HARMONIC AND PERCUSSIVE SOUND SEPARATION BASED ON MIXED PARTIAL DERIVATIVE OF PHASE SPECTROGRAM

Natsuki Akaishi, Kohei Yatabe, Yasuhiro Oikawa

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

ABSTRACT

Harmonic and percussive sound separation (HPSS) is a widely applied pre-processing tool that extracts distinct (harmonic and percussive) components of a signal. In the previous methods, HPSS has been performed based on the structural properties of magnitude (or power) spectrograms. However, such approach does not take advantage of *phase* that contains useful information of the waveform. In this paper, we propose a novel HPSS method named *MipDroP* that relies only on phase and does not use information of magnitude spectrograms. The proposed MipDroP algorithm effectively examines phase through its mixed partial derivative and constructs a pair of masks for the separation. Our experiments showed that MipDroP can extract percussive components better than the other methods.

***Index Terms*—** Short-time Fourier transform, time-frequency masking, phase derivative, instantaneous frequency, group delay.

1. INTRODUCTION

Harmonic and percussive sound separation (HPSS) is a processing method that decomposes a sound signal into harmonic (sinusoidal) and percussive (impulsive) components as illustrated in Fig. 1. HPSS is widely applied as pre-processing of, e.g., tempo estimation [1], chord recognition [2], collection of musical sounds [3], and music emotion recognition [4]. Due to its importance in various applications, several HPSS methods have been proposed [5–19] based on anisotropic smoothness (AS) [5–7], non-negative matrix factorization (NMF) [13,14], median filtering (MF) [8], kernel additive model (KAM) [9, 10], phase difference (PD) [11], instantaneous phase correction (iPC) [12], and deep neural network (DNN) [15–17].

The standard methodology of HPSS is based on horizontal and vertical smoothness of magnitude in the time-frequency domain. As shown in the right part of Fig. 1, the magnitude spectrogram of harmonic components exhibits a horizontal pattern, while that of percussive components exhibits a vertical pattern. This fact leads to the derivation of the conventional HPSS methods including AS, MF, and KAM. Specifically, the horizontal and vertical patterns are extracted by optimization-based anisotropic smoothing (AS) [5–7], directional median filtering (MF) [8], or horizontal and vertical proximity kernels (KAM) [9, 10]. While these methods are widely accepted in various applications, they ignore some information of spectrograms, i.e., phase. Since phase determines the waveform in the time domain, it should be helpful in detecting the percussive components.

Although phase-aware signal processing has recently gained attentions [20–30], few HPSS methods use phase information. An earlier attempt in HPSS considered phase continuity to enhance a time-frequency mask [11]. Latterly, a phase-aware method was proposed to extract harmonic components based on phase smoothness [12]. While these methods showed potential of considering phase in HPSS, they only used the time-directional relation of phase (i.e., instantaneous frequency). The frequency-directional relation of phase (e.g., group delay) has not been considered in HPSS.

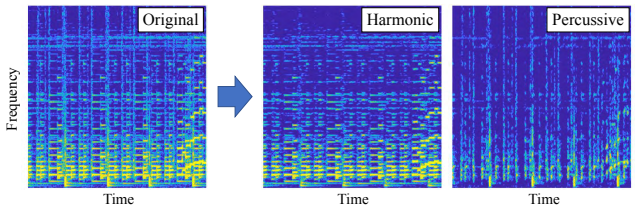

Fig. 1. Example of magnitude spectrograms before and after HPSS.

In this paper, we propose an HPSS method named **MipDroP** (**Mi**xed **p**artial **D**er**i**vative **o**f **P**hase) that simultaneously considers both time- and frequency-directional relations of phase. The proposed MipDroP algorithm constructs a time-frequency mask according to the values of mixed partial derivative of phase. Then, the obtained mask is iteratively refined to improve the quality of separated percussive components. Comparison with existing methods showed that the performance of MipDroP was comparable for harmonic components and was superior for percussive components.

2. RELATED WORKS

2.1. Magnitude-based HPSS methods

The standard HPSS methods have been derived based on directional smoothness of magnitude spectrograms [5–10]. Roughly speaking, these methods smooth the magnitude spectrogram to estimate the harmonic and percussive components for separation.

Let $\mathbf{X}$ be the spectrogram of input signal. Its magnitude $|\mathbf{X}|$ is smoothed in horizontal (time) and vertical (frequency) directions to estimate the harmonic and percussive components, respectively, as

$$\widetilde{\mathbf{H}} = \mathrm{horizontalSmoothing}_{\theta}(|\mathbf{X}|), \tag{1}$$

$$\widetilde{\mathbf{P}} = \mathrm{verticalSmoothing}_{\theta}(|\mathbf{X}|), \tag{2}$$

where $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{P}}$ are tentative estimates of the harmonic and percussive components, respectively, θ denotes parameters of the smoothing processes, and the absolute value $|\cdot|$ is applied element-wise. These smoothing processes may be independently performed by filtering [8] or jointly performed by optimization [5–7, 9, 10].

The smoothed magnitude spectrograms are used for separating the components. A typical method is the time-frequency masking:

$$\mathbf{H} = \mathbf{M}_{\mathrm{H}} \odot \mathbf{X}, \qquad \mathbf{P} = \mathbf{M}_{\mathrm{P}} \odot \mathbf{X}, \tag{3}$$

where $\odot$ represents the element-wise multiplication, and $\mathbf{M}$ is a mask calculated using the tentative estimates $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{P}}$. The separation masks can be obtained, e.g., by comparison of the magnitude (binary masking) or by the following formula (soft masking),

$$\mathbf{M}_{\mathrm{H}} = |\widetilde{\mathbf{H}}|^{\gamma}/(|\widetilde{\mathbf{H}}|^{\gamma}+|\widetilde{\mathbf{P}}|^{\gamma}), \qquad \mathbf{M}_{\mathrm{P}} = |\widetilde{\mathbf{P}}|^{\gamma}/(|\widetilde{\mathbf{H}}|^{\gamma}+|\widetilde{\mathbf{P}}|^{\gamma}), \tag{4}$$

where $\gamma > 0$ is a parameter for adjusting the masks, and the power and division are considered element-wise. To obtain the separated components, the spectrograms $\mathbf{H}$ and $\mathbf{P}$ given in Eq. (3) are inverted into the time domain. This type of HPSS methods relies only on magnitude spectrograms and ignores the phase information.

2.2. Phase-assisted HPSS methods

Some HPSS methods use phase information as assistance of the separation. Since such phase-assisted methods are more related to the proposed method, these methods are reviewed here one by one.

In [11], phase evolution of a sinusoidal component is considered for mask refinement. The phase of a sinusoid, say $\sin(\omega t + \phi)$, evolves at a constant rate (i.e., frequency) ω. In the time-frequency domain, this rate can be given by time-directional difference of the phase at adjacent time-frequency bins. In contrast, time-directional difference of the phase of a non-sinusoidal component does not obey such a constant-rate rule. Thus, a sinusoidal component can be distinguished from the other components by comparing the phase difference with the expected rate. The HPSS method proposed in [11] firstly constructs a mask for harmonic components based on the magnitude, and then the mask is refined using the phase difference. This mask is used for extracting harmonic components, and the remained components are treated as percussive components.

In [12], phase smoothing is considered based on the sinusoidal model. Similar to that in the previous paragraph, the HPSS method proposed in [12] assumes the constant-rate evolution of the phase of a sinusoidal component. This assumption is handled through time-derivative of the phase and a convex optimization algorithm. Although harmonic components can be effectively extracted by the assumed model, percussive components are handled as the residual regularized by an energy-based penalty function.

While these HPSS methods are based on phase information, they take advantage of phase only for harmonic components. Even though the phase is closely related to waveform in the time domain, it has not been considered for extracting percussive components.

3. PROPOSED METHOD

To fully take advantage of phase information, we propose to use mixed partial derivative of the phase for HPSS. The proposed algorithm is named MipDroP and will be introduced in this section.

3.1. Mixed partial derivative of phase spectrogram

To consider derivatives of phase, let the short-time Fourier transform (STFT) be defined in the continuous setting as follows:

$$\mathrm{STFT}(y)(\omega, t) = \int_{-\infty}^{\infty} y(\tau) w(\tau - t) e^{-i\omega(\tau - t)} d\tau \qquad (5)$$

where $y \in L^2(\mathbb{R})$ is the signal to be transformed, $w \in L^2(\mathbb{R})$ is a window function, and $i = \sqrt{-1}$. Hereafter, the spectrogram of y is written as Y $(= \mathrm{STFT}(y))$ for notational convenience.

The first-order partial derivatives of the phase are well-known as *instantaneous frequency*, $\mathrm{IF}(Y)(\omega, t) = (\partial \operatorname{Arg}(Y)/\partial t)(\omega, t)$, and (local) *group delay*, $\mathrm{GD}(Y)(\omega, t) = -(\partial \operatorname{Arg}(Y)/\partial \omega)(\omega, t)$, where $\operatorname{Arg}(Y)$ denotes the phase of the spectrogram Y. These quantities can be numerically computed by finite-difference approximation [31] or using analytic formula [32]. The second-order derivatives of phase are given in a similar manner and can be numerically computed [33]. Among their three variations, $\partial^2/\partial t^2$, $\partial^2/\partial t \partial \omega$, and $\partial^2/\partial \omega^2$, the mixed-partial derivative $\partial^2/\partial t \partial \omega$ is considered in this paper because of the following property.

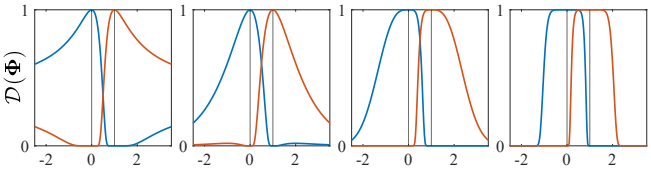

Fig. 2. Example of the proposed mask-generating functions (blue: $\mathcal{D}_{\mathrm{H}}$, red: $\mathcal{D}_{\mathrm{P}}$). From left to right, the shape parameters are varied as $(a, b) = (2, 2)$, $(2, 1)$, $(4, 2)$, and $(10, 1)$. The vertical lines are drawn to emphasize 0 and 1, which are important as in Eq. (6).

The mixed-partial derivative of phase provides cues for discriminating harmonic and percussive components [34]. The instantaneous frequency of a sinusoidal component is constant because the phase evolves at a constant rate. For spectrogram, this is true for both time and frequency directions because the window function spreads the component in the time-frequency domain. Hence, the frequency-directional derivative of instantaneous frequency is small for a sinusoidal components, i.e., $(\partial\,\mathrm{IF}(Y)/\partial\omega)(\omega, t) \approx 0$ if the bin (ω, t) is dominated by a sinusoidal component. In contrast, the group delay of a pulsive component appears as a time-directional linear function that takes zero at the center of gravity of the windowed segment. That is, the time-directional derivative of group delay is constant, i.e., $(\partial\,\mathrm{GD}(Y)/\partial t)(\omega, t) \approx 1$ [due to the definition of Eq. (5)]. Since these quantities $\partial\,\mathrm{IF}(Y)/\partial\omega$ and $\partial\,\mathrm{GD}(Y)/\partial t$ are the same mixed derivative, we can obtain

$$\frac{\partial^2}{\partial t \partial \omega} \operatorname{Arg}(Y)(\omega, t) \approx \begin{cases} 0 & (\text{if } Y(\omega, t) \text{ is sinusoidal}) \\ 1 & (\text{if } Y(\omega, t) \text{ is impulsive}) \end{cases}, \qquad (6)$$

where the details of this relation can be found in [34].

3.2. Mask-generating functions

To convert the mixed partial derivative into time-frequency masks, we propose a mask-generating function as a part of the proposed method. The relation in Eq. (6) indicates that harmonic components can be extracted by remaining the bins whose mixed partial derivative are close to 0 and suppressing those close to 1. On the contrary, percussive components can be extracted by remaining those close to 1 and suppressing those close to 0. Therefore, we propose the following functions to generate the time-frequency masks,

$$\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi}) = \exp(-|\mathbf{\Phi}|^a / |\mathbf{\Phi} - 1|^b), \qquad (7)$$

$$\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi}) = \exp(-|\mathbf{\Phi} - 1|^a / |\mathbf{\Phi}|^b), \qquad (8)$$

where $\mathbf{\Phi}$ represents the mixed partial derivative of phase, and $a, b > 0$ are shape parameters. As shown in Fig. 2, these functions drawn by blue ($\mathcal{D}_{\mathrm{H}}$) and red ($\mathcal{D}_{\mathrm{P}}$) treat 0 and 1 exclusively.

These masks can be directly applied to HPSS as illustrated in Fig. 3. $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi})$ can be used as a mask for extracting harmonic components, and $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi})$ can be used for extracting percussive components. However, as will be shown in the experimental section, these masks may not perform well in practice, which might be because Eq. (6) only holds under the ideal (unrealistic) condition. Therefore, we propose an iterative algorithm that performs well in practice.

3.3. Proposed HPSS algorithm: MipDroP

To improve the performance of HPSS based on $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi})$ and $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi})$ (Fig. 3), we propose a mask refinement method consisting of three processes: (1) time-frequency masking using $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi})$ and $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi})$,

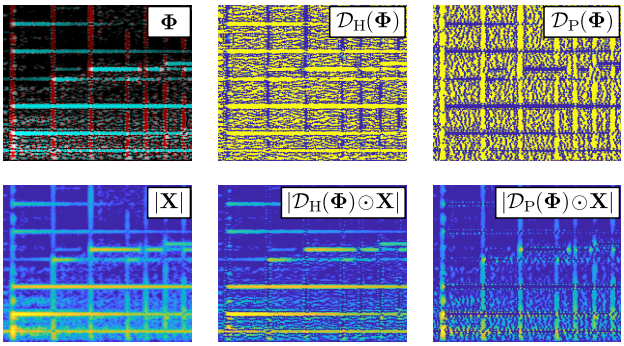

Fig. 3. Example of phase derivative and masking. From top left to bottom right, mixed partial derivative of phase $\mathbf{\Phi}$, mask for harmonic components $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi})$, that for percussive components $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi})$, magnitude spectrogram $|\mathbf{X}|$, and masked spectrograms $|\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi})\odot\mathbf{X}|$ and $|\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi})\odot\mathbf{X}|$. The mixed partial derivative $\mathbf{\Phi}$ is colored so that bins with $\mathbf{\Phi}\approx 0$ becomes blue and bins with $\mathbf{\Phi}\approx 1$ becomes red.

(2) spectrogram smoothing by STFT, and (3) Wiener-like filtering. These processes are iteratively applied as follows.

The proposed algorithm is summarized in **Algorithm 1**. Note that two similar processes applied to both harmonic and percussive components are written in a single line to shorten the number of lines. This algorithm is named **MipDroP** because "**Mi**xed **p**artial **Der**ivative **o**f **P**hase" is the core of the algorithm. Here, each step of the MipDroP algorithm is described one by one.

At first, the target signal $\mathbf{x}$ is inputted into the algorithm. Its spectrogram $\mathbf{X}$ is computed by STFT (1st line), and the variables for both harmonic and percussive components $\mathbf{h}^{[k]}$ and $\mathbf{p}^{[k]}$ are initialized by the inputted signal $\mathbf{x}$ (2nd line). Then, starting from these initial values, the algorithm is iterated K times (3rd to 10th lines), where k represents the iteration count.

The 4th to 6th lines are the time-frequency masking described in Section 3.2 and Fig. 3. The spectrograms of the harmonic and percussive components $\mathbf{H}$ and $\mathbf{P}$ are calculated by STFT (4th line). The mixed partial derivatives of their phase are also calculated numerically (5th line). Then, using the time-frequency masks generated by $\mathcal{D}_{\mathrm{H}}$ and $\mathcal{D}_{\mathrm{P}}$ in Eqs. (7) and (8), the harmonic and percussive components are enhanced by removing the other components (6th line). For the first iteration ($k=1$), these steps are the same as those in Fig. 3 because the variables are initialized as $\mathbf{h}^{[1]}=\mathbf{x}$ and $\mathbf{p}^{[1]}=\mathbf{x}$.

The 7th to 9th lines aim to refine the masks for improving the HPSS performance. The first step of the refinement is the inverse STFT (iSTFT) followed by STFT (7th line). This process improves smoothness of the masked spectrograms as shown in Fig. 4, which fills the bins having very small values caused by the masking. This smoothing effect is related to *spectrogram consistency* (see [27]). Then, these smoothed spectrograms are used for generating Wiener-like masks $|\widetilde{\mathbf{H}}|^2/(|\widetilde{\mathbf{H}}|^2+|\widetilde{\mathbf{P}}|^2)$ and $|\widetilde{\mathbf{P}}|^2/(|\widetilde{\mathbf{H}}|^2+|\widetilde{\mathbf{P}}|^2)$ in Eq. (4). By applying the Wiener-like masks to the inputted signal, harmonic and percussive components are separated (8th and 9th lines). Hence, the 4th to 7th lines (mixed-partial-derivative-based masking and spectrogram smoothing) have a role of obtaining tentative estimates of the harmonic and percussive components for the Wiener filtering.

These procedures are iterated K times to improve the Wiener-like masks in the 8th and 9th lines. This is because the mixed partial derivative can provide effective information only when the harmonic and percussive components are not mixed. That is, Eq. (6) only holds for the bins that are dominated by either a harmonic or percussive component. In the 5th line of the proposed method, the mixed partial derivatives are computed from the components separated in the previous iteration. Therefore, $\mathbf{\Phi}_{\mathrm{H}}$ and $\mathbf{\Phi}_{\mathrm{P}}$ provide more useful information as the iteration proceeds, and hence the masks become more reliable in the later iterations. This iterative mask refinement can improve the quality of separated percussive components that usually have less energy than harmonic components.

Algorithm 1 MipDroP algorithm

Input: $\mathbf{x}$
Output: $\mathbf{h}^{[K+1]}, \mathbf{p}^{[K+1]}$
1: $\mathbf{X}=\mathrm{STFT}(\mathbf{x})$
2: $(\mathbf{h}^{[1]}, \mathbf{p}^{[1]})=(\mathbf{x},\mathbf{x})$
3: **for** $k=1,2,\ldots,K$ **do**
4: $\quad(\mathbf{H},\mathbf{P})=(\mathrm{STFT}(\mathbf{h}^{[k]}),\ \mathrm{STFT}(\mathbf{p}^{[k]}))$
5: $\quad(\mathbf{\Phi}_{\mathrm{H}},\mathbf{\Phi}_{\mathrm{P}})=(\partial^2\,\mathrm{Arg}(\mathbf{H})/\partial t\partial\omega,\ \partial^2\,\mathrm{Arg}(\mathbf{P})/\partial t\partial\omega)$
6: $\quad(\hat{\mathbf{H}},\hat{\mathbf{P}})=(\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi}_{\mathrm{H}})\odot\mathbf{H},\ \mathcal{D}_{\mathrm{P}}(\mathbf{\Phi}_{\mathrm{P}})\odot\mathbf{P})$
7: $\quad(\widetilde{\mathbf{H}},\widetilde{\mathbf{P}})=(\mathrm{STFT}(\mathrm{iSTFT}(\hat{\mathbf{H}})),\ \mathrm{STFT}(\mathrm{iSTFT}(\hat{\mathbf{P}})))$
8: $\quad\mathbf{h}^{[k+1]}=\mathrm{iSTFT}(\mathbf{X}\odot|\widetilde{\mathbf{H}}|^2/(|\widetilde{\mathbf{H}}|^2+|\widetilde{\mathbf{P}}|^2))$
9: $\quad\mathbf{p}^{[k+1]}=\mathrm{iSTFT}(\mathbf{X}\odot|\widetilde{\mathbf{P}}|^2/(|\widetilde{\mathbf{H}}|^2+|\widetilde{\mathbf{P}}|^2))$
10: **end for**

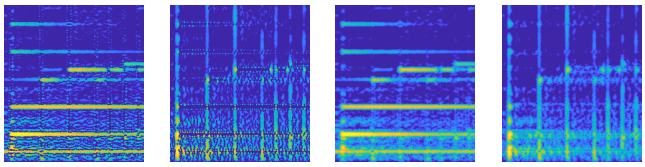

Fig. 4. Example of the spectrogram smoothing in the 7th line of Algorithm 1. The masked spectrograms in Fig. 3 are shown in the left half, and their smoothed versions are shown in the right half.

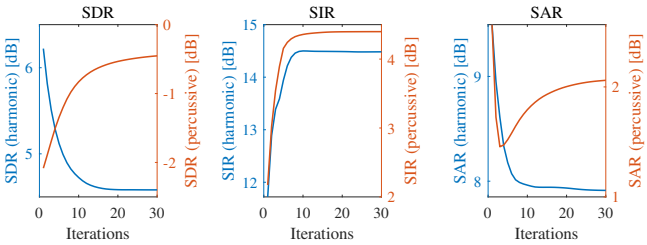

Fig. 5. Separation performance of the proposed method for each iteration. The performance for harmonic (blue) and percussive (red) components are illustrated using different scales.

4. EXPERIMENTS

In this paper, four experiments were performed. After the properties of the proposed method are shown by three experiments, comparison with the existing methods is provided at the end of this section. All the methods were applied to the 10 audio tracks[1] used in the literature of HPSS [11]. The sampling rate was 44.1 kHz, and STFT was calculated using a 10 000-sample-long Chebyshev window with 256-sample shifting. The number of frequency bins was the same as the window length. All the results in this section are given as the median of SDR, SIR, and SAR over the 10 audio tracks.

[1]These audio tracks can be downloaded from the following webpage. https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/phase_based_harmonic_percussive_separation.html

Table 1. Variations of the proposed algorithm tested in Section 4.2.

Procedure	Line #	Full	(a)	(b)	(c)
STFT & Compute derivative	4–5	✓	✓	✓	✓
Compute masks using $\mathcal{D}_{\mathrm{H}}$ and $\mathcal{D}_{\mathrm{P}}$	6	✓	✓	✓	✓
Apply masks $\mathcal{D}_{\mathrm{H}}$ to $\mathbf{H}$ and $\mathcal{D}_{\mathrm{P}}$ to $\mathbf{P}$	6	✓	✓	✓	
Perform iSTFT & STFT	7	✓			
Apply Wiener-like masks	8–9	✓		✓	✓

Table 2. Separation performance of the proposed method and its variations in Table 1 at the 1st and 10th iterations.

		1st iteration				10th iteration			
		Full	(a)	(b)	(c)	Full	(a)	(b)	(c)
Har.	SDR	**6.22**	−13.0	6.12	6.18	4.73	−12.9	4.89	**5.56**
	SIR	11.7	**14.3**	11.2	10.0	14.5	**22.7**	13.3	9.33
	SAR	**9.49**	−12.0	9.34	10.2	7.96	−11.7	8.03	**10.1**
Per.	SDR	**−2.08**	−20.9	−2.39	−4.12	**−0.83**	−22.1	−1.34	−5.61
	SIR	**2.17**	−4.36	1.57	−1.01	**4.35**	−4.15	4.07	−4.48
	SAR	2.57	−16.0	2.42	**4.12**	1.79	−14.7	1.46	**6.92**

4.1. Effect of iteration

First, the effect of iteration was investigated. The separation performance up to the 30th iteration is shown in Fig. 5, where the shape parameters were set to $a = 10$, $b = 1$. By iterating the algorithm, SIR of both components increased and saturated around the 10th iteration. SDR and SAR of the percussive components increased, but those of the harmonic components decreased. The results of the proposed algorithm can be adjusted by the number of iterations.

4.2. Importance of each procedure

Second, the variations of the proposed algorithm given in Table 1 were tested. "Full" represents the proposed algorithm, and the others (a,b,c) lack some of the procedures (those without marks). The results are summarized in Table 2, where the shape parameters were set to $a = 10$, $b = 1$. When the Wiener filtering was omitted (a), SDR and SAR for both harmonic and percussive components marked the worst scores. This result indicates that the Wiener filtering is mandatory to obtain proper results. Comparing the proposed algorithm (Full) with that lacking the spectrogram smoothing (b), they performed similarly for harmonic components, but the proposed algorithm performed better for percussive components. Thus, the spectrogram smoothing in the 7th line is effective for extracting percussive components. This process can be omitted if the processing speed is more important than the performance. Finally, directly using $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi}_{\mathrm{H}})$ and $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi}_{\mathrm{P}})$ for the Wiener filtering (c) performed well for the harmonic components but worse for the percussive components. Therefore, using $\mathcal{D}_{\mathrm{H}}(\mathbf{\Phi}_{\mathrm{H}})$ and $\mathcal{D}_{\mathrm{P}}(\mathbf{\Phi}_{\mathrm{P}})$ as time-frequency masks is important for extracting percussive components.

4.3. Effect of shape parameters

Third, the effect of shape parameters a and b was investigated. They were varied from 1 to 10, and hence 100 combinations were tested. The results are summarized in Fig. 6, where the number of iterations K was set to 10 for all conditions. As the colorbars indicate, the performance may vary about 1 dB by the shape parameters. In the figure, some trade-off between the performance for harmonic and percussive components can be seen: the proposed method performs well for harmonic components if $a < b$ and for percussive components if $a > b$. The maximum SDR was attained at the bottom right (harmonic) and top left (percussive) in the range given in Fig. 6 (their specific values are given in Table 3). This result suggests that the performance of the proposed method can be adjusted by selecting a combination of the shape parameters.

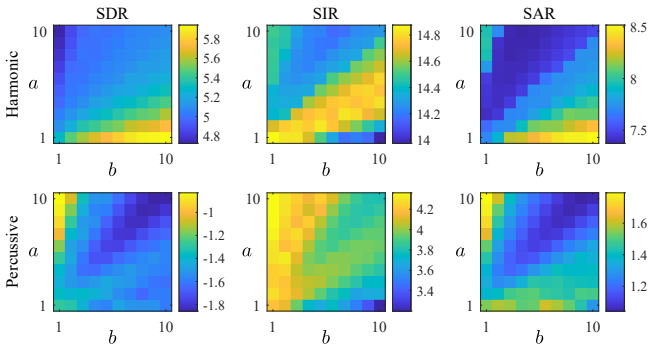

Fig. 6. Separation performance for each combination of the shape parameters a and b. The color represents SDR, SIR, or SAR in dB.

Table 3. Comparison of the separation performance. For reference, the proposed method with shape parameters $(a, b) = (1, 10)$ and $(a, b) = (10, 1)$ are also presented at the right end.

		MF [8]	KAM [9]	PD [11]	iPC [12]	Prop.	$(a,b) = (1,10)$	$(a,b) = (10,1)$
Har.	SDR	3.73	3.74	−7.40	**6.60**	4.66	5.91	4.58
	SIR	8.67	**23.5**	6.78	11.6	14.7	13.2	14.5
	SAR	**12.0**	3.77	−6.53	9.52	8.01	8.74	7.91
Per.	SDR	−4.79	−3.30	−11.1	−1.62	**−0.38**	−1.56	−0.45
	SIR	−0.08	−0.09	−2.01	2.25	**4.47**	2.87	4.39
	SAR	1.67	**4.58**	−6.85	3.41	2.07	1.14	2.06

4.4. Comparison with existing methods

Finally, the proposed method was compared with the existing methods, MF [8], KAM [9], PD [11], and iPC [12] (see Section 1 for the abbreviation). The results are summarized in Table 3, where the shape parameters were set to $a = 15$, $b = 1$, and the number of iterations was $K = 30$. For harmonic components, the proposed method achieved the second-best SDR and SIR, where the best SDR and SIR were obtained by iPC and KAM, respectively. This result indicates that the performance of the proposed method is balanced for harmonic components compared to the other method. For percussive components, the proposed method achieved the best SDR and SIR. It can be said that the proposed method is able to extract percussive components well when the parameters were selected properly. Note that, although the superiority of the proposed method was shown in terms of SDR and SIR, its SAR was in the middle of the compared methods. Improving SAR is remained as a future work.

5. CONCLUSION

In this paper, we proposed the HPSS method that performs well especially for percussive components. It effectively uses mixed partial derivative of phase to design the Wiener-like masks. The masks are iteratively refined by computing the phase derivative from tentative separation results obtained in the previous iteration. The future directions of the research include improvement of separation quality by enhancing the phase of separated components.

6. REFERENCES

[1] B. McFee and D. P. W. Ellis, "Better beat tracking through robust onset aggregation," in *ICASSP*, 2014, pp. 2154–2158.

[2] J. Reed, Y. Ueda, S. Siniscalchi, Y. Uchiyama, S. Sagayama, and C. Lee, "Minimum classification error training to improve isolated chord recognition," in *ISMIR*, 2009, pp. 609–614.

[3] C. Guedes, K. Ganguli, C. Plachouras, S. Senturk, and A. J. Eisenberg, "Mapping timbre space in regional music collections using harmonic-percussive source separation (HPSS) decomposition," in *2nd Int. Conf. Timbre*, 2020.

[4] H. Liu, Y. Fang, and Q. Huang, "Music emotion recognition using a variant of recurrent neural network," in *MMSSA*, 2018, pp. 15–18.

[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 *EUSIPCO*, 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 *ICASSP*, 2012, pp. 465–468.

[7] H. Tachibana, N. Ono, H. Kameoka, 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, 2014.

[8] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *DAFx*, 2010, pp. 246–253.

[9] D. FitzGerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *ISSC/CIICT*, 2014, pp. 35–40.

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

[11] E. Cano, M. Plumbley, and C. Dittmar, "Phase-based harmonic/percussive separation," in *Interspeech*, 2014, pp. 1628–1632.

[12] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-aware harmonic/percussive source separation via convex optimization," in *ICASSP*, 2019, pp. 985–989.

[13] C. Dittmar, P. López-Serrano, and M. Müller, "Unifying local and global methods for harmonic-percussive source separation," in *ICASSP*, 2018, pp. 176–180.

[14] J. Park and K. Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints," in *ISMIR*, 2015, pp. 148–154.

[15] C. Lordelo, E. Benetos, S. Dixon, S. Ahlback, and P. Ohlsson, "Adversarial unsupervised domain adaptation for harmonic-percussive source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 81–85, 2021.

[16] W. Lim and T. Lee, "Harmonic and percussive source separation using a convolutional auto encoder," in *EUSIPCO*, 2017, pp. 1804–1808.

[17] K. Drossos, P. Magron, S. I. Mimilakis, and T. Virtanen, "Harmonic-percussive source separation with deep neural networks and phase recovery," in *IWAENC*, 2018, pp. 421–425.

[18] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *ISMIR*, 2014, pp. 611–616.

[19] R. Füg, A. Niedermeier, J. Driedger, S. Disch, and M. Müller, "Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," in *ICASSP*, 2016, pp. 445–449.

[20] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.

[21] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Commun.*, vol. 81, pp. 1–29, 2016.

[22] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, Wiley, 2016.

[23] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of complex spectrogram via phase conversion," *Acoust. Sci. & Tech.*, vol. 40, no. 3, pp. 170–177, 2019.

[24] K. Yatabe and Y. Oikawa, "Phase corrected total variation for audio signals," in *ICASSP*, 2018, pp. 656–660.

[25] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Low-rankness of complex-valued spectrogram and its application to phase-aware audio processing," in *ICASSP*, 2019, pp. 855–859.

[26] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa, "Joint amplitude and phase refinement for monaural source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 1939–1943, 2020.

[27] K. Yatabe, "Consistent ICA: Determined BSS meets spectrogram consistency," *IEEE Signal Process. Lett.*, vol. 27, pp. 870–874, 2020.

[28] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2020, no. 1, pp. 1–35, 2020.

[29] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 37–50, 2021.

[30] K. Kobayashi, Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-recovery algorithm for harmonic/percussive source separation based on observed phase information and analytic computation," *Acoust. Sci. & Tech.*, vol. 42, no. 5, pp. 261–269, 2021.

[31] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 360–371, 2006.

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

[33] D. J. Nelson, "Instantaneous higher order phase derivatives," *Digit. Signal Process.*, vol. 12, no. 2, pp. 416–428, 2002.

[34] S. A. Fulop and K. Fitz, "Separation of components from impulses in reassigned spectrograms," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1510–1518, 2007.