Pitch Extraction Using Fourth-Root Spectrum in Noisy Speech

Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura

Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama 338-8570, Japan
E-mail: {saifur, sugiura, shima}@sie.ics.saitama-u.ac.jp

Abstract In this paper, we present the use of the fourth-root spectrum instead of the log spectrum for pitch extraction in noisy environments. To obtain clear harmonics, lifter and clipping operations are performed. When the resulting spectrum is transformed into the time domain by the discrete Fourier transform, pitch detection is robust against narrow-band noise. When the same spectrum is amplified by power calculation and transformed into the time domain, pitch detection is robust against wide-band noise. These properties are investigated through exhaustive experiments in various noises. The required computational time is also studied.

Keywords: pitch, fourth-root spectrum, log spectrum, lifter, clipping, power calculation, narrow-band noise, wide-band noise

1. Introduction

The pitch period is defined as the inverse of the fundamental frequency of the excitation source from a voiced speech signal. The pitch period (in short, pitch) or fundamental frequency is a prominent parameter of speech and highly applicable for speech-related systems such as speech analysis-synthesis, speech coding, speech enhancement, and speaker identification systems. The performance of these systems is significantly affected by the accuracy of pitch or fundamental frequency extraction. In this study, we treat pitch and fundamental frequency as having the same meaning, while pitch is inherently interpreted as the perception of fundamental frequency.

Pitch extraction has proven to be a difficult task, even for speech in a noise free environment [1][2]. A clean speech waveform is not really periodic; it is quasi-periodic and highly nonstationary. On the other hand, when a speech signal is corrupted by noise, it is difficult to maintain the reliability and accuracy of pitch extraction algorithms. Under noisy conditions, the periodic structure of the speech signal is destroyed so that pitch extraction becomes an extremely complex task. Among the conventional pitch extraction methods, the autocorrelation function (ACF) [3] is straightforward to compute in the time domain and shows robustness against wide-band random noises such as white noise. The ACF corresponds to a correlation function between the input speech signal and its delayed version in the time domain, but it is also obtained by the inverse Fourier transform of the power spectrum of the speech signal. The ACF is, however, affected by the characteristics of the vocal tract. To reduce the effect of the vocal tract, many algorithms have been developed that rely on the properties of the correlation function [4]-[13]. For example, YIN [4] focuses on the relationship between the conventional ACF and the difference function, and utilizes a cumulative mean function of the difference function to reduce the error rate in pitch extraction. The average magnitude difference function (AMDF) [5] is a simplified version of the ACF, which treats the difference between the speech signal and its delayed version. In [6], the AMDF was combined with linear predictive analysis to eliminate the effect of the vocal tract. Correntropy [7] has similar properties to the ACF and correntropy has a kernel function to transform the original signal into a high-dimensional reproducing kernel Hilbert space (RKHS) in a nonlinear manner. This transformation preserves the characteristics of the periodic signal. Higher-order statistics [8] are also used to enhance the resolution of pitch extraction. However, the performance of correntropy and higher order-statistics in noisy environments is unsatisfactory. In [9], harmonic sinusoidal autocorrelation (HSAC) was proposed. The symmetric average magnitude sum function (SAMSF) was utilized to generate a periodic impulse train to extract the pitch. The resulting pitch extractor based on least squares and optimum value finding (searching) is too complex to implement because it re-
quires post-processing. In [10], dominant harmonic reshaping from the normalized autocorrelation function (NACF) [11] of noisy speech was performed and the empirical mode decomposition (EMD) of the resulting NACF waveform was implemented where an iterative operation could not be avoided. The method in [10] is also complicated and results in a long computation time. In [12], the auditory filterbank decomposed the speech signal into subbands. Then, the NACF was applied to the subband signals, which were encoded to extract the pitch. The NACF reduces the variations in signal amplitude more than the ACF does. The approach in [12] is very effective, but it inherently relies on a sophisticated post-processing technique to compensate for the pitch extraction errors.

In highly noisy environments, the two correlation-based methods, ACF and AMDF, are inferior to the weighted autocorrelation function (WAF) [13]. The WAF also focuses on the ACF, but it is weighted by the inverse of the AMDF, resulting in an excellent pitch extractor in noisy environments. Most of the ACF-based pitch extraction methods are effective in white noise. However, the pitch extraction performance of the ACF-based methods is degraded when clean speech is corrupted by color noise.

In the frequency domain, one of the most widely used techniques employs the cepstrum (CEP), which was originally proposed in [14] and improved in [15]. In the CEP method, the pitch is extracted by applying the inverse Fourier transform to the log-amplitude spectrum, which is also effective. The logarithmic function involved in the CEP shifts the vocal tract characteristics to low-quefrency parts. Utilizing high-quefrency parts, we can extract the pitch without being affected by the characteristics of the vocal tract. The modified CEP (MCEP) in [16] further involves the liftering and clipping operations on the log spectrum, which is used to remove the characteristics of the vocal tract as well as to eliminate the unnecessary notches of spectral valleys that correspond to noise in the log spectrum. The MCEP also removes the high-frequency components to increase the pitch extraction accuracy. The ACF of the log spectrum (ACLOS) [17] also utilizes the liftering and clipping operations on the log spectrum. Then, the ACF is applied to the resulting log spectrum. The ACLOS emphasizes the periodicity of harmonics in the spectrum.

The CEP-based methods clearly express the harmonic structure of the speech signal under no-noise conditions. However, in noisy environments, the CEP-based methods do not always perform well because the speech peaks are affected by the noise peaks in the frequency domain. A spectral harmonic technique was proposed in [18]. In this method, a bank of bandpass lifters is used to flatten the spectrum. The ACF is applied in the spectrum domain to extract the pitch periodicity by reducing the effect of vocal tract characteristics. This approach may be effective but the overall procedure is too complex to implement.

Recently, two sophisticated approaches have been proposed [19][20]. The pitch estimation filter with amplitude compression (PEFAC) [19] is a frequency domain pitch extraction method, which utilizes subharmonic summation [21] in the log frequency domain. The PEFAC also includes an amplitude compression technique to enhance its noise robustness. On the other hand, BaNa [20] considers noisy speech peaks and provides a hybrid pitch extraction method that selects the first five spectral peaks in the amplitude spectrum of the speech signal. BaNa calculates the ratios of the frequencies of the spectral peaks with tolerance ranges and accurately extracts the pitch of the speech signal.

Although deep neural network (DNN)-based approaches exist [22][23] as a recent approach to pitch extraction, they typically require a tremendously long time for learning owing to the huge data size.

In this paper, we propose the use of the fourth-root (FROOT) spectrum of noisy speech for pitch extraction. Motivated by the fact that the MCEP method is very simple to implement but provides an excellent pitch extraction performance in noisy environments, the MCEP method is improved. In the proposed method, which is referred to as the FROOT method, the fourth-root spectrum is used instead of the log spectrum in the MCEP method. The idea of the FROOT method has been reported in a conference [24], where a preliminary experiment was conducted and only limited results for narrow-band noise were shown. In this paper, we further extend the FROOT method for wideband noise and investigate the performance of both the FROOT and extended FROOT methods in various noises. In wide-band noise, the noise energy is distributed over a wide range of frequencies. In this case, the FROOT method is corrupted in the high-
frequency domain by the noise characteristics. However, the extended FROOT method additionally utilizes the fourth-power calculation (fourth-power spectrum) to present clear harmonics and emphasizes the pitch peak in the frequency domain, simultaneously suppressing the noise components included in noisy speech. In this paper, the extended FROOT method is referred to as the FROOT+ method.

The remainder of this paper is organized as follows. Section 2 describes the principle of the FROOT and FROOT+ methods. In Sec. 3, we first show preliminary experiments. After that, we compare the FROOT and FROOT+ methods with conventional methods through experimental results and then discuss the performance and processing time for each method. Finally, we conclude this paper in Sec. 4.

2. FROOT and FROOT+ Methods

Let us assume that the clean speech signal \(x(n)\) is corrupted by noise, \(w(n)\). The noisy speech signal \(y(n)\) is expressed as

\[ y(n) = x(n) + w(n) \] (1)

Figure 1 shows a block diagram of the FROOT and FROOT+ methods. When the fourth-power calculation in parentheses is included, Figure 1 corresponds to the FROOT+ method. When this part is not included, it corresponds to the FROOT method. In the FROOT and FROOT+ methods, firstly we apply a low-pass filter (LPF) to the noisy speech signal because the LPF can eliminate the noise characteristics to increase the accuracy of pitch extraction. The LPF is often applied before the analysis of speech signals and filters out the high-frequency components of the noisy speech signal. We use an LPF with the telephone line cut-off frequency.

After windowing, we calculate the fourth-root spectrum. Here, we considered different spectral shapes of a speech signal as shown in Fig. 2. From Fig. 2, we can observe that the periodicity of the log spectrum is destroyed by the noise. On the other hand, the fourth-root spectrum emphasizes the pitch harmonics in the low-frequency region as well as reduces the noise effect. For this reason, the fourth-root spectrum is used in the FROOT and FROOT+ methods.

However, the fourth-root spectrum is sometimes affected by vocal tract characteristics. To overcome this problem, the operation of flattening is effective. Therefore, we apply a lifter to the fourth-root spectrum by multiplying a filter in the quefrency domain and then converting the liftering result back to the frequency domain. Basically, the vocal tract information is present at the lower part in the quefrency domain. At the higher part in the quefrency domain, the pitch information is present. Therefore, we apply a high-pass lifter (HPL) to eliminate the effect of the vocal tract information and simultaneously eliminate the noise components contained at the lower part in the quefrency domain. The cutoff quefrency level of the HPL should be small to reduce the effect of the vocal tract characteristics. Experimentally, we found that the cutoff quefrency level of 2.5 [ms] (25 samples for the sampling rate of the NTT database) for the HPL preserves the high periodicity more reliably than that with a higher cutoff quefrency level at the lifter output. Some examples are shown in Fig. 3. Therefore, when the FROOT and FROOT+ methods were used in the experiments in Sec. 3, the cutoff quefrency level of 2.5 [ms] for the HPL was used. However, after the lifter operation, we observed that noise components are present between the harmonics. Therefore, a clipping operation is also applied to the lifter output, which reduces the effect of the noise using an accurate clipping threshold level. The selection of the clipping threshold level is described in Sec. 3.

After the above process, in the FROOT+ method, a power calculation is performed (in the FROOT method, this part is omitted). Figure 4 shows an example of which power factor is suitable for the clipping output to reduce the noise components in the FROOT+ method. In Fig. 4, we observe that the noise components are reduced by increasing the power factor. However, as the power factor increases, the effect of the formant characteristics of the vocal tract sometimes also increases. Therefore, undesired peaks arise. From Fig. 4, we selected four as the power factor value for the FROOT+ method, which is the most effective value for reducing the noise. This is the reason why the fourth-power calculation is drawn in Fig. 1.
After this process, for both the FROOT and FROOT+ methods, the inverse discrete Fourier transform (IDFT) is applied and the resulting spectrum is transformed into the time domain, where a peak corresponding to the pitch peak is detected.

Figure 5 illustrates how to extract the pitch period by using the FROOT and FROOT+ methods in narrow-band noise (car interior noise) and in wide-band noise (white noise). In the narrow-band noise (Fig. 5(a)), we observe that the energy level of the first three peaks provides almost the same amplitude in the low-frequency region of the fourth-root spectrum. The pitch information exists in this region, but some peaks are undesired. When the fourth-power calculation is applied to the clipping output, the undesired peaks are enhanced. This leads to the FROOT+ method producing a pitch detection error. However, the FROOT method gives correct pitch detection without undesired peaks. In contrast, in the wide-band noise (Fig. 5(b)), the harmonic peaks maintain their periodicity in the low-frequency region at the clipping output. When the fourth-power calculation is applied to the clipped spectrum, the noise effect is suppressed. Otherwise, the noise components remain in a wide frequency range. Therefore, the FROOT+ method accurately detects the pitch peak. The FROOT method leads to a detection error in this case.

3. Experiments

We conducted experiments on speech signals.

3.1 Experimental conditions

Speech signals were taken from two databases: NTT [25] and KEELE [26]. In the NTT database, which was developed by NTT Advanced Technology Corporation, the speech materials are 11 [s] long and are spoken by four male and four female Japanese speakers for each sentence; the speech signals were sampled at a rate of 10 [kHz]. From the KEELE database, we utilize five male and five female English speech signals. The total length of the ten speakers' speeches is about 6 [m]. The speech signals were sampled at a rate of 16 [kHz]. To generate noisy speech signals, we added different types of noise to the speech signals in both databases. White noise with zero mean and unit variance was generated by a computer and added to the speech signals with amplitude adjustment. Pink, babble, factory, HF channel, car interior, and military vehicle noises were taken from the NOISEX-92 database [27] with a sampling frequency of 20 [kHz], and train noise was taken from the Japanese Electronic Industry Development Association (JEIDA) noise database [28] with a sampling frequency of 8 [kHz]. These noises were resampled with sampling frequencies of 10 [kHz] and 16 [kHz], respectively, when they were added to the speech data in the NTT and KEELE databases. The SNR was set to -5, 0, 5, 10, 20, and ∞ [dB], and the other experimental conditions for pitch extraction were

- frame length: 51.2 [ms], except for BaNa;
- frame shift: 10.0 [ms];
- window function: Hanning;
Fig. 5 Step-by-step processing for FROOT and FROOT+ methods at (a) SNR=0 [dB] (car interior noise) and (b) SNR=0 [dB] (white noise)

- band limitation of LPF: 3.4 [kHz];
- DFT (IDFT) length: 1024 points for the NTT database and 2048 points for the KEELE database except for BaNa.

The following pitch extraction error $e(l)$ based on Rabiner's rule [2] was used for the evaluation of pitch extraction accuracy:

$$e(l) = F_{est}(l) - F_{true}(l)$$  \hspace{1cm} (2)

where $l$ is the frame number and $F_{est}(l)$ and $F_{true}(l)$ are the fundamental frequency extracted from the noisy speech signal and the ground truth fundamental frequency at the $l$th frame, respectively. If $|e(l)| > 10$[%] from the ground truth fundamental frequency, we classified the error as a gross pitch error (GPE) and calculated the GPE rate (as a percentage) over all the voiced frames included in the speech data. Otherwise, we classified the error as a fine pitch error (FPE) and calculated the mean value of the absolute errors. We detected and assessed only voiced parts in sentences for pitch extraction. To extract the pitch, we used the search range of $f_{max} = 50$ [Hz] and $f_{min} = 400$ [Hz], which corresponds to the fundamental frequency range of most people.

The ground truth information for the fundamental frequency at each frame is included in the KEELE database, while the true fundamental frequencies at each frame in the NTT database were measured in [17] by observing the speech waveforms carefully, which are used here. Therefore, the $F_{true}(l)$ values in (2) are known a priori in the evaluation.

3.2 Preliminary experiments

For the FROOT and FROOT+ methods, it is important to set a constant parameter for the clipping threshold level, $\eta$, which is expressed as

$$\eta = \alpha_{min} + C(\alpha_{max} - \alpha_{min})$$  \hspace{1cm} (3)

where $\alpha_{min}$ and $\alpha_{max}$ are respectively the minimum and maximum values of the fourth-root spectrum after the lifter operation, and $C$ denotes a constant parameter. We conducted preliminary experiments to determine the optimal value of the clipping threshold level. For this purpose, we used the NTT database, because the size of its speech data is smaller than that of the KEELE database. We employed male and female speech signals corrupted by white noise. By adjusting the amount of noise to be added, a range of SNR of -5
Fig. 6 Relationship between clipping constant level (C) and GPE at different SNRs (male speakers)

Fig. 7 Relationship between clipping constant level (C) and GPE at different SNRs (female speakers)

[dB] to 20 [dB] was investigated. Additionally, clean speech was also investigated. Figures 6 and 7 show the relationship between the clipping threshold level and average GPE rate of the FROOT+ method for four male and four female speakers, respectively. Here, we changed the clipping threshold level from 0 to 0.9. In Figs. 6 and 7, we observe that setting C = 0.6 - 0.7 for male and female speakers gives low GPE rates at almost all SNR levels.

In accordance with the results in Figs. 6 and 7, we select the constant parameter C = 0.6 commonly for both male and female speech signals to ensure a high extraction accuracy in the FROOT and FROOT+ methods.

3.3 Performance comparison

The pitch extraction performance of the conventional methods (YIN [4] and BaNa [20]) and the FROOT and FROOT+ methods was investigated in noisy environments. In [29], BaNa was assessed as the best pitch extractor in noisy environments among nine methods that were compared. YIN was the second-best method in [22], where the best one was DNN-based. In this paper, we consider eight types of noise, which are classified into two categories depending on their characteristics: wide-band noise and narrow-band noise. White, pink, babble, train, factory, and HF channel noises correspond to wide-band noise. Car interior and military vehicle noises correspond to narrow-band noise. The noise characteristics are discussed in detail in Sec. 3.4. For the FROOT and FROOT+ methods, except for the frame length and the number of DFT(IDFT) points for BaNa. Specifically, for BaNa, the frame length was set as 60 [ms] and the number of DFT (IDFT) points was $2^{16}$ in accordance with [20] (this is the best setting for BaNa). The source code used to implement BaNa was taken from [30]. We implemented the YIN method on the basis of the algorithm described in [4]. In particular, for the YIN method, to confirm the validity of our code, we used the same parameter settings and GPE evaluation criteria as those in [22], and confirmed that the performance of our implemented YIN method provides a similar average GPE rate to the YIN method in [22] for white and babble noises in the KEELE database.

For pitch extraction, we cannot ignore the fact that the extraction performance is largely dependent on the speaker’s characteristics, especially for low or high pitches [1][2][17], which are typical characteristics of male and female speech, respectively. Additionally, different natures of additive noise such as wide-band or narrow-band, flat-spectral or not flat-spectral, and time-invariant or time-variant produce different results for pitch extraction [19][20][22][23]. This is due to the nonuniform phenomena invoked in a complex combination of speech harmonics, formant characteristics and the noise shape created in a framed voiced speech. Therefore, it is important to investigate the pitch extraction performance separately on male and female speech and separately on each noise type. For this reason, we precisely show the result for each case and discuss it later.

(A) NTT database case

Figures 8 and 9 show the average GPE rates of the four male and four female speech signals in the NTT database, respectively, with different noises. Each plot
was obtained under each SNR level from -5 [dB] to ∞ [dB] (clean speech). From Fig. 8, it is observed that in the case of wide-band noise, the average GPE rate of the FROOT+ method is lower than those of the other methods for the white, train, and HF channel noises at low SNRs. At high SNRs (>10 [dB]), the FROOT+ and FROOT methods have similar performance characteristics. At low SNRs of pink and factory noises, BaNa provides a lower error rate than the other methods. At high SNRs (>5 [dB]) of pink and factory noises, the FROOT+ and FROOT methods have similar performance characteristics but provide lower GPE rates than BaNa. In the babble noise case, the FROOT+ method has a lower GPE rate than the YIN method and BaNa, and competitive performance with the FROOT method. On the other hand, in the case of narrow-band noise, the FROOT method provides a lower GPE rate than the other methods at almost all SNR levels except for BaNa at low SNRs (<5 [dB]) of car interior noise.

Fig. 8 GPE for four male speakers with different types of noise under different SNR levels in NTT database
From Fig. 9, the FROOT+ method has significantly better performance than the FROOT method in the wide-band noise case. However, BaNa has a lower GPE rate than the other methods in the wide-band noises except for pink noise. BaNa is still also better in the narrow-band noises, although the FROOT method has better performance than BaNa at low SNRs of car interior noise. In the pink noise case, the FROOT+ method has better performance than the conventional and FROOT methods at all SNRs.

Figures 10 and 11 show the average FPE for male and female speech data, respectively, in the NTT database. The FPE represents the degree of variation in detecting the pitch. The average FPE for all methods ranges approximately from 0.8 [Hz] to 6.2 [Hz]. In Fig. 10, we observe that the FPE of the FROOT+ method is better than those of most of the other methods but not the best. The YIN method has excellent performance at low SNRs (<15 [dB]) in the case of wide-band noise and the FROOT method has the best performance at
high SNRs (>15 [dB]). In the narrow-band noise case, the FROOT method is the best, and the FROOT+ method is typically the second best. On the other hand, in Fig. 11, we observe that BaNa performs better than the other methods at low SNRs (<5 [dB]) in wide-band noise. At high SNRs, the YIN method has the best performance and the FROOT and FROOT+ methods, and BaNa have similar performance characteristics. In the narrow-band noise case, the FROOT method is the best.

(B) KEELE database case
To validate the performance of the FROOT and FROOT+ methods in a more reliable manner, we also employed the KEELE database. Figures 12 and 13 show the average GPE rates for male and female speakers, respectively. The KEELE database provides the ground truth values of the fundamental frequency, which are obtained from laryngograph signals. We analyzed them and found that some discontinuities are present. Therefore, the ground truth values are not
Fig. 11 FPE for four female speakers with different types of noise under different SNR levels in NTT database

particularly accurate. This is reflected in the resulting GPE rates. In Figs. 12 and 13, the GPE rates of the clean speech are clearly higher than those of the clean speech in Figs. 8 and 9. This is due to the lower accuracy of the ground truth values in the KEELE database.

Figure 12 indicates a tendency similar to that in Fig. 8 for all methods. Figure 13 is also similar to Fig. 9 from a performance comparison aspect, although BaNa has comparatively low performance in the babble and car interior noise cases.

The average FPE performance characteristics for male and female speech data are shown in Figs. 14 and 15, respectively. Figure 14 is also similar to Fig. 10, but the FROOT and FROOT+ methods behave similarly, giving the best performance in almost all cases. However, Fig. 15 indicates a different tendency from Fig. 11. In particular, the performance of BaNa deteriorates and BaNa has the worst performance in all cases. However, the relationship between the performance characteristics of the FROOT, FROOT+ and
YIN methods in Fig. 15 is similar to that in Fig. 11.

(C) Summary

Through the results in Figs. 8-15, we observe that the performance of each method has a similar tendency for both speech databases. To summarize, in the wide-band noise case, the FROOT+ method provides a low GPE rate in various types of noise over a wide range of SNRs, although BaNa is advantageous for female speech. Regarding the FPE performance, the FROOT method is superior to BaNa. In the narrow-band noise case, the FROOT method has excellent performance in terms of GPE and FPE.

3.4 Discussion

We next discuss the performance of each method. Figure 16 shows long-term spectra of the different noises we employed. The spectra of the narrow-band noises (car interior and military vehicle noises) have the greatest amplitude in the frequency range of less...
than 200 [Hz], producing narrow-band peaks. On the other hand, the spectra of the wide-band noises (white, pink, babble, train, factory, and HF channel noises) are comparatively evenly spread.

The YIN [4] method is an ACF-based method. For such a method, pitch extraction is robust against wide-band random noises such as white noise but weak against narrow-band noises such as periodic noise. This is consistent with the results in Figs. 8-15. Car interior and military vehicle noises create sharp peaks in the low-frequency region as shown in Fig. 16. These peaks produce clear periodicity in the noise waveform, resulting in the degradation of the GPE rate. HF channel noise has wide-band characteristics. However, a typical peak exists at around 2600 [Hz]. When SNR is high, the peak is negligible. However, when SNR becomes lower, the peak increases in magnitude and is expected to produce periodicity in the noise waveform. This is considered to be the reason why the GPE rate of the YIN method is often severely degraded at low SNRs.
the HF channel noise case as shown in Figs. 8, 9, 12, and 13. In BaNa, the pitch of speech is found from some candidates and post-processing is also incorporated to accomplish accurate pitch extraction. BaNa is capable of overcoming the movement of distorted peaks in noisy cases by estimating the pitch by calculating the harmonic number with a permitted margin. Female speech consists of fewer harmonics in the first formant range and the energy of the voice speech is concentrated at these harmonics; thus, female speech is less affected by noise. In this case, BaNa is advantageous, as shown in Figs. 9 and 13 regardless of the noise type. In contrast, in the male speech, the speech energy is spread over many harmonics and is highly affected by noise. In this case, the performance of BaNa degrades, since the choice of more spectral harmonic peaks must be considered. Actually, the performance of BaNa is comparatively low as shown in Figs. 8 and 12, but it is still excellent at low SNRs by relying on the post-processing algorithm. However, a huge number of points of FFT
is required to find each harmonic peak accurately in BaNa. The computation of several candidate pitches and that of the post-processing including the Viterbi algorithm are complex, resulting in a long computation time as shown in Sec. 3.5. On the other hand, for the FROOT and FROOT+ methods, the fourth-root spectrum makes the periodicity in the harmonic structure clear. In this process, the spectral peak of narrow-band noise is suppressed. Since this effect is combined with the following lifter and clipping operations, the FROOT method is robust against car interior and military vehicle noises as shown in Figs. 8, 9, 12 and 13. However, the FROOT+ method uses the fourth-power calculation after the clipping operation. In this process, the remaining spectral peak of narrow-band noise is enhanced again. Therefore, the performance of the FROOT+ method is lower than that of the FROOT method. However, in the wide-band noise that we employed, unnecessary peaks typically arise in the high-frequency region as noise components. These are sup-

Fig. 15 FPE for five female speakers with different types of noise under different SNR levels in KEELE database
pressed by the fourth-power calculation as shown in Fig. 4. Therefore, the fourth-power calculation can be effectively applied to the clipped spectrum to reduce the noise effect in the case of wide-band noise.

3.5 Processing time

In Table 1, we show the processing time per second of data for each method in the NTT database. We tested all methods on a PC with an Intel (R) Core(TM) i5-6400K CPU with 4 [GHz] clock speed and 8 [GB] of memory. For evaluation, we used five trials for each method then calculated the average processing time required to obtain reliable measurements. The computational time of the YIN method is reasonable because it uses the squared difference function to identify the pitch. BaNa has the longest processing time because of the large FFT size used to achieve a high frequency resolution. The processing times of the FROOT+ and FROOT methods are similar and shorter than those of the other methods, since the clipping and liftering operations are directly applied to the fourth-root spectrum.

4. Conclusion

In this paper, we proposed the use of the fourth-root spectrum to deal with the problem of pitch extraction from noise-corrupted speech signals. The FROOT and FROOT+ methods were derived from the liftered and clipped version of the fourth-root spectrum. The FROOT method in a simple manner by embedding the fourth-power calculation after the liftered and clipped spectrum calculation. The FROOT+ method reduces the effect of vocal tract characteristics as well as suppresses the non-pitch peaks in the frequency domain, enhancing the pitch peak in the wide-band noise. On the other hand, the FROOT method behaves similarly to the FROOT+ method but results in a pitch extractor that is strongly robust against narrow-band noise. Through experiments, we confirmed that both methods are efficient and effective for extracting the pitch in a wide range of noise types when selected in accordance with the noise characteristics such as wide-band and narrow-band noises.

References

[1] W. Hess: Pitch Determination of Speech Signals, Springer-Verlag, 1983.
[2] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal: A comparative performance study of several pitch detection algorithms, IEEE Trans. Acoust. Speech and Signal Process., Vol. 24, No. 5, pp. 339-417, 1976.
[3] L. R. Rabiner: On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech and Signal Process., Vol. 25, No. 1, pp. 24-33, 1977.
[4] A. Cheveigne and H. Kawahara: YIN: A fundamental frequency estimator for speech and music, J. Acoust. Soc. Am., Vol. 111, No. 4, pp. 1917-1930, 2002.
[5] M. J. Ross: Average magnitude difference function pitch extractor, IEEE Trans. Acoust. Speech and Signal Process., Vol. 22, No. 5, pp. 353-362, 1974.
[6] C. K. Un and S. Yang: A pitch extraction algorithm based on LPC inverse filtering and AMDF, IEEE Trans. Acoust. Speech and Signal Process., Vol. 25, No. 6, pp. 565-572, 1977.
[7] J. W. Xu and J. C. Principle: A pitch detector based on a generalized correlation function, IEEE Trans. Audio Speech and Lang. Process., Vol. 16, No. 8, pp. 1420-1432, 2008.
[8] A. Moreno and J. A. R. Fonollosa: Pitch determination of noisy speech using higher order statistics, Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), 1992.
[9] C. Shahnaz, W. Zhu and M. O. Ahmad: Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme, IEEE Trans. Audio Speech and Lang. Process., Vol. 20, No. 1, pp. 322-335, 2012.
[10] S. K. Roy, M. K. I. Molla, K. Hirose and M. K. Hasan: Harmonic modification and data adaptive filtering based approach to robust pitch estimation, Int. J. Speech Tech., Vol. 14, No. 4, pp. 339-349, 2011.
[11] D. Talkin: A robust algorithm for pitch tracking (RAPT), in speech coding and synthesis, W. B. Kleijn and K. K. Paliwal, Eds., pp. 496-518, Elsevier, 1995.
[12] S. Lin: Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-Bernoulli filter, IEEE Trans. Speech and Lang. Process., Vol. 27, No. 4, pp. 827-841, 2019.
[13] T. Shimamura and H. Kobayashi: Weighted autocorrelation for pitch extraction of noisy speech, IEEE Trans. Speech and Audio Process., Vol. 9, No. 7, pp. 727-730, 2001.
[14] A. M. Noll: Cepstrum pitch determination, J. Acoust. Soc. Am., Vol. 41, No. 2, pp. 293-309, 1967.
[15] S. Ahnadi and A. S. Spanias: Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Trans. Speech and Audio Process., Vol. 7, No. 3, pp. 333-338, 1999.
[16] H. Kobayashi and T. Shimamura: A modified cepstrum method for pitch extraction, Proc. IEEE Asia-Pacific Int. Conf. Circuits and Systems Microelectronics and Integrating Systems (APCCAS), 1998.
[17] N. Kunieda, T. Shimamura and J. Suzuki: Pitch extraction by using autocorrelation function on the log spectrum, Electron. Commun. Jpn., Part 3, Vol. 83, No. 1, pp. 90-98, 2000.
[18] M. Lahat, R. J. Niederjohn and D. A. Krubsack: A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech, IEEE Trans. Acoust. Speech and Signal Process., Vol. 35, No. 6, pp. 741-750, 1987.
[19] S. Gonzalez and M. Brookes: PEFAC - A pitch estimation algorithm robust to high levels of noise, IEEE/ACM Trans. Audio Speech and Lang. Process., Vol. 22, No. 2, pp. 518-530, 2014.
[20] N. Yang, H. Ba, W. Cai, I. Demirkol and W. Heinzelman: BaNa : A noise resilient fundamental frequency detection algorithm for speech and music, IEEE/ACM Trans. Audio Speech and Lang. Process., Vol. 22, No. 12, pp. 1833-1848, 2014.
[21] D. J. Hermes: Measurement of pitch by subharmonic summation, J. Acoust. Soc. Am., Vol. 83, No. 1, pp. 257-264, 1988.
[22] D. Wang, C. Yu and J. H. L. Hansen: Robust harmonic features for classification-based pitch estimation, IEEE/ACM Trans. Audio Speech and Lang. Process., Vol. 25, No. 5, pp. 952-964, 2017.
[23] K. Han and D. Wang: Neural network based pitch tracking in very noisy speech, IEEE/ACM Trans. Audio Speech and Lang. Process., Vol. 22, No. 12, pp. 2158-2168, 2014.
[24] S. Motegi and T. Shimamura: Fundamental frequency extraction in narrow band noise, Proc. Spring Conf. Acoust. Society of Japan, pp. 277-278, 2011.
[25] “20 Countries Language Database,” NTT Advanced Technology Corp., Japan, 1988.
[26] F. Plante, G. Meyer and W. Ainsworth: A fundamental frequency extraction reference database, Proc. Eurospeech, pp. 837-840, 1995.
[27] A. Varga and H. J 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. 247-251, 1993.
[28] S. Itahashi: Creating speech copora for speech science and technology, IEICE Trans. Fundam. of Electron. Commun., and Computer Sci., Vol. E 74, No. 7, pp. 1906-1910, 1991.
[29] L. Sukhostat and Y. Imamverdiyev: A comparative analysis of pitch detection methods under the influence of different noise conditions, J. Voice, Vol. 29, No. 4, pp. 410-417, 2015.
[30] WCNG, Wireless Communication Networking Group, [Online] Available: http://www.ecc.rochester.edu/projects/wcng/code.html

Md. Saiful Rahman received his B.Sc.(Honours) and M.Sc. degrees in information and communication engineering from University of Rajshahi, Bangladesh in 2006 and 2007, respectively. In 2012, he joined as a Lecturer in the Department of Information and Communication Technology, Comilla University, Comilla, Bangladesh, where he is currently serving as an Assistant Professor. He is currently pursuing his Ph.D. degree in the field of speech processing at the Graduate School of Science and Engineering at Saitama University, Japan. His research interests include speech analysis and digital signal processing.

Yosuke Sugiura received his B.E., M.E., Ph.D. degrees from Osaka University, Osaka, Japan in 2009, 2011, and 2013, respectively. In 2013, he joined Tokyo University of Science, Tokyo, Japan. In 2015, he joined Saitama University, Saitama Japan, where he is currently an Assistant Professor. His research interests include digital signal processing, adaptive filter theory, and speech information processing.

Tetsuya Shimamura received his B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1986, 1988, and 1991, respectively. In 1991, he joined Saitama University, Saitama, Japan, where he is currently a Professor. He was a visiting researcher at Loughborough University, U.K. in 1995 and at Queen’s University of Belfast, U.K. in 1996, respectively. Prof. Shimamura is an author and coauthor of six books. He serves as an editorial member of several international journals and is a member of the organizing and program committees of various international conferences. His research interests are in digital signal processing and its application to speech, image, and communication systems. (Received February 21, 2020; revised April 11, 2020)