A Pitch Detection Algorithm Based on Windowless Autocorrelation Function and Modified Cepstrum Method in Noisy Environments

Mirza A. F. M. Rashidul Hasan

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

Summary
This paper proposes a new pitch detection algorithm of speech signals in noisy environment. The performance of the cepstrum method is effected due to the formant effect and the presence of spurious peaks introduced in noisy condition. In our proposed method, we firstly employ windowless autocorrelation function instead of its speech signal for obtaining the cepstrum. The windowless autocorrelation function is a noise-reduced version of the speech signal where the periodicity is more apparent with enhanced pitch peak. Secondly the modified cepstrum method is applied to windowless autocorrelation function which utilizes clipping and band pass filtering operation on log spectrum. The performance of the proposed pitch detection method is compared in terms of gross pitch error with the other related methods. Experimental results on male and female voices in white and color noises shows the superiority of the proposed method over some of the related methods under low levels of signal to noise ratio.

Key words:
Pitch Detection; Cepstrum; Windowless Autocorrelation Function; White Noise; Color noise

1. Introduction

The pitch detection (fundamental frequency, F0) is a critical problem in the acoustic characterization of speech signal [1]. Accurate pitch detection plays an important role in speech processing and has a wide spread of applications in speech related systems. For example, it is found in speech communications [2], automatic speaker recognition [3], analysis of speech perception [4], and in the assessment of speech disorders [5]. For this reason, recently many numerous methods to detect the pitch of speech signals have been proposed but accurate and efficient pitch detection is still a challenging task [6, 7]. The speech signal is not always strongly periodic and the presence of noise generates a degraded performance of pitch detection algorithms. Numerous methods have been proposed in the literature to address this problem. In general, they can be categorized into three classes: time-domain, frequency-domain, and time-frequency domain algorithms. Due to the extreme importance of the problem, the strength of different methods has been explored [8].

A large number of pitch detection algorithms perform satisfactorily with clean speech. Among the reported methods, the autocorrelation based approaches are very popular for their simplicity, low computational complexity and better performance in noise. The autocorrelation function (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 in noisy or even in noiseless conditions. This sometimes makes true peak selection a difficult task. This motivates the researchers to propose numerous modifications on the ACF method. Some significant improvements are proposed in [9-13]. Takagi et al. [9] used ACFs from multiple windows, Shimamura et al. [10] weighted ACF by average magnitude difference function, YIN [11] used a difference function, Talkin [12] used a normalized cross correlation based method and Hasan et al. [13] reshaped the signal to emphasize the true pitch peak. The methods are successful in white Gaussian noise, but robustness against formant structure is still not achieved. Markel [14] and Itakura et al. [15] utilized auto-regressive (AR) inverse filtering to flatten the signal spectrum. This AR preprocessing step has effects on emphasizing the true period peaks in ACF. However, for high-pitched speech or in white Gaussian noise, the process of AR estimation is itself erroneous [16]. Further, though color noise are also encountered in practice, performance improvement of the above ACF based methods in color noise is not satisfactory.

The cepstrum (CEP) method is one of the traditional methods to detect the pitch, which makes use of spectral characteristics of speech signals. The CEP method is able to accurately detect the pitch with little affections of vocal tract [17]. It can detect an accurate pitch of clean speech signal, but is not effective in noisy environments. Though the CEP method is sensitive to additive noise, empirically it is seen that the CEP method performs relatively better in color noise than other classical methods do. It is therefore expected that if the robustness of the CEP method against additive noise can be improved, it can be very useful in pitch determination. Toward this end, Andrews et al. proposed a subspace based method to reduce noise effects

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

[18], and Kobayashi et al. discussed a modified cepstrum (MCEP) method for pitch extraction [19]. Ahmadi et al. derived a statistical approach to improve the performance of the CEP method [20].

In this paper we propose a pitch detection method that utilizes the windowless ACF of the signal instead of the signal itself [21] and modified cepstrum method [19]. The windowless ACF of the signal is a noise compensated equivalent of the signal in terms of periodicity which improves signal to noise ratio (SNR) greater than 10 dB [22]. The modified cepstrum method utilizes the clipping and band pass filtering operation on log spectrum. Then, application of the MCEP method on the SNR improved signal removes the effect of predominant formant structure and also removes unnecessary frequency component in the frequency domain and provides better pitch determination. The proposed method thus combines the advantage of both the ACF and CEP methods. In our proposed method, pitch detection is robust in white and color noise cases.

The rest of this paper is organized as follows. Section 2 describes the background information of ACF and CEP methods. Section 3 introduces the proposed pitch detection algorithm which utilizes windowless autocorrelation function and MCEP method. Section 4 compares the performance of the proposed method with existing methods in terms of gross pitch error. Finally Section 5, we conclude the paper.

2. Background Information

The voiced speech can be expressed as a periodic signal $s(n)$ as follows:

$$s(n) = \sum_{i=0}^{\alpha} a_i \cos(2\pi i f_0 n + \theta_i) \quad (1)$$

where $f_0 = 1/T_0$ is the fundamental frequency and T_0 is the pitch period. The ACF is a popular measure for pitch period that can be expressed as

$$R_{SS}(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} s(n)s(n+\tau) \quad (2)$$

for $s(n)$, n = 0, 1, 2,..., N-1. By using (1), (2) can be expressed for a very long data segment approximately as

$$R_{SS}(\tau) = \frac{1}{2} \sum_{n=0}^{\alpha} a_n^2 \cos(2\pi f_0 n \tau) \quad (3)$$

The $R_{ss}(\tau)$ exhibits local maxima at nT_0 and provides pitch period candidates (Fig.1(b)). The main advantage of this method is its noise immunity. However, effect of formant structure can result in the loss of a clear peak in $R_{ss}(\tau)$ at the true pitch period. The performance of the conventional ACF method is significantly degraded at low SNR (Fig.2(b)). Methods have been proposed to improve the pitch period detection by emphasizing the true peak in ACF [9-14].

Cepstrum of $s(n)$ can be obtained as

$$C(n) = \frac{1}{M} \sum_{k=0}^{M-1} \log|S(k)| e^{j2\pi kn/M} \quad (4)$$

where $S(k)$ is the Discrete Fourier Transform (DFT) of $s(n)$ with M frequency points. The amplitude of $S(k)$, $|S(k)|$, in (4) can be expressed as

$$|S(k)| = \sum_{n=0}^{L} A_n [\delta(k - nf_0/F_s) + \delta(k + nf_0/F_s)] \quad (5)$$

where the harmonic amplitude $A_n = Na_n/2$ (N is assumed to be the input speech signal length), L is the number of harmonics and δ is the Kronecker delta function. $|S(k)|$ has maxima at integer multiples of f_0/F_s. Thus $C(n)$ tends to have local maxima at nT_0 (n=1, 2, 3, ...), which provides detect of pitch period as shown in Fig.1(c). The advantage of CEP method is that the logarithm operation compresses the spectral diversity of $|S(k)|$ which leads to more distinct periodic peaks and results in robustness against predominant formant structure of $|S(k)|$. However, when the speech is affected by noise, the nonlinear log operation introduces speech correlated noise products which change the algebraic structure assumed in the cepstrum processing. This can be shown in

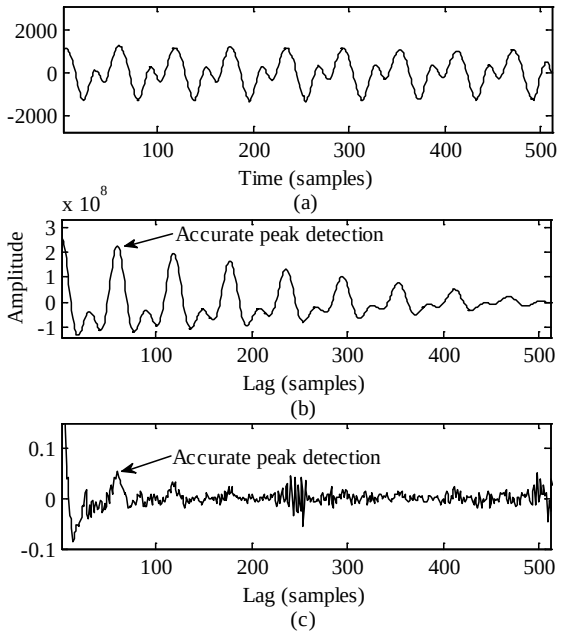

Fig 1. ACF and CEP: (a) Clean speech signal, (b) ACF of clean speech signal in (a), (c) CEP of clean speech signal in (a).

$$C'(n) = \frac{1}{M}\sum_{k=0}^{M-1} \log\left|S(k)+V(k)\right| e^{j2\pi kn/M} \tag{6}$$

where $V(k)$ corresponds to the DFT of additive noise. According to (6), addition of $\log|V(k)|$ can destroy the periodicity of $\log|S(k)|$ at low SNRs. The cepstrum obtained from the speech signal used in Fig.1(a) after corrupting with white noise at 0 dB SNR, is shown in Fig.2, where it fails to detect the true peak. The error in Fig.2(c) comes from the log operation which is used to deconvolve the multiplicative process of the vocal tract and excitation. The nonlinear log operation introduces speech correlated noise products which change the algebraic structure assumed in the cepstrum processing.

3. Proposed Method

According to ACF in (3), clearly the periodicity of $s(n)$ and that of $R_{ss}(\tau)$ are similar. When $s(n)$ is corrupted by additive noise $v(n)$, the noisy signal is given by

$$x(n) = s(n) + v(n) \tag{7}$$

When $v(n)$ is white Gaussian uncorrelated with $s(n)$, (3) can be written as

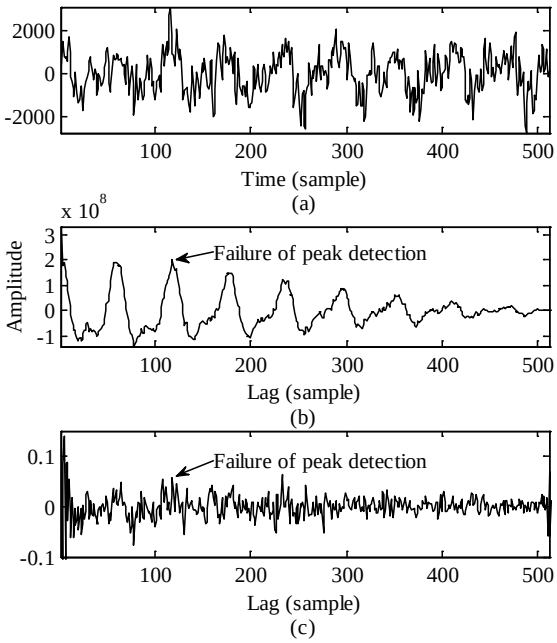

Fig 2. ACF and CEP in noise: (a) Noisy speech signal at 0 dB SNR, (b) ACF of noisy speech signal in (a), (c) CEP of noisy speech signal in (a).

$$R_{xx}(\tau) = \begin{cases} R_{ss}(\tau) + \sigma_v^2 & for\ \tau = 0, \\ R_{ss}(\tau) & for\ \tau \neq 0, \end{cases} \tag{8}$$

where δ_v^2 is the noise variance. According to (8), only the first lag is affected by the noise presence. In this paper, we aim to utilize $R_{xx}(\tau)$ as the input signal with modification for pitch determination. The modification is performed because $R_{xx}(\tau)$ is computed using a finite length of speech segment. $R_{xx}(\tau)$ can be enhanced in terms of periodicity by defining it in a windowless condition as exploited in [22], where the signal outside the window is not considered as zero. Thus the number of additions in the averaging process is always common. This results in almost similar amplitude correlation peaks even as the lag number increases. The windowless ACF can be defined for the noisy signal $x(n)$ as

$$R_{xw}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{9}$$

for $x(n)$, n = 0, 1, 2, . . . , 2N-1. In this case, an N length sequence of $R_{xw}(\tau)$, τ = 0, 1, 2, . . . , N-1 is obtained and all of them are utilized (in (10)) instead of an N length signal of $x(n)$, n = 0, 1, 2, . . . , N-1 as used in (4). For the ACF in (2), when $(n+\tau)>N$, $s(n+\tau)$ becomes zero. However, in (9), $x(n+\tau)$ is not zero outside N. This modification makes $R_{xw}(\tau)$ more stronger in periodicity with emphasized peaks as shown in Fig 3.

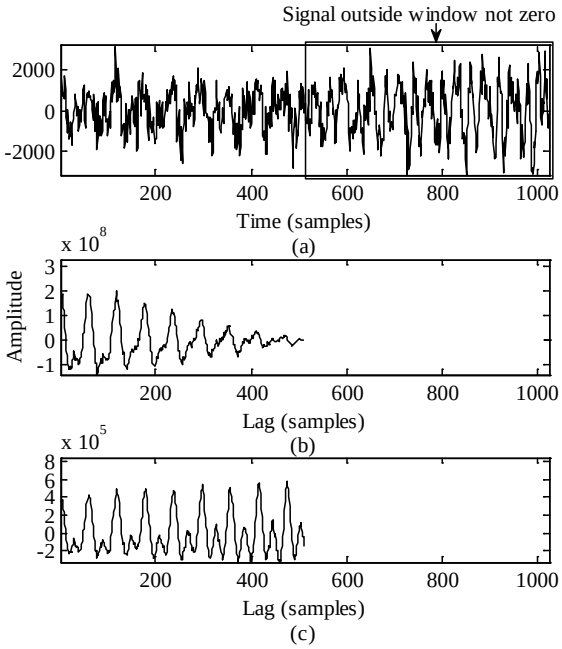

Fig 3. ACF and windowless ACF: (a) Noisy speech signal at 0 dB SNR, (b) ACF of the first half of the signal in (a)(which is the same as Fig. 2(a)), (c) Windowless ACF of signal in (a).

The sequence $R_{xw}(\tau)$ is then used for further processing instead of the signal itself. After that we apply clipping and band limitation on log spectrum of the sequence $R_{xw}(\tau)$ [19]. Lifter is carried out to remove the characteristics of vocal tract on spectrum because this is effective to perform the following clipping operation. The clipping is carried out to mostly remove unnecessary peaks which are noisy affection on spectral valleys. After the clipping and band limitation operation on log spectrum, the operation to remove high frequency components corrupted by noise is carried out. A spectral flattening logarithm operation is applied on the DFT of $R_{xw}(\tau)$, inverse DFT of which results in a time-domain sequence, $C_w(n)$, as

$$C_w(n) = \frac{1}{M}\sum_{k=0}^{M-1} \log\left|P_{xw}(k)\right| e^{j2\pi kn/M} \qquad (10)$$

where $P_{xw}(k)$ is the DFT of $R_{xw}(\tau)$.

In the windowless condition, $R_{xw}(\tau)$ and $R_{ss}(\tau)$ are similar except at τ=0, thus the DFT of $R_{xw}(\tau)$ and that of $R_{ss}(\tau)$ differ only in the DC value. Therefore $|P_{xw}(k)|$ can be written as

$$\left|P_{xw}(k)\right| = \sum_{n=0}^{L} B_n\left[\delta(k - nf_0/F_s) + \delta(k + nf_0/F_s)\right] \qquad (11)$$

Where B_n is the harmonic amplitude. For n = 0, $B_0 = \frac{N}{2}(\frac{a_0^2}{2} + \delta_v^2)$,otherwise $B_n = \frac{a_n^2}{2}.\frac{N}{2}(n = 1,2,3,....)$. Though $|P_{xw}(k)|$ in (11) is expanded due to the squared amplitude (*i.e.*, $\frac{Na_n^2}{4}$),the log operation in (10) compresses the diversity of $|P_{xw}(k)|$ and the resulted sequence C_w (n) is quite similar as that in clean speech case of (4). To summarize, this modification makes (10) a cepstral-like method but with added robustness against additive noise. Thus, (10) combines the advantage of both the ACF and MCEP methods. When our method detects the pitch, an interpolation is used. Specially, the interpolation is carried out by using the Lagrange method based on three points around the peak. The proposed cepstrum derived from the windowless ACF of the noisy speech in Fig.2(a)is shown in Fig.4. This time the obtained cepstrum is very similar with that in Fig.1(c) (clean speech case) and the pitch peak is accurately determined. Fig.5 represents a block diagram of the proposed pitch detection method.

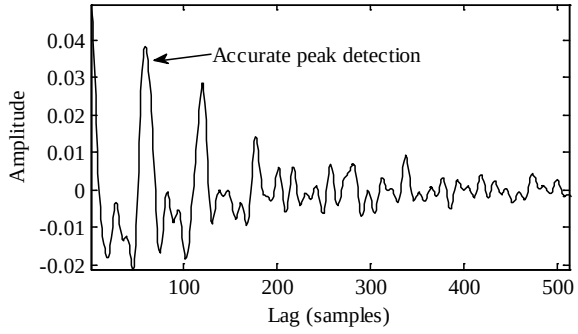

Fig 4. Proposed cepstrum of noisy speech.

4. Experimental Results and Performance

To assess the proposed method, natural speech signals spoken by three Japanese male and three female speakers are examined. Speech materials are 11 sec-long sentences spoken by every speaker sampled at 10 kHz rate with a band limitation of 3.4 kHz, which are taken from NTT database [23]. 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.

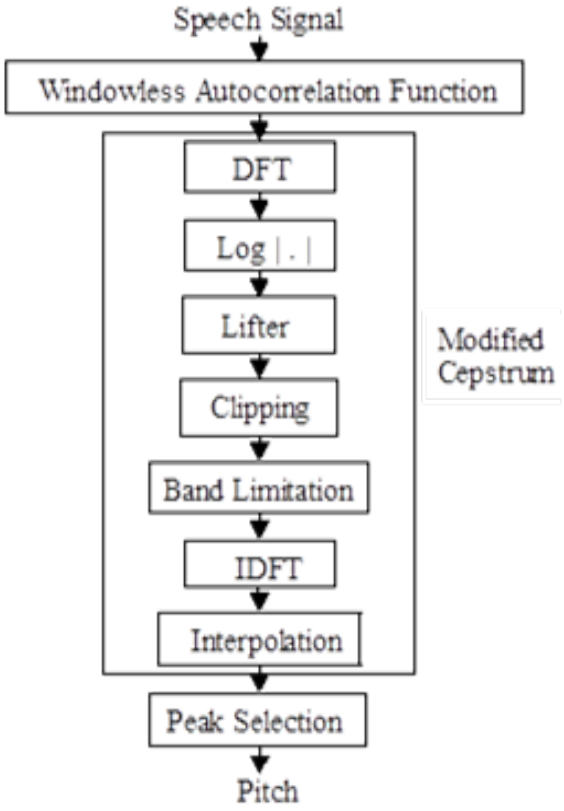

Fig 5. Block diagram of the proposed method.

The simulations were performed after adding additive noise to these speech signals. The evaluation of accuracy of the extracted fundamental frequency is carried out by using

$$e(l) = F_t(l) - F_e(l) \tag{12}$$

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) [11, 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,

$$GPE(\%) = \frac{F_{GPE}}{F_V} \times 100 \tag{13}$$

As metrics, the GPE(%) provide a good description of the performance of a fundamental frequency estimation method. The experimental conditions are tabulated in Table 1. We attempt to extract the pitch information of clean and noisy speech. Additive white Gaussian noise, exhibition noise and train noise are used, which are taken from the Japanese Electronic Industry Development Association (JEIDA) Japanese Common Speech Corporation.

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

The performance of the proposed method (PRO) is compared with a well-known autocorrelation based method, YIN [11], and cepstrum based methods, MCEP [19] and CEP [17]. The Matsig (a Mathlab library for signal processing) implementation of YIN algorithm is used here [24] without any changes. The threshold value 0.1 of YIN algorithm is assumed with respect to the global minimum instead of zero. Pitch is determined from every 51.2 ms frame at 10 ms interval. The pitch range is set to 50 to 400 Hz. The number of FFT point is 2048 and the SNR varies from a high value of ∞ dB to a very low value of -5 dB. 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 male speaker from the reference database and also the fundamental frequency contours obtained from the other fundamental frequency estimation method in Fig 6. This figure 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 female speech corrupted by the white noise at an SNR of 0 dB. In Fig 7 it is clear that the proposed method is able to give a smoother contour. The fundamental frequency contours in Fig 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.

Pitch estimation error in percentage, which is the average of GPEs for male and female speakers, are shown in Figs 8 and 9, respectively. The experimental results show that the CEP method provide less accurate result at all SNRs and all noise cases. The YIN method provide relatively accurate result than CEP method except low value of SNRs in exhibition noise case. The MCEP method provide better result than the CEP and the YIN methods in color noise cases and the MCEP method is competitive with the YIN method in white noise case. On the contrary, the proposed method gives far better results for both white and color noises in different types of SNR conditions. In particular, it is evident from Figs 8 and 9 that, for the level of SNR from 10 dB to -5 dB, the percentage GPE values resulting from the proposed method are very small but the YIN, CEP and MCEP methods give relatively higher values of percentage GPE in this range.

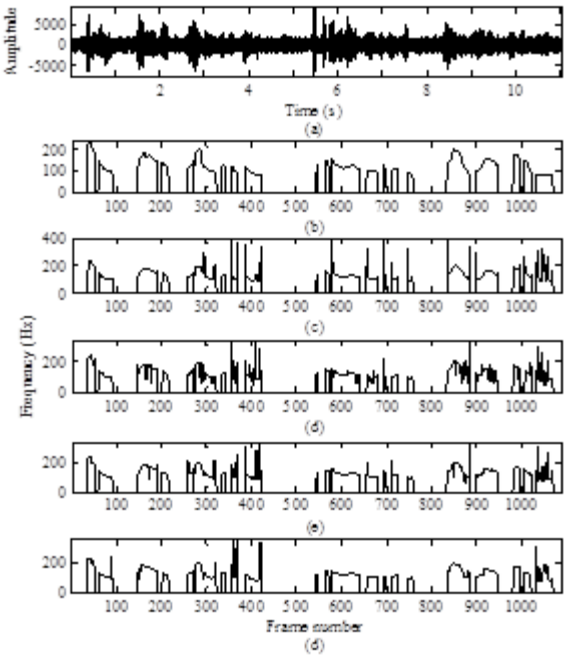

Fig 6. (a) Noisy speech signal for male speaker in white noise at an SNR 0dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) YIN, (d) CEP, (e) MCEP, and (f) PRO.

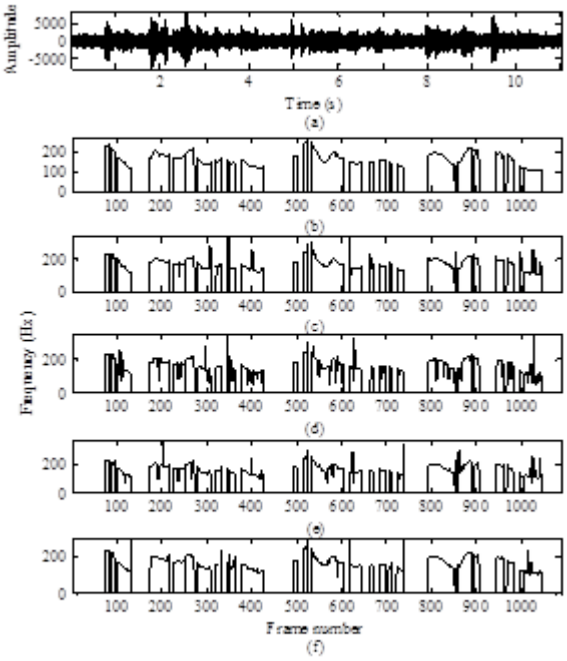

Fig 7. (a) Noisy speech signal for female speaker in white noise at an SNR 0dB, (b) True fundamental frequency of signal (a), Fundamental frequency contours extracted by (c) YIN, (d) CEP, (e) MCEP, and (f) PRO.

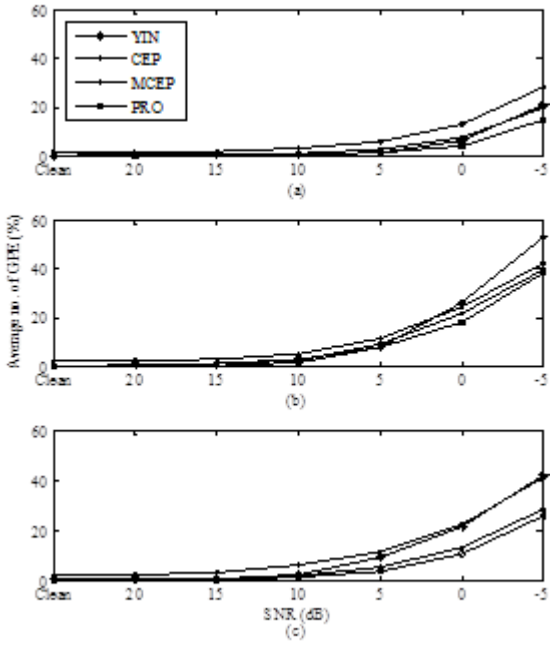

Fig 8. Comparison of percentage of average gross pitch error (GPE) for three male speakers in different noises. (a) white noise, (b) exhibition noise, and (c) train noise at various SNR conditions.

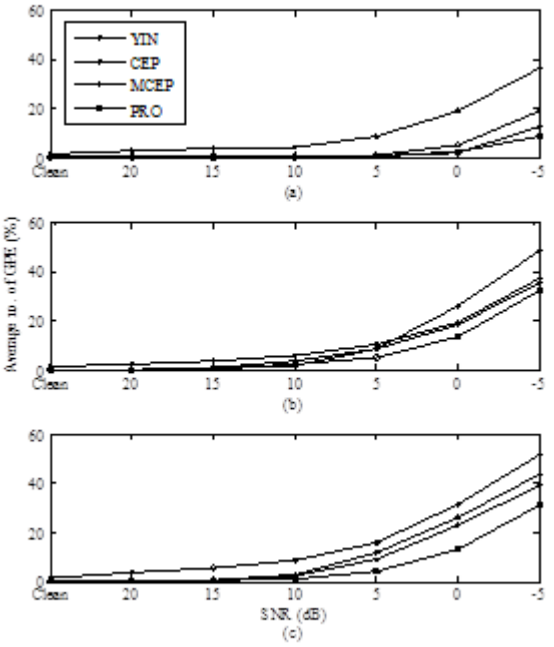

Fig 9. Comparison of percentage of average gross pitch error (GPE) for three female speakers in different noises. (a) white noise, (b) exhibition noise, and (c) train noise at various SNR conditions.

5. Conclusion

In this paper, an efficient pitch detection algorithm using windowless autocorrelation function and modified cepstrum method was introduced which leads to robustness against additive noise as well as effect of formant structure. Experimental results indicate that the proposed method outperforms existing methods such as YIN, CEP, and MCEP in terms of GPE (in percentage) for a wide range of SNR varying from ∞ dB to -5 dB. Especially the performance of the proposed method in low SNR cases is noticeable higher both in white and color noise cases than that of the other three methods. This is because windowless autocorrelation function is a noise reduced version of speech signal and application of MCEP improves the pitch detecting by utilizing the clipping and band pass filtering on log spectrum. These results suggest that the proposed method can be a suitable candidate for detecting pitch information both in white and color noise conditions with very low levels of SNR as compared with other related methods.

References

[1] M. Christensen, and A. Jakobsson, "Multi-pitch Estimation", Morgan and Claypool, (2009).
[2] A. S. Spanias, "Speech Coding: A Tutorial Review", Proceedings of the IEEE, vol. 82, no. 10, (1994), pp. 1541-1582.
[3] H. Beigi, "Fundamental of Speaker Recognition", Springer, (2011).
[4] I. R. Titze, "Principles of Voice Production", National Center for Voice and Speech, Iowa City, USA, Second Edition, (2000).
[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear Speech Analysis Algorithms Mapped to a Standard Metric Achieve Clinically Useful Quantification of Average Parkinson's Disease Symptom Severity", Journal of the Royal Soc Interface, vol. 8, (2011), pp. 842-855.
[6] W. Hess, "Pitch Determination of Speech Signals", Springer-Verlag, (1983).
[7] L. R. Rabiner, and R. W. Schafer, "Theory and Applications of Digital Speech Processing", First Edition, Prentice Hall, (2010).
[8] P. Veprek, and M. S. Scordilis, "Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques", Speech Communication, vol. 37, (2002), pp. 249-270.
[9] T. Takagi, N. Seiyama, and E. Miyasaka, "A Method for Pitch Extraction of Speech Signals Using Autocorrelation Functions Through Multiple Window Lengths", Electronics and Communications in Japan, vol. 83, no. 2, (2000), pp.67-79.
[10] T. Shimamura, and H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech", IEEE Trans on Speech and Audio Processing, vol. 9, no.7, (2001), pp.727-730.
[11] A. Cheveigne, and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music", Journal of Acoust Soc Am, vol. 111, no. 4, (2002), pp. 1917-1930.
[12] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", In: Speech Coding and Synthesis edited by W. B. Kleijn, and K. K. Paliwal, Elsevier, (1995).
[13] 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, (2006), pp.1010-1018.
[14] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans on Audio and Electroacoustics, AU-20, no. 5, (1972), pp. 367-377.
[15] F. Itakura, and S. Saito, "Speech Information Compression Based on the Maximum Likelihood Spectral Estimation", Journal Acoust Soc Japan, vol. 27, no. 9, (1971), pp. 463-472.
[16] W. J. Hess, "Pitch and Voicing Determination", In Advances in Speech Signal Processing edited by S. Furui, and M. M. Sondhi, Marcel Dekker, (1992).
[17] A. M. Noll, "Cepstrum Pitch Determination", Journal of Acoust Soc Am, vol. 41, no. 2, (1967), pp. 293-309.
[18] M. S. Andrews, J. Picone, and R. D. Degroat, "Robust Pitch Determination via SVD Based Cepstral Methods", Proceedings of Acoustics, Speech and Signal Processing, ICASSP-90, (1990), pp. 253-256.
[19] H. Kobayashi, and T. Shimamura, "An Extraction Method of Fundamental Frequency using Clipping and Band Limitation on Log Spectrum", IEICE Trans, J82-A, vol. 7, (1999), pp. 1115-1122.
[20] S. Ahmadi, and A. S. Spanias, "Cepstrum-based Pitch Detection using a New Statistical V/UV Classification Algorithm", IEEE Trans on Speech and Audio Processing, vol. 7, no. 3, (1999), pp. 333-338.
[21] M. A. F. M. R. Hasan, M. S. Rahman, and T. Shimamura, "Windowless Autocorrelation Based Cepstrum Method for Pitch Extraction of Noisy Speech", Journal of Signal Processing, vol. 16, no. 3, (2012), pp.231-239.
[22] J. Suzuki, "Speech Processing by Splicing of Autocorrelation Function", Proceedings of Acoustic, Speech and Signal Processing, ICASSP-76, (1976), pp. 713-716.
[23] Multilingual Speech Database for Telephometry. NTT Advance Technology Corporation, Japan, (1994)
[24] Online:http://www.sourceforge.net/projects/matsig, accessed on April 30, (2010)

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.