Pitch Estimation Algorithm for Narrowband Speech Signal using Phase Differences between Harmonics

Yuya HOSODA*[1], Arata KAWAMURA†, and Youji IIGUNI*
* Graduate School of Engineering Science, Oska University, Japan
† Faculty of Information Science and Engineering, Kyoto Sangyo University, Japan
[1]hosoda@sip.sys.es.osaka-u.ac.jp

***Abstract*—This paper proposes a pitch estimation algorithm for a narrowband speech signal using phase differences between harmonics. A narrowband speech signal has an incomplete harmonic structure due to bandwidth limitation, which degrades the pitch estimation accuracy. In this paper, we focus on the fact that phase differences between harmonics are constant by approximating a speech signal based on a sinusoidal model. Since the narrowband speech signal has a partial harmonic structure, the proposed method selects the pitch such that the phase differences between harmonics are constant on the narrow bandwidth. Also, we take a frame additive average method for the phase differences between harmonics to improve the robustness against noise. Experimental results show that the proposed method estimates the pitch from the narrowband speech signal under noisy environments more accurately than the traditional methods.**

I. Introduction

A speech signal on the public switching telephone network is limited to a narrowband speech signal with the bandwidth of 0.3–3.4 kHz. The bandwidth limitation due to a low sampling rate results in speech quality degradation [1]. An artificial bandwidth extension (ABE) approach is an effective speech enhancement method for the bandwidth limitation, where the missing upper bandwidth of 3.4–7 kHz and the missing lower bandwidth of 0–0.3 kHz have been reconstructed using the existing narrow bandwidth [2]–[5]. As the ABE approach for the missing lower bandwidth, the methods using sinusoidal synthesis have been established [4][5]. Sinusoidal synthesis generates sinusoidal waves on the missing lower bandwidth using a pitch estimated from the narrowband speech signal. A pitch estimation algorithm is critical to reconstruct the harmonic structure on the missing lower bandwidth [5]. Also, a speech signal on the public switching telephone network may suffer noisy environments. Hence, the ABE approach for the missing lower bandwidth needs a pitch estimation algorithm robust against noise.

Researchers have worked on pitch estimation algorithms for a long time [6]–[15]. Time-domain pitch estimation algorithms obtains a peak of the autocorrelation function as a pitch with a short computation cost [6][7]. The RAPT algorithm [6] is used for the ABE approach for the missing lower bandwidth [5], but it is challenging to the robustness against noise. Also, frequency-domain pitch estimation algorithms estimate a pitch from the amplitude or power spectra [8]–[11]. Here, comb-filter [10] or frame additive average [11] methods have been employed to improve the robustness against noise. Parametric pitch estimation algorithms model a speech signal as a sum of sinusoidal waves with a harmonic structure and estimate a pitch by taking iterative updates for model parameters and noise statics [12][13]. Recently, pitch estimation algorithms using machine learning have been devised, showing the robustness against noise with pre-training a speech model [14][15].

Most of the pitch estimation algorithms assume that a speech signal has a complete harmonic structure. However, a narrowband speech signal has lost several harmonics due to the bandwidth limitation because male speakers have a pitch range of 50–150 Hz and female speakers have a pitch range of 120–400 Hz. As a result, a pitch estimation algorithm may incorrectly estimate a rational multiple of the pitch or sub-harmonics from the narrowband speech signal, termed as 'octave error' [16][17]. Also, pitch estimation algorithms using machine learning require considerable computation cost and speech data to train a speech model with different sampling rates and frequency bandwidths.

In this paper, we propose a pitch estimation algorithm for a narrowband speech signal using phase differences between harmonics. We focus on the fact that the phase differences between harmonics are constant and can be theoretically derived using the pitch by approximating a speech signal based on a sinusoidal model [18][19]. Even for the narrowband speech signal with a partial harmonic structure, the approximation is valid on the narrow bandwidth. The proposed method thus estimates a pitch such that the phase differences between harmonics are constant on the narrow bandwidth. First, pitch candidates are selected using the YIN algorithm [7] without pre-training a speech model. We then calculate the phase differences between harmonics on the narrow bandwidth for each pitch candidate. When a pitch candidate is correct, the phase differences between harmonics are constant. The proposed method calculates similarities between the theoretical phase difference and the phase differences between harmonics of the pitch candidates. Moreover, we enhance the robustness against noise by a frame additive average method for the similarities. Finally, the proposed method determines the pitch using the Viterbi algorithm, considering the temporal smoothness for the pitch variation. Experimental results show that the proposed method estimates the pitch from the narrowband speech signal under noisy environments more accurately than the traditional methods.

II. Proposed Algorithms

A. Pitch Candidates Selection

First, the proposed method selects pitch candidates from a narrowband speech signal using the YIN algorithm [7]. Let $x_l(n)$ $(n = 0, \ldots, N-1)$ be a narrowband speech signal at the l-th frame, where N denotes the number of frame samples. Given a lag index τ $(\tau = 0, 1, \ldots, N/2)$, an autocorrelation function $r_l^t(\tau)(t = 0, 1, \ldots, N/2 - \tau)$ is defined as

$$r_l^t(\tau) = \sum_{v=0}^{N/2-1} x_l(v+t)x_l(v+t+\tau). \quad (1)$$

Using the autocorrelation function, a difference function $d_l(\tau)$ is also defined by

$$\begin{aligned} d_l(\tau) &= \sum_{v=0}^{N/2-1} \Big(x_l(v) - x_l(v+\tau)\Big)^2 \\ &= r_l^0(0) + r_l^\tau(0) - 2r_l^0(\tau). \end{aligned} \quad (2)$$

The difference function is then normalized as a cumulative mean normalized difference function

$$d'_l(\tau) = \begin{cases} 1 & \tau = 0 \\ \frac{\tau d_l(\tau)}{\sum_{v=1}^{\tau} d_l(v)} & \tau = 1, \ldots, N/2 \end{cases}. \quad (3)$$

When a narrowband speech signal is approximately periodic at a period $T = \tau'/F_\mathrm{s}$ with a sampling rate F_s, the cumulative mean normalized difference function becomes locally minimum at $\tau = \tau'$. Let $\hat{\tau}_l$ denote the smallest lag index among the local minima of the cumulative mean normalized difference function. The YIN algorithm outputs the pitch $\hat{f0}_l^{\mathrm{YIN}} = F_\mathrm{s}/\hat{\tau}_l$, but may incorrectly estimate a rational multiple of the correct pitch or sub-harmonics due to the bandwidth limitation and noisy environments. In this paper, we obtain pitch candidates from the local minima of the cumulative mean normalized difference function. The proposed method then selects a pitch among the pitch candidates using phase differences between harmonics.

B. Phase Difference between Harmonics

Based on a sinusoidal model, the relation between the pitch and the phase difference has been discussed using Short-Time Fourier Transform (STFT) [18][19]. A sinusoidal model assumes that a speech signal consists of multiple sinusoidal waves with a harmonic structure and that the pitch changes little over time. Given a frequency index $k(k = 0, 1, \ldots, N-1)$, STFT $X_l(k)$ is defined as

$$X_l(k) = \sum_{n=0}^{N-1} x_l(n)w(n)\mathrm{e}^{-j\frac{2\pi n}{N}k}, \quad (4)$$

where $w(n)$ denotes a window function, and $j = \sqrt{-1}$. A phase spectrum $\phi_l(k)$ is then given as

$$\phi_l(k) = \frac{X_l(k)}{|X_l(k)|}, \quad (5)$$

where $|\cdot|$ denotes an absolute operator. Given a pitch $f0_l$, the frequency index for the h-th harmonic k_l^h is defined as

$$k_l^h = \arg\min_k \Big|k - \frac{h \cdot f0_l}{F_\mathrm{s}} N\Big|. \quad (6)$$

The relational expression for the phase spectrum of the h-th harmonic between the $(l-1)$-th and l-th frames is then given as

$$\phi_l(k_l^h) \simeq \phi_{l-1}(k_{l-1}^h) + 2\pi\frac{h \cdot f0_l}{F_\mathrm{s}} M, \quad (7)$$

where M denotes a frame shift.

Let $\Phi_l^h = \phi_l(k_l^h) - \phi_l(k_{l-1}^h)$ denote the phase difference between the l-th and $(l-1)$-th frames. Using Eq.(7), we theoretically derive the relational expression for the phase difference between the h-th and $(h-1)$-th harmonics as

$$\Phi_l^{h+1} - \Phi_l^h \simeq 2\pi\frac{f0_l}{F_\mathrm{s}} M. \quad (8)$$

It can be seen that the phase difference between harmonics is constant with the pitch regardless of the harmonic number. The proposed method thus selects a pitch among the pitch candidates such that the phase differences between harmonics are constant on the narrow bandwidth.

C. Pitch Selection using Phase Difference between Harmonics

Let $\hat{f0}_l(j)$ $(j \in \mathcal{P}_l)$ be a pitch candidate, where $\mathcal{P}_l$ denotes the set of the pitch candidates. We define the phase difference for each pitch candidate as $\Phi_{l,j}^h$. Let $\hat{f0}_l(j')$ denote a correct pitch candidate. From Eq.(8), the phase differences between harmonics for the correct pitch candidate coincide with the theoretical phase difference, following as

$$\Phi_{l,j'}^{h+1} - \Phi_{l,j'}^h - 2\pi\frac{\hat{f0}_l(j')}{F_\mathrm{s}} M \simeq 0. \quad (9)$$

The proposed method thus evaluates the pitch candidates using the similarity between the theoretical phase difference and the phase differences between harmonics on the narrow bandwidth. Let $h_{l,j}^{\mathrm{NB}}$ denote the lowest harmonic number of the harmonic over 300 Hz for each pitch candidate. The similarity of the phase difference between harmonics $G_l^h(j)$ $(h \geq h_{l,j}^{\mathrm{NB}})$ is defined as

$$G_l^h(j) = \cos\left(\Phi_{l,j}^{h+1} - \Phi_{l,j}^h - 2\pi\frac{\hat{f0}_l(j)}{F_\mathrm{s}} M\right). \quad (10)$$

When the pitch candidate is correct, $G_l^h(j)$ approaches 1, and vice versa.

We examine the validity of the similarity of the phase differences between harmonics on the pitch estimation algorithm for the narrow bandwidth signal. We used a narrowband speech signal with the pitch of 216 Hz, where the pitch candidates ranged of 50–400 Hz. Figure 1 (a)(b)(c) shows the similarities of the phase difference between the $h_{l,j}^{\mathrm{NB}}$-th and $(h_{l,j}^{\mathrm{NB}}+1)$-th, the $(h_{l,j}^{\mathrm{NB}}$+1)-th and $(h_{l,j}^{\mathrm{NB}}+2)$-th, and the $(h_{l,j}^{\mathrm{NB}}+2)$-th and $(h_{l,j}^{\mathrm{NB}}+3)$-th harmonics, respectively. The similarities for the correct pitch candidate at 216 Hz were more than 0.90.

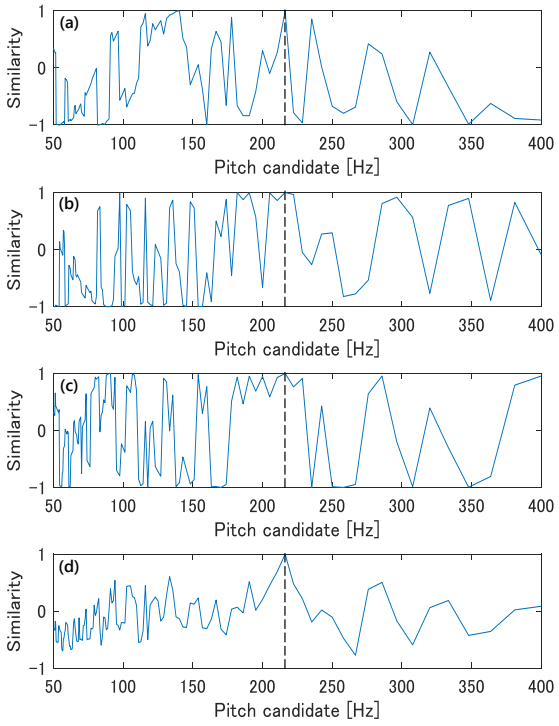

Fig. 1. Similarity of the phase differences between harmonics. (a) Between the $h_{l,j}^{\mathrm{NB}}$-th and $(h_{l,j}^{\mathrm{NB}}+1)$-th harmonics. (b) Between the $(h_{l,j}^{\mathrm{NB}}+1)$-th and $(h_{l,j}^{\mathrm{NB}}+2)$-th harmonics. (c) Between the $(h_{l,j}^{\mathrm{NB}}+2)$-th and $(h_{l,j}^{\mathrm{NB}}+3)$-th harmonics. (d) After the additive average method with $H=4$. Dash line denotes a pitch (216 Hz).

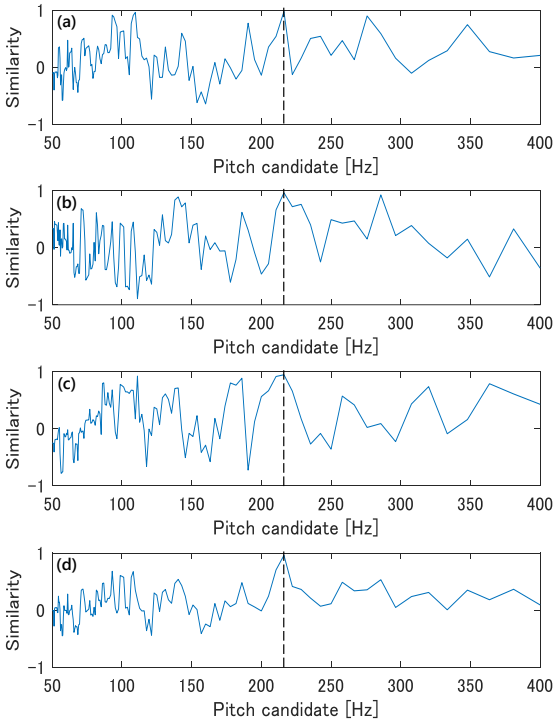

Fig. 2. Similarity of the phase differences between harmonics under a noisy environment. (a) After the additive average method at the l-th frame. (b) After the additive average method at the $(l-1)$-th frame. (c) After the additive average method at the $(l-2)$-th frame. (d) After the frame additive average method with $L=3$. A Vehicle noise signal was added to a narrowband speech signal at 0dB SNR.

However, the similarities for other pitch candidates were also more than 0.90 because a phase spectrum wraps at 0–2π. The proposed method thus takes an additive average method for the several similarities. We calculate the similarity after the additive average method, following as

$$\hat{G}_l(j) = \frac{1}{H} \sum_{h=h_{l,j}^{\mathrm{NB}}}^{h_{l,j}^{\mathrm{NB}}+H-1} G_l^h(j), \tag{11}$$

where H denotes the number of the similarities taken by the additive average method. In this paper, we set $H=4$. Figure 1 (d) shows the similarity after the harmonic additive average method. It can be seen that the maximum of the similarity after the additive average corresponded to the pitch. The similarity of the phase differences between harmonics is therefore valid for the pitch estimation algorithm for the narrow bandwidth signal.

The robustness against noise is challenging for the pitch estimation algorithm. The proposed method takes a frame additive average method for the several similarities to enhance the robustness against noise. The proposed method calculates a similarity after the frame additive average method as

$$\tilde{G}_l(j) = \frac{1}{L} \sum_{l'=l-L+1}^{l} \hat{G}_{l'}(j), \tag{12}$$

where L denotes the number of the similarities taken by the frame additive average method. In this paper, we set $L=3$.

We examine the validity of the similarity after the frame additive average method using a Vehicle noise signal at 0dB Signal-to-Noise ratio (SNR). Figure 2 shows the similarities after the additive average method at the l-th, $(l-1)$-th, and $(l-2)$-th frames, and the similarity after the frame additive average method. Before the frame additive average method, the similarities of not only the correct pitch candidate but also others were more than 0.9 due to noise. On the other hand, the peak of the similarity after the frame additive average method corresponded to the correct pitch candidate. Therefore, the similarity after the frame additive average method is valid for the pitch estimation algorithm under noisy environments.

The proposed method selects a pitch among the pitch candidates according to the similarity after the frame additive average. Note that, even for the rational multiple of the pitch, the similarity after the frame additive average method also becomes significant because Eq.(8) is established for the rational multiple of the pitch. As a result, the proposed method may incorrectly select the rational multiple of the correct pitch. We thus introduce the Viterbi algorithm that selects a pitch considering the temporal smoothness of the pitch variation.

TABLE I
GPE OF PITCH ESTIMATION ALGORITHMS UNDER NOISY ENVIRONMENTS AT 0 dB SNR.

	Engine noise			Vehicle noise			Cockpit noise			Bubble noise		
	Male	Female	Average	Male	Female	Average	Male	Female	Average	Male	Female	Average
YIN	0.547	0.580	0.563	0.327	0.286	0.307	0.415	0.403	0.409	0.478	0.488	0.483
SRH	0.946	0.419	0.682	0.333	0.211	0.272	0.502	0.292	0.397	0.505	0.343	0.424
PEFAC	0.681	0.489	0.585	0.409	0.259	0.334	0.465	0.317	0.391	0.499	**0.333**	0.416
PROP	**0.311**	**0.378**	**0.345**	**0.245**	**0.209**	**0.227**	**0.282**	**0.287**	**0.285**	**0.354**	0.392	**0.373**

TABLE II
OER OF PITCH ESTIMATION ALGORITHMS UNDER NOISY ENVIRONMENTS AT 0 dB SNR.

	Engine noise			Vehicle noise			Cockpit noise			Bubble noise		
	Male	Female	Average	Male	Female	Average	Male	Female	Average	Male	Female	Average
YIN	0.424	0.431	0.428	0.257	0.216	0.237	0.312	0.278	0.295	0.292	0.346	0.319
SRH	0.864	0.199	0.531	0.144	**0.065**	0.105	0.295	**0.062**	0.178	0.200	**0.087**	0.144
PEFAC	0.635	0.275	0.455	0.334	0.113	0.223	0.382	0.094	0.238	0.306	0.124	0.215
PROP	**0.063**	**0.188**	**0.126**	**0.096**	0.110	**0.103**	**0.085**	0.117	**0.101**	**0.077**	0.208	**0.143**

D. Viterbi Algorithm

The Viterbi algorithm calculates a Viterbi score for each pitch candidate using the similarity after the frame additive average method. When there were the pitch candidates in the past frame, the proposed method calculates the Viterbi score

$$\delta_l(j) = \left[\max_{i \in \mathcal{P}_{l^*}} \delta_{l-1}(i) \cdot a(i,\ j)\right] \cdot (\tilde{G}_l(j) + 1), \tag{13}$$

where l^* denotes the latest frame where the pitch candidates existed, and $a(i,\ j)$ denotes a transition probability between the pitch candidates at the l^*-th and l-th frames. The proposed method designs a transition probability as well as the method of Abel et al. [5]:

$$a_l(i,\ j) = \max\left(1 - \left|\frac{\hat{f}0_l(j) - \hat{f}0_l(i)}{\Delta f}\right|^{\beta},\ 0\right). \tag{14}$$

Here, $\Delta f = 285$ Hz and $\beta = 0.3679$ have been optimized beforehand [5]. Finally, the proposed method outputs the pitch as $\hat{f}0_l^{\text{PROP}} = \hat{f}0_l(j^*)$ with $j^* = \arg\max_{j \in \mathcal{P}_l} \delta_l(j)$. When there were no pitch candidates in the past frame, the proposed method outputs the pitch using the YIN algorithm such as $\hat{f}0_l^{\text{PROP}} = \hat{f}0_l^{\text{YIN}}$.

III. EXPERIMENT AND RESULTS

A. Experiment Setup

We conducted an experiment to validate the performance of the proposed method. In the experiment, 100 sentences generated from 10 male and 10 female speakers were selected from the PTDB-TUG database [20], where a speech signal has been coded in 16 bits with $F_{\text{s}} = 16$ kHz. We pre-processed a speech signal using a modified mobile station input filter to assume that the speech signal was passed through the public switching telephone network as well as the method of Abel et al. [5]. First, the speech signal was high-pass filtered with an infinite impulse response filter whose cutoff frequency was 300 Hz with 3 dB attenuation. We then employed a second high-pass filter, which was attenuated by 80 dB at 200 Hz. The high-pass filtered speech signal was low-pass filtered with an infinite impulse response filter whose cutoff frequency was 3.4 kHz with 50 dB attenuation and then down-sampled at $F_{\text{s}} = 8$ kHz. Finally, we obtain a narrowband speech signal by encoding and decoding the low-pass filtered speech signal with G.711 [21]. Four noise signals (Engine, Vehicle, Cockpit, and Bubble noise signals) were used from the NOISE-X database [22] and added to a speech signal before pre-processing. The SNR level for each noise was set from -10 dB to 20 dB in steps of 5 dB.

In this paper, a pitch range was limited to 50–400 Hz. Since at least twice periods have to be considered to capture the minimum pitch of 50 Hz, we set the frame length and the frame shift as 40 ms and 10 ms, respectively. We calculated the phase differences between harmonics using STFT with Hamming window. Also, we assumed that the voiced active frames has been known to evaluate the accuracy of the pitch estimation algorithm with the phase differences between harmonics in this paper. The proposed method (**PROP**) was compared with the YIN algorithm (**YIN**) [7], the Summation of Residual Harmonics algorithm (**SRH**) [9] and the Pitch Estimation Filter with Amplitude Compression algorithm (**PEFAC**) [10].

B. Evaluation Metrics

We evaluated the accuracy of the pitch estimation algorithm using Gross Pitch Error (GPE) [7]. GPE is the rate of the frames on voiced sounds where the relative error for the pitch detection is higher than 20%, following as

$$\text{GPE} = \frac{N_{\text{E}}}{N_{\text{V}}}. \tag{15}$$

Here, N_{E} and N_{V} denote the number of the frames where the the relative error of the estimated pitch is higher than 20% and the number of the frames on the voiced sounds, respectively. Also, we used the rate for octave error (OER) on the estimated pitch, following as

$$\text{OER} = \frac{N_{\text{OE}}}{N_{\text{V}}}, \tag{16}$$

where N_{OE} denotes the number of the frames where the error of the estimated pitch is greater by more than one octave.

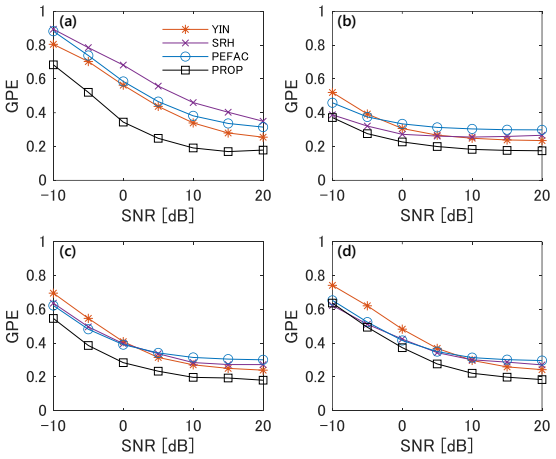

Fig. 3. GPE of pitch estimation algorithms under noisy environments. (a) Engine noise signal. (b) Vehicle noise signal. (c) Cockpit noise signal. (d) Bubble noise signal.

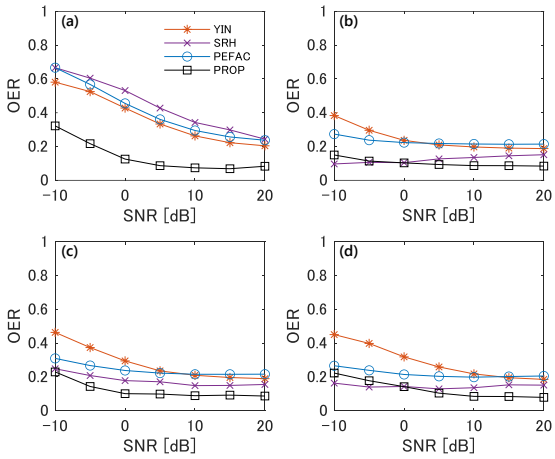

Fig. 4. OER of pitch estimation algorithms under noisy environments. (a) Engine noise signal. (b) Vehicle noise signal. (c) Cockpit noise signal. (d) Bubble noise signal.

C. Results

Table I shows GPE of the pitch estimation algorithms under noisy environments at 0 dB SNR. Since the time-domain pitch estimation algorithm was sensitive to noise, **YIN** recorded the largest GPE for female speakers in all noisy environments. Although **PEFAC** recorded the lowest GPE for female speakers under the Bubble noise signal, the frequency-domain pitch estimation algorithms increased GPE for male speakers in all noisy environments because multiple harmonics have been missed. Significantly, under the Engine noise signal, the rate of the frames for the pitch with a relative error of more than 20% was over 60%. The proposed method achieved the lowest GPE of less than 0.40 for both male and female speakers under all noisy environments. It can be seen that the proposed method robustly estimates the pitch under noisy environments regardless of male or female speakers, even when the part of the harmonic structure has been lost.

Table II shows OER of the pitch estimation algorithms under noisy environments at 0dB SNR. For **YIN**, octave error occurred in more than 20% frames under all noisy environments. **SRH** avoided octave error for female speakers and recorded the lowest OER under the Vehicle, Cockpit, and Bubble noise signals. However, OER for the frequency-domain pitch estimation algorithms increased for male speakers because several harmonics have been lost. Significantly, **SRH** resulted in octave error in more than 60% frames under the Engine noise signal. The proposed method suppressed OER regardless of male and female speakers, and octave error occurred in less than 20% frames under all noisy environments on average. However, for female speakers under the Bubble noise signal, the proposed method resulted in octave error in more than 10% higher frames than **SRH**. Since a Bubble noise signal is a non-stationary noise consisting of multiple speech signals, a narrowband speech signal under the Bubble noise signal not only has lost several harmonics but also has mixed other harmonic structures. As a result, it is challenging for the proposed method to select the correct pitch using the mixed phase difference between harmonics. Therefore, the proposed method avoids octave error under noisy environments without harmonic structures.

Figure 3 shows GPE for pitch estimation algorithms under noisy environments. The proposed method achieved the lowest GPE at all SNR levels. However, under the Vehicle and Bubble noise signals at -10 dB SNR, GPE for the proposed method was comparable to one for **SRH**. These results imply that the partial harmonic structure on the narrow bandwidth has been lost due to noise. Figure 4 shows OER for pitch estimation algorithms under noisy environments. The proposed method achieved the lowest OER at all SNR levels under the Engine and Cockpit noise signals. Under the Vehicle and Bubble noise signals, **SRH** recorded the lowest OER at less than 0 dB SNR. These results confirm that the proposed method is an effective pitch estimation algorithm for the narrowband speech signal when the partial harmonic structure exists in the narrow bandwidth.

IV. Conclusion

This paper proposed a pitch estimation algorithm for a narrowband speech signal, which selected a pitch such that phase differences between harmonics on the narrow bandwidth were constant. We verified the proposed pitch estimation algorithm under noisy environments. The proposed method achieved the lowest GPE under noisy environments regardless of male or female speakers compared with the traditional methods. Also, we showed that the proposed method avoided octave error when the partial harmonic structure existed in the narrow bandwidth. In the future, the pitch estimation algorithm robust to noise with a harmonic structure will avoid octave error more effectively. The code of the proposed method is available at https://github.com/Yuya-Hosoda/Works.

REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, Chichester, West Sussex, U.K.:Wiley, 2006.

[2] 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, 2018.

[3] Y. Dong, Y. Li, X. Li, S. Xu, D. Wang, Z. Zhang, S. Xiong, "A Time-Frequency network with channel attention and non-local modules for artificial bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.(ICASSP)*, 2020, pp.6954–6958.

[4] 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 Language Process.*, vol.20, no.8, pp.2219–2231, 2012.

[5] J. Abel and T. Fingscheidt, "Sinusoidal-based lowband synthesis for artificial speech bandwidth extension," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol.27, no.4, pp.765–776, 2019.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, 1995, pp.495–518.

[7] 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.

[8] A. Camacho, and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol.124, no.3, pp.1638–1652, 2008.

[9] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. INTERSPEECH*, 2011, pp.1973–1976.

[10] S. Gonzalez and M. Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol.22, no.2, pp.518–530, 2014.

[11] F. Huang, and T. Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *IEEE Trans. Audio, Speech, Language Process.*, vol.21, no.1, pp.99–109, 2012.

[12] A. E. Jaramillo, A. Jakobsson, J. K. Nielsen, M. G. Christensen, "Robust Fundamental Frequency Estimation in Coloured Noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.(ICASSP)*, 2020, pp.741–745.

[13] B. G. Quinn, J. K. Nielsen, M. G. Christensen, "Fast algorithms for fundamental frequency estimation in autoregressive noise," *Signal Process.*, vol.180, no.107860, 2021.

[14] 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.(ICASSP)*, 2018, pp.161–165.

[15] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol.28, pp.1118–1128, 2020.

[16] C. Wang and S. Seneff "Robust pitch tracking for prosodic modeling in telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.(ICASSP)*, 2000, pp.1343–1346.

[17] 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.

[18] 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, 2014.

[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.*, vol.26, no.9, pp.1559–1569, 2018.

[20] 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.

[21] International Telecommunications Union, *Pulse code modulation (PCM) of voice frequencies*, ITU-T G.711, 1988.

[22] 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 Communication*, Vol.12, No.3, pp.247–252, 1993.