Volume 6, June 2017

Jahangirnagar University

Journal of Information Technology (JIT)

ISSN 2227-1279 (Print Version), 2304-3237 (Online Version)

Institute of Information Technology

Jahangirnagar University, Dhaka, Bangladesh

Jahangirnagar University

Journal of Information Technology

A Research Publication of Institute of Information Technology,
Jahangirnagar University

Volume 6, June 2017

Editorial Board

Editor

M. Mesbahuddin Sarker

Members

Jesmin Akhter
Risala Tasin Khan
Dr. Mohammad Abu Yousuf
Dr. Mohammad Shahidul Islam

The Jahangirnagar University Journal of Information Technology is peer-reviewed online and printed scholarly journal, which gives a space for researchers by publishing articles and reviews of current work in the field of Information and Communication Technology. It is published annually in June.

Copyright @2017 by Jahangirnagar University Journal of Information Technology (JIT).

All rights reserved,

However permission is granted to quote and photocopy any part or full of an article for educational or research purpose to individuals, institutions and libraries with an appropriate citation in the reference and/or customary acknowledgment of the journal.

ISSN 2227-1279 (Print Version), 2304-3237 (Online Version)

Annual fees for institution USD 25 (postage included) or individual USD 10 (postage included) or BDT 200 (local researchers).

Further information on subscriptions and the JU Journal of Information Technology (JIT) can be found at
http://www.juniv.edu/jujit/

Cover and Website:

Mr. A. N. M. Hasanat Tanvir, IIT, Jahangirnagar University

Published by:

Institute of Information Technology, Jahangirnagar University, Dhaka-1342

Institute of Information Technology of Jahangirnagar University is pleased to present the sixth volume of JU Journal of Information Technology (JIT). We have received 15 papers for publication, 9 papers being accepted (66.6%).

We continuously strive to publish original research that contains elements of technical novelty. The journal focus on analytical and simulation based work on Information and Communication Technology (ICT), In order to publish a high quality journal editorial board seeks excellent contributor containing original research work or review. Our editorial board work tirelessly to provide contributors with a prompt and through double blind review process.

We would like to extend our heartfelt thanks to every author, reviewer and reader for their dedication and support to JIT. We strongly believe that together, we shall elevate the journal to ever high levels of quality, impact and reputation.

Editorial Board

JU Journal of Information Technology

The following is a list of reviewers who have given their precious times and expertise for the publications of volume 6 of Jahangirnagar University Journal of Information Technology (JIT).

- Md. Mustafizur Rahman, Dept. Mathematics, Bangladesh University of Engineering and Technology (BUET),
- ATM Mahbubur Rahman, Dept. of CSE, Dhaka International University, Bangladesh.
- Md. Sheikh Sadi, Dept. of CSE, Khulna University of Engineering and Technology (KUET), Bangladesh.
- Mohamed Ruhul Amin, Department of Electronics and Communications Engineering, East West Universty, Bangladesh.
- Md. Mustafizur Rahman, Department of CSE, University of Dhaka, Bangladesh.
- Haris Gacanin, aleatel-Lucent, Antwerp, Belgium.
- Md. Abdur Razzaque, Department of Computer Science and Engineerig, University of Dhaka, Bangladesh.
- Tapan Kumar Godder, Department of ICE, Islamic University, Bangladesh.
- Muhammad Shorif Uddin, CSE, Jahangirnagar University, Bangladesh.
- Shahdat Hossain, Dept. of Biotechnology and Genetic Engineering, Jahangirnagar University, Bangladesh.
- Sanjeewa P. Herath, Electrical Engineering at McGill University, Canada.
- Md. Imdadul Islam, CSE, Jahangirnagar University, Bangladesh.
- Sharif Akhteruzzaman, Department of Genetic Engineering and Bio-Technology, University of Dhaka, Bangladesh.
- Mohammad Motiur Rahman, Dept. of CSE, Mawlana Bhashani Science and Technology University, Bangladesh.
- Mostofa Kamal Nasir, Dept. of CSE, Mawlana Bhashani Science and Technology University, Bangladesh.
- Md. Whaiduzzaman, Institute of Information Technology, Jahangirnagar University, Bangladesh.
- Mufti Mahmud, Dept. of Biomedical Sciences, University of Padova, Italy.
- Md. Mijanur Rahman, Department of CSE, Jatiya Kabi Kazi Nazrul Islam University, Bangladesh.
- Muhammad Arifur Rahman, University of Sheffield, United Kingdom.
- Mohammed Shahedur Rahman, Dept. of Biotechnology and Genetic Engineering, Jahangirnagar University, Bangladesh.
- Kazi M. Ahmed, Telecommunications, AIT, Thailand.
- Shamim Ahmed, Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Bangladesh.
- Md. Fazlul Karim Patwary, IIT, Jahangirnagar University, Bangladesh.
- Mohammad Abu Yousuf, IIT, Jahangirnagar University, Bangladesh.

JAHANGIRNAGAR UNIVERSITY JOURNAL OF INFORMATION TECHNOLOGY (JIT)

VOLUME 6, 2017

C O N T E N T S

Jahangirnagar University

Journal of Information Technology (JIT)

Volume 6, June 2017

CONTENTS

JJIT, Vol. **6**, June, 2017, 1-17

ANALYSIS OF EEG SIGNALS AND EXTRACTION OF VISUAL EVOKED POTENTIAL FROM THE EEG SIGNALS: COMPARATIVE STUDY

MOHAMMAD ABU YOUSUF[1] AND MOHAMMAD BADRUL ALAM MIAH[2]

[1]*Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh*
[2]*Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail, Bangladesh*

Abstract

An electroencephalogram (EEG) is a test that measures and records the electrical activity of the brain. The objective of this paper is to extract single trial Visual Evoked Potential (VEP) from a highly noisy brain activity. Generally, the desired VEP is corrupted by background electroencephalogram (EEG). The common method for separating VEP from EEG is to use signal averaging. But the averaging method is not good always to give appropriate VEP in single trial EEG. That‟s why different analysis has to be accomplished in order to extract smooth VEP from the given specific EEG signals. In this paper, different approaches are used to analysis EEG signals and to extract VEP from the specific EEG signals. Principal Component Analysis (PCA) is applied to remove artifacts from multichannel EEG and to extract VEP from EEG signals. Independent Component Analysis (ICA), is also applied to single-trial multichannel EEG signals and VEP has been extracted from the noisy signals.

Keywords : Electrocardiogram (ECG), Visual Evoked Potentials (VEP), Independent Component Analysis (ICA), Principal Component Analysis (PCA).

1. Introduction

The Electrocardiogram (ECG) is the most commonly used biomedical signal [1]. The electroencephalogram (EEG) is a set of data measured by electrodes placed on the scalp and is always under the influences of artifacts. The occurrence of artifacts, such as eye blinks, muscle activity, line noise, and pulse signals, in electroencephalographic recordings obscures the underlying processes and makes EEG analysis difficult. Large amounts of data must often be discarded because of contamination by these artifacts. Therefore, the noise removal is of the prime necessity to make easier data interpretation and representation, and to recover the signal that matches perfectly a brain functioning [2]. To overcome this difficulty, signal processing techniques are used to remove artifacts from the EEG data of interest.

Visual Evoked Potentials (VEP) represents the EEG response to stimulation triggered by a series of flashes presented to a human subject‟s eye. Basically, VEP are the signals

generated in the brain in response to visual stimulus. Like many neural signals, VEP measurements are very weak signals and strongly corrupted by background noise. Its analysis has become very useful for neuropsychological studies and clinical purposes. In a hospital, a visual evoked potential test remains as the only objective test [3] to assess the physiology (i.e., conduction) of the visual or optical pathway from the retina to the occipital cortex of the brain. The VEP signal is embedded in the ongoing EEG with additive noise causing difficulty in detection and analysis of this signal. Furthermore, SNR of VEP to EEG is very low, which complicates the situation further. The traditional method of ensemble averaging is commonly used to extract the meaningful VEP signals from a noisy background. This technique is based on averaging most of the signals recorded during the test until a clean plot of the VEP is obtained. This often requires the acquisition of several unit brain responses [4] [5]. While averaging of EEG, some important features of the response, it is known that some information is lost in the process [6]. Boaz Sadeh and Galit Yovel describe a methodology for combining three tools: TMS, EEG, and fMRI to extract VEP from EEG [7]. M. Z. Yusoff explained an optimization-and Karhunen-Loeve Transform (KLT)- based approach to estimate the latencies of single-trial visual evoked potentials (VEPs) which are highly corrupted by colored electroencephalogram (EEG) noise [8].

In this paper, we analyze the EEG signals and applied different approaches to extract the VEP from the very noisy signals. At first, the traditional method of ensemble averaging is applied to extract smooth VEP from the noisy EEG signals. PCA is a statistical technique commonly employed to reduce the dimension of the feature set [9]. It has also been used to reduce noise from biomedical signals [10]. PCA approach is also applied in this paper to analysis EEG signals and to extract VEP. A new linear decomposition tool, ICA is also applied to single-trial multichannel EEG signals and VEP has been extracted from the noisy signals. ICA is suitable for performing blind source separation on EEG data because it is plausible that EEG data recorded at multiple scalp sensors are linear sums of temporally independent components arising from spatially fixed, distinct or overlapping brain or extra-brain networks. Results show that ICA can separate artifactual, stimulus-locked, response-locked, and non-event-related background EEG activities into separate components, taxonomy not obtained from conventional signal averaging approaches.

2. VEP Model

It is assumed that a VEP is actually a “known" waveform which can be artificially produced. The created VEP will then be added to much higher power “colored noise” that represents EEG and other background noise. The resultant waveform will be treated as a composite signal that needs to be processed and extracted using the developed technique to get back the desired VEP. Thus, the following model is defined.

$$y = x + n \tag{1}$$

where, $\boldsymbol{y}$ is the M-dimensional vector of the corrupted (noisy) VEP signal; $\boldsymbol{x}$ is the M-dimensional vector of the original (clean) VEP signal; $\boldsymbol{n}$ is the M-dimensional vector of the additive EEG noise which is assumed to be uncorrelated with $\boldsymbol{x}$.

3. Methodology & Results

In this section, we discuss different approaches that we have used to remove artifacts and to extract VEP. These approaches include Averaging, ICA and PCA. EEG data were recorded from 31 scalp electrodes, 29 placed at locations based on a modified International 10-20 system, one placed below the right eye (VEOG), and one placed at the left outer canthus (HEOG). Fig. 1 shows the signals where each channel length is taken as 10000 from the original length of 130240.

The top two PCs represent CG and EOG artifacts. Before extraction process of VEP from EEG data, we apply a band pass filter such as Butterworth filter where the frequency band is from 0.1 to 30 Hz. We know that VEP lies in low frequency band where the high frequency components are considered as noise.

3.1. *Averaging Approach*

The very common method for extracting VEP from EEG is to use signal averaging. Fig.2 represents the average template of 29 channels. For that purpose, for every channel we have followed two steps:

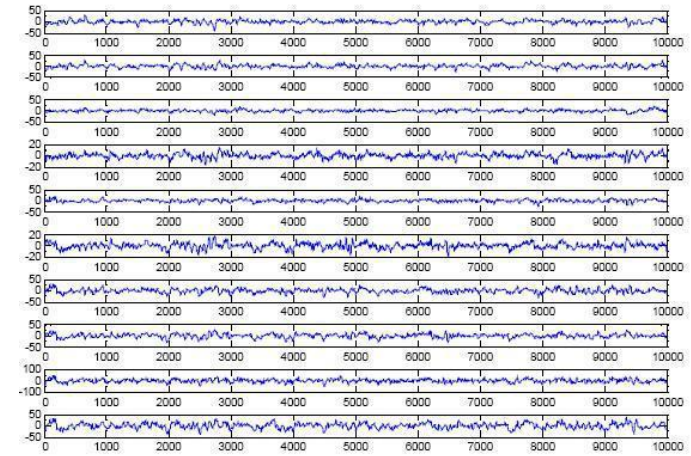
a) From every trigger point collect 300 samples and store them in a buffer and
b) Do the average of buffer.

3.2 PCA Approach

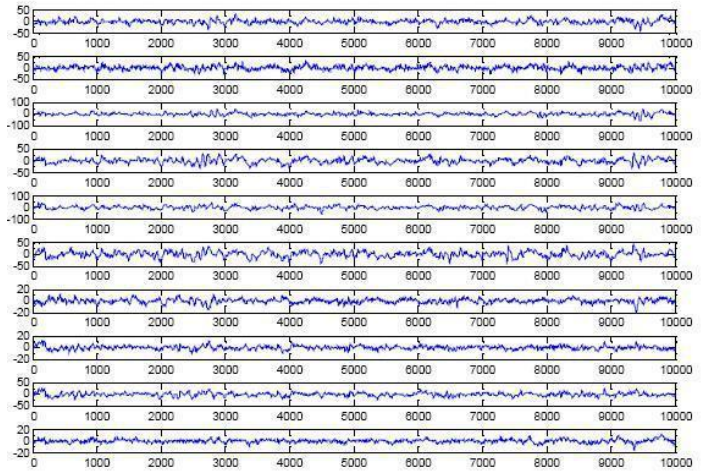
PCA is popular method to approximate original data with lower dimensional feature space. The fundamental approach is to compute the eigenvectors of the covariance data matrix Q and then approximation is done using the linear combination of top eigenvectors. The covariance matrix of the samples and the principle components of the covariance matrix can be calculated respectively as

$$E^T Q E = A \qquad (2)$$

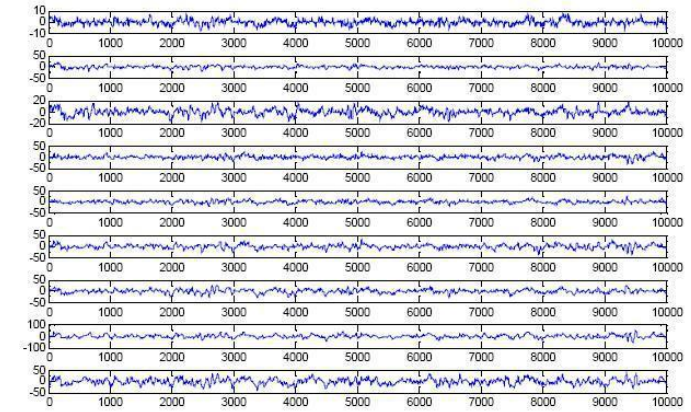
Where E represents the matrix of orthonormal eigenvectors and A diagonal matrix of the eigenvalues. Usually, the eigenvalues that are about to zero values carry negligible variance and hence can be excluded. So, the m eigenvectors corresponding to the largest eigenvalues can be used to define the subspace.



(a)



(b)



(c)

Fig. 1: (a) Channel 1-10, (b) Channel 11-20, and (c) Channel 21-29 from the EEG

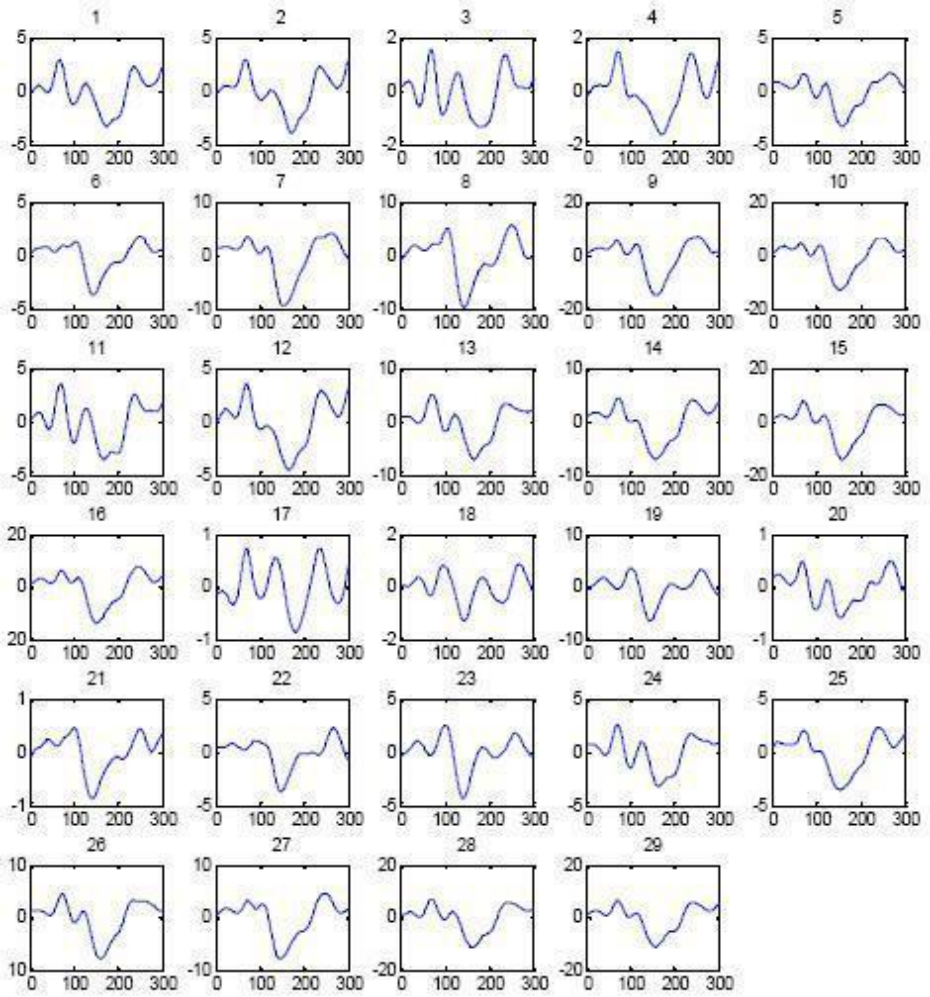


Fig. 2: Average template of 29 channels.

The proposed algorithm using PCA is divided into two parts:

1. Removing noise from the signal
2. Identification of each VEP for each signal

3.2.1: Removing noise from the signal

Steps for removing noise from the signal are as follows:

i) Load all signal:

In first step load all 29 signals

ii) Apply low pass filter on all signal:

Butterworth low pass filter is applied for filtering. Because the Butterworth filter provides the best Taylor series approximation to the ideal low pass filter. Here cutoff frequency as 40 Hz is used.

iii) Get the filtered signal:

After applying low pass filter filtered signal is available.

iv) Apply PCA to all filtered signal:

a) At first covariance matrix is calculated.
b) PCA on the p-by-p covariance matrix is performed and returns the principal component coefficients. The values of principal component coefficients are in order of decreasing component variance.
c) After applying PCA on covariance matrix, it also returns a vector containing the principal component variances, that is, the eigenvalues of covariance matrix. I got 29 eigenvalues for 29 eigenvectors. The eigenvalues are in decreasing order.
d) In all eigenvalues, the first tow eigenvalues are very large than compared to the others eigenvalues. I assumed that these two contains the noise. Then the signals for these two eigenvectors are reconstructed and then subtracted

v) Get output signal:

After applying PCA noise free output signal is available.

Fig. 3 and Fig. 4 show the eigenvalues of principle components and top two principle components that represents the CG and EOG artifacts. As the top two PCs represent noise, that‟s why we remove these two PCs from every channel to remove these artifacts.

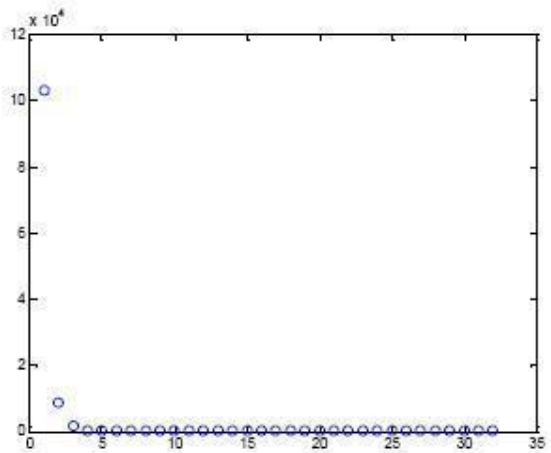


Fig. 3: Eigenvalues corresponding to the 32 eigenvectors.

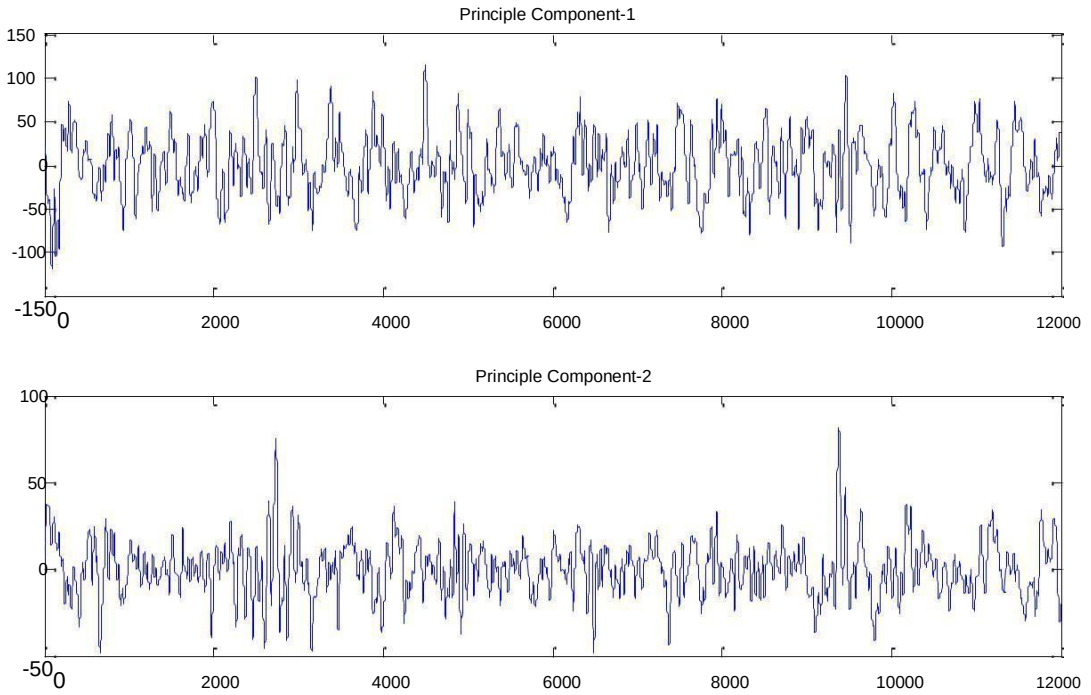


Fig. 4: Top two PCs after analyzing PCA over 29 channels.

3.2.2: Identification of each VEP for each signal

Using the output singles, perform the following operations:

a) Select one signal.

b) Divide signal data into "I" blocks.

c) Average all "I" blocks to get a template "t".

d) Starting from beginning of the signal, slide the template across the signal.

e) Calculate the cross correlation in every successive point to find a set of best matches for the visual evoked potential (VEP) identification in that channel.

f) Do step from 1 to 5 for other signal also.

Fig. 5 and Fig. 6 depicts the effect of PCA and the VEP of first 40 trials of channel 9 and 10 respectively.

3.3: ICA Approach

Basically, ICA algorithm is a blind source separation problem where the objective is to decompose an observed signal into a linear combination of some unknown independent signals.

The basic idea is to represent a set of random observed variables using basis function where the components are statistically independent. The ICA algorithm finds the statistically independent basis. If S is collection of basis and X is collection of input channels then the relation between X and S is modeled as shown in equation 3.

$$X = MX \tag{3}$$

where M represents an unknown linear mixing matrix of full rank. It is assumed that the sources are independent of each other and the mixing matrix is invertible and now based on these ICA algorithm tries to find out the mixing matrix M or the separating matrix so that

$$U = WX \tag{4}$$

$$U = WMS \tag{5}$$

where U is an estimation of the independent sources. The estimation problem of finding M can be simplified by a pre-whitening of the observed vectors X. In order to do that, first X is linearly transformed to a matrix Y such that

$$Y = RX \tag{6}$$

$$R = A^{\frac{1}{2}} \tag{7}$$

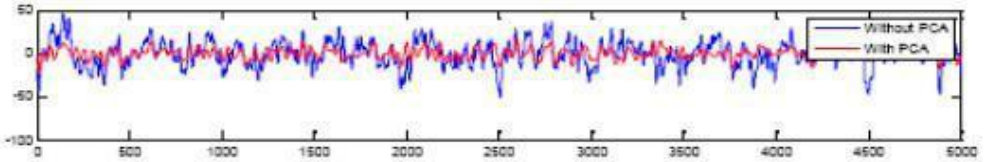
where the correlation of Y is an unit matrix. A substitutes the diagonal matrix of the eigenvalues and E orthonormal eigenvectors of the covariance matrix of X. After transforming the sample vectors X to Y, we have

$$Y = RX \tag{8}$$

$$Y = RMS \tag{9}$$

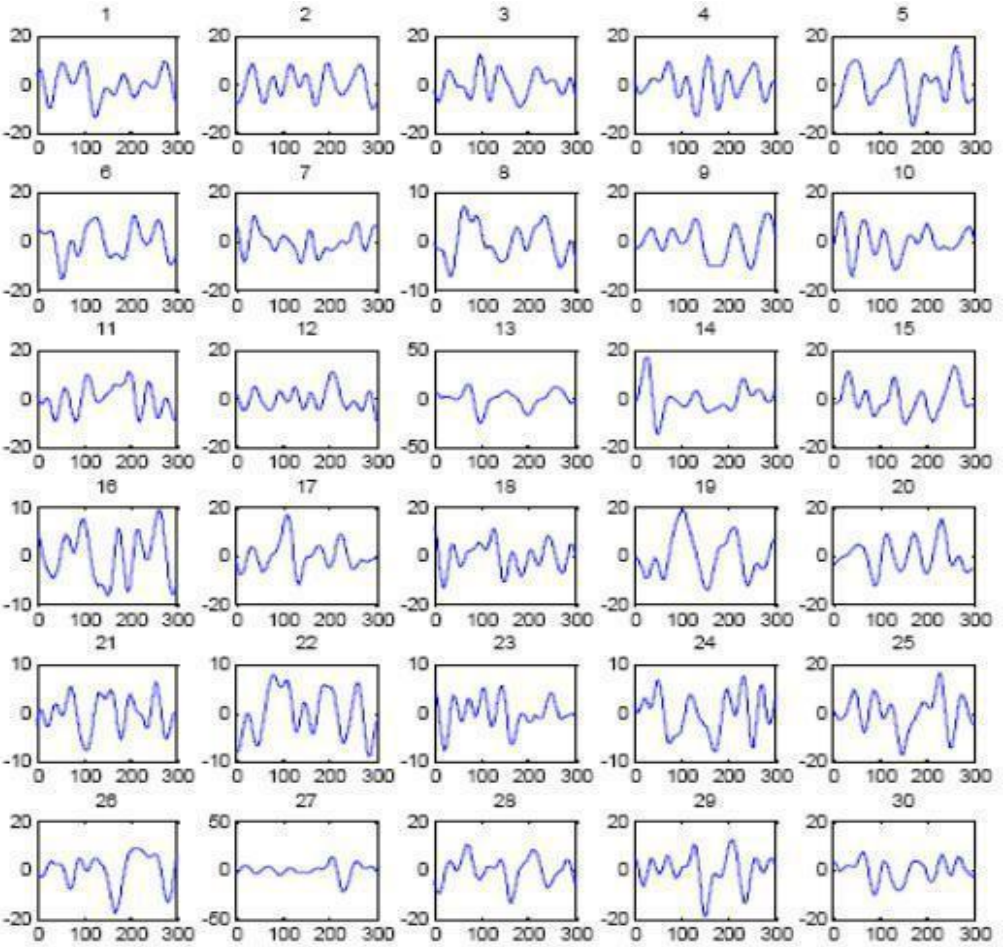
$$Y = BS \tag{10}$$

where B is the estimation of the unknown mixing matrix. ICA is applied on the transformed matrix Y now. In brief, The ICA algorithm learns the weight matrix W, the inverse of mixing matrix M, is used to recover the set of independent basis S. Normally, in ICA approach, most of the papers used ICA and selected the ICs that include the noise to remove and reconstruct the signal to get VEP.



Time (msec)

(a)



Time (msec)

(b)

Fig. 5: Channel-9 before and after PCA and (b) Forty epochs of channel-9 after subtracting top 2 PCs

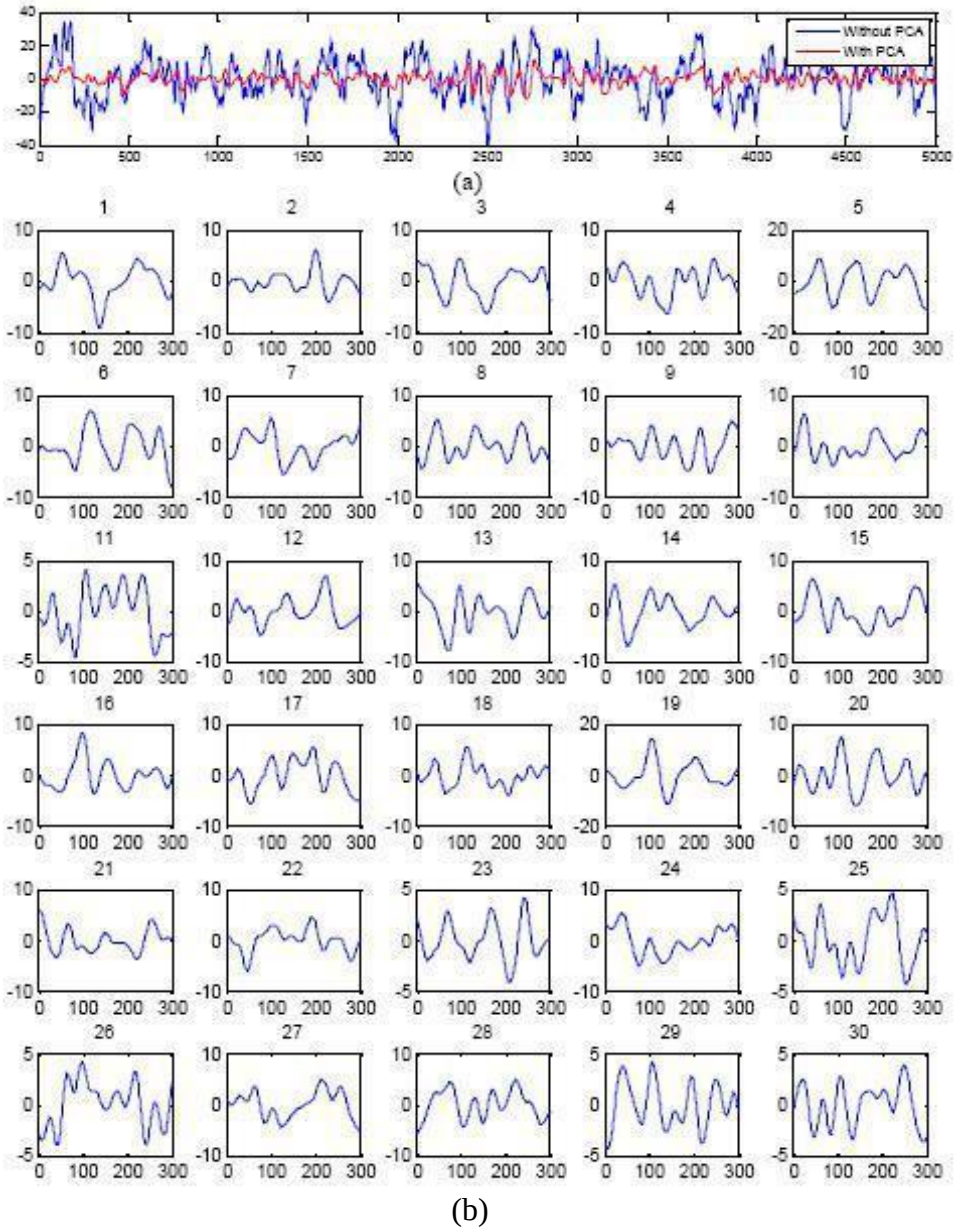


Fig. 6: (a) Channel-10 before and after PCA and (b) Forty epochs of channel-10 after subtracting top 2 PCs

Fig.7 shows 29 ICs after applying ICA over twenty nine channels. We already have come to know that, the two top PCs represent CG and EOG artifacts. As a result we can use them here as reference to remove the ICs from IC space. For that purpose, we have chosen the correlation approach. First, the top PC is taken and the correlation coefficients are measured with every IC. If the correlation coefficients are greater than a threshold i.e. 0.25

here, the ICs are selected to be removed. We have applied the same technique for EOG artifacts i.e. the second PC is selected as reference this time. Thus we have selected 8 ICs to remove. Fig. 7, Fig. 8 and Fig. 9 represents 1 to 10, 11 to 20, 21 to 29 ICs respectively after ICA over the first 29 EEG channels. Fig.10 and Fig. 11 depicts the effect of ICA and the VEP of first 40 trials of channel 9 and 10 respectively.

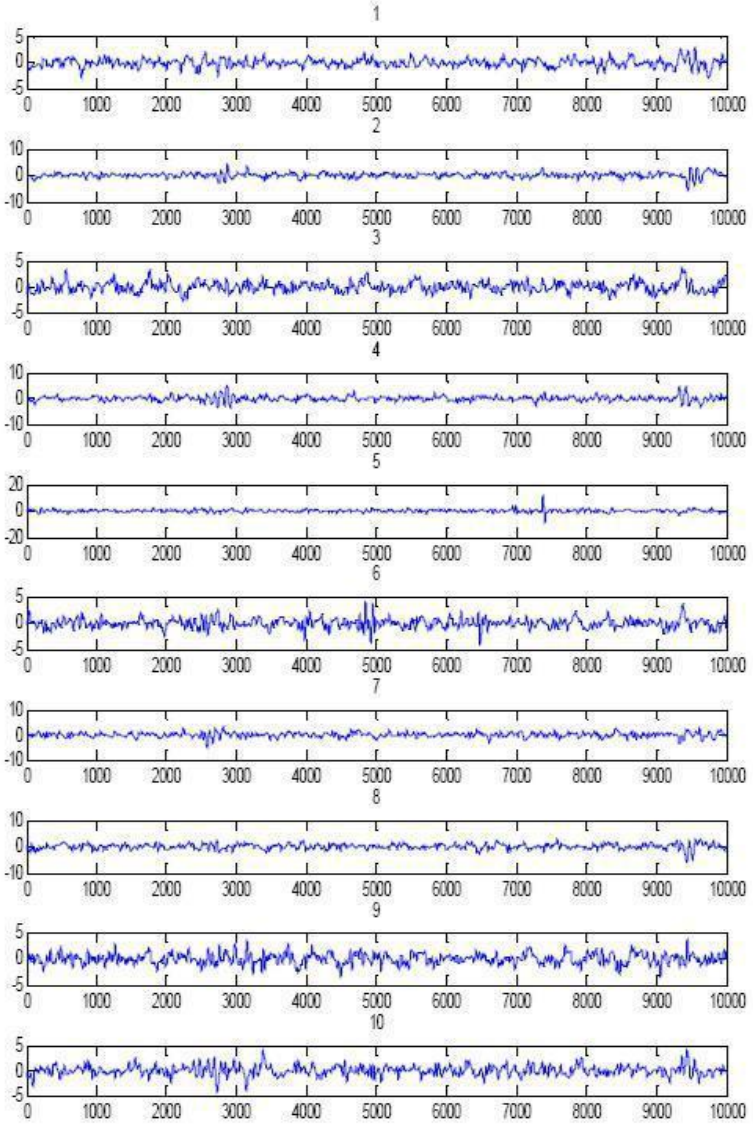


Fig. 7: IC 1-10 after applying ICA on 29 EEG channels.

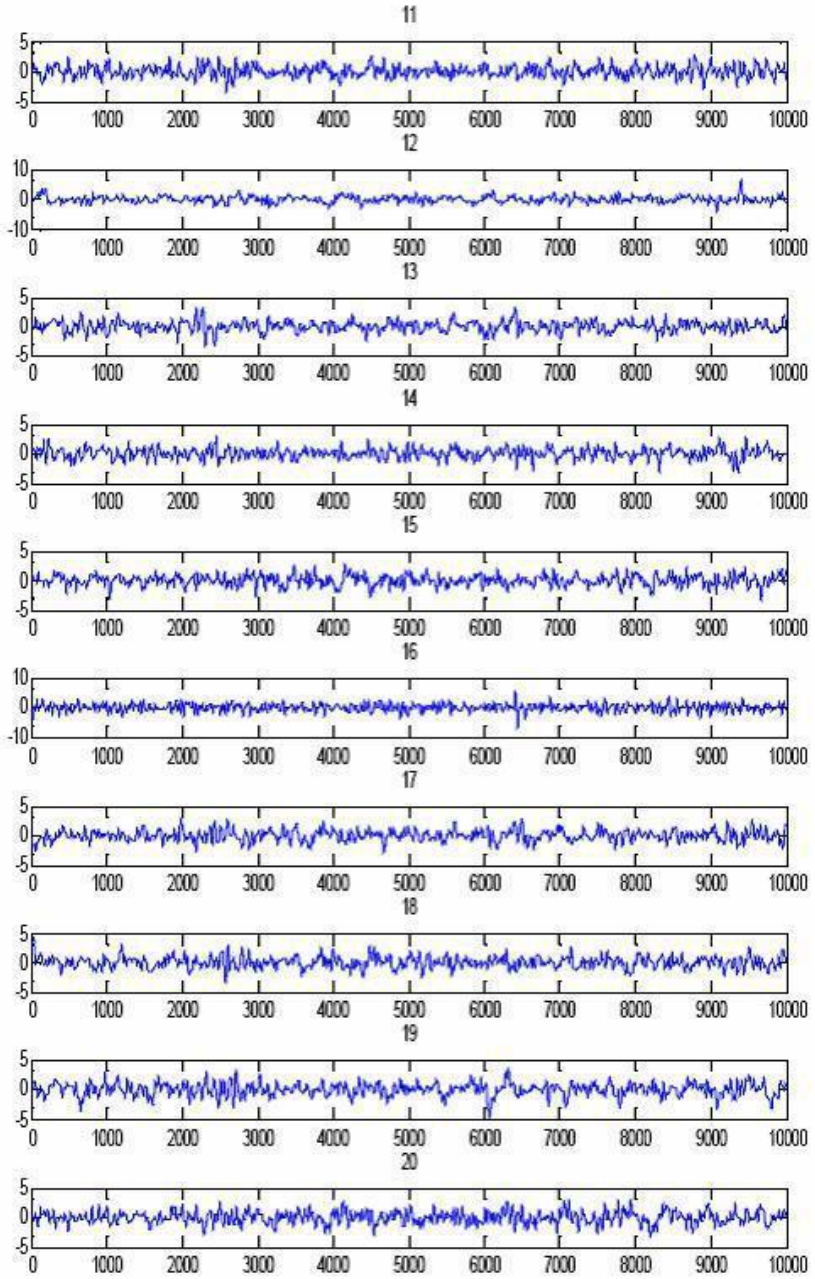


Fig. 8: IC 11-20 after applying ICA on 29 EEG channels.

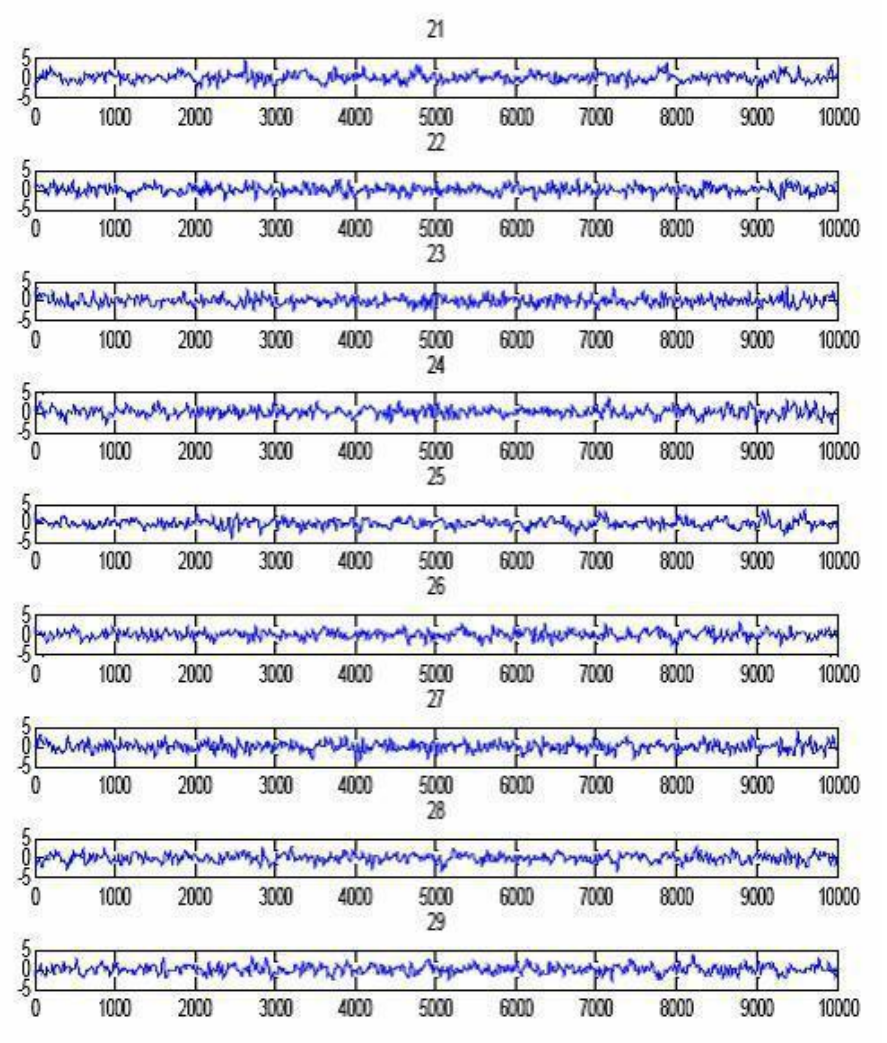


Fig. 9: IC 21-29 after applying ICA on 29 EEG channels.

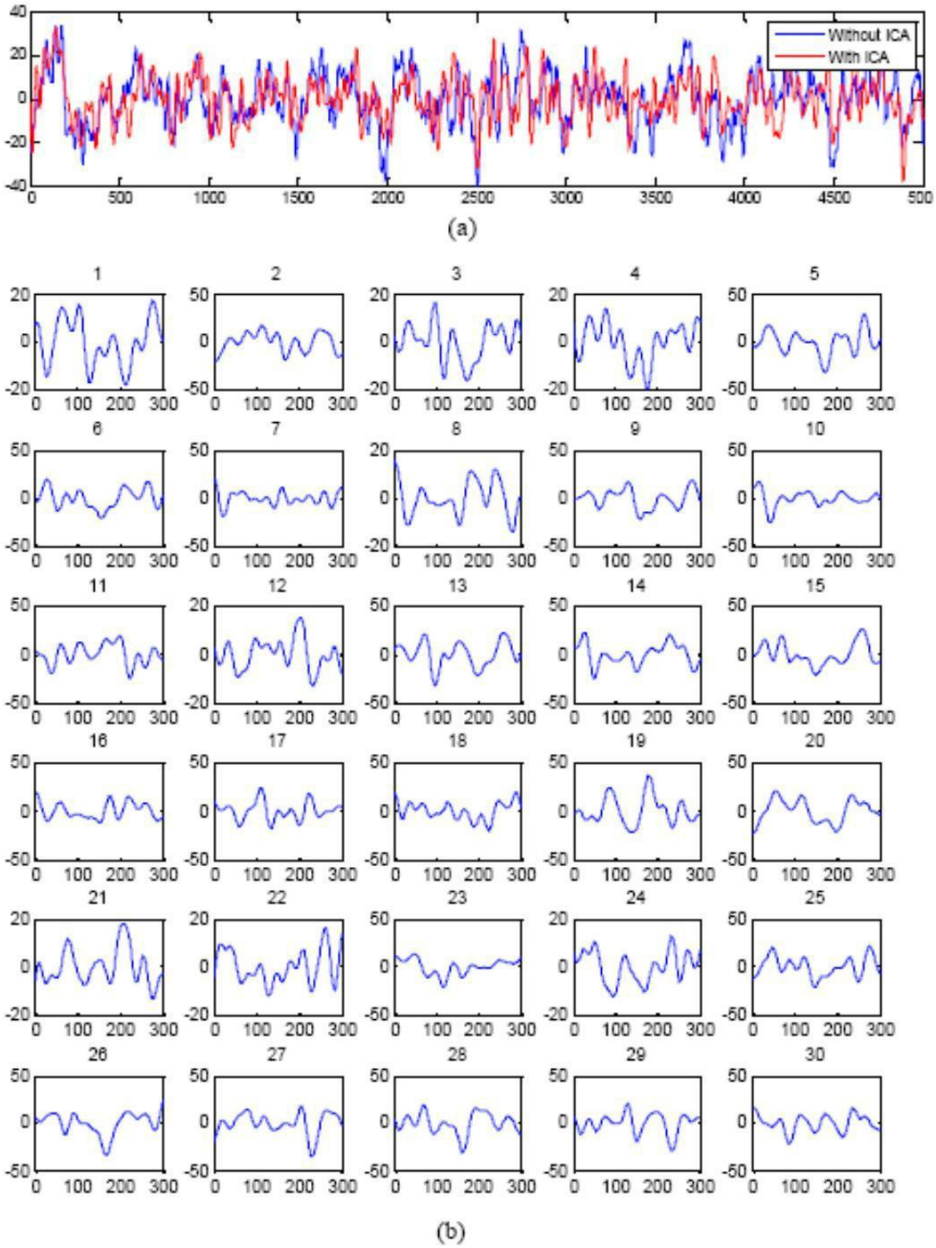


Fig. 10: (a) The channel-9 before and after applying ICA and (b)Thirty epochs of channel 9 after applying ICA.

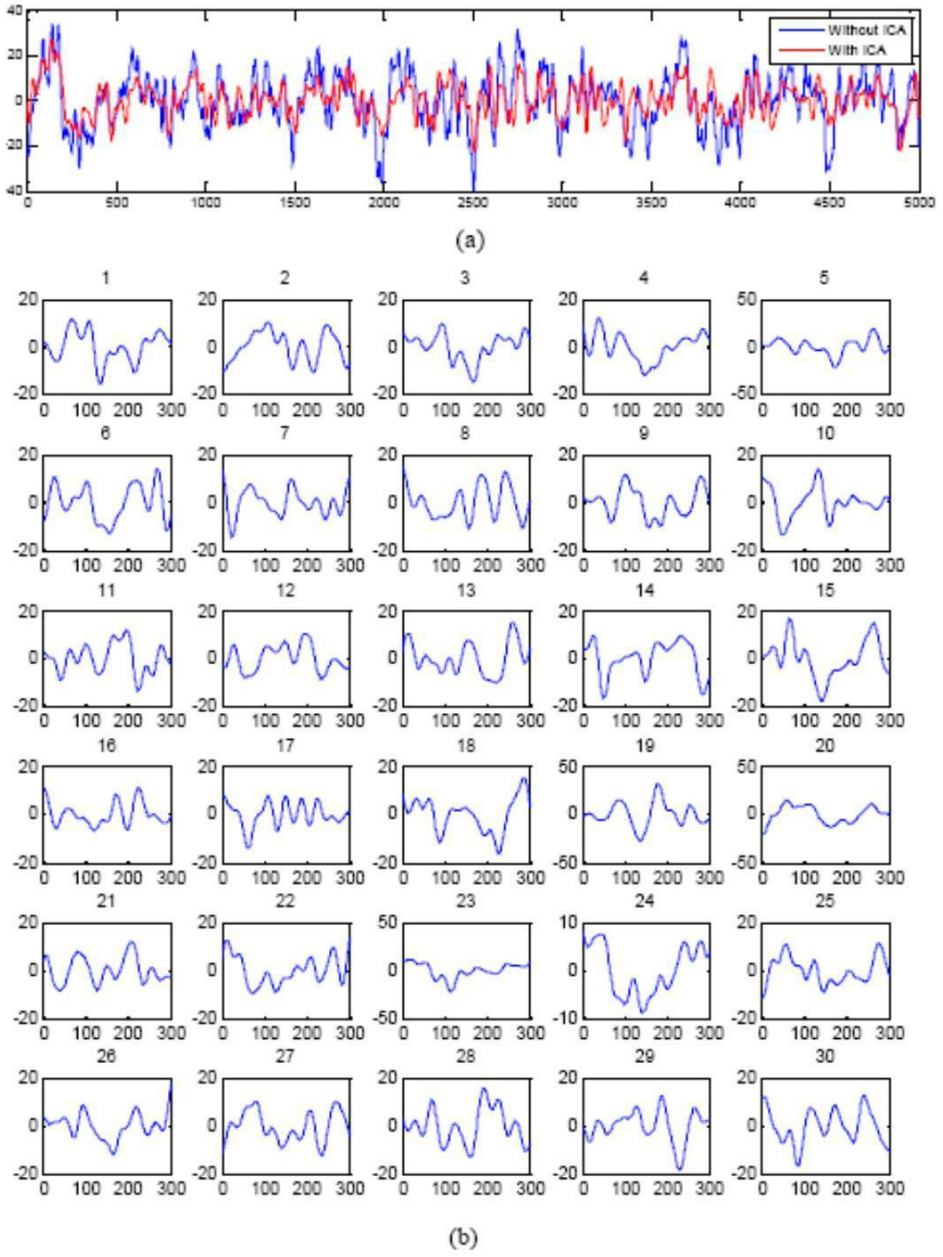


Fig. 11: (a) The channel-10 before and after applying ICA and (b) Thirty epochs of channel-10 after applying ICA.

4. Discussion and Conclusions

This work addresses the different approaches to analysis EEG signals and VEP has been extracted from the noisy signals. The most common method that we applied for extracting VEP from EEG is to use signal averaging. One of the disadvantages of the averaging method resides in the fact that the obtained signal's shape presents some amplitude distortions as the number of the averaged recorded single VEP increases. The method of ensemble average results therefore in a considerable loss of information.

PCA approach is also applied to remove background artifacts and to extract VEP from the EEG signals. The EEG signals consist of two parts: noise and EEG. Therefore, using PCA, it is possible to separate EEG part (i.e. signal part) from noise and background artifacts using the fact that the noise subspace will consist of principal components (PCs) with eigenvalues chosen below a certain threshold and eigenvalues with PCs above this threshold represent the EEG signal subspace. However, PCA cannot completely separate eye artifacts from brain signals, especially when they have comparable amplitudes.

Finally, we apply a new linear decomposition tool, ICA to multichannel single-trial EEG records to derive spatial filters that decompose single-trial EEG epochs into a sum of temporally independent and spatially fixed components arising from distinct or overlapping brain or extra-brain networks. But the problem of this conventional ICA approach is that ICA yields the ICs whose number matches the given number of channels. Then users must manually identify which ICs represent what sources. The results obtained with all of these approaches are encouraging which may be used for wide range of any clinical purposes.

References

1. Rangayyan RM. Biomedical Signal Analysis-A Case study Approach, Wiley-IEEE Press, New York, 2002.
2. P. LeVan, E. Urrestarrazu, J. Gotman, “A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification”, *Clinical Neurophysiology*, vol. 117(4), pp. 912-927, Apr. 2006.
3. L. Huszar, “Clinical Utility of Evoked Potential,” eMedicine, April 18, 2006, retrieved from http://www.emedicine.com/neuro/topic69.htm.
4. K.H Chiappa. Evoked Potentials in Clinical Medecine. Raven Press, New York, 1990.
5. C. M. Epstein, Introduction to EEG and Evoked Potentials. Lippincott Williams & Wilkins, 1983.
6. J. Zhang, C. Zheng, “Extracting Evoked Potentials with the Singularity Detection Technique”, IEEE Engineering in Medicine and Biology, pp 155-161, Sept. 1997.
7. B. Sadeh, G. Yovel, “Extracting Visual Evoked Potentials from EEG Data Recorded During fMRIguided Transcranial Magnetic Stimulation”, Journal of Visualized Experiments, pp 1-8, May 2014.

8. M. Z. Yusoff , “Estimation of visual evoked potential latencies using Karhunen Loeve Transform method”, 18th European Signal Processing Conference (EUSIPCO-2010), Denmark, pp. 1577-1581, August 23-27, 2010.
9. Gupta CN, Palaniappan R, Rajan S, Swaminathan S, Krishnan SM. “Segmentation and Classification of Heart Sounds”, Proceedings of 18th Annual Canadian Conference on Electrical and Computer Engineering, Saskatchewan, Canada, pp. 1674 – 1677, May, 2005
Canada, 2005; 1678–1681.
10. Sharmilakanna P, Palaniappan R. “EEG artifact reduction in VEP using 2-stage PCA and N4 analysis of alcoholics.”, Proceedings of 3rd International Conference on Intelligent Sensing and Information Processing, Bangalore, India, pp. 2-7, 2005.

JJIT, Vol. **6**, June, 2017, 19-26

SAVITZKY-GOLAY FILTER CONTROLLED SPECKLE REDUCTION ANISOTROPIC DIFFUSION FILTER FOR ULTRASOUND IMAGE

MOHAMMAD MOTIUR RAHMAN

Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University

Abstract

Speckle reduction anisotropic diffusion (SRAD) is the novel and widely used method for speckle noise reduction from ultrasound images. But edge preseravation of SRAD from speckled ultraound image is not satifactory. The instantaneous coefficient of variation (ICOV) of SRAD is responsible for edge detection. In this paper ICOV of SRAD is modified by Savitzky-Golay based filter. Qualitaive and quantitative analysis of the proposed method exhibits better result than conventional SRAD filter.

Keywords: SRAD, ICOV, Savitzky-Golay Filter, Speckle noise, Ultrasound Image

1. Introduction

Constructive and destructive interference effects characterize the ultrasound echoes from nonspecular reflections. Because the sound is reflected in all directions, there are many opportunities for waves to travel different pathways. The wave fronts that return to the transducer may constructively or destructively interfere at random. The random interference pattern is known as “speckle”[1]. This speckle degrades ultrasound image quality which is difficult for physicians to accurate decision about diagnosis.

Two important and frequently used methods to reduce speckle noise[] from ultrasound image are adaptive and diffusion techniques. An adaptive filter [2][3] is a digital filter that has self adjusting characteristics. It is capable of adjusting its coefficients automatically to adapt the input signal via an adaptive algorithm. Lee [4][5], Kuan [6], Frost[7] are the most important adaptive filters to remove speckle noise from US image. Diffusion filtering, which models the diffusion process, is an iterative approach of spatial filtering in which image intensities in a local neighborhood are utilized to compute new intensity values. Perona-Malik[8] and Speckle Reduction Anisotropic Diffusion (SRAD) are the most important diffusion filters to remove speckle noise from US image.

2. Speckle Reducing Anisotropic Diffusion (SRAD)

Speckle reduction anisotropic diffusion (SRAD) [9] is now the state of the art method for speckle noise reduction from ultrasound and SAR images. The strength of SRAD is that it combines the statistical information of speckle noise into anisotropic diffusion framework

to smooth homogeneous speckle regions while preserving image features. In SRAD the Lee and Frost filters can be cast as partial differential equations, and then SRAD formulation is derived by allowing edge-sensitive anisotropic diffusion within this context. Just as the Lee and Frost filters utilize the coefficient of variation in adaptive filtering, SRAD exploits the instantaneous coefficient of variation, which is shown to be a function of the local gradient magnitude and Laplacian operator. Given an intensity image $I_0(x,y)$ having finite power and no zero values over the image support **Ω**, the output image $I(x,y;t)$ is evolved according to the following PDE:

$$\partial t(x,y;t)/\partial t = div\big(c(q)\nabla I(x,y;t)\big)$$
$$I(x,y;0) = I_0(x,y), \quad (\partial I(x,y;t)/\partial \vec{n})|_{\partial\Omega} = 0$$

Where t represents diffusion time, $\partial\Omega$ denotes the border of Ω, $\vec{n}$ is the outer normal to the $\partial\Omega$, and

$$c(q) = \frac{1}{1+[q^2(x,y;t) - q_0^2(t)]/[q_0^2(t)\big(1+q_0^2(t)\big)]}$$

Where $q(x,y;t)$ is the instantaneous coefficient of variation (ICOV) determined by

$$q(x,y;t) = \sqrt{\frac{\left(\frac{1}{2}\right)(\nabla I/I)^2 - (1/4^2)(\nabla^2 I/I)^2}{[1+(1/4)(\nabla^2 I/I)]^2}}$$

And $q_0(t)$ is the factor of speckle or speckle scale function. The ICOV $q(x,y;t)$ serves as the edge detector in speckled imagery. The function exhibits high values at edges or on high contrast features and produces low values in homogeneous regions. The speckle scale function $q_0(t)$ effectively controls the amount of smoothing applied to the image by SRAD. It is estimated using

$$q_0(t) = \frac{\sqrt{var\,[z(t)]}}{\overline{z(t)}}$$

Where $var[z(t)]$ and $\overline{z(t)}$ are the intensity variance and mean over a homogeneous area at t, respectively. For calculating $q_0(t)$ SRAD needs a homogeneous area of the processed image, it is not trivial for a computer. So the $q_0(t)$ can be approximated by

$$q_0(t) = q_0 \ \exp[-\rho t]$$

So the proposed iteration formula of diffusion is defined with

$$I_{i,j}^{n+1} = I_{i,j}^{n} + \frac{\Delta t}{4} d_{i,j}^{n}$$

And is defined with

$$d_{i,j}^{n} = c_{i+1,j}^{n}\left(I_{i+1,j}^{n} - I_{i,j}^{n}\right) + c_{i,j}^{n}\left(I_{i-1,j}^{n} - I_{i,j}^{n}\right) + c_{i,j-1}^{n}\left(I_{i,j+1}^{n} - I_{i,j}^{n}\right) + c_{i,j}^{n}\left(I_{i,j-1}^{n} - I_{i,j}^{n}\right)$$

Where ρ is a constant to slow down the decrease of q_0 while the algorithm is iterating.

SRAD can preserve edges even enhance edges; however this character or function highly depends on the precision of edge detecting. If the edge is not detected, the edge will not be enhanced and even smoothed.

3. Savitzky- Golay Filter

The fundamental idea is to fit a different polynomial to the data surrounding each data point. The smoothed points are computed by replacing each data point with the value of its fitted polynomial [10]. Numerical derivatives come from computing the derivative of each fitted polynomial at each data point. While fitting polynomials for these purposes is obvious, the surprising part is that the polynomial coefficients can be computed with a linear filter. For smoothing, only one coefficient of the polynomial is needed, so the whole process of least squares fitting at every point becomes a simple process of applying the appropriate linear filter at every point. The size of the smoothing window is NXN where N is odd, and order of the polynomial to fit is K, where N> k +1. The general smoothing equation is

$$g_{x,y} = \sum_{j=-n}^{n} \sum_{i=-n}^{n} C_{i,j} \; f_{x+i,y+j}$$

N is equal to N-1/2. C is the convolution matrix and fxy is the original data. The advantages of the Savitzky-Golay filter has over moving average and other filters is its ability to preserves higher moments in the data and thus reduce smoothing on peak heights.

3.1: The Proposed Model

Most existing anisotropic diffusion models use intensity gradient to discriminate variance caused by noise or by image features. Because of the strong artifacts introduced by speckle, a gradient operator is not an effective edge detection scheme for ultrasound images. S-G

filter fits the polynomial to image data. For each filter the mask coefficients are constant and it operates on the overall image. The PDE of the proposed model has the form of

$$\partial t(x,y;t)/\partial t = div\big(c(q)\nabla I(x,y;t)\big)$$

$$I(x,y;0) = I_0(x,y)$$

$$c(q) = \frac{1}{1 + [\,q^2(x,y;t) - q_0^2(t)]/[q_0^2(t)\big(1 + q_0^2(t)\big)]}$$

The ICOV $q(x,y;t)$ serves as the edge detector in speckled imagery. In this case The ICOV function $q(x,y;t)$ is defined by

$$q(x,y;t) = \sqrt{\frac{\sigma \times (N \times N)}{Degree\ of\ the\ polynomial}}$$

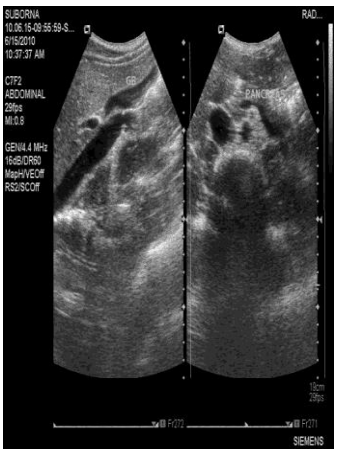
After operating the mask we get the mask value related to the centre

$$if\ Mask\ value < \sqrt{\frac{\sigma \times (N \times N)}{Degree\ of\ the\ polynomial}}$$

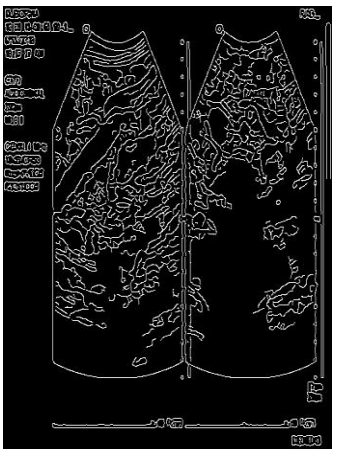
$$then$$

$$Mask\ value = median\ of\ mask\ coefficient$$

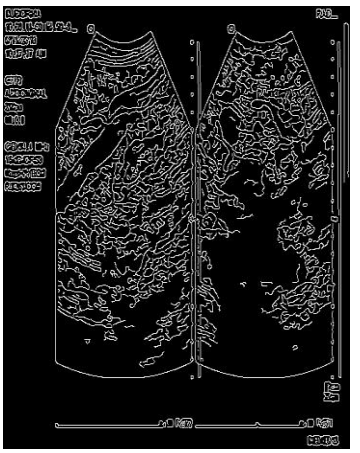
Improvement in real time US Image



Abdominal Image



Edges after SRAD Filtering



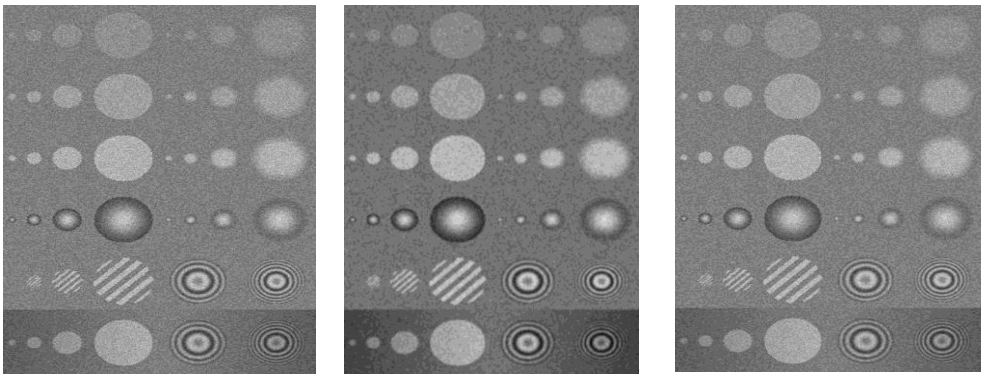
Edges after S-G based SRAD Filtering

We applied canny edge detector to both SRAD and S-G based SRAD filtered images. From the output it is clear that the proposed model edge detection capability is more robust than the state of the art SRAD.

4. Experimental Analysis

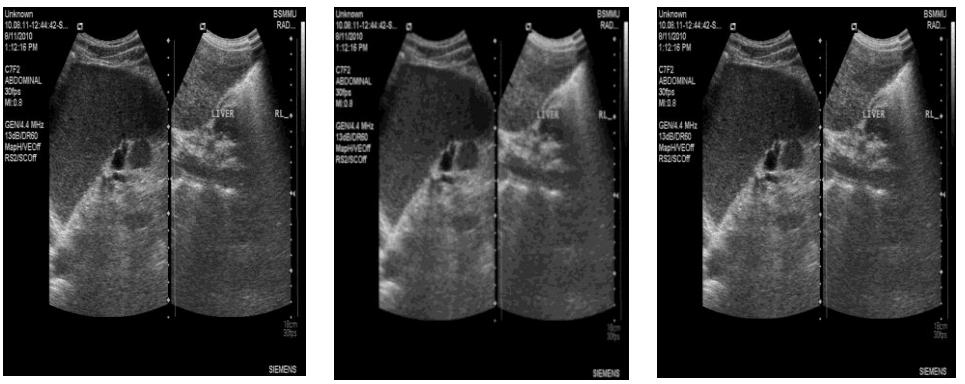
4.1 Qualitative analysis

Image 1- Phantom



Noisy SRAD Proposed SRAD

Image 2- Liver



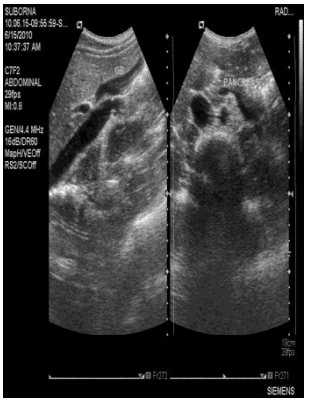
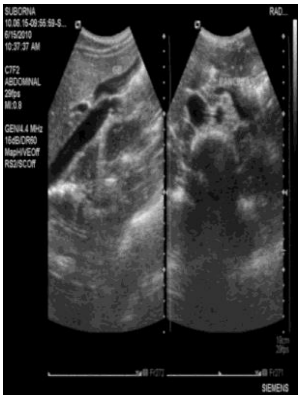
Noisy SRAD Proposed SRAD

Image 3- Kidney



Noisy SRAD Proposed SRAD

Image4- Abdominal


Noisy SRAD Proposed SRAD

4.2 : Quantitative Analysis

Quantitaive analysis of conventional SRAD and proposed Savitzky Golay filter based SRAD are given in the following tables.

Methods	Synthetic (Phantom) and Real Ultra Sound Image											
	Phantom			US Image Liver			US Image Kidney			US Image Abdominal		
	MSE	SNR	EPF	MSE	SNR	EPF	MSE	SNR	EPF	MSE	SNR	EPF
SRAD	.003	18.16	.463	.0018	15.67	.760	.0023	16.71	.743	.002	15.86	.765
Proposed	1.88	21.25	.998	2.538	17.90	.998	2.996	17.23	.997	2.74	17.59	.997

5. Conclusion

From qualitative and quantitative analysis it is obvious that SRAD has excellent ability to remove noise from simulated phantom image and from real time ultrasound images but it performs poorly to preserve edges. The proposed Savitzky-Golay based SRAD filter performs well to preserve edges as well as remove noise with respect to conventional SRAD.

References

1. Wanger R.F, Smith S.W,Sandrik J.M, and Lopez M. H(1983), "Statistics in speckle in ultrasound B scans ," IEEE Trans. sonics Ultrason., vol.30,no. 3, pp.156-163.
2. D.T. kaun, A.A. Sawchuk, T.C. Strand and P. Chave P [1987]," Adaptive restoration of images with speckle," IEEE Trans. ASSP., vol. 35, no.3,pp.373-383.
3. T. Loupas, W.N Mcdicken and P.L Allan," An adaptive Weighted Median Filter for Speckle Suppression in Medical Ultrasonic Images," IEEE Tranaction on circuits and System, vol.36, no.1,pp.129-135.
4. Kuan D. T, Sawchuk A. A, Strand T. C, and P. Chave P (1987), "Adaptive restoration of images with speckle," IEEE Trans. ASSP., March, vol. 35,no. 3, pp. 373-383.
5. Lee J. S (1980), "Digital Image Enhancement and Noise Filtering by Use of Local Statistics," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 2, pp.165-168.
6. Lee J. S, (1981), "Refined filtering of image noise using local statistics," Computer Vision, Graphics, and Image Processing. Vol.15, pp.380-389.
7. Frost S. Victor, Josephine Abbott Stiles, Shanmugan K. S, and Julian C. Holtzman (1982), "A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise," IEEE Trans. on Pattern Analysis and Machine Intelligence, March, Vol. PAMI-4, No.2, pp.157-166.

8. Perona P and Malik J (1998), “Scale space and edge detection using anisotropic diffusion”. In Proc. IEEE Computer Society Workshop on Computer Vision, Miami Beach, FL, November, pp. 16-22.
9. Yu Y. and Acton S (2002), “Speckle reducing anisotropic diffusion”, IEEE Trans. Image Processing, vol.11, pp. 1260-1270.
10. A. Savitzky, M.J.E Golay [1964],” Smoothing and differentiation of data by simplified least square procedures,” Anal chem., vol. 36, pp 1627-1639.

JJIT, Vol. **6**, June 2017, 27-41

WORLDWIDE ELECTRONIC VOTING SYSTEM – AN ALGORITHM FOR E-VOTING DATABASE MANAGEMENT

M. MESBAHUDDIN SARKER, TANVIR AHMED SIDDIQUE, MOST. SHAHERA KHATUN, SYED MOHAMMAD RAKIB AND MD. RASHEDUZZAMAN RIAD

Institute of Information Technology, Jahangirnagar University, Bangladesh.

Abstract

Electronic voting is often seen as a tool for making the electoral process more efficient and for increasing trust in its management. The advantage of this system is it provides less cost, improve accessibility for voters with disabilities, faster results, greater accuracy and low risk of mechanical and human errors. There are many different electronic voting and counting technologies being used globally. The variety of offered technologies used makes it difficult to easily categorize them. In this paper, we have developed an algorithm for electronic voting system, which is applicable for knowing the status of a country's electoral system.

Keywords: Database, DRE, Electoral, VVPAT.

1. Introduction

Electronic voting means to either aid or take care of the chores of casting and counting votes.
This technology can include punched cards, optical scan voting systems and specialized voting kiosks (including self-contained direct-recording electronic voting systems, or DRE). It can also involve transmission of ballots and votes via telephones, private computer networks, or the Internet.In general, two main types of e-Voting can be identified [Buchsbaum, 2004]:

i. E-voting which is physically supervised by representatives of governmental or independent electoral authorities (e.g. electronic voting machines located at polling stations).

ii. Remote e-voting via the internet (also called i-voting) where the voter votes at home or without going to a polling station [Zissis, 2011].

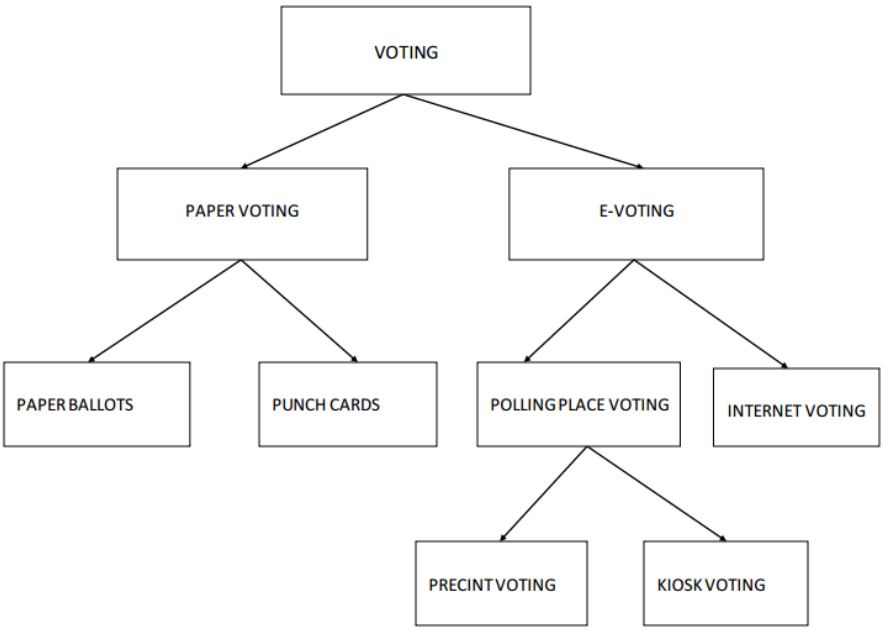


Fig. 1: Different Types of Voting Systems

2. Methodology

We have listed 252 countries from the different sources such as website, journals, online text and blogand categorize them according to their types of electoral system such as Internet voting, EVM, polling place votingetc. Then we developed an algorithm and coded the algorithm with C-language. Necessary figure, diagram and tables are presented and finally implemented and simulated.

3. Related Works

Voting technologies have a surprisingly long history. In the United States, mechanical lever voting machines were first used for elections in 1892 and were commonly used in U.S. elections until the 1990s [Kimball, 2004]. Electronic technologies began to appear in the 1960s with punch card counting machines. After then, DRE voting machines, ballot scanning machines and Internet voting began to appear. In 1990s and the first decade of the new millennium, an increasing number of countries around the world also started to adopt these technologies. Currently, there are four major types of e-voting around the world that are worth keeping an eye on: Brazil's homegrown direct recording electronic (DRE) setup, Australia's open-source software, Estonia's Internet voting, and a Spanish startup's efforts to expand what's been called "crypto-voting" [Drew et. al.]. Each of these approaches has its own unique set of problems, but the primary obstacles they present for many voting officials and computer scientists is their lack of ability to verify source code and expense. Also much insecurity has been found in commercial voting machines, such as using a

default administration password [Schneier, Feldman, 2015].Cases have also been reported of machines making unpredictable, inconsistent errors. Key issues with electronic voting are therefore the openness of a system to public examination from outside experts, the creation of an authenticatable paper record of votes cast and a chain of custody for records [Kobie, TC-J, 2015]. However, there has been contention, especially in the United States, that electronic voting, especially DRE voting, could facilitate electoral fraud and may not be fully auditable. In addition, electronic voting has been criticized as unnecessary and expensive to introduce. Several countries have cancelled e-voting systems or decided against a large-scale rollout, notably the Netherlands and the United Kingdom [Hern, 2015].

In India EVMs manufactured in 1989-90 were used on experimental basis for the first time in 16 Assembly Constituencies in the States of Madhya Pradesh, Rajasthan and NCT of Delhi at the General Elections to the respective Legislative Assemblies held in November, 1998 [ECI].Where in Bangladesh, election commission produced its own electronic voting machine, piloted in several stages, first in Chittagong city corporation election and then in Narayanganj city corporation election, which proved to be very successful. However, electronic voting in Bangladesh is yet to be implemented nationally, in spite of successful pilots [Meftaul, 2011].

4. Worldwide Voting Process: Some Statistics

Recent research has shown that 31 countries around the world have used non-remote electronic voting machines for binding political elections at some point [Esteve et. al., 2012]. Some of these countries have experimented with EVMs and then decided not to continue with their use, in some cases after using them for many years. EVMs are being used in 20 countries, with six of these countries still piloting the technology. Globally, every different trends are seen in different regions. Europe and North America can be seen as moving away from the use of EVMs, while South America and Asia show increasing interest in using electronic voting technologies.

Statistics on the global use of electronic voting are as follows:

- 31 countries around the world use or have used EVMs for binding political elections
- 4 use e-voting nationwide – India, Brazil, Bhutan and Venezuela
- 11 use in some parts of country
- 5 have pilots ongoing
- 8 piloted and did no continue
- 3 used for a number of elections and then discontinued – Germany, Netherlands and Paraguay

Table-1: No. of countries with electronic voting system

No.	Types of E-voting	No. Countries
1	EVM	31
2	Nationwide	4
3	Partially	11
4	Pilots	5
5	Piloted but not continued	8
6	Used, now discontinued	3
	Total:	62

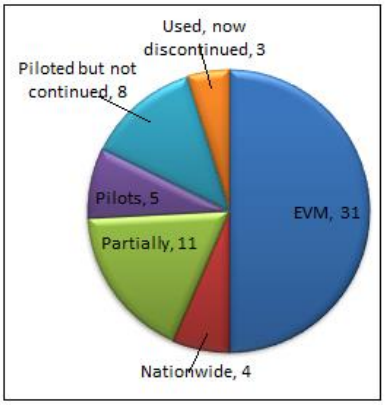


Fig. 2: No. of countries with electronic voting system

Table-2: No. of countries and percentage of electronic voting system

	Types of Voting	No. of Countries	Percentage
1	Manually marking of ballots.	209	81.32%
2	Mechanical voting machine	2	0.78%
3	Punch card	1	0.39%
4	Electronic voting machine	18	7.4%
5	Internet	5	1.95%
6	Other	9	3.50%
7	No information available	9	3.50%
8	Not applicable	3	1.16%

Source: http://aceproject.org/epic-en/CDTable?question=VO011

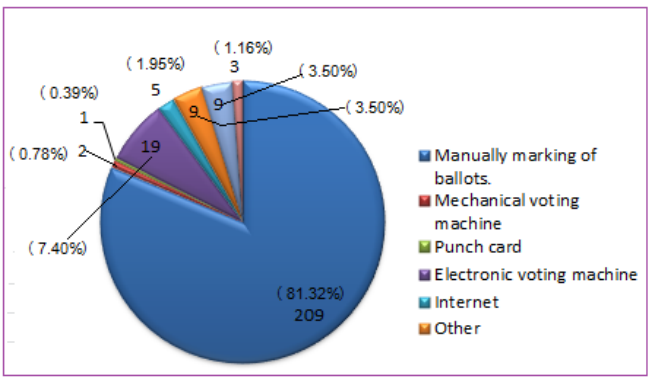


Fig. 3: No. of countries and percentage of electronic voting system

The following is a list of examples of electronic voting from elections around the world. Examples include polling place voting electronic voting and Internet voting, these are: Brazil, Venezuela, Paraguay, Belgium, USA, Spain, Canada, Switzerland, Bhutan, Germany, India, The Netherlands, Panama. However, listed the following countries where election commissions have implemented electronic voting as a first user.

Table-3: Remote and polling place electronic voting used by each country

Country	Remote e-Voting	Polling Place e-Voting
Australia		X
Austria	X	
Bangladesh		X
Belgium		X
Brazil		X
Canada	X	X
Denmark		
Estonia	X	
France	X	X
Germany		X
India		X
Ireland		X
Netherlands	X	X
Norway		X
Portugal	X	X
Spain	X	X
Switzerland	X	X
UK	X	X
USA	X	X

Source: http://www.tiresias.org/research/guidelines/evoting_projects.htm

Table-4: World-widefirst electronic voting user

Country	Type of Voting	Year	Purpose/Event
Australia	CyberVote; Electronic Voting;	1995 2001 2006	 Australian Parliamentary election Victorian State election
Bangladesh	Electronic Voting Machine (EVM)	2010-2012	Different City Corporation and municipality elections
Belgium	Electronic Voting;	1991 and 999	General and Municipal election
Brazil	Electronic Voting By EVM;	1996 2000,2002	Santa Catarina State Nationwide

Country	**Type of Voting**	**Year**	**Purpose/Event**
Canada	Electronic Voting; Optical scan and Internet voting;	1990 At present	Municipal level Nationwide
Estonia	Internet voting;	2005 2007 2009 2011	Nationally local election; World first national election; Local municipal election; Parliamentary Election;
EU	CyberVote by using fixed and mobile internet terminals;	2000	Trials were performed in Sweden, France, and Germany
Finland	Internet-enabled DER	2008	Three municipalities (Karkkila, Kauniainen and Vihti) election
France	Remote Internet voting; Remote e-Voting and touch screen e-Voting over the Internet;	2003 2007	French citizens living in the USA, National presidential primary,
Germany	EVM (from Nedap); Optical scan voting based on digital paper;	2005 2006 2008	Bundestag election Cottbus municipality election Hamburg state election
India	EVM	1982 2003 2004, 2009	Kerala state election All state elections Parliamentary election
Ireland	EVM (from Nedap);	2002	General election for 3 constituencies (pilot basis)
Italy	EVM (from Nedap);	2006	Cremona municipality election
Kazakhstan	Electronic Voting; Indirect recording electronic voting by using Smart card;	2004 2005 2007	Parliamentary elections; Presidential election; Parliamentary elections;
Netherlands	EVM (from Nedap);	2006	Parliamentary election
Norway	EVM	2003	3 municipalities election
Philippines	Optical Scan Voting		Nationwide
Romania	Electronic Voting	2003	Limited basis
Switzerland	Internet voting	2009	Swiss citizens living abroad
United Kingdom	Optical Scan Voting Optical Character Recognition	2000 2004	London Mayoral and Assembly elections; London Mayoral, Assembly and European Parliamentary elections;
USA	EVM	1960	For experimental basis

Country	Type of Voting	Year	Purpose/Event
	Optical Scan Voting	1964	Presidential election in 7 counties)
Scotland	Optical Scan Voting	2007	Scottish Parliament and Scottish council elections.

Source: :Wikipedia (http://en.wikipedia.org/wiki/Electronic_voting_examples)

5. Proposed Algorithm and Flowchart

1. Start
2. while (True) {
3. Take a string input t ;
4. If (t == “First 3 letters of any country in Database”)
5. { fopen(“Filename.txt, “r”) //open the file at read node//
6. fgets (array name, size, associated file pointer)
7. print the information with puts function
8. print : “Enter next choice”
9. else if (t==”End”)
   break ;
10. else
    Print: “Sorry we don’t have enough data about this country..”}
11. Print: “Enter your next choice.”

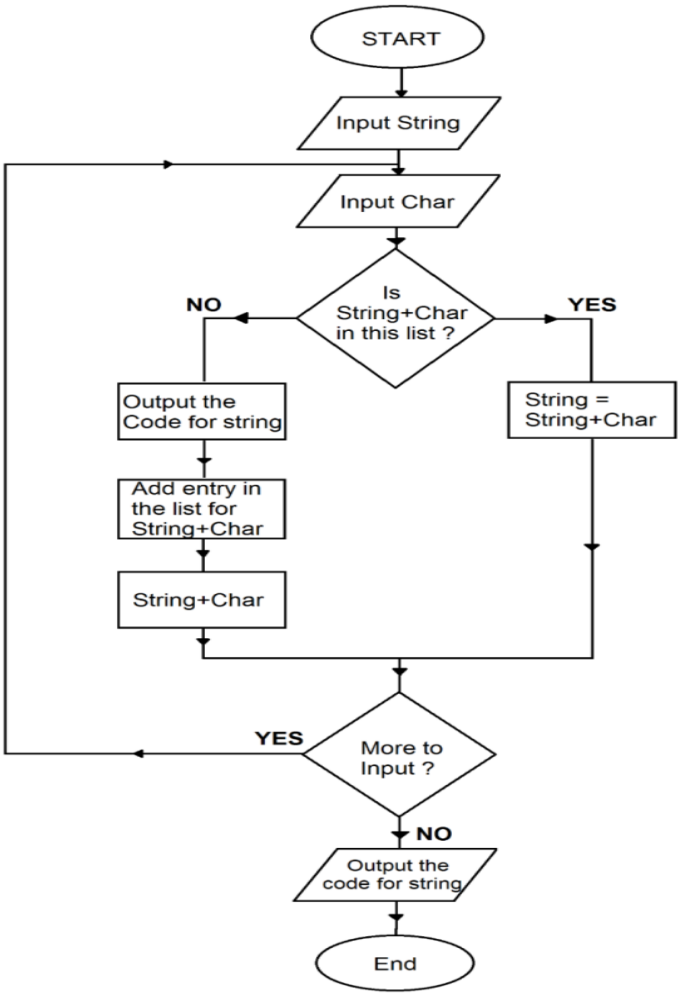


Fig. 3: Flowchart of proposed electronic voting system

6. Country list with different electoral process

1. EVM: Argentina, Belgium, Brazil, Bhutan, Ecuador, France, Guam, India, Mongolia, New Caledonia, Peru, Philippines, Portugal, Paraguay, Singapore, United States of America, Venezuela, Wallis and Futuna
2. INT: Armenia, Australia, Canada, Switzerland, Estonia
3. MMB: Andorra, United Arab Emirates, Afghanistan, Antigua and Barbuda, Anguilla, Albania, Armenia, Angola, Argentina, American Samoa, Austria, Australia, Aruba, Azerbaijan, Bosnia and Herzegovina, Barbados, Bangladesh, Belgium, Burkina Faso, Bulgaria, Bahrain, Burundi, Benin, Bermuda, Bolivia, Bahamas, Bhutan, Botswana,

Belarus, Belize, Canada, Congo (Kinshasa), Democratic Republic of the, Central African Republic, Congo (Brazzaville), Switzerland, Cote de Ivoire, Cook Islands, Chile, Cameroon, China, Colombia, Costa Rica,Cuba, Cape Verde, Cyprus, Czech Republic, Germany, Djibouti, Denmark, Dominica, Dominican Republic, Algeria, Ecuador, Estonia, Egypt, Eritrea, Spain, Ethiopia, Finland, Fiji, Falkland Islands (Malvinas), Micronesia, Federated States of, France, Gabon, United Kingdom of Great Britain and Northern Ireland, Grenada, Georgia, French Guiana, Ghana, Gibraltar,Greenland, Guinea, Guadeloupe, Greece, Guatemala, Guam, Guinea-Bissau,Guyana, Hong, Kong, Honduras, Croatia, Haiti, Hungary, Indonesia, Ireland, Israel, India, Iraq, Iran, Iceland, Italy, Jamaica, Jordan, Japan, Kenya,Kyrgyzstan, Cambodia, Kiribati, Comoros, Saint Kitts, and Nevis, Korea, Democratic People's Republic of, Korea, Republic of, Kosovo, Kuwait, Cayman Islands, Kazakhstan, Lao People's Democratic Republic, Lebanon, Saint Lucia,Liechtenstein, Sri Lanka, Liberia, Lesotho, Lithuania, Luxembourg, Latvia, Libya, Morocco, Monaco, Moldova, Republic of Montenegro, Madagascar,Marshall Islands, Macedonia, The Former Yugoslav Republic of, Mali, Burma, (Myanmar), Mongolia, Martinique, Mauritania, Malta, Mauritius, Maldives,Malawi, Mexico, Malaysia, Mozambique, Namibia, Niger, Nigeria, Nicaragua, Netherlands, Norway, Nepal, Nauru, Niue, New Zealand, Oman, Panama,Peru, Papua New Guinea, Philippines, Pakistan, Poland, Puerto Rico,Palestine, Portugal, Palau, Paraguay, Reunion, Romania, Serbia, Russia,Rwanda, Solomon Islands, Seychelles, Sudan, Sweden, Singapore, Saint Helena, Slovenia, Slovakia, Sierra Leone, San Marino, Senegal, Suriname,South Sudan, Sao Tome and Principe, El Salvador, Syria, Swaziland, Turks and Caicos Islands, Chad, Togo, Thailand, Tajikistan, Timor-Leste,Turkmenistan, Tunisia, Tonga, Turkey, Trinidad and Tobago, Tuvalu, Taiwan,Tanzania, Ukraine, Uganda, United States of America, Uruguay, Uzbekistan,Saint Vincent and the Grenadines, Viet Nam, Samoa, Yemen, South Africa,Zambia, Zimbabwe, Zanzibar.
4. MVM: Bulgaria, United States of America
5. NOINFO: Cyprus (North), Western Sahara, Equatorial Guinea, Comoros, Somalia, Holy See (Vatican City State), Virgin Islands, British, U.S.
6. NOVOTE: Saudi Arabia, Brunei Darussalam, Qatar
7. OTH: Switzerland, Gambia, New Caledonia, French Polynesia, Swaziland, Taiwan, Vanuatu, Wallis and Futuna
8. PC: United States of America

Source: http://aceproject.org/epic-en/CDTable?question=VO011

7. Simulation

Enter your choice in short: (EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE)
(UPPER CASE LETTERS ONLY!)
To stop this program type END

Type :EVM

FOLLOWING 19 COUNTRIES HAVE ELECTRONIC VOTING SYSTEM

Argentina, Belgium, Brazil, Bhutan, Ecuador, France, Guam, India, Mongolia, New Caledonia Peru, Philippines, Portugal, Paraguay, Singapore, United States of America, Venezuela, Wallis, Futuna

Enter your choice in short: (EVM, INT, MVM, MMB, PC, OTH, NOINFO, NOVOTE)
Type: INT

FOLLOWING 5 COUNTRIES HAVE INTERNET VOTING SYSTEM
Armenia, Australia, Canada, Switzerland, Estonia

Enter your choice in short: (EVM, INT, MVM, MMB, PC, OTH, NOINFO, NOVOTE)
To stop the process, Type : END

8. Conclusion and Future Work

Electronic voting is not in its testing phase as several countries have used the technology for more than 10-15 years. However, every country's electoral requirements are different, and systems have to be tailored to these differences.Moreover, pilots are being conducted in Mexico and Peru, and specific tests to electronic voting systems' functionality have been conducted in Argentina and Ecuador, whereas some countries have more than one system.The Voter Verified Paper Audit Trail (VVPAT) is becoming the standard for reliability in electronic voting, especially for polling station based electronic voting. It is noted that VVPAT is being used in nearly 800,000 polling stations across India, with relative ease. This machine minimizes invalid votes and is accessible to illiterate voters through buttons with symbols rather than stamping/marking. However, our algorithm can be connected with web/internet so that any information of a country (i.e., country name, type of voting system, percentage etc.) can be known automatically. For this, in future, may need to develop the algorithm with web based application.

References

1. Buchsbaum, T. (2004). "E-voting: International developments and lessons learnt". Proceedings of Electronic Voting in Europe Technology, Law, Politics and Society. Lecture Notes in Informatics. Workshop of the ESF TED Programme together with GI and OCG.

2. Drew Springall, Travis Finkenauer, Zakir Durumeric, Jason Kitcat, Harri Hursti, Margaret MacAlpine, J. Alex Halderman, "Security Analysis of the Estonian Internet Voting System", Open Rights Group, U.K (estoniaevoting.org.).

3. Esteve, JordiBarrat I, Ben Goldsmith and John Turner. International Experience with E-Voting. Norwegian E-Vote Project. IFES, June 2012.

4. : Election Commission of India : http://eci.nic.in/eci_main1/evm.aspx

5. lman, Halterman&Felten (2015). "Security Analysis of the Diebold AccuVote-TS Voting :hine". Usenix. Retrieved 3 December 2015.

6. n, Alex (2015): "Should Britain introduce electronic voting?". The Guardian. Retrieved 3 ember 2015.

7. Kimball W. Brace (2004): "Overview of Voting Equipment Usage in United States, Direct Recording Electronic (DRE) Voting", Election Data Services Inc. to the United States Election Assistance Commission May 5, 2004.

8. Kobie, Nicole (2015): "Why electronic voting isn't secure". The Guardian. Retrieved 3 December 2015.

9. Meftaul Islam (2011): EVM and Digital Bangladesh,Dept. of International Relations, Jahangirnagar University, http://www.thedailystar.net/news-detail-215643.

10. Schneier, Bruce (2015): "What's Wrong With Electronic Voting Machines?" Schneier on Security. Retrieved 3 December 2015.

11 TC-J (2015):"Wichita State mathematician says Kansas voting machines need to be audited to check accuracy". Topeka Capital-Journal (TC-J). Retrieved 3 December 2015.

12 Zissis, D., Lekkas (April 2011): "Securing e-Government and e-Voting with an open cloud computing architecture". Government Information Quarterly. **28** (2): 239–251.

Other Sources

1. "i-Voting". e-Estonia.

2. Philippines concerning grid power requirements for various needs including i-voting

3. "Switzerland's new legislation on internet voting". electoralpractice.ch.

4. http://aceproject.org/epic-en/CDTable?question=VO011

5. Wikipedia (http://en.wikipedia.org/wiki/Electronic_voting_examples)

6. http://www.tiresias.org/research/guidelines/evoting_projects.htm

https://arstechnica.co.uk/features/2016/11/internet-based-and-open-source-how-e-voting-is-working-around-the-globe/

Appendix

Program Documentation

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <strings.h>
#include<iostream>
using namespace std;

int main()
{
printf("#####################  VOTING PROCESS WORLDWIDE
######################\n");
      FILE *in1,*in2,*in3,*in4,*in5,*in6,*in7,*in8,*in9;
      char
evm[99999],in[99999],mvm[99999],mmb[99999],pc[99999],oth[99999],noinfo[99999],no
vote[99999],t[9];
printf("                    ELECTRONIC VOTING MACHINE(EVM)\n");
printf("                       INTERNET VOTING(INT)\n");
printf("                     MECHANICAL VOTING MACHINE(MVM)\n");
printf("                    MANUALLY MAKING OF BALLOTS(MMB)\n");
printf("                        PUNCH CARD(PC)\n");
printf("                         OTHERS(OTH)\n");
printf("                      NO INFORMATION! (NOINFO)\n");
printf("                      NOT APPLICABLE! (NOVOTE)\n");
printf("\n============================\n=======================\n");
printf("Enter your choice in short:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE)\n");
printf("(UPPER CASE LETTERS ONLY!)\nTo stop this program type END \n");
while(1)
   {
gets(t);
if(strcmp(t,"EVM")==0)
      {
printf("\nFOLLOWING 19 COUNTRIES HAVE ELECTRONIC VOTING
SYSTEM\n\n");
         in1 = fopen("EVM.txt","r");
while(!feof(in1))
         {
fgets(evm,99999,in1);
            //puts(evm);
cout<<evm;
         }
```

```
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"INT")==0)
      {
printf("\nFOLLOWING 5 COUNTRIES HAVE INTERNET VOTING SYSTEM\n");
         in2 = fopen("INT.txt","r");
while(!feof(in2))
         {
fgets(in,99999,in2);
            //puts(in);
cout<<in;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"MVM")==0)
      {
printf("\nFOLLOWING 2 COUNTRIES HAVE MECHANICAL VOTING
MACHINE\n\n");
         in3 = fopen("MVM.txt","r");
while(!feof(in3))
        {
fgets(mvm,99999,in3);
            //puts(mvm);
cout<<mvm;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"MMB")==0)
      {
printf("\nFOLLOWING 209 COUNTRIES HAVE MANUALLY MAKING OF
BALLOTS\n\n");
         in4 = fopen("MMB.txt","r");
while(!feof(in4))
         {
fgets(mmb,99999,in4);
            //puts(mmb);
cout<<mmb;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"PC")==0)
```

```
      {
printf("\nFOLLOWING 1 COUNTRY HAS PUNCH CARD VOTING SYSTEM\n\n");
         in9 = fopen("PC.txt","r");
while(!feof(in9))
         {
fgets(pc,99999,in9);
            //puts(mmb);
cout<<pc;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"OTH")==0)
      {
printf("\nFOLLOWING 9 COUNTRIES USE OTHER PROCESS\n\n");
         in5 = fopen("OTH.txt","r");
while(!feof(in5))
         {
fgets(oth,99999,in5);
           // puts(oth);
cout<<oth;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"NOINFO")==0)
      {
printf("\nWE HAVE NO INFORMATIONS OF FOLLOWING 10 COUNTRIES\n\n");
         in6 = fopen("NOINFO.txt","r");
while(!feof(in6))
         {
fgets(noinfo,99999,in6);
            //puts(noinfo);
cout<<noinfo;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"NOVOTE")==0)
      {
printf("\nFOLLOWING 3 COUNTRIES HAVE NO VOTING SYSTEM\n\n");
         in7 = fopen("NOVOTE.txt","r");
while(!feof(in7))
         {
fgets(novote,999,in7);
```

```
            //puts(novote);
cout<<novote;
         }
printf("\n\nENTER YOUR NEXT CHOICE:
(EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE) \n");
      }
else if(strcmp(t,"END")==0) break;
else
      {
printf("\nSORRY!THE KEYWORD IS NOT CORRECT!\n");
printf("PLEASE TRY AGAIN: (EVM,INT,MVM,MMB,PC,OTH,NOINFO,NOVOTE)
\n");
      }
   }
return 0;
}
```

JJIT, Vol. **6**, June, 2017, 43-55

NUMERICAL SIMULATION OF MHD FREE CONVECTIVE FLOW PAST A VERTICAL CONE IN PRESENCE OF HEAT GENERATION

SREEBASH C. PAUL AND MOUSUMI MUKHERJEE

Department of Arts and Sciences, Ahsanullah University of Science and Technology (AUST), Dhaka 1208, Bangladesh

Abstract

Magnetohydrodynamic (MHD) free convective flow from a vertical circular cone maintained at variable surface heat flux in presence of heat generation is considered. Non-linear partial differential equations which govern the boundary layer of the flow are reduced to non-similar boundary layer equations and solved numerically by using finite difference method with Keller-Box scheme.The solutions are presented in terms of local skin friction, local Nusselt number, velocity and temperature profile for different values of magnetic parameter and heat generation parameter.It is found that the value of both the local skin friction and the local Nusselt number decreases as the magnetic parameter increases but for increasing values of heat generation parameter the value of local skin friction increases whereas the value of local Nusselt number decreases.

Key words: Free convection, Heat transfer, Magnetohydrodynamic (MHD), Heat generation.

1. Introduction

Many free convection process occur in environments such as closed containers and environmental chambers with heated walls. Free convection associated witha vertical circular cone with uniform or non uniform surface heat flux is of interest of many researchers. Hering and Grosh [1] obtained similarity solutions for free convection from the vertical. They showed that the similarity solutions to the boundary layer equations for a cone exist when the wall temperature is a power function of distance along a cone ray. Latter, Hering [2] extended this analysis to investigate for low Prandlt number. On the other hand the problem of Hering and Grosh [1] has been extended by Roy [3] for the case of high Prandlt number. Na and Chiou [4] investigated the effect of slenderness over a slender cone with constant wall heat flux for the natural convection flow. An integral method is applied by Alamgir [5] to study the overall heat transfer from vertical cones in laminar natural convection.The investigation of the natural convection flow from a heated vertical permeable circular cone is performed byHossain and Paul [6, 7]. They show the effect of transpiration parameter, surface heat flux gradient and the Prandtl number on laminar free convection flow from a vertical cone.

The free convection flows of fluid in presence of magnetic field and heat generation have received much attention by many researchers of fluid mechanics. The study of flows of the

fluid which is electrically conducting and moving in a magnetic fieldis known as Magnetohydrodynamic(MHD). The temperature distribution may be altered by the possible heat generation effects. This may occur in such applications related to nuclear reactor cores, fire and combustion modeling.The flowing fluid carry the interaction of the magnetic field and the moving electric charge and the moving fluid induced a force. The influence of magnetic field on the boundary layer is exerted only through induced forces within the boundary layer itself, with no additional effects arising from the free stream pressure gradient. Kuiken[8] studied the problem of MHD free convection in a strong cross field. MHD free convection flow and mass transfer through a porous medium had been investigated by Rapits and Kafoussias [9]. Sparrow and Cess [10] studied the effect of magnetic field on free convection heat transfer. However, Hossian*etal*[11-12] discussed both the forced and free convection boundary layer flow of an electrically conducting fluid in presence of magnetic field.

The heat transfer characteristics in the laminar boundary layer of viscous fluid over a stretching sheet with viscous dissipation and internal heat generationhave been investigated by Vajravelu and Hadjinoiaou [13]. In this study, the volumetric rate of heat generation, q''' [W.m^2], was considered as

$$q''' = \begin{cases} Q_0(T - T_\infty) \text{for} T \geq T_\infty \\ 0 \qquad\qquad\quad \text{for } T < T_\infty \end{cases}$$

whereQ_0 is the heat generation constant. They also reported that the above relationhaving T_∞ as the onset temperature is valid as an expression of the state of some exothermic process.Following,Vajravelu and Hadjinoiaou [13], effects of heat generation or absorption on hydrodynamic flow with heat and mass transfer over a flat plate had been investigated by Chamkha and Camille [14]. Also the effect of the conjugate conduction-natural convection heat transfer along a thin vertical plate with non-uniform heat generation have studied by Mendez and Trevino [15]. Latter,Molla*et al* [16] has investigated the MHD natural convection flow on a sphere in presence of heat generation.The natural convection flow from a horizontal circular cylinder with uniform heat flux in presence of heat generation has been studied by Molla*et al* [17].

The effect of heat generation on free convection flow from a vertical circular cone with variable surface heat flux in presence of MHD has not been studied yet. In the present study, it is proposed to investigate the MHD free convection flow from a vertical circular cone with non uniform heat flux in presence of heat generation. At first the governing equations of flow are transformed into the local non-similarity boundary layer equations. The transformed boundary layer equations are solved numerically using implicit finite difference method along with the Keller-Box method. After solving these equations, the obtaining numerical solutions are represented in terms of velocity profile, temperature profile, local skin friction, local Nusselt number for different values of magnetic parameter and heat generation parameter.Comparison of numerical results of present work with other published data has been shown in table.

2. Governing Equations

A steady two-dimensional free convective MHD laminar flow of a viscous incompressible fluid having temperature T, measured from a vertical circular cone with non uniform surface heat flux. A magnetic field of strength β_0 acts normal to the circular cone.The physical coordinates (*x*, *y*) are chosen such that *x* is measured from the leading edge in the stream wise direction and *y* is measured normal to the surface of the cone. The coordinate system and flow configuration are shown in fig.1.

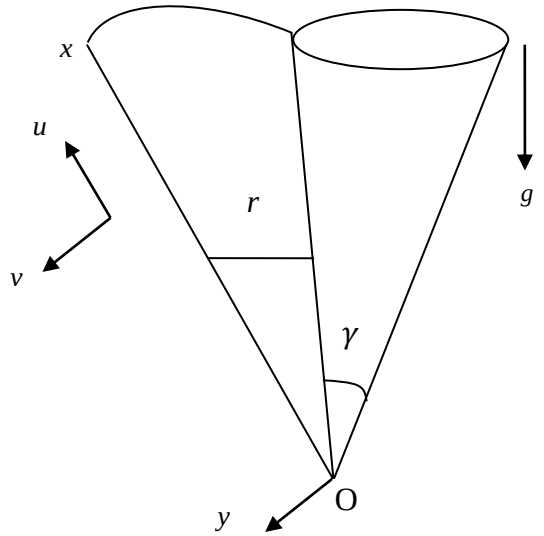


Figure 1: Physical model and co-ordinates

Following the boundary layer approximations the flow is governed by the following equations:

$$\frac{\partial(ur)}{\partial x} + \frac{\partial\ (\vartheta r)}{\partial y} = 0 \tag{1}$$

$$u\frac{\partial u}{\partial x} + \vartheta\frac{\partial u}{\partial y} = \nu\frac{\partial^2 u}{\partial y^2} + g\beta\cos\gamma(T - T_\infty) - \frac{\alpha_0\beta_0^2}{\rho}u \tag{2}$$

$$u\frac{\partial T}{\partial x} + \vartheta\frac{\partial T}{\partial y} = \alpha\frac{\partial^2 T}{\partial y^2} + \frac{Q_0}{\rho c_p}(T - T_\infty) \tag{3}$$

With the boundary conditions

$$u = 0, \vartheta = -V, q = -k\left(\frac{\partial T}{\partial y}\right) aty = 0$$
$$u = 0, T = T_\infty asy \to \infty \tag{4}$$

where u, ϑ are the fluid velocity components in the x and y directions, respectively, ν the kinematic viscosity, g is the acceleration due to gravity, β is the coefficient of volume expansion, α is the thermal diffusivity, γ is the cone apex half-angle, ρ is the density,α_0 is the electrical conduction,β_0 is the strength of the magnetic field,c_p is the specific heat at constant pressure, T is the temperature of the fluid and T_∞ is the ambient fluid temperature.The amount of heat generated or absorbed per unit volume is $Q_0(T - T_\infty)$, Q_0 being a constant, which may take either positive or negative. The source term represents the heat generation when $Q_0 > 0$ and the heat absorption when $Q_0 < 0$. V represents the transpiration velocity of the fluid through the surface of the cone which is positive for suction and negative for injection of fluidthrough the surface of the cone. In the present paper, we shall consider only suction case and V is taken as positive throughout.

To make the above equations dimensionless, the following transformations are introduced,

$$\psi = \nu r Gr_x^{\frac{1}{5}}\left[f(\xi,\eta) + \frac{1}{2}\xi\right], T - T_\infty = \frac{qx}{\kappa} Gr_x^{-\frac{1}{5}}\theta(\xi,\eta), \eta = \frac{y}{x} Gr_x^{\frac{1}{5}},$$
$$\xi = \frac{Vx}{\nu} Gr_x^{-\frac{1}{5}}, Gr_x = \frac{g\beta q x^4 \cos\gamma}{\kappa\nu^2}, q_w \approx x^m \tag{5}$$

whereGr_x is the local Grashof number, ξ is the dimensionless suction parameter, η is the pseudo-similarity variable and ψ is the stream function defined by

$$u = \frac{1}{r}\frac{\partial\psi}{\partial y} \text{and} \vartheta = -\frac{1}{r}\frac{\partial\psi}{\partial x} \tag{6}$$

Substituting the transformations (5)along with the equation (6) into the equations (1) to (4), we obtained the non-similarity system of equations as follows:

$$f''' + \frac{9}{5}ff'' - \frac{3}{5}f'^2 + \theta - Mf' = \frac{1}{5}\xi\left(f'\frac{\partial f'}{\partial\xi} - f''\frac{\partial F}{\partial\xi}\right) \tag{7}$$

$$\frac{1}{Pr}\theta'' + \frac{9}{5}f\theta' - \frac{1}{5}\theta f' + Q\theta = \frac{1}{5}\xi\left(f'\frac{\partial\theta}{\partial\xi} - \theta'\frac{\partial f}{\partial\xi}\right) \tag{8}$$

Corresponding boundary conditions transforms to

$$f = f' = 0, \quad \theta' = -1 \text{ at } \eta = 0$$
$$f' = 0, \quad \theta = 0 \; as\eta \to \infty \tag{9}$$

where $f(\xi,\eta)$ and $\varphi(\xi,\eta)$ are respectively the dimensionless stream function and temperature function of the fluid in the boundary layer region. $M = \frac{\alpha_0\beta_0^2x^2}{\rho v}Gr_x^{-\frac{2}{5}}$ and $Q = \frac{Q_0x^2}{\nu\rho c_p}Gr_x^{-\frac{2}{5}}$ are magnetic field and heat generation parameter respectively.

Now by the definition of skin friction, C_f and Nusselt number, we know that

$$C_f = \frac{\tau_w}{\rho U^2}$$

where the local shear stress,

$$\tau_w = \mu\left(\frac{\partial u}{\partial y}\right)_{y=0} \qquad (10)$$

and $Nu_x = \frac{q_w x}{\kappa(T_w - T_\infty)}$, where the surface heat flux, $q_w = -k\left(\frac{\partial T}{\partial y}\right)_{y=0}$ (11)

Using the transformation (5) into the above expression (10) and (11), we can calculate the values of the local skin-friction coefficient and Nusselt number from the following relations:

$$\frac{1}{2}C_f Gr_x^{\frac{1}{5}} = f''(\xi,0) \qquad (12)$$

$$\frac{Nu_x}{Gr_x^{\frac{1}{5}}} = \frac{1}{\theta(\xi,0)} \qquad (13)$$

3. Numerical Methods

In the present study, we employ an efficient solution technique, known as implicit finite difference method together with Keller-boxelimination technique introduced by Keller [18].In this method, equations (7) to (9) are converted into the following system of first order equations with dependent variables $u(\xi,\eta)$, $v(\xi,\eta)$ and $p(\xi,\eta)$ as

$$\frac{\partial f}{\partial \eta} = u,\ \frac{\partial u}{\partial \eta} = v,\frac{\partial \theta}{\partial \eta} = p \qquad (14)$$

$$v' + p_1 f v - p_2 u^2 + p_3\theta - p_4 u + p_5 v = p_0\xi\left(u\frac{\partial u'}{\partial \xi} - v\frac{\partial f}{\partial \xi}\right) \qquad (15)$$

$$\frac{1}{Pr}p' + p_1 f p - p_7 u\theta + p_5 p + p_6\theta = p_0\xi\left(u\frac{\partial \theta}{\partial \xi} - p\frac{\partial f}{\partial \xi}\right) \qquad (16)$$

$$f = 0, f' = 0, \theta' = -1\ at\eta = 0 \text{ and} f' = 0, \theta = 0\ at\eta \to \infty \qquad (17)$$

Where

$$p_1 = \frac{m+9}{5}, p_2 = \frac{2m+3}{5}, p_0 = \frac{1-m}{5},\ p_5 = \xi, p_4 = M, p_6 = Q, p_3 = 1, p_7 = \frac{4m+1}{5} \qquad (18)$$

We now consider the net rectangle on the (ξ , η) plane and denote the net points by

$$\xi^0 = 0, \xi^n = \xi^{n-1} + \kappa_i;\ n = 1,2,3, \ldots\ldots\ldots\ldots\ldots\ldots, \mathrm{N} \tag{19}$$

$$\eta_0 = 0, \eta_j = \eta_{j-1} + h_j;\ \ \mathrm{j} = 1,2,3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathrm{J};\ \eta_j = \eta_\infty \tag{20}$$

Here *n* and *j* are just sequence of numbers on the (ξ , η) plane, $\kappa_n$and $h_j$be the variable mesh widths.

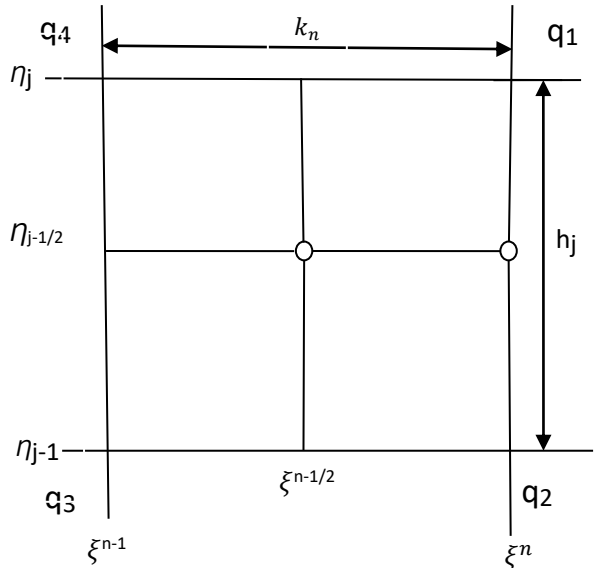


Figure2: Net rectangle of the difference approximation

We approximate the quantities (f , u ,v , θ , p) at points (ξ^n, η_j) of the net by ($f_j^n, u_j^n, v_j^n, p_j^n, \theta_j^n$), which we call net function.It is also employed the notation g_j^n for any net function quantities midway between net points shown in Figure 2 as follows

$$\begin{aligned} &\xi^{n-\frac{1}{2}} = \frac{1}{2}(\xi^n + \xi^{n-1}),\ \ \eta_{j-\frac{1}{2}} = \frac{1}{2}(\eta_j + \eta_{j-1}), \\ &g_j^{n-\frac{1}{2}} = \frac{1}{2}\left(g_j^n + g_j^{n-1}\right),\ g_{j-\frac{1}{2}}^n = \frac{1}{2}\left(g_j^n + g_{j-1}^n\right) \end{aligned} \tag{21}$$

Now we write the difference equations that are to approximate equations (14) to (17) by considering one mesh rectangle. We start by writing the finite difference approximation of the equations (14) using backward difference quotients and average about the midpoint ($\xi^n, \eta_{j-\frac{1}{2}}$) to obtained,

$$\frac{f_j^n - f_{j-1}^n}{h_j} = u_{j-\frac{1}{2}}^n, \frac{u_j^n - u_{j-1}^n}{h_j} = v_{j-\frac{1}{2}}^n, \frac{\phi_j^n - \phi_{j-1}^n}{h_j} = p_{j-\frac{1}{2}}^n \tag{22}$$

Similarly, equations(15)–(17) can be expressed in finite difference form, by approximating thefunctions and their derivatives bycentral differences about the midpoints, $(\xi^{n-\frac{1}{2}}, \eta_{j-\frac{1}{2}})$giving the following non-linear difference equations:

$$\frac{v_j^n - v_{j-1}^n}{h_j} + (p_1^n + \alpha_n)(fv)^n_{j-\frac{1}{2}} - (p_2^n + \alpha_n)(u^2)^n_{j-\frac{1}{2}} - p_4^n\left(u^n_{j-\frac{1}{2}}\right) + \alpha_n\left(v^{n-1}_{j-\frac{1}{2}} f^n_{j-\frac{1}{2}} - v^n_{j-\frac{1}{2}} f^{n-1}_{j-\frac{1}{2}}\right) + p_3^n\left(\theta^n_{j-\frac{1}{2}}\right) + p_5^n\left(\mathrm{v}^n_{j-\frac{1}{2}}\right) = R^{n-1}_{j-\frac{1}{2}} \tag{23}$$

$$\frac{1}{Pr}\frac{p_j^n - p_{j-1}^n}{h_j} + (p_1^n + \alpha_n)(fp)^n_{j-\frac{1}{2}} - (p_7^n + \alpha_n)(u\theta)^n_{j-\frac{1}{2}} + p_5^n p^n_{j-\frac{1}{2}} + p_6^n \theta^n_{j-\frac{1}{2}} - \alpha_n\left(u^{n-1}_{j-\frac{1}{2}}\theta^n_{j-\frac{1}{2}} - u^n_{j-\frac{1}{2}}\theta^{n-1}_{j-\frac{1}{2}} - p^{n-1}_{j-\frac{1}{2}} f^n_{j-\frac{1}{2}} + p^n_{j-\frac{1}{2}} f^{n-1}_{j-\frac{1}{2}}\right) = T^{n-1}_{j-\frac{1}{2}} \tag{24}$$

$$f_0^n = 0, \qquad u_0^n = 0, \qquad p_0^n = 0, \qquad u_J^n = 1, \qquad \theta_J^n = 0 \tag{25}$$

Where

$$\alpha_n = \frac{p_0 \xi^{n-\frac{1}{2}}}{\kappa_n} \tag{26}$$

$$R^{n-1}_{j-\frac{1}{2}} = -L^{n-1}_{j-\frac{1}{2}} + \alpha_n\left[(fv)^{n-1}_{j-\frac{1}{2}} - (u^2)^{n-1}_{j-\frac{1}{2}}\right] \tag{27}$$

$$L^{n-1}_{j-\frac{1}{2}} = \left[\frac{v_j^n - v_{j-1}^n}{h_j} + p_1(fv)_{j-\frac{1}{2}} - p_2(u^2)_{j-\frac{1}{2}} + p_3\theta_{\mathrm{j}-\frac{1}{2}} - p_4\mathrm{u}_{\mathrm{j}-\frac{1}{2}} + p_5\mathrm{v}_{\mathrm{j}-\frac{1}{2}}\right]^{n-1} \tag{28}$$

$$T^{n-1}_{j-\frac{1}{2}} = -M^{n-1}_{j-\frac{1}{2}} + \alpha_n\left[(fp)^{n-1}_{j-\frac{1}{2}} - (u\theta)^{n-1}_{j-\frac{1}{2}}\right] \tag{29}$$

$$M^{n-1}_{j-\frac{1}{2}} = \left[\frac{1}{Pr}\frac{p_j^n - p_{j-1}^n}{h_j} + p_1(fp)_{j-\frac{1}{2}} + p_6\theta_{\mathrm{j}-\frac{1}{2}} - \mathrm{p}_7(u\theta)_{\mathrm{j}-\frac{1}{2}} + p_5\mathrm{p}_{\mathrm{j}-\frac{1}{2}}\right]^{n-1} \tag{30}$$

If we assume $f_j^{n-1}, u_j^{n-1}, v_j^{n-1}, \theta_j^{n-1}$ and p_j^{n-1} to be known for $0 \le j \le J$, equations (23) to (25) are a system of $5J+5$ equations for the solutions of $5J+5$ unknowns $(f_j^n, u_j^n, v_j^n, \theta, p_j^n)$, $j = 0,1,2,\ldots\ldots\ldots,J$. These nonlinear systems of algebraic equations are to be linearized by Newton's Quasi linearization method and solved in a very efficient manner by using the Keller-box method (Keller and Cebeci [19], Cebeci and Bradshaw [20]).

4. Results and Discussions

The aim of this present work to investigate the effect of heat generation as well as MHD on free convection flow from a vertical permeable circular cone maintained at non-uniform surface heat flux. The results are presented in terms of the velocity, $f'(\xi, \eta)$, and temperature, $\theta(\xi, \eta)$ profiles, the local skin friction, $\frac{1}{2}C_f Gr_x^{1/5}$, and the local Nusselt number, $\frac{Nu_x}{Gr_x^{1/5}}$,for different values of pertinent parameters such as heat generation parameter,*Q*, and magnetic field parameter, *M*, in Figures 3-6.

For code validation,the present results are compared with the results of Hossain and Paul [6]. The numerical values of the local skin friction, $\frac{1}{2}C_f Gr_x^{1/5}$ and local Nusselt number,$\frac{Nu_x}{Gr_x^{1/5}}$are depicted in Table 1 for Pr = 0.1 and *m* = 0.5 at different values of ξ = 0, 2, 4. Here, the parameters *M* and *Q* are ignored to make the numerical data comparable with [6]. It is evident from Table 1that the present results agreed well with the results of Hossain and Paul [6].

Table 1: Numerical values of skin Friction, $f''(\xi, 0)$and Nusselt number, $\frac{1}{\theta(\xi,0)}$for Prandtl number,Pr = 0.1, heat flux gradient, m = 0.5, Magnetic Parameter, M = 0, and heat generation parameter, Q = 0 .

ξ	$F''(\xi, 0)$		$1/\theta(\xi, 0)$	
	Present	Hossain and Paul [6]	Present	Hossain and Paul [6]
0	2.28882	2.29051	0.33168	0.33174
2	2.82802	2.83194	0.41472	0.41568
4	2.73784	2.74038	0.51619	0.51729

4.1: Velocity and temperature profile

In figure 3(a-b), the effect of varying Magnetic parameter, *M* (= 0.0 , 2.0 , 4.0), on the velocity,$f'(\xi, \eta)$, and temperature , $\theta(\xi, \eta)$,distribution against η at ξ = 0.0 , 2.0 , 4.0 for Prandlt number, Pr = 0.3, heat flux gradient, *m* = 0.5, heat generation parameter, *Q* = 0.1 areshown.

From the figure 3(a-b) it is seen that the velocity decreases but the temperature increases as the magnetic parameter *M* increases. Also when the suction parameter,ξ, increases then both the velocityand the temperature decreses. Figure3(a-b) also shows that for large suction parameter the velocity and the temperature profile becomes closer for different values of *M*. In these figures we observe that at each value of ξ as well as *M*, there exist local maxima in velocity profiles within the boundary layer region. It can also be seen that as ξ increases, the local maxima of the veocity profiles become closer to the surface of the cone.

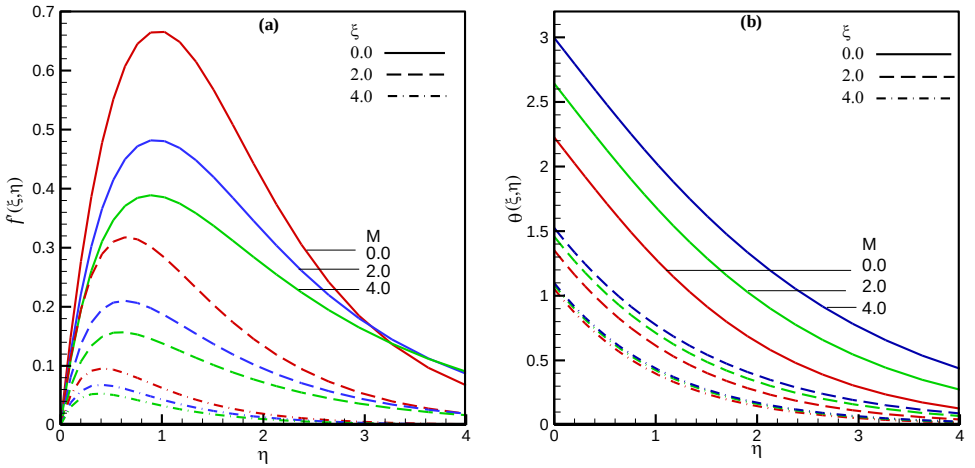


Figure 3: (a) Velocity profile and **(b)** Temperature profile against similarity variable η, for different values Magnetic parameter, M = 0.0, 2.0, 4.0 , and the suction parameter,ξ = 0.0, 2.0, 4.0 , while Pr = 0.3, m = 0.5, Q = 0.1.

Now the effect of varying heat generation parameter, Q = 0.0, 0.5, 1.0 , on the velocity, $f'(\xi, \eta)$, and temperature , θ (ξ, η),distribution against the similarity variable, η at ξ = 0.0, 2.0, 4.0 while Prandlt number, Pr = 0.3, heat flux gradient, m = 0.5, magnetic parameter, M = 0.3, are shown in the following figure 4(a-b).

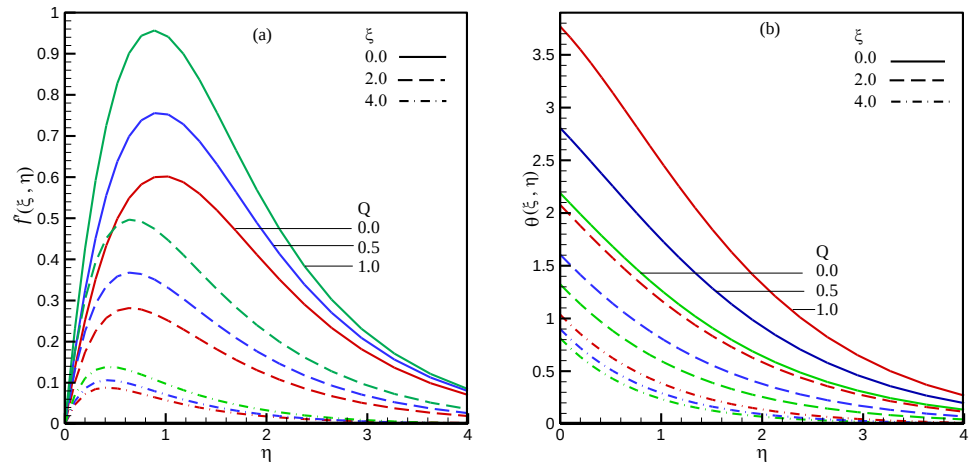


Figure 4: (a) Velocity profile and **(b)** Temperature profile against similarity variable η, for different values heat generation parameter, Q = 0.0, 0.5, 1.0 , and the suction parameter, ξ = 0.0, 2.0, 4.0, while Pr = 0.3, m = 0.5, M = 0.3.

Figure 4(a-b) shows that when the heat generation parameter Q increases, then both the velocity$f'$and the temperatureθincrease. Also from here we see that, for increasing values of the suction parameter ξ the velocityf'as well as temperatureθ is decreasing. It is also observed that at each value of ξ as well as Q , there exists local maxima in velocity profilef' within boundary layer region. Observing these local maxima, it is clear that as ξ increases, the local maxima of the veocity profilesf' become closer to the surface of the cone from which we can conclude that the velocity bounday layer decreases.

4.2: Skin friction and Nusselt Number

The variation of the local skin friction,$\frac{1}{2}C_f Gr_x^{1/5}$, and local Nusselt number,$\frac{Nu_x}{Gr_x^{1/5}}$, for different values of Magnetic parameter, M (= 0.0,1.0,2.0,3.0,4.0), against the suction parameter, ξ, while the heat generation parameter, Q = 0.1, heat flux gradient, m = 0.5, and Prandlt number, Pr = 0.3, are illustrated in figure 5(a-b).

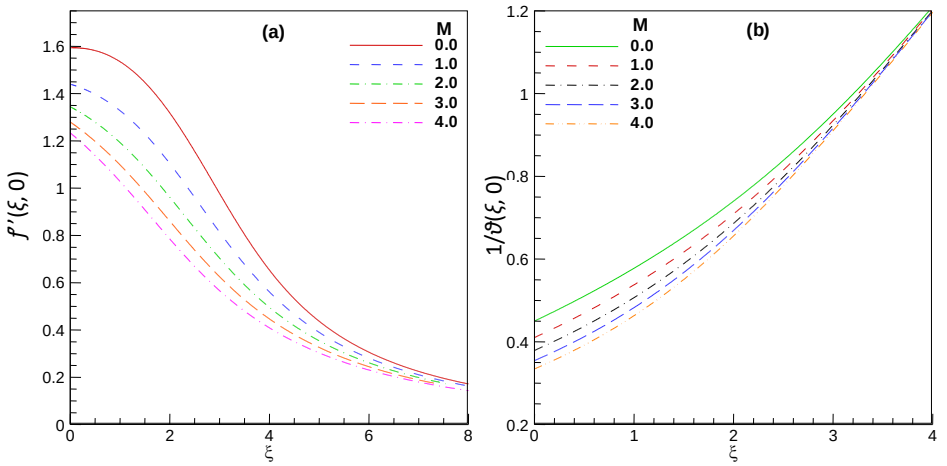


Figure 5: (a) Skin-friction and **(b)**Nusselt number against suction parameter, ξ, for different values of Magnetic parameter, M =0.0, 1.0, 2.0, 3.0, 4.0, while Q = 0.1, m = 0.5, Pr = 0.3.

Figure 5(a) shows the distribution of local skin friction against the suction parameter ξ,for different values of Magnetic parameter, M = 0.0, 1.0, 2.0, 3.0 and 4.0. Here it is seen that the skin frictiondecreases owing to increases the value of Magnetic parameter M. And also for increasing values the suction parameter $\xi$the skin friction decreases to the asymptotic value. Figure 5 (b) shows the distribution of local Nusselt number against the suction parameterξ for the different values of Magnetic parameter, M = 0.0, 1.0, 2.0, 3.0 and 4.0. From this figure we observe that the value of local Nusselt number increases as ξ increases and also for increasing values of Magnetic parameter$M$ the value of Nusselt number decreases.The magnetic field acts against the flow and reduces the skin friction as well as the rate of heat transfer.

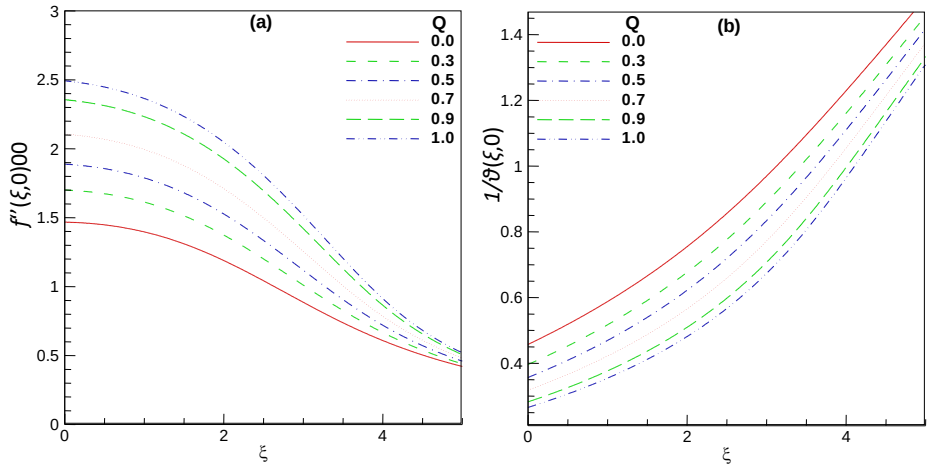


Figure 6: (a) Skin-friction and **(b)**Nusselt number against suction parameter ξ for different values of heat generation parameter, Q =0.0, 0.3, 0.5, 0.7, 0.9, 1.0, while Pr=0.3, m = 0.5, M = 0.5.

In figure 6 (a-b) we have presented the variation of local skin friction,$\frac{1}{2}C_f Gr_x^{1/5}$, and local nusselt number,$\frac{Nu_x}{Gr_x^{1/5}}$, against suction parameter, ξ, for varying heat generation parameter, Q (= 0.0, 0.3, 0.5, 0.7, 0.9, 1.0), while Prandlt number, Pr = 0.3, heat flux gradient, m = 0.5, and heat generation parameter, M = 0.5 are fixed.

From figure 6(a) we observe that the value of local skin frictionincreases with the increase in the value of heat generation parameter$Q$. On the other hand as the suction parameter ξ increases the value of skin friction decreases to the asymptotic value. From figure 6(b) it is seen that the value of local Nusselt number decreases as the value of heat generation parameter$Q$ increases. And also when increasing the value of suction parameter ξ, the value of Nusselt number increases.

5. **Conclusions**

In this work, we have investigated numerically the effect of heat generation and MHD on free convective flow from a vertical circular cone maintained at non-uniform surface heat flux. The results are presented in terms of the velocity,$f'(\xi,\eta)$, and temperature,$\theta(\xi,\eta)$ profiles, local skin friction,$\frac{1}{2}C_f Gr_x^{1/5}$, and local Nusselt number,$\frac{Nu_x}{Gr_x^{1/5}}$, for different values of pertinent parameters such as heat generation parameter,Q, magnetic parameter, M. From this investigation, we may conclude the followings:

1. The velocity decreases for increasing values magnetic parameter whereas the velocity increases asheat generation parameter increases. But the temperature increases for both the increasing values of the magnetic parameter and heat generation parameter.

2. The velocity boundary layer decreases for the increasing values of both magnetic parameter and heat generation parameter.
3. The value of local skin friction and the local Nusselt number decreases as the magnetic parameter increases. Again when the suction parameter increases, the local skin friction decreases but the Nnusselt number increases.
4. The value of local skin friction increases whereas the local Nusselt number decreases as the heat generation parameter increases. While the suction parameter increasing then skin friction decreasing but the Nusselt number increasing.

References

1. Hering, R.G. and Grosh, R.J, Laminar free convection from an isothermal cone, Int. J. Heat Mass transfer, 5, 1059-1068, 1962.

2. Hering, R. G, Laminar free convection from a non isothermal cone at low Prandlt numbers. Int. J. Heat Mass transfer, 8, 1333-1337, 1965.

3. Roy, S, Free convection over a slender vertical cone at high Prandlt numbers. ASME J. Heat Transfer, 101, 174-176, 1974.

4. Na, T. Y. andChiou, J. P, Laminar natural convection over a frustum of a cone. Appl.Sci. Res., 35, 409-421, 1979.

5. Alamgir M, The overall heat transform from vertical cones in laminar natural convection: an approximate method. ASME J. Heat Transfer, 101, 174-176, 1989.

6. Hossain, M. A. and Paul, S. C, The natural convection flow from a heated vertical permeable circular cone with non uniform surface temperature, ActaMechanica,151, 103-114, 2001.

7. Hossain, M.A. and Paul, S.C, The natural convection flow from a heated vertical permeable circular cone with non uniform surface heat flux, Heat and Mass Transfer,37, 167-173, 2001.

8. Kuiken, H. K,Magnetohydrodynamic free convection in a strong cross field,Journal of Fluid Mechanics, 4(1), 21-38, 1970.

9. Rapit, A. and Kafoussias, N, MHD free convection flow and mass transfer through a porous medium bounded by an infinite vertical porous plate with constant heat flux, Canadian Journal of Physics, 60(12), 1725-1729, 1982.

10. Sparrow, E. M.andCess, R.D. The effect of magnetic field on free convection heat transfer, Int. J. Heat Mass transfer, 3, 267-274,1961

11. Hossain, A. and Ahmed, M, MHD forced and free convection boundary layer flow near the leading edge, Int. J. Heat Mass transfer, 33(3), 571-575,1990.

12. Hossain, M. A., Das, S.K. and Pop, I, Heat transfer response of MHD free convection flow along a vertical plate to surface temperature oscillation, Int. J. Non linear Mechanics, 33(3), 541-553, 1998.

13. Vajravelu,K. andHadjinoiaou, A, Heat transfer characteristics in the laminar boundary layer of viscous fluid over a stretching sheet with viscous dissipation or frictional heating and internal heat generation. Int. Comm. Heat Mass Transfer, 20, 417-430, 1993.

14. Chamkha, A. J. and Camille, I, Effects of heat generation or absorption and thermophoresis on hydrodynamic flow with heat and mass transfer over a flat plate. Int. J. Numer. Meth. Heat fluid flow, 10(4), 432-438, 2000.

15. Mendez, F. and Trevino,C, The conjugate conduction-natural convection heat transfer along a thin vertical plate with non-uniform heat generation, Int. J. Heat Mass Transfer, 43, 2739-2748, 2000.

16. Molla, Taher, Chowdhury, and Hussain, The MHD natural convection flow on a sphere in presence of heat generation, Int. J. Nonlinear Analysis: Modeling and control, 10(4), 349-365, 2005.

17. Molla, M., Paul, S. C. and Hossain, M. A, Natural convection flow from a horizontal circular cylinder with uniform heat flux in presence of heat generation. Applied Mathematical Modelling, 33, 3226-3236, 2009.

18. Keller, H.B, Numerical methods in boundary layer theory. Annual Rev. Fluid Mech., 10, 417-433, 1978.

19. Keller, H. B. and Cebeci, T, Accurate numerical methods for boundary layer flows. Part I. Two dimensional laminar flows. Proc. Int. Conf. on Numerical Methods in fluid dynamics. (Published as Lecture Notes in Physics. Springer), 1971.
20. Cebeci, T. and Bradshaw, P, Physical and computational aspects of convective heat transfer. NY; Springer, 1984.

JJIT, Vol. **6**, June, 2017, 57-68

K-MEANS CLUSTERING MANIFESTED PROTEIN MOTIFS SHARE SIMILAR CHEMICAL PROPERTIES

ABDULLAH ZUBAER AND MD. FAZLUL KARIM PATWARY

Institute of Information Technology, Jahangirnagar University, Savar, Dhaka-1342

Keywords: motif; k-means-clustering; PROSITE; pattern

Running Title: Protein Motifs Sharing Similar Properties

Abstract

Protein sequence motif is a portion of a protein sequence that remains conserved throughout the process of evolution. Motifs play significant roles in protein structure and function. Previous research reported several techniques for searching motifs and classifying proteins, but no recognizable research found on the chemical nature of motifs. The current study focused on the amino acid composition and chemical properties of the motifs collected from different proteins from different prokaryotes and eukaryotes. We have developed a python script for the conversion of motif pattern to standard amino acid sequence and use them to compute pI, Aliphatic Index and GRAVY with a bioinformatics tool ProtParam. K-means cluster analysis was adopted for clustering the data to find out natural groups among the motifs. Two R scripts were executed to identify the optimal number of centroid and to generate k-means cluster and PCA graph. The results showed that the motifs, although from different proteins from different organism, maintain similar chemical properties. The current methodology and the outcomes of this study will be beneficial for the future study on motifs and their evolution.

I. Introduction

Protein sequence analysis is one of the leading research issues in Bioinformatics. Protein sequence analysis and classification in different families were a major challenge in past few decades. Classifying proteins into families was conducted using special sequence signals that are found to be conserved, in addition to the protein overall sequence and domain matching [1]. Those conserved sequences are called "motif". Motifs are characteristic and conserved small peptide sequence. They are important for certain function or specific structural signatures. On the basis of motif patterns, protein classification has been done and all the proteins are categorized into several protein families. Motifs can be found in functionally active domain or in a structural domain of a protein. The database representation of motifs can be a pattern or a matrix. Patterns are series of notations including amino acid one-letter-code, whereas matrices are probability

scores for each amino-acid at each position of a fixed-length motif. In PROSITE, both of the formats are found [2].

However, most of the previous research on protein motif was focused on protein classification and new motif finding. There were a lot of motif finding methods and databases had been developed such as EXTREME [3], kmerHMM [4], PMS, eMotif, PROSITE [5], MEME [6] etc. A lot of motif finding web applications are available such as BLOCK-maker, ELM [7], MEME SUITE [6], SCOPE [8], RSAT, ScanProsite [9] etc. Different software and methods use different algorithms and strategies. For instance, kmerHMM uses HMM based method where EXTREME uses an EM algorithm. Moreover, some software use multiple algorithms simultaneously, such as SCOPE. Except those searching and motif identifying algorithm developments, some applied studies have been done. Watson et al. [10] showed that motifs are important for protein function prediction rather than only using sequence matching approach. There is no previous research has been recognized on data mining of motif patterns and their properties.

We have collected protein motif patterns from virus, bacteria, and eukaryotes. Some motifs are only specific to virus or bacteria or eukaryotes, some can be found in more than one group, some also are ubiquitous. We have used several computational and programming tools such as PROSITE, ProtParam, Python, R etc. to study those motif patterns, their chemical properties and used k-means cluster analysis to find out a relationship among them.

II. Methodology

Data Collection

Expert Protein Analysis System (ExPASy) is a bioinformatics portal of SIB Swiss Institute of Bioinformatics [11], which contains the PROSITE [12]. PROSITE is a database of protein families and protein domains which is built on a huge number of protein grouped on the basis on sequence similarity [5], It is currently consists of hundreds of protein motif patterns and profiles (matrix) specific for more than a thousand protein families or domains from different forms of life such as viruses, bacteria, archaea and eukaryotes. We have collected protein motif patterns from PROSITE database browsing by taxonomic scope. Data collection was based on diversity of organisms. We have included motif pattern from viruses, bacteria and eukaryote (including protists, fungi, plants and animals). Collected data categorized into seven groups as following.

Table 1: Number of data collected and categorized into group on the basis of their source organism.

Source Organism	**Number of Data**
Virus only	02
Bacteria only	24
Eukaryotes only	16
Virus and Bacteria	09
Virus and Eukaryotes	11
Bacteria and Eukaryotes	16
Virus, Bacteria and Eukaryotes	09
Total =	87

Data collection was randomized, given that the availability of the motif pattern data (NOT motif profile matrix) in PROSITE database.

Conversion of "pattern" sequence into standard amino acid sequence

Representation of protein motif pattern is different from the standard IUPAC one-letter-code amino acid sequence (protein primary structure notation). In addition to one-letter-code for amino acid, motif patterns use some special notations such as "[ ]", "{ }", " - ", "x". The pattern format data collected from PROSITE cannot be used directly because of its incompatibility with most of the bioinformatics and data analysis software. So, we have developed a python- (a general purpose programming language; https://www.python.org/) script to convert the protein motif pattern into IUPAC amino acid sequence.

Calculating pI, Aliphatic Index and GRAVY

ProtParam tool [13] is a well-recognized online based bioinformatics tool for protein primary structure analysis. It calculates various physico-chemical properties of protein such as Theoretical pI (isoelectric point), Extinction coefficient, Half-life, Instability index, Aliphatic index and GRAVY (Grand Average of Hydropathy). As we are working with a part of a protein or a short string of amino acids, Extinction coefficient, Instability index and Half-life will not signify anything, because light absorption, instability of protein in test tube or *in vivo* half-life cannot be a real value of a fraction of a protein. So we measured pI, Aliphatic index and GRAVY for our study.

K-means clustering

K-means clustering uses a simple learning algorithm for cluster analysis. It aims for partitioning ***n*** number of observations into ***k*** number of clusters where each observation

belongs to the nearest cluster-mean or centroid. In K-means clustering, clusters are defined based on Euclidean distances so as to reduce the variability of individuals within a cluster, while maximizing the variability between clusters [14].

We have used an R (statistical software) implementation of this algorithm which is named as "**kmeans**" and built in R. Moreover, we have used an external code [15]. The code was used to find out most efficient value for ***k*** in clustering. It also offers a z-score optimization and PCA clustering. The script (and documentation) is available at "http://www.mattpeeples.net/kmeans.html#help" and should run in R.

III. Results and Discussion

Motif-pattern Data from Virus, Bacteria and Eukaryotes

PROSITE is a database for protein domain and motif (can be pattern or matrix). We have collected motif-patterns from PROSITE and the collected data was tabulated in a Microsoft Excel file and then transferred into the table (*see Supplementary Table 3*). Motif name, Accession in PROSITE database, Function or Description of the motif and Pattern data were collected and categorized according to their source organism. Some motifs are specific to one organism group; some others belong to more than one group.

Python script for the conversion of "pattern" sequence into standard amino acid sequence

Python is a high-level programming language that is very popular now-a-days for its wonderful usability in general-purpose computing although it supports for object-oriented, imperative, functional programming and more. Python is designed for it excellent readability and easy syntax to express concepts in fewer lines of code. It uses an interpreter rather than compiler and make the code executable instantly. This scripting nature of python language made it popular in genomics and bioinformatics research.

Pattern data from PROSITE is not an accepted format to be used any bioinformatics software such as BLAST, ProtScale, ProtParam etc. So it was very essential to convert the pattern into a well-accepted sequence format. We have developed a python script to convert the pattern sequence to standard one-letter-code amino acid sequence. It takes input of a pattern from user and give an amino-acid sequence in output. The python code we developed is in the following:

```
#!/usr/bin/python
my_motif = raw_input('enter a motif pattern: ')
#e.g, awv-{ky}-x(4)-x-g[kt]s
def motif2seq (motif):
        import random
        motif = motif.lower()
```

```
        i = 0; j = 0; k = 0; l = 0; m = 0
        s = ''; t = ''; u = ''; v = ''
        x = len(motif)
        amino_acid = 'arndceqghilkmfpstwyv'
        while (i < x):
                if motif[i] in amino_acid:
                        s = s + motif[i]
                elif motif[i] == '{':
                        i = i + 1
                        while motif[i] != '}':
                                t = t + motif[i]
                                i = i + 1
                                w = amino_acid
                        for j in range(len(t)):
                                u = t[j]
                                ami_rp = w.replace(u,'')
                        k = random.randint(0, len(ami_rp)-1)
                        random_aa = ami_rp[k]
                        s = s + random_aa
                elif motif[i] == '[':
                        i = i + 1
                        while motif[i] != ']':
                                v = v + motif[i]
                                i = i + 1
                        l = random.randint(0, len(v)-1)
                        rand_aa = v[l]
                        s = s + rand_aa
                elif motif[i] == 'x':
                        m = random.randint(0, len(amino_acid)-1)
                        rand_a = amino_acid[m]
                        s = s + rand_a
                else:
                        pass
                i += 1
        return ss
n = motif2seq(my_motif).upper()
print n
```

Calculated pI, Aliphatic Index and GRAVY

Aliphatic Index, pI and GRAVY are the basic characteristic properties for peptides. As motifs are peptides so that we can calculate their amino acid properties. We have used ProtParam for calculating Aliphatic Index, pI and GRAVY for motifs. The outputs of ProtParam for each of the motifs are tabulated below. In this tabulation we have added an extra two letter identifier before each accession number. The identifiers are - “vo” for virus

only, “bo” for bacteria only, “eo” for eukaryote only, “vb” for virus and bacteria, “ve” for virus and eukaryote, “be” bacteria and eukaryote, and “ag” for all groups.

Table 2: Motif accession number with identifier to recognize their source group and the calculated Aliphatic Index GRAVY and pI for each motif is listed here.

Accession	Aliphatic Index	GRAVY	pI	Accession	Aliphatic Index	GRAVY	pI
vo_PS00555	19.33	−0.973	3.56	eo_PS00411	114.17	0.825	4
vo_PS00418	67.69	−0.269	10.84	eo_PS00128	117	1.95	5.5
bo_PS01328	134.62	0.385	5.96	eo_PS00612	0	−0.063	7.76
bo_PS01157	0	−1.15	6.74	eo_PS00613	141.82	0.082	9.6
bo_PS01092	70.91	−0.191	6.1	eo_PS00265	0	−0.708	10.28
bo_PS01093	52	−1.187	8.6	eo_PS01107	35.45	−0.218	11
bo_PS00274	12.5	−1.913	10	eo_PS01190	78	−1.49	6.56
bo_PS00159	48.75	−0.15	6	vb_PS00694	32.22	−1.367	8.5
bo_PS00160	39	0.37	8.59	vb_PS00695	12.5	0.062	4
bo_PS00968	91.33	0.553	6.74	vb_PS00588	49.29	−0.657	9.31
bo_PS00969	48.57	−0.35	4.63	vb_PS00576	132.14	0.75	9.7
bo_PS00594	90.71	0.779	6.74	vb_PS01307	57.5	0.308	4
bo_PS00949	74.71	−0.941	6.8	vb_PS00277	54.44	−0.611	6.74
bo_PS01139	128.12	0.719	6.06	vb_PS00278	52.31	−1.146	4.83
bo_PS00494	90.71	0.221	6.78	vb_PS00875	37.14	−0.01	6
bo_PS00363	62.94	−1.312	3.92	vb_PS00922	103.75	0.181	4.13
bo_PS00364	30	−0.208	3.8	ve_PS00406	58	0.15	3.67
bo_PS60030	54.44	0.833	5.5	ve_PS00141	162.5	1.775	3.8
bo_PS01303	78	0.47	5.52	ve_PS00237	62.94	0.135	8.05
bo_PS00146	83.57	−0.093	9.53	ve_PS00520	83.57	−0.743	10.03
bo_PS00743	97.14	−0.157	5.21	ve_PS00031	24.38	0.237	7.83
bo_PS00744	11.11	0.056	8.6	ve_PS00239	0	−2.7	3.93
bo_PS00336	146.25	0.188	6	ve_PS00240	83.57	0.171	6.73
bo_PS00337	80	1.282	8.27	ve_PS00915	52.73	−1.082	4.68
bo_PS01324	59.44	−0.511	7.8	ve_PS00916	115.62	−0.369	4.41
bo_PS00330	83.57	−0.093	3.77	ve_PS00472	86.67	0.611	7.88

Accession	Aliphatic Index	GRAVY	pI	Accession	Aliphatic Index	GRAVY	pI
eo_PS00305	61.82	0.009	3.56	ve_PS00299	124.44	0.944	5.99
eo_PS00796	123.64	0.382	8.79	be_PS00574	89.17	−0.242	6.46
eo_PS00797	97.5	−0.695	3.71	be_PS00367	70.91	−1.145	4.21
eo_PS00262	9.09	1.109	5.49	be_PS00506	32.22	−0.211	5.08
eo_PS00242	48.75	0.3	8.75	be_PS00549	154.17	1.608	5.28
eo_PS00243	60	0.823	3.79	be_PS00605	42.14	−0.85	4.14
eo_PS00252	65	−0.408	6.92	be_PS00449	49	−0.63	9.47
eo_PS00253	52.31	0.569	5.99	be_PS00798	37.69	−0.977	4.14
eo_PS00424	141.82	1.164	3.67	be_PS00060	129.44	1.328	3.8
be_PS00913	115.62	0.219	6.72	ag_PS01132	32.5	−1	4.37
be_PS00440	97.78	−0.05	4.18	ag_PS00122	22.31	0.285	5.52
be_PS00439	149.33	0.5	4.37	ag_PS01095	43.33	0.222	3.67
be_PS00073	55.71	−1.107	6.06	ag_PS00522	38.67	−0.353	3.77
be_PS00072	53.64	−0.2	3.49	ag_PS00191	85	0.887	6.74
be_PS00786	86.67	−0.244	5.24	ag_PS00460	45.33	−0.333	8.96
be_PS00785	89.17	0.542	4.35	ag_PS00108	78	−0.44	6.74
be_PS00895	137.5	1.35	5.52	ag_PS00640	68.82	−0.094	5.97
ag_PS00139	8.33	−0.275	4				

K-means Cluster Analysis

K-means clustering uses ***k*** number of center or mean value and every instance of the input table is assigned or cluster against one of them according to their Euclidian distance from each center. We have taken Table 2 as our input, which have Aliphatic Index, GRAVY and pI as variables, and every instance was put according to the Accession Number of protein motif patters. The ***ods*** table file (Libre Office Cacl program file format similar to Microsoft Office Excel file format) was converted into ***csv*** file format to make the table compatible for **R** input. The conversion of file was done by a Linux-based tool "**unoconv**".

This command generated a ***csv*** file which was used as an input for the following **R** script run in Linux Terminal.

```
x <- read.csv("file.csv", header=TRUE, row.names=1)
# run K-Means
km <- kmeans(x, 3, 15)
```

```
# print components of km
print(km)
# plot clusters
plot(x, col = km$cluster)
```

In this script, we have set 3 centers or **k = 3**, and the number of iteration was set to 15. The right choice of **k** is often ambiguous, and determining the value of **k** is a problem in data clustering what can be resolved by different methods such as elbow method, silhouette method etc [16].

The R script, (kmeans.R) from Pepples, was used for similar analysis where the script calculate the best fit value for **k** in k-means cluster analysis. It used elbow method which calculates and compares the sum of squared error (SSE) for a number of cluster solutions. A plot of the SSE against a number of sequential cluster levels can provide a useful graphical way to choose an appropriate cluster level (*see Supplementary Figure 1*). The most suitable value can be found where the elbow forms. After running the "kmeans.R" script, it produced the following images on the basis of that the value of **k** can be determined.

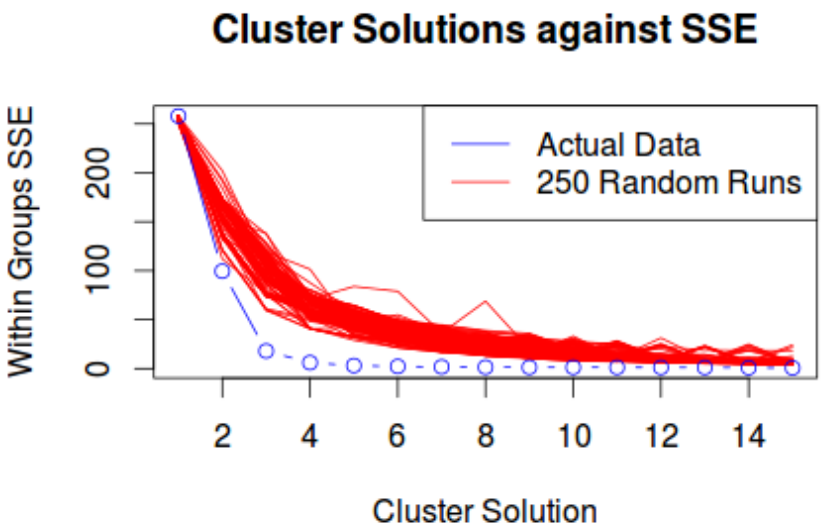


Figure 1: This plot indicates for the best value of **k** by using elbow method. The SSE for the actual data does decrease faster than the 250 randomized data sets. This suggests that the data set has structure and clusters are present. There is somewhat of a reduction in the rate of SSE decrease at about the 3 cluster solution. The "elbow" in this plot is very clear and obvious that k=3 will be the best value.

From the figure 1, we can determine that the best value for **k** is 3. So by using k=3, the first R script generated output with three clusters. k-means clustering with 3 clusters of sizes are 37, 33, 17. That means, there are three cluster formed, first one consist of 37 instances, second one 33 instances, and the third one 17 instances. The cluster means are as following:

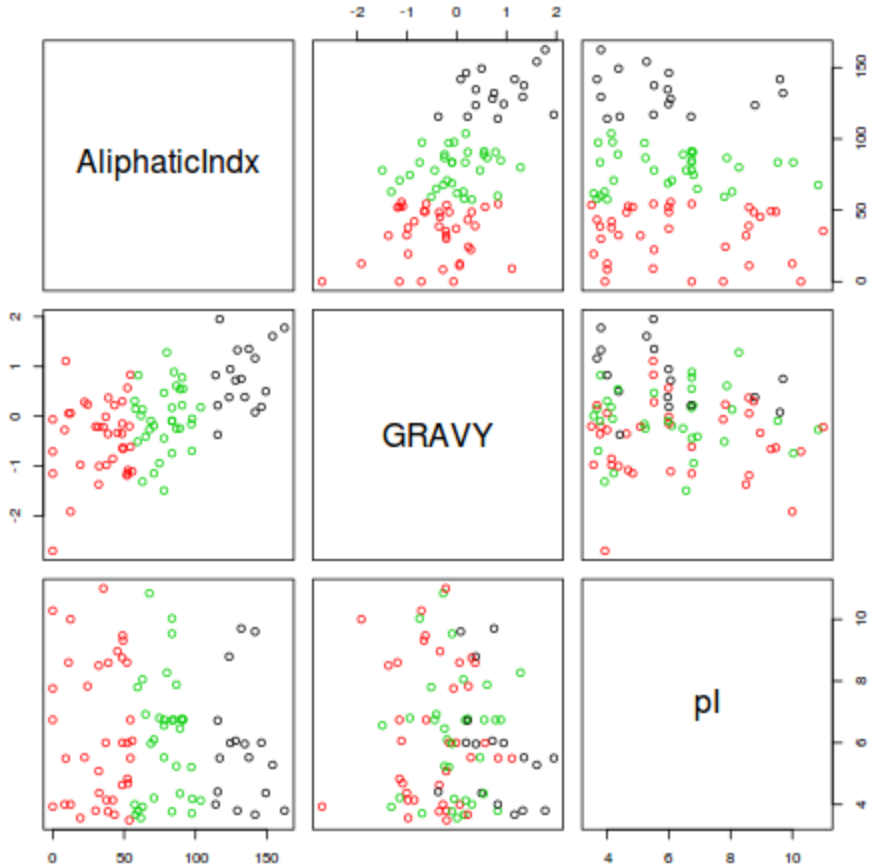


Figure 2: Three centroid k-means clustering graphs using three variables- Aliphatic Index, GRAVY and pI. There are 6 graphs generated by R plot built with the combinations of three variables.

Proteins from bacteria only, virus only, eukaryote only, bacteria and virus, bacteria and eukaryote, virus and eukaryote, and virus-bacteria-eukaryote are gone mixed up. Motifs from those distinct groups formed cluster together to form three clusters. Every distinct organism-group members are found in all three clusters (*see Supplementary Table 1 and 2*). That means the motifs as well as proteins are not organism dependent. They have some other characteristics that are maintained by the motifs. To enhance the confidence of this result we have used kmeans.R script to make a PCA cluster.

The kmeans.R script produces the following clusters. The script conducts a principal components analysis (PCA) on the original data set. Each sample is then displayed on a scatter plot of the first two principal axes of the PCA with the clusters outlined. If the clusters are strong at the selected level, there should not be substantial overlap in the distributions of the cluster outlines on the PCA plot. It is important to note that PCA plots may not be particularly useful for K-means analysis of data sets with a large number of samples or a large number of variables. There is no hard and fast rule for this, but the percent of the variability explained by the PCA provides some clue as to the potential utility of this approach. Principal component analysis (PCA) is a statistical procedure that uses a transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables used in the working dataset.

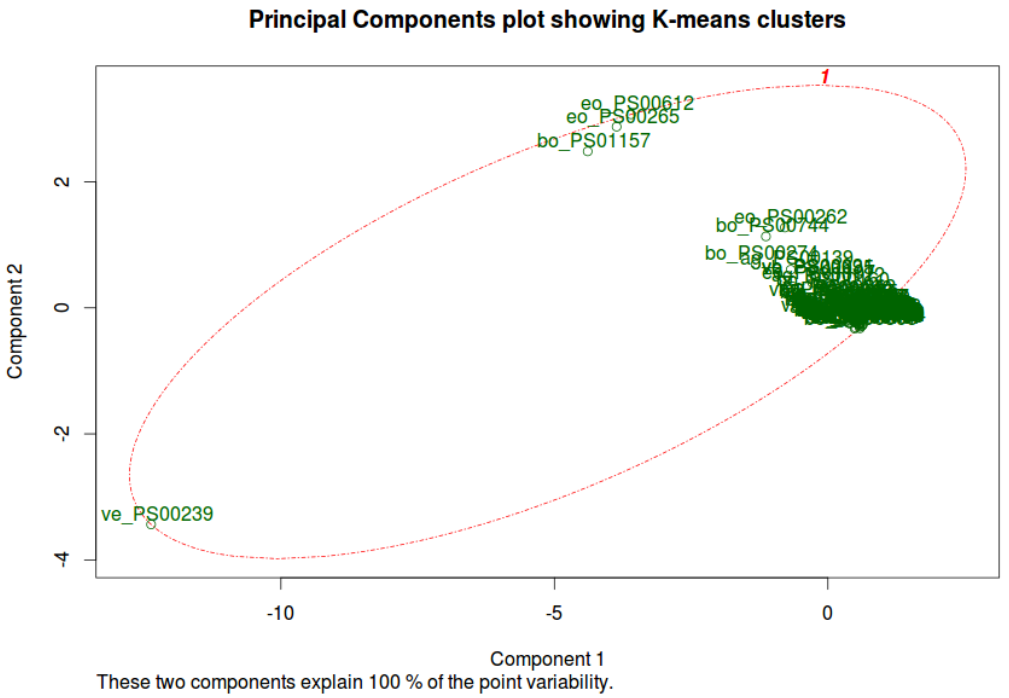


Figure 3: PCA graph for three variables of motif characteristics. The graph showed that there are three cluster formed by the motifs where almost all motifs from different source of organism clustered together.

From both k-means cluster and PCA graph proved that there is no significant impact or relation of the organisms on motifs. But motifs maintain some special features that are organism independent. However, all motifs maintain some common characteristics as

almost all of them cluster together. They contain similar pI, aliphatic index and GRAVY even they are from different source.

IV. Conclusion

Motifs are conserved string of amino acids reside in a protein and help it to have a particular function. Motifs are important for protein classification and identifying protein function. The current study investigated the impact of source organism on motifs amino acid properties. Moreover, it was a search for natural groups among the motifs collected from different prokaryotes and eukaryotes. The study found that there is no effect of organism on its protein motif, thus the evolution of motifs are independent of the organism. Interestingly, motifs maintain a similar amino acid properties and chemical properties regardless the source organism.

The methodology used in this study will be helpful for further study in this area, as a python script has been developed and also some R scripts were used to modulate data mining analysis. For further research, other than considering source-organism, protein localization (such as membrane bound or soluble protein) or other important criteria can be considered and analyze the motifs' properties and evolution by using more elaborate data mining and statistical methods.

References

1. Yun, X., (2007). BioPM:An Efficient Algorithm for Protein Motif Mining. *1st International Conference on Bioinformatics and Biomedical Engineering . Wuhan. 6-8 July 2007. IEEE*, pp394 – 397. 10.1109/ICBBE.2007.104.

2. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J.A., Hofmann, K., Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res., 30 (1),* 235-238. doi:10.1093/nar/30.1.235

3. Quang, D., and Xie, X. (2014). EXTREME: an online EM algorithm for motif discovery. *Nucl. Acids Res. 30 (12),* 1667-1673. doi:10.1093/bioinformatics/btu093

4. Wong, K.C., Chan, T.M., Peng, C., Li, Y., Zhang, Z. (2013). DNA motif elucidation using belief propagation. *Nucleic Acids Res., 41(16),* e153. doi:10.1093/nar/gkt574. Epub 2013 Jun 29.

5. Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform., 3*, 265-274.

6. Bailey, L.T., Williums, N., Misleh, C., Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucl. Acid Res., 34 (suppl 2),* W369-W373. doi: 10.1093/nar/gkl198.

7. Dinkel, H., Roey, K.V., Michael, S., Kumar, M., Uyar, B., Altenberg, B., et al. (2016). ELM 2016--data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.,44(D1),* D294-300. doi:10.1093/nar/gkv1291.

8. Carlson, J.M., Chakravarty, A., DeZiel, C.E., Gross, R.H. (2007). SCOPE: a web server for practical de novo motif discovery. *Nucl. Acids Res., 35 (suppl 2),* W259-W264. doi: 10.1093/nar/gkm310.

9. De Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucl. Acids Res., 34,* W362-5.

10. J. D., Laskowski, R. K., Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Curr Opin Struct Biol.,15*: 275-284.

11. Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D.,Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucl. Acids Res., 40(W1),* W597-W603.

12. Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I. (2012) New and continuing developments at PROSITE. *Nucleic Acids Res.,* doi:10.1093/nar/gks1067

13. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook. *Humana Press,* pp. 571-607

14. Kintigh, K. W., and Ammerman, A.J. (1982). Heuristic Approaches to Spatial Analysis in Archaeology. *American Antiquity, 47,* 31-63.

15. Peeples, M. A. (2011). R Script for K-Means Cluster Analysis. [online]. Available: http://www.mattpeeples.net/kmeans.html.

16. David, J., Ketchen, Jr., Shook, C.L. (1996). The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal, 17 (6),* 441–458.

JJIT, Vol. **6**, June, 2017, 69-79

DISTRIBUTION OF CLOUD DATA CENTER WORLDWIDE : A RESPONE TIME APPROACH

MD WHAIDUZZAMAN[1], QI HAN[2]

[1]*Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh.*

[2]*Mobile Cloud Computing Research Lab, Faculty of Computer Science and Information Technology University of Malaya, Kuala Lumpu, Malaysia*

Abstract

A Data Center Network (DCN) is a vital component of cloud data centers. It consists of a large number of servers and switches that are virtualized and centrally connected with high speed communication links, which provide resources via on demand access to the users. Owing to the lack of network bandwidth and scalability as well as the high cost of establishing new data centers, it is important to select a suitable physical location for the data centers. The design of a DCN and its deployment worldwide have significantly affected its overall performance. In addition, data center virtualization is expected to achieve flexible control facility, low cost, scalable, efficient resource utilization, and energy efficiency. Therefore, cost efficiency, robustness, energy efficiency, and quick provisioning response time are required. In this paper, we focused on the global distribution of data centers from the User Base (UB) and the performance based on user response time. The findings indicate that the response time and its various patterns depend on the proximity of the physical location of the UB and the data center as well as their worldwide distribution.

Keywords : Cloud computing, data center network, data center distribution, response time

1. Introduction

Cloud computing is a network-based computing model that provides services on demand. Amazon, Google, Salesforce.com, and other corporations have established large data centers around the world to achieve massive computing tasks and data storage [1] [2]. Since the advent of cloud computing, data volume in the Internet has tremendously increased. The International Data Corporation (IDC) reported that the size of data generated worldwide in 2011 reached up to 1.8 ZB (1.8 trillion GB), and that such data are expected to increase 50-fold (reaching 35.2 ZB in 2020) in the next decade . The deployed data-management systems and processing mechanisms in data center networks (DCNs), such as BigTable , Dryad and MapReduce are responsible for managing and processing these massive data [3]. Currently, data centers are experiencing increased data communication traffic and tremendous growth of user traffic demand [4]. Recently, data centers have received considerable attention as cost-effective infrastructure for storing large volumes of data and hosting large-scale service applications [5]. Such large companies as Amazon, Google, Facebook, and Yahoo! routinely utilize data centers for storage, web search, and large-scale computations [6]. With the advent of cloud

computing, service hosting in data centers plays a crucial role in the future of the IT industry [7].

Reasonable cost and elastic utilization in accordance with prevailing business requirements are among the important factors considered by enterprises investing in DCNs. In recent years, cloud computing has gained tremendous attention from the industry and the academia. This is because a cloud computing provider offers a large pool of high-performance computing and storage resources that are shared among end users [8]. Users then subscribe to cloud computing services and receive computing and storage resources allocated on demand from the pool [9]. However, several enterprises still have misgivings about the cloud computing service models. For example, the network part of the data center has not been commoditized yet. The enterprise-class network equipment is expensive and is not designed to accommodate Internet-scale services in data centers. As a result, the use of enterprise class equipment limits end-to-end network capacity and leads to non-agility and the creation of fragmented server pools. Similarly, the increased number of servers requires a high end-to-end aggregate bandwidth [10].

The expansion of cloud computing has transformed the traditional data center, thus creating a new generation of cloud data centers. A cloud DCN can contain hundreds to thousands of servers in an economy of scale [11]. Thus, findings ways by which to efficiently integrate the exponentially increasing servers with a fault-tolerant high availability and usability, as well as significant aggregate bandwidth requirements for the DCN remains a challenge [12]. Considering that cloud-based DCN is more scalable and is simple to organize and operate, its cost is lower compared with the traditional DCN [13]. The architectural design of the DCN significantly affects its total performance; therefore, the design and the distribution of the data center has to be scalable, cost efficient, robust, and energy efficient [14].

Driven by these limitations, new trends geared towards virtualizing DCNs in addition to server virtualization have emerged. Similar to server virtualization, network virtualization aims to create multiple virtual networks (VNs) in addition to a shared physical network substrate , which can allow each VN to be implemented and managed independently. This model has a tremendous effect on user QoS performance when the user base (UB) is distributed worldwide. In this setup, the different users' geographical locations across the world must be considered, that is, every user deserves the same quality of services because that they paid for the same price for resource utilization [15]. Hence, it is necessary to ensure that the user service provisioning experience is consistent for all the users worldwide. For this reason, the distribution of data centers worldwide has a serious research impact in the user service experience perspective. In this paper, we aim to establish the data center, the UB distance, and their relevant response time in the global positioning perspective, as well as the distribution infrastructure model between users and the data center. The paper finds that the variations in response time depend on the physical location of the data center and the UB distance from the data center [16].

The rest of the paper is organized as follows: Section 2 describes the fundamental background of DCN; Section 3 discusses the experimental setup, simulation results with graphs and their significance; and Section 4 provides the conclusion.

2. Background

In this section, we present the related terminologies that are relevant to the study of DCNs [3].

Data Center: This consists of servers, storage, network devices, power distribution, and cooling systems.

Data Center Network: A DCN refers to the communication infrastructure shared in a data center, and includes the network topology, routing equipment, and protocols.

Virtualized Data Center: This is a center wherein the hardware can be virtualized through the use of software or firmware. A hypervisor segments the equipment into multiple isolated and independent virtual instances.

Virtual Network: This refers to a set of virtual networking resources. Thus, a VN is a component of a VDC.

User Base: This component refers to a group of users and generates traffic to represent the users.

Datacenter Controller: This component controls the data center activities.

VmLoad Balancer: This component refers to the load balance policy employed by data centers when serving allocation requests.

Cloud AppService Broker: This component refers to the service brokers that handle traffic routing between user bases and data centers.

Service Response Time: The service response time of the cloud provider indicates the total amount of time consumed to respond to a user request, including provisioning time, VM booting time, and IP address assigning time. Service response time indicates how fast a response is given to a user request in terms of service availability.

Maximum Response Time: This refers to the maximum promised response time by the service provider.

Average Response Time: This represents the total time taken to serve requests divided by the total number of requests.

Minimum Response Time: This refers to the lowest possible response time by the service provider.

The DCNs still rely on traditional TCP/IP protocol stack, resulting in the following limitations [3].

• no performance isolation;
• increased security risks;
• poor application deployment ability;
• limited management flexibility; and
• lack of support for network innovation

The recent emerging trend is geared towards virtualizing DCNs in addition to server virtualization. Network virtualization, which is akin to server virtualization, can create multiple VNs in addition to a shared physical network, thus allowing each VN to be managed independently.

3. Simulation of DC and UB Responses From Different Regions

For the current experiment, we used the simulation software CloudAnalyst [17, 18] . Several UBs worldwide and a DC are considered for the simulation. UB represents the component models of a group of users and generates traffic to represent the users. Table 1 shows the response time from the UBs in the different regions to the DC. The subsequent graphs depict the average, minimum, and maximum response times from the different regions around the world. UB1 is nearest to the DC, and UB5 is situated at the farthest location from the DC. Fig. 1 provides the graph of the response times.

A. Simulation Configuration

Six UBs are defined to represent the six main regions of the world with parameters. The size of the virtual machines used to host applications in the experiment is 100 MB. The virtual machines have 1 GB of RAM memory and 10 MB of available bandwidth. The simulated hosts have x86 architecture, virtual machine monitors Xen, and a Linux operating system. Each simulated data center hosts 20 virtual machines, each having 2 GB of RAM and 100 GB of storage. Each machine has 4 CPUs, and each CPU has a capacity power of 10000 MIPS. Time shared policy is used to schedule resources to VMs, and the simulation time is set to 10 hours. Users are grouped by a factor of 1000, and requests are grouped by a factor of 100.

B. Simulation Results

Table 1 lists the different response times and the UB relation, which are obtained using CloudAnalyst. Table 1 includes the data on the average, maximum, and minimum response times from the UB to the different data centers around the world. The simulation graphs

are shown in Figs. 1(a) to 1(f) for UserBase1 to UserBase6, respectively. All response times are measured in milliseconds.

Table 1: Data for UB to DC in different places in the world

UserBase (UB)	Avg (ms)	Min (ms)	Max (ms)
UB1	50.10	36.61	65.34
UB2	200.21	146.15	262.16
UB3	300.15	217.61	403.68
UB4	500.47	367.63	650.22
UB5	499.89	362.53	667.65
UB6	200.14	142.12	264.23

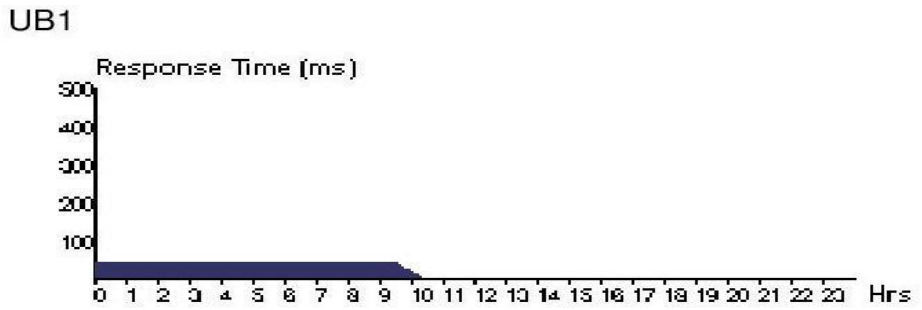


Figure 1 (a) : UB1 to DC average response time.

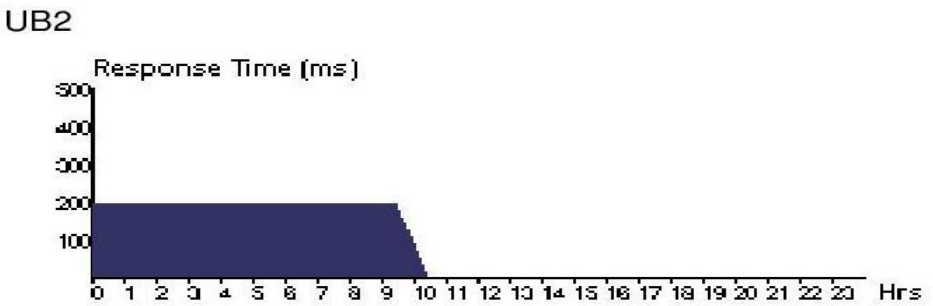


Figure 1 (b) : UB2 to DC average response time.

UB 3

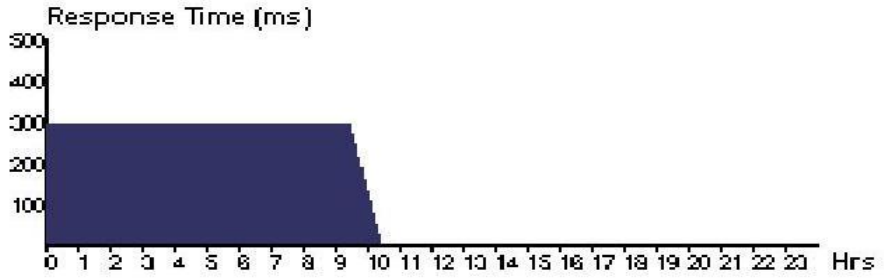


Figure 1 (c) : UB3 to DC average response time.

UB 4

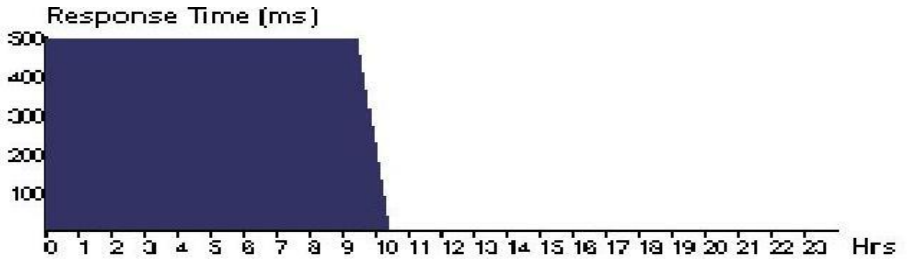


Figure 1 (d) : UB4 to DC average response time

UB5

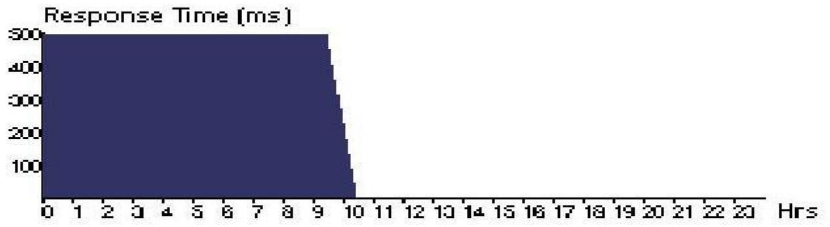


Figure 1 (e) : UB5 to DC average response time.

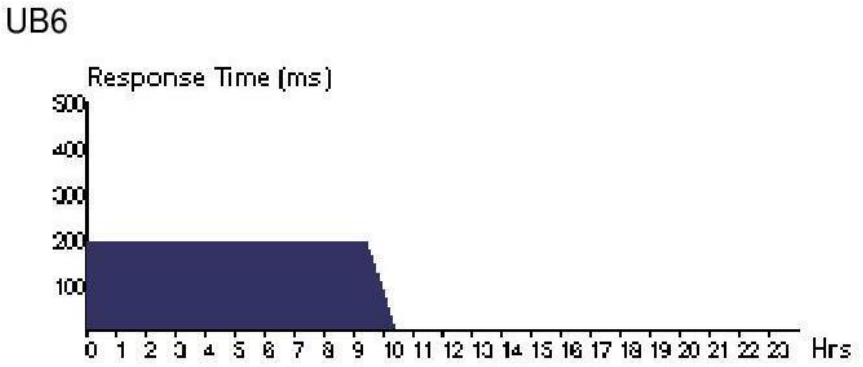


Figure 1 (f) : UB6 to DC average response time.

The above graphs indicate a significant trend in achieving the different response times and data center distributions worldwide. UB1 is the closest to the DC, and thus, less response time is needed. Conversely, UB5 is situated at the farthest point from DC, and thus, more response time is required. The results also show the improvement of the quality of services, with the time differences between the average and maximum response times for UB1 and UB6 being 15.4 and 64.09 milliseconds, respectively. Moreover, the difference between the average response time of the UB closest to the DC (UB1) and that farthest from the DC (UB6) is 150.04 milliseconds. By analyzing these time variations, the global DC positioning can be designed, and thus, uniformly distributed worldwide. The proximity of DC to the user and the peak time can also help us predict the cloud user usability patterns. In turn, this allows us to design DC locations and implement the conceptual model of the DC monitoring and distribution systems. These figures also show the ideal global distribution of DCs and UBs, which can make up a reasonable solution for efficient large-scale resource provisioning for their different response times. In turn, this can help us design a dynamic geographical location distribution of data centers and monitoring systems worldwide.

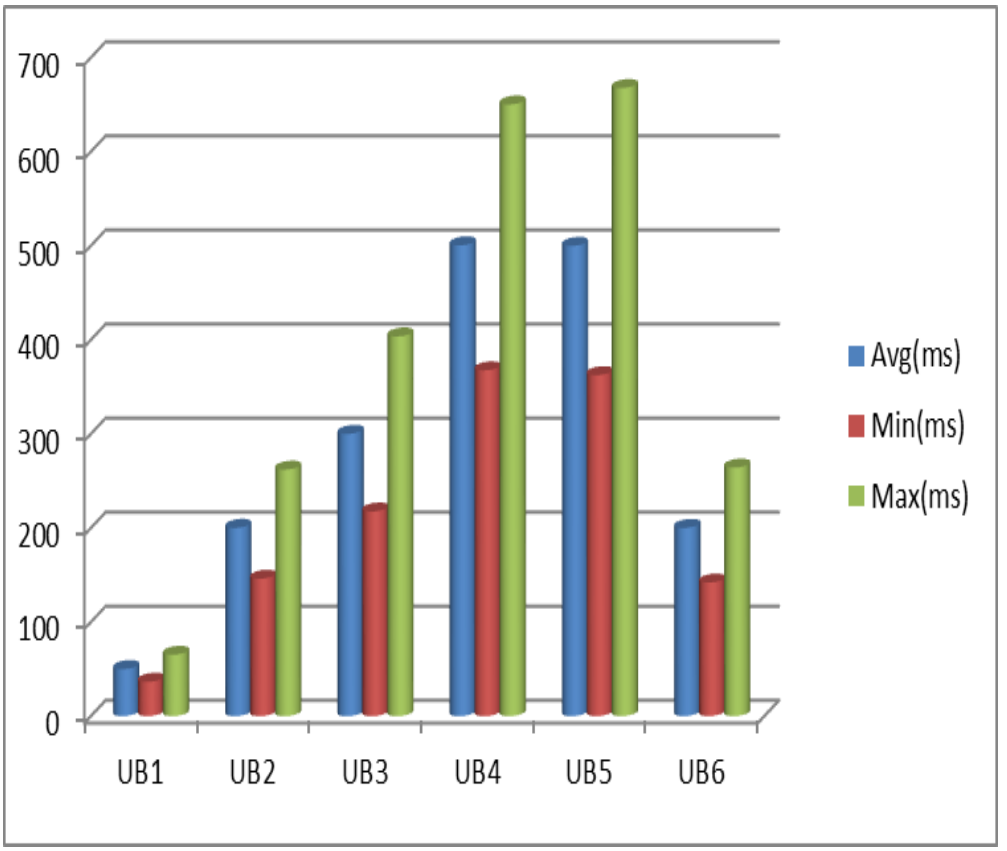


Figure 2 : Graphical presentation of UserBases to DataCenter Avgerage, Minimum and Maximum response time.

By checking the different response time graphs in the same place, as shown in Fig. 2, we can compare the time responses at the same place. As can be seen, the X-axis shows the different response times in milliseconds, the Y-axis, and the different UBs. The graph of UB to DC presented in Fig. 3 illustrates the average response time in milliseconds.

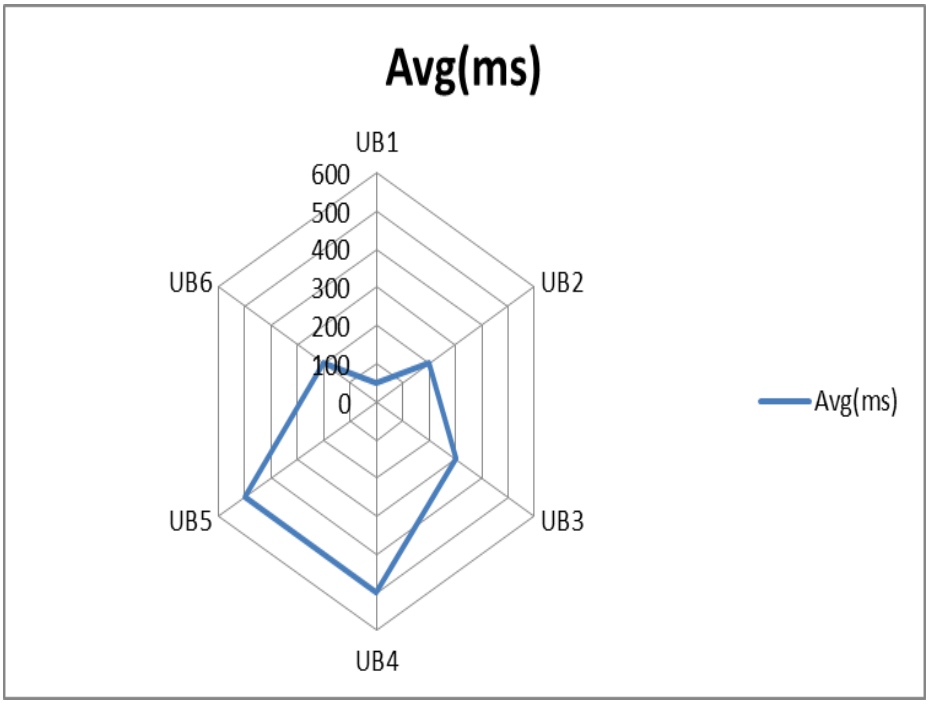


Figure 3 : Graphical presentation of UserBase to DataCenter Average response time.

4. Conclusion

At present, data centers have turned into a commercial infrastructure for storing data and hosting large-scale network applications via provisioning different services. However, traditional DCN architectures are not appropriate for multi-tenant data center environments. Data centers must provide uniform distributions worldwide according to the user traffic by employing deployable virtualization technology with a scalable design. In addition, the cloud computing applications should reduce the infrastructure cost, improve management flexibility, and reduce energy consumption. In this paper, breakthroughs in recent DCN basics have been reviewed, and the global distribution schemes of the cloud data center are also discussed. Moreover, the user response time and the location of UBs are considered to evaluate traffic pattern, congestion, and network performance. In addition, CloudAnalyst is used to show the difference in response times among different UBs and data center distributions, as well as their comparative effects worldwide. Through these results, worldwide data centers can be modeled and redistributed based on user service experience. Finally, this assessments can ensure the best utilization of traffic pattern, agility, availability and the service provisioning with QoS of data centers distribution across the globe.

References

1. D.-y. Xu, S.-l. Yang, and R.-p. Liu, "A mixture of HMM, GA, and Elman network for load prediction in cloud-oriented data centers," *Journal of Zhejiang University SCIENCE C,* vol. 14, pp. 845-858, 2013/11/01 2013.
2. S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, "Cloud-based augmentation for mobile devices: Motivation, taxonomies, and open challenges," *Communications Surveys & Tutorials, IEEE,* vol. PP, pp. 1-32, 2013.
3. M. F. Bari, R. Boutaba, R. Esteves, L. Z. Granville, M. Podlesny, M. G. Rabbani, *et al.*, "Data Center Network Virtualization: A Survey," *Communications Surveys & Tutorials, IEEE,* vol. 15, pp. 909-928, 2013.
4. S. Abolfazli, Z. Sanaei, A. Gani, F. Xia, and L. T. Yang, "Rich Mobile Applications: Genesis, taxonomy, and open issues," *Journal of Network and Computer Applications*.
5. Y. Cheng, Z.-y. Wang, J. Ma, J.-j. Wu, S.-z. Mei, and J.-c. Ren, "Efficient revocation in ciphertext-policy attribute-based encryption based cryptographic cloud storage," *Journal of Zhejiang University SCIENCE C,* vol. 14, pp. 85-97, 2013/02/01 2013.
6. X.-b. Li, Y.-l. Lei, H. Vangheluwe, W.-p. Wang, and Q. Li, "A multi-paradigm decision modeling framework for combat system effectiveness measurement based on domain-specific modeling," *Journal of Zhejiang University SCIENCE C,* vol. 14, pp. 311-331, 2013/05/01 2013.
7. M. Shiraz, A. Gani, R. H. Khokhar, and R. Buyya, "A Review on Distributed Application Processing Frameworks in Smart Mobile Devices for Mobile Cloud Computing," *Communications Surveys & Tutorials, IEEE,* vol. 15, pp. 1294-1313, 2013.
8. M.-w. Tang and X.-x. Wang, "Resource allocation algorithm with limited feedback for multicast single frequency networks," *Journal of Zhejiang University SCIENCE C,* vol. 13, pp. 146-154, 2012/02/01 2012.
9. M. Whaiduzzaman, A. Gani, N. B. Anuar, M. Shiraz, M. N. Haque, and I. T. Haque, "Cloud Service Selection using Multi-Criteria Decision Analysis," *The Scientific World Journal,* 2013.
10. Z. Sanaei, S. Abolfazli, A. Gani, and R. Buyya, "Heterogeneity in mobile cloud computing: Taxonomy and open challenges," *Communications Surveys & Tutorials, IEEE,* vol. PP, pp. 1-24, 2013.
11. D. Cheun, H. La, and S. Kim, "A taxonomic framework for autonomous service management in Service-Oriented Architecture," *Journal of Zhejiang University SCIENCE C,* vol. 13, pp. 339-354, 2012/05/01 2012.
12. J. Zhang, X.-j. Chen, J.-h. Li, and X. Li, "Task mapper and application-aware virtual machine scheduler oriented for parallel computing," *Journal of Zhejiang University SCIENCE C,* vol. 13, pp. 155-177, 2012/03/01 2012.
13. E. Ahmed, M. Shiraz, and A. Gani, "Spectrum-aware Distributed Channel Assignment for Cognitive Radio Wireless Mesh Networks," *Malaysian Journal of Computer Science,* vol. 26, 2013.

14. M. Whaiduzzaman, M. Sookhak, A. Gani, and R. Buyya, "A survey on vehicular cloud computing," *Journal of Network and Computer Applications*.

15. M. Shiraz, S. Abolfazli, Z. Sanaei, and A. Gani, "A study on virtual machine deployment for application outsourcing in mobile cloud computing," *The Journal of Supercomputing*, vol. 63, pp. 946-964, 2013.

16. R. Enayatifar, H. J. Sadaei, A. H. Abdullah, and A. Gani, "Imperialist competitive algorithm combined with refined high-order weighted fuzzy time series (RHWFTS–ICA) for short term load forecasting," *Energy Conversion and Management*, vol. 76, pp. 1104-1116, 12// 2013.

17. B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "CloudAnalyst: A CloudSim-Based Visual Modeller for Analysing Cloud Computing Environments and Applications," in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, 2010, pp. 446-452.

18. R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, pp. 23-50, 2011.

JJIT, Vol. **6**, June, 2017, 81-96

FUNDAMENTAL FREQUENCY DETECTION METHOD IN NOISY ENVIRONMENT

MIRZA A. F. M. RASHIDUL HASAN

Department of Information and Communication Engineering, Faculty of Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.

Abstract

In this article a new method for efficient fundamental frequency detection algorithm in noisy environments of a speech signal is proposed. The proposed method introduces a greatly enhanced correlation based algorithm which employs the autocorrelation function and YIN methods. In this proposed method, at first we introduced the autocorrelation function with original signal and then apply again autocorrelation function with its previous output signal. Finally this autocorrelation function is weighted by the reciprocal of the YIN 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 methods. A comprehensive evaluation of the fundamental frequency estimation, considering GPE, in all type speaker cases in white noise the proposed method gives better results than the other related methods in different types of signal to noise ratio (SNR) conditions. Especially at SNR=0dB and SNR=-5dB, the proposed method gives far better results than the other related methods.

Keywords: *Fundamental Frequency, Pitch, Autocorrelation Function, White Noise.*

1. Introduction

Fundamental frequency *i.e.*, pitch period is the key parameters of speech related research. In speech signal, there are two major categories- 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. [3,4,5,6,7]. At present there are so many fundamental frequency detection algorithms have been established, but significant and accurate fundamental frequency detection algorithm is still lacking.

Three categories of fundamental frequency detection algorithms (FFDAs) in the text: time honored [8,9,10], frequency honored [11,12], and time-frequency honored [13,14]. Due to the extreme importance of accurate fundamental frequency detection problem, the muscles of different FFDAs have been searched [15,16], and several fundamental frequency reference databases have been developed to facilitate fair comparison of different FFDAs

on a common platform [17]. Correlation based processing is known to be comparatively robust against noise. Among the reported method, the time honored method i.e., the autocorrelation function (ACF) [8] method is classified into correlation based approaches and are well accepted for their plainness and well performance in the presence of noise. Correlation based processing also includes the YIN method [18]. This article, we propose a new approach of fundamental frequency extraction method, which uses autocorrelation function weighted by the inverse of YIN method. The characteristics of the YIN are very similar with those of the ACF. The YIN produces a valley, while the autocorrelation produces a peak. However, Both functions essentially have the same periodicity. The proposed method utilizes the feature that in a noisy environment, the noise components included in the autocorrelation function and YIN behave independently (and are uncorrelated each other). This feature will be validated in this article. By such uncorrelated properties, the peak of the autocorrelation function is emphasized in a noisy environment when the autocorrelation function is combined with the inversed YIN. As a result, it is expected that the accuracy of fundamental frequency extraction for the ACF and also YIN is improved.

This article is organized as follows. In Section 2, we intensively discuss the problem of some conventional time honored methods. The proposed FFDA is presented in Section 3. In Section 4, we verify the effectiveness of our method by comparing with other methods based on experimental results and finally we conclude our work in Section 5.

2. Fundamental Frequency Detection Algorithms (FFDAs)

Fundamental frequency is an auditory perceptual property that allows the ordering of sounds on frequency domain. Most of the time honored fundamental frequency period estimation methods use ACF.

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

$$n(m) = s(m) + w(m) \tag{1}$$

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

$$R_{nn}(\tau) = \begin{cases} R_{SS}(\tau) + \sigma_V^2 & for\ \tau = 0, \\ R_{SS}(\tau) & for\ \tau \neq 0, \end{cases} \tag{2}$$

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

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

where 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 ACF 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 ACF 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 conventional ACF method is significantly degraded at low SNR which is shown in Fig. 2. On the other hand, YIN is based on the difference function, which attempts to minimize the difference between the waveform and its delayed duplicate instead of minimizing the product. The difference function is defined as

$$h_{ss}(\tau) = \sum_{m=0}^{M-1} \left| s(m) - s(m+\tau) \right|^2 \qquad (4)$$

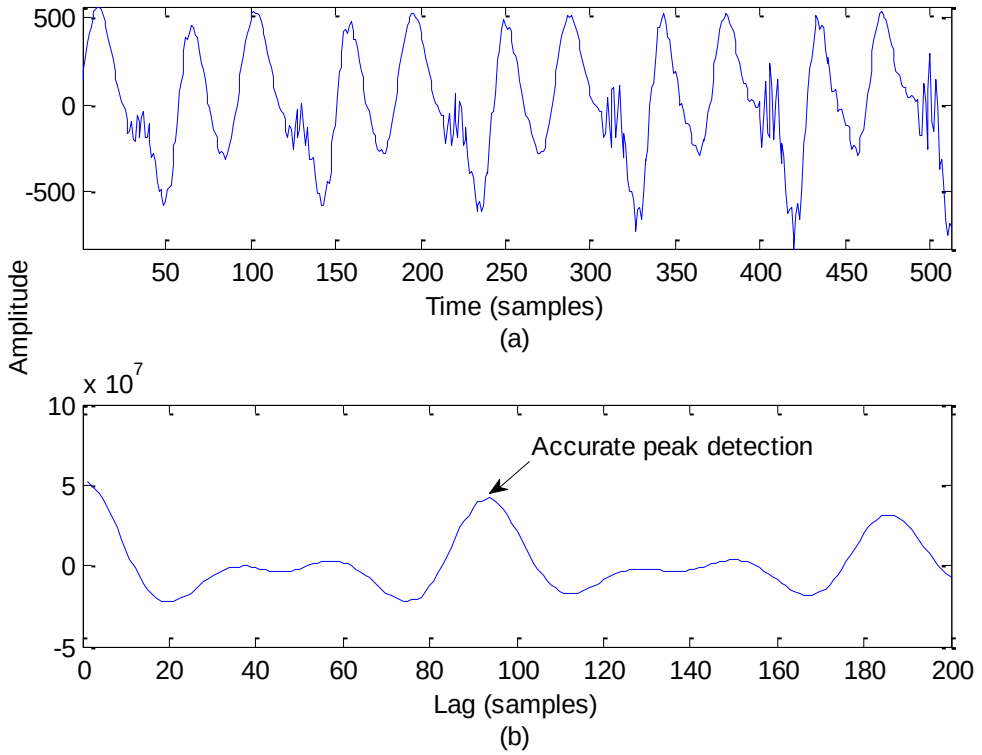


Fig. 1: (a) Clean speech signal of a female speaker, (b) Autocorrelation function of signal in (a).

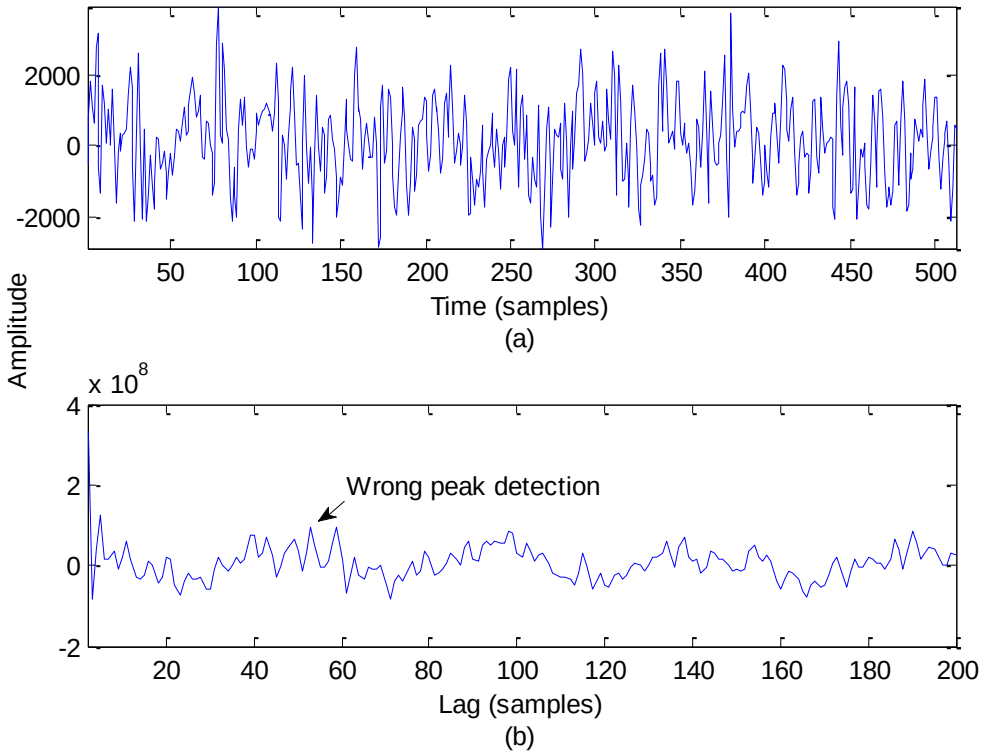


Fig. 2 : (a) Noisy speech signal of a female speaker (which is the same frame as Fig. 1(a)) at SNR of -5dB, (b) Autocorrelation function of signal in (a).

The YIN has the characteristic that when $s(m)$ is similar with $s(m+\tau)$, $h_{ss}(\tau)$ becomes small. This means that if $s(m)$ has a period of T, $h_{ss}(\tau)$ produces a deep valley at $\tau = T$. Therefore, $1/h_{ss}(\tau)$ makes a peak at $\tau = T$. The fundamental frequency period is identified as the value of the lag at which the minimum YIN occurs (Fig. 3). This algorithm has many advantages, but the probability of double misjudge and half misjudge is very high when noise is added (Fig. 4).

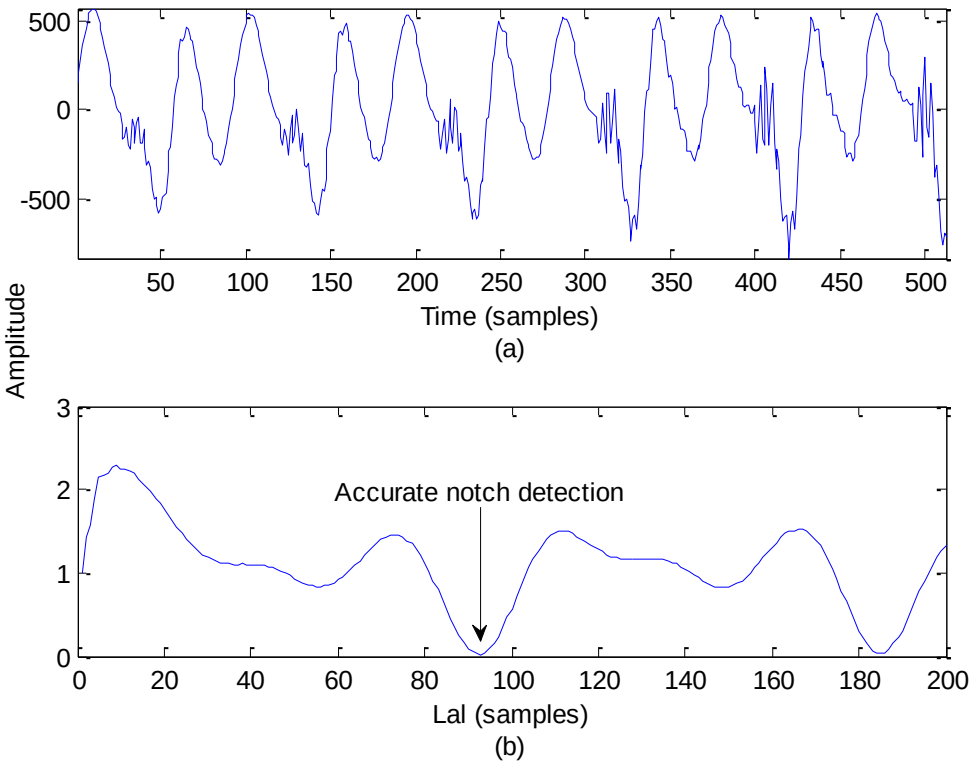


Fig. 3: (a) Clean speech signal of a female speaker (which is the same frame as Fig. 1(a)), (b) YIN of signal in (a).

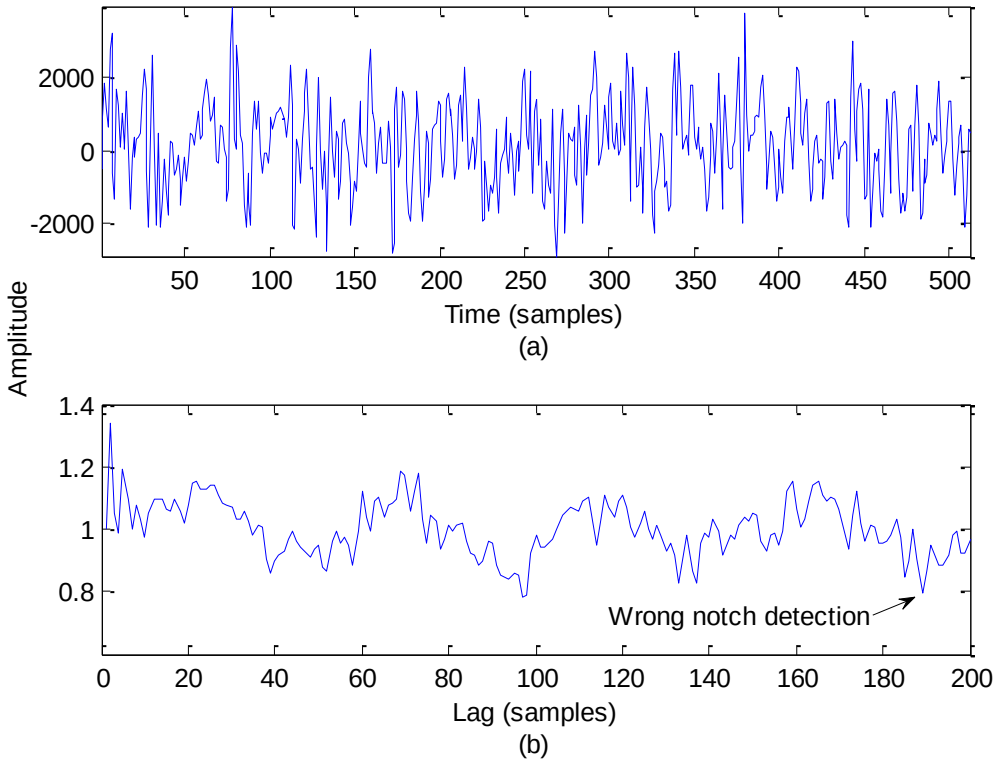


Fig. 4 : (a) Noisy speech signal of a female speaker (which is the same frame as Fig. 1(a)) at SNR of -5dB, (b) YIN of signal in (a).

3. Proposed Method

The ACF is weighted by the inverse of an Average Magnitude Difference Function (AMDF) [19] is used for fundamental frequency extraction [20] and is defined as

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

where $R_{ss}(\tau)$ and $\xi_{ss}(\tau)$ denotes the ACF and AMDF of signal $s(m)$ respectively, l is a small positive constant. It is expected to give maximum peak at $\tau = mT$ (ACF) & deep notches at $\tau = mT$ (AMDF), and therefore the true fundamental frequency peak in $\varphi_{ss}(\tau)$ is emphasized (Fig. 5). This methods provides a good performance in the presence of clean signal (Fig. 6). 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. 7. To overcome this situation,in our proposed method, firstly we introduced the autocorrelation function with original signal and then apply again autocorrelation function with its previous output signal (Fig. 8). In figure 8 implies that the

output signal is smoother and more prominent then previous one. Finally the output of autocorrelation function is weighted by the reciprocal of the YIN in our proposed method and is define as

$$p_{SS}(\tau) = \frac{R_{SS}(\tau)}{h_{SS}(\tau) + l} \tag{6}$$

Where $R_{ss}(\tau)$ and $h_{ss}(\tau)$denotes the ACF and YIN of signal $s(m)$ respectively. It is expected that the true peak is more emphasized and as a result the accurate fundamental frequency is detected i.e. the errors are decreased (Fig. 9).

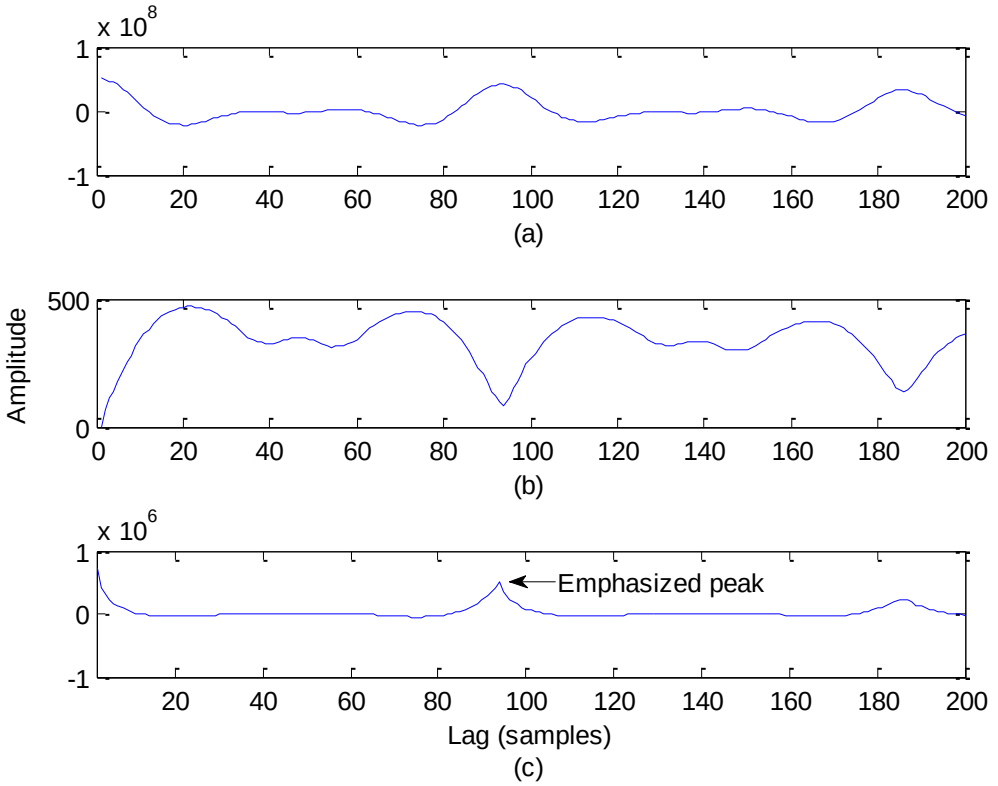


Fig. 5 : Pitch peak detection using (a) Autocorrelation function method, (b) Average magnitude difference function method, (c) Weighted autocorrelation function method.

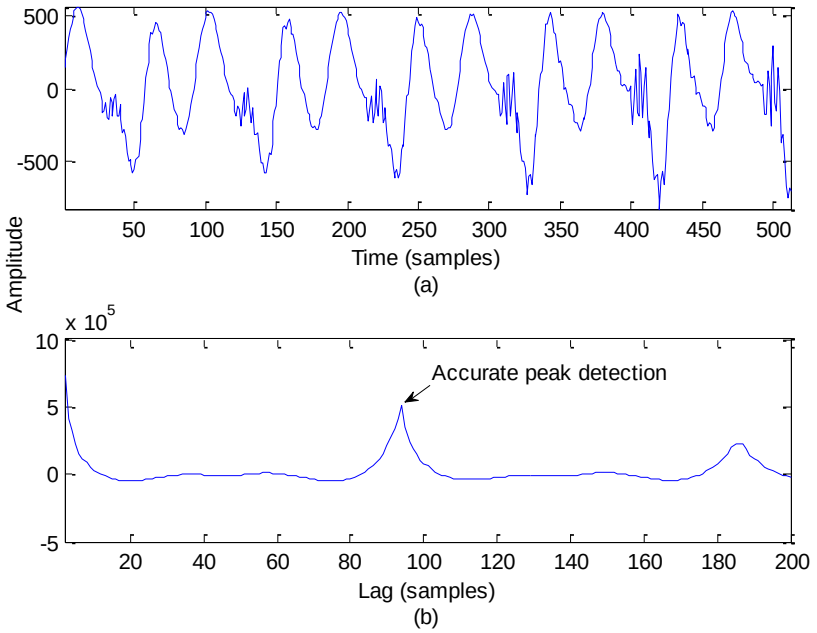


Fig. 6 : (a) Clean speech signal of a female speaker (which is the same frame as Fig. 1(a)), (b) Pitch peak detection using Weighted autocorrelation function method.

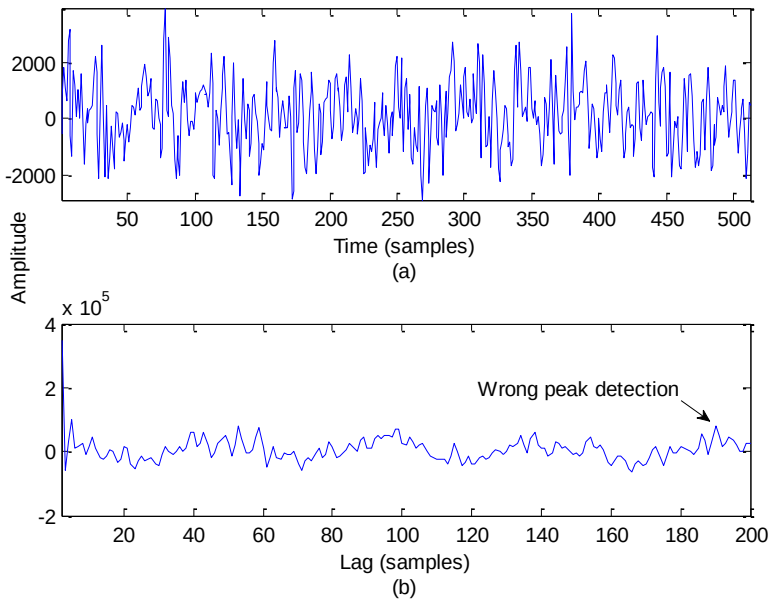


Fig. 7 : (a) Noisy speech signal of a female speaker (which is the same frame as Fig. 1(a)) at SNR of -5dB, (b) Pitch peak detection using Weighted autocorrelation function method.

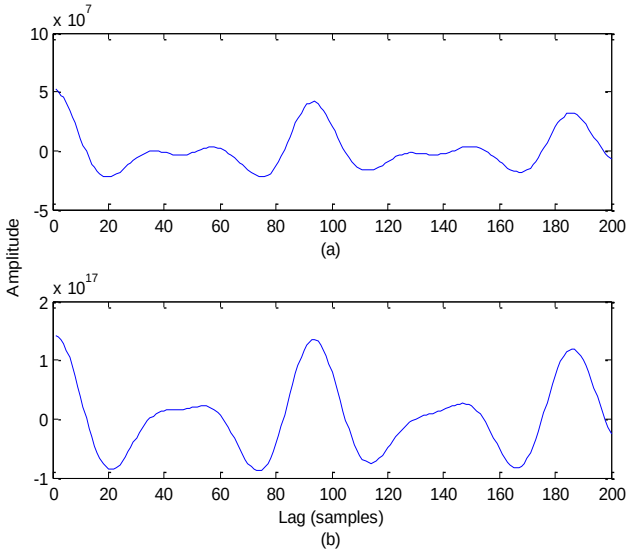


Fig. 8 : (a) Autocorrelation function of speech signal, (b) Autocorrelation function of signal in (a).

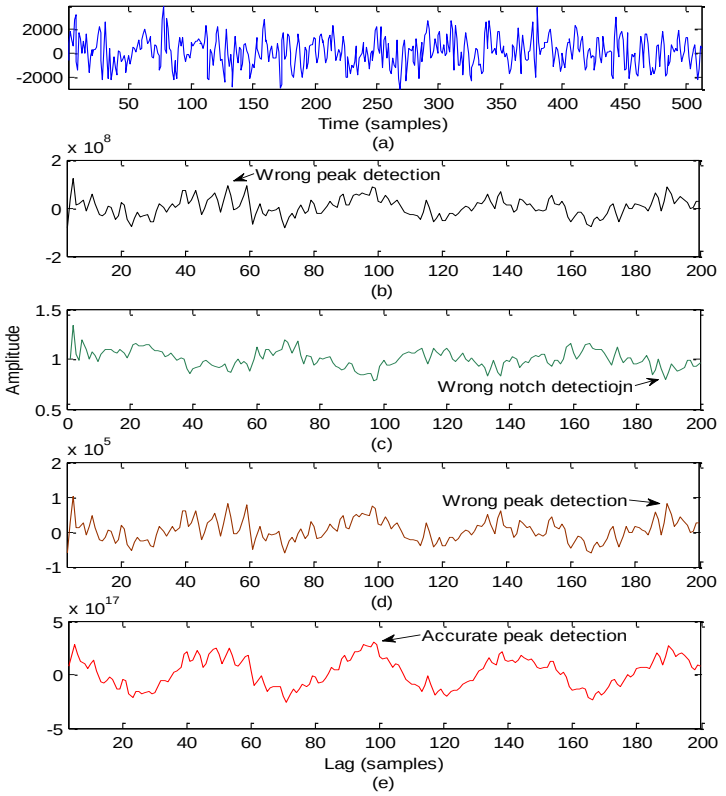


Fig. 9 : Pitch peak detection in noisy speech signal of a female speaker (which is the same frame as Fig. 1(a)) at SNR of -5dB using (b) Autocorrelation function method, (c) YIN, (d) Weighted autocorrelation function method, and (e) Proposed method.

4. Experimental Results

To evaluate the proposed method, there are two Japanese female and male speakers 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 [21]. 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. For performance evaluation, we 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

$$e(q) = F_t(q) - F_e(q) \tag{7}$$

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) [18,22]. 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

$$GPE(\%) = \frac{F_{GPE}}{F_v} \times 100 \tag{8}$$

where 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

$$FPE_{mean} = \frac{1}{M_i} \sum_{j=1}^{M_i} e(q_j) \tag{9}$$

where 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 attempt 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 ACF, YIN, and weighted autocorrelation method, WACF [8,18,20]. In WACF and proposed method the constraint τ is set to 0.5. 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 ACF,YIN,WACF and the proposed

method. In order to assess the pitch assessment performance of the proposed method, we sketch a reference pitch contour for SNR -5dB 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.

Fig. 10 indicates that in compare to the different method, the proposed method performs a better smoother pitch contour even at an SNR of -5dB. Fig. 11 indicates a judgment of the pitch contour considering the male speech corrupted by the white noise at an SNR of -5dB. 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 four 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.

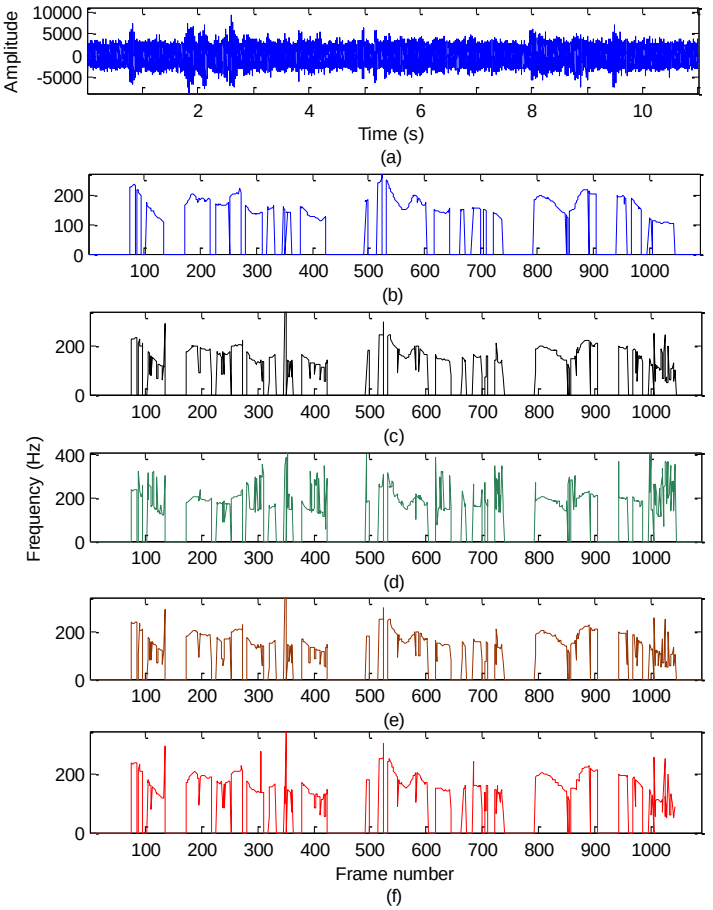


Fig. 10 : (a) Noisy speech signal for female speaker in white noise at an SNR -5dB, (b) True pitch of signal (a), Pitch contours extracted by (c) ACF (d) YIN, (e) WACF, and (f) Proposed method.

These simulation outcomes indicate that the proposed method is better performs to the ACF, YIN, and WACF method in most of the cases. The proposed method performs more strongly compared with the ACF, YIN, and WACF method at low SNR (0dB, -5dB).

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 four methods are sketch which is shown in Fig. 13. Average FPEs for all methods range around from 1.5 Hz ~ 7Hz. Fig. 13, implies that the FPE values resulting from the proposed method are not affect the SNR condition i.e. FPE values almost constant but the ACF, YIN, and WACF method present relatively higher values of FPE when noise is increased.. From the experimental results it is establish that the range of FPEs is also within the satisfactory limit and consistently satisfactory at other SNRs.

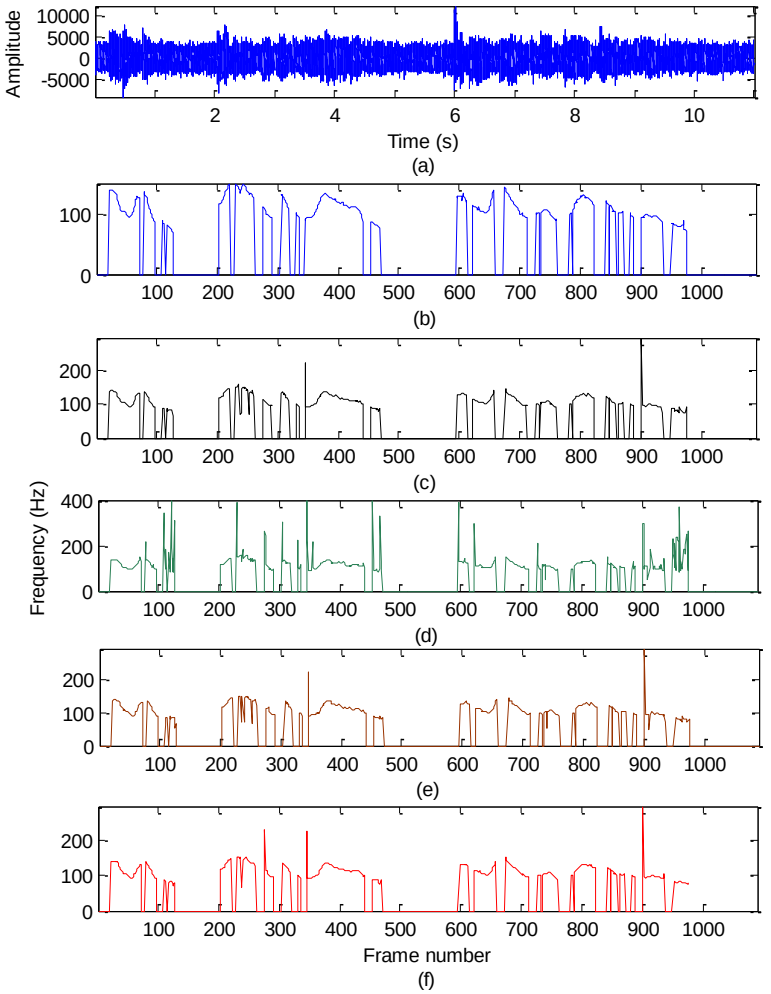


Fig. 11: (a) Noisy speech signal for male speaker in white noise at an SNR -5dB, (b) True pitch of signal (a), Pitch contours extracted by (c) ACF, (d) YIN, (e) WACF, and (f) Proposed method.

5. Conclusion

Perfect fundamental frequency estimation is a tricky problem in speech analysis particularly in noisy environments. In this article, we proposed a correlation based method by utilizing the autocorrelation function is weighted by the reciprocal of the YIN. 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 ACF, YIN, and WACF for a wide range of signal to noise ratio varying from -5dB to 20dB and clean. Considering GPE, in all type speaker cases the proposed method gives better results than the other methods in different types of SNR conditions. Especially at SNR=0dB and SNR=-5dB, the proposed method gives far better results than the other method. 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 candidate for extracting fundamental frequency information in white noise environments with very low levels of SNR as compared with other allied methods. Color noises are found very difficult to handle for fundamental frequency determination compared to white noise. In future we will extend our research to develop new fundamental frequency estimation methods which will be particularly robust against color noise.

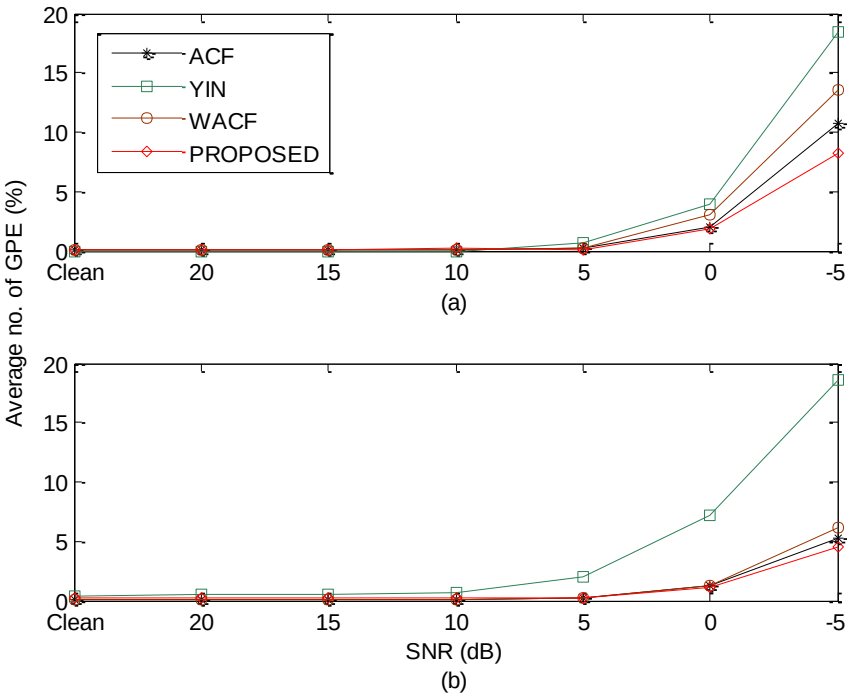


Fig. 12 : Percentage of average gross pitch error (GPE) in white noise for different speakers under various SNR conditions; (a) Female speakers, (b) Male speakers.

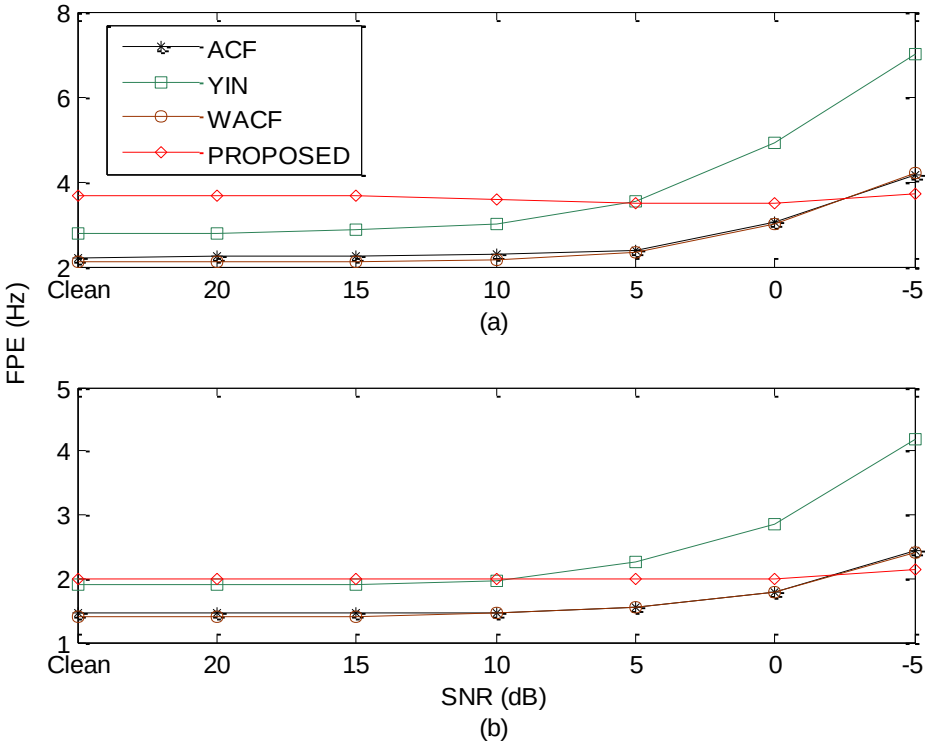


Fig. 13 : Comparison of average performance results in terms of mean fine pitch error (FPE) for different speakers under various SNR conditions; (a) Female speakers, (b) Male speakers.

References

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

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

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

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

5. M. Tamura, T. Masuko, K. Takuda, and T. Kobayashi, "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. A. A. Razak, M. I. Z. Abidin, and R. Komiya, “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. A.S.M.M. Jameel, S. A. Fattah, R. Goswami, W. P. Zhu, and M. O. Ahmad, “Noise robust formant frequency estimation method based on spectral model of repeated autocorrelation of speech,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1357-1370, 2017.

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

9. M. A. F. M. R. Hasan, and T. Shimamura, “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.

10. M. A. F. M. R. Hasan “Correlation based fundamental frequency extraction method in noisy speech signal,” *International Journal of Computer Science, Engineering and Information Technology*, vol. 7, no. 1, pp. 1-12, 2017.

11. A. M. Noll, “Cepstrum pitch determination,” *Journal of Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293-309, 1967.

12. S. Ahmadi, and A. S.Spanias, “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.

13. 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, pp. 231-239, 2012.

14. M. A. F. M. R. Hasan, “A pitch detection algorithm based on windowless autocorrelation function and modified cepstrum method in noisy environment,” *International Journal of Computer Science and Network Security*, vol. 17, no. 2, pp. 106-112, 2017.

15. 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*, vol. ASSP-24, no. 5, pp. 399-417, 1976.

16. P. Veprek, M. S. Scordilis, “Analysis, enhancement and evaluation of five pitch determination techniques,” *Speech Communication*, vol. 37, pp. 249-270, 2002.

17.] F. Plante, G. Meyer, and W. A. Ainsworth, “A pitch extraction reference database”, In Proc. EUROSPEECH, pp. 837-840, 1995.

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

19. 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*, vol. ASSP-22, no. 5, pp. 353-362, 1974.

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

21. NTT, "Multilingual Speech Database for Telephometry," NTT Advance Technology Corp., Japan, 1994.

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

JJIT, Vol. **6**, June, 2017, 97-108

BANGLA SPELL CHECKER: A DISTANCE AND PRIOR PROBABILITY BASED APPROACH

MUNTASIR WAHED[1], M M ABID NAZIRI[1], MOHAMMAD SHOYAIB[2] AND MUHAMMAD ASIF HOSSAIN KHAN[1]

[1]Department of Computer Science and Engineering University of Dhaka, Dhaka, Bangladesh
[2]Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

Abstract

A spell checker is a tool that can identify a misspelled word in a document. A working spell checker that is highly accurate and can offer the desired correction within the first few positions of the suggestion list is a necessary tool for any language. However, very fewworks have been done for the development of a fully functional Bangla spell checker with all of the aforementioned qualities. Most of the existing works use distance (edit-distance/phonetic-distance/both) based methods for generating spelling suggestions. However, in many cases, several candidate correction words remain at the same distance from the wrong word, and some of these candidates must be discarded as only a limited number of corrections can be accommodated in a suggestion list. If the plucking is done at random, it is highly possible that the actual correction word will be purged in this process. In this paper, we focus on improving the overall accuracy of the proposed spell checker while striving to put the actual correction word at the top of the suggestion list. For this purpose, we propose two methods: PEP and ERPP, both of which use two types ofdistances and prior probability for generating the results. Experimental results show that one of the proposed methods can achieve as high as 97.62% accuracy while putting the accurate suggestion in the first position of the suggested list in 80.19% of the cases. It is noteworthy to mention that the proposed method is about seven times faster than the state-of-the-art method considered in the experiments.

Index terms - *Edit distance, Phonetic distance, Prior probability, Bangla spell checker.*

I. Introduction

A spell checker detects misspelled words and gives suggestions to correct the error(s). Instead of suggesting a single correct word, it often suggests a list of possible corrections to the user. The user then selects the appropriate word from the list. Assuming that the correct word is in the suggestion list, finding the desired word within the top positions of the suggested list is one of the most desirable criteria from a user perspective, as it saves time and effort of finding the actual correction. Again, offering these suggestions with little or no delay is also demanding.

Hundreds of spell checker algorithms have been proposed for many different languages. Various methods have been adopted for constructing these spell checkers. Some of these methods are language independent; i.e. can be adopted for any language, whereas others exploit features specific to a particular language. These methods can be broadly classified into several categories such as edit distance based methods [1] [2], phonological encoding based methods [3] [4] [5], stemming based methods [6] [7], n-gram based methods [8] [9] [10] etc. A combination of these methods has also been used by researchers [11] [12].

Bangla is the state language of Bangladesh. It is also the second most spoken language in India. More than 250 million people all over the world speak in Bangla. It is the seventh largest spoken language in the world. Though Bangla is a sweet spoken language, its script is quite complicated, which makes it difficult to write properly. The language has 50 letters in the alphabet, of which 11 are vowels and 39 are consonants. There are no upper or lower case letters. Some of the challenges imposed by the set of rules of Bangla language are: Conjuncts or *Juktakkhor*, example: ণ্ট, জ্ঞ, *Phala*, example: *ma-phala*(গ্ম), *ba-phala*(দ্বা), *ra-phala*(দ্র), *Kar*, example: *a-kar*(া), *u-kar* ু(), *Matra*, example: ভাল, খারাপ, Conjuncts with counter-intuitive pronunciations, such as ক্ষ = ক + ষ is pronounced as কখ instead of কষ, Phonetic similarities, example: স(s), শ(sh) and ষ(Sh), Context based pronunciation, example: ক্ষমা is spelled as *khoma*, where রক্ষা is spelled as *rokkha*, Modifiers: such as □, ং, ঃ [13] etc.

There are different kinds of spelling errors. They can be classified into two broad categories: non-word error and real-word error [14]. If a misspelled word is not a valid word of that language, i.e. not found in any dictionaries, it's known as a non-word error. For example:ঘড়. On the other hand, if a word is a valid word according to the dictionary, but not appropriate for the context, it's known as a real-word error. For example, আমবিইপরছি - here পরছি is not an incorrect word considering the dictionary, but in this context, the correct spelling would be পড়ছি, since the first one means "wearing" and the latter one means "reading", and we can read a book, but can't wear it. The Error Pattern Analysis is also described in [15] with further details. In this paper, we mainly focus on developing a spell checker for non-word error.

To date, many spell checkers have been developed for different languages. Analysis of these spell checking mechanisms reveal that they can be broadly classified into five categories:

i. Typological error checking methods (ex. Edit Distance)
ii. Phonological error checking methods (ex. Soundex)
iii. Stemming methods
iv. n-gram based methods
v. Hybrid method: combination of two or more of these and others.

Although many industry-grade spell checkers are available for English, such robust spell checker is yet to be developed for Bangla [13]. Only a handful of works have been published in the field of Bangla spell checker. Abdullah *et al.* proposed a spell checking mechanism which falls into the first category [1]. They used an ad-hoc approach to detect and correct spelling mistakes focusing on typological error. Besides this, few works are found based on Phonological error checking (second category). B. B Chaudhuri [16] proposed a method that involves a reversed dictionary and grouping based on phonetic similarities. A Soundex based spell checker with some modifications was proposed in [3]. However, their implementation had some limitations. For example, this method is unable to detect vowels accurately and *phalas* and conjuncts are not handled properly. Hence, they further modified the algorithm for phonetic errors [11]. A word based spell checker that uses stemming approach was proposed in [7]. In 2009, a language independent spell checker was presented in [10] that used an enhanced n-gram model, which includes the usage of n-gram statistics and lexical resources. Khan *et al.* [17] also usedan n-gram model on Bangla. They calculated the unigram, bigram, trigram, and quadgram respectively and saved them on discs to use the n-gram method. Inclusion of n-gram statistics also incorporates the context information which usually improves the overall accuracy. However, in this paper we mainly focus on the development of the basic performance of a spell checker without any context information.Two or more of the aforementioned methods are combined in hybrid approaches. A clustering based approach proposed in [18] falls in this category. They divide the dictionary based on phonetic similarities and find medoids for each cluster using a combination of edit distance and phonetic distance. They match the incorrect word to the clusters' medoids and choose the clusters which have the minimum distance. Words from those clusters are then selected when they satisfy a particular threshold. Some other methods are also adopted in various research works, such as the one done by Abdullah et al. [19]. They generate a finite state machine to recognize only a specific set of strings, which is a valid set of suggestions.

All of the aforementioned methods mainly focus on improving the accuracy of their list of suggested words. However, from the user perspective, there are other important metrics to assess the performance of a spell checker. For example, a user would like to see the intended correct word at the top of the suggested list. Hence, the position of the desired correction in the list should be considered as an important performance measure for a spell checker. Moreover, the speed at which the list of suggested corrections is constructed should also be considered as a performance metric. To the best of our knowledge, none of the aforementioned methods considered all these performance metrics.

In this paper, we propose a Bangla spell checker that takes these three metrics into consideration. The proposed algorithm increases the probability of the correct word occurring at the start of the suggestion list; while improving the accuracy at the same time. We propose a second algorithm that reduces the search space, thus significantly reducing the time required for generating the suggested corrections while at the same time preserving the overall accuracy and providing correct guesses at the top of the list.

The rest of the paper is organized as follows: section II introduces the proposed methods. After that, section III presents experimental results and discussions. Finally, section IV concludes the paper.

II. Proposed Method

To generate better Bangla spelling suggestion, we propose two methods namely Phonetic and Edit Distance with Prior Probability (PEP) and Edit Distance Refined with Phonetic Distance and Prior Probability (ERPP) in this paper. The first one (PEP) improves the overall prediction and is more accurate in putting the desired correction in the first position compared to the state-of-the-art methods. ERPP further improves the performance by generating the suggestion list at a much reduced time. Both of these methods use Edit Distance (*ed*), Phonetic Distance (*pd*) and Prior Probability (*pp*). We first discuss on Prior Probability calculation and then PEP, followed by ERPP.

Calculation of Prior Probability

In most of the cases, there are many possible correct words for a given misspelled word. For example, if a user typed 'জম' it is not possible to be sure whether the user wanted to type 'জমা', 'জমা' or 'জমে'. The appropriate word can be chosen if the probability of a correct word (c) given a wrong word (w),$P(c|w)$can be calculated. In that case, among all the possible correct words at a minimum distance from the wrong word, we could simply choose the word that have the maximumprobability$P(c|w)$. This posterior probability can be calculated using the standard Bayesian Theorem given in equation (1),

$$P(c|w) = P(c)\frac{P(w|c)}{P(w)} \quad (1)$$

Where, $P(w|c)$is the likelihood, $P(c)$is the prior and $P(w)$ is the normalization constant that is same for every candidate c.Ignoring$P(w)$, equation (1) becomes

$$P(c|w) \propto P(c)\,P(w|c) \quad (2)$$

But it is not possible to calculate $P(w|c)$ since it is very hard to obtain enough data regarding which wrong word w the user typedwhile trying to type the given correct word c. Thus, we have no other choice but to use only the prior probability $P(c)$is to approximate $P(c|w)$. It is natural that such an approximation is very crude and might give very poor results. However, in addition to some other metrics (described below), $P(c)$ can be used for generating better results. Furthermore, to prevent the probability of the rare words being zero, we perform Laplace Smoothing.

Phonetic and Edit Distance with Prior Probability (PEP)

For generating spelling suggestion we first propose Phonetic and Edit Distance with Prior Probability (PEP) that first takes an input word and matches it with the dictionary. If the word does not exist in the dictionary, it is detected as a misspelled word. PEP then finds a particular phonetic similar group $(PSG) \in PG$ corresponding to the word. For getting faster performance, the whole dictionary is first divided into phonetic similar groups $PG = \{pg_1, pg_2, \ldots pg_n\}$ based on the phonetic representation of the first letter of a word [18]. For example, all the words starting with 'ক' and 'খ' are mapped to the same group, and so on. For this, we define a function namely (wrong word W) , which maps the word to its respective PSG based on its first letter. After that, each word in the same PSG as the misspelled word is checked against the misspelled word. Edit distance is a measure that quantifies the amount of dissimilarity between two strings; i.e. the minimum number of operations required to transform the one string to the other, where the operations might be insertion, deletion or substitution of one character in the string [2].

The distance (d_1) between a misspelled word and the word from its phonetic similar group is calculated using equation (3)

$$d_1 = \alpha \times ed(w_1, w_2) + \beta \times ed(DM_1, DM_2) \quad (3)$$

where, DM_i= Double metaphone encoding for w_i , w_1 is a word from the PSG, w_2 is the given wrong word, α and β are two user defined parameters ($\alpha + \beta = 1$). The function *ed* returns the edit distance between its two parameters.

Algorithm 1: Generating Suggestions

Input: Wrong word *W*, δ and Phonetic Similar Group $PG \leftarrow \{pg_1, pg_2, \ldots pg_n\}$

Output: List of Possible Suggestions $PS \leftarrow \{ps_1, ps_2, \ldots ps_n\}$

Begin
  $PS \leftarrow \varphi$
  $minEd \leftarrow \infty$
  $PSG \leftarrow ps(W)$
  for each $w_i \in PSG$
    $d_1 \leftarrow distance\ between\ W\ and\ w_i$ $(using\ equation\ (3))$
    $minEd \leftarrow min(d_1, minEd)$
  End for
  $\tau \leftarrow minEd + \delta$
  for each $w_i \in PSG$
    $d_1 \leftarrow distance\ between\ W\ and\ w_i$ $(using\ equation\ (3))$
    if $(d_1 \leq \tau)$
    then
      $PS \leftarrow PS \cup w_i$
    endif

16. $sort(PS)$ /*sorts PS based on d_1. Prior probability is used to break ties.*/
17. **return** top *k* elements of *PS*
18. **End**

Equation (3) results in several words with different d_1 for a particular wrong word *W*. However, we are interested to get a short list of suggestions. To achieve this we define a threshold τ and discard the words with $d_1 > \tau$. Here, τ is chosen empirically such that the correct word belongs to the list. The definition of τ is given in equation (4)

$$\tau = \min(d_1) + \delta \quad (4)$$

where δ is a user defined constant. By changing δ, it is possible to control the size of the initial list of possible suggestions. If a candidate word c_i has less distance from a wrong word *W* compared to c_j, then c_i has a higher probability of being chosen compared to c_j. When d_1 is same for more than one candidate words, we consider prior probability to break the ties. Finally, the first *k* words are returned. The whole procedure is shown in Algorithm 1. However, this method looks for words in a large search space and the calculation of phonetic distance for each of word is very costly. Hence, we modify PEP and propose our second method.

Edit Distance Refined with Phonetic Distance and Prior Probability (ERPP)

To reduce the search space and get an improved suggestion list we modify PEP and propose Edit Distance Refined with Phonetic Distance and Prior Probability (ERPP). In PEP, the initial list is generated using equation (3), whereas, in ERPP, the initial list is generated using equation (5), which does not use the phonetic distance.

$$d_2 = ed(w_1, w_2) \quad (5)$$

The initial list is made based on edit distance only (using equation (5)), alleviating the need for performing phonetic encoding for each word in *PSG*. Thus the time to generate the initial list is reduced notably. Once the initial list is obtained, d_1 is calculated for each of the selected words using equation (3). Then the words are sorted according to their d_1 values and the first *k* words are returned. Let us consider Example 1 for a better illustration of ERPP.

Example 1

Suppose a user wanted to type ‘অঙ্গ’, but instead, he typed ‘অংগ'. The program tries to find the existence of‘অংগ’ in the dictionary but fails and thus declares it as a misspelled word. ‘অংগ’ belongs to the phonetic group ‘অ’, so the method calculates its distance d_2 with each of the words of that group. The words that satisfy the threshold τ is then obtained as a primary list of 30 words such as অংশ, অংশী, অংশু, অংস, অগম, অগা, অঘর, অঙ্গজ, অঙ্গদ, অঙ্গী, অনড়, অনুগ, অঙ্গ and so on. For each word, d_1 is calculated using

equation (3). Now the program sorts the list based on $d_1$and the prior probabilities of the words and presents a list of three words shown in Table 1. As mentioned in Algorithm 1, the *sort* function uses the prior probabilities of two words to break the tie when they have the samed_1. For this particular example, we consider $\alpha = \beta = 0.5$ and $k= 3$.

Table-1: Suggestion list for the misspelled word 'অংগ'

Word	*ed* with অংগ	*pd* with অংগ	d_1	Prior Probability
অঙ্গ	2	0	1	0.000282666
অংশ	1	1	1	0.000211999
অংস	1	1	1	8.8333e-06

Note that, all of three words had the same $d_1$from অংগ, but অঙ্গwas moved to the top of the list because its prior probability is higher.

III. Experimental Results and Discussion

In this section, we first introduce the datasets that we use for evaluation purpose followed by the description of the evaluation metrics. Finally, we present experimental results along with necessary discussions.

A. Dataset Description

To validate our proposed algorithms we use three datasets. The first dataset was taken from the book [20], referred to as the "Book Dataset". It is a list of commonly occurring mistakes which are accumulated for educational purposes as it is a reference grammar book for Bangla. 914 out of 980 words are taken from here, leaving cases where a spelling mistake or a suggestion is a phrase containing more than one word. The second dataset, obtained from Wikipedia [21], was created for research purposes in the field of Bangla language processing and all 631 words were considered. We refer to this dataset as the "Wiki Dataset". The third dataset was taken from a blog [22] and is referred to as the "SM Dataset". It is a blog post, about commonly mistaken words and their correct forms. There are 323 words in this dataset, but we took 255 out of them for the same reason as the Book Dataset. We also combined all the three datasets and made a combined datasets referred to as the "Combined Dataset" consisting 1800 words.

For calculating the prior probabilities we use a dataset obtained from crawling news websites (ex- ProthomAlo), literature (ex- *RabindraRochonaboli*) and others which resulted in around 4.2 million words, with 127,000 unique words.

B. Implementation Details

For evaluating the results of different methods we use four different metrics namely accuracy (%), first guess rate (FGR), first three guess rate (FTGR) and the average time required (in millisecond) for generating the suggestions for a misspelled word. The definition of these metrics are given in equation (6), equation (7) and equation (8)

$$Accuracy = \frac{Number\ of\ words\ for\ which\ target\ words\ were\ present\ in\ suggestion\ list}{Number\ of\ misspelled\ words} x100\% \quad (6)$$

$$FGR = \frac{Number\ of\ times\ the\ correct\ suggestion\ was\ at\ the first\ position\ of\ suggestion\ list}{Numberofmisspelledwords} x100\% \quad (7)$$

$$FTGR = \frac{Number\ of\ times\ the\ correct\ suggestion\ was\ with\ in\ the\ first\ three\ positions\ of\ suggestion\ list}{Numberofmisspelledwords} x \quad (8)$$

C. Result and Analysis

Using the three equations (6)-(8) and the datasets mentioned above, the performances of PEP and ERPP have been shown in Table 2 - 5 for Book Dataset, Wiki Dataset, SM Dataset and for the Combined Dataset respectively. We also compare our methods with the state-of-the-art method described in [18] and the method is referred to as PMM (P. Mandal's Method) here.

Table-2: Performance Analysis using Book Dataset

Method	Accuracy (%)	FGR (%)	FTGR (%)	Time/Word (ms)
PMM	94.20	68.93	87.96	665
PEP	95.08	77.35	92.69	705
ERPP	96.50	77.24	93.33	85

Table-3: Performance Analysis using Wiki Dataset

Method	Accuracy (%)	FGR (%)	FTGR (%)	Time/Word (ms)
PMM	95.88	72.90	90.33	845
PEP	96.83	80.35	94.93	903
ERPP	97.62	80.19	95.40	129

Table-4: Performance Analysis using SM Dataset

Method	Accuracy (%)	FGR (%)	FTGR (%)	Time/Word (ms)
PMM	90.20	62.75	84.71	713
PEP	90.59	67.45	89.02	733
ERPP	94.12	67.45	91.76	91

Table-5: Performance Analysis usingthe Combined Dataset

Method	Accuracy (%)	FGR (%)	FTGR (%)	Time/Word (ms)
PMM	94.22	69.67	88.56	747
PEP	95.06	77.00	93.06	813
ERPP	96.56	76.89	93.83	97

Analyzing Table 2 – 5, it can be observed that our methods outperform PMM [18] in terms of accuracy for all the datasets. This is because there might exist more than *k* words in the respective PSG having the samed_1. PMM took top *k* words from this list for suggestion without any further processing. For this reason, a more likely word can be left out of the top *k* suggestions even after having the same distance from the given word, whereas our proposed methods sort the list based on the prior probabilities to keep the more likely words on top, resulting in a more accurate output.

In Table 2 - 5, First Guess Rate (FGR) means getting the accurate word in the first position of the suggestion list, where again our results are better due to the use of prior probabilities. In addition, First Three Guess Rate (FTGR) mean getting the results within the first three positions of the suggestion list, where our proposed methods perform better for the same reason.

Among the three datasets, the result of Wiki Dataset is the best for all the methods. This is because the other two datasets mainly focus on the grammatical mistakes in day-to-day usage, which contains some words that are grammatically wrong. But the Wiki Dataset contains words that are generally misspelled because of typical errors, not grammatical rules. For example, let us consider the word "জগ□". This word is written as "জগত" in the dictionary. So a dataset can indicate that the word is spelled wrong and expect a correct word, even though the spell checker didn't detect any mistakes at all. These types of controversial words are relatively few in the Wiki Dataset.

Our second proposed method (ERPP) runs significantly faster than the other two methods (Table 2 - 5, last column). This is because PEP uses the double metaphone encoding to the whole list of phonologically grouped words, whereas the ERPP only runs the double metaphone encoding on a much smaller list. This list consists of only those words which satisfy the threshold τ. The search space is thus reduced to a significant extent. To clearly visualize the pros and cons of the proposed PEP and ERPP let us consider the examples presented in Table 6.

Table-6: Illustrative examples to compare among the Methods

Case	Wrong Word	Correct Word	Suggested Correctly By	Not Correctly Suggested by
1.	পৈত্রিক	পৈতৃক	PMM, PEP	ERPP
2.	ঘুর্ন্নি	ঘূর্ণি	PEP, ERPP	PMM
3.	অধীনস্ত	অধীন	ERPP, PEP	PMM

The main reason for PMM and PEP giving accurate results in case 1 is because these methods consider phonetic and edit distance at the same time to make the initial list of suggestions. However, in order to reduce the search space, ERPP only considers edit distance to generate the initial list. So the correct suggestion was pushed back because there were some dictionary words having less edit distance from the wrong word than the actual target word and for this reason ERPP fails in this case.

On the other hand, PMM performed worse in case 2 and 3, because there were too many words with less or same edit distance than the actual correct word. So the actual correct

word was pushed back further than *k* (=10) in the list and was not returned as a suggestion. Whereas, PEP or ERPP considered prior probability of the word and thus moved the correct word among top ten positions.

IV. Conclusion

Users in general like to see the correct word at the top of the suggestion list. With the rapid growth of smartphones and other devices, the efficiency of the algorithms, in terms of memory and time has also become an important metric. We combine all these observations and propose ERPP, where we combine typological correction, phonological correction, and the prior probability. ERPP provides better accuracy, improved efficiency in terms of time, and more relevant suggestions at the top of the list. However, there are some cases where ERPP fails to generate accurate suggestions (Table 6), which leads us to believe that there is a lot of room for improvement. Besides these, we do not incorporate context information with ERPP. We believe incorporation of context information will further enhance the performance of ERPP which we will address in future.We aim at continuing this research and improve ERPP to obtain an industrial grade Bangla spell checker that would provide even faster and more accurate results.

Acknowledgement

This work is supported by the University Grants Commission, Bangladesh under the Dhaka University Teachers Research Grant No – Reg/Admin-3/80894-C.

References

1 Abdullah, Al-Mahmud, A.B., &Rahman, A. (2003), Spell Checker for Bangla Language: An Implementation Perspective, Proc. 6th International Conference on Computer and Information Technology, Dhaka, Bangladesh.

2 Ristad, E. S., &Yianilos, P. N. (1998),Learning string-edit distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5), pp. 522-532.

3 UzZaman, N., & Khan, M. (2004), A Bangla Phonetic Encoding for Better Spelling Suggestion, Proc. 7th International Conference on Computer and Information Technology, Dhaka.

4 UzZaman, N., & Khan, M. (2005). A double metaphone encoding for Bangla and its application in spelling checker., Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference, pp. 705-710

5 Knuth, D. E. (1982), The Art of Computer Programming, Vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts, 2nd edition.

6 Islam,M. Z., Uddin, M. N.,& Khan, M. (2007), A Light Weight Stemmer for Bengali and its Use in Spelling Checker, Proc. 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA07), Irbid, Jordan.

7 Bhowmik, K., Chowdhury, A. Z., &Mondal, S. (2014), Development of A Word Based Spell Checker for Bangla Language, Doctoral dissertation, Department of Computer Science and Engineering, Military Institute of Science and Technology.

8 Khan, N.H., Khan, M.F., Islam, M.M., Rahman, M.H., &Sarker B. (2014), Verification of Bangla Sentence Structure using NGram, Global Journal of Computer Science and Technology: A Hardware & Computation, vol. 14.

9 Huong, N. T. X., Dang, T. T., & Le, A. C. (2015), Using large n-gram for Vietnamese spell checking, In Knowledge and Systems Engineering, Springer, Cham. pp. 617-627

10 Ahmed, F., Luca, E. W. D., &Nürnberger, A. (2009), Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness, Polibits, (40), pp. 39-48.

11 UzZaman, N. & Khan, M. (2006), A Comprehensive Bengali Spelling Checker, In the Proceeding of the International Conference on Computer Processing on Bengali (ICCPB), Dhaka, Bangladesh.

12 Priya, M., &Kalpana (2016), R. Log Posterior Approach in Learning Rules Generated using N-Gram based Edit distance for Keyword Search, Journal of Intelligent Systems.

13 Mandal, P., &Hossain, B. M. (2017), A Systematic Literature Review on Spell Checkers for Bangla Language, International Journal of Modern Education and Computer Science (IJMECS), 9(6), pp. 40.

14 Kukich, K. (1992), Techniques for automatically correcting words in text, ACM Computing Surveys (CSUR), 24(4), pp. 377-439.

15 Kundu, P., &Chaudhuri, B. B. (1999), Error pattern in Bangla text, IJDL, International journal of Dravidian linguistics, 28(2), pp. 49-88.

16 Chaudhuri, B. B. (2001), Reversed word dictionary and phonetically similar word grouping based spell-checker to Bangla text, In Proc. LESAL Workshop, Mumbai.

17 Khan, N. H., Saha, G. C., Sarker, B., &Rahman, M. H. (2014), Checking the Correctness of Bangla Words using N-Gram, International Journal of Computer Application, 89(11).

18 Mandal, P., &Hossain, B. M. (2017), Clustering-based Bangla spell checker, In Imaging, Vision & Pattern Recognition (icIVPR), 2017 IEEE International Conference on, pp. 1-6.

19 Abdullah, M. M., Islam, M. Z., & Khan, M. (2007), Error-tolerant Finite-state Recognizer and String Pattern Similarity Based Spelling-Checker for Bangla, In Proceeding of 5th International Conference on Natural Language Processing (ICON).

20 Mahmud, H. (2015), উচ্চতরস্বনরি্ভরবশিুদ্ধভাষাশকি্ষা,The Atlas Publishing House.

21 https://bn.wikipedia.org/s/1ylp (Accessed: 27th July, 2017)

22 http://www.somewhereinblog.net/blog/Zobair7/28973213 (Accessed: 28th July, 2017)

JJIT, Vol. **6**, June, 2017, 109-118

AN EFFICIENT APPROACH TO OPTIMIZE THE PROFIT OF A TEA GARDEN BY USING BRANCH-AND-BOUND METHOD

ABU HASHAN MD MASHUD[1], NHM.A.AZIM[2], ROWSHON ARA BEGUM[3] AND KANIJ FATEMA[4]

[1]*Dept of Mathematics,Hajee mohammad Danesh Science and Technology university, Dinajpur-5200.*

[2]*School of Business Studies, Southeast University, Dhaka-1213, Bangladesh.*

[3]*Department of Mathematics, Eden Mohila College, Azimpur, Dhaka.*

[4]*Department of Finance and Banking, Daffodil International College, Dhaka.*

Abstract

In this paper we formulate a new problem as a linear programming and integer programming problem and optimize the profit based on the construction of a tea garden problem. It describes a new idea about how to optimize profit and focuses on the practical aspects of modeling and the challenges of providing a solution to a complex real life problem. Finally a comparative study is carried out among Graphical method, Simplex method and Branch-and-bound method.

Key Words: Integer programming, Tea garden, Graphical method, Simplex method and Branch-and-bound method.

1. Introduction

A special type of algorithm commonly used for discrete and combinatorial optimization problems, as well as in mathematical optimization [3,7] is known as Branch and bound method. It consists of a systematic inventory [6, 1] of candidate solutions by means of state space search: the set of candidate solutions is thought of as forming a rooted tree with the full set at the root. This algorithm explores branches of this tree, which symbolize subsets of the result set. The branch is checked against upper and lower estimated bounds on the optimal solution before enumerating the candidate solutions, and is superfluous if it cannot generate a better result than the best one originate so far by the algorithm.

The algorithm mainly depends on the proficient inference of the lower and upper limits of a area/branch of the investigate space and approaches comprehensive details as the size (*n*-dimensional volume) of the area tends to zero. The cutting plane method developed by R.E Gomory is also called fractional algorithm, while the branch-and-bound method is a search technique was first anticipated by A. H. Land and A. G. Doig in 1960 [2] for the discrete programming, and now has become the most commonly used tool for solving NP-

hard optimization problems[9]. The name "branch and bound" first used in the work of Little *et al.* [5] on the traveling salesman problem.

Integer programming models are used in a wide variety of apphcations including logistics, infrastructure planning, activity scheduling, resource assignment, planning, supply chain design, auction design, and many others.

Applying integer programming to a real life application stated in this paper basically involves two phases; first one needs to create a model for the problem to be solved, and then one applies a solution method to find a good or an optimal solution to the problem described by the model. These two phases are of course not isolated from each other, but the natures of the challenges involved in each phase differ. In the first phase all aspects that need to be considered when solving the problem must be stipulated and quantified in order to create a foundation for the construction of a mathematical model describing the problem stated in this paper. When faced with a complex real life problem the challenge is to construct a model of manageable size that mirrors the reality sufficiently well and yields a solution practical interest.

Integer programming adds additional constraints to linear programming [4]. An integer program begins with a linear program, and adds the requirement that some or all of the variables take on integer values. This seemingly innocuous change greatly increases the number of problems that can be modeled, but also makes the models more difficult to solve. In fact, one frustrating aspect of integer programming [8] is that two seemingly similar formulations for the same problem can lead to radically different computational experience: one formulation may quickly lead to optimal solutions, while the other may take an excessively long time to solve. There is no single method like the simplex method, of solving all types of integer programming problems [8]. A number of integer algorithms have been developed to solve the various types of integer programming problems. Unfortunately none of them is uniformly computationally efficient. Commonly used techniques for solving integer programming problem are: the cutting plane method and branch-and-bound technique.

The first portion of this paper describes the sources of data and procedure, second portion describes the mathematical formulation and solution, lastly a comparative study has been carried out to illustrate the validity of the suggested result.

2. Collection of Data

The data for the paper were collected from the tea company named “Double A tea Estate” which is located in Vitorgor, Panchagar.The actual expenditure incurred on material, labor and transportation cost were collected from the “Double A tea Estate”. This data are used to formulate linear programming problem and solving it by Simplex method, Graphical method and Branch-and-bound method in order to maximize our profit.

3. Solution procedure

Constructing tea garden is a huge task. It's very difficult task to arrange total money in starting of large project. Sometimes we are failure to do so. In this situation we start our work with limited budget. Here we arrange such type of model problem to complete maximum work with our limited budget.

Construction of a tea garden starts its work with limited budget. After completing design and land requisition we have 2 lacks to complete the work. So what will be the maximum area totally completed in this limited budget. Here we use data, which is collected from a company named "Double A tea State".

Here we use two variables x_1, x_2 where x_1 represents the number of units of material per biggha. (1 unit=Tk.31880), x_2 represents the number of units of labor and transportation per biggha.(1 unit=Tk.13900).Here we have to calculate two types of cost such as Material cost, Labor and Transportation cost.

Now we divide our amount as total Tree cost never exceed Tk.62,000,total Fertilizer cost never exceed Tk.30,000 and total water supply cost never exceed Tk.108000. All costs are collected from "Double A tea State" on 20 march 2013, Material cost means purchasing pipe , a pump, and boring equipment for water supply.

According to the company the Material cost Tk. 31880 need for every biggha of tea garden and Labor and Transportation cost need Tk. 13900 for per biggha. The total amount can be divided in the following way:

Tree	**Fertilizer**	**Water Supply**
32%	14%	54%

Mathematical analysis by Branch-and–bound method for maximizing profit of a tea garden

We formulate the above problem in terms of two variables:

Basic things	**Cost per Biggha**	**Material Cost**	**Labor & Transportation Cost**
Tree	14300	11400	2900
Fertilizer	6480	3480	3000
Water Supply	25000	17000	8000
Total	45780	31880	13900

Table: Mathematical Formulation for Constructing Tea Garden.

By the reason of limited money we start our project work with Tk.200000, this money distributes as follows that total Tree cost never exceed Tk.62,000 ,total Fertilizer cost never exceed Tk.30,000 and total Water supply cost never exceed Tk.108,000 in this limited budget we have to find the maximum completed area.

Formulation:

The given problem can be expressed as:

$$\text{Max: } Z = 31880x_1 + 13900x_2$$

Subject to the constraints:

$$11400\,x_1 + 2900x_2 \le 62000$$
$$3480x_1 + 3000x_2 \le 30000$$
$$17000x_1 + 8000x_2 \le 108000$$
$$x_1, x_2 \ge 0$$

Where, x_1 =Number of units of material.

x_2 =Number of units of Labor & transportation.

Solution:

To simplify the computations, the model can be put as:

$$\text{Max: } Z' = Z/1000 = 31.88x_1 + 13.9x_2$$

Subject to the constraints:

$$11.4x_1 + 2.9x_2 \le 62$$
$$3.48x_1 + 3x_2 \le 30$$
$$17x_1 + 8x_2 \le 108$$
$$x_1, x_2 \ge 0$$

In figure 8-1 the ILP solution space is shown by dots. The associated LP solution space, LP0, is defined by dropping the integer constraints. The optimum LP0 solution is given in figure 8-1 as

$$(x_1 = 4.36, x_2 = 4.22, Z_{\max} = 197654.8).$$

The B&B procedure is based on dealing with the LP problem only. Since the optimum LP solution:

$$(x_1 = 4.36, x_2 = 4.22, Z_{\max} = 197654.8)$$

does not satisfy the integer requirements, the B&B algorithm calls for “modifying” the LP solution space in a manner that should eventually allow us to identify the ILP optimum.

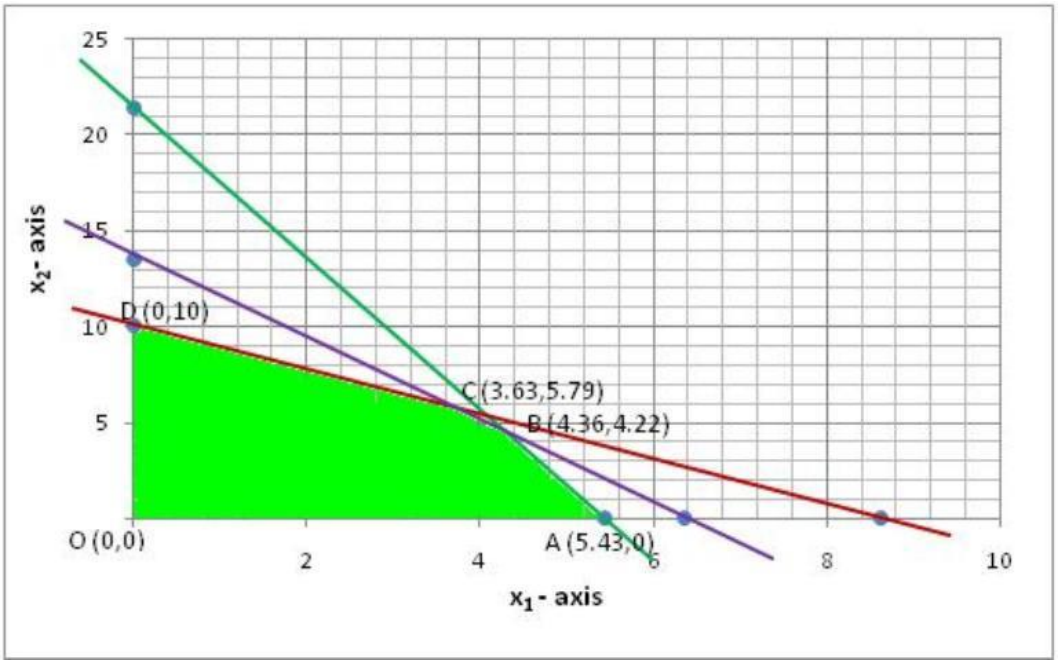


Figure: 8-1

First, we select one of the variables whose current value at the optimum LP0 solution violates the integer requirements. Selecting x_1(=4.36) arbitrarily, we observe that the region (4<x<5) of the LP0 solution space cannot, by definition include any feasible ILP solutions. We can thus modify the LP solution space by eliminating this non promising region, which, in essence, is equivalent to replacing the original LP0 space with two LP spaces, LP1 and LP2, defined as follows:

1. $\text{LP1Space} = \text{LP0Space} + (x_1 \le 4)$.
2. $\text{LP2Space} = \text{LP0space} + (x_1 \ge 5)$.

Figure 8-2 shows LP1 and LP2 graphically. You will notice that the two spaces contain the same integer feasible points of the ILP model. This means that from the standpoint of the original ILP problem, dealing with LP1 and LP2 is the same as dealing with the original LP0.The main difference is that the selection of the new bounding constraints $(x_1 \le 4 \, and \, x_1 \ge 5)$ will now improve the chance of forcing the optimum extreme points of LP1 and LP2 toward satisfying the integer requirements. Additionally, the fact that the bounding constraints are in the "immediate vicinity" of the continuous LP0 optimum will increase their chances of producing "good" integer solution.

As can be seen in Figure 8-2, since the new restrictions $(x_1 \le 4 \, and \, x_1 \ge 5)$ are mutually exclusive, LP1 and LP2 must be dealt with as two separate linear programs. This

dichotomization gives rise to the concept of branching in the B&B algorithm. In effect, branching signifies partitioning a current solution space into mutually exclusive subspace.

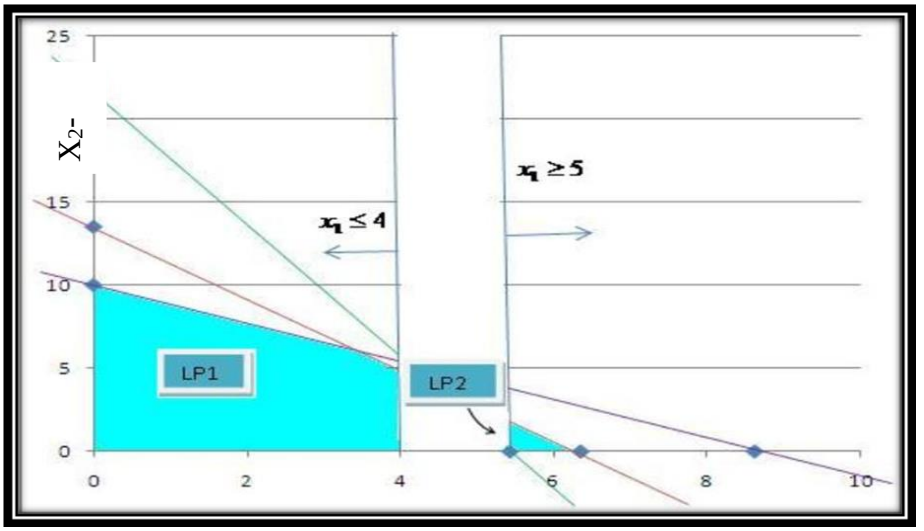


Figure: 8-2

Figure 8-3 demonstrates the creation of LP1 and LP2 from LP.

The associated branches are defined by the constraints in $x_1 \leq 4 \, and \, x_1 \geq 5$ which case x_1 is referred to as the branching variable.

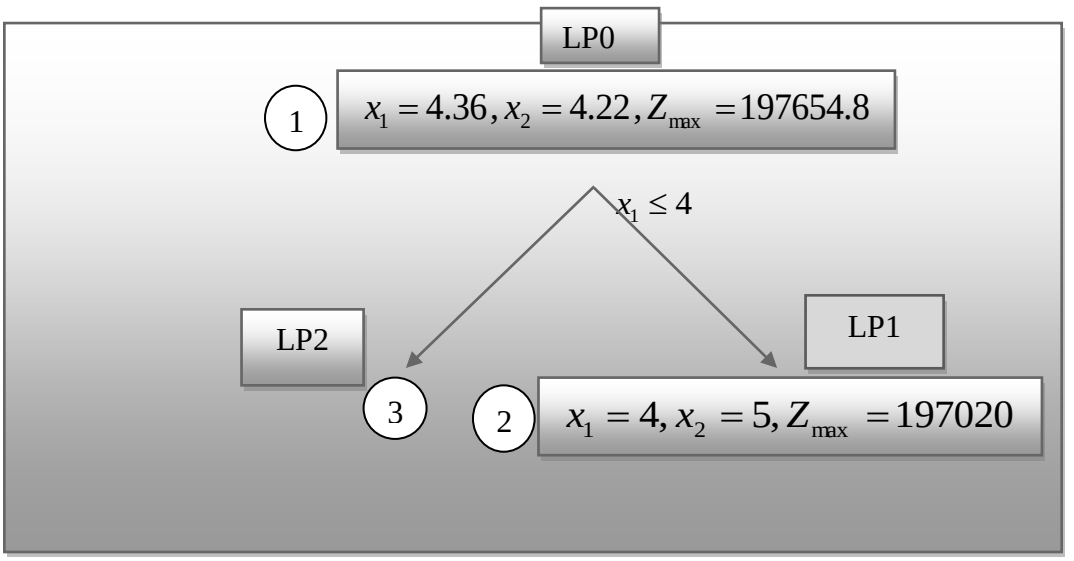


Figure: 8-2

We know that the optimum ILP must lie in either LP1 or LP2.However, in the absence of the graphical solution space; we have no way of determining where the optimum may be. Consequently, our only option is to investigate both problems. We do so by working with one problem at a time (LP1 or LP2).Suppose that we arbitrarily select LP1 associated with $x_1 \leq 4.$

In effect, we solve the following problem:

$$\begin{aligned} 11.4x_1+2.9x_2 &\leq 62 \\ 3.48x_1+3x_2 &\leq 30 \\ 17x_1+8x_2 &\leq 108 \\ x_1 &\leq 4 \\ x_1, x_2 &\geq 0 \end{aligned}$$

As stated above, LP1 is the same as LP0 with the added upper bound restriction $x_1 \leq 4$

We can thus apply the primal upper-bounding algorithm to solve the problem. This will yield the new optimum solution

$x_1=4$, $x_2=5$, and $Z=197020$.

Since the solution happens to satisfy the integer requirements, we say that LP1 has been **fathomed**, which means that LP1 cannot produce any better ILP solutions and hence need not be investigated any further.

The attainment of a feasible integer solution at an early stage of the computations is crucial to enhancing the efficiency of the B&B algorithm. Such a solution sets a lower bound on the optimum objective value of the ILP problem, which in turn can be used to automatically discard any unexplored problems (such as LP2) that do not yield a better integer solution. In terms of our example, LP1 produces the lower bound Z=197020.This means that any improved integer solution must have Z-value higher than 197020.

However, since the optimum solution of the LP0 problem has Z=197509 and since all the coefficients of the objective function happen to be integers, it follows that no sub problem emanating from LP0 can produce a value of Z that is better than 197020.As a result, we can, without further investigation, discard LP2.In this case LP2 is said to be fathomed because it cannot yield a better integer solution.

In our example,LP1 and LP2 are fathomed by conditions 1 and 2,respectively.Since there are no more sub problems to be investigated, the procedure ends and the optimum integer solution of the ILP problem is the one associated with the current lower bound: namely,
$x_1=4$, $x_2=5$, and $Z=197020$.

If we investigate the procedure outlined above, we will discover that a number of questions remain unanswered:

i) At LP0, could we have selected x_2 as the branching variable in place of x_1?
ii) When selecting the next sub problem to be investigated, could we have solved LP2 first instead of LP1?

The answer to both questions is "yes" but the ensuing computational details could differ dramatically. We illustrate this point by referring to figure 8-3 .Suppose that we decided to investigate LP2 first.

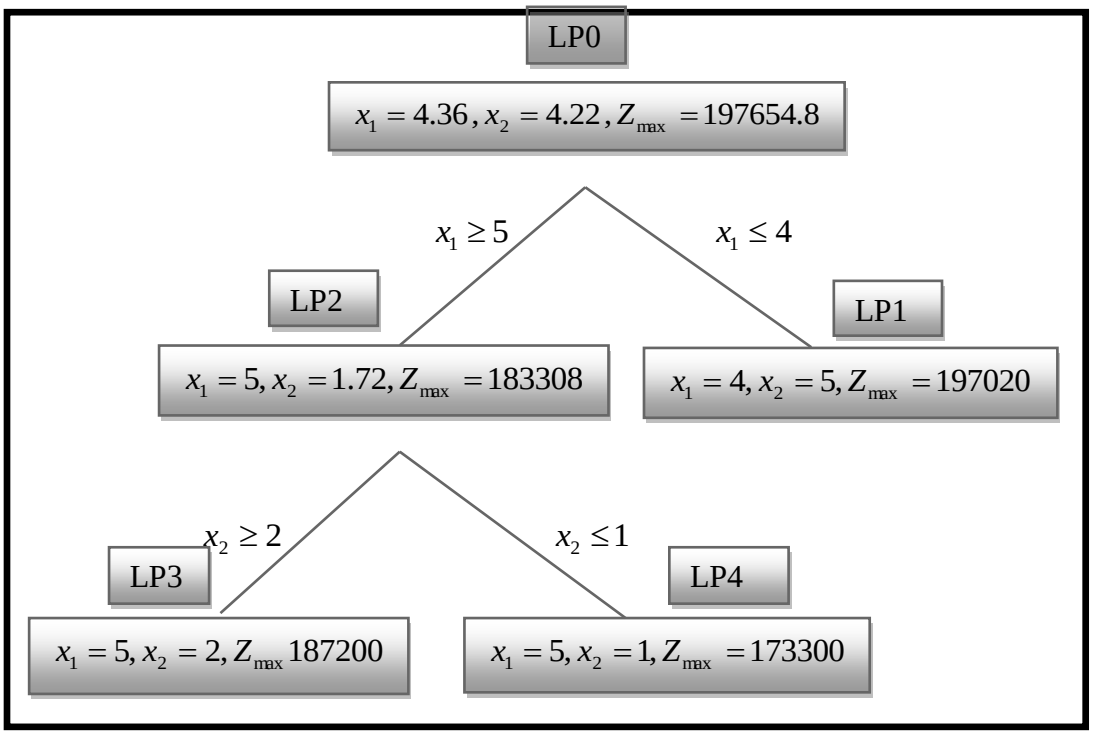


Figure : 8-4

Since x_2=1.72 is non-integer, LP2 must be investigated further by creating LP3 and LP4 using the respective branches $x_2 \geq 2$ *and* $x_2 \leq 1$.This means that

1. LP3 Space = LP0 + $(x_1 \geq 5)$ + $(x_2 \leq 1)$.
2. LP4 Space = LP0 + $(x_1 \geq 5)$ + $(x_2 \geq 2)$.

At this point we have three sub problems to choose from: LP1, LP2, LP3 and LP4.Selecting LP4 arbitrarily, we discover that it has a feasible solution $x_1 = 5, x_2 = 2, Z = 187200$.

Next, we select LP3 it's also have a solution which is :

$$x_1 = 5, x_2 = 1, Z_{max} = 173300.$$

Unfortunately, this lower bound is both "too weak "and "too late" to be useful. The only remaining node is LP1, is fathomed next with Z=197020, which immediately sets a new lower bound. Since there are no more sub problems to be investigated, the last lower bound associates the optimum ILP solution with LP1.

4. Comparison of Results

In order to establish maximum profit, we formulate the linear programming model to describe the problem stated in the methodology. Then the resulting linear programming model was solved by using Simplex method, Graphical method and Branch-and-bound method. The summary of the optimal solution for linear programming models formulate is as shown in below:

Methods	Variables	Total Amount(Tk.)	Comments
Simplex method	$x_1 = 4.36, \quad x_2 = 4.22$	197654.8	Not perfect
Graphical method	$x_1 = 4.37, \quad x_2 = 4.24$	198251.6	Not perfect
Branch-and-bound method	$x_1 = 4, \quad x_2 = 5$	197020	More Perfect

5. Conclusion

In this paper our main goal is to formulate the construction of a tea garden problem as a Linear programming and Integer Programming problem to optimize the profit and cost of a tea garden. Then we have used different methods to optimize the profit and cost such as simplex method, branch-and-bound methods, graphical method and suggesting that Branch-and-bound method gives the maximum advantages within limited budget and resources.

References

1. Abu Hashan Md Mashud, Md. Al-Amin Khan, M. Sharif Uddin and M. Nazrul Islam, (2018), A non-instantaneous inventory model having different deterioration rates with stock and price dependent demand under partially backlogged shortages , *Pages: 49-64* . DOI: 10.5267/j.uscm.2017.6.003.

2. A. H. Land and A. G. Doig (1960), "An automatic method of solving discrete programming problems". Econometrica. 28 (3). pp. 497–520.

3. Clerc, M. (1999), 'The swarm and queen: towards a deterministic and adaptive particle swarm optimization', *Proceedings of IEEE Congress on Evolutionary Computation*, Washington, DC, USA, pp.1951–1957.

4. G.B. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, N J, 1963.

5. Little, John D. C.; Murty, Katta G.; Sweeney, Dura W.; Karel, Caroline (1963), "An algorithm for the traveling salesman problem" *(PDF)*. Operations Research. 11 (6): 972–989.

6. Shaikh, A.A.,Mashud, A.H.M., Uddin, M.S. and Khan, M.A-A. (2017), 'Non-nstantaneous deterioration inventory model with price and stock dependent demand for fully backlogged shortages under inflation', *Int. J. Business Forecasting and Marketing Intelligence*, Vol. 3, No. 2, pp.152–164.

7. Sun, J., Feng, B. and Xu, W.B. (2004a*)*, 'Particle swarm optimization with particles having quantum behaviour', *IEEE Proceedings of Congress on Evolutionary Computation*, pp.325–331.

8. Taha, H.A. Natarajan, A.M., Balasubramanie, P., Tamilarasi A., and Operation Research: An Introduction, Pearson Education, Eighth Edition, 2008.

9. Xi, M., Sun, J. and Xu, W. (2008), 'An improved quantum-behaved particle swarm optimization, algorithm with weighted mean best position', *Applied Mathematics and Computation*, Vol. 205, No. 1, pp.751–759.
.