INVITED REVIEW

Speech enhancement using harmonic-structure-based phase reconstruction

Yukoh Wakabayashi^{1,2,*}¹*Tokyo Metropolitan University, 6-6, Asahigaoka, Hino, 191-0065 Japan*²*Ritsumeikan University, 1-1-1, Noji-Higashi, Kusatsu, 525-8577 Japan*

Abstract: Recent work has shown that phase information is useful for further improving the performance of speech enhancement, source separation, and speech synthesis. In the speech enhancement field, the combination of amplitude and phase estimations improves the perceived quality more than only amplitude estimation. In this paper, we review two harmonic-structure-based phase estimation methods with temporal and frequency constraints on the harmonic speech phase. In addition, we describe important parameters for phase estimation, such as the frame shift length and window function of the short-time Fourier transform. Subjective experiments using listening tests and future work for phase processing are briefly described.

Keywords: Phase-aware signal processing, Phase reconstruction, Speech enhancement, Harmonic structure, Time-frequency constraint

PACS number: 43.60.Uv [doi:10.1250/ast.40.162]

1. INTRODUCTION

In general, acoustic and speech signal processing has not handle phase spectrum because it was regarded as an additional information for constructing a time-domain signal. In the acoustic and speech signal processing field, speech enhancement and noise reduction have been studied for a long time. In these processes, it is important to reconstruct the perceived quality of speech which is degraded by ambient noise. It has been reported that phase reconstruction in the time-frequency domain brings less improvement to speech quality than amplitude reconstruction, and auditory perception is insensitive to the phase change [1]. Since these reports, amplitude estimation has been the main target in acoustic and speech signal processing. In addition, the phase does not have straightforward features, while the amplitude spectrum has a harmonic structure and its energy concentrates at low frequencies. In statistical analysis, the amplitude is modeled using the Rayleigh or Gamma distributions, but the phase is uniformly distributed, which complicates its modeling. Another reason for neglecting phase processing includes the statistical analysis by Ephraim *et al.* and Lotter *et al.* [2,3], who respectively showed the unprocessed phase is the optimal estimate in terms of the minimum mean square error (MMSE) and the maximum a posteriori (MAP) criteria under the assumption of an uniformly distributed phase. Even speech enhancement methods

without phase processing, such as spectral subtraction [4], can improve the perceived quality sufficiently.

The Griffin and Lim algorithm (GLA) [5] may be the most well-known method among the few phase reconstruction methods available. The GLA calculates phases from amplitudes by iteratively performing the short-time Fourier transform (STFT) and inverse STFT (ISTFT). The iterative calculation explores consistent phases of given amplitudes using the redundancy of the STFT with the overlapped frames; that is, a sample of the time-domain signal is included in multiple frames. The amplitude-required phase estimation accurately reconstructs the signal when the given amplitude is a desired one, while it has been reported that this phase estimation using estimated amplitudes does not improve the perceived quality [6]. As a result, speech enhancement with the GLA has not been efficient.

However, recent research shows the importance of phase. Paliwal [7] showed that introducing different window functions for analyzing amplitudes and phases improves an objective measure, the perceptual evaluation of speech quality (PESQ) [8]. Kazama *et al.* [9] have showed that the perceived quality depends on the speech phase and the frame length through listening tests. The dependence of the perceived quality on the phase observed in these previous studies has motivated further research on phase-aware processing, namely, source separation with phase information and phase estimation for speech enhancement. Complex nonnegative matrix factorization (NMF), which is NMF augmented to the complex-valued

*e-mail: wakayuko@tmu.ac.jp, wakayuko@fc.ritsumei.ac.jp

domain [10] is a well-known example. Advances in deep learning have accelerated phase reconstruction for speech synthesis with deep neural networks (DNNs) [11] and the time-frequency mask estimation using phase information with DNNs [12]. A textbook on phase processing was also published recently [13].

As described in survey papers [14,15], phase estimation for single-channel speech enhancement includes an estimation method based on the harmonic structure of speech [6,16–19]. This approach does not handle phases of all the STFT time-frequency block at the same time but separates them into harmonic and non-harmonic components and estimates each phase with different equations: firstly harmonic phases are estimated, and then non-harmonic phases are estimated using the harmonic phases. Mowlae *et al.* [6,16] formulated a harmonic phase decomposition model which consists of a linear phase component and an unwrapped component. They found that the variance of the unwrapped phase component of speech is small and proposed a harmonic phase estimation method which reduces the variance of noisy unwrapped phase component. Gerkmann *et al.* [17] modeled speech as the simple summation of multiple sinusoidal signals and formulated the temporal behavior of the harmonic phase and the effect of the harmonic phase on the non-harmonic phase in the time-frequency domain. In our previous work [19], we estimated speech phase using the time-frequency relationship between harmonic phases of speech, the so-called phase distortion feature. In this paper, we describe harmonic-structure-based phase estimation, specifically by explaining the following two methods in detail: the STFT phase improvement method by Gerkmann [17] and the phase distortion averaging method [19].

2. FORMULATION

In this paper, we assume that a speech signal is corrupted by an additive noise.

$$x(t) = s(t) + d(t), \quad (1)$$

where $x(t)$ is the observed signal with a time sample index t , $s(t)$ is the speech sample, and $d(t)$ is the additive noise. Then, the noisy speech spectrum at the n th frame is represented using the STFT with the window function $w(t)$, the frame length N , and the hop size L as follows:

$$X_{n,k} = \sum_{\tau=0}^{N-1} w(\tau)x(\tau + Ln)(F_N)^{k\tau}, \quad (2)$$

where k is the frequency bin index and $F_N = \exp(-j2\pi/N)$ with $j = \sqrt{-1}$. When we let the phase of $X_{n,k}$ be $\phi_{n,k}$, the following equation is obtained from Eq. (1) and the linearity of the STFT:

$$|X_{n,k}|e^{j\phi_{n,k}} = |S_{n,k}|e^{j\psi_{n,k}} + |D_{n,k}|e^{j\zeta_{n,k}}, \quad (3)$$

Fig. 1 Flowchart of phase-aware speech enhancement.

where $S_{n,k}$ and $D_{n,k}$ are the spectra of $s(t)$ and $d(t)$, respectively, and $\psi_{n,k}$ and $\zeta_{n,k}$ are the phase spectra of these signals, respectively. $\hat{A}$ represents the estimation of symbol A . The goal of speech enhancement is estimating the complex-valued spectrum of speech which consists of the amplitude estimate $|\hat{S}_{n,k}|$ and the phase estimate $\hat{\psi}_{n,k}$, given a noisy amplitude $|X_{n,k}|$ and noisy phase $\phi_{n,k}$, as in Fig. 1. The time-domain speech estimate $\hat{s}(t)$ is reconstructed using the ISTFT and the overlap add method as follows:

$$\hat{s}(t) = \sum_n \tilde{w}(t - nL)\tilde{s}_n(t - nL), \quad (4)$$

$$\tilde{s}_n(\tau) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{S}_{n,k}(F_N)^{-k\tau}, \quad (5)$$

$$\hat{S}_{n,k} = |\hat{S}_{n,k}|e^{j\hat{\psi}_{n,k}}, \quad (6)$$

where $\tilde{w}(\tau)$ is the synthesis window function.

In addition, we define the symbols for the harmonic component. When $f_{0,n}$ [Hz] and $f_{h,n} = (h+1)f_{0,n}$ are the fundamental frequency and the h th harmonic frequency at the n th frame, respectively, a frequency bin index with respect to $f_{h,n}$ is defined as

$$k_h = \underset{k}{\operatorname{argmin}} |k - \kappa_h|, \quad (7)$$

$$\kappa_h = \frac{f_{h,n}}{F_s} K, \quad (8)$$

where F_s is the sampling frequency, and K is the discrete Fourier transform length, which is set as $K = N$ generally. As shown in Fig. 2, κ_h is a non-integral value in the frequency bin scale.

3. HARMONIC-STRUCTURE-BASED PHASE RECONSTRUCTION

The harmonic-structure-based method analyzes the harmonic component of speech, compared with the GLA,

Fig. 2 Harmonic component indexes in the continuous and discrete representations.

Fig. 3 Conceptual diagram of harmonic-structure-based phase reconstruction.

which uses the redundancy of the STFT and DNN-based methods. This approach is consolidated into the following three points:

- 1) Speech is modeled as the summation of multiple sinusoidal signals,
- 2) Harmonic phases rotate under a constraint, and
- 3) Non-harmonic phases are affected by the nearest harmonic phase.

The sinusoidal model is formulated as

$$s(t) = \sum_{h=0}^{H-1} A_h(t) \cos\left(\int_0^t \omega_h(t') dt' + \psi_{h,t}^0\right), \quad (9)$$

where $A_h(t)$ is the harmonic amplitude, $\omega_h(t) = 2\pi f_h(t)/F_s$ is the normalized angular frequency, and $\psi_{h,t}^0$ is the initial phase. The harmonic-structure-based method models speech as a signal which consists of sinusoids with a fundamental frequency and its harmonics. This model simplifies the phase estimation problem by separating the harmonic and non-harmonic phase estimations. Figure 3 illustrates the concept of this approach. This diagram shows that the phase at the h th harmonic component k_h , as illustrated by red lines, is firstly estimated, and then the non-harmonic phases are estimated using the harmonic phase. In Sect. 3.1, we explain a method that considers the temporal rotation of the harmonic phase. Section 3.2 describes a method that considers the relationship between different harmonic phases, as illustrated by the yellow arrows in Fig. 3.

3.1. Method Using a Temporal Constraint between Harmonic Phases

The STFT phase improvement (STFTPI) method proposed by Gerkmann *et al.* [17] estimates the speech phase with a temporal constraint between harmonic phases. The sinusoidal model shown in Eq. (9) is represented in the time-frequency domain with the STFT as follows:

$$S_{n,k} = \sum_{\tau=0}^{N-1} w(\tau) \sum_{h=0}^{H-1} A_{k_h} \cos(\omega_h \cdot (nL + \tau) + \psi_{h,n}^0) (F_N)^{k\tau}, \quad (10)$$

where it is supposed that the harmonic frequency is constant in an analysis frame and thereby the integration term of Eq. (9) becomes $\omega_h \cdot t$. Here, it is additionally supposed that the spectral leakage in a harmonic component from other harmonic components is negligibly small. Then, the spectrum at k_h in Eq. (10) is approximated as follows:

$$\begin{aligned} S_{n,k_h} &\approx A_{k_h} \sum_{\tau=0}^{N-1} w(\tau) e^{j(\omega_h(nL+\tau) + \psi_{h,n}^0)} (F_N)^{k\tau} \\ &= A_{k_h} e^{j(\omega_h nL + \psi_{h,n}^0)} \sum_{\tau=0}^{N-1} w(\tau) e^{-j2\pi \frac{k_h - \kappa_h}{N} \tau} \\ &= A_{k_h} e^{j(2\pi \frac{f_{h,n}}{F_s} nL + \psi_{h,n}^0)} W_{k_h - \kappa_h}, \end{aligned} \quad (11)$$

where W is the spectrum of the window function. The phase term in Eq. (11) is formulated as

$$\psi_{n,k_h} = 2\pi \frac{f_{h,n}}{F_s} nL + \psi_{h,n}^0 + \phi_{k_h - \kappa_h}^W, \quad (12)$$

where ϕ^W is the phase property of the window function. Equation (12) and the assumption that the initial phase $\psi_{h,n}^0$ is not invariant at each frame, in other words, $\psi_{h,n}^0 = \psi_{h,n-1}^0$, lead to the following algorithms of phase estimation:

$$\hat{\psi}_{n,k_h} = \hat{\psi}_{n-1,k_h} + 2\pi \frac{f_{h,n}}{F_s} L, \quad (13)$$

$$\hat{\psi}_{n,k_h+\delta} = \hat{\psi}_{n,k_h} - \phi_{k_h - \kappa_h}^W + \phi_{k_h - \kappa_h + \delta}^W, \quad (14)$$

where $\delta \in [-\kappa_0/2, \kappa_0/2]$ is the integer. Equation (13) shows that the harmonic frequency $f_{h,n}$ and the frame shift L resolve the harmonic phase, and the non-harmonic phase is calculated from the harmonic phase through the phase property of the window ϕ^W , as shown in Fig. 4. The authors of [17] used the noisy harmonic phase ϕ_{n_0,k_h} as the initial phase along the time frame in Eq. (13), $\hat{\psi}_{n_0,k_h}$. Note that ϕ_{Ω}^W is not calculated from the DFT but rather the discrete time Fourier transform (DTFT) because Ω is not an integer. Let the M th-order cosine window be $w(t)$, which is involved in the Hann, Hamming, and Blackman windows, for calculating the phase property of the window. Then, $w(t)$ is defined as

$$w(t) = \sum_{m=0}^{M-1} a_m \cos(2\pi t m / N), \quad (15)$$

where a_m is the m th coefficient and N is the window length. In the case of the Blackman window, $M = 3$, $a_0 = 0.42$, $a_1 = -0.5$, and $a_2 = 0.08$. The DTFT of the cosine window is represented as

Fig. 4 Temporal relationship between harmonic phases in the STFT representation.

$W(\Omega)$

$$\begin{aligned}
 &= \sum_{t=-\infty}^{\infty} w(t) \exp(-jt\Omega) \\
 &= \sum_{m=0}^{M-1} \left[\frac{a_m}{2} \exp\left(-j\frac{N-1}{2}\Omega\right) \sin\left(\frac{N}{2}\Omega\right) \right. \\
 &\quad \cdot \left. \left\{ \frac{\exp\left(-j\frac{\pi}{N}m\right)}{\sin\left(\frac{1}{2}\left(\Omega - \frac{2\pi}{N}m\right)\right)} + \frac{\exp\left(j\frac{\pi}{N}m\right)}{\sin\left(\frac{1}{2}\left(\Omega + \frac{2\pi}{N}m\right)\right)} \right\} \right]. \quad (16)
 \end{aligned}$$

This representation is used to calculate the phase property of the window as

$$\phi_{\Omega}^W = L W\left(2\pi \cdot \frac{\Omega}{N}\right). \quad (17)$$

The STFTPI method estimates the speech phase such that it follows the sinusoidal model, which generates a buzzy and unnatural enhanced speech which contains beep sounds at high frequencies. As a result, it degrades the perceived quality.

3.2. Method with Time-frequency Constraint on Harmonic Phase

This section explains our previous method based on phase distortion averaging [19]. This method considers not only a temporal constraint on the harmonic phase but

also a constraint between harmonics in comparison with the STFTPI method. Figure 5 illustrates the concept of the constraint, namely, rotations of harmonic phases in a time frame. Firstly, it is supposed that the phase in the fundamental frequency component, ψ_{n,k_0} , rotates θ radians in the time frame Δ . In this case, the h th harmonic phase ψ_{n,k_h} rotates $(h+1)\theta$ radians. They are formulated as

$$\psi_{n+\Delta,k_0} = \psi_{n,k_0} + \theta, \quad (18)$$

$$\psi_{n+\Delta,k_h} = \psi_{n,k_h} + (h+1)\theta. \quad (19)$$

Then, the following phase differences are defined:

$$\Phi_{n,h} = \psi_{n,k_h} - \psi_{n,k_{h-1}} - \psi_{n,k_0}. \quad (20)$$

The temporal fluctuation of the difference $\Phi_{n,h}$ in Δ is shown as

$$\begin{aligned}
 \Phi_{n+\Delta,h} &= \{\psi_{n,k_h} + (h+1)\theta\} \\
 &\quad - \{\psi_{n,k_{h-1}} + h\theta\} - \{\psi_{n,k_0} + \theta\} \\
 &= \psi_{n,k_h} - \psi_{n,k_{h-1}} - \psi_{n,k_0} \\
 &= \Phi_{n,h}. \quad (21)
 \end{aligned}$$

These equations show that $\Phi_{n,h}$ is time-independent. That is to say, harmonic phases of a speech over the voiced intervals rotate with the constraint that $\Phi_{n,h}$ is constant temporally. This difference is called “phase distortion (PD),” and it is discussed in [20]. The PD has relationships between both the harmonics and temporal fluctuations. Note that the application of the PD’s temporal constancy to speech enhancement requires determining a constant value for the PD. In other words, harmonic-wise parameters must be determined over a voiced interval. It is difficult to determine these values because initial phases depend on them. Our previous method assumed that noisy harmonic phases are close to speech harmonic phases because the concentration of speech energy in harmonic components brings a high signal-to-noise ratio (SNR) to these components. By using this assumption, the method estimates harmonic phases with temporally smoothed PD features of noisy phases. The algorithm for phase estimation is as follows:

Fig. 5 Rotation of harmonic phases in the complex coordinate system.

$$\hat{\psi}_{n,k_h} = \tilde{\Phi}_{n,h} + \phi_{n,k_{h-1}} + \phi_{n,k_0}, \quad (22)$$

$$\tilde{\Phi}_{n,h} = \angle \exp \left(j \sum_{n'=n-\Psi}^{n+\Psi} \Phi_{n',h} \right), \quad (23)$$

$$\Phi_{n,h} = \phi_{n,k_h} - \phi_{n,k_{h-1}} - \phi_{n,k_0}, \quad (24)$$

where Ψ is the parameter for temporal smoothing. Equations (23) and (24) temporally smooth the noisy PD. Equation (22) shows that the replacement of the noisy PD with the smoothed PD estimates a speech harmonic phase that is similar to the noisy phase and that it smoothly fluctuates in the PD domain. Here, it is supposed that the SNR in the fundamental frequency component k_0 is high as follows:

$$\psi_{n,k_0} = \phi_{n,k_0}. \quad (25)$$

Non-harmonic phases are estimated with window phase compensation, as described in Sect. 3.1, as follows:

$$\hat{\psi}_{n,k_h+\delta} = \hat{\psi}_{n,k_h} - \phi_{k_h-\kappa_h}^W + \phi_{k_h-\kappa_h+\delta}^W. \quad (26)$$

Thus, the difference between this method and the STFTPI method is harmonic phase estimation.

4. EVALUATION OF PHASE RECONSTRUCTION

This section explains the perceived quality measure for evaluating phase estimation and briefly shows the experimental results and the dependence of it on the analysis frame shift and window functions, which play an important role in conducting evaluations.

4.1. Perceived Quality Measure

PESQ [8] is a frequently used measure for speech enhancement with phase reconstruction. This objective measure was originally used for evaluation of band-limited speech (0.3–3.4 kHz) in telecommunications. To handle a wider speech band up to 7 kHz, an augmented measure, wideband PESQ [21], has also been invented. Two phase reconstructions described in Sect. 3 consider only harmonics up to 4 kHz, that is, the number of harmonics $H = \lfloor 4,000/f_{0,m} \rfloor$, and therefore it is valid to use the PESQ to evaluate them. In addition, short-time objective intelligibility (STOI) [22] and unwrapped phase RMSE (UnRMSE) [13] have been used for evaluating quality. The STOI is difficult to discern during the performance of estimation accuracy unless the estimation accuracy is greatly improved. UnRMSE is defined as the squared error of unwrapped harmonic phases which are weighed by their amplitudes. It has been reported that this measure is correlated with subjective scores of speech quality [23]. To obtain reliable evaluation results, it is desirable to conduct subjective evaluations through listening tests.

Fig. 6 PESQ improvements relative to noisy speech as a function of frame shift, with a sampling frequency of 16 kHz and frame length of 512 samples (32 ms).

4.2. Frame Shift Length

General acoustic and speech signal processing employs 1/2 of the frame length as a frame shift length. In comparison to this approach, recent research on phase processing often uses a frame shift 1/4 or 1/8 of the frame length. The use of a small frame shift is equivalent to increasing the redundancy of the time-frequency representation along the time axis, which results in more detailed analysis of temporal phase fluctuations. The aforementioned harmonic-structure-based methods use harmonic phase fluctuations along with time as an important key to phase reconstruction. Then, the use of 1/2 of the frame length degrades the estimation accuracy due to the difficulty of identifying this key. The GLA with a small frame shift also improves phase estimation accuracy more than GLA with a large shift. Figure 6 illustrates average PESQ improvements in two noisy environments as a function of the frame shift by our previous method described in Sect. 3.2. As shown, the use of the 1/4 and 1/8-frame shifts improves the PESQ by more than one 1/2-frame shift.

4.3. Window Function

It has been reported that it is useful to employ the mismatched window approach, which selects two different windows for amplitude and phase analyses [7]. Also, it is desirable that the window function for phase analysis has a high dynamic range to satisfy the assumption that a harmonic phase does not affect other harmonic phases, as described in Sect. 3.1. Thus, the Blackman window has been preferred to the Hann and Hamming windows. In [6], they evaluated the difference between the use of the Blackman window and other windows. The experiments described in the next section employed the mismatched window approach, which uses the Hamming and Blackman windows for amplitude and phase analysis, respectively. In this case, however, it is impossible to use the optimal synthesis window [5] for perfect reconstruction in Eq. (4) in the synthesis step; thus, we used the Blackman window

Fig. 7 Subjective scores of perceived quality obtained by listening tests.

as the synthesis window in [19]. In the preliminary experiments that we conducted, the use of the Blackman window for phase analysis improved PESQ overall, while its use for amplitude analysis conversely degraded PESQ due to the low-frequency resolution.

4.4. Evaluations of Perceived Quality

This section introduces the subjective scores that we obtained through listening tests. Figure 7 shows the average subjective score assigned by 12 subjects, who were students in their 20s with normal hearing. In this experiment, the amplitude was estimated with the MMSE gain estimator [2], and the phase was estimated with the STFTPI method (Sect. 3.1) and the PD averaging method (Sect. 3.2). The fundamental frequency $f_{n,0}$ was estimated by the pitch estimation filter with amplitude compression (PEFAC) [24]. The unprocessed phase case and the oracle phase case were also used. All subjects wore headphones and evaluated the total speech quality using a five-level quality scale based on paired comparison. Each subject listened to 120 paired sources, a reference source and an object source presented alternately, and evaluated the perceived quality of the object source in comparison to the reference source on a scale from -2 to $+2$. A score of $+2$ indicated that the perceived quality of the object source was much better than that of the reference source, whereas a score of -2 indicated the opposite. A score of 0 means that their quality was identical.

As shown, our previous method of PD averaging largely improved the perceived quality more than the unprocessed phase case and the STFTPI method. The results of an analysis of variance (ANOVA) by F-test show that these methods are statistically significant. Some subjects believed that the proposed method improved auditory quality more than the other methods, and their opinions are reflected in the subjective scores. The spectrograms illustrated in Fig. 8 support the validity of the opinions. As shown, the harmonic structures in (c) and (d) are seen more clearly than that in (b). The harmonics at

Fig. 8 Spectrograms of noisy speech, three enhanced speech signals, and clean speech.

the higher frequencies of 2–4 kHz in the 0.3 to 0.9 s interval of (c) are enhanced excessively in comparison with the clean speech in (e). This excessive enhancement degraded the perceived speech quality.

5. SUMMARY

In this paper, we explained two harmonic-structure-based phase estimations that consider the temporal and frequency constraints of harmonic phases in the STFT domain. We briefly summarized subjective experimental results through listening tests.

Future work will involve considering a complementary estimation method for phase and amplitude. The methods

in this paper estimate both spectra independently. However, these two spectra are known to have a tight relationship, demonstrated, for example, by the synchronization of power spectra and group delay spectra. This relationship implies that the two spectra can be estimated in a complementary way. To invent a complementary method, it is necessary to analyze the fluctuations of group delay and instantaneous frequency, which are both features of phase rotation. It is expected that handling complex-valued spectra will become a mainstream method for acoustic and speech signal processing. For example, some complex-valued spectral estimators [25,26] have already been proposed. Considering phase information is essential to further develop acoustic and speech signal processing. It is hoped that this paper will be of some benefit for this development.

ACKNOWLEDGEMENT

This work was partly supported by the SECOM Science and Technology Foundation and the JSPS KAKENHI Grant Number JP18H06482.

REFERENCES

- [1] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, **67**, 1586–1604 (1979).
- [2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, **32**, 1109–1121 (1984).
- [3] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, **2005**, 1110–1126 (2005).
- [4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, **27**, 113–120 (1979).
- [5] D. W. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, **32**, 236–243 (1984).
- [6] P. Mowlae and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23**, 1283–1294 (2015).
- [7] K. Paliwal, K. Wójcicki and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, **53**, 465–494 (2011).
- [8] ITU-T, "Perceptual evaluation of speech quality (PESQ)," *ITU-T Rec. P. 862* (2001).
- [9] M. Kazama, S. Gotoh, M. Tohyama and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility," *J. Acoust. Soc. Am.*, **127**, 1432–1439 (2010).
- [10] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **21**, 971–982 (2013).
- [11] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, pp. 286–290 (2018).
- [12] G. Wichern and J. L. Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, pp. 396–400 (2018).
- [13] P. Mowlae, J. Kulmer, J. Stahl and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice* (John Wiley & Sons, Chichester, 2017).
- [14] T. Gerkmann, M. Krawczyk and J. Le Roux, "Phase processing for single-channel speech enhancement," *IEEE Signal Process. Mag.*, **32**, 55–66 (2015).
- [15] P. Mowlae, R. Saeidi and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Commun.*, **81**, 1–29 (2016).
- [16] P. Mowlae and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23**, 1521–1532 (2015).
- [17] 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).
- [18] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, "Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5560–5565 (2017).
- [19] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **26**, 1559–1569 (2018).
- [20] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP J. Audio Speech Music Process.*, **2014**, 38, 1–16 (2014).
- [21] ITU-T, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *ITU-T Rec. P. 862.2* (2005).
- [22] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **19**, 2125–2136 (2011).
- [23] A. Gaich and P. Mowlae, "On speech intelligibility estimation of phase-aware single-channel speech enhancement," *Proc. Interspeech*, pp. 2553–2557 (2015).
- [24] S. Gonzalez and M. Brookes, "PEFAC — A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 518–530 (2014).
- [25] P. Mowlae, J. Stahl and J. Kulmer, "Iterative joint MAP single-channel speech enhancement given non-uniform phase prior," *Speech Commun.*, **86**, 85–96 (2017). [Online]. Available: <https://doi.org/10.1016/j.specom.2016.11.008>
- [26] M. Krawczyk and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24**, 2251–2262 (2016).

Yukoh Wakabayashi received B.E. and M.E. degrees from Osaka University, Osaka, Japan, in 2008 and 2010, respectively, and a Ph.D. degree from Ritsumeikan University, Kusatsu, Japan, in 2017. He joined Rohm, Inc., Kyoto, Japan, in 2010, and was an Assistant Researcher with Kyoto University from 2012 to 2014. He was a recipient of the JSPS Research Fellowship

for Young Scientists DC2 from 2016 to 2017. He is currently an Assistant Professor with the Faculty of Systems Design, Tokyo Metropolitan University, Tokyo, Japan, and an affiliate Assistant Professor with Ritsumeikan University. His research interests include acoustic signal processing, array signal processing, and speaker diarization. He is a member of the Institute of Electronics, Information and Communication Engineers and Acoustical Society of Japan.