Complex-Domain Pitch Estimation Algorithm for Narrowband Speech Signals

Yuya Hosoda, *Member, IEEE*, Arata Kawamura, *Member, IEEE*, and Youji Iiguni, *Member, IEEE*

***Abstract*—We propose a complex-domain pitch estimation algorithm for narrowband speech signals, which utilizes a complex spectrum containing both amplitude and phase spectrum information. Traditional frequency-domain pitch estimation algorithms assume that a speech signal has a harmonic structure; they estimate a pitch by calculating the distance between the adjacent peaks of the amplitude spectrum corresponding to harmonics. However, only a few peaks can be detected from narrowband speech signals because of their limited bandwidth, resulting in pitch estimation errors. In this article, phase differences between harmonics are utilized as an additional cue for pitch estimation. The phase difference between harmonics refers to the two-step phase difference between successive analysis frames and between the lowest-order harmonic and other harmonics, which is theoretically derived using the pitch. When the phase spectrum for the higher-order harmonic is shifted by the theoretical value, it agrees with the phase spectrum for the lowest-order harmonic. Therefore, for each pitch candidate, the proposed method calculates the shifted phase spectra and combines them with the amplitude spectrum to generate complex spectra. When a pitch candidate is correct, the cumulative sum of the complex spectra is added in the same direction, which emphasizes harmonics even for narrowband speech signals. Results of the objective evaluation show that the proposed method accurately estimates the pitch from narrowband speech signals.**

***Index Terms*—Narrowband speech signal, phase difference, pitch estimation, sinusoidal model.**

I. Introduction

In A public switched telephone network, speech signals are encoded using speech coding techniques to enable communication with low latency. Narrowband speech coding techniques such as G.711 [1] and Adaptive Multi-Rate [2] limit the bandwidth of the decoded speech signal to a narrowband of 300–3400 Hz, whereas wideband speech coding techniques such as Enhanced Voice Services [3] limit it to a wideband of 50–14400 Hz. When understanding unknown words or names, bandwidth limitation reduces speech intelligibility on the phones, such as unvoiced or plosive utterances [4]. Consequently, narrowband speech provides poorer speech intelligibility of syllables than wideband speech. Old public switched telephone networks, where system renovation is not practical because of the effort and cost required, may only support narrowband speech coding techniques [5].

Artificial bandwidth extension (ABE) is a speech enhancement approach for narrowband speech signals, which reconstructs the missing lower spectrum (0–300 Hz) and upper spectrum (3400–7000 Hz) using the existing narrowband spectrum of 300–3400 Hz [6], [7], [8], [9], [10], [11], [12]. Because of the missing lower-order harmonics, the narrowband speech signal has a partial harmonic structure. ABE for the missing lower spectrum reconstructs the harmonic structure by generating multiple sinusoidal waves corresponding to the missing lower-order harmonics; this is referred to as sinusoidal synthesis [10], [11], [12]. A fundamental frequency (or pitch) is critical for sinusoidal synthesis to accurately reconstruct the harmonic structure. Whereas a fundamental frequency is a physical property of the speech signal, corresponding to the lowest-order harmonics, a pitch is the perceived property corresponding to the relative highness of the harmonic. In speech signal processing, fundamental frequency and pitch are considered to have the same meaning. In this article, the term 'pitch' is used. For narrowband speech coding techniques [1], a pitch must be estimated from the narrowband speech signal. Because artifacts such as human voices and car sounds may be mixed in the narrowband speech signal, pitch estimation algorithms should be robust to noise.

Researchers have worked on pitch estimation for a long time [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. Because speech signals have pseudo-periodicity with a short frame length, time-domain pitch estimation algorithms estimate a pitch using an autocorrelation function [13], [14], [15], [16]. ABE for the missing lower spectrum has adopted time-domain pitch estimation algorithms because of their low computational complexity [11], [12]. In addition, pitch estimation errors can be prevented by postprocessing, using pitch state transition tracks based on hidden Markov models [15] or the Ornstein–Uhlenbeck process [16]. However, time-domain pitch estimation algorithms may suffer from pitch estimation errors in noisy environments. Kim et al. employed a deep convolutional neural network to estimate the pitch, which directly operates on the time-domain signal [17]. Segal et al. introduced a pitch estimation algorithm that deployed an encoder and multiple decoders to represent signal processing filterbanks [18]. Although their robustness can be

Manuscript received 2 January 2022; revised 20 October 2022 and 13 May 2023; accepted 15 May 2023. Date of publication 22 May 2023; date of current version 26 May 2023. This work was supported by the Ono Charitable Trust for Acoustics. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hema A Murthy. *(Corresponding author: Yuya Hosoda.)*

Yuya Hosoda is with the Center for IT-based Education, Toyohashi University of Technology, Aichi 441-8122, Japan (e-mail: hosoda.yuya.ho@tut.jp).

Arata Kawamura is with the School of Faculty of Information Science and Engineering, Kyoto Sangyo University, Kyoto 603-8555, Japan (e-mail: kawamura@cc.kyoto-su.ac.jp).

Youji Iiguni is with the Graduate School of Engineering Science, Osaka University, Osaka 560-0043, Japan (e-mail: iiguni@sys.es.osaka-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3278488

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

enhanced by training predictive models from speech signals with noise, their applications are limited because retraining predictive models that is required for handling different sampling rates or frame lengths is time-consuming.

Parametric pitch estimation algorithms estimate a pitch using a harmonic model without training predictive models [19], [20], [21], [22], [23]. The harmonic model represents the speech signal as the sum of sinusoidal waves, with a harmonic structure, and noise. Most appropriate model parameters are calculated using nonlinear least squares (NLS) [19], [20]. In addition, a pitch tracking method that considers the temporal smoothness of the model parameters can reduce the pitch estimation error in noisy environments [21], [22]. However, because parametric pitch estimation algorithms have focused on speech signals with a harmonic structure, they have difficulty accurately estimating the model parameters from narrowband speech signals with a partial harmonic structure. A parametric pitch estimation algorithm for speech signals with a partial harmonic structure has also been devised [23], but iterative updates are required to estimate the model parameters. The calculation time must be reduced as much as possible because public switched telephone networks require real-time processing.

Frequency-domain pitch estimation algorithms estimate a pitch from the amplitude spectrum using the harmonic summation algorithm with a short processing time [24], [25], [26], [27], [28]. Since harmonics correspond to peaks on the amplitude spectrum, a pitch is defined as the distance between adjacent peaks. The harmonic summation algorithm calculates the distance using the cumulative sum of the amplitude spectra in frequency indices corresponding to harmonics [24], [25], [26]. The summation of residual harmonics (SRH) algorithm [25] suppresses the effects of vocal tract resonances and noise using the flat amplitude spectrum of the excitation signal. In addition, the pitch estimation error can be reduced using the Viterbi algorithm that considers the pitch transition from past analysis frames [27], [28]. However, only a few peaks can be detected from narrowband speech signals due to their limited bandwidth, and therefore overtones or subharmonics may be obtained; these are called 'octave error.' Octave errors are an inherent problem of pitch estimation algorithms, even for wideband speech signals, and they are observed especially in narrowband speech signals with missing lower harmonics. Thus, it is difficult to estimate the pitch from narrowband speech signals using only the amplitude spectrum.

We have focused on the phase spectrum as another cue for pitch estimation for narrowband speech signals. The phase spectrum has not received as much attention in speech signal processing as the amplitude spectrum. Recently, it has been reported that the phase spectrum is also closely related to harmonic structure, using the short-time Fourier transform (STFT) representation [35]. Phase-aware speech enhancement methods have been devised based on models that separate the phase spectrum into linear and unwrapped phases [36], [37] and approximate a speech signal using the sum of sinusoidal waves with a harmonic structure [43], [44]. Pitch estimation algorithms using the phase spectrum have also been devised, which utilize the phase difference between successive analysis frames [29], [30] and group delay [30], [31], [32], [33]. Note that harmonics are the key to pitch estimation, as known in frequency-domain pitch estimation algorithms. Our previous study [34] utilized the phase difference between harmonics, which is defined as the two-step phase difference between successive analysis frames and between the lowest-order harmonic and other harmonics. While group delay, which is useful for speech signal processing, differentiates the phase spectrum for frequency, the phase difference between harmonics focuses only on harmonics and is theoretically derived using the pitch. By checking whether the phase differences between harmonics agree with the theoretical value in each pitch candidate, the pitch can be obtained even when only a few peaks can be detected on the amplitude spectrum. However, the phase spectrum may rapidly fluctuate at the beginning and end of voiced active frames or deteriorate in severely noisy environments, resulting in pitch estimation errors.

In this article, we propose a complex-domain pitch estimation algorithm for narrowband speech signals. The word "complex-domain" denotes a complex-valued mathematical representation such as a complex spectrum containing both amplitude and phase spectrum information, which is utilized in other studies [38], [39] The proposed method is inspired by the idea that the amplitude spectrum can be interpreted as the complex spectrum with the phase spectrum all zeros. That is, frequency-domain pitch estimation algorithms using the amplitude spectrum, which stably work even at the beginning and end of voiced active frames in noisy environments, have degrees of freedom for the phase spectrum. The proposed method thus introduces phase differences between harmonics as an additional cue for pitch estimation. Our previous study showed that the phase difference between harmonics could be derived using the pitch [34]. When the phase spectrum for the higher-order harmonic is shifted by the theoretical value, it agrees with the phase spectrum for the lowest-order harmonic. Therefore, for each pitch candidate, the proposed method calculates the shifted phase spectra and combines them with the amplitude spectrum to generate complex spectra. When a pitch candidate is correct, the cumulative sum of the complex spectra is added in the same direction, emphasizing harmonics even for narrowband speech signals. Finally, a pitch is obtained using the Viterbi algorithm with the cumulative sum. The proposed method can be applied to wideband speech signals as well as narrowband speech signals. The pitch estimation using both the amplitude and phase spectra is more effective for narrowband speech signals with poor pitch estimation cues.

The proposed method is related to several works [40], [41], [42]. Das et al. developed a complex-domain pitch estimation algorithm using the extended complex Kalman filter, which requires prior knowledge of the type of signal [40], [41]. Drugman et al. estimated the spectral boundary separating periodic and aperiodic components derived from amplitude and phase spectra because the phase spectrum conveys relevant information about harmonics [42]. Contributions of this work are summarized as follows:

- We propose a complex-domain pitch estimation algorithm using amplitude and phase spectra without prior knowledge.
- Experimental results show that the proposed method is suitable for narrowband speech signals, not depending on gender, and

suppresses pitch estimation errors even in noisy environments at 0 dB.
- We discuss that the computation complexity of the proposed method is comparable to that of the online pitch estimation algorithm.

II. Relationship Between the Phase Spectrum and Harmonics

This section discusses the relationship between the phase spectrum and harmonics using the STFT representation. In this article, the same STFT representation as phase estimation methods using a sinusoidal model [43], [44] is used. The sinusoidal model assumes that a speech signal can be represented as the sum of sinusoidal waves with a harmonic structure. Let n denote a sample index. Here, a speech signal using the sinusoidal model is denoted as

$$s(n) = \sum_{h=0}^{H-1} 2\,A^h(n) \cos\left(\frac{2\pi \cdot h \cdot f(n)}{F_\mathrm{s}} n + \Omega^h\right), \quad (1)$$

where H is the number of harmonics, $2\,A^h(n)$ is a real-valued amplitude, $f(n)$ is a pitch, F_s is the sampling rate, and Ω^h is the initial phase. With a frequency index k, the STFT representation at the lth frame is defined as

$$S_l(k) = \sum_{n=0}^{N-1} s(lM+n) w(n) \mathrm{e}^{-j\frac{2\pi k}{N} n}, \quad (2)$$

where N is the number of samples to be analyzed, M is the hop size, and $w(n)$ is the analysis window. Pitch and amplitude may only sometimes be stable, particularly in the case of female speech signals toward the end of a word or phrase. Nonetheless, the speech signal can be considered quasi-stationary in a short-time analysis frame, such as between 30 and 80 ms, where the amplitude and pitch are nearly constant [26], [27]. In this work, we set the analysis frame at 40 ms; thus, the assumption is valid such that $A^h(lM) \simeq \cdots \simeq A^h(lM+N-1) = A^h_l$ and $f(lM) \simeq \cdots \simeq f(lM+N-1) = f_l$. By substituting (1) into (2), we have

$$S_l(k) = \sum_{n=0}^{N-1} w(n) \sum_{h=0}^{H-1} A^h_l \left(\mathrm{e}^{j\varphi^h_l(lM+n)} + \mathrm{e}^{-j\varphi^h_l(lM+n)}\right) \mathrm{e}^{-j\frac{2\pi k}{N} n}, \quad (3)$$

where

$$\varphi^h_l(n) = \frac{2\pi \cdot h \cdot f_l}{F_\mathrm{s}} n + \Omega^h. \quad (4)$$

The STFT representation analyzes a signal using band-pass filters with N frequency bands, which are determined by the analysis window. As seen in (3), the output of each band-pass filter contains the analysis results of H harmonic components. It is assumed that the frequency resolution of the STFT representation is sufficiently high, and only a single harmonic is present in a given frequency index. Here, frequency index can only handle integer values, while harmonics include non-integer values. By following the traditional methods [43], [44], the nearest frequency index to the hth ($h = 1, \ldots, H$) harmonic is selected as follows:

$$k^h_l = \arg\min_k |\kappa^h_l - k|, \quad (5)$$

where

$$\kappa^h_l = \frac{h \cdot f_l}{F_\mathrm{s}} N. \quad (6)$$

In addition, it is assumed that the sideband attenuation of the band-pass filters is sufficiently large such that spectrum leakage can be neglected. With these assumptions, the STFT representation for the harmonic (3) is reduced to

$$\begin{aligned} S_l(k^h_l) &\simeq \sum_{n=0}^{N-1} w(n) \cdot A^h_l \mathrm{e}^{j\varphi^h_l(lM+n)} \cdot \mathrm{e}^{-j\frac{2\pi k^h_l}{N} n} \\ &= A^h_l \mathrm{e}^{j\left(\frac{2\pi \cdot h \cdot f_l}{F_\mathrm{s}} lM + \Omega^h\right)} \sum_{n=0}^{N-1} w(n) \mathrm{e}^{-j\frac{2\pi(k^h_l - \kappa^h_l)}{N} n} \\ &= A^h_l \mathrm{e}^{j\left(\frac{2\pi \cdot h \cdot f_l}{F_\mathrm{s}} lM + \Omega^h\right)} W\left(k^h_l - \kappa^h_l\right), \end{aligned} \quad (7)$$

where $W(k)$ denotes the discrete Fourier transform representation of the analysis window. By letting $\phi^\mathrm{W}(k)$ denote the phase spectrum of $W(k)$, the phase spectrum is given as follows:

$$\phi^\mathrm{S}_l(k^h_l) \simeq \frac{2\pi \cdot h \cdot f_l}{F_\mathrm{s}} lM + \Omega^h + \phi^W\left(k^h_l - \kappa^h_l\right). \quad (8)$$

The phase spectrum for the harmonic shifts over time, including the initial phase and the window function. Let us focus on the phase difference between successive frames. Because a pitch slowly changes on voiced active frames, it can be approximated such that $f_l \simeq f_{l-1}$. Similarly, the approximations $k^h_l \simeq k^h_{l-1}$ and $\kappa^h_l \simeq \kappa^h_{l-1}$ are given. The phase difference is then

$$\begin{aligned} \Phi^h_l &= \phi^\mathrm{S}_l(k^h_l) - \phi^\mathrm{S}_{l-1}(k^h_{l-1}) \\ &= \frac{2\pi \cdot h \cdot f_l}{F_\mathrm{s}} M. \end{aligned} \quad (9)$$

The phase difference is theoretically derived using the pitch, not depending on the initial phase or window function. The presence of the harmonic structure can be determined from the phase spectrum continuity based on the phase difference. However, the phase spectrum is unstable due to noise, so the phase difference between the lowest-order harmonic and other harmonics is further calculated. By using (9), the phase difference between harmonics is defined as follows:

$$\begin{aligned} \Psi^h_l &= \Phi^h_l - \Phi^1_l \\ &= \frac{2\pi \cdot (h-1) \cdot f_l}{F_\mathrm{s}} M. \end{aligned} \quad (10)$$

By substituting (9) into (10), we have the relationship between the phase spectrum and harmonics as follows:

$$\phi^\mathrm{S}_l(k^h_l) = \phi^\mathrm{S}_l(k^1_l) + \Theta^h_l, \quad (11)$$

with the phase lead

$$\Theta^h_l = \phi^\mathrm{S}_{l-1}(k^h_{l-1}) - \phi^\mathrm{S}_{l-1}(k^1_{l-1}) + \frac{2\pi \cdot (h-1) \cdot f_l}{F_\mathrm{s}} M. \quad (12)$$

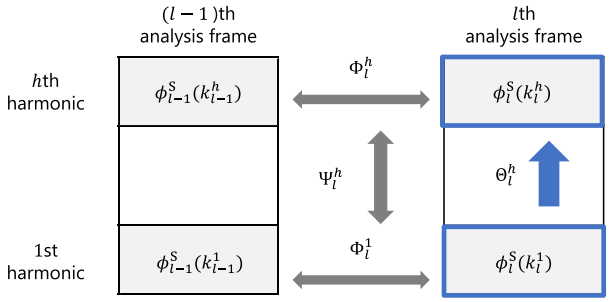

Fig. 1. Relationship between the phase spectrum and harmonics.

Fig. 1 shows the relationship between the phase spectrum and harmonics. In the correct pitch candidate, phase spectra between higher and the lowest-order harmonics exhibits a linear relationship. The validity of the phase difference between harmonics for pitch estimation is discussed in the following section.

III. Phase Differences Between Harmonics

This section discusses the validity of the phase difference between harmonics for pitch estimation. In a natural environment, a speech signal can suffer from noise. In addition, narrowband speech signals can be obtained through a band-pass filter whose cutoff frequencies are 300 and 3400 Hz. A narrowband speech signal is defined as follows:

$$x(n) = g\left(s(n) + e(n)\right), \tag{13}$$

where $e(n)$ is an additive noise signal and $g(\cdot)$ is a band-pass filtering function. Here, the sampling delay of the band-pass filtering has been compensated.

The jth pitch candidate is represented as $f(j)$ $(j \in \mathcal{P})$, where $\mathcal{P}$ denotes the set of indices of the pitch candidates. In this article, the pitch is within 50–400 Hz. When a pitch candidate is correct, the phase spectra for the higher-order harmonics, which are shifted based on the phase difference between the harmonics, agree with the phase spectrum for the lowest-order harmonic. Note that a narrowband speech signal has lost several lower-order harmonics due to its limited bandwidth. The frequency index containing the hth harmonic for the jth pitch candidate in the narrowband is redefined as follows:

$$k^h_j \simeq \frac{h \cdot f(j) + \check{f}(j)}{F_{\mathrm{s}}} N, \tag{14}$$

considering the nearest integer with the highest harmonic in the missing lower band

$$\check{f}(j) = \left\lfloor \frac{300}{f(j)} \right\rfloor \cdot f(j), \tag{15}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. For notational simplicity, we put k^h_j as k^*. Since harmonics possess greater amplitude than other frequencies, phase spectra encompassing them exhibit a high signal-to-noise ratio (SNR). Indeed, phase enhancement methods [43], [44] reconstruct the degraded phase spectrum from phase spectra encompassing harmonics even in noisy environments at 0 dB. Based on this knowledge, we work on a robust pitch estimation algorithm using phase spectra encompassing

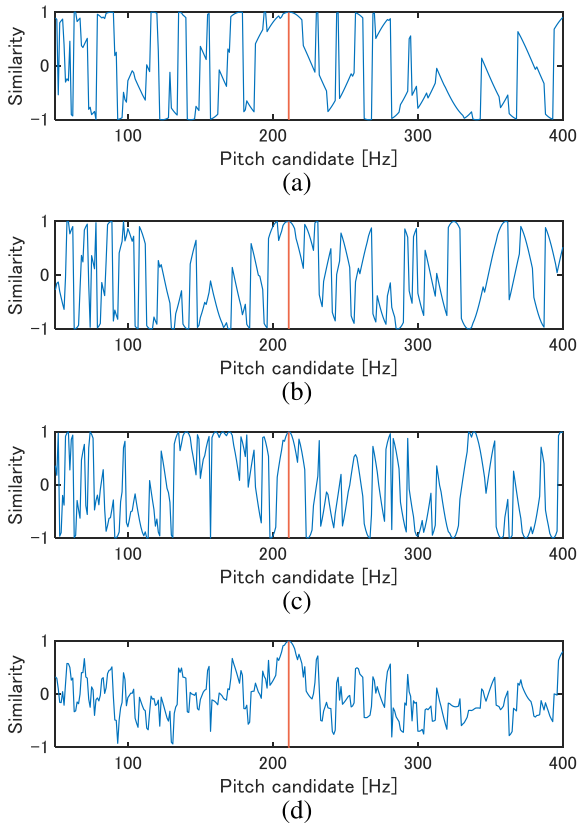

Fig. 2. Similarity between the phase spectra for the lowest-order harmonic and other harmonics in a clean environment. (a) $Z^2_l(j)$. (b) $Z^3_l(j)$. (c) $Z^4_l(j)$. (d) Average of $Z^2_l(j)$, $Z^3_l(j)$, and $Z^4_l(j)$. The red line depicts a pitch of 211 Hz.

harmonics. Let $\phi_l(k)$ be the phase spectrum of the narrowband speech signal. The phase spectrum for the harmonic is then given as

$$\phi_l(k^*) \simeq \phi^{\mathrm{S}}_l(k^*) + \phi^{\mathrm{g}}(k^*), \tag{16}$$

where $\phi^{\mathrm{g}}(k)$ denotes the phase shift of the band-pass filtering.

By using (11) and (12), the shifted phase spectrum based on the phase difference between harmonics in each pitch candidate is generated as follows:

$$\hat{\phi}_l(k^*) = \phi_l(k^*) - \Theta^h_l(j), \tag{17}$$

with the phase lead

$$\Theta^h_l(j) = \phi_{l-1}(k^*) - \phi_{l-1}(k^1_j) + \frac{2\pi \cdot (h-1) \cdot f(j)}{F_{\mathrm{s}}} M. \tag{18}$$

When a pitch candidate is correct such that $f(j) = f_l$, we have $\hat{\phi}_l(k^*) = \phi_l(k^1_j)$. Hence, a pitch can be obtained from narrowband speech signals using the shifted phase spectrum without detecting the peak of the amplitude spectrum.

Let us examine the validity of the shifted phase spectrum for pitch estimation. In this discussion, the similarity between the phase spectra in the frequency indices k^1_j and k^h_j is measured as follows:

$$Z^h_l(j) = \cos\left(\hat{\phi}_l(k^*) - \phi_l(k^1_l)\right). \tag{19}$$

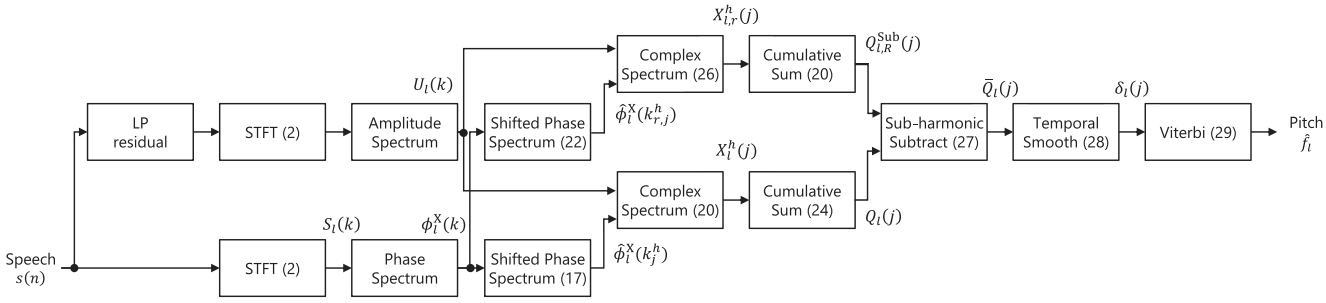

Fig. 3. Schematic representation of the proposed method. The number in each block corresponds to the expression number.

Fig. 2(a), (b), and (c) shows $Z^2_l(j)$, $Z^3_l(j)$, and $Z^4_l(j)$, respectively. Here, a clean narrowband speech signal with a pitch of 211 Hz is used. Similarities approach 1 at several pitch candidates because the phase spectrum is constrained to its principal value of $[0, 2\pi)$; this is called 'phase wrapping.' Due to phase wrapping, the shifted phase spectra can coincide with the phase spectrum for the lowest-order harmonic even in incorrect pitch candidates. As shown in Fig. 2(d), the averaged similarity records a maximum at the correct pitch candidate, resulting in the suppression of the phase wrapping effect using several shifted phase spectra. However, the phase spectrum may rapidly fluctuate at the beginning and end of voiced active frames or deteriorate in severely noisy environments. Moreover, (17) can be established for overtones, resulting in octave errors. To solve these issues, the complex-domain pitch estimation algorithm using the phase and amplitude spectra is discussed in the following section.

IV. Complex-Domain Pitch Estimation Algorithm

In this section, we propose a complex-domain pitch estimation algorithm using the phase and amplitude spectra. Fig. 3 shows a schematic representation of the proposed method. The proposed method obtains a cue for pitch estimation from the amplitude spectrum of the excitation signal using the harmonic summation algorithm [25]. The excitation signal is given by the linear predictive coding technique as the linear prediction (LP) residual [45]. The proposed method combines the shifted phase spectrum with the amplitude spectrum of the excitation signal to generate a complex spectrum. Let $U_l(k)$ denote the amplitude spectrum of the excitation signal. The proposed method generates the complex spectrum as follows:

$$X^h_l(j) = U_l(k^*) \cdot \mathrm{e}^{j\hat{\phi}_l(k^*)}. \tag{20}$$

The proposed method then obtains the distance between the adjacent peaks from the complex spectrum using the harmonic summation algorithm. The harmonic summation algorithm calculates the cumulative sum of the complex spectra

$$Q_l(j) = \left| \sum_{h=1}^{H-1} X^h_l(j) \right|. \tag{21}$$

Fig. 4 shows the vector diagrams of the harmonic summation algorithm. When the shifted phase spectra are the same for the correct pitch candidate, each complex spectrum is added in the same direction and the cumulative sum of the complex spectra then coincides with that of the amplitude spectra. Otherwise, the cumulative sum of the complex spectra is smaller. Therefore, the introduction of the shifted phase spectrum emphasizes peaks of the amplitude spectrum.

Let us examine the validity of the cumulative sum of the complex spectra for pitch estimation. Here, a male narrowband speech signal with a pitch of 103 Hz is used and a white noise signal is added to it at an SNR of 0 dB. Fig. 5(a), (b), and (c) shows the averaged similarity between the phase spectra for harmonics, the cumulative sum of the amplitude spectra, and the cumulative sum of the complex spectra, respectively. The averaged similarity has some peaks at incorrect pitch candidates because the phase spectrum can be degraded by noise. The pitch estimation using only the phase spectrum becomes unstable in severely noisy environments. The cumulative sum of the amplitude spectra suppresses the noise effect and also has peaks at several pitch candidates. Because the peaks on the amplitude spectrum are not easily detected from narrowband speech signals with a partial harmonic structure, the pitch estimation error cannot be suppressed using only the amplitude spectrum. Conversely, the cumulative sum of the complex spectra recorded a peak at the correct pitch candidate. These results confirm that the complex spectrum is an efficient cue for pitch estimation for narrowband speech signals in noisy environments. However, there are still peaks at the pitch candidates for overtones, resulting in the octave error.

When a pitch candidate is an overtone, the amplitude spectrum also has a peak at the subharmonic. To prevent the octave error, the subharmonic subtraction algorithm [25] subtracts the cumulative sum of the amplitude spectra for subharmonics from that for harmonics. However, narrowband speech signals also have missed subharmonics in the lower spectrum. In this article, the subharmonic subtraction algorithm is extended in the complex-domain.

When a pitch candidate is an overtone such that $f(j) = R \cdot f_l$ $(R \in \mathbb{N} - \{1\})$, (11) is also valid. The proposed method introduces an additional cue into the subharmonic subtraction algorithm as to whether the shifted phase spectra agree with the phase spectrum for the lowest-order harmonic. Let $(h + r/R) \cdot f(j) + \hat{f}(j)$ $(r = 1, \ldots, R-1)$ denote a subharmonic for the Rth overtone at a pitch candidate. With the frequency index containing the subharmonic $k^h_{r,j}$, the proposed method shifts

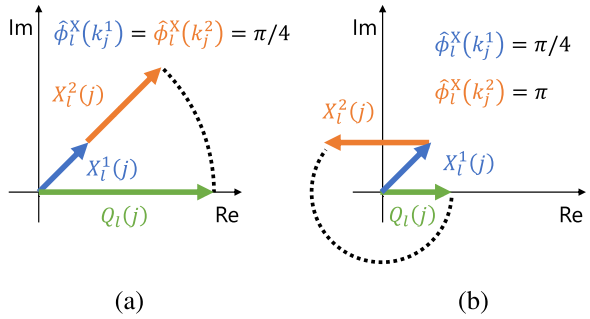

Fig. 4. Vector diagrams of the harmonic summation algorithm in the complex-domain. (a) The phase spectra are the same. (b) The phase spectra are different.

the phase spectrum for the subharmonic as follows:

$$\hat{\phi}_l^{\mathrm{X}}\left(k_{r,j}^h\right) = \phi_l^{\mathrm{X}}\left(k_{r,j}^h\right) - \Theta_{l,r}^h(j), \tag{22}$$

with the phase lead

$$\Theta_{r,j}^h = \phi_{l-1}^{\mathrm{X}}(k_{j,r}^h) - \phi_{l-1}^{\mathrm{X}}(k_j^1) + \frac{2\pi \cdot (R(h-1)+r) \cdot (f(j)/R)}{F_{\mathrm{s}}} L. \tag{23}$$

When a pitch candidate is an overtone, $\hat{\phi}_l^{\mathrm{X}}\left(k_{r,j}^h\right) = \phi_l^{\mathrm{X}}(k_j^1)$. The proposed method calculates the cumulative sum of the complex spectra for the subharmonics as

$$Q_{l,R}^{\mathrm{Sub}}(j) = \left| \sum_{r=1}^{R-1} \sum_{h=1}^{H_R} X_{l,r}^h(j) \right|, \tag{24}$$

where

$$H_R = \left\lfloor \frac{H}{R-1} \right\rfloor \tag{25}$$

$$X_{l,r}^h(j) = U_l\left(k_{r,j}^h\right) \cdot \mathrm{e}^{j\hat{\phi}_l^{\mathrm{X}}\left(k_{r,j}^h\right)}. \tag{26}$$

Consequently, we apply the subharmonic subtraction algorithm to (21):

$$\hat{Q}_l(j) = Q_l(j) - \sum_R Q_{l,R}^{\mathrm{Sub}}(j). \tag{27}$$

In this article, the second and third overtones ($R = 2, 3$) are considered. Fig. 5(d) shows the cumulative sum of the complex spectra with the subharmonic subtraction algorithm. It can be seen that the maximum cumulative sum corresponds to the correct pitch candidate, whereas the others have been attenuated. The octave error can be suppressed using the subharmonic subtraction algorithm in the complex-domain.

In addition, the proposed method enhances robustness using the temporal smoothing process. Here, the scale of $\hat{Q}_l(j)$ is normalized because the power of the narrowband speech signal changes over time. Let $\tilde{Q}_l(j)$ denote the normalized cumulative sum. The normalized cumulative sum over L frames is averaged as follows:

$$\bar{Q}_l(j) = \frac{1}{L} \sum_{l-L+1}^{l} \tilde{Q}_l(j). \tag{28}$$

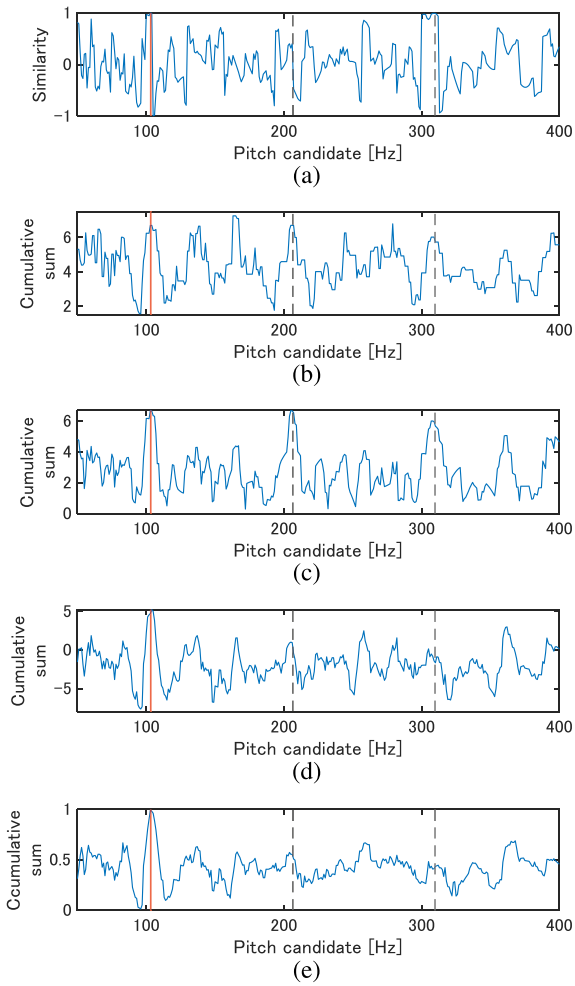

Fig. 5. Similarity and cumulative sum of the spectra with the white noise signal at an SNR of 0 dB. (a) Averaged similarity between the phase spectra for harmonics. (b) Cumulative sum of the amplitude spectra. (c) Cumulative sum of the complex spectra. (d) Cumulative sum of the complex spectra with the subharmonic subtraction algorithm. (e) Cumulative sum of the complex spectra with the subharmonic subtraction algorithm and the temporal smoothing process. The red line and the black dashed lines depict a pitch of 103 Hz and overtones of 206 and 309 Hz, respectively.

Fig. 5(e) shows the cumulative sum of the complex spectra with the subharmonic subtraction algorithm and the temporal smoothing process. The noise effect can be suppressed using the temporal smoothing process, emphasizing the peak at the correct pitch candidate.

Finally, the pitch is determined among the pitch candidates using the Viterbi algorithm that calculates a Viterbi score for each pitch candidate by considering the pitch transition from the previous frames. The Viterbi score is defined as

$$\delta_l(j) = \left[\max_{i \in \mathcal{P}} \delta_{l^*}(i) \cdot a(i,\ j) \right] \cdot \bar{Q}_l(j), \tag{29}$$

where $a(i,\ j)$ is the transition probability between the ith and jth pitch candidates and l^* denotes the latest voiced active frame. Consequently, the proposed method outputs the pitch $\hat{f}_l = f_l(j_l^*)$, where $j_l^* = \arg\max_{j \in \mathcal{P}} \delta_l(j)$.

V. Experimental Evaluation

Objective experiments were conducted to validate the performance of the proposed method. We randomly selected 800 speech signals of 10 male and 10 female speakers from PTDB-TUG [46]. The speech signals had been coded in 16 bits with a sampling rate of 16000 Hz. According to the preprocessing scheme expounded by Pulakka et al. [10], we simulated narrowband speech signals on the public switched telephone network using modified mobile station input filters. First, speech signals were high-pass filtered via an infinite-impulse-response filter whose cutoff frequency was 300 Hz with 3 dB attenuation. We employed a zero-phase digital filter to circumvent phase distortion in this work. Since the first high-pass filter was insufficient in attenuating frequencies below 300 Hz, a second high-pass filter was implemented to achieve an 80 dB attenuation at 200 Hz. We employed a finite impulse response filter as the second high-pass filter following Abel's preprocessing scheme [11]. The high-pass filtered speech signal was lower-pass filtered with an infinite-impulse-response filter whose cutoff frequency was 3400 Hz with 50 dB attenuation, and then downsampled at a sampling rate of 8000 Hz. Finally, narrowband speech signals were obtained by encoding and decoding the downsampled speech signal using G.711 [1].

White, cockpit, destroyer, factory, and babble noise signals were used from NOISEX-92 [47]. The noise signals had been coded in 16 bits with a sampling rate of 19980 Hz. The noise signal was also preprocessed using the modified mobile station input filter and added to the narrowband speech signal at different SNR levels, ranging from -10 to 20 dB in steps of 5 dB.

The online process should shorten an analysis frame, but the spectral resolution is also reduced. The proposed method (**PROP**) analyzed the narrowband speech signal with a frame length of 40 ms ($N = 320$) and a hop size of 10 ms ($M = 80$). In this case, the distance between the center frequencies of the adjacent band-pass filters in the STFT analysis was 25 Hz; the frequency resolution was sufficient to capture the lowest pitch of 50 Hz. In addition, a Hanning window with a frame length of 40 ms was adopted, so that spectrum leakage to the other frequency bands could be neglected. The proposed method set $H = 4$ and $L = 3$ in this article.

The performance of the pitch estimation algorithms was evaluated by measuring the gross pitch error (GPE) [14]. The GPE is defined as

$$\mathrm{GPE} = \frac{N_{\mathrm{E}}}{N_{\mathrm{V}}}, \tag{30}$$

where N_{E} denotes the number of frames in which the relative error of the estimated pitch is higher than 20% and N_{V} denotes the number of voiced active frames. In this work, we introduced the octave error rate (OER) as a metric to assess octave errors:

$$\mathrm{OER} = \frac{N_{\mathrm{OE}}}{N_{\mathrm{V}}}, \tag{31}$$

where N_{OE} denotes the number of frames in which the relative error of the estimated pitch is greater than one octave. We also define the proportion of octave errors to total pitch estimation errors:

$$p_{\mathrm{OE}} = \frac{N_{\mathrm{OE}}}{N_{\mathrm{E}}}, \tag{32}$$

Here, it is assumed that the number and positions of voiced active frames are known, to evaluate the maximum performance of the pitch estimation algorithms.

The proposed method was compared with the following traditional pitch estimation algorithms: **YIN** [14], the pitch estimation filter with amplitude compression (**PEFAC**) algorithm [26], **SRH** [25], the parametric pitch estimation algorithm (**NLS**) [19], [22], and the pitch estimation algorithm using phase differences between harmonics (**PD**) [34]. The frame length of **PEFAC** and **SRH** was 90 ms with a hop size of 10 ms, and the frame length of the others was 40 ms.

Fig. 6 shows the original (red line) and estimated (white line) pitch tracks for a male narrowband speech signal in a clean environment. Fig. 6(a) depicts a spectrogram of the original speech signal, and Fig. 6(b)–(g) depict spectrograms of the narrowband speech signals. The narrowband speech signal lacks the lower-order harmonics because of its limited bandwidth, resulting in the pitch estimation error for **YIN**, **PEFAC**, and **SRH**. Specifically, **PEFAC** and **SRH** suffered from the octave error because only a few peaks can be detected from the amplitude spectrum in the narrowband. In addition, the pitch estimation error was observed for **NLS** because the model parameters were not fully estimated from the narrowband speech signal with a partial harmonic structure. **PD** suppressed the pitch estimation error without detecting peaks on the amplitude spectrum but was unstable at the beginning and end of voiced active frames where the phase spectrum rapidly fluctuated. **PROP** accurately estimated the pitch from the narrowband speech signal using both the amplitude and phase spectra.

The performance of the pitch estimation algorithms was also evaluated for narrowband speech signals in clean and noisy environments. In this experiment, to verify the effectiveness of the shifted phase spectrum and the subharmonic subtraction algorithm, two variants of the proposed method were added: **PROP w/o PD** estimates the pitch from the cumulative sum of the amplitude spectra with the subharmonic subtraction algorithm. **PROP w/o SH** estimates the pitch from the cumulative sum of the complex spectra without the subharmonic subtraction algorithm.

Table I shows the resulting GPE. In the traditional pitch estimation algorithms, the GPE exceeded 0.17 in a clean environment because of the limited bandwidth. Significantly, because the frequency-domain and parametric pitch estimation algorithms focus on wideband speech signals with a harmonic structure, the GPE was higher in the male narrowband speech signals with more missing lower-order harmonics than in the female ones. **PD** achieved lower GPE than **YIN**, **SRH**, and **PEFAC** without detecting the peak on the amplitude spectrum but higher GPE than **NLS** because of the instability at the beginning and end of the voiced active frames. Conversely, **PROP** achieved the lowest GPE for the male and female narrowband speech signals. Consequently, the proposed method precisely estimates the pitch from narrowband speech signals, not depending on

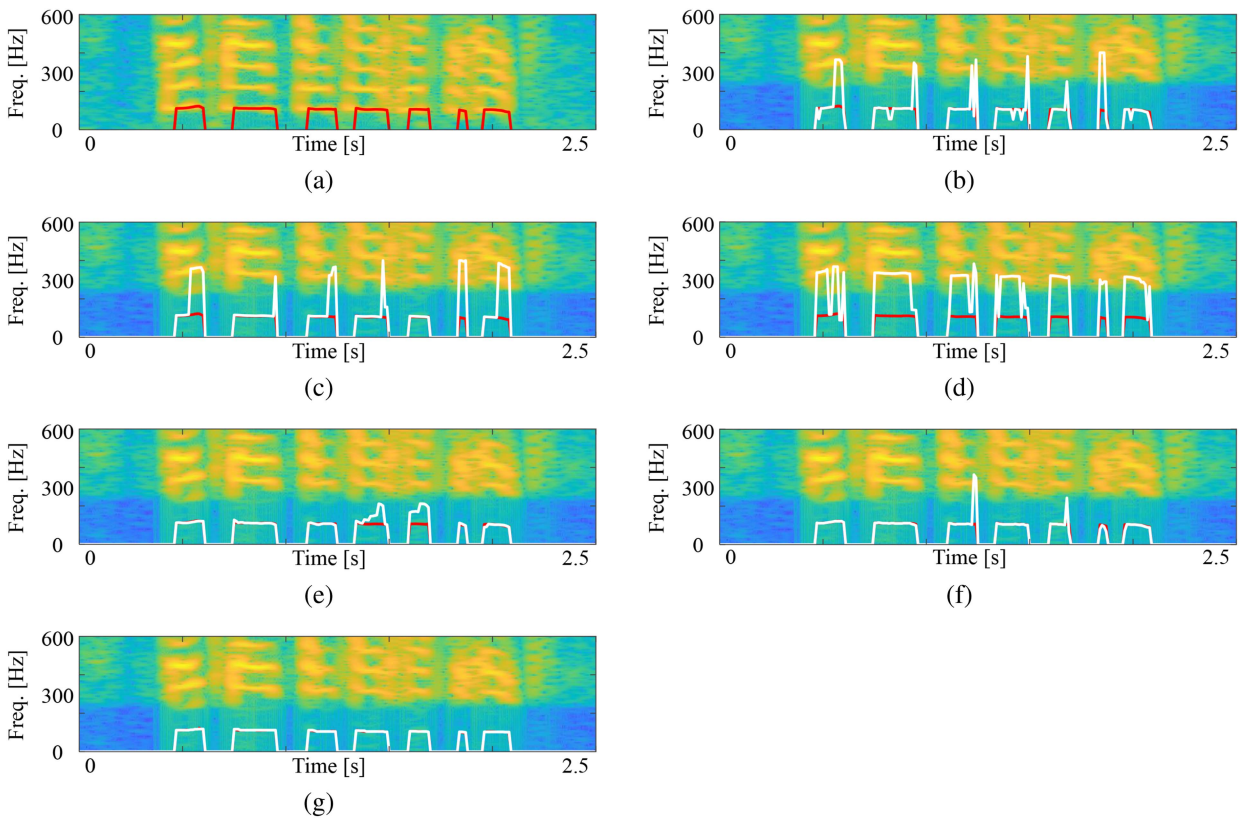

Fig. 6. Pitch track for a male narrowband speech signal in a clean environment. (a) Original. (b) **YIN**. (c) **PEFAC**. (d) **SRH**. (e) **NLS**. (f) **PD**. (g) **PROP**. The red and white lines show the original and estimated pitch tracks, respectively. The upper left panel depicts a spectrogram of the original speech signal, and the others depict the spectrograms of the narrowband speech signals.

gender. Compared with **PROP w/o PD**, **PROP** reduced the GPE by 0.039. These results imply that the pitch estimation error for narrowband speech signals can be suppressed using the amplitude and phase spectra.

Table II shows the resulting OER. In a clean environment, **PROP** recorded the lowest OER at 0.033, whereas **YIN**, **SRH**, and **PEFAC** recorded more than 0.174. The proposed method is an efficient pitch estimation algorithm for narrowband speech signals regardless of whether they are male or female. Compared with **PROP w/o SH**, **PROP** reduced the OER by 0.027. These results show that the subharmonic subtraction algorithm in the complex-domain is effective in preventing the octave error for narrowband speech signals.

In noisy environments, the GPE and OER increased compared to those in a clean environment. In **PD**, the GPE more than doubled because of the deteriorated phase spectrum. **PROP**, the pitch estimation algorithm using both the amplitude and phase spectra, achieved the lowest GPE in all noise signals, which is not a high SNR. **NLS**, employing a harmonic model with additive white Gaussian noise, reduced the GPE for female speech signals in White noise. Nevertheless, **PROP** mitigated pitch estimation errors more effectively for the others. These results show that the proposed method provides stable pitch estimation even in noisy environments. For the OER, **PROP** achieved the lowest OER for the white noise signal, but **NLS** outperformed **PROP** for the other noise signals. Since the non-stationary noise signals suddenly distort the phase spectrum, the proposed method using the phase spectrum has difficulty preventing the octave error. Here, **PD** and **NLS** recorded less than 0.08, whereas **YIN**, **SRH**, and **PEFAC** recorded more than 0.22. Therefore, the proposed method suppressed the octave error with the same level of performance as the state-of-the-art method, **NLS**. In Destroyer, Factory, and Babble noise, the lowest p_{OE} of **PROP** was 0.257, while that of **NLS** was 0.179. **NLS**, which did not capture the harmonic structure as shown in Fig. 6(e), had pitch estimation errors of less than one octave. In **PROP**, the proportion of octave errors increased, despite reducing the pitch estimation error compared to **NLS**. Since we set the extensive pitch range to accommodate both male and female speakers, octave errors are likely to occur. Hence, the postprocessing for the pitch range limitation based on speaker identification methods [48], [49] will further suppress pitch estimation errors.

The proposed method assumes that the phase spectra encompassing lowest and higher harmonics should be related linearly. The relationship between phase spectra can be checked by comparing **PROP w/o PD** and **PROP**. If the relationship between the phase spectra is not linear, the performance of **PROP** will decline compared to **PROP w/o PD** that omits the phase information. As demonstrated in Table I, **PROP** recorded the lowest GPE in Clean speech, signifying a linear relationship between the phase spectra. In noisy environments, **PROP** registered lower GPE with White, Cockpit, and Destroyer noise

TABLE I
GPE OF THE PITCH ESTIMATION ALGORITHMS IN CLEAN AND NOISY ENVIRONMENTS AT AN SNR OF 0 dB

	Clean			White noise			Cockpit noise		
	MALE	FEMALE	ALL	MALE	FEMALE	ALL	MALE	FEMALE	ALL
YIN	0.252	0.246	0.249	0.378	0.443	0.411	0.426	0.495	0.460
PEFAC	0.383	0.211	0.297	0.474	0.274	0.374	0.510	0.355	0.432
SRH	0.642	0.202	0.422	0.623	0.358	0.491	0.738	0.451	0.594
NLS	0.223	0.117	0.170	0.296	**0.120**	0.208	0.264	0.213	0.238
PD	0.206	0.146	0.176	0.360	0.469	0.414	0.402	0.470	0.436
PROP	**0.099**	**0.092**	**0.095**	**0.170**	0.135	**0.152**	**0.208**	**0.173**	**0.191**
PROP w/o PD	0.172	0.097	0.134	0.272	0.139	0.205	0.307	0.181	0.244
PROP w/o SH	0.132	0.102	0.117	0.252	0.128	0.190	0.340	0.180	0.260
	Destroyer noise			Factory noise			Babble noise		
	MALE	FEMALE	ALL	MALE	FEMALE	ALL	MALE	FEMALE	ALL
YIN	0.456	0.513	0.484	0.444	0.506	0.475	0.478	0.561	0.519
PEFAC	0.509	0.339	0.424	0.480	0.315	0.398	0.516	0.366	0.441
SRH	0.780	0.507	0.643	0.748	0.482	0.615	0.802	0.500	0.651
NLS	0.329	0.286	0.307	0.341	0.207	0.274	0.364	0.316	0.340
PD	0.407	0.510	0.458	0.413	0.533	0.473	0.414	0.546	0.480
PROP	**0.291**	**0.234**	**0.262**	**0.264**	0.190	**0.227**	**0.320**	0.280	**0.300**
PROP w/o PD	0.389	0.249	0.319	0.382	**0.187**	0.285	0.420	**0.277**	0.349
PROP w/o SH	0.415	0.261	0.338	0.383	0.198	0.291	0.438	0.276	0.357

TABLE II
OER OF THE PITCH ESTIMATION ALGORITHMS IN CLEAN AND NOISY ENVIRONMENTS AT AN SNR OF 0 dB

	Clean			White noise			Cockpit noise		
	MALE	FEMALE	ALL	MALE	FEMALE	ALL	MALE	FEMALE	ALL
YIN	0.180	0.168	0.174	0.198	0.308	0.253	0.256	0.330	0.293
PEFAC	0.335	0.110	0.223	0.406	0.124	0.265	0.446	0.160	0.303
SRH	0.596	0.089	0.343	0.541	0.135	0.338	0.652	0.143	0.398
NLS	0.057	0.055	0.056	0.114	0.057	0.086	**0.042**	0.069	0.055
PD	0.106	0.083	0.095	0.150	0.319	0.234	0.191	0.310	0.250
PROP	**0.025**	**0.041**	**0.033**	**0.047**	**0.054**	**0.051**	0.050	0.057	**0.053**
PROP w/o PD	0.041	0.043	0.042	0.079	0.060	0.070	0.074	**0.050**	0.062
PROP w/o SH	0.064	0.056	0.060	0.117	0.060	0.088	0.150	0.067	0.108
	Destroyer noise			Factory noise			Babble noise		
	MALE	FEMALE	ALL	MALE	FEMALE	ALL	MALE	FEMALE	ALL
YIN	0.226	0.340	0.283	0.224	0.355	0.289	0.216	0.372	0.294
PEFAC	0.350	0.114	0.232	0.341	0.128	0.235	0.315	0.129	0.222
SRH	0.678	0.128	0.403	0.650	0.171	0.411	0.672	0.151	0.411
NLS	**0.043**	0.067	**0.055**	**0.060**	0.055	**0.058**	**0.057**	0.076	**0.066**
PD	0.125	0.343	0.234	0.136	0.377	0.257	0.104	0.365	0.234
PROP	0.077	0.060	0.068	0.074	0.058	0.066	0.076	0.079	0.077
PROP w/o PD	0.111	**0.058**	0.084	0.110	**0.053**	0.082	0.107	**0.069**	0.088
PROP w/o SH	0.188	0.078	0.133	0.183	0.078	0.130	0.170	0.091	0.130

at 0 dB, while **PROP w/o PD** reported superior GPE for female speech signals in Factory and Babble noise. Since the harmonic structure was deteriorated, the linear relationship between the phase spectra could not be valid. Hence, the robustness toward the non-stationary noise is a challenge for the proposed method. In this article, we estimated the pitch using the deteriorated phase spectrum. Therefore, the preprocessing using phase enhancement methods [50], [51] further improves the performance.

Figs. 7 and 8 show the GPE and OER results in noisy environments with different SNR levels. **PROP** suppressed pitch estimation errors more efficiently within the 20 to -5 dB range. These results confirm that the proposed method is a robust pitch estimation algorithm for narrowband speech signals. In severely noisy environments at an SNR of -10 dB, **PROP** was comparable to **NLS**. The deteriorated phase spectrum caused pitch estimation errors. The preprocessing using phase enhancement methods [50], [51] will improve the performance in severely noisy environments. For the white noise signal, **PROP** also achieved the lowest OER regardless of the SNR level. For the non-stationary noise signals, **NLS** outperformed **PROP** at SNRs below 0 dB. The sudden phase spectrum distortion due to the non-stationary noise signals caused octave errors. Here, **PROP** achieved OER less than or equal to that of NLS at SNRs of 0 dB or higher. The proposed method prevented the octave error more efficiently compared to the state-of-the-art methods in slightly noisy environments.

Finally, we discuss the complexity of the pitch estimation algorithms. Let P denote the number of pitch candidates. The state-of-the-art method **NLS** consists of the harmonic model estimation process $[\mathcal{O}(N \log N) + \mathcal{O}(HN)]$ and the pitch tracking process $[\mathcal{O}(P^2)]$. The complexity of **NLS** is then $\mathcal{O}(N \log N) + \mathcal{O}(P^2)$ because of $H \ll N$. As shown in Fig. 3, the proposed method processes the STFT representation $[\mathcal{O}(N \log N)]$, the shifted phase spectrum calculation $[\mathcal{O}(H)]$, the cumulative summation algorithm $[\mathcal{O}(H)]$, the subharmonic subtraction algorithm $[\mathcal{O}(RH)]$, the temporal smoothing process $[\mathcal{O}(L)]$. and the Viterbi algorithm $[\mathcal{O}(P^2)]$. Because of $R \ll N$, the complexity of the proposed method is $\mathcal{O}(N \log N) + \mathcal{O}(P^2)$, which is the same as that of **NLS**. **NLS** is an online pitch estimation algorithm for practical

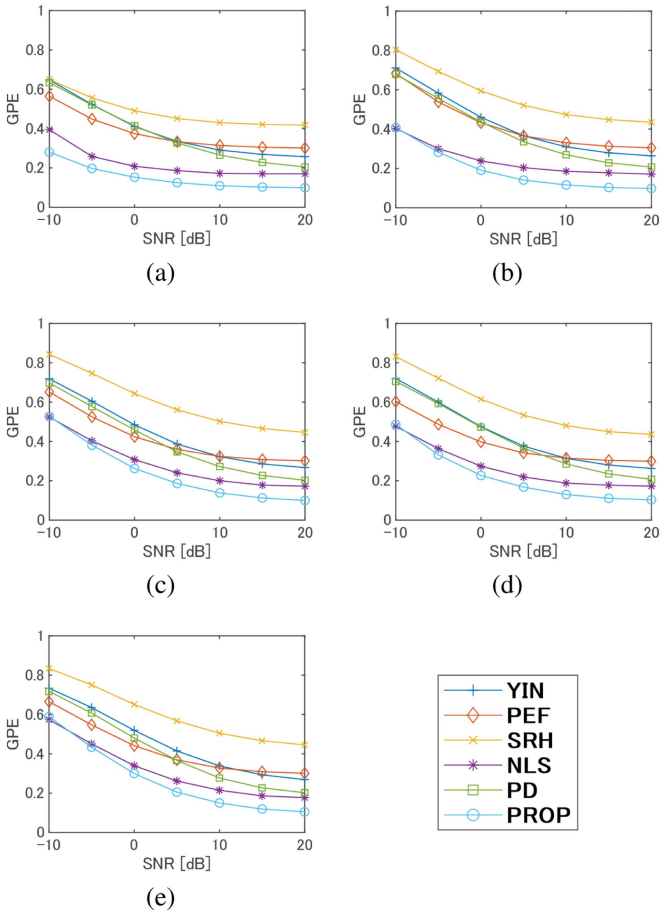

Fig. 7. GPE in noisy environments with different SNR levels. (a) White noise. (b) Cockpit noise. (c) Destroyer noise. (d) Factory noise. (e) Babble noise.

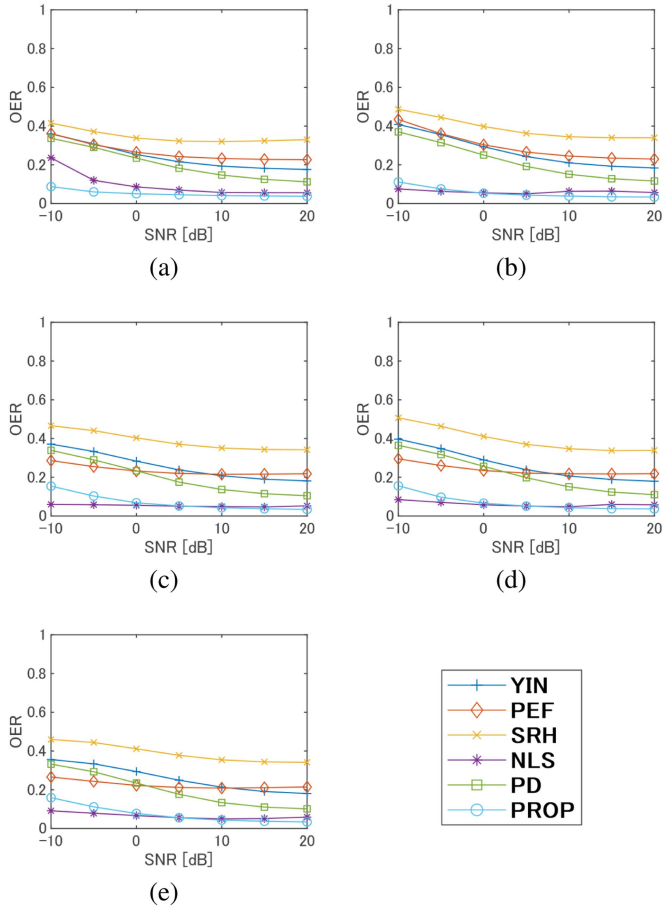

Fig. 8. OER in noisy environments with different SNR levels. (a) White noise. (b) Cockpit noise. (c) Destroyer noise. (d) Factory noise. (e) Babble noise.

applications, and thus the proposed method is also applicable to online processing.

VI. Conclusion

We proposed a complex-domain pitch estimation algorithm for narrowband speech signals and verified its performance through simulation experiments. The proposed method achieved the lowest GPE and OER in a clean environment for both male and female narrowband speech signals. In noisy environments, the proposed method also recorded the lowest GPE even at an SNR of -10 dB and OER less than or equal to that of the state-of-the-art method at SNRs of 0 dB or higher. These results implied that the proposed method was a robust pitch estimation algorithm for narrowband speech signals, not depending on gender. In severely noisy environments at SNRs below 0 dB, the octave error was observed because the phase spectrum was suddenly distorted due to the non-stationary noise signals. Future work includes speaker identification methods [48], [49] and phase enhancement methods [50], [51] to prevent the octave error in severely noisy environments, and the glottal closure instants detection algorithm [52], [53] to detect the voiced active frame. The code of the proposed method is available at https://github.com/Yuya-Hosoda.

References

[1] *Pulse Code Modulation (PCM) of Voice Frequencies*, Standard ITU-T G.711, Int. Telecommun. Union, Geneva, Switzerland, 1988.
[2] *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (SD-ACELP)*, Standard ITU-T G.729, Int. Telecommun. Union, Geneva, Switzerland, 2012.
[3] S. Bruhn et al., "Standardization of the new 3GPP EVS codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5703–5707.
[4] P. Vary and R. Martin, "Bandwidth extension (BWE) of speech signals," *Digital Speech Transmission: Enhancement, Coding Error Concealment*. Chichester, U.K.: Wiley, 2006, pp. 361–387.
[5] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?," *IEEE Commun. Mag.* vol. 44, no. 5, pp. 106–111, May 2006.
[6] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.
[7] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 691–695.
[8] Y. Hosoda, A. Kawamura, and Y. Iiguni, "Phase reconstruction for artificial bandwidth extension toward musical instrument sound signal," in *Proc. IEEE 29th Eur. Signal Process. Conf.*, 2021, pp. 71–75.
[9] Y. Hosoda, A. Kawamura, and Y. Iiguni, "Speech bandwidth extension using data hiding based on discrete hartley transform domain," *Circuit, Syst., Signal Process.*, vol. 41, pp. 2290–2307, 2022.
[10] H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo, and P. Alku, "Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and a gaussian mixture model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2219–2231, Oct. 2012.

[11] J. Abel and T. Fingscheidt, "Sinusoidal-based lowband synthesis for artificial speech bandwidth extension," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 765–776, Apr. 2019.
[12] Y. Hosoda, A. Kawamura, and Y. Iiguni, "Artificial bandwidth extension for lower bandwidth using sinusoidal synthesis based on first formant location," *IEICE Trans. Fundame.*, vol. E105-A, no. 4, pp. 664–672, 2022.
[13] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Proc. Speech Coding Synth.*, 1995, pp. 495–518.
[14] A. E. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.
[15] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2014, pp. 659–663.
[16] S. Lin, "Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-bernoulli filter," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 827–841, Apr. 2019.
[17] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2018, pp. 161–165.
[18] Y. Segal, M. Arama-Chayoth, and J. Keshet, "Pitch estimation by multiple octave decoders," *IEEE Signal Process. Lett.*, vol. 28, no. 7, pp. 1610–1614, Jul. 2021.
[19] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.
[20] B. G. Quinn, J. K. Nielsen, and M. G. Christensen, "Fast algorithms for fundamental frequency estimation in autoregressive noise," *Signal Process.*, vol. 180, 2021, Art. no. 107860.
[21] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 12, no. 1, pp. 76–87, Jan. 2004.
[22] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust bayesian pitch tracking based on the harmonic model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1737–1751, Nov. 2019.
[23] Z. Zhou, M. G. Christensen, J. R. Jensen, and S. Zhang, "Parametric modeling for two-dimensional harmonic signals with missing harmonics," *IEEE Access*, vol. 7, pp. 48671–48688, 2019.
[24] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2002, pp. I-333–I-336.
[25] T. Drugman and A. Abeer, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.
[26] S. Gonzalez and M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.
[27] D. Wang, C. Yu, and J. H. L. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 952–964, May 2017.
[28] M. Khadem-hosseini, S. Ghaemmaghami, A. Abtahi, S. Gazor, and F. Marvasti, "Error correction in pitch detection using a deep learning based classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 3, pp. 990–999, Mar. 2020.
[29] F. Charpentier, "Pitch detection using the short-term phase spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 1986, pp. 113–116.
[30] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the fourier transform," *J. Acoust. Soc. Amer.*, vol. 94, no. 2, pp. 662–667, 1993.
[31] R. Rajan and H. A. Murthy, "Two-pitch tracking in co-channel speech using modified group delay functions," *Speech Commun.*, vol. 89, pp. 37–46, 2017.
[32] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Low-complexity pitch estimation based on phase differences between low-resolution spectra," in *Proc. Interspeech*, 2017, pp. 2316–2320.
[33] E. Loweimi, J. Barker, and T. Hain, "On the usefulness of the speech phase spectrum for pitch extraction," in *Proc. Interspeech*, 2018, pp. 696–700.
[34] Y. Hosoda, A. Kawamura, and Y. Iiguni, "Pitch estimation algorithm for narrowband speech signal using phase differences between harmonics," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc*., 2021, pp. 920–925.
[35] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[36] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 598–602, May 2015.
[37] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1521–1532, Sep. 2015.
[38] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.
[39] R. Gu, S. -X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett*., vol. 28, no. 4, pp. 1370–1374, Apr. 2021.
[40] O. Das, J. O. Smith, and C. Chafe, "Real-time pitch tracking in audio signals with the extended complex Kalman filter," in *Proc. Int. Conf. Digit. Audio Effects*, 2017, pp. 118–124.
[41] O. Das, J. O. Smith, and C. Chafe, "Improved realtime monophonic pitch tracking with the extended complex Kalman filter," *J. Audio Eng. Soc.*, vol. 68, no. 1/2, pp. 78–86, 2020.
[42] T. Drugman and Y. Stylianou, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE Signal Process. Lett*., vol. 21, no. 10, pp. 1230–1234, Oct. 2014.
[43] 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.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.
[44] 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.*, vol. 26, no. 9, pp. 1559–1569, Sep. 2018.
[45] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
[46] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. InterSpeech*, 2011, pp. 1509–1512.
[47] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–252, 1993.
[48] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 4, pp. 1370–1384, Apr. 2020.
[49] C. Zhang, M. Yu, C. Weng, and D. Yu, "Towards robust speaker verification with target speaker enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2021, pp. 6693–6697.
[50] L. Zhang, M. Wang, Q. Zhang, X. Wang, and M. Liu, "PhaseDCN: A phase-enhanced dual-path dilated convolutional network for single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 6, pp. 2561–2574, Jun. 2021.
[51] L. Thieling, D. Wilhelm, and P. Jax, "Recurrent phase reconstruction using estimated phase derivatives from deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2021, pp. 7088–7092.
[52] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech*, 2009, pp. 2891–2894.
[53] J. Matoušek and D. Tihelka, "Using extreme gradient boosting to detect glottal closure instants in speech signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2019, pp. 6515–6519.

Yuya Hosoda (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University, Osaka, Japan, in 2017, 2019, and 2022, respectively. He is currently an Assistant Professor with the Center for IT-based Education, Toyohashi University of Technology, Toyohashi, Japan. His research interests include acoustic and speech signal processing.

Arata Kawamura (Member, IEEE) received the B.E. and M.E. degrees from Tottori University, Tottori, Japan, in 1995 and 2001, respectively, and the D.E. degree from Osaka University, Osaka, Japan, in 2005. He was an Assistant Professor from 2003 to 2012, and an Associate Professor from 2012 to 2018 with Osaka University. He is currently a Professor with the Faculty of Information Science and Engineering, Kyoto Sangyo University, Kyoto, Japan. His research interests include acoustic and speech signal processing.

Youji Iiguni (Member, IEEE) received the B.E. and M.E. degrees from Kyoto University, Kyoto, Japan, in 1982 and 1984, respectively, and the D.E. degree from Kyoto University in 1989. He was an Assistant Professor with Kyoto University from 1984 to 1995, and an Associate Professor with Osaka University, Osaka, Japan. Since 2003, he has been a Professor with Osaka University. His research focuses on systems analysis.