International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 45 No. 1 June 2024, pp. 36-47

Modified Method for Fundamental Frequency Detection of Voiced/Unvoiced Speech Signal in Noisy Environment

Md. Arifur Rahman [(1)], Md. Mahfuz Alam[(1)], Md. Firoz Ahmed[(1)], and M. A. F. M. Rashidul Hasan[* (1)]

[(1)]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.
Email : *mirzahasanice@gmail.com
uzzalruapee@gmail.com; mahfuzalam5611@gmail.com; firozice01@gmail.com

Abstract – An efficient fundamental frequency detection method is introduced in this paper. The method is based on time domain fundamental frequency detection method. In our proposed method, instead of the original speech signal, we employ its center clipping signal for obtaining the modified autocorrelation function and this function is weighted by the reciprocal of the average magnitude difference function for fundamental frequency detection. The performance of the proposed fundamental frequency detection method is compared in terms of gross pitch error and fine pitch error with the other related method. A comprehensive evaluation of the fundamental frequency estimation results on female and male voices in white noise show the superiority of the proposed method over three related method under low levels of signal to noise ratio (SNR).

Keywords – Fundamental frequency, Pitch, Center Clipping, White Noise.

I. INTRODUCTION

Fundamental frequency period is the important parameters of speech recognition and speech synthesis. fundamental frequency detection has been focus in the field of audio processing research. The speech signal can be classified into two general categories, voiced and unvoiced speech. A voice sound is one in which the vocal cords of the speaker vibrate as the sound is made, and unvoiced sound is one where the vocal cords do not vibrate. The fundamental frequency has great importance in many areas. Fundamental frequency is one of the oldest, yet unsolved topic among the researchers of speech and music [1, 2]. Accurate fundamental frequency detection is essential to areas such as automatic speech recognition, speaker identification, low-bit rate coding, speech enhancement using harmonic model, speech synthesis, and to more recent topic of speaker emotion recognition etc. [3, 4, 5, 6]. Recently many Fundamental frequency estimation algorithms have been proposed, but accurate and efficient fundamental frequency estimation is still a challenging task.

There are three types of fundamental frequency detection algorithms (FDAs) in the literature: time domain [7, 8], frequency domain [9, 10], and time-frequency domain [11]. Due to the extreme importance of accurate fundamental frequency detection problem, the strengths of different FDAs have been explored [12, 13], and several fundamental frequency reference databases have been developed to facilitate fair comparison of different FDAs on a common platform [14]. Among the reported method, the time domain method i.e., autocorrelation function (ACF) [7] based approaches are very popular for their simplicity, low computational complexity, and good performance in the presence of noise. The ACF is, however, 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 under noisy or even under noiseless conditions. This sometimes makes true peak selection a difficult task. This motivates researches to propose numerous modifications on the ACF method. One significant improvements are proposed in Sondhi [15] used center clipping ACF. On the other hand, a well known method, average magnitude difference function (AMDF) has the advantage of low computation and high precision and the calculation cost needed less than that of ACF [16]. But when the

Corresponding Author: M. A. F. M. Rashidul Hasan

magnitude or the fundamental frequency period of speech signal changes rapidly, AMDF method will decreased apparently in fundamental frequency estimation accuracy. Correlation based processing method is known to be comparatively robust against noise. In this paper, we proposed a correlation based fundamental frequency detection method, where the centre clipping signal is used for modified ACF (MACF) and this MACF is weighted by the inverse of an AMDF [17]. The proposed method utilizes the feature that in a noisy environment, the noise components included in the MACF and AMDF behave independently. By such type of uncorrelated properties, the peak of the MACF is emphasized when the MACF is combined with inversed AMDF [18].

The paper is organized as follows: Section 2 describes some basic fundamental frequency detection algorithms that include time domain processing. Section 3 presents the proposed fundamental frequency detection algorithm, and Section 4 gives experimental results. Finally, the paper is concluded in Section 5.

II. Fundamental Frequency Detection Algorithms

Fundamental frequency is the lowest frequency component of a signal that excites to a system i.e., vocal system. The fundamental frequency is the smallest repeating unit of a signal. One such period describes the periodic signal (i.e., voiced part of speech) completely. The fact that variations in voiced signal are so evident suggests that the time domain method should be capable in detecting fundamental frequency period of a voiced signal. Most of the time domain fundamental frequency period estimation methods use ACF.

Let $x(n)$ and $v(n)$ denote speech signal and uncorrelated white Gaussian noise with zero mean and variance σ^2_v, respectively. Therefore, the noisy signal $y(n)$ is then given by

$$y(n) = x(n) + v(n) \tag{1}$$

Based on the assumption that speech and noise are uncorrelated, the ACF $R_{yy}(k)$ of $y(n)$ can be expressed as

$$R_{yy}(k) = \begin{cases} R_{xx}(k) + \sigma_v^2 & for\ k = 0, \\ R_{xx}(k) & for\ k \neq 0, \end{cases} \tag{2}$$

where $R_{xx}(k)$ is the ACF of the clean speech signal $x(n)$ estimated as

$$R_{xx}(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+k) \tag{3}$$

here N is the total number of samples in a window of the speech under analysis and k is the lag index. The choice of window length N for calculating $R_{xx}(k)$ has conflicting requirements:

- N should be as small as possible to show time variation;
- N should be large enough to cover at least 2 periods so that periodicity can be captured by $R_{xx}(k)$.

In Eq. (3), $R_{xx}(k)$ essentially exhibits peaks at the periodicity (T) of $x(n)$ (i.e., at $k = lT$, where l is an integer). The basic idea behind the ACF based methods is to use the location of the second largest peak (at $k = T$) relative to the largest peak (at $k = 0$) to obtain an estimate of the pitch period (Figure 1). The main advantage of ACF method is its noise immunity. However, it effects the formant structure which result in the loss of a clear peak in $R_{xx}(k)$ at the true pitch period. The performance of the conventional ACF method is significantly degraded at low SNR (Figure 2).

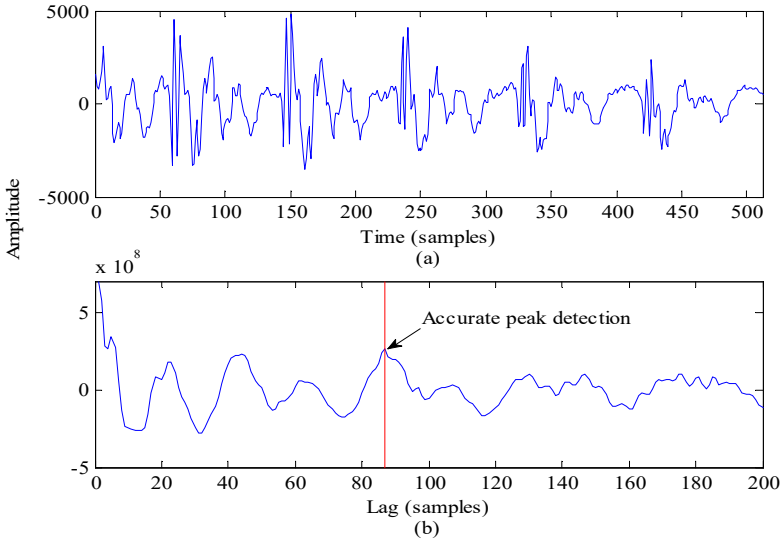


Figure 1. (a) Clean speech signal of a male speaker, (b) ACF of signal in (a). The vertical line indicates the correct fundamental frequency value.

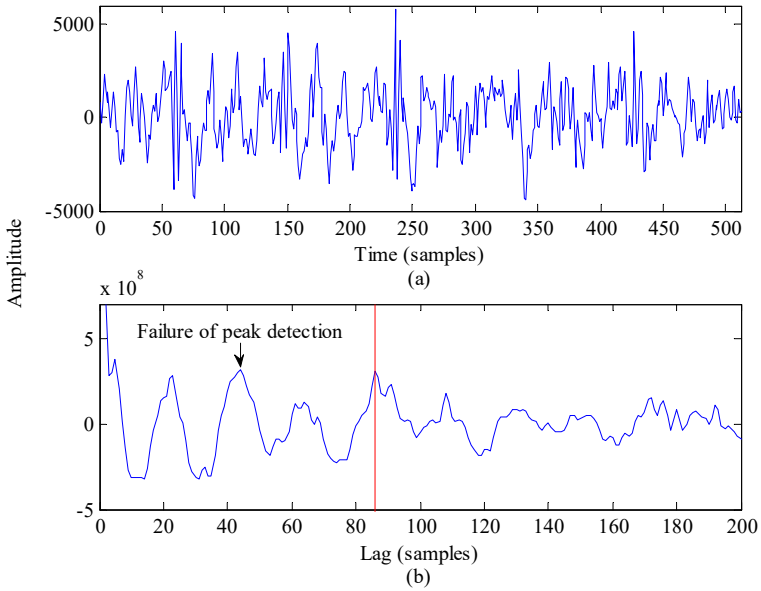


Figure 2. (a) Noisy speech signal of male speaker (which is the same frame as Figure 1(a)) at an signal to noise ratio of -5dB, (b) ACF of signal in (a). The vertical line indicates the correct fundamental frequency value.

The average magnitude difference function (AMDF) is another type of autocorrelation analysis. Instead of correlating the input signal at various delays, a difference signal is formed between the delayed signal and original, and at each delay value the absolute magnitude is taken. The AMDF is describe by

$$\xi_{xx}(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left|x(n) - x(n+k)\right| \qquad (4)$$

where $x(n+k)$ are the samples time shifted on k samples. The difference function is expected to have a strong local minimum if the lag k is equal to or very close to the fundamental frequency. AMDF has advantage in relatively low computational cost and simple implementation. Unlike the autocorrelation function, the AMDF calculations require no multiplications. This is a desirable property for real time applications. For each value of delay, computation is made over an integrating window of N

samples. The fundamental frequency period is identified as the value of the lag at which the minimum AMDF occurs (Figure 3). This algorithm has many advantages, but the probability of double misjudge and half misjudge is very high when noise is added (Figure 4).

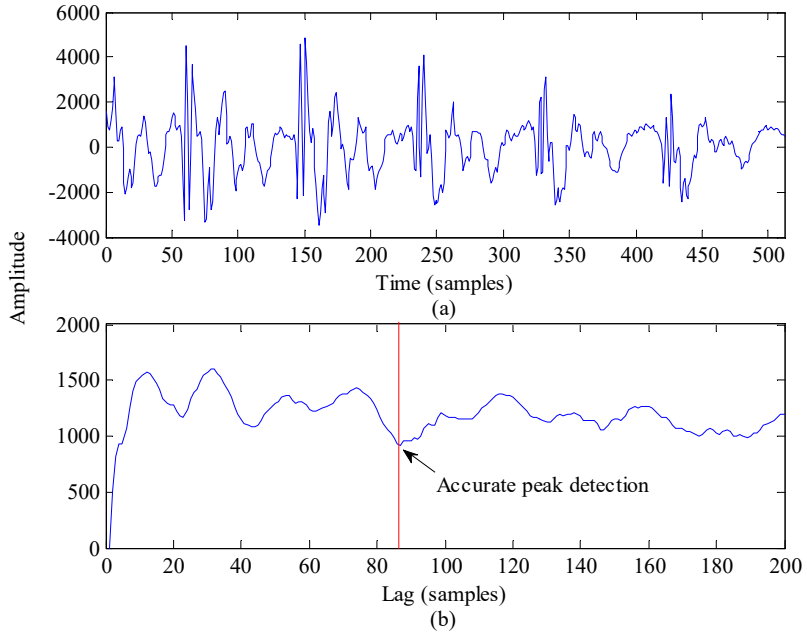


Figure 3. (a) Clean speech signal of a male speaker (which is the same frame as Figure 1(a)), (b) AMDF of signal in (a). The vertical line indicates the correct fundamental frequency value.

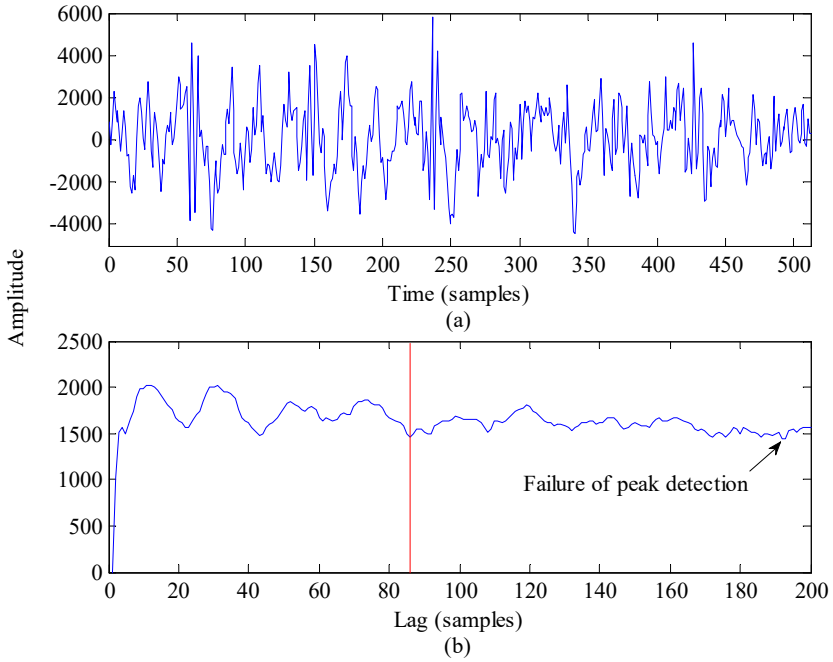


Figure 4. (a) Noisy speech signal of a male speaker (which is the same frame as Figure 1(a)), (b) AMDF of signal in (a). The vertical line indicates the correct fundamental frequency value.

III. PROPOSED METHOD

The MACF weighted by the inverse of an AMDF is used for fundamental frequency extraction and is defined as

$$\varphi_{xx}(k) = \frac{R_{xx}(k)}{\xi_{xx}(k) + \delta} \tag{5}$$

where $R_{xx}(k)$ and $\xi_{xx}(k)$ denotes the MACF and AMDF of signal $x(n)$ respectively, δ is a small positive constant. We consider the calculation procedures of power spectral density of conventional ACF are increased in MACF. It is expected to give maximum peak at $k = nT$ (MACF) & deep notches at $k = nT$ (AMDF), and therefore the true fundamental frequency peak in $\varphi_{xx}(k)$ is emphasized (Figure 5). But the main limitation of this method is that, it is very sensitive to the half or double pitch error in noisy case as shown in Figure 6.

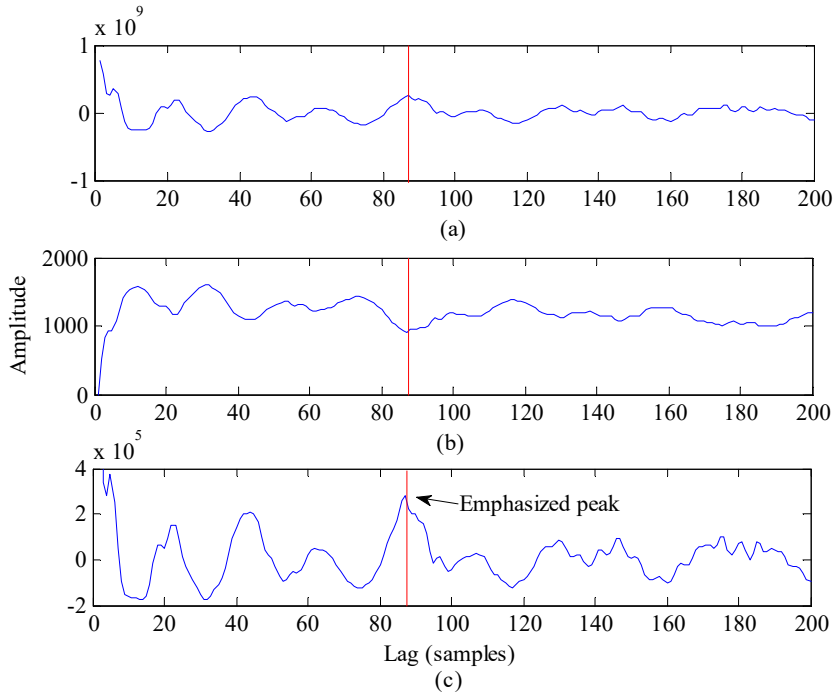


Figure 5. Fundamental frequency peak detection in clean speech signal (which is the same frame as Figure 1(a)) using (a) ACF method, (b) AMDF method, (c) Weighted autocorrelation function method. The vertical line indicates the correct fundamental frequency value.

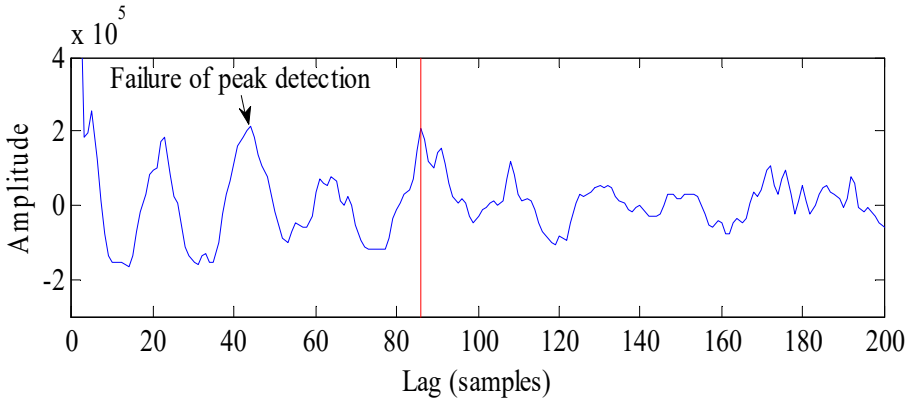


Figure 6. Fundamental frequency peak detection in noisy speech signal (which is the same frame as Figure 1(a)) using Weighted autocorrelation function method. The vertical line indicates the correct fundamental frequency value.

For fundamental frequency detection, speech signal is usually pre-processed to make the periodicity more prominent and to suppress other distracting features. Such techniques are often called spectrum flattening. Center clipping is the most popular spectrum flattening technique [15] and we used this technique in our proposed method. Center clipping technique can be expressed as

$$x'(n) = C_L\{x(n)\} = \begin{cases} (x(n) - C_L), & x(n) \geq C_L \\ 0, & |x(n)| < C_L \\ (x(n) + C_L), & x(n) \leq -C_L \end{cases} \quad (6)$$

where $x'(n)$ is the center clipping signal of speech signal $x(n)$ and C_L is the clipping level. A choice of C_L 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.

In our proposed method, instead of the speech signal $x(n)$, we employ its center clipping signal $x'(n)$ for obtaining the MACF and using this MACF weighted by $1/\xi_{xx}(k)$. It is expected that the true peak is more emphasized (Figure 7), and as a result the errors of fundamental frequency extraction are decreased. The correlation based proposed method is given by

$$\varphi_{xx\_cl}(k) = \frac{R_{xx\_cl}(k)}{\xi_{xx}(k) + \delta} \quad (7)$$

where $R_{xx\text{-}cl}(k)$ is the MACF of signal $x'(n)$.

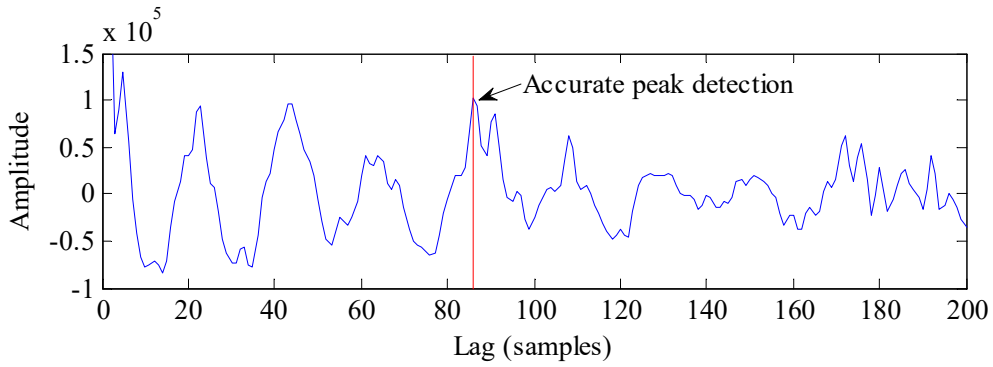


Figure 7. Fundamental frequency peak detection in noisy speech signal (which is the same frame as Figure 1(a)) using Proposed method. The vertical line indicates the correct fundamental frequency value.

IV. Experimental Results

To assess the proposed method, natural speeches spoken by three Japanese male and three Japanese female speakers are examined. Speech materials are 11 sec-long sentences spoken by every speaker sampled at 10 kHz rate, which are taken from NTT database [19]. The reference file of the fundamental frequency of speech is constructed by computing the fundamental 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, a criterion considered in our experimental work is gross pitch error (GPE). The evaluation of accuracy of the extracted fundamental frequency is carried out by using

$$e(l) = F_t(l) - F_e(l) \quad (8)$$

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) [20,21,22]. 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,

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

The mean FPE is calculated by

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

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

Table I. Experimental Parameter Specification

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	1024
SNRs (dB)	∞, 20, 15, 10, 5, 0, -5

We attempt to extract the pitch information of clean and noisy speech signals. 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 is compared with a well-known method, AMDF [16], weighted autocorrelation method, WACF [18], and conventional method, CEP [9]. For the implementation of the WACF, the parameter k in [18] is set to 0.1 and for proposed method, the parameter C_L is set to 10% of the maximum magnitude of signal. As the fundamental frequency range is known to be 50-500 Hz for most male and female speakers and our sampling frequency is 10 KHz, the setting of lag number (i.e., 200) is commonly used for the AMDF, WACF, CEP and the proposed method. 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 8.

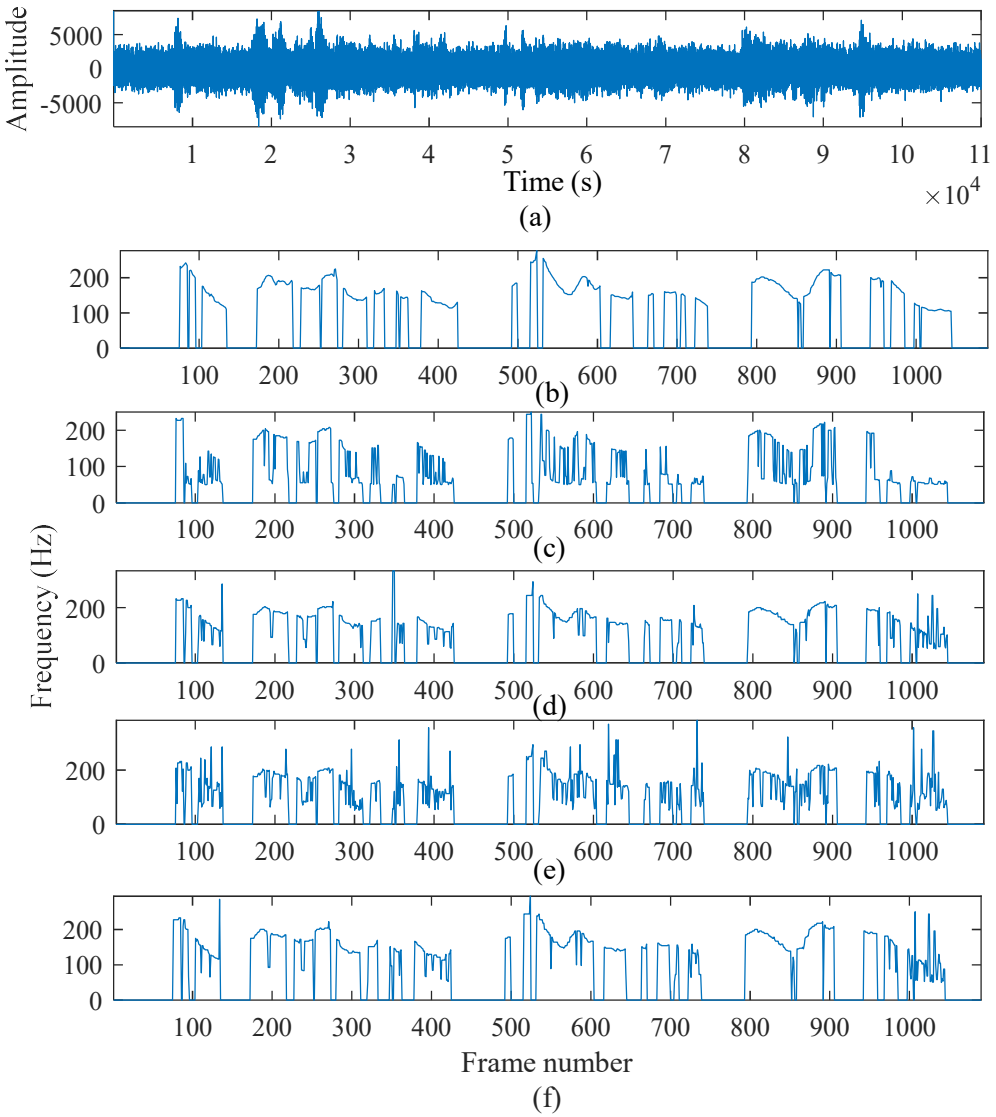
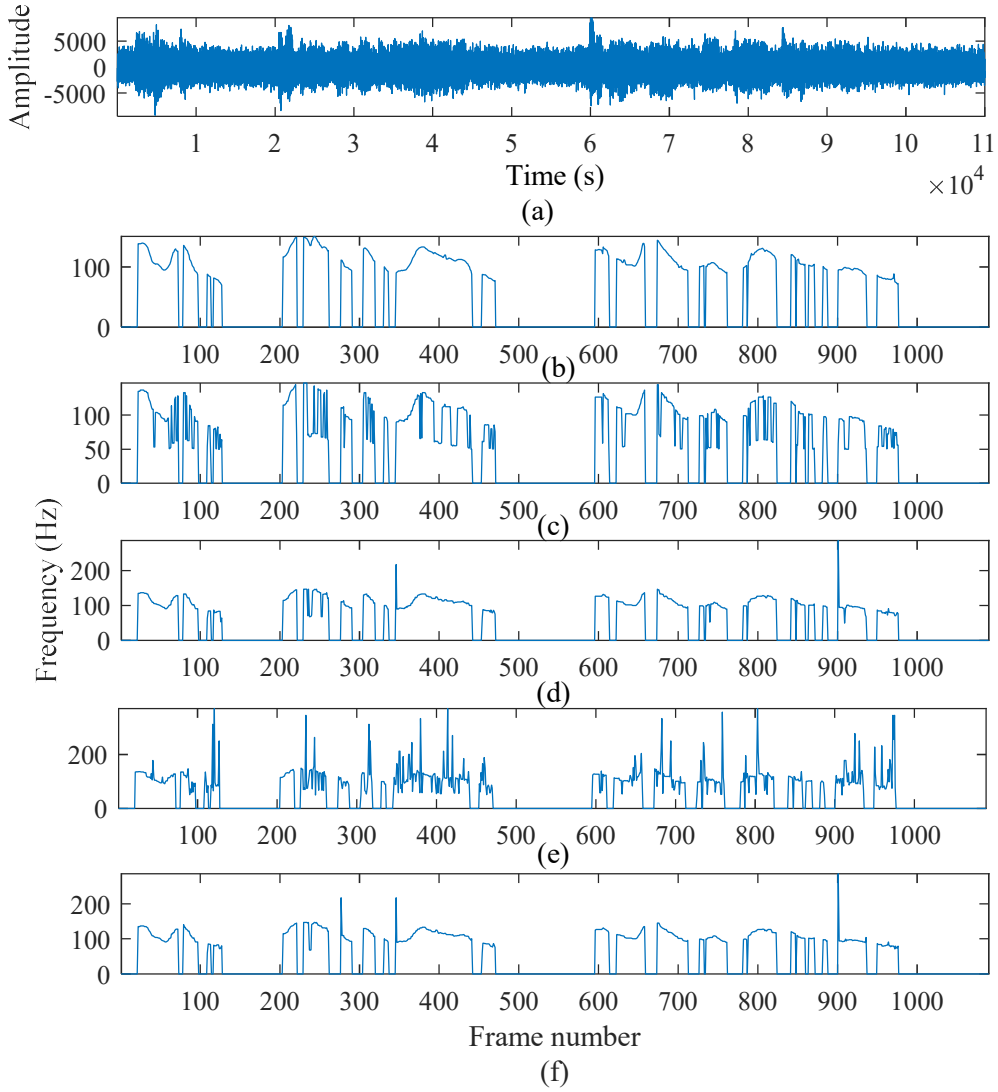


Figure 8. (a) Noisy speech signal for female speaker in white noise at an SNR -5 dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) AMDF (d) WACF, (e) CEP and

(f) Proposed method.

Figure 8 shows that in contrast to the other method, the proposed method yields a relatively smoother pitch contour even at an SNR of -5 dB. Fig. 9 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 -5 dB. In Fig. 9 it is clear that the proposed method is able to give a smoother contour even in the presence of white noise. The pitch contours in Figs. 8 and 9 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.



Figur 9. (a)) Noisy speech signal for male speaker in white noise at an SNR -5 dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) AMDF (d) WACF, (e) CEP and

(f) Proposed method.

Fundamental frequency estimation error in percentage, which is the average of GPEs for white noise, is shown in Figure 10. These figures 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 AMDF, WACF and CEP methods in almost all cases. Particularly, at low SNR (0 dB, -5 dB), the proposed method performs more robustly compared with the other three methods.

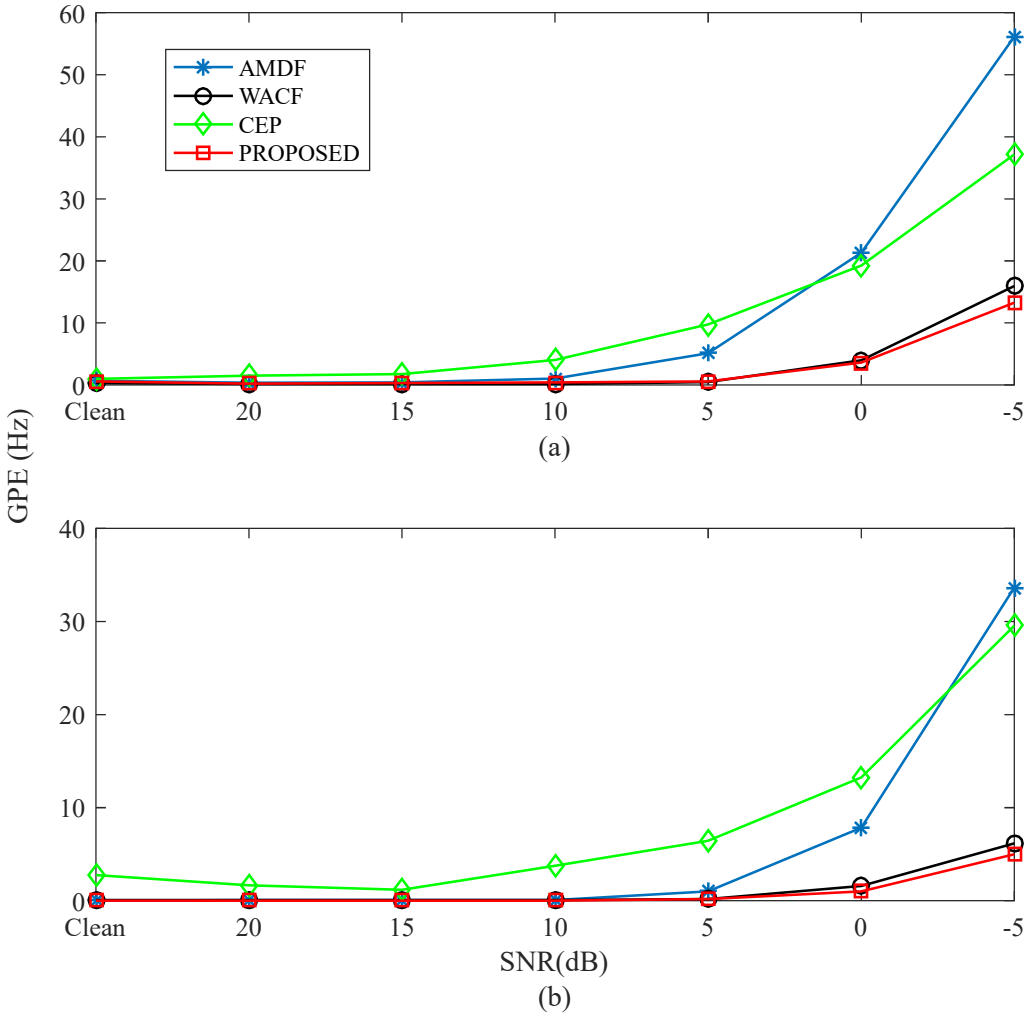


Figure 10. Percentage of average gross pitch error (GPE) in white noise for different speakers under various signal to noise ratio conditions; (a) Female speakers, (b) Male speakers.

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. 11, the FPE values resulting from the four methods are plotted, respectively. Average FPEs for all methods range approximately from 1 Hz ~ 9.2Hz. From the simulation results it is found that the value of FPEs is also within the acceptable limit and consistently satisfactory at other SNRs.

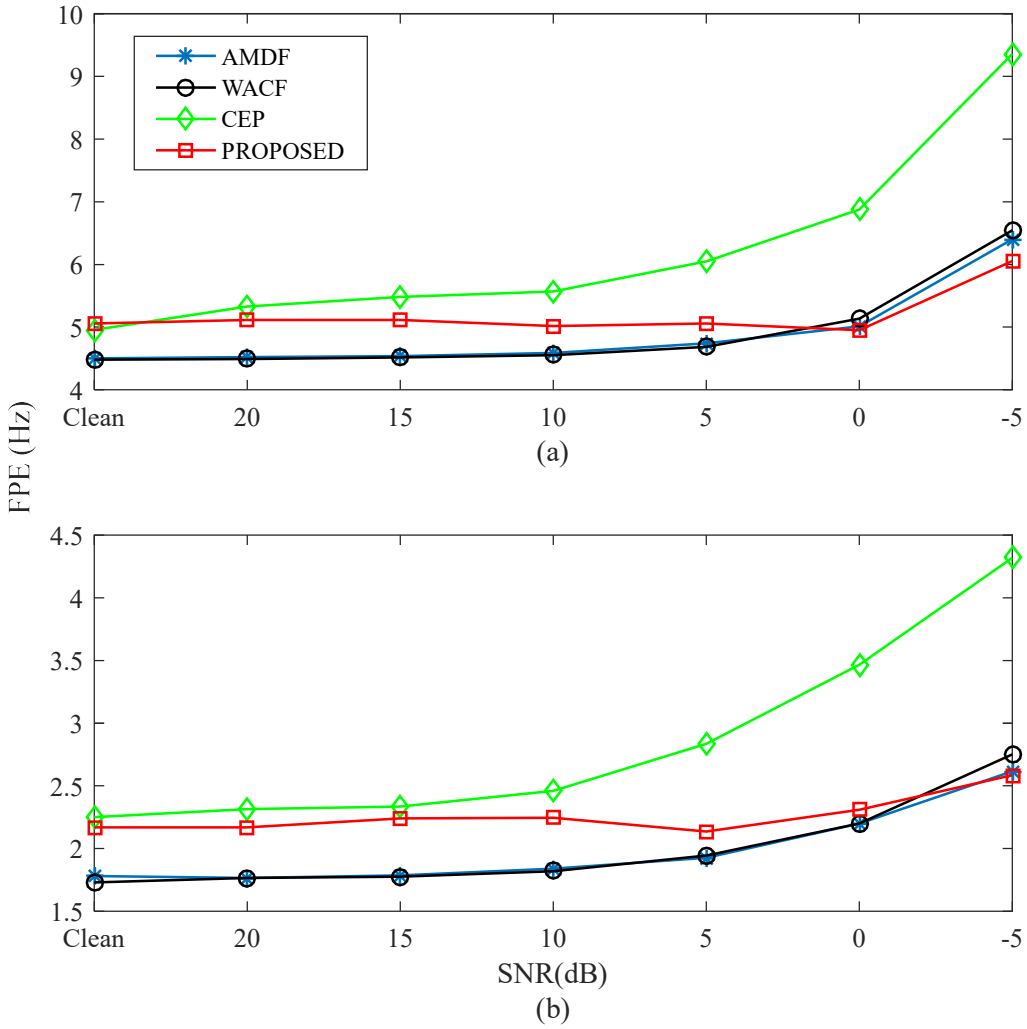


Figure 11. Comparison of average performance results in terms of mean fine pitch error (FPE) for different speakers under various signal to noise ratio conditions; (a) Female speakers, (b) Male speakers.

V. Conclusion

In this paper, an efficient fundamental frequency estimation using correlation-based method was introduced which leads to robustness against white noise. Simulation results indicate that the proposed method provides better performance in terms of GPE (in percentage) compared with the existing method such as AMDF, WACF, and CEP. The competitive values 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 in different noises conditions with very low levels of SNR as compared with other related method.

References

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

[2] Rabiner L. R., and Schafer R. W., Theory and Applications of Digital Speech Processing, 1st ed., Prentice Hall, 2010.

[3] Beigi H., Fundamental of Speaker Recognition, Springer, 2011.

[4] Rosenberg A. E., and Sambur M. R., "New Techniques for Automatic Speaker Verification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 2, pp. 169-176, 1975.

[5] Tamura M., Masuko T., Takuda K., and Kobayashi T., "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR", In Proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'01), pp. 805-808, 2001.

[6] Razak A. A., Abidin M. I. Z., and Komiya R., "Emotion pitch variation analysis in Malay and English voice samples", In Proc. 9th Asia-Pacific Conference on Communications (APCC'03), vol. 1, pp. 108-112, 2003.

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

[8] Hasan M. A. F. M. R., and Shimamura T., "An efficient pitch estimation method uning windowless and normalized autocorrelation functions in noisy environment," *International Journal of Circuits, Systems and Signal Processing*, Issue 3, vol. 6, pp. 197-204, 2012

[9] Noll A. M., "Cepstrum pitch determination," *Journal of Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293-309, 1967.

[10] Ahmadi S., and Spanias A. S., "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, 1999.

[11] Hasan M. A. F. M. R., Rahman M. S., and Shimamura T., "Windowless autocorrelation based Cepstrum method for pitch extraction of noisy speech," *Journal of Signal Processing*, vol. 16, no. 3, pp. 231-239, 2012.

[12] Rabiner L. R., Cheng M. J., Rosenberg A. M., and McGonegal C. A., "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 5, pp. 399-417, 1976.

[13] Veprek P., Scordilis M. S., "Analysis, enhancement and evaluation of five pitch determination techniques," *Speech Communication*, vol. 37, pp. 249-270, 2002.

[14] Plante F., Meyer G., and Ainsworth W. A., "A pitch extraction reference database", In Proc. EUROSPEECH, pp. 837-840, 1995.

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

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

[17] Rumy S. A., Hasan M. A. F. M. R., Yasmin R., and Rahman M. S., "A method for pitch detection of speech signal in noisy environment," In Proc. 1st National Conference on Intelligent Computing and Information Technology, pp. 90-94, 2013.

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

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

[20] Cheveigne A., and Kawahara H., "YIN. a fundamental frequency estimation for speech and music," *Journal of Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917-1930, 2002.

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

[22] Mirza A. F. M. Rashidul Hasan, "A pitch detection algorithm based on windowless autocorrelation function and modified cepstrum method in noisy environments", *International Journal of Computer Science and Network Security (IJCSNS)*, vol.17, no. 2, pp. 106-112, 2017.