PHASE-AWARE HARMONIC/PERCUSSIVE SOURCE SEPARATION VIA CONVEX OPTIMIZATION

Yoshiki Masuyama, Kohei Yatabe and Yasuhiro Oikawa

Department of Intermedia Art and Science, Waseda University, Tokyo, Japan

ABSTRACT

Decomposition of an audio mixture into harmonic and percussive components, namely harmonic/percussive source separation (HPSS), is a useful pre-processing tool for many audio applications. Popular approaches to HPSS exploit the distinctive source-specific structures of power spectrograms. However, such approaches consider only power spectrograms, and the phase remains intact for resynthesizing the separated signals. In this paper, we propose a phase-aware HPSS method based on the structure of the phase of harmonic components. It is formulated as a convex optimization problem in the time domain, which enables the simultaneous treatment of both amplitude and phase. The numerical experiment validates the effectiveness of the proposed method.

Index Terms— Music decomposition, sinusoidal model, instantaneous frequency, temporal smoothness, primal-dual splitting.

1. INTRODUCTION

Audio source separation, decomposing an audio mixture into each source, is one of the fundamental tools for audio signal processing. In particular, harmonic/percussive source separation (HPSS), which decomposes an audio mixture into harmonic components (e.g., guitar and piano) and percussive components (e.g., drums), has gained much attention as a pre-processing tool for many music-information retrieval (MIR) tasks including chord estimation [1] and tempo estimation [2]. For instance, extracted percussive components are useful cues for tempo estimation while the harmonic components are crucial for chord estimation. HPSS is also helpful for audio remixing [3] and time-scale modification [4].

One of the main approaches to HPSS is to take advantage of the *anisotropic smoothness* of power spectrograms (i.e., power spectrograms of harmonic components are continuous in the time direction, and those of percussive components are continuous in the frequency direction as shown in Fig. 1) [5–7]. Based on the anisotropic smoothness, HPSS was formulated as an optimization problem of minimizing the ℓ_2 norm of the gradient of power spectrograms in [5]. Considering the same assumption, the method presented in [8] applies the time-/frequency-directional median filtering (MF) to the power spectrogram of the audio mixture, which was further developed into the kernel additive model (KAM) [9–11]. Although these methods have been successfully applied to HPSS, they have a limitation because the degraded phase from the audio mixture is still utilized for resynthesizing the separated time domain signals.

Recent literature has shown the importance of phase in audio source separation [12, 13] and audio denoising [14–17]. These studies utilize a model of phase for harmonic components, called *sinusoidal model*. This model claims that the phase evolution of harmonic components can be predicted from their instantaneous frequencies. More recently, this model was also applied to HPSS [18, 19]. In [18], the real-valued time-frequency mask for extracting harmonic components was constructed based on the sinusoidal model. However, the phase is not modified through time-frequency masking, and thus the degraded phase from the audio mixture is utilized for resynthesizing back to the time-domain. On the other hand, [19] utilizes the sinusoidal model for recovering the phase after applying a time-frequency mask obtained by a deep neural network. However, simultaneous modification of both amplitude and phase has not been presented for HPSS.

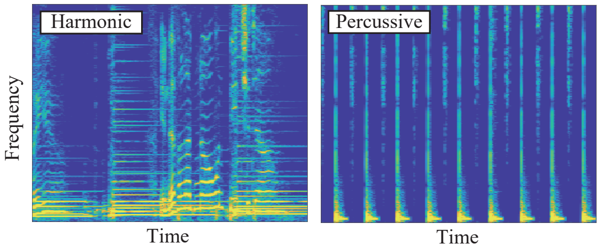

Fig. 1. Example of spectrograms of the harmonic and percussive components. The harmonic component is continuous in the time direction (left), and that of percussive components is continuous in the frequency direction (right). The time-frame-wise sparse structure is apparent in that of the percussive component.

In this paper, we propose a phase-aware HPSS method, which treats both amplitude and phase simultaneously, through convex optimization in the time domain. For harmonic components, the proposed method assumes the time-directional smoothness of the complex-valued spectrogram with the recently proposed phase modification [15]. This assumption can be interpreted as the unification of the two conventional HPSS methodologies: the anisotropic smoothness and sinusoidal model. For the percussive components, the time-frame-wise sparsity is assumed on their complex-valued spectrograms, which does not require any assumptions for its phase structure. The effectiveness of the proposed method was confirmed by the signal-to-distortion ratio (SDR).

2. PREVIOUS WORKS

In this section, we briefly revisit two approaches to HPSS, the anisotropic smoothness and sinusoidal model, since the proposed method combines these two approaches as described in Section 3.

2.1. HPSS based on anisotropic smoothness

One main approach to HPSS is the anisotropic smoothness of power spectrograms. It assumes that a power spectrogram of harmonic components $\mathbf{H} \in \mathbb{R}_+^{K\times T}$ and that of percussive components $\mathbf{P} \in \mathbb{R}_+^{K\times T}$ have the following relations:

$$H_{\omega,\tau} \approx H_{\omega,\tau\pm1}, \tag{1}$$

$$P_{\omega,\tau} \approx P_{\omega\pm1,\tau}, \tag{2}$$

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

where $\omega = 1, \ldots, K$ and $\tau = 1, \ldots, T$ are frequency and time indices, respectively. Eq. (1) indicates that the power spectrogram of the harmonic components varies slowly, while Eq. (2) claims that of the percussive components is smooth in the frequency direction. Based on these assumptions, the following optimization-based HPSS method was proposed in [5]:

$$\begin{array}{cl} \min\limits_{\mathbf{H},\mathbf{P}} & \frac{1}{2\sigma_{\mathrm{h}}^2} \|D_\tau(\mathbf{H})\|_{\mathrm{Fro}}^2 + \frac{1}{2\sigma_{\mathrm{p}}^2} \|D_\omega(\mathbf{P})\|_{\mathrm{Fro}}^2 \\ \text{s.t.} & H_{\omega,\tau} + P_{\omega,\tau} = |X_{\omega,\tau}|^{2\gamma}, \quad H_{\omega,\tau} \geq 0, \quad P_{\omega,\tau} \geq 0 \end{array}, \quad (3)$$

where D_τ and D_ω are the time and frequency-directional differences (i.e., discrete approximation of directional derivatives), σ_{h} and σ_{p} are parameters to adjust smoothness of harmonic and percussive spectrograms, and $\mathbf{X} \in \mathbb{C}^{K \times T}$ is the complex-valued spectrogram of the audio mixture to be separated. γ is a hyperparameter for range compression $(0 < \gamma \leq 1)$, and it will be set to 1 in the rest of this paper for simplicity. By minimizing the energy of directional derivatives of the spectrograms, this model attempts to find spectrograms which are smooth in each direction. In the experimental section, this optimization-based method will be referred to as *Ono's*.

While this optimization-based method is simple and effective, the assumption of additivity of power spectrograms holds only approximately (it can be justified only by some statistical sense [20]). Furthermore, it considers only power spectrograms, and thus the phase information is ignored. Although some extensions based on the anisotropic smoothness of power spectrograms have been proposed [8–11], the degraded phase of the audio mixture is utilized for resynthesizing the separated time domain signals, which often causes audible artifacts as mentioned in [12].

2.2. HPSS based on sinusoidal model

Another recent approach to HPSS is based on the sinusoidal model. Let the short-time Fourier transform (STFT) of a signal $\mathbf{x} \in \mathbb{R}^L$ be

$$\mathscr{F}(\mathbf{x})_{\omega,\tau} = \sum_{l=0}^{L-1} x_{l+a\tau}\, g_l\, \mathrm{e}^{-2\pi j \omega b l / L}, \quad (4)$$

where $\boldsymbol{g} \in \mathbb{R}^L$ is a window, $j = \sqrt{-1}$, a and b are the time and frequency shifting steps, and index overflow is treated by zero-padding. Considering a sinusoid given by

$$s_l = A\, \mathrm{e}^{2\pi j f l / L + \phi}, \quad (5)$$

where $A \in \mathbb{R}_+$, $f \in [0, L/2)$, and $\phi \in [0, 2\pi)$ are the amplitude, frequency, and initial phase, respectively. Its phase spectrogram $\boldsymbol{\phi}$ (with appropriate unwrapping) has the following relation:

$$\phi_{\omega,\tau} = \phi_{\omega,\tau-1} + 2\pi a v_{\omega,\tau-1}, \quad (6)$$

where $v_{\omega,\tau}$ is the instantaneous frequency at each time-frequency bin. This sinusoidal model has been studied in phase vocoders [21], and applied to audio signal processing tasks recently [13, 16, 17, 22].

More recently, the sinusoidal model was also applied to HPSS [18, 19]. The phase-based masking (PM), which constructs the time-frequency mask based on the relation among phases in successive time frames, was presented in [18]. Although this method considers phase information, the phase of the audio mixture is still utilized for resynthesizing the separated time domain signals. In contrast, the method presented in [19] utilizes the sinusoidal model for modifying phase. Specifically, the time-frequency mask is estimated by a deep neural network, and phases of separated signals are estimated by the specific algorithm based on the sinusoidal model [13]. It significantly depends on the time-frequency mask estimation, and phase information is utilized just in the post-processing.

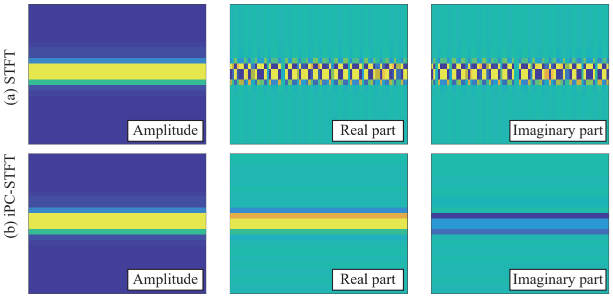

Fig. 2. Illustration of a complex-valued spectrogram calculated by (a) the usual STFT and (b) iPC-STFT.

3. PROPOSED HPSS METHOD

In this section, we propose a phase-aware HPSS method through convex optimization where the phase-aware smoothness in the time direction is assumed for harmonic components as a unification of the aforementioned approaches: anisotropic smoothness and sinusoidal model. On the other hand, the time-frame-wise sparsity is considered for the complex-valued spectrograms of percussive components as a phase insensitive prior. The proposed method directly separates the time-domain signal, which enables the simultaneous modification of amplitude and phase.

At first, we review the phase-aware smoothness of harmonic signals introduced in the previous study [15]. Then, based upon that, we formulate the proposed method and discuss the relation to the conventional HPSS approaches reviewed in the previous section.

3.1. Time-directional smoothness of instantaneous phase corrected complex-valued spectrogram

As shown in Eq. (6), the phase of a sinusoid has the relation among successive time frames. It indicates the phase in the next frame is predictable from the phase and instantaneous frequency in the current frame. Based on the relation of phase, a model of a complex-valued spectrogram of harmonic components was recently studied for audio signal processing [14, 15]. Since the amplitude of the sinusoid is constant, its complex-valued spectrogram satisfies the following relationship:

$$\mathscr{F}(\mathbf{x})_{\omega,\tau} = \mathscr{F}(\mathbf{x})_{\omega,\tau-1}\, \mathrm{e}^{2\pi j f a / L}. \quad (7)$$

Therefore, the complex-valued spectrogram of a sinusoid takes the same value in each sub-band if its phase evolution is eliminated.

In order to eliminate such a phase evolution, *instantaneous phase corrected STFT* (iPC-STFT) was proposed in [15]:

$$\mathscr{F}_{\mathrm{iPC}}(\mathbf{x}) = \mathbf{E} \odot \mathscr{F}(\mathbf{x}), \quad (8)$$

where $\mathbf{E}$ is the instantaneous phase correction matrix defined by

$$E_{\omega,\tau} = \prod_{\eta=0}^{\tau-1} \mathrm{e}^{-2\pi j v_{\omega,\eta} a / L}, \quad (9)$$

with $E_{\omega,0} = 1$ for all ω, and $\odot$ is the Hadamard product. This matrix eliminates the phase evolution of a sinusoid as in Eq. (6). While real and imaginary parts of the complex-valued spectrogram calculated by the usual STFT vary owing to the phase evolution, those of iPC-STFT are constant in each sub-band thanks to the instantaneous phase correction as illustrated in Fig. 2. Namely, the complex-valued spectrogram of the sinusoid is smooth in each sub-band as

$$\mathscr{F}_{\mathrm{iPC}}(\mathbf{s})_{\omega,\tau} = \mathscr{F}_{\mathrm{iPC}}(\mathbf{s})_{\omega,\tau-1}. \quad (10)$$

Although this equation considers a single sinusoid, the time-directional smoothness of the complex-valued spectrogram calculated by iPC-STFT is also reasonable for harmonic signals consisting of a sum of sinusoids. In [15], based on this characteristic of iPC-STFT, a simple prior for harmonic signals was proposed, which penalizes the time-derivative of the complex-valued spectrogram calculated by iPC-STFT for enhancing harmonic components. For more details about iPC-STFT, we refer readers to [15].

Note that the instantaneous frequency $v_{\omega,\tau}$ is not known and must be estimated in advance. It can be estimated by the direct time-differential of phase as

$$v_{\omega,\tau} = b\omega - \mathrm{Im}\left[\frac{\tilde{\mathscr{F}}(\mathbf{x})_{\omega,\tau}}{\mathscr{F}(\mathbf{x})_{\omega,\tau}}\right], \tag{11}$$

where $\tilde{\mathscr{F}}$ is the usual STFT whose window is the time-derivative of the original window $\boldsymbol{g}$, and $\mathrm{Im}[z]$ is the imaginary part of z [23].

3.2. Proposed optimization-based HPSS method

As described in the previous subsection, complex-valued spectrograms of harmonic components calculated by iPC-STFT have the distinctive structure. Utilizing iPC-STFT, we propose a phase-aware HPSS method through the following convex optimization:

$$\begin{array}{ll} \displaystyle\min_{\mathbf{x}_\mathrm{h},\mathbf{x}_\mathrm{p}} & \displaystyle\frac{1}{2}\left\|\mathbf{W}\odot D_\tau(\mathbf{X}_\mathrm{h})\right\|_\mathrm{Fro}^2 + \lambda\left\|\mathbf{X}_\mathrm{p}\right\|_{2,1} \\ \text{s.t.} & \mathbf{x} = \mathbf{x}_\mathrm{h} + \mathbf{x}_\mathrm{p}, \;\; \mathbf{X}_\mathrm{h} = \mathscr{F}_\mathrm{iPC}(\mathbf{x}_\mathrm{h}), \;\; \mathbf{X}_\mathrm{p} = \mathscr{F}(\mathbf{x}_\mathrm{p}) \end{array}, \tag{12}$$

where $\mathbf{x}_\mathrm{h}$, $\mathbf{x}_\mathrm{p}$, and $\mathbf{x}$ are the time-domain signals of harmonic components, percussive components, and the audio mixture, respectively, $\lambda > 0$ is the hyperparameter which adjusts the amount of harmonic and percussive components, and $\mathbf{W} \in \mathbb{R}_+^{K\times T}$ is a weight constructed in advance. Note that the the instantaneous phase correction matrix $\mathbf{E}$ in iPC-STFT is calculated from the the audio mixture, and thus iPC-STFT is treated as the fixed linear operator in the proposed formulation.

The first term induces the time-directional smoothness of harmonic components as an extension of Eq. (10). Since power spectrogram of harmonic components is smooth in the time-direction as in Eq. (1), the time-directional smoothness of iPC-STFT spectrogram of general harmonic components is also reasonable. The weight $\mathbf{W}$, which adjusts the time-directional smoothness around each time-frequency bin, is given by

$$W_{\omega,\tau} = \kappa \,/\, \max(\,\kappa,\; |\tilde{X}_\mathrm{h}|_{\omega,\tau}), \tag{13}$$

where $\kappa > 0$ is a small number for adjusting the weight, and $|\tilde{X}_\mathrm{h}|$ is the normalized amplitude of pre-estimated harmonic components which can be obtained by any existing method as [5, 8, 10]. This weight takes a small value when the amplitude of harmonic components are large, and thus harmonic components with large amplitude are not penalized so much. Note that this additional weight was not introduced in the previous study [15], and thus it is one of the contributions of this paper.

The second term is the $\ell_{2,1}$-norm which induces group sparsity. Here, the time-frame-wise sparsity is promoted by defining it as

$$\|\mathbf{X}\|_{2,1} = \sum_{\tau=1}^{T}\left(\sum_{\omega=1}^{K}|X_{\omega,\tau}|^2\right)^{\frac{1}{2}}. \tag{14}$$

This penalty function concentrates the energy into a few time frames and enhances impulsive components. Such time-frame-wise sparsity of percussive components can be seen in Fig. 1.

3.3. Relation to the conventional methods

The proposed method is related to the method based on anisotropic smoothness [5] described in Section 2.1. While the first term in Eq. (3) only considers the power spectrogram of harmonic components, that of Eq. (12) treats both amplitude and phase through the operation in the complex domain (note that the squared amplitude of $\mathbf{X}_\mathrm{h}$ corresponds to $\mathbf{H}$). Thus, the proposed method given by Eq. (12) can be interpreted as a phase-aware extension of Eq. (3). For the percussive components, the proposed method assumes the time-frame-wise sparsity, while Eq. (3) considered smoothness in the frequency direction. Considering the frequency-directional smoothness for the complex-valued spectrogram of percussive components may not be easy. Thus, the proposed method utilizes $\ell_{2,1}$-norm, which is insensitive to phase, instead of penalizing the frequency-derivative of the complex-valued spectrogram.

As in Eq. (12), the proposed method directly treats the time-domain signals, and its constraint claims that the separated components satisfy the perfect reconstruction property in the time-domain as in a recent audio source separation method [24]. Some of the conventional HPSS methods (e.g., anisotropic smoothness based methods [5, 7] and non-negative matrix factorization based methods [25, 26]) assume additivity of power spectrograms, but it requires some statistical assumptions as discussed in [20]. In contrast, the constraint in the proposed method (additivity in the time-domain) is always justified. To the best of our knowledge, this is the first study applying the perfect reconstruction constraint to the separated time domain signals in HPSS.

3.4. Primal-dual splitting algorithm for proposed HPSS method

In order to solve the convex optimization problem given by Eq. (12), in this paper, a primal-dual splitting algorithm [27] is adopted because it can handle some priors tied with linear operators with a constraint. A primal-dual splitting (PDS) method [27] is one of the convex optimization algorithms for solving the following problem:[1]

$$\min_{\mathbf{x}}\;\; \Theta(\mathbf{x}) + \Upsilon_1\big(\mathscr{L}_1(\mathbf{x})\big) + \Upsilon_2\big(\mathscr{L}_2(\mathbf{x})\big), \tag{15}$$

where Θ and Υ_m are proper lower-semicontinuous convex functions, and $\mathscr{L}_m$ is a linear operator ($m \in \{1, 2\}$). A PDS algorithm solves this problem by iterating the following procedure:

$$\tilde{\mathbf{x}} = \mathrm{prox}_{\mu_1\Theta}\big(\mathbf{x} - \mu_1\big(\mathscr{L}_1^*(\mathbf{y}_1^{[n]}) + \mathscr{L}_2^*(\mathbf{y}_2^{[n]})\big)\big), \tag{16}$$

$$\mathbf{z}_m = \mathbf{y}_m^{[n]} + \mathscr{L}_m^*(2\tilde{\mathbf{x}} - \mathbf{x}^{[n]}) \quad (\forall m), \tag{17}$$

$$\tilde{\mathbf{y}}_m = \mathbf{z}_m - \mu_2\,\mathrm{prox}_{1/\mu_2\Upsilon_m}(\mathbf{z}_m/\mu_2) \quad (\forall m), \tag{18}$$

$$(\mathbf{x}^{[n+1]}, \mathbf{y}_m^{[n+1]}) = \alpha(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}_m) + (1-\alpha)(\mathbf{x}^{[n]}, \mathbf{y}_m^{[n]}) \quad (\forall m), \tag{19}$$

where $\mathscr{L}_m^*$ is the adjoint operator of $\mathscr{L}_m$, n is the iteration index, $\mu_1 > 0$, $\mu_2 > 0$, and $0 < \alpha < 2$. The important feature of this procedure is that the minimization of each function is handled separately through the proximity operator [28]:

$$\mathrm{prox}_{\rho\Psi(\mathbf{Y})} = \arg\min_{\mathbf{X}}\; \Psi(\mathbf{X}) + \frac{1}{2\rho}\left\|\mathbf{Y}-\mathbf{X}\right\|_\mathrm{Fro}^2. \tag{20}$$

To apply this PDS algorithm to the proposed method in Eq. (12), it should be reformulated to the following equivalent problem:

$$\min_{\mathbf{x}_\mathrm{h},\mathbf{x}_\mathrm{p}}\;\; \iota_{\mathbf{x}}(\mathbf{x}_\mathrm{h},\mathbf{x}_\mathrm{p}) + \frac{1}{2}\left\|\mathscr{L}_\mathrm{h}(\mathbf{x}_\mathrm{h})\right\|_\mathrm{Fro}^2 + \lambda\left\|\mathscr{F}(\mathbf{x}_\mathrm{p})\right\|_{2,1}, \tag{21}$$

[1] This primal-dual splitting algorithm is a simplified version chosen for easier explanation of the proposed algorithm. We refer readers to [27] for the general form which can handle wider range of problems.

Algorithm 1 Proposed HPSS algorithm solving Eq. (21)

```
**Input**: $\mathbf{x}, \mathbf{x}_{\mathrm{h}}^{[0]}, \mathbf{x}_{\mathrm{p}}^{[0]}, \mathbf{Y}_{\mathrm{h}}^{[0]}, \mathbf{Y}_{\mathrm{p}}^{[0]}, \lambda, \mu_1, \mu_2, \alpha$
**Output**: $\mathbf{x}_{\mathrm{h}}^{[n+1]}, \mathbf{x}_{\mathrm{p}}^{[n+1]}$
**for** $n = 1, 2, \ldots$ **do**
    $(\tilde{\mathbf{x}}_{\mathrm{h}}, \tilde{\mathbf{x}}_{\mathrm{p}}) = P_{\mathbf{x}}\big(\mathbf{x}_{\mathrm{h}}^{[n]} - \mu_1 \mathscr{L}_{\mathrm{h}}^{*}(\mathbf{Y}_{\mathrm{h}}^{[n]}),\ \mathbf{x}_{\mathrm{p}}^{[n]} - \mu_1 \mathscr{F}^{*}(\mathbf{Y}_{\mathrm{p}}^{[n]})\big)$
    $\mathbf{z}_{\mathrm{h}} = \mathbf{y}_{\mathrm{h}}^{[n]} + \mathscr{L}_{\mathrm{h}}(2\tilde{\mathbf{x}}_{\mathrm{h}} - \mathbf{x}_{\mathrm{h}}^{[n]})$
    $\mathbf{z}_{\mathrm{p}} = \mathbf{y}_{\mathrm{p}}^{[n]} + \mathscr{F}(2\tilde{\mathbf{x}}_{\mathrm{p}} - \mathbf{x}_{\mathrm{p}}^{[n]})$
    $\tilde{\mathbf{y}}_{\mathrm{h}} = \mathbf{z}_{\mathrm{h}} - \mu_2\, \mathrm{prox}_{(1/\mu_2)\|\cdot\|_{\mathrm{Fro}}^2}(\mathbf{z}_{\mathrm{h}}/\mu_2)$
    $\tilde{\mathbf{y}}_{\mathrm{p}} = \mathbf{z}_{\mathrm{p}} - \lambda\mu_2\, \mathrm{prox}_{(1/\lambda\mu_2)\|\cdot\|_{2,1}}(\mathbf{z}_{\mathrm{p}}/\lambda\mu_2)$
    $(\mathbf{x}_{\mathrm{h,p}}^{[n+1]}, \mathbf{y}_{\mathrm{h,p}}^{[n+1]}) = \alpha(\tilde{\mathbf{x}}_{\mathrm{h,p}}, \tilde{\mathbf{y}}_{\mathrm{h,p}}) + (1-\alpha)(\mathbf{x}_{\mathrm{h,p}}^{[n]}, \mathbf{y}_{\mathrm{h,p}}^{[n]})$
**end for**
```

where $\mathscr{L}_{\mathrm{h}}(\cdot) = \mathbf{W} \odot D_\tau\big(\mathscr{F}_{\mathrm{iPC}}(\cdot)\big)$, and $\iota_{\mathbf{x}}(\cdot,\cdot)$ is the indicator function of the perfect reconstruction constraint given by

$$\iota_{\mathbf{x}}(\mathbf{x}_{\mathrm{h}}, \mathbf{x}_{\mathrm{p}}) = \begin{cases} 0 & (\mathbf{x} = \mathbf{x}_{\mathrm{h}} + \mathbf{x}_{\mathrm{p}}) \\ \infty & (\text{otherwise}) \end{cases}. \tag{22}$$

Applying the PDS algorithm to the reformulated problem in Eq. (21) yields Algorithm 1, where the choice of μ_1 and μ_2 can be automated as in [27], and the proximity operators involved in the algorithm can be analytically calculated as follows [28,29]:

$$P_{\mathbf{x}}(\mathbf{x}_{\mathrm{h}}, \mathbf{x}_{\mathrm{p}}) = (\mathbf{x}_{\mathrm{h}}, \mathbf{x}_{\mathrm{p}}) + (\mathbf{x} - \mathbf{x}_{\mathrm{h}} - \mathbf{x}_{\mathrm{p}})/2, \tag{23}$$

$$\mathrm{prox}_{\rho\|\cdot\|_{\mathrm{Fro}}^2}(\mathbf{X}) = \mathbf{X}/(1+\rho), \tag{24}$$

$$(\mathrm{prox}_{\rho\|\cdot\|_{2,1}}(\mathbf{X}))_\tau = (1 - \rho/\|\mathbf{X}_\tau\|_2)_+\ \mathbf{X}_\tau, \tag{25}$$

where $\mathbf{X}_\tau$ is the K-dimensional vector at the τth time frame. We stress that this algorithm does not require the inverse of linear operators, $\mathscr{L}_{\mathrm{h}}$ and $\mathscr{F}$, but only require applying them and their adjoint, and thus we can avoid the huge computation for calculating their inverse.

4. NUMERICAL EXPERIMENT

The proposed method was applied to separations of 10 audio tracks[2] which was presented in the previous study [18]. The sampling rate was 44100 Hz, and STFT was calculated by the canonical tight window of the Hann window of 4096 samples with 1024 sample shifting. The proposed method was compared with four conventional methods: Ono's [5], MF [8], KAM [10], and PM [18]. In each method, the hyperparameters were set to suggested values in each original paper. Separation performance was evaluated by the average of BSS Eval measures: SDR, signal-to-interference ratio (SIR), signal-to-artifacts ratio (SAR) [30].

For the proposed method, the amplitude spectrogram of harmonic components should be estimated in advance for calculating the weight $\mathbf{W}$. Here we utilized MF as a simple and fast HPSS method for prior estimation of the harmonic components. Although MF does not treat phase, it is modified through the proposed method. For calculating iPC-STFT, the instantaneous frequency of the harmonic components is also required. Since its oracle information is not available, we calculated it from the audio mixture (Prop-mix). In order to evaluate the potential of the proposed method, iPC-STFT with the instantaneous frequency calculated from the oracle harmonic components was also compared (Prop-ora). In both cases, the instantaneous frequency was calculated by Eq. (11). The hyperparameters, λ and κ, were experimentally determined to be 0.5 and 0.001. Algorithm 1 was iterated 100 times with $\mu_1 = 1$, $\mu_2 = 0.25$, and $\alpha = 0.5$. While an arbitrary choice is allowable, MF was utilized for estimating the initial value.

Table 1. Mean scores over 10 audio tracks. The average (Ave.) of the harmonic (Har.) and percussive (Per.) components are also presented. Bold font indicates the highest (excluding Prop-ora).

		Ono's [5]	MF [8]	KAM [10]	PM [18]	Prop-mix	Prop-ora
Har.	SDR	5.8	8.6	4.9	−8.6	**9.3**	10.3
	SIR	11.2	15.1	**23.1**	6.1	12.3	13.8
	SAR	7.6	10.2	5.1	−7.5	**15.4**	15.4
Per.	SDR	−8.1	−4.2	−4.7	−12.1	**−3.8**	−2.7
	SIR	−2.8	−1.3	−2.3	−3.2	**1.7**	2.8
	SAR	−1.9	3.5	**4.2**	−6.7	1.6	2.6
Ave.	SDR	−1.1	2.0	0.1	−10.4	**2.8**	3.8
	SIR	4.2	6.9	**10.4**	1.5	7.0	8.3
	SAR	2.8	6.9	4.6	−7.1	**8.5**	9.0

4.1. Results

The experimental results are summarized in Table 1. MF outperformed other conventional methods in terms of SDR, which served as the initial value of the proposed method. Since the time-frequency mask was estimated solely from the phase information, PM resulted in the lowest performance. In contrast, the proposed method, which simultaneously treats both amplitude and phase, outperformed conventional methods in terms of SDR. We observed that KAM reduced much components, which significantly improved SIR of harmonic components but induced low SDR and SAR. As a result of taking phase into account, the proposed method achieved higher scores than Ono's which is the non phase-aware version of the proposed method as discussed in Section 3.3.

Comparing Prop-ora with Prop-mix, we confirmed that the accurate estimation of the instantaneous frequency of harmonic components can improve the performance of the proposed method. This is because the first term in Eq. (12) takes a smaller value for harmonic components and penalizes percussive components more by utilizing the appropriate instantaneous frequency. The instantaneous frequency estimation given by Eq. (11) is one of the simplest methods. More specific methods for the accurate estimation of the instantaneous frequency of harmonic components should be considered, which is a future work.

5. CONCLUSION

In this paper, a phase-aware HPSS method through convex optimization was proposed. Based on two HPSS approaches (anisotropic smoothness and sinusoidal model), the proposed method assumes the smoothness of the complex-valued spectrogram of harmonic components calculated by iPC-STFT in the time direction. On the other hand, the time-frame-wise sparsity of percussive spectrograms was considered as a phase insensitive prior. Furthermore, the proposed method considers the perfect reconstruction constraint in the time domain instead of power spectrograms. Through the experiment, the effectiveness of the proposed method was validated in terms of SDR. The experimental results indicated the accurate estimation of the instantaneous frequency of harmonic components can improve the performance of the proposed method, which is included in our future works.

[2] It can be downloaded from `https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/phase_based_harmonic_percussive_separation.html`

6. REFERENCES

[1] J. Reed, Y. Ueda, S. Siniscalchi, Y. Uchiyama, S. Sagayama, and C. Lee, "Minimum classification error training to improve isolated chord recognition," in *Int. Soc. Music Inf. Retrieval (ISMIR)*, Oct. 2009, pp. 609–614.

[2] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stajylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *IEEE Int. Conf. Acoust., Speech. Signal Process. (ICASSP)*, 2012, pp. 421–424.

[3] C. Dittmar and M. Müller, "Reverse engineering the amen break — score-informed separation and restoration applied to drum recordings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1535–1547, Sept. 2016.

[4] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109, Jan. 2014.

[5] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2008, pp. 1–4.

[6] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 465–468.

[7] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2059–2073, Dec. 2014.

[8] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Int. Conf. Digit. Audio Effects (DAFX-10)*, Sept. 2010, pp. 1–4.

[9] 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. 4293–4310, Aug. 2014.

[10] D. Fitzgerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *IET Irish Signals Syst. Conf.*, Jan. 2014, pp. 35–40.

[11] C. Dittmar, P. López-Serrano, and M. Müller, "Unifying local and global methods for harmonic-percussive source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 176–180.

[12] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: An insightful benchmark," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2015, pp. 81–85.

[13] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1095–1105, June 2018.

[14] I. Bayram and M. E. Kamasak, "A simple prior for audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1190–1200, June 2013.

[15] K. Yatabe and Y. Oikawa, "Phase corrected total variation for audio signals," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 656–660.

[16] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Model-based phase recovery of spectrograms via optimization on Riemannian manifolds," in *Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Sept. 2018, pp. 126–130.

[17] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Low-rankness of complex-valued spectrogram and its application to phase-aware audio processing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019.

[18] E. Cano, M. Plumbley, and C Dittmar, "Phase-based harmonic/percussive separation," in *Ann. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sept. 2014, pp. 1628–1632.

[19] K. Drossos, P. Magron, S. I. Mimilakis, and T. Virtanen, "Harmonic-percussive source separation with deep neural networks and phase recovery," in *Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Sept. 2018, pp. 422–425.

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

[21] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[22] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: Application to audio signal restoration," in *Eur. Signal Process.Conf. (EUSIPCO)*, Aug. 2015, pp. 1–5.

[23] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. on Signal Process.*, vol. 43, no. 5, pp. 1068–1089, May 1995.

[24] H. Kameoka, "Multi-resolution signal decomposition with time-domain spectrogram factorization," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2015, pp. 86–90.

[25] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURSIP J. Audio, Speech, Music Process.*, vol. 2014, no. 1, pp. 26, July 2014.

[26] C. Laroche, M. Kowalski, H. Papadopoulos, and G. Richard, "Hybrid projective nonnegative matrix factorization with drum dictionaries for harmonic/percussive source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1499–1511, Sept. 2018.

[27] N. Komodakis and J. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 31–54, Nov. 2016.

[28] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Opt.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[29] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harm. Anal*, vol. 27, no. 3, Jan. 2009.

[30] 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, 2006.