Phase-recovery algorithm for harmonic/percussive source separation based on observed phase information and analytic computation

Kenji Kobayashi*, Yoshiki Masuyama†, Kohei Yatabe‡ and Yasuhiro Oikawa§

Department of Intermedia Art and Science, Waseda University, 3–4–1, Ohkubo, Shinjuku-ku, Tokyo, 169–8555 Japan

(Received 31 October 2020, Accepted for publication 14 May 2021)

Abstract: Phase recovery is a methodology of estimating a phase spectrogram that is reasonable for a given amplitude spectrogram. For enhancing the signals obtained from the processed amplitude spectrograms, it has been applied to several audio applications such as harmonic/percussive source separation (HPSS). Because HPSS is often utilized as preprocessing of other processes, its phase recovery should be simple. Therefore, practically effective methods without requiring much computational cost, such as phase unwrapping (PU), have been considered in HPSS. However, PU often results in a phase that is completely different from the true phase because (1) it does not consider the observed phase and (2) estimation error is accumulated with time. To circumvent this problem, we propose a phase-recovery method for HPSS using the observed phase information. Instead of accumulating the phase as in PU, we formulate a local optimization model based on the observed phase so that the estimated phase remains similar to the observed phase. The analytic solution to the proposed optimization model is provided to keep the computational cost cheap. In addition, iterative refinement of phase in the existing methods is applied for further improving the result. From the experiments, it was confirmed that the proposed method outperformed PU.

Keywords: Sinusoidal model, Instantaneous frequency, Local phase matching, Nonconvex optimization, Analytic global solution

1. INTRODUCTION

Harmonic/Percussive source separation (HPSS) is a widely utilized preprocessing method in audio applications [1–8], such as automatic transcription [1], chord detection [2], and beat tracking [3]. It aims to separate harmonic and percussive components (denoted by $s_h$ and $s_p$, respectively) from their mixture $s_{obs} = s_h + s_p$. This separation is usually performed in the time-frequency domain, where the signals are represented using complex numbers that comprise amplitude and phase. Although separation of amplitude is the main interest in HPSS, phase is also critical to recovering the waveform in the time domain from the separated components in the time-frequency domain.

Phase recovery, which tries to obtain a phase spectrogram that is reasonable with respect to the given amplitude, has been applied in signal enhancement [9–14], generation [15–18], and separation [19–21]. It has also been applied to HPSS after the estimation of the amplitude spectrograms [22,23] for improving the quality of separation. A popular phase-recovery method is temporal linear phase unwrapping (PU) [24], which recursively calculates the subsequent phase from the instantaneous frequency based on the sinusoidal model, as described in Sect. 2. PU is a computationally efficient method, which makes it popular in HPSS because HPSS is mainly used as preprocessing whose computation should be cheap. In HPSS, it is usually applied to the harmonic components, and the phase of the percussive components is intact or iteratively estimated using enhanced harmonic components [23].

However, PU often results in a phase that is completely different from the true phase. PU is originally proposed for a problem of phase recovery from an amplitude spectrogram, i.e., phase is not given in observation. Therefore, the generated phase usually does not match with the true phase unless some other specific information on phase is additionally given. Its recursive nature accumulates the estimation error with time, which may also reduce the reliability. Yet, the observation in HPSS contains the phase which should include the important information of the
signals. By considering the observed phase, it should be possible to obtain a phase-recovery method that is suitable for HPSS.

In this study, a phase-recovery method that considers the observed phase is proposed for HPSS. It is formulated as a simple optimization problem using the observed phase and instantaneous frequency. Because the formulation is locally defined, it does not accumulate the estimation error. The analytic solution to the proposed formulation is provided so that the computational cost does not considerably increase compared to PU. In addition, an iterative refinement algorithm as in [25] is applied to the proposed method for further improvement. The effectiveness of the proposed method was tested by numerical experiments, and the results indicated that the proposed method outperformed PU.

The rest of the paper is organized as follows. In Sect. 2, the sinusoidal model (which is the basis of PU and the proposed method) and PU are briefly reviewed. Then, the proposed method is described in Sect. 3. After the iterative refinement method used in the experiments is reviewed in Sect. 4, the effectiveness of the proposed method is investigated through experiments in Sect. 5. Finally, this paper is concluded in Sect. 6.

2. SINUSOIDAL-MODEL-BASED PHASE RECOVERY

A popular methodology of phase recovery is based on the sinusoidal model [23–31]. It locally considers a signal \( s = [s_1, \ldots, s_L]^T \) as the sum of \( N \) sinusoids as follows:

\[
s_l = \sum_{n=0}^{N-1} A_n e^{2\pi j f_n l + j \phi_n},
\]

where \( A_n, f_n, \) and \( \phi_n \) denote the amplitude, frequency, and initial phase of the \( n \)th sinusoid, respectively, and \( j = \sqrt{-1} \) denotes the imaginary unit. When this signal model is valid for the \( r \)th and \((r+1)\)th windowed segments (and when the frequencies are sufficiently separated), the phase spectrogram of the signal, \( \phi \), obeys the following relation:

\[
\phi_{\xi,\tau+1} = \phi_{\xi,\tau} + 2\pi a \phi_{\xi,\tau},
\]

where \( a \) is the window-shifting step, \( \phi_{\xi,\tau} \) is the instantaneous frequency at the \((\xi,\tau)\)th time-frequency bin, and \( \xi \) and \( \tau \) are the frequency and time indices, respectively. This relation shows that the phase of a harmonic signal can be estimated from that of the adjacent bin via the instantaneous frequency.

PU is one of the most well-known phase-recovery methods based on the above-mentioned sinusoidal model. Starting from some initial values \( \phi_{\xi,0} \) and \( \phi_{\xi,\tau} \), it recursively estimates the subsequent phase \( \phi_{\xi,\tau+m} \) \((m > 0)\) by directly calculating Eq. (2). Such model-based phase-recovery methods were utilized for source separation [25], speech synthesis [26], and enhancement [27–30]. To calculate the phase in this manner, the instantaneous frequency at each time-frequency bin, \( \phi_{\xi,\tau} \), is required.

There are several methods for estimating the instantaneous frequency. When the available information is only amplitude spectrum (which is an often-encountered situation in, e.g., speech synthesis), a popular choice is the quadratic interpolation method [32], which estimates the instantaneous frequency by finding the peak positions of the amplitude spectrogram. However, when the phase information is available as in HPSS, the instantaneous frequency can be directly calculated by differentiating the phase as follows [33]:

\[
\phi_{\xi,\tau} = \frac{\partial \phi_{\xi,\tau}}{\partial \tau} = \Im \left[ \frac{1}{\delta_{\xi,\tau}} \frac{\partial S_{\xi,\tau}}{\partial \tau} \right],
\]

where \( \tau \) denotes the time, \( \Im[\cdot] \) denotes the imaginary part, and \( S_{\xi,\tau} \) denotes the \((\xi,\tau)\)th time-frequency bin of the complex spectrogram of the signal. The derivative on the right-hand side can be easily calculated using a differential window [34] or approximated using a finite difference method [35,36].

3. PHASE RECOVERY METHOD USING OBSERVED PHASE INFORMATION

In this section, we propose a local phase-matching method based on the sinusoidal model to consider the observed phase in the harmonic phase recovery. Since usual HPSS algorithms output amplitude spectrograms of harmonic and percussive components, their phase should be recovered to obtain better waveforms. The proposed method aims to recover the phase of harmonic component because it can be modeled well by the sinusoidal model. Unlike PU, the proposed method considers the observed phase which is available for HPSS. To do so, we formulate an optimization problem for local phase matching and provide its analytic solution for efficient computation. For enhancing the percussive component, the iterative phase refinement method [25] is applied afterward, which is described in the next section.

3.1. Formulation

Based on the sinusoidal model, the phase of the harmonic component at the \((\xi,\tau)\)th time-frequency bin can be predicted based on the adjacent bins at \((\xi,\tau-1)\) and \((\xi,\tau+1)\) as follows:

\[
\phi_{\xi,\tau} = \phi_{\xi,\tau-1} + 2\pi a \phi_{\xi,\tau},
\]

\[
\phi_{\xi,\tau+1} = \phi_{\xi,\tau+1} - 2\pi a \phi_{\xi,\tau+1},
\]

where \( \phi_{\xi,\tau} \) denotes the observed phase at the \((\xi,\tau)\)th bin, \( \phi_{\xi,\tau} \) denotes the instantaneous frequency calculated from the observed signal as in Eq. (3), and \( \phi_{\xi,\tau} \) and \( \phi_{\xi,\tau} \) are the
time-forward and time-reversed predictions of the phase based on the previous and next bins, respectively. If the signal comprises only harmonic components well-described by the sinusoidal model, the predicted phase values, \( \phi_{\xi,\tau}^+ \) and \( \phi_{\xi,\tau}^- \), should be the same as the observed phase value, \( \phi_{\text{obs},\xi,\tau} \). In reality, however, they do not coincide because of the perturbation due to the percussive components. Using \( \phi_{\text{obs},\xi,\tau}^+ \) and \( \phi_{\text{obs},\xi,\tau}^- \), the proposed method aims to estimate the phase of the harmonic component, denoted by \( \psi_{\xi,\tau} \).

To estimate the phase in a computationally efficient manner, an optimization problem is formulated time-frequency-bin-wise. As the phase is an an interval variable [37], the following is a natural measure for quantifying the closeness of the two phase values \( \phi_1 \) and \( \phi_2 \):

\[
\mathcal{D}(\phi_1, \phi_2) = 1 - \cos(\phi_1 - \phi_2),
\]

which takes the minimal value of 0 when the phase values are the same up to the addition of integer multiples of 2\pi. Using this measure, we formulate a scalar-valued optimization problem for estimating the phase at each time-frequency bin as follows:

\[
\psi_{\xi,\tau} = \arg \min_x \lambda^+ \mathcal{D}(x, \phi_{\text{obs},\xi,\tau}^+) + \lambda \mathcal{D}(x, \phi_{\text{obs},\xi,\tau}^-) + \lambda^- \mathcal{D}(x, \phi_{\text{obs},\xi,\tau}),
\]

(7)

where \( \lambda^+, \lambda, \lambda^- > 0 \) denote weights that will be described in the next subsection. Regarding the weighted least squares method, a solution to this problem can be interpreted as the weighted average in terms of phase, i.e., this optimization problem attempts to merge \( \phi_{\text{obs},\xi,\tau}^+ \) and \( \phi_{\text{obs},\xi,\tau}^- \) based on the weights that are selected by their reliability.

The concepts of PU and the proposed method are depicted in Fig. 1. Although PU in Eq. (2) propagates the phase in the forward direction only, the proposed method in Eq. (7) considers both forward and backward directions. The estimated result of PU depends on the previously estimated phase, while the proposed method estimates each phase independently of each other, which enables the parallel implementation. In addition, the proposed method considers the observed phase in contrast to PU.

3.2. Selection of Weights

The weights \( \lambda^+, \lambda, \lambda^- \) should be selected such that they appropriately reflect the reliability of the phase values. Based on the Wiener filter for source separation, we propose to utilize bin-wise weights, \( \lambda = \lambda_{\xi,\tau} \), which are calculated using the separated amplitude spectrogram \( A_{\xi,\tau} \) as follows:

\[
\lambda_{\xi,\tau}^+ = \frac{(A_{\xi,\tau}^h)^2}{\sum_{n=-1}^{1}(A_{\xi,\tau+n}^h)^2 + (A_{\xi,\tau+n}^p)^2},
\]

(8)

\[
\lambda_{\xi,\tau} = \frac{(A_{\xi,\tau}^h)^2}{\sum_{n=-1}^{1}(A_{\xi,\tau+n}^h)^2 + (A_{\xi,\tau+n}^p)^2},
\]

(9)

\[
\lambda_{\xi,\tau}^- = \frac{(A_{\xi,\tau}^p)^2}{\sum_{n=-1}^{1}(A_{\xi,\tau+n}^h)^2 + (A_{\xi,\tau+n}^p)^2},
\]

(10)

where \( A_{\xi,\tau}^h \) and \( A_{\xi,\tau}^p \) represent the separated amplitudes of the harmonic and percussive components, respectively. Note that the proposed method imposes no restriction on the selection of these weights, and any other rules are acceptable for them. These formulae were empirically selected based on our preliminary experiments.

3.3. Analytic Solution to Eq. (7)

Because Eq. (7) is a scalar-valued optimization problem, it can be easily solved as follows. The objective function without the constant terms can be written as

\[
g(x) = -\lambda^+ \cos(x - \phi^+) - \lambda \cos(x - \phi_{\text{obs}}) - \lambda^- \cos(x - \phi^-),
\]

(11)

where the subscripts (\( \xi \) and \( \tau \) of \( \lambda_{\xi,\tau} \) and \( \phi_{\xi,\tau} \)) are omitted in this subsection for convenience. Its derivative with respect to \( x \) is obtained as follows:

\[
\frac{dg}{dx}(x) = \sqrt{\alpha_x^2 + \alpha_c^2} \sin(x - \text{atan2}(\alpha_c, \alpha_x)),
\]

(12)

where \text{atan2}(\cdot, \cdot) denotes the two-argument arctangent, and \( \alpha_x \) and \( \alpha_c \) are the coefficients calculated by

\[
\alpha_x = \lambda^+ \cos(\phi^+) + \lambda \cos(\phi_{\text{obs}}) + \lambda^- \cos(\phi^-),
\]

(13)

\[
\alpha_c = \lambda^+ \sin(\phi^+) + \lambda \sin(\phi_{\text{obs}}) + \lambda^- \sin(\phi^-).
\]

(14)

The zero-crossing points of this derivative are located at \( x = \text{atan2}(\alpha_c, \alpha_x) + \pi k \), where \( k \) is an arbitrary integer. Therefore, the solution to Eq. (7) is either \( \text{atan2}(\alpha_c, \alpha_x) + 2\pi k \) or \( \text{atan2}(\alpha_c, \alpha_x) + \pi + 2\pi k \) because a zero-crossing point of Eq. (12) is either the minimum or maximum point.
To distinguish these two, the second derivative of the objective function,
\[
\frac{d^2 \tilde{g}}{dx^2}(x) = \alpha_x \cos(x) + \alpha_c \sin(x),
\]
is considered. As it is negative at the maxima, the solution to Eq. (7) should be modified by adding \( \pi \) whenever the second derivative is negative. Thus, by introducing an indicator function \( \chi \) for the second-order optimality,
\[
\chi(x) = \begin{cases} 
1 & (\alpha_x \cos(x) + \alpha_c \sin(x) < 0) \\
0 & \text{otherwise}
\end{cases},
\]
the solution to Eq. (7) can be expressed as follows:
\[
\psi_{x_0} = \text{atan2}(\alpha_c, \alpha_x) + \pi \chi(\text{atan2}(\alpha_c, \alpha_x))
\]
up to the 2\( \pi \) ambiguity whose determination is not important for phase recovery.

The proposed method is summarized in Algorithm 1. For each time-frequency bin, the proposed method requires 17 multiplications, 13 additions, 3 divisions, 8 trigonometric evaluations, 1 arctan operation, and 1 comparison. Although the number of computations of the proposed method is greater than that of PU (1 multiplication with the pre-calculated constant 2\( \pi a \) and 1 addition), it should be a reasonable amount for supporting the computational efficiency. Note that the proposed method can be computed in parallel for each bin (in contrast to PU that requires previously estimated results for calculating the current segment). In addition, as will be discussed at the end of Sect. 5, the number of computations of the proposed method can be reduced to 5 multiplications, 8 additions, 8 trigonometric evaluations, 1 arctan operation, and 1 comparison when the window-shifting step is sufficiently small.

4. ITERATIVE REFINEMENT OF PHASE

In addition to the proposal in the previous section, we propose applying the following iterative phase refinement method for enhancing the percussive components. In [25], an iterative phase-refinement method was formulated as an optimization problem for determining complex spectrograms \( X^h \) and \( X^p \) as follows:
\[
\min_{X^h, X^p} \| S^{obs} - X^h - X^p \| _2 \quad \text{s.t.} \quad \| X^h \| = A^h, \quad \| X^p \| = A^p,
\]
where \( S^{obs} \) is the observed complex spectrogram, \( \| \cdot \| \) is the Euclidean norm, and \( \| \cdot \| \) is the element-wise absolute value. An iterative method for solving this problem is summarized in Algorithm 2, where Arg(\( \cdot \)), exp(\( \cdot \)), and \( \odot \) are the element-wise complex argument, element-wise exponential function, and element-wise multiplication, respectively. \( \mu^h \) and \( \mu^p \) are the element-wise step sizes that are calculated as follows:

Algorithm 1 Proposed Phase-Recovery Method

**Input:** Phase spectrogram \( \phi^{obs} \), instantaneous frequency \( \nu^{obs} \), estimated amplitude spectrograms \( \{A^h, A^p\} \)
**Output:** Estimated phase spectrogram \( \psi \)
1: Calculate \( \{h^+, \phi^h\} \) from \( \{\phi^{obs}, \nu^{obs}\} \) by Eqs. (4), (5)
2: Calculate \( \{A^+, A^-\} \) from \( \{A^h, A^p\} \) by Eqs. (8)–(10)
3: Calculate \( \{\alpha_x, \alpha_c\} \) from \( \{\phi^h, \phi^{obs}, \phi^-, A^+, A^-\} \) by Eqs. (13), (14)
4: Calculate \( \psi \) from \( \{\alpha_x, \alpha_c\} \) by Eqs. (16), (17)

Algorithm 2 Iterative Phase Refinement

**Input:** Complex spectrogram \( S^{obs} \), estimated phase spectrogram \( \psi \), estimated amplitude spectrograms \( \{A^h, A^p\} \)
**Output:** Refined phase spectrograms \( \{\psi^h, \psi^p\} \)
1: \( X_0^h = A^h \odot \exp(j\psi) \)
2: \( X_0^p = A^p \odot \exp(j\text{Arg}[S^{obs}]) \)
3: for \( k = 0 \) to \( K - 1 \) do
4: \( Y^h = X_k^h + \mu^h \odot (S^{obs} - X_k^h - X_k^p) \)
5: \( Y^p = X_k^p + \mu^p \odot (S^{obs} - X_k^h - X_k^p) \)
6: \( X_{k+1}^h = A^h \odot \exp(j\text{Arg}[Y^h]) \)
7: \( X_{k+1}^p = A^p \odot \exp(j\text{Arg}[Y^p]) \)
8: end for
9: \( \psi^h = \text{Arg}[X_{K}^h] \)
10: \( \psi^p = \text{Arg}[X_{K}^p] \)

\[
\mu^h = \frac{(A^h)^2}{(A^h)^2 + (A^p)^2},
\]
\[
\mu^p = \frac{(A^p)^2}{(A^h)^2 + (A^p)^2},
\]
where the square and division are performed element-wise. This algorithm can assist the phase recovery when the window-shifting step is not small, as shown below.

5. NUMERICAL EXPERIMENTS

5.1. Experimental Conditions

As the target of HPSS, we utilized DSD100, which is a dataset of musical signals proposed for the 2016 signal separation evaluation campaign [38]. DSD100 is provided as two sets, Dev and Test, where each set contains 50 audio tracks (sampled at 44,100 Hz) that comprise 4 components (bass, drums, vocals, and other). In the experiments, drums was considered the percussive component, while a mixture of the other three components was used as the harmonic component. All 50 songs from the Test set were utilized, and a 10-second-long segment was selected for each song. The signals were down-mixed to monaural by averaging the left and right channels [23]. The short-time Fourier transform (STFT) and instantaneous frequency were calculated using the LTFAT toolbox [39] with a 4096-sample Hann window, whose shifting step was
set to 2,048, 1,024, and 512 samples (corresponding to the 1/2-, 1/4-, and 1/8-shifting, respectively). For the evaluation, the source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) were calculated using the BSS Eval toolbox [40].

5.2. Effect of Iterative Refinement

Here, we investigated the effect of the iterative phase refinement in Algorithm 2. The iterative refinement was applied to the phase spectrograms obtained using PU and the proposed method, where the amplitude spectrograms, $A^b$ and $A^p$, were obtained by employing the ideal Wiener filter calculated using the true amplitudes. The mean and standard deviation (SD) for all 50 songs for each iteration were calculated and summarized in Fig. 2. Note that the 0th iteration at the left end represents the results of PU and the proposed method without iterative refinement.

From the figure, it can be seen that the proposed method without iteration (left end) tends to perform better when the window-shifting step is smaller. Specifically, the results for 1/8-shifting (7/8-overlap) in the bottom row suggest that the iterative refinement is not necessary for the proposed method when the window shifting is sufficiently small. This should be because the predictions of the phase based on the instantaneous frequency become better for such situations. However, when the window shifting was large, the iterative algorithm refined the results, as seen in the top row. For such situations, the proposed method achieved better scores than those achieved by PU at the same iteration, or the proposed method obtained scores similar to those of PU with fewer iterations.

At the right end of the figure, SDR for the first 10 iterations is shown. For the 7/8- and 3/4-overlap conditions, no iteration or few iterations are required for the proposed method. In contrast, PU requires 50 iterations for Algorithm 2 to achieve the same performance. This result illustrates the computational efficiency of the proposed method when they are combined with the iterative refinement. While the proposed method itself needs more operations as discussed at the end of Sect. 3.3, the proposed method is much efficient compared to PU in total because the number of required iterations of the iterative refinement for the proposed method can be very small. Although the number of required iterations increased for the 1/2-overlap case, the proposed method performed better than PU when the number of iterations is small.

An example of a phase recovered using the proposed method is depicted in Fig. 3, where the so-called baseband phase difference [27] is shown instead of the phase for better visibility. By comparing the bottom-left and bottom-right figures, it can be seen that the recovered phase was more similar to the original phase than the observed (mixed) phase at the bottom center.

5.3. Application to Several HPSS Methods

The proposed method is compared with other phase-recovery methods. For the amplitude estimation of HPSS, the median filtering [41] and kernel additive model (KAM)
were utilized in addition to the ideal Wiener filter. The parameters of the median filter and KAM were set to the suggested values provided in their original paper. As phase-recovery methods, PU and PU with iterative refinement (PUIter) were applied. Two computationally expensive methods, the manifold-optimization-based iterative method (ManOpt) [37] and the multiple input spectrogram inversion (MISI) [19], were also applied as reference. ManOpt [37] is a sinusoidal-model-based method that considers the inter-frequency relation of a sinusoid, i.e., the proposed method can be interpreted as a simplified version of ManOpt. MISI [19] is an iterative method that involves STFT and the inverse STFT within each iteration, which requires much computational time compared to the other methods. For the manifold optimization (ManOpt), the algorithm in [37] was iterated 300 times. For the iterative refinement and MISI, the number of iterations was set to 10 or 50. In addition, to investigate the effect of the weights, the proposed method without iterative refinement or weights (NoWeight) performed comparable to the other methods. Therefore, for small window-shifting step, the computational cost of the proposed method can be further reduced by setting all the weights to 1 (this simplification eliminates 12 multiplications, 5 additions, and 3 divisions from the computation listed at the end of Sect. 3).

Since PU itself cannot perform well, it requires the iterative refinement which tends to be computationally expensive than the proposed method without the iterative refinement. By comparing PUIter10 (or PUIter50) with Prop (or NoWeight), they tended to perform similar for the 3/4- and 7/8-overlap conditions. In such case, the proposed method without the iterative refinement is advantageous in terms of computational efficiency because of its non-iterative nature provided by the proposed analytic computation. For the 1/2-overlap condition, the advantage of computational efficiency is diminished because performance of the proposed method without the iterative refinement decreased in that case. Even so, PropIter10 was able to perform similar to PUIter50 (as also confirmed by Fig. 2), which indicates the computational efficiency of the proposed method. Note that, even though ManOpt performed well, it is computationally far expensive compared to the proposed method (see the algorithm in [37] that requires several hundreds iterations whose computation per iteration is expensive than the proposed method). Overall, the proposed method is balanced in terms of performance and computational efficiency, which confirms its suitability as a preprocessing method.

6. CONCLUSIONS

In this study, a phase-recovery method for HPSS was proposed. It uses the observed phase information to improve the quality of the recovered phase. By formulating an optimization problem and providing its analytic solution, the proposed method can be computed efficiently as indicated by the easily counted number of computations per time-frequency bin. We experimentally confirmed that the proposed method performed well especially when the window-shifting step was small. However, when the
window-shifting step is not sufficiently small, the proposed method requires the iterative refinement, which is not very efficient. Therefore, some modification of the proposed method for the larger window shifting case is desired, which is left as a future work. In this work, a basic proposal and investigation were presented. Some real-world applications as well as a fast parallel implementation of the proposed method should also be considered as the future works.

### REFERENCES

[1] E. Tsunoo, N. Ono and S. Sagayama, “Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals,” *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 185–188 (2009).

[2] J. Reed, Y. Ueda, S. Siniscalchi, Y. Uchiyama, S. Sagayama and C. Lee, “Minimum classification error training to improve isolated chord recognition,” *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, pp. 609–614 (2009).

[3] A. Gkiokas, V. Katsouros, G. Carayannis and T. Stajilakis,
“Music tempo estimation and beat tracking by applying source separation and metrical relations,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 421–424 (2012).

[4] J. Driedger, M. Müller and S. Ewert, “Improving time-scale modification of music signals using harmonic-percussive separation,” IEEE Signal Process. Lett., 21, 105–109 (2014).

[5] 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., 24, 1535–1547 (2016).

[6] S. Sawada, S. Fukayama, M. Goto and K. Hirata, “Drums: A drum pattern transfer system preserving global pattern structure,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 391–395 (2019).

[7] B. Sharma, C. Gupta, H. Li and Y. Wang, “Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 396–400 (2019).

[8] F. Arabnezhad and B. Nasersharif, “Acoustic scene classification using binaural representation and classifier combination,” Proc. Conf. Comput. Knowl. Eng. (ICCKE), pp. 351–355 (2019).

[9] P. Mowlae and R. Saedie, “On phase importance in parameter estimation in single-channel speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 7462–7466 (2013).

[10] D. Wang and J. Lim, “The unimportance of phase in speech enhancement,” IEEE Trans. Acoust. Speech Signal Process., 30, 679–681 (1982).

[11] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” IEEE Signal Process. Mag., 32(2), pp. 55–66 (2015).

[12] P. Mowlae and R. Saedie, “Iterative closed-loop phase-aware single-channel speech enhancement,” IEEE Signal Process. Lett., 20, 1235–1239 (2013).

[13] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton and R. A. Saurous, “Differentiable consistency constraints for improved deep speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 900–904 (2019).

[14] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, “Deep Griffin–Lim iteration,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 61–65 (2019).

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” Proc. Conf. Int. Conf. Speech Commun. Assoc. (INTERSPEECH), pp. 4006–4010 (2017).

[16] S. Takaki, H. Kameoka and J. Yamagishi, “Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis,” Proc. Conf. Int. Conf. Speech Commun. Assoc. (INTERSPEECH), pp. 1128–1132 (2017).

[17] Y. Masuyama, K. Yatabe and Y. Oikawa, “Griffin–Lim like phase recovery via alternating direction method of multipliers,” IEEE Signal Process. Lett., 26, 184–188 (2019).

[18] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, “Phase reconstruction based on recurrent phase unwrapping with deep neural networks,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 826–830 (2020), doi: 10.1109/ICASSP40776.2020.9053234.

[19] D. Gunawan and D. Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures,” IEEE Signal Process. Lett., 17, 421–424 (2010).

[20] P. Mowlae, R. Saedie and R. Martin, “Phase estimation for signal reconstruction in single-channel speech separation,” Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), pp. 1548–1551 (2012).

[21] J. Le Roux and E. Vincent, “Consistent Wiener filtering for audio source separation,” IEEE Signal Process. Lett., 20, 217–220 (2013).

[22] Y. Masuyama, K. Yatabe and Y. Oikawa, “Phase-aware harmonic/percussive source separation via convex optimization,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 985–989 (2019).

[23] K. Drossos, P. Magron, S. I. Mimilakis and T. Virtanen, “Harmonic-percussive source separation with deep neural networks and phase recovery,” Proc. Eur. Signal Process. Conf. (EUSIPCO), pp. 421–425 (2018).

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

[25] P. Magron, R. Badeau and B. David, “Model-based STFT phase recovery for audio source separation,” IEEE/ACM Trans. Audio Speech Lang. Process., 26, 1095–1105 (2018).

[26] G. T. Beauregard, M. Harish and L. Wyse, “Single pass spectrogram inversion,” Proc. IEEE Int. Conf. Digit. Signal Process. (DSP), pp. 427–431 (2015).

[27] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” IEEE/ACM Trans. Audio Speech Lang. Process., 22, 1931–1940 (2014).

[28] J. Kulmer and P. Mowlae, “Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 5063–5067 (2015).

[29] M. Pirott, J. Stahl, P. Mowlae, V. I. Vorobiov, S. Y. Barysenka and A. G. Davydov, “Phase estimation in single-channel speech enhancement using phase invariance constraints,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 5585–5589 (2017).

[30] Y. Masuyama, K. Yatabe and Y. Oikawa, “Low-rankness of complex-valued spectrogram and its application to phase-aware audio processing,” IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 855–859 (2019).

[31] K. Yatabe and Y. Oikawa, “Phase corrected total variation for audio signals,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 656–660 (2018).

[32] M. Abe and J. O. Smith, “Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks,” Audio Eng. Soc. Conv., pp. 28–31 (2004).

[33] Y. Katabe, Y. Masuyama, T. Kusano and Y. Oikawa, “Representation of complex spectrogram via phase conversion,” Acoust. Sci. & Tech., 40, 170–177 (2019).

[34] S. Fenet, R. Badeau and G. Richard, “Reassigned time-frequency representations of discrete time signals and application to the constant-Q transform,” Signal Process., 132, 170–176 (2017).

[35] D. J. Nelson, “Cross-spectral methods for processing speech,” J. Acoust. Soc. Am., 110, 2575–2592 (2001).

[36] S. A. Fulop and K. Fitz, “Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications,” J. Acoust. Soc. Am., 119, 360–371 (2006).

[37] Y. Masuyama, K. Yatabe and Y. Oikawa, “Model-based phase recovery of spectrograms via optimization on Riemannian manifolds,” Proc. Int. Workshop Acoust. Signal Enhanc. (IWAENC), pp. 126–130 (2018).

[38] A. Litjhus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N.
Ito, N. Ono and J. Fontecave, “The 2016 signal separation evaluation campaign,” \textit{Latent Var. Anal. Signal Sep.}, pp. 323–332 (2017).

[39] P. L. Søndergaard, B. Torrésani and P. Balazs, “The linear time frequency analysis toolbox,” \textit{Int. J. Wavelets Multi-resolut. Inf. Process.}, 10(4), 1250032 (2012).

[40] E. Vincent, R. Gribonval and C. Evotte, “Performance measurement in blind audio source separation,” \textit{IEEE Trans. Audio Speech Lang. Process.}, 14, 1462–1469 (2006).

[41] D. Fitzgerald, “Harmonic/percussive separation using median filtering,” \textit{Proc. Int. Conf. Digit. Audio Eff. (DAFx-10)}, pp. 6–10 (2010).

[42] C. Dittmar, P. López-Serrano and M. Müller, “Unifying local and global methods for harmonic-percussive source separation,” \textit{Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)}, pp. 176–180 (2018).

Kenji Kobayashi received the B.E., and M.E., degrees from Waseda University in 2018 and 2020, respectively. He currently works for Casio Computer Co., Ltd. as an engineer.

Yoshiki Masuyama received his B.E. and M.E. degrees from Waseda University in 2019 and 2021, respectively. He is currently pursuing a Ph.D. degree in the Graduate School of Systems Design, Tokyo Metropolitan University.

Kohei Yatabe received the B.E., M.E., and Ph.D. degrees from Waseda University in 2012, 2014, and 2017, respectively. He is currently an Assistant Professor of the Department of Intermedia Art and Science, Waseda University.

Yasuhiro Oikawa received his B.E., M.E., and Ph.D. degrees in Electrical Engineering from Waseda University in 1995, 1997, and 2001, respectively. He is a professor of the Department of Intermedia Art and Science, Waseda University. His main research interests are communication acoustics and digital signal processing of acoustic signals. He is a member of ASJ, ASA, IEICE, IPSJ, VRSJ, and AIJ.