International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 7, Issue 5 (June 2013), PP. 23-30

An Auto-Correlation Based Speech Enhancement Algorithm

Lalchhandami[1], Maninder Pal [2]
[1,2]Department of Electronics & Communication Engineering
Maharishi Markandeshwar University, Mullana (Ambala), INDIA

Abstract:- Enhancement of speech signals recorded using signal channel devices such as mobile phones is of prime interest. It is because for these devices, it is not possible to record noise signals separately, and the surrounding background noises are picked up by their microphone simultaneously with the speech signal. This may even completely fade-in the speech signal, depending upon the signal-to-noise ratio (SNR). Therefore to address this problem, number of algorithms and techniques has been developed. However, the existing methods are not able to perform homogenously across all noise types. The auto-correlation function of a noisy speech signal is usually confined to lower time lag and is very small or zero for higher time lag. Therefore, the higher-lag auto-correlation coefficients are relatively robust to additive noise distortion. This paper is focused on enhancing the noisy speech signal from single channel devices by using only the higher-lag auto-correlation coefficients. The efficiency of the algorithm is evaluated in terms of energy, zero crossings and intelligibility of speech signal.

Keywords:- Single channel speech enhancement; speech processing; spectral subtraction auto-correlation and Speech signals SNR

I. INTRODUCTION

Speech enhancement is a topic of interest for last many years. In particularly, enhancement of speech signals recorded using signal channel devices such as mobile phones is of prime interest in this paper. It is because for these devices, it is not possible to record noise signals separately, and the surrounding background noises are picked up by its microphone simultaneously with the speech signal. This may even completely fade-in the speech signal, depending upon the signal-to-noise ratio (SNR). Therefore to address this problem, number of algorithms and techniques has been developed. The key techniques include: suppression of noise using the periodicity of speech or noise, enhancement based on perceptual criteria and subtractive-type algorithms such as spectral subtraction and Wiener filter [1], [6]. Among these, the spectral subtraction algorithm is the oldest speech enhancement algorithm. In this approach, the degraded speech signal is enhanced by subtracting an estimate of the average noise spectrum from a noisy speech spectrum and the noise spectrum is estimated or updated during the silence period of the speaker. However, the existing methods are not able to perform homogenously across all noise types. It is because all of these speech enhancement systems are based on certain assumptions and constraints that are typically dependent on the application and the environment. This paper is focused on reducing the continuous background noises from the speech signals recorded using single channel microphone based devices. For this purpose, auto-correlation function in time domain and frequency domain is used. The magnitude of auto-correlation coefficients is usually large between 2ms and 12ms, as the human pitch period is typically constrained between these values. However, the same is generally not true for noisy speech signals. The auto-correlation function of a noisy speech signal is usually confined to lower time lag and is very small or zero for higher time lag. The higher-lag auto-correlation coefficients are relatively robust to additive noise distortion [5]. Therefore, this paper is focused on enhancing the noisy speech signal from single channel devices by using only the higher-lag auto-correlation coefficients. The efficiency of the algorithm is evaluated in terms of energy, zero crossings and intelligibility of speech signal. This is presented below.

II. SYSTEM MODEL

As discussed above, the major constraint of single channel methods in speech enhancement is that there is no reference signal for the noise available. Therefore, the power spectral density of the noise needs to be estimated based on the available noisy speech signal only and this makes it a challenging task. In all single channel enhancement techniques, the noisy speech signal is given by

$$y(n) = x(n) + d(n) \quad (1)$$

Where, $x(n)$ represents the clean speech signal, $d(n)$ is the uncorrelated additive noise and $y(n)$ represents the degraded noisy speech signal. To enhance such noisy speech signals, an auto-correlation function based speech enhancement is presented in this paper. Its principle is based on traditional Spectral Subtraction method. In this Spectral Subtraction method, the degraded speech signals are enhanced by subtracting the estimate of the average noise spectrum from a noisy speech spectrum [1]. The noise spectrum is estimated during the periods when the signal is absent; which is usually very difficult to do in practice. In addition, it is also assumed that speech and noise is additive and uncorrelated. An estimate of the clean signal $\hat{x}(n)$ is recovered from the noisy signal $y(n)$ by assuming that there is an estimate of the power spectrum of noise $\left|\widehat{D}_k\right|^2$, which is obtained by averaging over multiple frames of a known noise segment. An estimate of the short-time squared magnitude spectrum of the clean signal using this method can be obtained as follows:

$$\left|\hat{X}_k\right|^2 = \begin{cases} |Y_k|^2 - \left|\widehat{D}_k\right|^2, & \text{if } |Y_k|^2 - \left|\widehat{D}_k\right|^2 \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Where, $|Y_k|$ is the magnitude of the noisy signal spectrum, $\hat{X}_k$ is the estimate of the clean frequency spectrum for a given frame, and $|\widehat{D}_k|$ is the magnitude of noise spectrum estimate during non-speech activity. To recover the signal, the magnitude spectrum estimate is finally combined with the phase of the noisy signal, using Inverse Discrete Fourier Transform (IDFT) as follows:

$$\hat{x}(n) = \left|\hat{X}_k\right| \mathrm{e}^{j\varphi(y,k)} \quad (3)$$

Although the spectral subtraction algorithm can be easily implemented to effectively reduce noise present in the corrupted signal; yet, it has several shortcomings. The major drawback of this method is the resulting musical noise, due to rapid coming and going of speech signals over successive frames. This is why this paper focuses on using auto-correlation function instead of power spectrum.

The autocorrelation function of a signal contains the same information about the signal as its power spectrum [2]. However, the main difference between the power spectrum and auto-correlation domain is that in the power spectrum domain, the information is presented as a function of frequency; while in the latter, it is presented as a function of time. More specifically, the higher-lag auto-correlation coefficients of the speech signal $x(n)$ usually contain information about the signal's power spectrum; whereas, the magnitude of higher-lag auto-correlation coefficients of the noise signal $d(n)$ is relatively small for some noise types. Therefore, the lower-lag auto-correlation coefficients can be discarded and only higher-lag auto-correlation coefficients of the noisy speech signal can be used for spectral estimate. The spectral estimation is done using only the higher-lag auto-correlation coefficients; and, the speech and noise signals can be separated without having to estimate the noise signal directly. The estimated power spectrum can then be used to enhance the corrupted noisy speech signals. It is to be noted that the auto-correlation function of a signal can be computed in time domain and frequency domain. This paper has focused on computing auto-correlation function in both domains in order to investigate their effect on enhancing speech signals.

III. IMPLEMENTATION PROCEDURE

The flowchart of the autocorrelation based speech enhancement algorithm is given in Fig. 1. As shown in Fig. 1, the first step in spectral estimation is to divide the noisy speech signal into overlapping short time frames. These short time frames are 64 ms (512 samples) long. A Hamming window with 50% overlapping is then applied on these short time frames. This process is defined by Eqn 4, where $x(n)$ is discrete speech signal and $w(n)$ is the window function shifted by m samples.

$$x_w(n) = x(n)w(m-n) \quad (4)$$

By taking the Fourier Transform of $x_w(n)$, a discrete Short-Time Fourier Transform can be computed as shown in Eqn 5.

$$X(w,k) = \begin{cases} \frac{1}{N} \sum_{n=w-N+1}^{w} x_w(n)\, e^{\frac{-j2\pi kn}{N}} & k = 0, \ldots, N-1 \\ 0 & otherwise \end{cases} \quad (5)$$

The resulting discrete STFT can be expressed as

$$X(w,k) = |X(w,k)| e^{j\varphi(w,k)} \quad (6)$$

The $|X(w,k)|$ component is the magnitude of the Short-Time Fourier Transform and the $\varphi(w,k)$ component is the phase. For speech recorded using single-channel based devices, the noisy signal $y(n)$ can be analytically expressed in frequency domain as:

$$Y_k e^{j\varphi(y,k)} = X_k e^{j\varphi(x,k)} + D_k e^{j\varphi(d,k)} \quad (7)$$

In the modification stage, the degraded signal in the frequency domain is modified in order to restore the

original speech signal. For this purpose, a biased auto-correlation sequence of the windowed time frames is then computed. As the power spectrum of an auto-correlation sequence has dynamic range twice as large as that of the signal's power spectrum, so if a Hamming window is applied to the positive side of the auto-correlation sequence, the dynamic range would be effectively reduced by half. Therefore, to address this problem, a new window of double dynamic range is used to process the one-sided auto-correlation sequence. For implementing this window, the dynamic range of Hamming window is made double by computing its auto-correlation sequence and resulting auto-correlation sequence. The resulting sequence is defined as the coefficients of the new high dynamic range window. To increase the noise reduction effect, the new window is placed over the higher-lag region of the auto-correlation sequence, which is taken to be greater than 2 ms (16 samples), as there is natural null in the average short-time auto-correlation of speech signals at approximately 2 ms lag. An estimate of the original speech signals power spectrum is then computed for each frame as the magnitude spectrum of the windowed higher-lag autocorrelation coefficients. This is done by pairing the modified magnitude spectra $\hat{X}_k$ with each of the original corresponding phase spectra $\varphi(y,k)$. The degraded phase spectra are then used to create the enhanced spectra. The enhanced speech frames are then synthesized using an inverse Fourier transform of the complex spectra as mentioned in Eqn 8.

$$\hat{x}(n) = IDFT\{\hat{X}_k e^{j\varphi(y,k)}\} \tag{8}$$

These frames are then overlapped and added to obtain the enhanced speech signal.

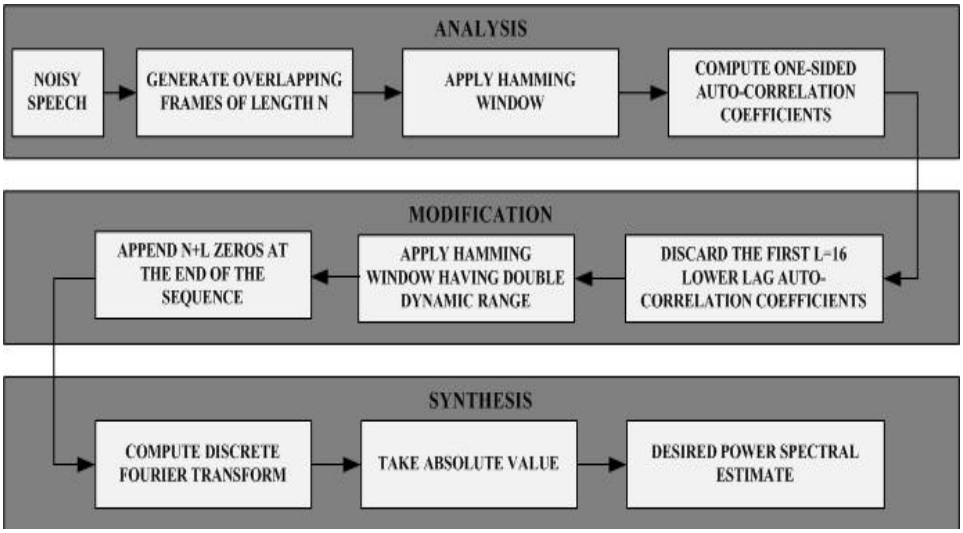


Fig. 1: Flowchart of power spectral estimate using higher-lag auto-correlation coefficients.

The block diagram of the speech enhancement using higher-lag auto-correlation coefficients is shown in Fig. 2.

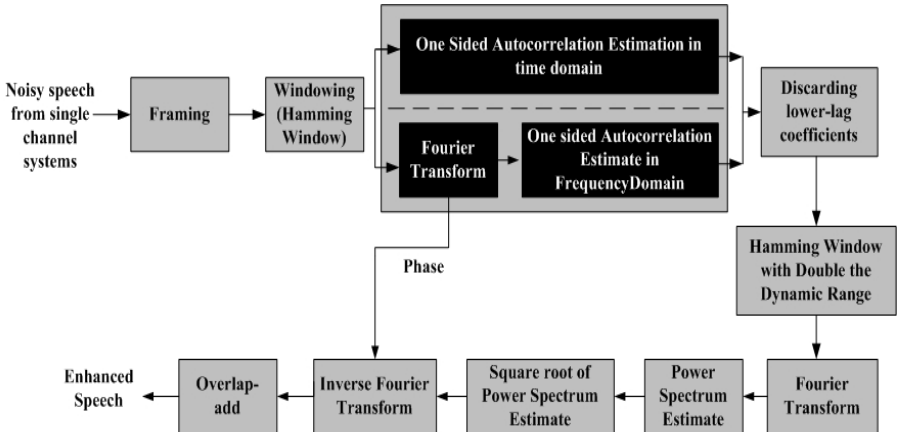


Fig. 2: Block diagram of the auto-correlation based speech enhancement algorithm.

IV. RESULTS & DISCUSSION

To evaluate the performance of the auto-correlation based speech enhancement algorithm, the same is implemented in MATLAB and extensive simulations have been performed. A small length of approximately

10 seconds of speech of a female person, in the age group 20 to 25 years, is recorded in the presence of noise. The noise is produced by two sources: running fan and car. The speech signals are recorded separately in both noise conditions. To record signals at different SNR's, the speed of fan is changed to change the level of noise. The mobile phone and sound card used for recording signals has a single microphone, so the speech signal and noises are both picked up by the microphone. These recorded signals are called noisy speech signals. These noisy speech signals are then processed by an auto-correlation based speech enhancement algorithm and the estimated cleaned signals are then stored and listened back to estimate their degree of similarity (intelligibility) with their original version. Auto-correlation function is computed by two methods: the generalized cross-correlation (GCC) method in frequency domain and using Matlab *xcorr* command in time domain. The results obtained for these two methods are then compared. These results are shown in Fig. 4 to Fig. 15, and the key conclusions drawn from the results are discussed below.

1. Fig. 4 & Fig. 5 shows the auto-correlation function and one-sided auto-correlation window respectively of one frame (i.e. 512 samples). For enhancing the speech signal, the lower lags of 16 samples (2 ms) of the one-sided auto-correlation function are discarded and only the higher-lag coefficients are retained. A Hamming window with wide range (width 512 samples) is applied to the higher-lag auto-correlation sequence. It is observed that this wide dynamic range Hamming window has peak in the middle frequencies and decrease to zero at the edges; thereby, reducing the effects of the discontinuities as a result of finite duration. It is also observed that this window captures all the harmonic peaks of the auto-correlation sequence, and therefore, captures all the formant structures with no spectral loss. This indicates that this window works well in the auto-correlation domain in enhancing the noisy speech signal.
2. Fig. 6 show the power spectrum estimate of a single frame (512 samples) of the speech signals; which is used to obtain the enhanced version of the signal. The estimated power spectrum is computed on a frame to frame basis, and need not necessarily be computed during the non-speech activity of the speaker, as required in the spectral subtraction algorithm. This is the key advantages of this algorithm.
3. Fig. 7, Fig. 12 & Fig. 13 shows the recorded speech signal with different levels of noises i.e. different SNRs and in the presence of various noise producing sources. The energies of the recorded and the recovered speech signals are computed and is observed that a high level of energy (above threshold) of the speech signal is approximately at the same time instance for both the recorded and the recovered signal. It is also observed that the zero-crossing rate is high during the speech period of the signal Thus, the intelligibility of the recovered signal is maintained.
4. In Fig. 7 to Fig. 15, the recorded speech signals are processed by the auto-correlation based speech enhancement method. The auto-correlation function is calculated using the GCC method in frequency domain and using *xcorr* Matlab command for time domain. It can be observed that the second method (*xcorr*) gives better results in reducing the noise from the speech signals; however, at the expense of small decrease in magnitude of the speech signal. The decrease in the magnitude of the recovered speech signal is uniform; thus, the algorithm maintains the intelligibility of the recovered speech signal.
5. Fig. 11, Fig.14 & Fig. 15 shows the spectrograms of the recorded noisy signal and the recovered signal obtained using the auto-correlation based algorithm. The spectrograms showed that the noise energy, at a particular frequency and time, is reduced to a considerable extent in the recovered speech signal; thereby, improving the quality of the recovered speech signals.
6. Fig. 12 & Fig. 14 shows the results of noisy speech signals corresponding to a female voice of approximately 10 seconds, recorded in a running car by using a microphone and sound card, and mobile phone. It is observed that the recovered signal using *xcorr* gives better results than the GCC method, as both musical and residual noises are observed in the signals recovered using GCC method. The continuous noises from the ambient environment are reduced to a significant extent in the recovered signal by the two methods; thereby, proving the effectiveness of the algorithm for continuous surrounding noises.
7. Fig. 13 & Fig. 15 shows the results for speech signals recorded using a single microphone (mono) based mobile phone, used by a person in a running car. The recorded signal is then processed by the auto-correlation based algorithm and is observed that the moving wind noise and engine noise of car are reduced in the recovered signal. It is also observed from the spectrogram that the energy of the noises lie in the audio frequency range indicating that the speech signals gets easily distorted by the ambient noises. The spectrogram of the recovered signal shows that the energy of these noises is reduced in the recovered signal. Thus, the algorithm works well for enhancing speech signals recorded using single channel devices in a noisy environment. This justifies that this algorithm is designed to minimize the noises produced by constant sources such as vehicles and is well suited for environments such as persons using mobile phones in cars, trains etc.

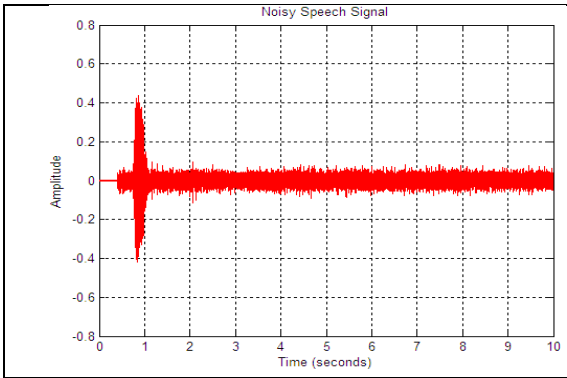


Fig. 3: Recorded speech (female speaking /a/ sound) signal of length 10 seconds in presence of noise produced by fan moving at constant speed.

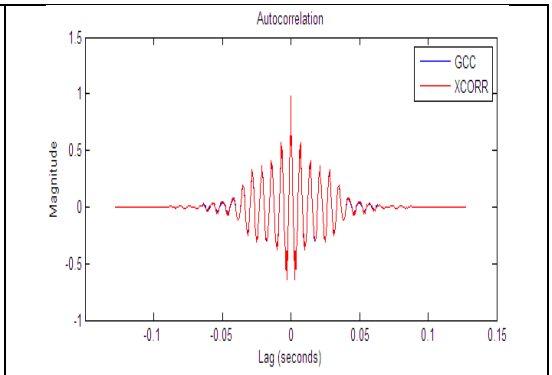


Fig. 4: An example of autocorrelation of a single speech frame (512 samples) using *xcorr* and generalized cross-correlation (GCC).

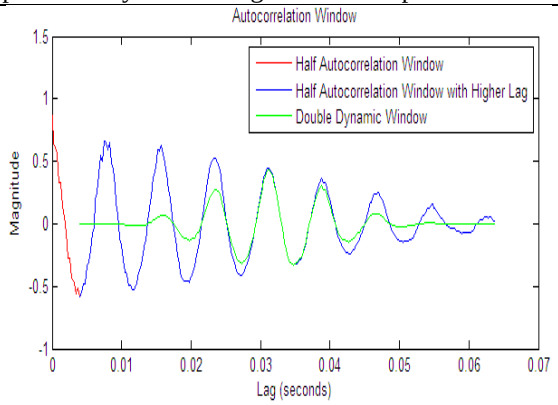


Fig. 5: An example of autocorrelation window (single sided) of single frame (512 samples).

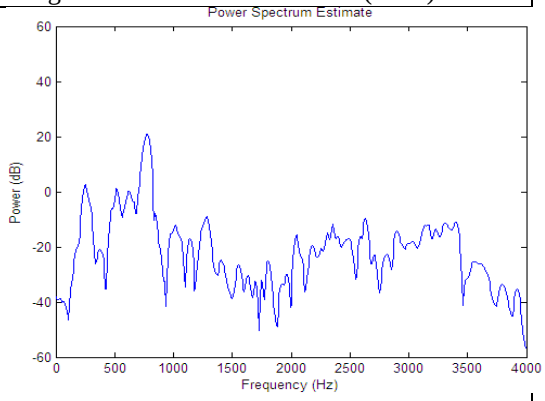


Fig. 6: An example of the power spectrum estimate of a single frame (512 samples).

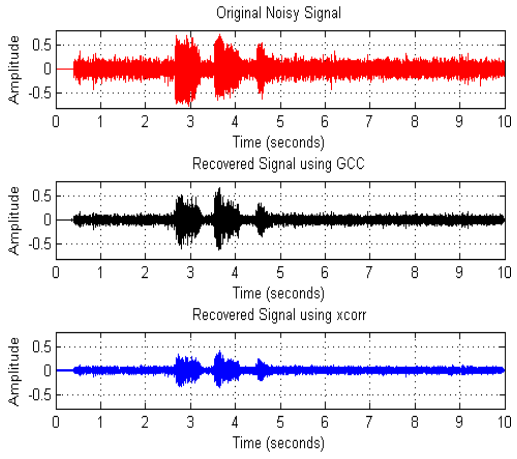


Fig. 7: The recorded speech (female speaking /a/, /b/ and /c/ sound) signal of length 10 seconds in presence of noise produced by a fan moving at constant speed, and the recovered signals using GCC (frequency domain) and *xcorr* (time domain).

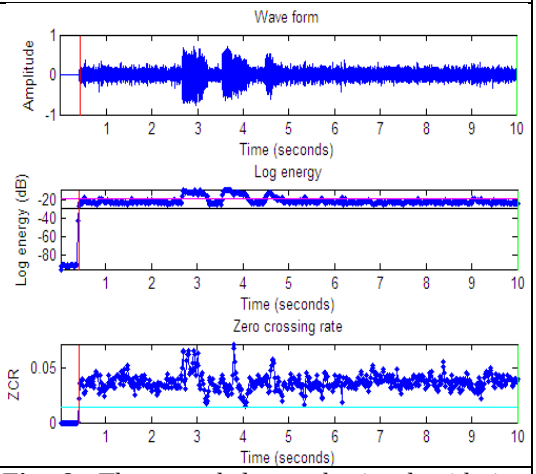


Fig. 8: The recorded speech signal with its energy (logarithmic scale) and zero crossing rate.

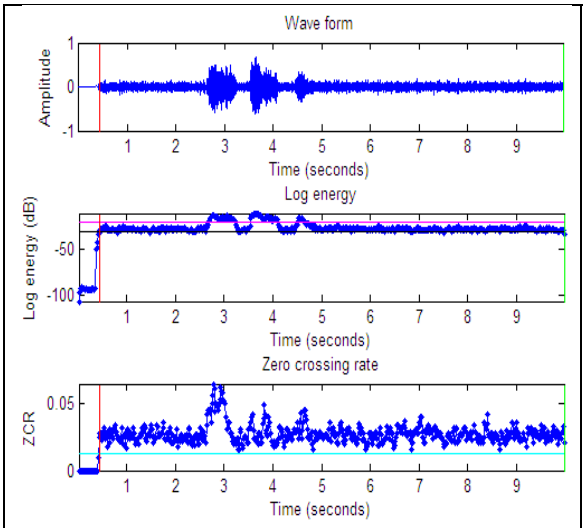


Fig. 9: The recovered speech signal (using GCC) with its energy (logarithmic scale) and zero crossing rate.

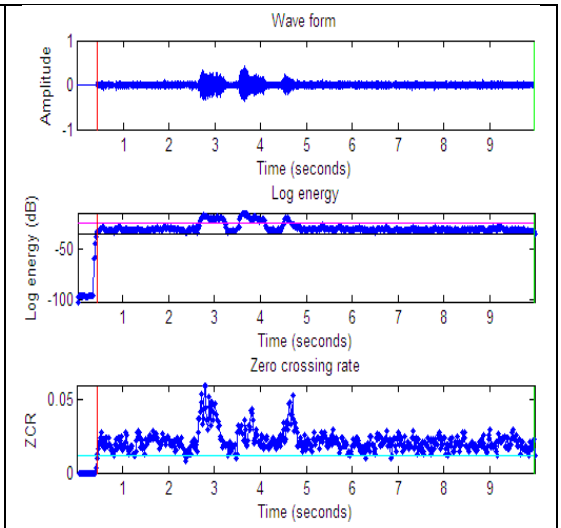
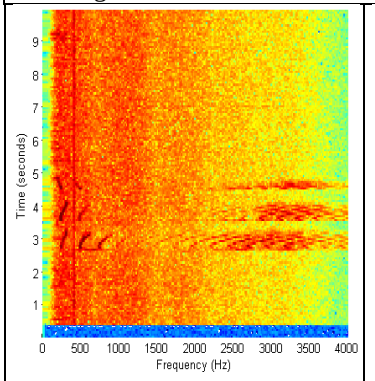
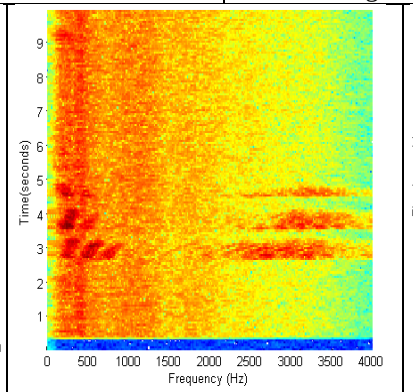


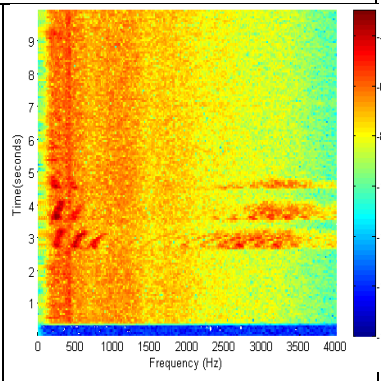
Fig. 10: The recovered speech signal (using *xcorr*) with its energy (logarithmic scale) and zero crossing rate.



a) The spectrogram of recorded noisy speech signal.



b) The spectrogram of recovered speech signal using GCC.



c) The spectrogram of recovered speech signal using *xcorr*.

Fig. 11: The spectrogram of a) recorded speech signal, b) recovered signal using GCC, and c) recovered signal using *xcorr*.

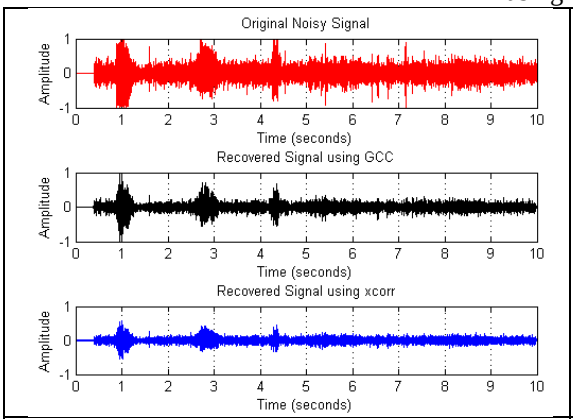


Fig. 12: The recorded speech signal of a female /a/, /b/ and /c/ sound of length 10 seconds in running car and the recovered signals using GCC (frequency domain) and *xcorr* (time domain).

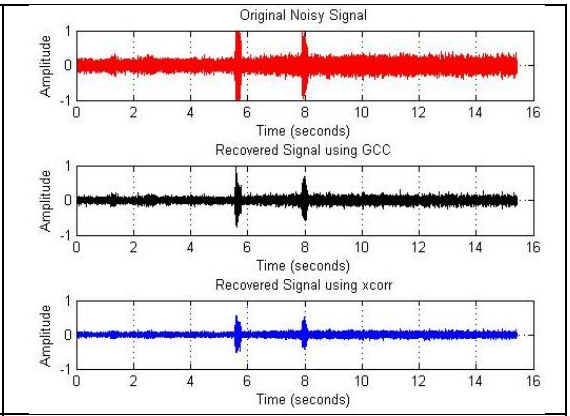
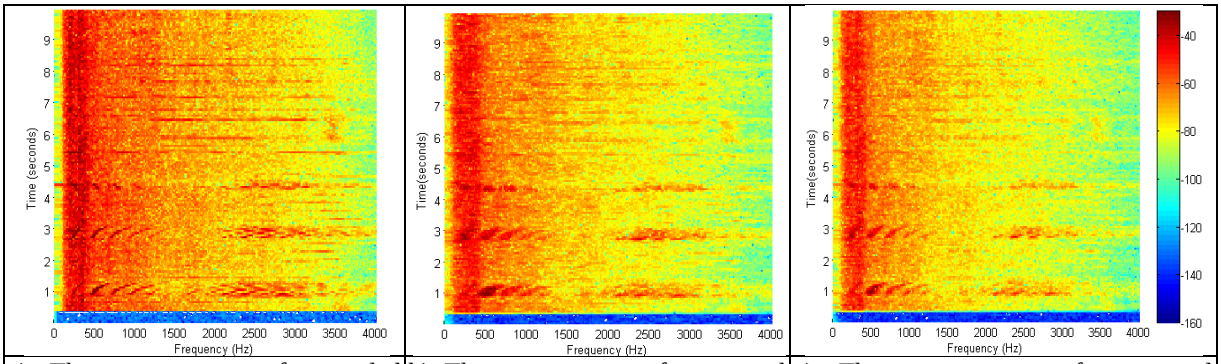
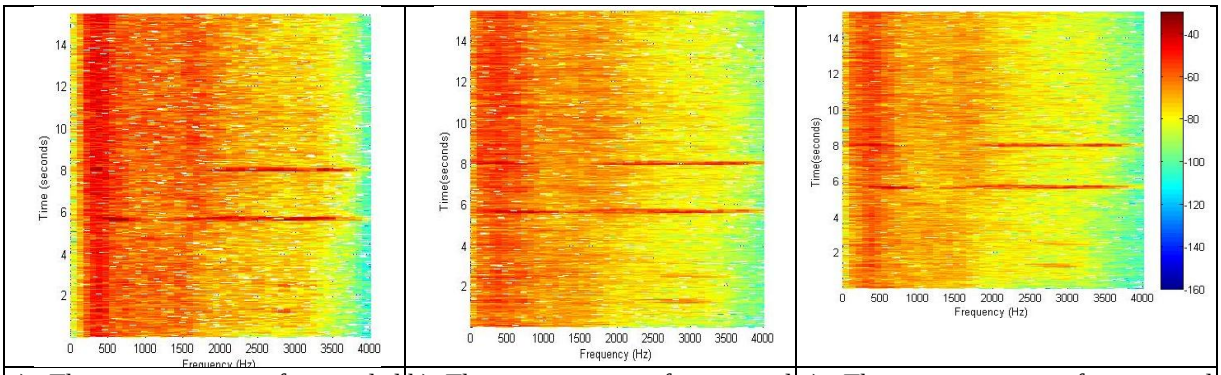


Fig. 13: Mobile phone recorded speech signal corresponding to alphabets /a/ and /b/, produced by male speaker in running car and the recovered signals using GCC (frequency domain) and *xcorr* (time domain).



a) The spectrogram of recorded noisy speech signal (Corresponding to Fig. 12).	b) The spectrogram of recovered speech signal using GCC.	c) The spectrogram of recovered speech signal using *xcorr*.

Fig. 14: The spectrogram of a) recorded speech signal (corresponding to Figure 12), b) recovered signal using GCC, and c) recovered signal using *xcorr*.



a) The spectrogram of recorded noisy speech signal (Corresponding to Fig. 12).	b) The spectrogram of recovered speech signal using GCC.	c) The spectrogram of recovered speech signal using *xcorr*.

Fig. 15: The spectrogram of a) recorded speech signal using using mobile phone in Fig. 13, b) recovered signal using GCC, and c) recovered signal using *xcorr*.

V. CONCLUSIONS

This paper presented a speech enhancement algorithm based on auto-correlation function. In this approach, it is expected that by discarding the more noise affected lower lag auto-correlation coefficients, an estimate of power spectrum of clean signal can be obtained from the magnitude spectrum of higher lag auto-correlation coefficients and finally the enhanced speech signal can be obtained using the overlap-add method. More specifically, it presented the results obtained on applying the auto-correlation based speech enhancement algorithm on the recorded speech signals with continuous background noises. The speech enhancement is performed by using auto-correlation in both time domain and frequency domain. From the results obtained, it is found that the time domain auto-correlation gives better results in reducing the noise from the speech signal and gives better quality recovered signals as compared to the frequency domain auto-correlation. With the auto-correlation based algorithm, the stationary noises (e.g., the moving fan noise and car noise) are reduced in the recovered speech signals; but, at the expense of small decrease in magnitude of the speech signal. However, the decrease in the magnitude of the recovered speech signal is uniform; thus, the algorithm maintains the intelligibility of the original speech signal. It is also observed from the results that the algorithm does not work well for non-stationary background noises, as these noises are still audible in the recovered signal. Therefore, it is concluded that this algorithm is capable of enhancing the speech signal of a single channel based devices and perform well for continuous or stationary noises.

REFERENCES

[1]. Z. Chen, "Simulation of Spectral Subtraction Based Noise Reduction Method," International Journal of Advanced Computer Science and Applications, Vol. 2, No. 8, 2011.

[2]. M. A. Hasan & T. Shimamura, "A Fundamental Frequency Extraction Method Based on Windowless and Normalized Autocorrelation Functions," Proc. 6th WSEAS Int. Conf. Circuits, Systems, Signal and Telecommunications, Cambridge, pp. 305-309, 2012.

[3]. M. Gabrea, "Two Microphones Speech Enhancement Systems based on Instrumental Variable Algorithm for Speaker Identification," Proc. Of the 24th Canadian Conference on Electrical and Computer Engineering, May 2011.

[4]. J. I. Hurtado & D. V. Anderson, "FFT-Based Block Processing in Speech Enhancement: Potential Artifacts and Solutions," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 8, November 2011.

[5]. B. J. Shannon, K. K. Paliwal, & C. Nadeu, "Speech enhancement based on spectral estimation from higher-lag autocorrelation", Proc. Interspeech, 2006.

[6]. L. D. Alsteris & K. K. Paliwal, "Importance Of Window Shape For Phase-Only Reconstruction Of Speech," International Conf. on Acoustics, Speech, and Signal Processing-ICASSP, Vol. 1, pp. I-573-6, 2004.

[7]. P. Basu, P. J. Wolfe, D. Rudoy, T. F. Quatieri & Bob Dunn, "Adaptive Short-Time Analysis-Synthesis for Speech Enhancement," International Conf. on Acoustics, Speech, and Signal Processing-ICASSP, pp. 4905-4908, 2008.

[8]. A. K. Swain & W. Abdulla, "Estimation of LPC Parameters of Speech Signals in Noisy Environment," TENCON, Vol. 1, pp. 139-142, 2004.

[9]. H. G. Kim, M. Schwab, N. Moreau & T. Sikora, "Speech Enhancement of Noisy Speech Using Log-Spectral Amplitude Estimator and Harmonic Tunneling," International Workshop on Acoustic Echo and Noise Control, Japan, September 2003.

[10]. K. A. Sheela and K. S. Prasad, "Less Computational Complex Method for Estimation of Non-Stationary Noise in Speech Using Low frequency Regions for Power Spectral Subtraction," National conference on Communications (NCC2007) IIT Kanpur, January 2007.

[11]. T. Fingscheidt, S. Suhadi, & S. Stan, "Environment-Optimized Speech Enhancement," IEEE Transactions on Audio, Speech & Language Processing, pp. 825-834, 2008.

[12]. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5, pp. 504-512, July 2001.

[13]. M. K. Hasan, S. Salahuddin & M. R Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," IEEE Signal Processing Letters, Vol. 11, No. 4, pp. 450-453, April 2004.

[14]. S. Ogata, S. Ebataz & T. Shimamuray, "Improved Model of SPAC (Speech Processing System by Use of Auto-Correlation Function) Utilizing Spectral Subtraction as Preprocessing," Proc. of ICA 2004: The 18th international congress on acoustics, Kyoto, Japan, pp. IV-3037 - IV-3040, April 2004.

[15]. Y. Hu, M. Bhatnagar & P. C. Loizou, "A Cross-Correlation Technique for Enhancing Speech Corrupted with Correlated Noise," Proc. of IEEE International Conference on Acoustic Speech Signal Processing, Vol. 1, Salt Lake City, UT, pp. 673-676, 2001.

[16]. J. B. Boldt, D. Ellis, "A Simple Correlation-Based Model of Intelligibility for Nonlinear Speech Enhancement and Separation," Proc. of the 17th European Signal Processing Conference EURASIP, 2009.

[17]. M. A. Hasan & T. Shimamura**, "**An Efficient Pitch Estimation Method Using Windowless and Normalized Autocorrelation Functions in Noisy Environments," International Journal of Circuits, Systems and Signal Processing, Vol. 6 No. 1, pp. 197-204, 2012.