Rajshahi University Journal of Science & Engineering **ISSN 2309-0952**
Vol. 43: 51-61, 2015

Fundamental Frequency Extraction of Noisy Speech Signals

Mirza A.F.M. Rashidul Hasan[*1], Rubaiyat Yasmin[1], Dipankar Das[1], M. M. Hoque[2], M. I. Pramanik[3] and M. S. Rahman[4]

[1]*Department of ICE, University of Rajshahi, Rajshahi 6205, Bangladesh.*
[2]*Department of CSE, Chittagong University of Engineering & Technology, Chittagong 4349, Bangladesh.*
[3]*Department of Information System, City University of Hong Kong, Hong Kong.*
[4]*Department of CSE, Shahjalal University of Science & Technology, Sylhet 3114, Bangladesh.*

*Corresponding Author: *mirza_iu@yahoo.com*

Abstract

In this paper, we proposed a correlation based method which is a new approach using the autocorrelation function is weighted by the reciprocal of the YIN and very useful for accurate fundamental frequency extraction. The autocorrelation function and also YIN is a popular measurement in estimating fundamental frequency in time domain. In our proposed method, instead of the original signal, we employ its center clipping signal for obtaining the autocorrelation function and this function is weighted by the reciprocal of the YIN for fundamental frequency detection. Comparative results on female and male voices in white and exhibition noise shows that the proposed method can detect fundamental frequency with better accuracy in terms of gross pitch errors as compared to other related methods.

Keywords: pitch extraction, domain, correlation, autocorrelation function, white noise, exhibition noise.

INTRODUCTION

Fundamental frequency estimation, also referred to as pitch detection, has been a popular research topic for many years and is still being investigated today. Accurate determination of pitch plays a vital role in acoustical signal processing and has a wide range of applications in related areas such as speech analysis synthesis, speech coding, speech recognition, speech enhancement, speech and speaker identification. In this connection, numerous pitch detection algorithms (PDAs) have been proposed in the literature [1], [2], [3]. In general, they can be broad categorized into three classes: time domain, frequency domain, and time frequency domain algorithms. Most of classical techniques perform satisfactorily with clean speech. However, performance improvement in noisy environments is still desired. For example, this is particularly true in speech enhancement systems, because in such systems the accuracy of pitch extraction is directly related with the quality of speech after the operations of enhancement. Also, speech communication systems often transmit pitch information. Thus, we need to extract the pitch of speech signals in practical noisy environments for most of the applications. Till now, unfortunately, we do not have a single method reliable and accurate for pitch extraction in noisy environments.

The time domain autocorrelation method appears to be one of the most popular PDAs for its simplicity, explanatory power, and physiological plausibility. Pitch extraction methods are classified into the following three categories: (i) waveform processing, (ii) spectral processing, and (iii) correlation processing. Correlation based processing is known to be comparatively robust against noise. The autocorrelation function method (ACF) is classified into correlation based category, and may be one which provides the best performance in noisy environments [4]. In [5], an integrated method for the ACF has been proposed. Correlation based processing also includes the YIN method [6]. This paper, we propose a new approach of pitch extraction method, which uses a center clipping autocorrelation function weighted by the inverse of YIN method. The characteristics of the YIN are very similar with those of the autocorrelation function. The YIN produces a valley, while the autocorrelation produces a peak. However, Both functions essentially have the same periodicity. The proposed method utilizes the feature that in a noisy environment, the noise components included in the autocorrelation function and YIN behave independently (and are uncorrelated each other). This feature will be validated in this paper. By such uncorrelated properties, the peak of the autocorrelation function is emphasized in a noisy environment when the autocorrelation function is combined with the inversed YIN. As a result, it is expected that the accuracy of pitch extraction for the ACF and also YIN is improved.

This paper is organized as follows. In Section II, we briefly describes the problem of some conventional time domain methods. The proposed PDA is presented in Section III. In Section IV, we verify the effectiveness of our method by comparing with other methods based on experimental results and we conclude our work in Section V.

PROBLEM DESCRIPTION

The autocorrelation function $L(\tau)$ of the speech signal $d(w)$ is generally defined as in (1)

$$L(\tau) = \frac{1}{W} \sum_{w=0}^{W-1} d(w)d(w+\tau) \qquad (1)$$

where W is the length of the underlying speech frame and τ is the lag number. In this paper we consider a clean speech signal which is plotted in Fig. 1(a). If $d(w)$ is periodic at pitch period T, $L(\tau)$ exhibits peak at $\tau = iT$, where $i = 0,1,2,3....$ As the value of τ increases, $L(\tau)$ tends to decrease which facilitates the use of the second peak (at $\tau = T$) for estimation of the pitch period. This function is illustrated in Fig. 1(b) for the clean speech signal. In clean signal the performance of ACF is better and this example agree that this method detected the pitch peak accurately. On the other hand, the average magnitude difference function (AMDF) is another type of autocorrelation analysis [7]. This method also better for pitch detection in clean speech. For example, the pitch peak is accurately estimated by the AMDF method in clean speech as shown in Fig. 1(c).

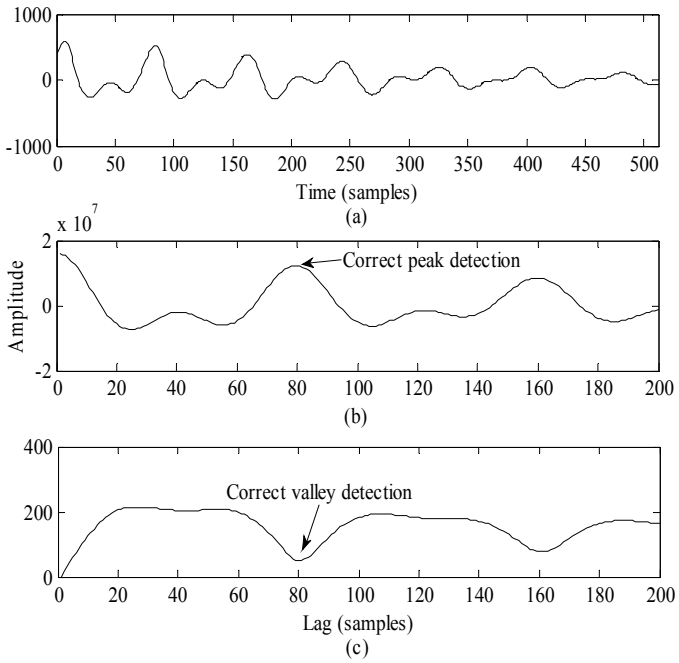

Fig. 1. (a) Clean speech frame, (b) Autocorrelation of clean speech in (a), and (c) AMDF of clean speech in (a).

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 AMDF obtained from the speech signal in Fig. 1(a), corrupted with white noise at signal to noise ratio (SNR) = 0 dB, is shown in Fig. 2, where both method fails to detect the true peak. Among many other improvements reported on the ACF method. Sondhi proposed a center clipping ACF based method [8]. Talkin proposed a normalized cross correlation based method [9]. Hasan proposed signal reshaping technique [10] for emphasizing the true peak. Shimamura proposed weighted autocorrelation (WAC) i.e., the ACF by the inverse average magnitude difference function [11].

The main shortcoming of WAC, however, is associated with the falling trend of minima in the AMDF. Dividing the ACF by such characteristic AMDF has the effect of enhancing the pitch peak in a lower proportion than the succeeding speaks and thereby making the algorithm vulnerable to double pitch error. To overcome this issue, we propose to utilize a correlation based processing i.e., the center clipping autocorrelation function is weighted by the reciprocal of the YIN.

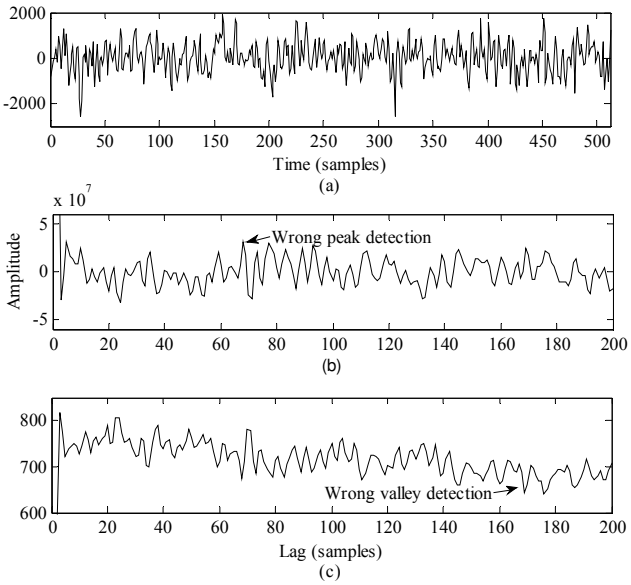

Fig. 2. (a) Noisy speech frame (SNR = 0 dB)(which is the same frame as Fig. 1(a)), (b) Autocorrelation of noisy speech in (a), and (c) AMDF of noisy speech in (a).

I. PROPOSED METHOD

The ACF is the correlation of waveform with itself. One would expect exact similarity at a time lag of zero, with increasing dissimilarity as the time lag increases. The definition of this functions as previously explain in (1). Let us assume that the *d(w)* signal is corrupted by additive white Gaussian noise, the noisy signal is given by

$$d(w) = e(w) + g(w) \tag{2}$$

for $w = 0,1,2,...,W\text{-}1$, where *e(w)* is a clean signal and *g(w)* denotes additive white Gaussian noise. In this case, we have an autocorrelation function given [11] by

$$\begin{aligned} L(\tau) &= \frac{1}{W}\sum_{w=0}^{W-1}(e(w)+g(w))(e(w+\tau)+g(w+\tau)) \\ &= \frac{1}{W}\sum_{w=0}^{W-1}((e(w)e(w+\tau)+e(w)g(w+\tau) \\ &\quad +g(w)e(w+\tau)+g(w)g(w+\tau)) \\ &= L_{ee}(\tau)+2L_{eg}(\tau)+L_{gg}(\tau) \end{aligned} \tag{3}$$

where $L_{ee}(\tau)$ is the ACF of $e(w)$, $L_{eg}(\tau)$ is the cross correlation of signal $e(w)$ and noise $g(w)$, and $L_{gg}(\tau)$ is the ACF of noise $g(w)$. For large W, if $e(w)$ does not correlate with $g(w)$, then $L_{eg}(\tau)=0$. Furthermore, if $g(w)$ is uncorrelated, then $L_{gg}(\tau) = 0$ except for $\tau = 0$. In such case, the relations

$$L(\tau) = L_{ee}(\tau) + L_{gg}(\tau)(\tau = 0)$$
$$L(\tau) = L_{ee}(\tau)(\tau \neq 0) \tag{4}$$

are valid. Based on these properties, the ACF provides robust performance against noise. The autocorrelation function with the period of T has 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 $\tau < T$. This situation, in some cases, leads to a double pitch error. For such reasons, if unnecessary peaks of ACF as shown in Fig. 2(b) are suppressed somehow, then it is expected that the accuracy of pitch extraction becomes higher.

For the purpose of emphasizing the true peak the ACF makes, we proposed a central clipping ACF weighted by an inversed YIN. Center clipping is the most popular spectrum flattening technique [8] and we used this technique in our proposed method. Center clipping technique can be expressed as

$$d'(w) = C_p\{d(w)\} = \begin{cases} (d(w) - C_p), & d(w) \geq C_p \\ 0, & |d(w)| < C_p \\ (d(w) + C_p), & d(w) \leq -C_p \end{cases} \tag{5}$$

where $d'(w)$ is the center clipping signal of speech signal $d(w)$ and C_p is the clipping level. A choice of C_p should be fulfill the following criterion:

- should be high enough to eliminate all distracting peaks, but
- cannot be too high so as not to lose desirable peaks.

On the other hand, YIN is based on the difference function, which attempts to minimize the difference between the waveform and its delayed duplicate instead of minimizing the product. The difference function is defined as in (6)

$$h(\tau) = \sum_{w=0}^{W-1} |d(w) - d(w+\tau)|^2 \tag{6}$$

The YIN has the characteristic that when $d(w)$ is similar with $d(w+\tau)$, $h(\tau)$ becomes small. This means that if $d(w)$ has a period of T, $h(\tau)$ produces a deep valley at $\tau = T$. Therefore, $1/h(\tau)$ makes a peak at $\tau = T$. Furthermore, the additive noise $g(w)$ included in $h(\tau)$ behaves independently with that included in $L(\tau)$ [11]. Hence using the center clipping ACF weighted by $1/h(\tau)$, it is expected that the true peak is emphasized, and as a result the errors of pitch extraction are decreased. The proposed method is given by

$$C(\tau) = \frac{L(\tau)}{h((\tau) + j)} \tag{7}$$

where j is a fixed number (j>0). The YIN is (6) provides at τ = 0, which invokes a divergence of the directly inversed YIN. For this reason, the denominator in (7) is stabilized by adding the number j.

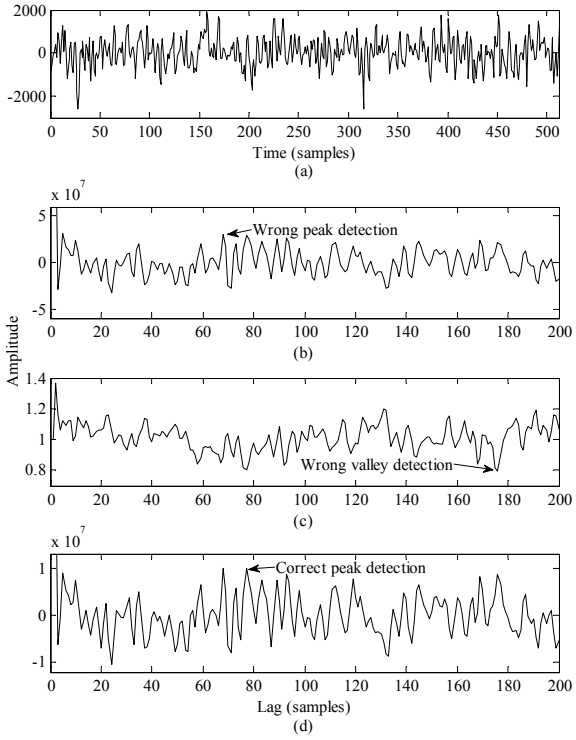

Fig. 3. (a) Noisy speech frame (SNR = 0 dB)(which is the same frame as Fig. 1(a)), (b) Autocorrelation of noisy speech in (a), (c) YIN of noisy speech in (a) and (d) Proposed of noisy speech in (a).

Fig. 3 shows the ACF, YIN and proposed methods obtained for a speech signal in Fig. 1(a) is corrupted by white noise (SNR = 0 dB). In this case, by picking the maximum amplitude of each function, the proposed method leads to the true pitch, while the ACF and YIN function does an erroneous one.

EXPERIMENTS AND RESULTS

To evaluate the proposed method, natural speech signals spoken by two Japanese female and male speakers are examined. Speech signals are sampled at a rate of 10 kHz, which are taken from NTT database [12]. The reference file is constructed by computing the pitch frequency every 10 ms using a semi automatic technique based on visual inspection. The simulations were performed after adding additive noise to these speech signals. For

the performance evaluation of the proposed method, criteria considered in our experimental work are: 1) gross pitch error (GPE); and 2) fine pitch error (FPE). The evaluation of accuracy of the extracted fundamental frequency is carried out by using

$$e(l) = F_t(l) - F_e(l) \tag{8}$$

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) [6], [10]. 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,

$$GPE(\%) = \frac{F_{GPE}}{F_v} \times 100 \tag{9}$$

The mean FPE is calculated by

$$FPE_m = \frac{1}{N_i} \sum_{j=1}^{N_i} e(l_j) \tag{10}$$

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 (%), and FPE_m provide a good description of the performance of a pitch estimation method. The experimental conditions are tabulated in Table I.

TABLE I. CONDITION OF EXPERIMENTS

Sampling frequency	10 kHz
Band limitation	3.4 kHz
Window function	Rectangular
Window size	51.2 ms
Frame shift	10 ms
Number of FFT points	2048
SNRs (dB)	∞, 20, 15, 10, 5, 0

We attempt to extract the pitch information of clean and noisy speech. All the candidate algorithms are applied in additive white Gaussian noise and exhibition noise. The noises are taken from the Japanese Electronic Industry Development Association (JEIDA) Japanese Common Speech Corporation. The performance of the proposed method (PROPOSED) is compared with a well known ACF, YIN and WAC method. In order to evaluate the pitch estimation performance of the proposed method, we plot a reference pitch contour for noisy speech in white noise speech of a female speaker from the reference database and also the pitch contours obtained from the four pitch estimation methods in Fig. 4.

Fig. 4 shows that in contrast to the other three methods, the proposed method yields a relatively smoother pitch contour even at an SNR of 0 dB. Fig. 5 shows a comparison of the pitch contour resulting from the four methods for the male speech corrupted by the

white noise at an SNR of 0 dB. In Fig. 5 it is clear that the proposed method is able to give a smoother contour even in the presence of white noise.

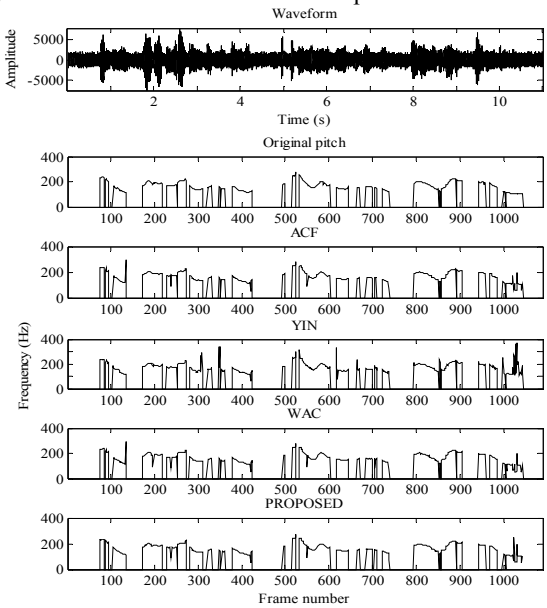

Fig. 4. Pitch contours extracted by ACF, YIN, WAC, and PROPOSED methods for female speech with white noise (SNR=0 dB).

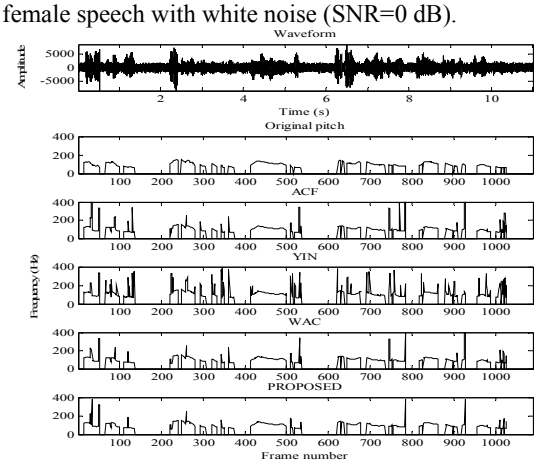

Fig. 5. Pitch contours extracted by ACF, YIN, WAC, and PROPOSED methods for male speech with white noise (SNR=0 dB).

The pitch contours in Figs. 4 and 5 obtained from the four methods have convincingly demonstrated that the proposed method is capable of reducing the double and half pitch

errors thus yielding a smooth pitch track. The number of GPEs found in determining the pitch using ACF, YIN, WAC and PROPOSED are summarized and after that the percentage average number of GPEs for female and male speakers are shown in Fig. 6 and 7, respectively.

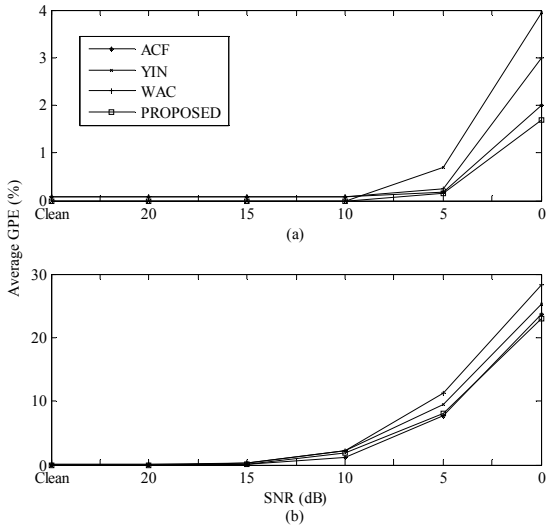

Fig. 6. Average performance results in terms of percentage of gross pitch error for female speakers in (a) white noise, (b) exhibition noise at various SNR conditions.

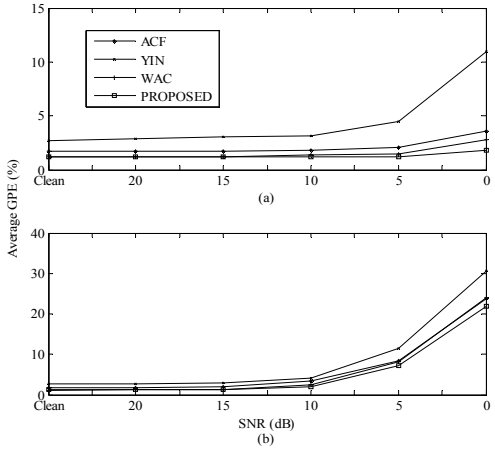

Fig. 7. Average performance results in terms of percentage of gross pitch error for male speakers in (a) white noise, (b) exhibition noise at various SNR conditions.

In all type speaker cases the proposed method gives better results than the other methods in different types of SNR conditions. Especially at SNR = 5 dB and SNR = 0 dB, the

proposed method gives far better results than the other methods. From these Figs., it is obvious that the proposed method outperforms the other related methods in different types of SNR conditions. In summary, the proposed method can estimate stable pitch with high accuracy not only in a noise free environment but also in heavy noisy conditions.

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 Figs. 8 and 9, the FPE values resulting from the four methods are plotted, respectively.

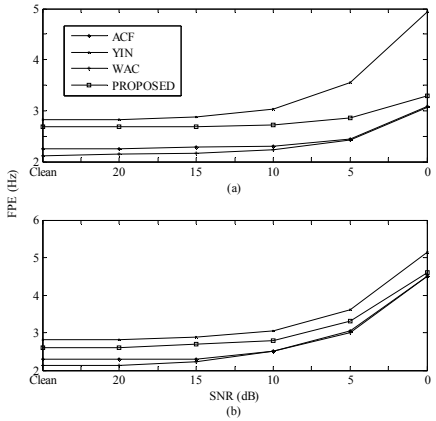

Fig. 8. Comparison of average performance results in terms of mean fine pitch error for female speakers in different noises: (a) white, and (b) exhibition noises at various SNR conditions.

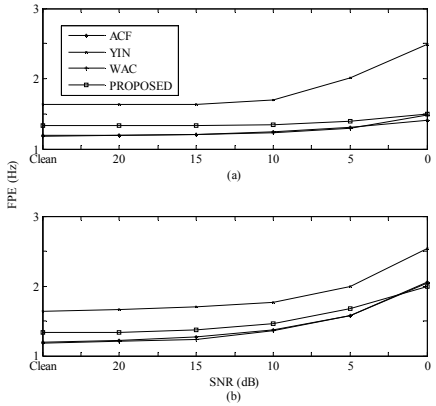

Fig. 9. Comparison of average performance results in terms of mean fine pitch error for male speakers in different noises: (a) white, and (b) exhibition noises at various SNR conditions.

Average FPEs for all methods range approximately from 1.0 Hz ~ 5.2Hz. From the simulation results it is found that the value of FPEs is also within the acceptable limit and consistently satisfactory.

CONCLUSION

Perfect fundamental frequency estimation is a tricky problem in speech analysis particularly in noisy environments. In this paper, we proposed a correlation based method by utilizing the center clipping autocorrelation function is weighted by the reciprocal of the YIN. Simulation results indicate that the proposed method provides better performance in terms of GPE (in percentage) compared with the existing methods such as ACF, YIN, and WAC for a wide range of SNR varying from 0 dB to ∞ dB. Especially the performance of the proposed method in low SNR cases is noticeable higher both in white and exhibition noise cases than that of the ACF, YIN, and WAC based methods. The competitive value of mean FPEs also indicate the accuracy of pitch extraction by the proposed method. These results suggest that the proposed method can be a suitable candidate for extracting pitch information both in white and color noise conditions with low levels of SNR as compared with other related methods.

REFERENCES

[1] W. Hess, Pitch Determination of Speech Signals. New York: Springer-Verlag, 1983.

[2] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing. New York: Prentice Hall, 2010.

[3] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGibegak, "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, pp. 399-418, 1976.

[4] K. A. Oh and C. K. Un, "A performance comparison of pitch extraction algorithms for noisy speech," in Proc. IEEE int. Conf. Acoustics, Speech, Signal Processing, pp. 18B4.1-18B4.4, 1984.

[5] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-25, no. 1, pp. 24-33, Feb. 1977.

[6] A. Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimation for speech and music," J. Acoustical Society of America, vol. 111, no. 4, pp. 1917-1930, Apr. 2002.

[7] M. J. Ross, H. L. Schafer, A. Cohen, R. F. B, and H. Manley, "Average magnitude difference function pitch extraction," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-22, no. 5, pp. 353-362, 1974.

[8] M. M. Sondhi, "New methods of pitch extraction," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 262-266, 1968.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995, pp. 496-518.

[10] M. K. Hasan, S. Hussain, M. T. Hossain and M. N. Nazrul, "Signal reshaping using dominant harmonic for pitch estimation of noisy speech," Signal Processing, vol. 86, pp. 1010-1018, May 2006.

[11] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. Speech and Audio Processing, vol. 9, no. 7, pp. 727-730, Oct. 2001.

[12] NTT, "Multilingual Speech Database for Telephometry," NTT Advance Technology Corp., Japan, 1994.