledgments

The authors thank the developers of MIR-1K dataset and LabROSA dataset and the organizers and the evaluators of MIREX 2010.

References

[1] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP Journal on Advances in Signal Processing*, 2007.

[2] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[3] W.-H. Tsai and H.-M. Wang, "Automatic identification of the sung language in popular music recordings," *Journal of New Music Research*, vol. 36, pp. 105–114, 2007.

[4] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, and T. Ogata, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. ISMIR*, 2005.

[5] M. Ryynämem, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Proceedings of IEEE International conference on Multimedia and Expo (ICME)*, 2008, pp. 1417–1420.

[6] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. WASPAA*, 2005, pp. 90–93.

[7] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, 2007.

[8] Y. Li and D.-L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, 2007.

[9] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA*, 2003, pp. 177–180.

[11] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. ISMIR*, 2005, pp. 337–344.

[12] T. Virtanenm, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc SAPA*, 2008, pp. 17–20.

[13] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. ISMIR*, 2006.

[14] 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. ICASSP*, 2012, pp. 57–60.

[15] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proc. ICASSP*, 2011, pp. 221–224.

[16] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. International Symposium Frontiers of Research Speech and Music (FRSM)*, 2007.

[17] M. Lagrange, L. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predomnant melodic source separation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 2, pp. 278–290, 2008.

[18] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram," in *Proc. ICASSP*, 2012, pp. 465–468.

[19] M. Goto, "Development of the RWC music database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, 2004, pp. I–553–556.

[20] Y. Horii, "Acoustic analysis of vocal vibrato: A theoretical interpretation of data," *J. Voice*, vol. 3, pp. 36–43, 1989.

[21] H. Mori, W. Odagiri, and H. Kasuya, "F0 dynamics in singing: Evidence from the data of a baritone singer," *IEICE Trans. Inf. Syst.*, vol. E87-D, pp. 1086–1092, 2004.

[22] J. Sundberg, *The Science of the singing voice*. Nothern Illinois University Press, 1987.

[23] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Advances in Music Information Retrieval*, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wieczorkowska, Eds. Springer, Feb. 2010, pp. 213–236.

[24] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. ISMIR*, 2008, pp. 139–144.

[25] 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 *Proc. EUSIPCO*, 2008.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," vol. 14, no. 4, pp. 1462–1469, 2006.

[27] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody extraction in music audio signals by melodic component enhancement and pitch tracking," in *MIREX2009*, 2009.

[28] ——, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Proc. ICASSP*, 2010, pp. 425–428.

[29] C.-L. Hsu, D. L. Wang, J.-S. 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, 2012.

[30] 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.

[31] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.

[32] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search," in *Proc. ICASSP*, 2006, pp. 253–256.

[33] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," in *Proc. ISMIR*, 2007, pp. 373–374.

[34] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc ICASSP*, 2008, pp. 169–172.

[35] J.-L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, 2010.

[36] 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, 2012.

[37] C.-L. Hsu, D. L. Wang, and J.-S. R. Jang, "A trend estimation algorithm for singing pitch detection in musical recordings," in *in Proc. ICASSP*, 2011, pp. 393–396.

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

[39] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 804–815, 2003.

[40] ——, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 244–266, 2008.

[41] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustic Science & Technology*, vol. 29, pp. 247–255, 2008.

[42] [Online]. Available: http://www.music-ir.org/mirex/wiki/2009:Audio_Melody_Extraction_Results

[43] [Online]. Available: http://unvoicedsoundseparation.googlepages.com/mir-1k

[44] [Online]. Available: http://labrosa.ee.columbia.edu/projects/melody/

[45] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, 2007.

[46] [Online]. Available: http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results

Hideyuki Tachibana (S'10) received the B.E. degree in mathematical engineering and information physics and the M.E. degree in information physics and computing from the University of Tokyo, Tokyo, Japan, in 2008 and 2010, respectively. He is currently pursuing a Ph.D. degree at the Department of Information Physics and Computing, Graduate School of Information Science and Technology, the University of Tokyo. His research interests include music signal processing. He is a student member of IEEE, ACM, Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and the Institute of Electronics, Information and Communication Engineers (IEICE).

Nobutaka Ono (M '02) received the B.E., M.S., and Ph.D degrees in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He joined the Graduate School of Information Science and Technology, the University of Tokyo, Japan, in 2001 as a Research Associate and became a Lecturer in 2005. He moved to the National Institute of Informatics, Japan, as an Associate Professor in 2011. His research interests include acoustic signal processing, specifically, microphone array processing, source localization and separation, music signal processing, audio coding and watermarking, and optimization algorithms for them. He is the author or co-author of more than 100 articles in international journal papers and peer-reviewed conference proceedings. Dr. Ono has been an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing since 2012 and also an Associate Editor of Acoustic Society and Technology since 2012. He was a Tutorial speaker at ISMIR 2010 and organized a special session in EUSIPCO 2013. He is a chair of SiSEC (Signal Separation Evaluation Campaign) evaluation committee in 2013. He is a Senior member of the IEEE Signal Processing Society, and a member of the Institute of Electronics, Information and Communications Engineers (IEICE), the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Institute of Electrical Engineers of Japan (IEEJ), and the Society of Instrument and Control Engineers (SICE). He received the Sato Paper Award and the Awaya Award from ASJ in 2000 and 2007, respectively, received the Igarashi Award at the Sensor Symposium on Sensors, Micromachines, and Applied Systems from IEEJ in 2004, and received the best paper award in IEEE International Symposium on Industrial Electronics (ISIE) in 2008.

Shigeki Sagayama (M' 82) received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics. He joined Nippon Telegraph and Telephone Public Corporation (currently, NTT) in 1974 and started his career in speech analysis, synthesis, and recognition at NTT Labs in Musashino, Japan. From 1990, he was Head of the Speech Processing Department, ATR Interpreting Telephony Laboratories, Kyoto, Japan where he was in charge of an automatic speech translation project. In 1993, he was responsible for speech recognition, synthesis, and dialog systems at NTT Human Interface Laboratories, Yokosuka, Japan. In 1998, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa. In 2000, he was appointed Professor at the Graduate School of Information Science and Technology (formerly, Graduate School of Engineering), the University of Tokyo. On his retirement from the University of Tokyo in 2013, he became a Project Professor at the National Institute of Informatics (NII). His major research interests include the processing and recognition of speech, music, acoustic signals, handwriting, and images. He was the leader of anthropomorphic spoken dialog agent project (Galatea Project) from 2000 to 2003. Prof. Sagayama received the National Invention Award from the Institute of Invention of Japan in 1991, the Chief Official's Award for Research Achievement from the Science and Technology Agency of Japan in 1996, and other academic awards including Paper Awards from the Institute of Electronics, Information and Communications Engineers, Japan (IEICEJ) in 1996 and from the Information Processing Society of Japan (IPSJ) in 1995. He is a member of the Acoustical Society of Japan, IEICEJ, and IPSJ.