BAUET JOURNAL
Published by
Bangladesh Army University of Engineering & Technology
Journal Homepage: http://journal.bauet.ac.bd

Improved Fundamental Frequency Detection Method of Speech Signal in Noisy Environment

S. K. Bose[1], M. A. Rahman[1], M. G. S. Bhuyan[2], Mirza A. F. M. R. Hasan[1,2]

[1]*Dept. of Information and Communication Engineering, Faculty of Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.*
[2]*Dept. of Computer Science and Engineering, Faculty of Electrical and Computer Engineering, Bangladesh Army University of Engineering & Technology, Natore-6431, Bangladesh*

Abstract: An efficient fundamental frequency detection algorithm is proposed in this article. The algorithm is based on time honored fundamental frequency detection algorithm. In this proposed method, instead of the original speech signal, at first we apply low pass filter with original signal and then we employ its center clipping signal for obtaining the autocorrelation function and the average magnitude difference function. Finally this autocorrelation 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 allied method. A inclusive assessment of the fundamental frequency estimation outcomes on male and female speech in white noise indicate the superiority of the proposed method over the allied method under low levels of signal to noise ratio (SNR).

Keywords: *Fundamental Frequency; Pitch; Filter; Center Clipping; White Noise*

Introduction: Fundamental frequency i.e., pitch period is the important parameters of audio processing research. There are two general categories of speech signal, one is voiced and another one is unvoiced speech. The voice sound means when the vocal cords of the speaker vibrate and unvoiced sound means when the vocal cords are not vibrate. Fundamental frequency detection is one of the oldest, yet unsolved topic among the researchers of speech signal [1,2]. There are so many areas where the accurate fundamental frequency detection are needed, such as speech synthesis, speech recognition, speech coding, speaker identification, and to more recent topic of speech related research etc. [3,4,5]. At present there are so many fundamental frequency detection methods have been established, but significant and accurate fundamental frequency detection method is still absent.

There are three types of fundamental frequency detection algorithms (FFDAs) in the literature: time honored [6,7], frequency honored [8,9], and time-frequency honored [10]. Due to the extreme importance of accurate fundamental frequency detection problem, the muscles of different FFDAs have been searched [11,12], and several fundamental frequency reference databases have been developed to facilitate fair comparison of different FFDAs on a common platform [13]. Among the reported method, the time honored method i.e., autocorrelation

Article history:
Received 20 April 2017
Received in revised form 23 May 2017
Accepted 05 June 2017,
Available online 30 June 2017

Corresponding author details:
E-mail: mirzahasanice@gmail.com
Tel: +8801715249842

function (AUT) [6] based approaches are well accepted for their plainness and well performance in the presence of noise. The AUT 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 continued in the AUT. Spurious peaks are also sometimes introduced in the spectrum under noisy or even under noiseless conditions. This sometimes makes true peak detection a tricky job. This motivates researches to propose numerous modifications on the AUT method. One significant improvements are proposed in Sondhi [14] used center clipping AUT. Alternatively, a well known algorithm, average magnitude difference function (AMDF) has the advantage of low computation and high accuracy and the estimate price desired less than that of AUT [15]. When the scale or the fundamental frequency period of speech signal varies quickly, AMDF method will reduced apparently in fundamental frequency estimation accuracy. Correlation based fundamental frequency detection method are used in our proposed method, where the filtered centre clipping signal is used for AUT and AMDF. And this AUT is weighted by the inverse of an AMDF.

This article is ordered as follows: Section 2 describes some basic fundamental frequency detection algorithms that include time honored processing. Section 3 presents the proposed fundamental frequency detection algorithm, and Section 4 provides some experimental outcomes. Lastly, the article is completed in Section 5.

Fundamental Frequency Detection Algorithms: Fundamental frequency is an auditory perceptual property that allows the ordering of sounds on frequency domain. 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 honored method should be capable in detecting fundamental frequency period of a voiced signal. Most of the time honored fundamental frequency period estimation methods use AUT.

Let *s(m)* and *w(m)* indicate speech signal and white Gaussian noise with zero mean and variance σ^2_v, respectively. Therefore, the noisy signal *n(m)* is then given by

$$n(m) = s(m) + w(m) \qquad \text{eq.1}$$

Based on the assumption that speech and noise are uncorrelated, the AUT $R_{nn}(\tau)$ of *n(m)* can be expressed as

$$R_{nn}(\tau) = \begin{cases} R_{SS}(\tau) + \sigma_v^2 & for\ \tau = 0, \\ R_{SS}(\tau) & for\ \tau \neq 0, \end{cases} \qquad \text{eq. 2}$$

here $R_{ss}(\tau)$ is the AUT of the noise free speech signal *s(m)* estimated as

$$R_{SS}(\tau) = \frac{1}{M} \sum_{m=0}^{M-1} s(m)s(m+\tau) \qquad \text{eq. 3}$$

Here, M is the total number of considering samples in a window of the speech and τ is the lag number. In eq. 3, $R_{ss}(\tau)$ essentially exhibits peaks at the periodicity (T) of $s(m)$ (i.e., at $\tau=iT$, where i is an integer). The AUT based methods is to use the location of the second largest peak (at $\tau=T$) relative to the largest peak (at $\tau=0$) to obtain an estimate of the fundamental frequency period (Fig. 1). The major progress of AUT method is its noise exemption. On the other hand, it effects the formant structure which outcome in the failure of a clear peak in $R_{ss}(\tau)$ at the accurate fundamental frequency period. The outcome of the usual AUT method is extensively corrupted at low SNR which is shown in Fig. 2.

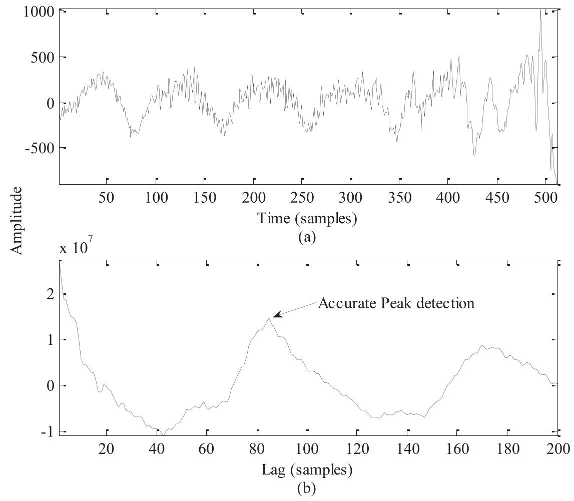

Fig. 1: (a) Noise free speech signal of a male speaker, (b) Autocorrelation function of signal in (a)

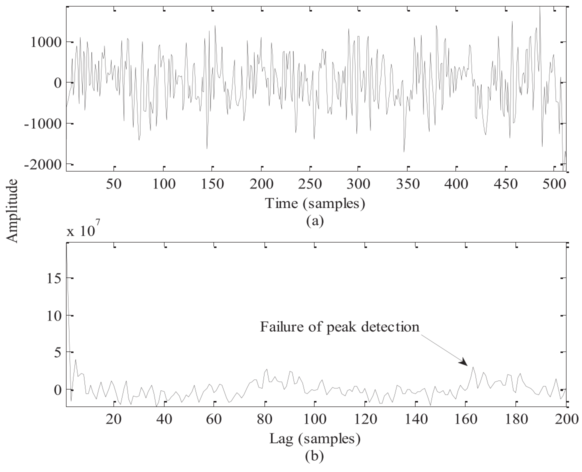

Fig. 2: (a) Noisy speech signal of male speaker (which is the same frame as Fig. 1(a)) at an SNR of 0dB, (b) Autocorrelation function of signal in (a)

The average magnitude difference function (AMDF) is one more type of autocorrelation study. The AMDF is as follow by

$$\xi_{SS}(\tau)=\frac{1}{M}\sum_{m=0}^{M-1}\left|s(m)-s(m+\tau)\right| \qquad \text{eq. 4}$$

Here, *s(m+τ)* are the samples time shifted on *τ* samples. The difference function is expected to have a strong local minimum if the lag *τ* is equal to or very close to the fundamental frequency. AMDF has benefit in comparatively low computational rate and easy implementation. Dissimilar 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 *M* samples. The fundamental frequency period is identified as the value of the lag at which the minimum AMDF occurs

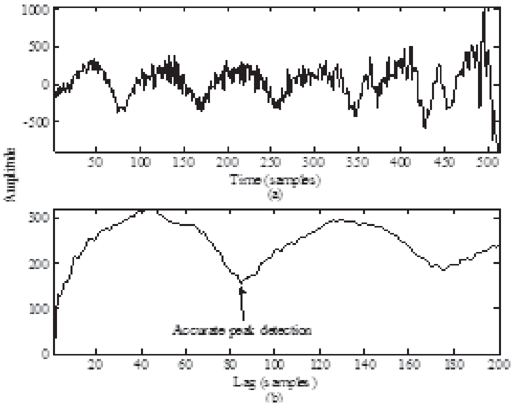

Fig. 3: (a) Noise free speech signal of a male speaker (which is the same frame as Fig. 1(a)), (b) Average magnitude difference function of signal in (a)

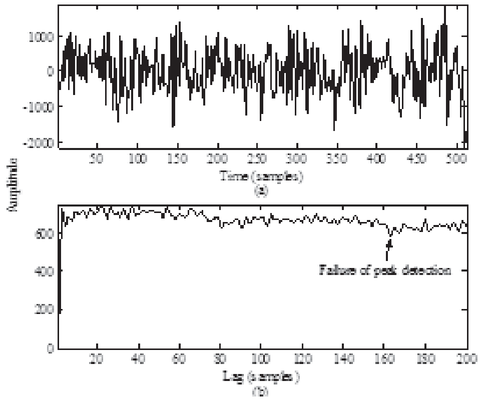

Fig. 4: (a) Noisy speech signal of a male speaker (which is the same frame as Fig. 1(a)), (b) Average magnitude difference function of signal in (a)

Proposed Method: The AUT weighted by the inverse of an AMDF is used for fundamental frequency extraction [16] and is defined as

$$\varphi_{SS}(\tau) = \frac{R_{SS}(\tau)}{\xi_{SS}(\tau) + l} \qquad \text{eq. 5}$$

Here $R_{ss}(\tau)$ and $\xi_{ss}(\tau)$ denotes the AUT and AMDF of signal *s(m)* respectively, *l* is a small positive constant. It is expected to give maximum peak at $\tau = mT$ (AUT) & deep notches at $\tau = mT$ (AMDF), and as a result the accurate fundamental frequency peak in $\varphi_{ss}(\tau)$ is highlighted (Fig. 5). Main limitation of this method is that, it is very sensitive to the half or double pitch error in noisy case as shown in Fig. 6. For fundamental frequency extraction, speech signal is generally pre-processed to build the periodicity more outstanding and to suppress other distracting features. Such techniques are often called spectrum flattening.

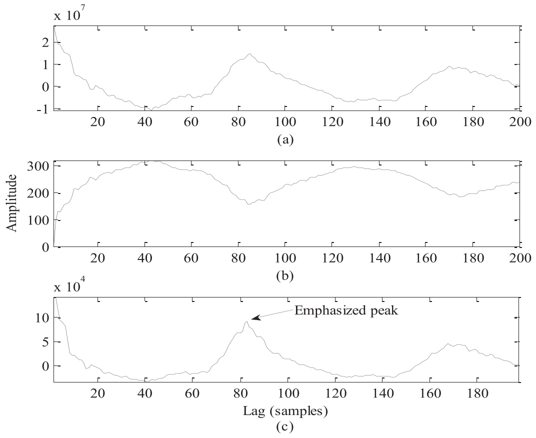

Fig. 5: Pitch peak detection in noise free speech signal (which is the same frame as Fig. 1(a)) using (a) Autocorrelation function method, (b) Average magnitude difference function method, (c) Weighted autocorrelation function method

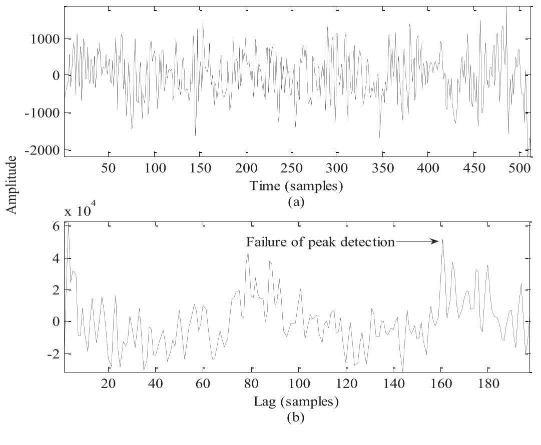

Fig. 6: Pitch peak detection in noisy speech signal (which is the same frame as Fig. 1(a)) using Weighted autocorrelation function method

Center clipping is the most well-liked spectrum flattening technique [14] and this technique is used in our proposed method. This technique is as follow by

$$s'(m) = C_C\{s(m)\} = \begin{cases} (s(m) - C_C), & s(m) \ge C_C \\ 0, & |s(m)| \langle C_C \\ (s(m) + C_C), & s(m) \le -C_C \end{cases} \qquad \text{eq. 6}$$

here *s´(m)* is the center clipping signal of speech signal *s(m)* and C_C is the clipping level.

Instead of the speech signal *s(m)*, we introduced low pass filter with original signal in our proposed method (Fig. 7) and then employ its center clipping signal *s´(m)* (Fig. 8) for obtaining the AUT and using this AUT weighted by $1/\xi_{ss}(\tau)$. Finally the outcome of this proposed method is that the accurate peak is more highlighted (Fig. 9), and as a end result the errors of fundamental frequency detection are reduced. The correlation based proposed method is as follow by

$$\varphi_{ss\_cc}(\tau) = \frac{R_{ss\_cc}(\tau)}{\xi_{ss\_cc}(\tau) + l} \qquad \text{eq. 7}$$

Here $R_{ss\text{-}cc}(\tau)$ and $\xi_{ss\text{-}cc}(\tau)$ is the AUT and AMDF of signal $s´(m)$ respectively.

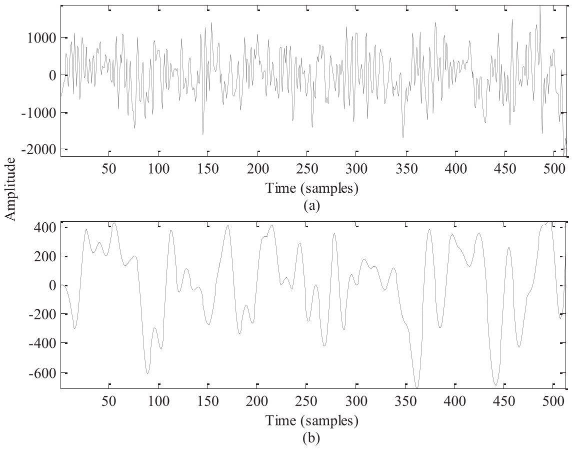

Fig. 7: (a) Noisy speech signal of a speaker, (b) Apply lowpass filter of signal in (a).

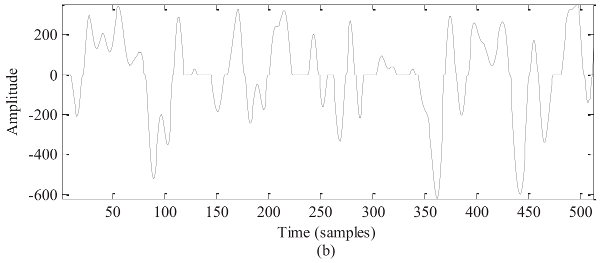

Fig. 8: Apply Center clipping technique of signal in Fig. 7(b).

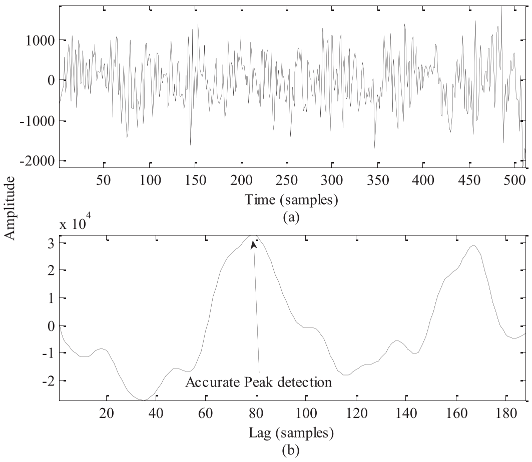

Fig. 9: Pitch peak detection in noisy speech signal (which is the same frame as Fig. 1(a)) using Proposed method.

Experimental Results: To evaluate the proposed method, there are two Japanese male and female speaker's speech are considered. Speech materials are 11 sec-long sentences spoken by every speaker sampled at 10 kHz rate, which are taken from NTT database [17]. 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. The assessment of the proposed method, we consider rectangular window, low pass filter, 51.2 ms each window, 1024 FFT point, SNRs are clean and from 20dB to -5dB interval 5dB. Also consider two criteria, one is gross pitch error (GPE) and another one is fine pitch error (FPE). The estimation of accurateness to detected fundamental frequency is carried out follow by

$$e(q) = F_t(q) - F_e(q) \qquad \text{eq. 8}$$

Here $F_t(q)$ is the accurate fundamental frequency, $F_e(q)$ is the detected fundamental frequency by each method, and $e(q)$ is the detected error for the q-th frame. If $|e(q)| > 20\%$, we identify the error as a gross pitch error (GPE) [18,19]. On the other hand we identify the error as a fine pitch error (FPE). The probable sources of the GPE are pitch doubling, halving and inadequate suppression of formants to affect the assessment. The percentage of GPE is as follow by

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

Here F_{GPE} is the number of frames yielding GPE and F_v is the total number of voiced frames. The mean FPE is as follow by

$$FPE_{mean} = \frac{1}{M_i} \sum_{j=1}^{M_i} e(q_j) \qquad \text{eq. 10}$$

Here q_j is the j-th interval in the utterance for which $|e(q_j)| \leq 20\%$ (fine pitch error), and M_i is the number of such intervals in the utterance.

We try to detect the fundamental frequency of noise free and noisy speech signals. Every method is applied in additive white Gaussian noise. The Japanese Electronic Industry Development Association (JEIDA) Japanese Common Speech Corporation provided the noise speech. The outcome of the proposed method is evaluated with AMDF and weighted autocorrelation method, WAUT [15, 16]. In WAUT, the constraint τ in [16] is set to 0.5 and in proposed method, the constraint C_C is set to 1% of the maximum magnitude of signal. As the fundamental frequency vary is known to be 50-500 Hz for most male and female speakers and our sampling frequency is 10 KHz, the setting of lag digit (i.e., 200) is usually used for the AMDF, WAUT and the proposed method. In order to assess the pitch assessment performance of the proposed method, we sketch a reference pitch contour for SNR 0dB noisy speech in white noise speech of a female speaker from the reference database and also the pitch contours obtained from the different pitch estimation method which is shown in Fig. 10. This figure indicates that in compare to the different method, the proposed method performs a better smoother pitch contour even at an SNR of 0 dB. Fig. 11 indicates a judgment of the pitch contour considering the male speech corrupted by the white noise at an SNR of 0 dB. Fig. 11 also performs a smoother contour even in the presence of white noise in our proposed method. From Figs. 10 and 11, it is obvious for three methods; our proposed method is able to reducing the double and half pitch errors thus yielding a smooth pitch path. Fundamental frequency detection inaccuracy in percentage, which is the average of GPEs for white noise are shown in Fig. 12. This figure implies that the proposed method provides distant enhanced outcomes for both female and male cases in various SNR environments.

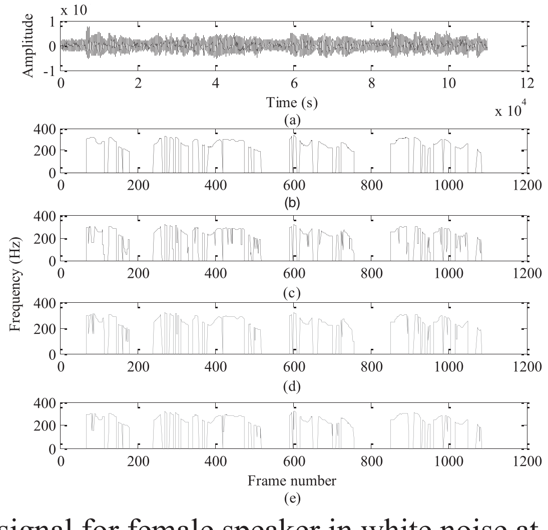

Fig. 10: (a) Noisy speech signal for female speaker in white noise at an SNR 0 dB (b) True pitch of signal (a), Pitch contours extracted by (c) AMDF (d) WAUT, and (e) Proposed method.

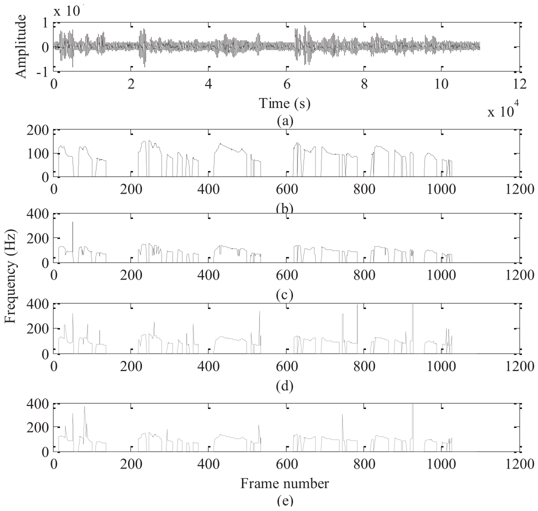

Fig. 11: (a) Noisy speech signal for male speaker in white noise at an SNR 0 dB, (b) True pitch of signal (a), Pitch contours extracted by (c) AMDF, (d) WACF, and (e) Proposed method.

These simulation outcomes indicate that the proposed method is better performs to the AMDF and WAUT method in most of the cases. The proposed method performs more strongly compared with the AMDF and WAUT method at low SNR (0 dB, -5 dB). The FPE signify a degree of the variation in extracted fundamental frequency. Mean of the errors (in Hz) was considered in FPE. Considering all the utterances of the female and male speakers, the FPE values ensuing from the three methods are sketch which is shown in Fig. 13. Average FPEs for all methods range around from 1.7 Hz ~ 5.6Hz. Fig. 13, implies that the FPE values resulting from the proposed method are tiny but the AMDF and WAUT method present relatively higher values of FPE in this scale. From the experimental results it is establish that the range of FPEs is also within the satisfactory limit and consistently satisfactory at other SNRs.

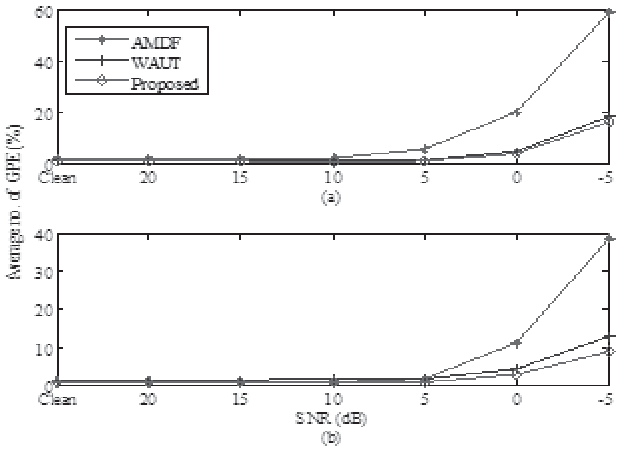

Fig. 12: Percentage of average gross pitch error (GPE) in white noise for different speakers under various SNR conditions; (a) Female speakers, (b) Male speakers.

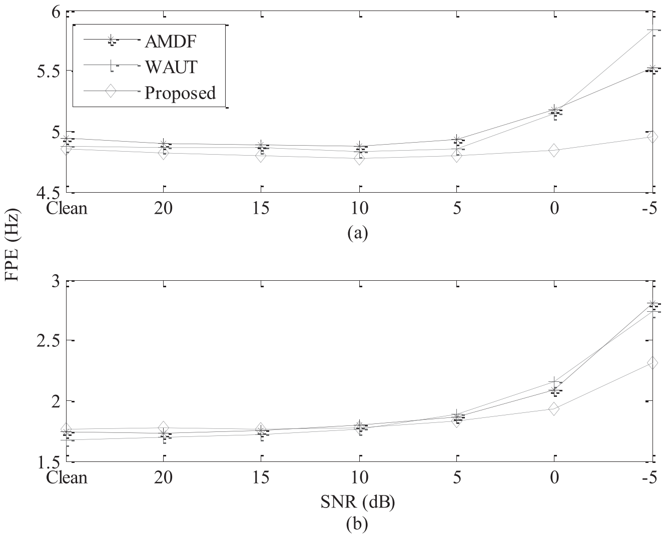

Fig. 13: Comparison of average performance outcomes in terms of mean fine pitch error (FPE) for different speakers under various SNR conditions; (a) Female speakers, (b) Male speakers.

Conclusions: This article, a competent fundamental frequency estimation using correspondence based method was launched which leads to robustness against white noise. Experimental outcomes indicate that the proposed method gives better performance in terms of GPE (in percentage) and FPE compared with the different method such as AMDF and WAUT. The competitive values of mean FPEs also point out the accurateness of pitch detection by the proposed method. These significant outcomes recommend that the proposed method can be appropriate contestant for extracting fundamental frequency information in various noises environments with very low levels of SNR as compared with other allied methods.

References:

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

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

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

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

[5] A. E. Rosenberg , and M. R. Sambur, New techniques for automatic speaker verification, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-23, 2 (1975) 169-176.

[6] L. R. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-25, 1(1977) 24-33.

[7] M. A. F. R. Hasan, and T. Shimamura, An efficient pitch estimation method using windowless and normalized autocorrelation functions in noisy environment, International Journal of Circuits, Systems and Signal Processing, 3, 6 (2012) 197-204.

[8] A. M. Noll, Cepstrum pitch determination, Journal of Acoust. Soc. Am., 41, 2 (1967) 293-309.

[9] S. Ahmadi, and A. S. Spanias, Cepstrum based pitch detection using a new statistical V/UV classification algorithm, IEEE Trans. Speech and Audio Processing, 7, 3 (1999) 333-338.

[10] 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, 16, 3 (2012) 231-239.

[11] L. R. Rabiner, M. J. Cheng, A. M. Rosenberg, and C. A. McGonegal, A comparative performance study of several pitch detection algorithms, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-24, 5 (1976) 399-417.

[12] P. Veprek, M. S. Scordilis, Analysis, enhancement and evaluation of five pitch determination techniques, Speech Communication, 37 (2002) 249-270.

[13] F. Plante, G. Meyer, and W. A. Ainsworth. A pitch extraction reference database. In: Proceedings of EUROSPEECH, 1995, p. 837-840.

[14] M. M. Sondhi, New methods of pitch extraction, IEEE Trans. Audio Electroacoust., AU-16(1968) 262-266.

[15] M. J. Ross, H. L. Schafer, A. R. F. B. Cohen and H. Manley, Average magnitude difference function pitch extraction, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-22, 5 (1974) 353-362.

[16] T. Shimamura, and H. Kobayashi, Weighted autocorrelation for pitch extraction of noisy speech, IEEE Trans. on Speech and Audio Processing, 9, 7 (2001) 727-730.

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

[18] A. Cheveigne, and H. Kawahara, YIN. a fundamental frequency estimation for speech and music, Journal of Acoust. Soc. Am., 111, 4 (2002) 1917-1930.

[19] 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, 86 (2006) 1010-1018.