CORRELATION BASED FUNDAMENTAL FREQUENCY EXTRACTION METHOD IN NOISY SPEECH SIGNAL

Mirza A. F. M. Rashidul Hasan

Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh

ABSTRACT

This paper proposed a correlation based method using the autocorrelation function and the YIN. The autocorrelation function and also YIN is a popular measurement in estimating fundamental frequency in time domain. The performance of these two methods, however, is effected due to the position of dominant harmonics (usually the first formant) and the presence of spurious peaks introduced in noisy conditions. The experimental results of computer simulations on female and male voices in different noises perform that the gross pitch errors are lower in proposed method as compared to other related method in different types of signal to noise ratio conditions.

KEYWORDS

Fundamental frequency extraction, correlation, autocorrelation function, white noise, exhibition noise

1. INTRODUCTION

Speech signals are composed of a sequence of sounds, which, for purposes of analysis and study, are assumed to be physical realizations of a discrete set of symbols. These sounds, and the transitions between them, serve as a symbolic representation of information. In real world fundamental frequency (i.e., pitch) estimation is very advanced research area. Accurate extraction of fundamental frequency plays a key role in acoustical signal processing and has a various applications in related fields like speech analysis synthesis, speech coding, speech recognition, speech enhancement, speech and speaker identification. In this connection, different fundamental frequency detection methods are proposed in literature [1,2,3]. Generally, fundamental frequency detection methods are three types, i.e., time domain [4, 5], frequency domain [6,7], and time frequency domain [8]. At present many fundamental frequency detection methods have been published, but more reliable method is still a challenging research.

The time domain autocorrelation function (ACF) method is very popular for their simplicity, low computational complexity and good performance in the presence of noise [4]. On the other hand, a very popular method, average magnitude difference function (AMDF) [9] has the benefit of low computation and high precision and the calculation cost needed less than that of ACF. The major disadvantages of ACF and AMDF are half pitch error and double pitch error when high noise is introduce with original signal. Correlation based methods are relatively perform good result against noisy conditions. We uses ACF method and YIN [10] method of this paper. The ACF and YIN produces a peak and a notch respectively. According to our proposed method, when the ACF is combined with the inversed YIN, the peak of the ACF is emphasized in a noisy condition. Finally, using this proposed method, we extract accurate fundamental frequency in high noisy case [11].

The article is organized as follows. Section 2 describe time domain methods and their drawbacks. Section 3 present the proposed method and Section 4 gives the experimental results. Finally, the article is concluded in Section 5.

DOI : 10.5121/ijcseit.2017.7101

2. PROBLEM DESCRIPTION

The ACF $R(\tau)$ of the voiced signal $y(m)$ is generally defined as in (1)

$$R(\tau) = \frac{1}{M} \sum_{m=0}^{M-1} y(m)\, y(m+\tau) \tag{1}$$

in (1) the underlying voiced frame size is M and the lag number is τ. If $y(m)$ is periodic at pitch period T, $R(\tau)$ exhibits peak at $\tau = iT$, where $i = 0,1,2,3....$ As the value of τ increases, $R(\tau)$ tends to decrease which facilitates the use of the second peak (at $\tau = T$) for estimation of the pitch period. Figure 1(b) represents the ACF of clean signal plotted in Figure 1(a). In clean signal the performance of ACF is better and this examples agree that this method detected the pitch peak accurately. On the other hand, YIN method also better for pitch detection in clean speech. For example, the pitch peak is accurately estimated by the YIN method in clean speech as shown in Figure 1(c).

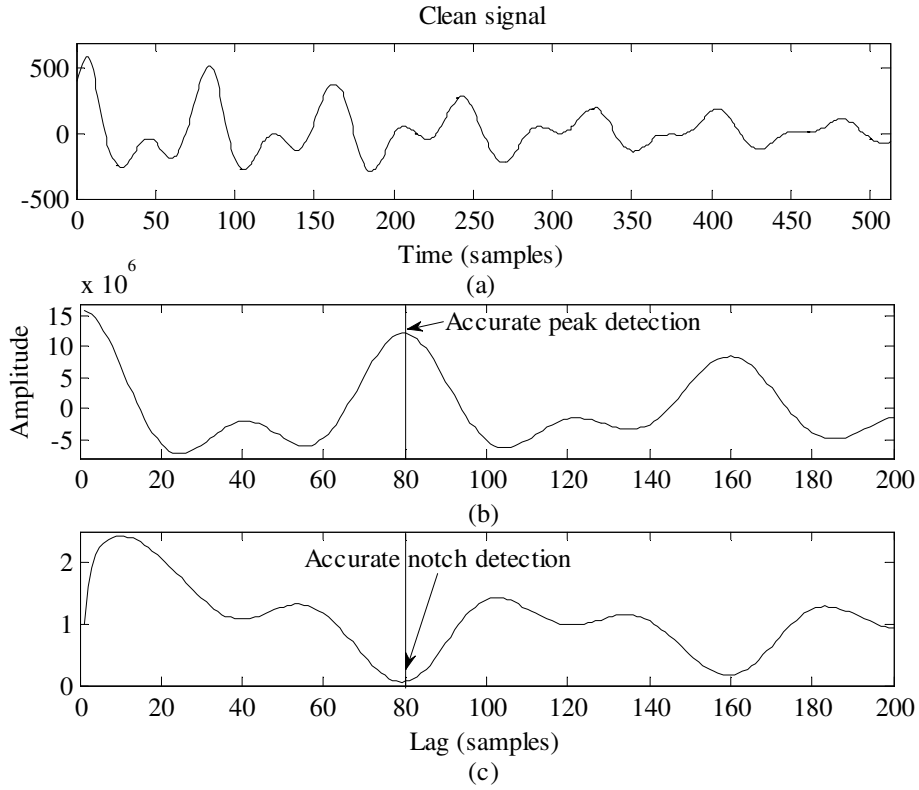

Figure 1. (a) Clean speech frame, (b) Autocorrelation of clean speech in (a), and (c) YIN of clean speech in (a). The vertical line indicates the correct fundamental frequency value.

The ACF is also the inverse Fourier transform of the power spectrum of the signal. Thus, if there is a distinct formant structure in the signal, it is maintained in the ACF. Spurious peaks are also sometimes introduced in the spectrum in noisy or even in noiseless conditions. This sometimes makes true peak selection a difficult task. The ACF and YIN obtained from the voice signal in Figure 1(a), corrupted with noise at signal to noise ratio (SNR) = 0 dB, is shown in Figure 2,

where both method fails to detect the true peak. Among many other improvements reported on the ACF method. Talkin proposed a normalized cross correlation based method [12]. Hasan proposed signal reshaping technique [13] for emphasizing the true peak. Shimamura proposed weighted the ACF [14] by the inverse average magnitude difference function [9].

The effect of falling trend of minima in the AMDF, enhancing the pitch peak in a lower proportion than the succeeding speaks and thereby making the method vulnerable to double pitch error. To overcome this issue, we propose to utilize a correlation based processing i.e., the ACF is divided by the inverse of the YIN.

3. PROPOSED METHOD

Let us assume that the *y(m)* signal is introduced by additive white Gaussian noise, the noisy signal is given by

$$y(m) = u(m) + v(m) \tag{2}$$

for $m = 0,1,2,...,M\text{-}1$, here *u(m)* and *v(m)* represent the noise less signal and noise respectively. Regarding this situation the ACF given [14] by

$$\begin{aligned} R(\tau) &= \frac{1}{M}\sum_{m=0}^{M-1}(u(m)+v(m))(u(m+\tau)+v(m+\tau)) \\ &= \frac{1}{M}\sum_{m=0}^{M-1}((u(m)u(m+\tau)+u(m)v(m+\tau) \ +v(m)u(m+\tau)+v(m)v(m+\tau)) \\ &= R_{uu}(\tau)+2R_{uv}(\tau)+R_{vv}(\tau) \end{aligned} \tag{3}$$

where $R_{uu}(\tau)$ is the ACF of signal *u(m)*, $R_{uv}(\tau)$ is the cross correlation of signal *u(m)* and noise *v(m)*, and $R_{vv}(\tau)$ is the ACF of noise *v(m)*. For large *M*, if *u(m)* does not correlate with *v(m)*, then $R_{uv}(\tau)=0$. Furthermore, if *v(m)* is uncorrelated, then $R_{vv}(\tau) = 0$ except for $\tau = 0$. Under this consideration, the relations

$$\begin{aligned} R(\tau) &= R_{uu}(\tau)+R_{vv}(\tau)(\tau=0) \\ R(\tau) &= R_{uu}(\tau)(\tau\neq 0) \end{aligned} \tag{4}$$

are valid. According to this (4), the ACF gives robust performance against noise. The ACF provide some peaks at the locations of *iT*. Although the maximum peak is located at $\tau = T$ except for the case of $\tau = 0$, in some cases, the peak located at $\tau = 2T$ becomes larger than that located at $\tau = T$. Then a half pitch error occurs. On the other hand, a peak is often made at $\tau < T$. That cases, leads to a double pitch error. Here is an example for these types of error in ACF as shown in Figure 2(b). The YIN is established on the principle of difference function. This method attempts to minimize the difference between the waveform and its delayed duplicate instead of minimizing the product. This method is defined as

$$d(\tau) = \sum_{m=0}^{M-1}\left|y(m) - y(m+\tau)\right|^2 \tag{5}$$

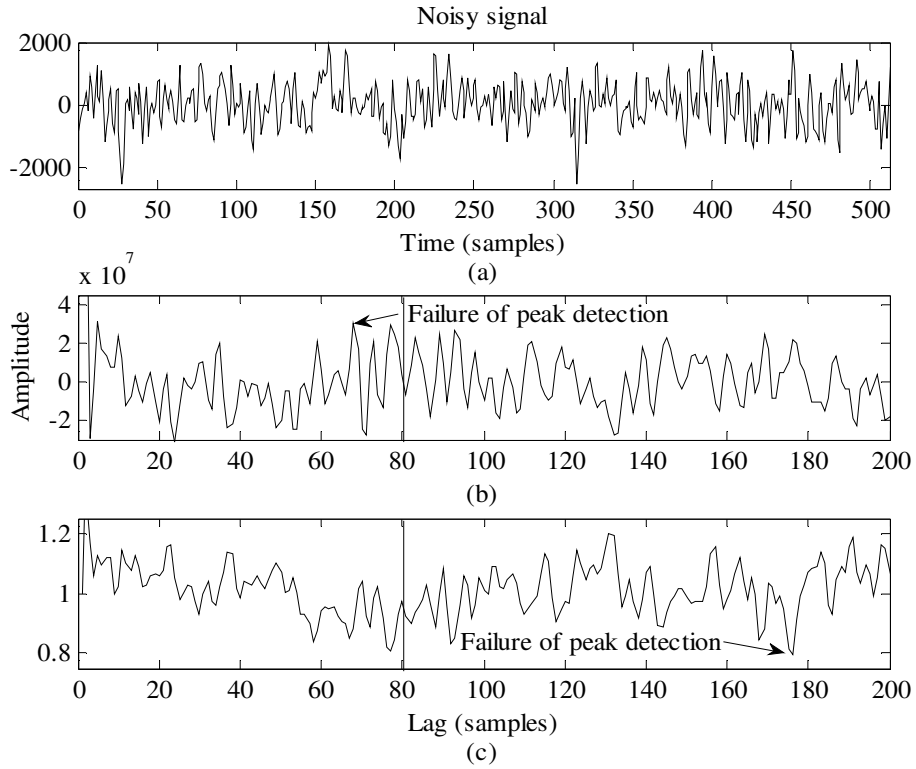

Figure 2. (a) Noisy speech frame (SNR = 0 dB)(which is the same frame as Figure 1(a)), (b) Autocorrelation of noisy speech in (a), and (c) YIN of noisy speech in (a). The vertical line indicates the correct fundamental frequency value.

In (5), $d(\tau)$ produces small when $y(m)$ is similar with $y(m+\tau)$. It means that if $y(m)$ has a period of T, $d(\tau)$ produces a deep notch at $\tau = T$. Therefore, $1/d(\tau)$ makes a peak at $\tau = T$. Hence using the ACF weighted by $1/d(\tau)$, it is expected that the true peak is emphasized (Figure 3), and finally the pitch errors are decreased.

The proposed method is define as

$$F(\tau) = \frac{R(\tau)}{d((\tau) + l)} \tag{6}$$

where l is a fixed number (l>0). The YIN is (5) provides at $\tau = 0$, which invokes a divergence of the directly inversed YIN. For this reason, the denominator in (6) is stabilized by adding the number l.

Figure 2 and Figure 4 shows the ACF, YIN and proposed methods obtained for a voiced speech signal in Figure 1(a) is corrupted by noise (SNR = 0 dB). In that condition, by picking the maximum amplitude of each method, the proposed method performs to the true fundamental frequency (Figure 4), on the other hand, the ACF and YIN method (Figure 2) does an erroneous one. Figure 5 represents a block diagram of the proposed fundamental frequency extraction method.

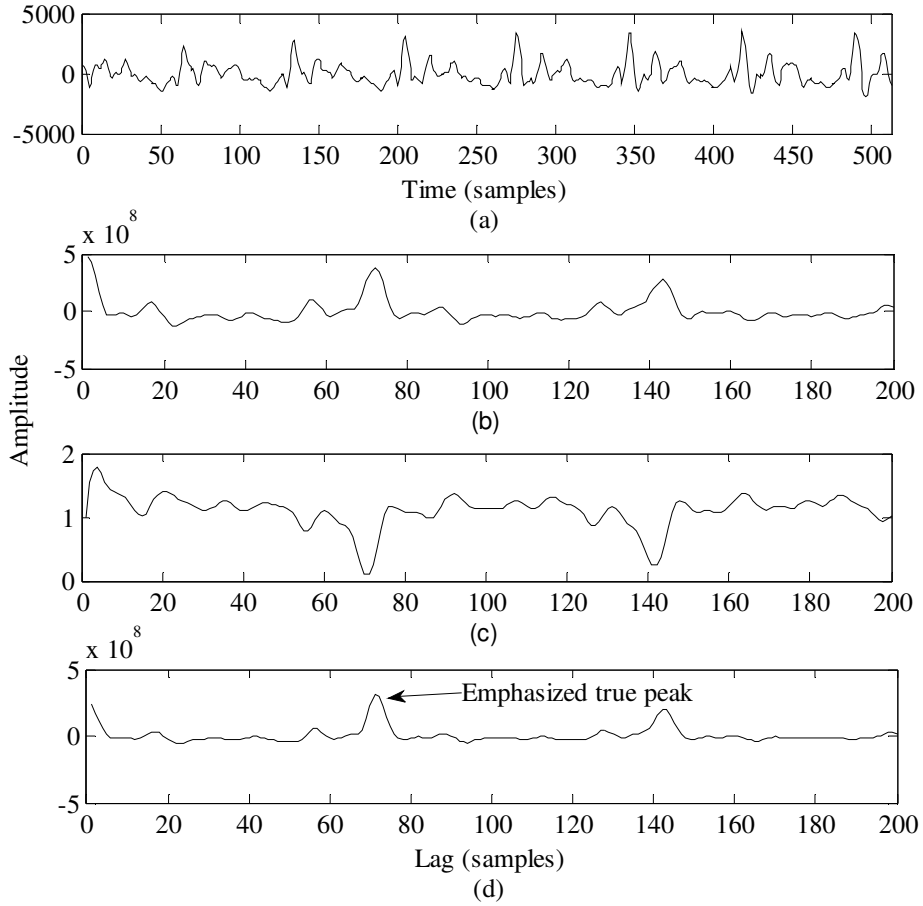

Figure 3. (a) Clean speech frame, (b) Autocorrelation of clean speech in (a), (c) YIN of clean speech in (a), and (d) Proposed method of clean speech in (a).

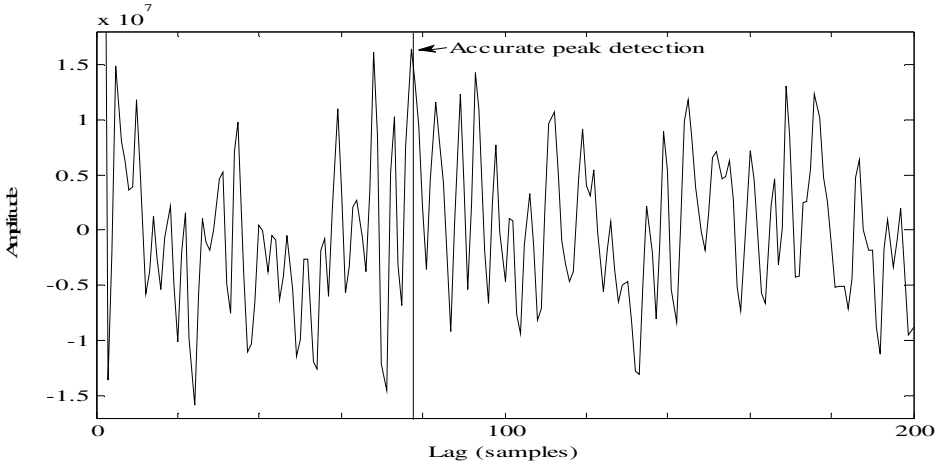

Figure 4. Performance of the proposed method of noisy speech.

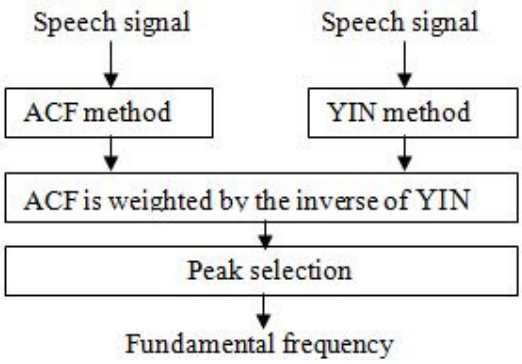

Figure 5. Block diagram of the proposed method.

4. EXPERIMENTS AND RESULTS

To evaluate the proposed method, natural speech signals spoken by two Japanese female and male speakers are examined. Speech signals are taken from NTT database [15]. The reference file is constructed by computing the pitch frequency every 10 ms using a semi automatic technique based on visual inspection. White Gaussian noise is added to these speech signals before the simulations. Fundamental frequency estimation error is calculated as the difference between the reference and estimated fundamental frequency. For the performance evaluation of the proposed method, criteria considered in our experimental work is gross pitch error (GPE). The evaluation of accuracy of the extracted fundamental frequency is carried out by using

$$e(l) = F_t(l) - F_e(l) \tag{7}$$

where $F_t(l)$ is the true fundamental frequency, $F_e(l)$ is the extracted fundamental frequency by each method, and $e(l)$ is the extraction error for the l-th frame. If $|e(l)| > 20\%$, we recognized the error as a gross pitch error (GPE)[10,13]. Otherwise we recognize the error as a fine pitch error (FPE). The possible sources of the GPE are pitch doubling, halving and inadequate suppression of formants to affect the estimation. The percentage of GPE, which is computed from the ratio of the number of frames (F_{GPE}) yielding GPE to the total number of voiced frames (F_v), namely,

$$GPE(\%) = \frac{F_{GPE}}{F_v} \times 100 \tag{8}$$

The mean FPE is calculated by

$$FPE_m = \frac{1}{N_i} \sum_{j=1}^{N_i} e(l_j) \tag{9}$$

where l_j is the j-th interval in the utterance for which $|e(l_j)| \leq 20\%$ (fine pitch error), and N_i is the number of such intervals in the utterance. As metrics, the GPE (%), FPE_m provide a good description of the performance of a fundamental frequency estimation method. The experimental conditions are tabulated in Table 1.

Table 1. Condition of Experiments.

Sampling frequency	10 kHz
Window function	Rectangular
Frame size	51.2 ms
Frame shift	10 ms
Number of FFT points	2048
SNRs (dB)	∞, 20, 15, 10, 5, 0, -5

We aim to detect the pitch information of noiseless and noisy speech. Noise is taken from the Japanese Electronic Industry Development Association (JEIDA) Japanese Common Speech Corporation. The outcomes of the proposed method (PROPOSED) is compared with a well known weighted autocorrelation (WAUT) method [14]. For the implementation of the WAUT, the parameter τ in [14] is set to 1. In our proposed method, we consider the YIN method assumes a threshold value 0.1 considering the value of the first dip (opposite to peak) as 0. Though the assumption is true for clean speech, it does not always hold for noisy cases. The thresholds is therefore redefined as [global minimum +0.1], which is in tune with the concept of threshold and is empirically confirmed to be suitable for all conditions. In order to evaluate the fundamental frequency estimation performance of the proposed method, we plot a reference fundamental frequency contour for noisy speech in white noise speech of a female speaker from the reference database and also the fundamental frequency contours obtained from the other fundamental frequency estimation method in Figure 6.

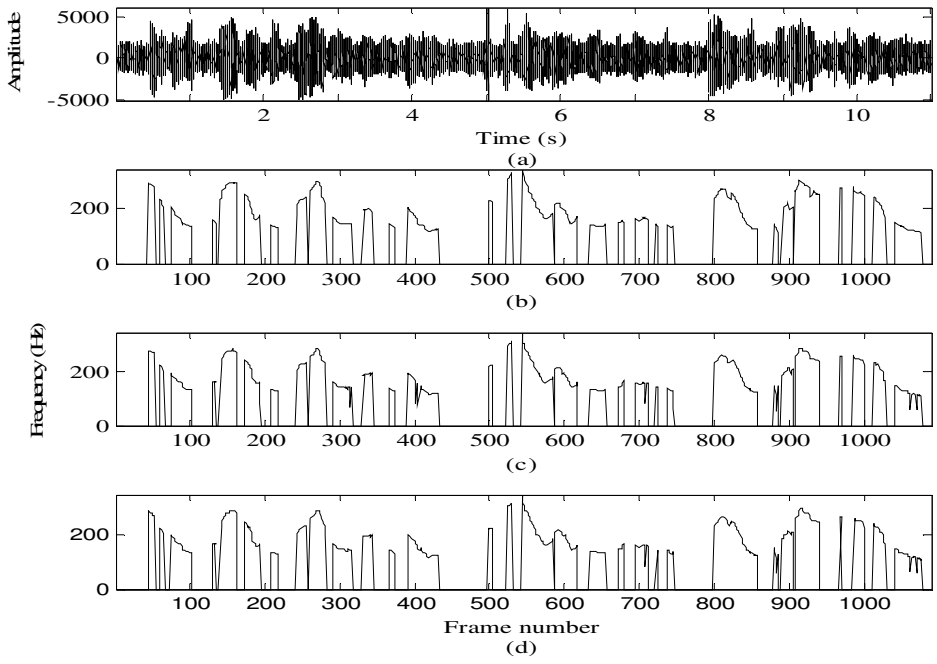

Figure 6. (a) Noisy speech signal for female speaker in white noise at an SNR 0 dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) WAUT, and (d) proposed method.

Figure 6 shows that in contrast to the other method, the proposed method yields a relatively smoother fundamental frequency contour even at an SNR of 0 dB. Figure 7 shows a comparison of the fundamental frequency contour resulting from the two methods for the male speech corrupted by the white noise at an SNR of 0 dB. In Figure 7 it is clear that the proposed method is able to give a smoother contour. The fundamental frequency contours in Figures 6 and 7 obtained from the two methods have convincingly demonstrated that the proposed method is capable of reducing the double and half fundamental frequency errors thus yielding a smooth fundamental frequency track.

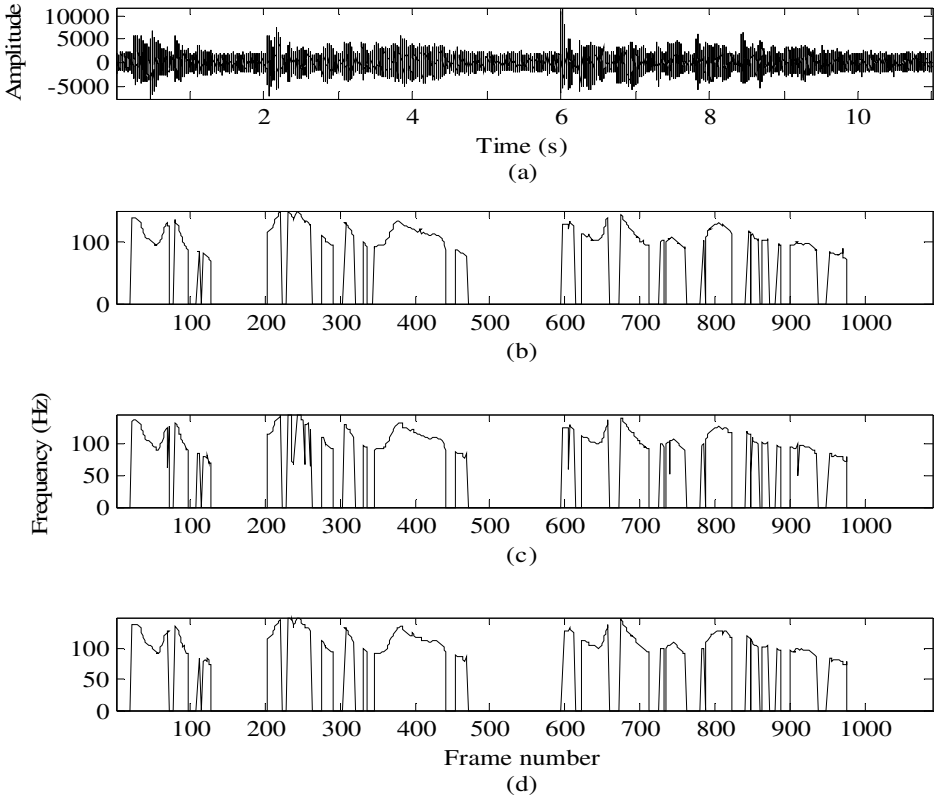

Figure 7. (a) Noisy speech signal for male speaker in white noise at an SNR 0 dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) WAUT, and (d) proposed method.

Fundamental frequency estimation error in percentage, which is the average of GPEs for white noise and exhibition noise, are shown in Figures 8 and 9, respectively. These figure implies that the proposed method gives far better results for both female and male cases in different types of noises in various SNR conditions. These experimental results show that the proposed method is superior to the WAUT method in almost all cases. Particularly, at low SNR (0 dB, -5 dB), the proposed method performs more robustly compared with the WAUT method.

The FPE indicates a degree of the fluctuation in detected fundamental frequency. For the FPE, mean of the errors (in Hz) was calculated. Considering all the utterances of the female and male

speakers, in Figures 10 and 11, the FPE values resulting from the two methods are plotted, respectively. Average FPEs for all methods range approximately from 1.5 Hz ~8 Hz. It is also seen from Figures 10 and 11 that in all noise cases, the FPE values resulting from the proposed method are small but the WAUT method give relatively higher values of FPE in this range. From the simulation results it is found that the value of FPEs is also within the acceptable limit and consistently satisfactory at other SNRs.

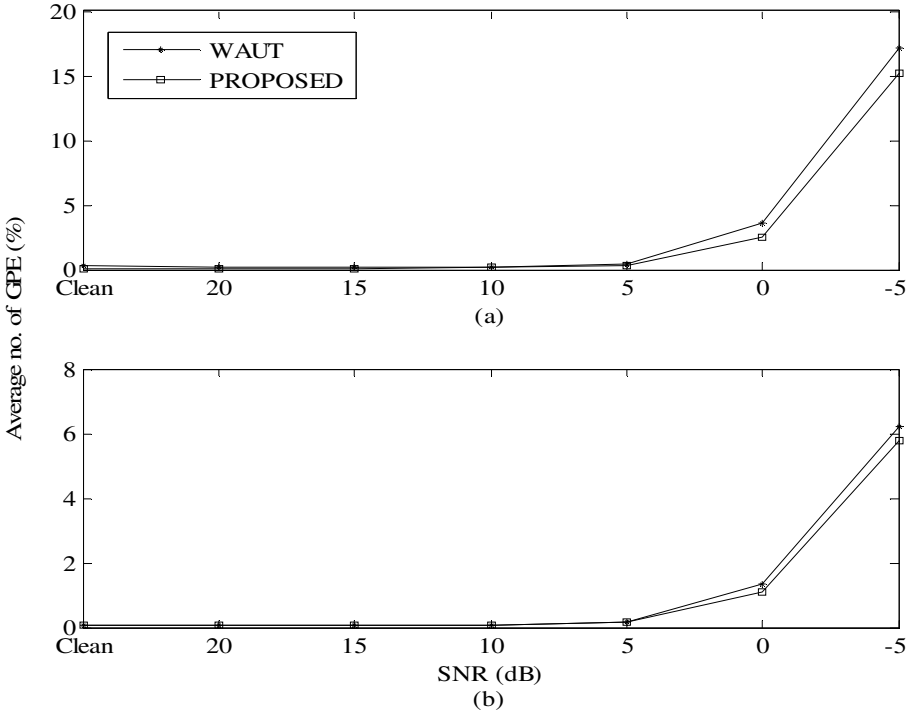

Figure 8. Comparison of percentage of average gross pitch error (GPE) in white noise at various SNR conditions. (a) Female speaker, and (b) Male speaker.

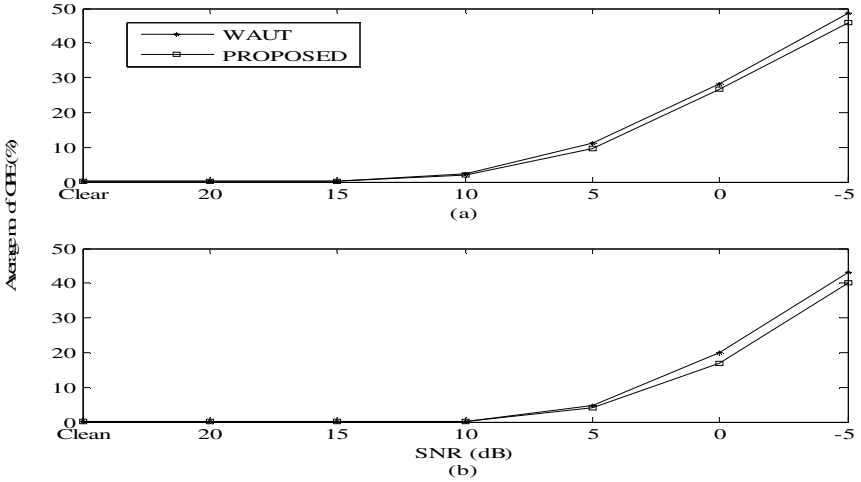

Figure 9. Comparison of percentage of average gross pitch error (GPE) in exhibition noise at various SNR conditions. (a) Female speaker, and (b) Male speaker.

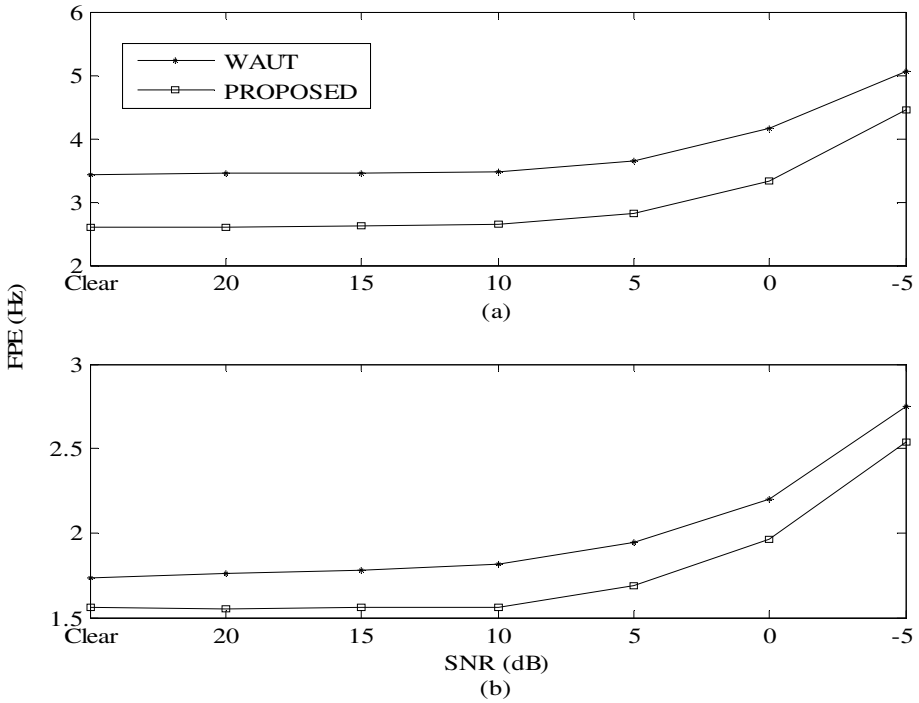

Figure 10. Comparison of average performance results in terms of mean fine pitch error (FPE) in white noise for different speaker under various SNR conditions; (a) Female speaker, (b) Male speakers.

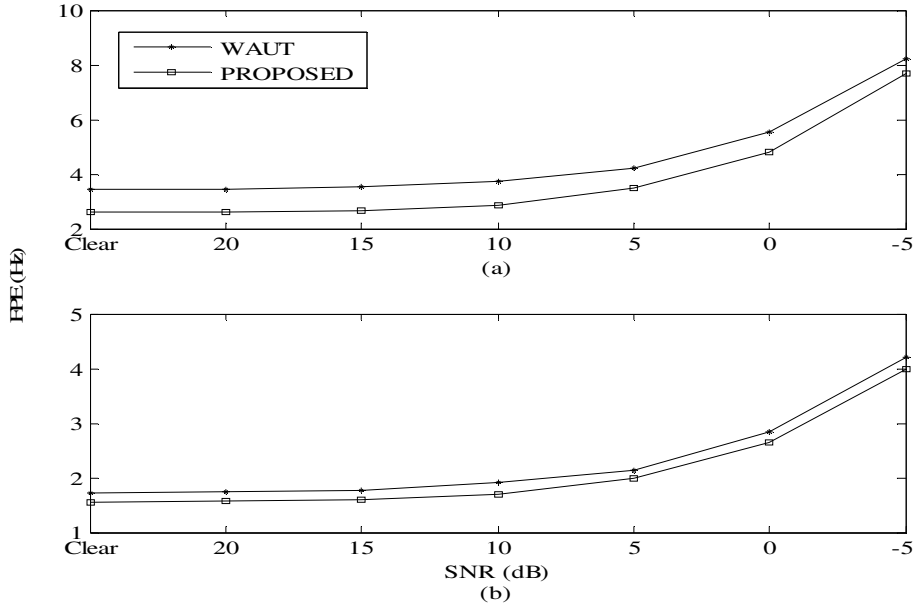

Figure 11. Comparison of average performance results in terms of mean fine pitch error (FPE) in exhibition noise for different speaker under various SNR conditions; (a) Female speaker, (b) Male speakers.

5. CONCLUSIONS

Perfect fundamental frequency estimation is a tricky problem in speech analysis particularly in noisy conditions. In this paper, we proposed a correlation based method by utilizing the autocorrelation function is weighted by the reciprocal of the YIN. The result of the proposed algorithm was evaluated and compared to the well known weighted autocorrelation method. The competitive values of mean FPEs also indicate the accuracy of fundamental frequency extraction by the proposed method can be a suitable candidate for extracting fundamental frequency information in different noises conditions with very low levels of SNR as compared with other related method. It was also shown that the results verified lower values of %GPEs which indicate that the proposed method is satisfactory to extract the fundamental frequency of voiced speech signals accurately in noisy conditions.

REFERENCES

[1] W. Hess (1983), Pitch Determination of Speech Signals. New York: Springer-Verlag.
[2] L. R. Rabiner and R. W. Schafer (2010), Theory and Applications of Digital Speech Processing. New York: Prentice Hall.
[3] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGibegak (1976), "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, pp. 399-418.
[4] L. R. Rabiner (1977), "On the use of autocorrelation analysis for pitch detection," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-25, no. 1, pp. 24-33.
[5] M. A. F. M. R. Hasan, and T. Shimamura (2012), "An Efficient Pitch Estimation Method Using Windowless and Normalized Autocorrelation Functions in Noisy Environment," International Journal of Circuits, Systems and Signal Processing, Issue 3, vol. 6, pp. 197-204.
[6] A. M. Noll (1967), "Cepstrum Pitch Determination," Journal of Acoust. Soc. Am., vol. 41, no. 2, pp. 293-309.
[7] S. Ahmadi, and A. S. Spanias (1999), "Cepstrum Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," IEEE Trans. Speech and Audio Processing, vol. 7, no. 3, pp. 333-338.
[8] M. A. F. M. R. Hasan, M. S. Rahman, and T. Shimamura (2012), "Windowless Autocorrelation Based Cepstrum Method for Pitch Extraction of Noisy Speech," Journal of Signal Processing, vol. 16, no. 3, pp. 231-239.
[9] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley (1974), "Average magnnitude difference function pitch extractor," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-22, pp.353-362.
[10] A. Cheveigne and H. Kawahara (2002), "YIN, a fundamental frequency estimation for speech and music," J. Acoustical Society of America, vol. 111, no. 4, pp. 1917-1930.
[11] M. A. F. M. R. Hasan, R. Yasmin, D. Das, M. S. Rahman (2014), "Correlation based pitch extraction method in speech signal" In Proc. The 9th International Forum on Strategic Technology (IFOST), pp.140-143.
[12] D. Talkin (1995), "A robust algorithm for pitch tracking (RAPT)," in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, pp. 496-518.
[13] M. K. Hasan, S. Hussain, M. T. Hossain and M. N. Nazrul (2006), "Signal reshaping using dominant harmonic for pitch estimation of noisy speech," Signal Processing, vol. 86, pp. 1010-1018.
[14] T. Shimamura and H. Kobayashi (2001), "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. Speech and Audio Processing, vol. 9, no. 7, pp. 727-730.
[15] NTT (1994), "Multilingual Speech Database for Telephometry," NTT Advance Technology Corp., Japan.

AUTHORS

Mirza A. F. M. Rashidul Hasan received the B. Sc. (Hons), M. Sc. and M. Phil. Degrees in Applied Physics and Electronic Engineering from University of Rajshahi, Bangladesh in 1992, 1993, and 2001 respectively. In 2006, he joined University of Rajshahi, Rajshahi, Bangladesh as a faculty member, where he is currently serving as an Associate Professor in the Department of Information and Communication Engineering. He was a visiting researcher at Waseda University, Japan from 2003 to 2004 and as a junior fellow of IWMI from 2006 to 2007. He received his Ph. D. degree in 2009 from the faculty of Applied Science and Technology, Islamic University, Kushtia, Bangladesh and also received his D. Engg. degree in 2012 from the Graduate School of Science and Engineering, Saitama University, Saitama, Japan. His current research interests are in digital signal processing and its applications to speech, image and communication systems.