ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 5, Issue 1, January 2017

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

An Approach of Speech Recognition System for Desktop Application

Prof. Shivashankar M. Rampur[1]
Department of Computer Science and Engineering
Malik sandal polytechnic, vijaypur
Vijaypur – India

Kavita S. Rampur[2]
Department of Electrical and Electronics Engineering
BLDEA's Engineering college, Vijaypur
Vijaypur – India

Abstract: ***Speech Recognition for desktop application is a system that allows the computer to identify and understand the commands spoken by a particular speaker based on individual information included in speech waves. The ultimate goal of work is to be able to produce a system that can recognize with expected accuracy, all commands that are spoken by particular person so that each operation of the application (power point) is accessed by corresponding commands. In the proposed work, Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) are the techniques used for speech recognition system. Speech signals are taken and sampled directly from microphone and then they are processed using LPC method for extracting the features of speech signal. These features are used to construct the database to train the Artificial Neural Network using Back propagation method. Further the testing is conducted for the particular person by giving the voice command and if command is match/found then corresponding operation of application(power point) get enabled. The overall recognition accuracy of the proposed system with static or standard database is about 91.75% when tested with minimum 100 voice sample commands of a particular speaker. But, for real-time system, the accuracy is depending on the environmental situation and it will gives around 85-86% accuracy in worth.***

Keywords: Speech Recognition System, Linear Predictive Coding (LPC) Feature Extraction Method, Back Propagation Neural Network.

I. INTRODUCTION

Speech recognition is nothing but recognizing persons from their speech. Generally, any two individual's voice will not be identical because of their larynx sizes, vocal tract shapes and other parts of their voice production organs are different. State-of-the-art speaker recognition system utilizes a number of features to reach highest accurate recognition. One of the important applications of speaker recognition is voice based interactive security system in which an individual's voice is used for authentication. Voice/speech recognition is very simple, robust and secure technology. To recognize the speaker's speech command efficiently different parameters of speech are used like pitch, amplitude pattern or power/energy. When an individual person speaks, their voice contains many levels of information. Speech recognition technique includes extracting unique features from their voice samples and makes these features as knowledge base and finally test individual person's voice by matching knowledge base and predict about voice whether matched or unmatched.

Speech recognition is categorized as identification and verification:

Speech identification is the task of identifying certified speaker who has provided the given vocalization. In this system, the vocalization/utterance of test phrase is matched with a set of certified speakers and the best match of the test vocalization is selected. Speech identification is further classified into two types i.e. Closed-set and Open-set.

Speech verification involves the task of accepting or rejecting an identity asserted by an individual. Speaker and anti-speaker model is calculated by verification score during testing. This verification score is compared to a threshold. If the

threshold is less than verification score, then speaker is accepted else rejected. Speech recognition is based on the principle that speech of speaker exhibits unique characteristics. However, speech signals in training and testing sessions may differ for many reasons such as age growth, illness (e.g. speaker may be suffer from cold and cough),speaking rate, speaking style etc.

II. SPEECH RECOGNITION SYSTEM DESIGN

Before going to discuss about feature extraction of speech signal. Let's design and implement the model for particular user authentication on application in which voice command of particular speaker is matched with stored commands from database and corresponding operations of power point application get enabled. Speech recognition system operates into two phases: Training phase and testing phase. Each speaker has to give voice commands as input samples in training phase, which uttering the commands "one', "two", "three", "four". Once speech acquisition is done successfully, preprocessing steps like noise removal, silence removal, pre-emphasis, framing and windowing is carried out. Later from LPC coefficient, unique Autocorrelation features are extracted from their speech and these feature vectors are used to construct the database, so that model can construct a reference model for the certified users of the system. In the testing phase, the input voice command is given from particular speaker through headphone and feature vectors are matched with trained Artificial Neural Network (ANN) classifier for voice command recognition.

The block diagram of proposed speech recognition system for desktop application is described in figure 1.

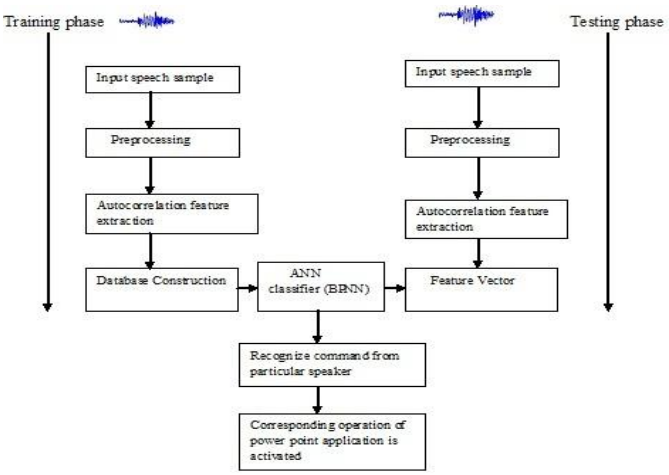


Fig. 1 the block diagram of proposed voice recognition system for desktop application

A. Feature Extraction Using Linear Predictive Coding

The feature extraction is the method or technique that is used to extract the relevant information from speech signal which is discriminative and more constant than original speech signal. Linear predictive coding (LPC) is also called as Autoregressive model. This method is used to compress the speech signal for efficient transmission and storage. For efficient speech signal features Linear Predictive Cepstral Coefficients (LPCC) are generally used and it is widely used in various areas such as extraction of speech features, speech analysis and user recognition.

Prediction of linear is a method or technique which is used for estimation of spectral and it provides the estimate of the poles (formants) of the vocal tract transfer function. The LPC encodes the speech signal by calculating the formants and removing their unwanted sound from the speech signal and finally calculating the frequency and intensity of remaining buzz. Every speech sample can be approximated as a linear combination of a few past speech signals through LPC coding. The linear prediction technique provides reliable, robust and accurate method for calculating the parameters. The LPC algorithm is shown in figure 2.

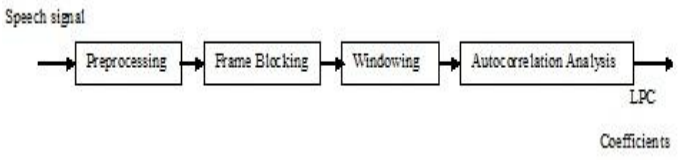


Fig.2 the block diagram of LPC algorithm

B. Calculating Autocorrelation Features

If we assume x(a, b) is the additive noise, v(a, b) is the speech signal of sound-free, h(n) indicates impulse response of the channel, and then the noised speech signal y(a, b) will be written as

$$y(a, b) = v(a, b) * h(n) + x(a, b)\ 0 \le a \le M-1,\ 0 \le b \le N-1,$$

Where * indicates the convolution operation, b indicates discrete time index of the frame, N shows frame length of speech, a indicates the index of frame. Number of frames is denoted by M. If x(a, b), v(a, b) and h(n) will be treated as uncorrelated. Feature of the autocorrelation of noised speech sample is given as:

$$cyy\ (b, k) = cxx(a, k)*h(k) *h(k) + cvv(a, k)\ 0 \le a \le M-1,\ 0 \le k \le N-1,$$

Where cyy (a, k) , cxx(a, k) and cvv (a, k) are the short-time autocorrelation sequences of the noised speech, very clean speech and noise respectively and k denoted by index of autocorrelation sequence within each frame. So that additive noise is assumed to be a stationary. The one-sided autocorrelation sequence of each frame will be estimated by using an unbiased estimator. But, for its calculation, may use a biased estimator. For the calculation of one-sided autocorrelation sequence the unbiased and biased estimators are used and given below:

$$cyy(a,k) = \frac{1}{N-K}\sum_{i=0}^{N-1-K} y(a,i)y(a,i+k)$$

$$cyy(a,k) = \sum_{i=0}^{N-1-k} y(a,i)y(a,i+k)$$

Figure 3 represents the speech signal sample and magnitude of autocorrelation spectrum and magnitude of differentiated autocorrelation spectrum for the same signal

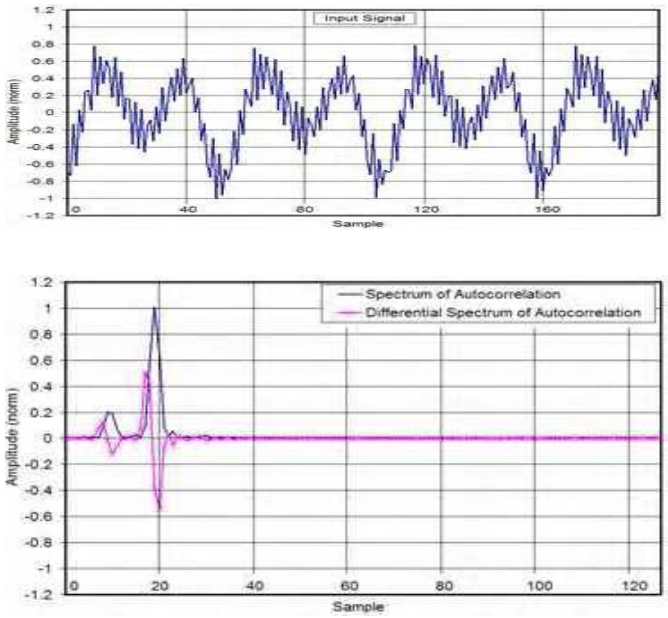
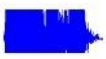


Fig.3 Representation of a speech signal sample and magnitude of the autocorrelation spectrum.

The basic effects of autocorrelation feature of noise on transparent (no disturbance) voice signal of autocorrelation are on its lower lags. So that removing the lower lags of the autocorrelation of disturb speech signal must gives to removing of components of main disturb speech signal. Autocorrelation's maximum index to be removed and that is found experimentally.

Once the calculation of autocorrelation features has done then next is description of speech signal and its feature values are explained. A speech signal is given as input to Autoregressive model from which 228 features are extracted from each command. The trained network is loaded and simulates the extracted features. This gives the match of command and activation of operations of power point applications. At very first step, speech sample acquisition is performed and given as input for preprocessing step, further for each command 228 autocorrelation features using LPC coefficients are extracted. For each user, all four commands are used for training phase. The following table describes the speech signal of particular command and their features. The sample input command and their associated autocorrelation features are depicted in table 1.

Table 1 Voice sample commands of a speaker and its autocorrelation features using LPC coefficient

SL.NO	COMMAND NAMES	SPEAKER VOICE SAMPLE COMMAND	AUTOCORRELATION FEATURES(C1 TO C128)
1	ONE		0.3194 0.3234 0.1605 0.0551 0.3191 0.0856 -0.1228 0.0142 0.1074 0.0028 -0.1713 -0.0993 1.0833 0.0206 -0.0180 0.0137 -0.0120 0.0083 -0.0154 0.0081
2	TWO		-0.1372 0.2776 0.2756 0.1651 - 0.1043 -0.0902 0.0015 0.0618 0.0364 -0.0813 0.8077 -0.5346 0.4517 -0.2127 1.4832 -1.3966 1.2117 -0.5605 0.3342 -0.2320
3	THREE		0.1722 -0.1834 0.3991 -0.1278 0.3914 -0.2261 -0.0185 0.0456 0.0195 -0.2446 0.3664 -0.3201 0.0754 -0.2371 0.0310 -0.0783 0.0138 -0.1087 -0.0954 -0.1071
4	FOUR		0.2315 -0.1040 0.3719 0.1739 0.2263 0.1149 0.0776 -0.0034 -0.0445 -0.0531 0.3133 -0.0721 -0.1070 -0.0440 0.0331 -0.0459 0.7093 0.1900 0.0185 -0.0332

From each speaker, take the 100 voice commands as input and perform the preprocess step, further extract the 22800 autocorrelation features using LPC method and train on those samples using BPNN and stored into database. Later, start to give

voice command for testing so that if command is matched or found then it's considered to be recognized and operations of power point application get activated.

In the proposed system, there are four commands like One, Two, Three and Four and One corresponds to open power point, two corresponds to add slide, three corresponds to Add image and Four corresponds to quit ppt.

C. Recognition Using Artificial Neural Network

The Artificial Neural Network is the biologically scoped classification algorithm. It is made up of a large number of highly interconnected processing elements i.e. neurons. In this paper, Back propagation neural network method is used for recognition of speech command.

Back propagation neural network is nothing but, a large number of simple elements i.e. neuron. Those are called as processing units and organized into several layers. Each layer consists of units and the previous layers are connected by all the units of present layer. But, all these connections may not be equal and each connection between units won't have a different weight/ strength. Connections associated with weights are encoded the knowledge of network. Neural networks consist of number of units those are called as nodes. Enters of data are at the inputs and passes through layer by layer in the network, till it comes at the outputs. There is no feedback between layers when it acts as a classifier for normal operation. Therefore, they are known as Feed forward neural networks. Back propagation means that common method through which networks will be trained. It takes the algorithm and changes the weights of network so that when training is finished, it results with required output for a particular input. Training method is the process through which the weight matrix of a neural network is automatically adjusted in order to produce expected results. Neural networks are usually nonlinear function with basic equation is F(x, w)=y, where in x indicates input vector represented to the network and weights of network is denoted by w and y indicates the respected output vector that is predicted by network. First vector's weight w is ordered by layer and then ordered by elements/neurons and finally it is by weights of every neuron and its bias. A simple neural network is shown in figure 4. which allow the signal to travel in only one direction i.e. from input to output.

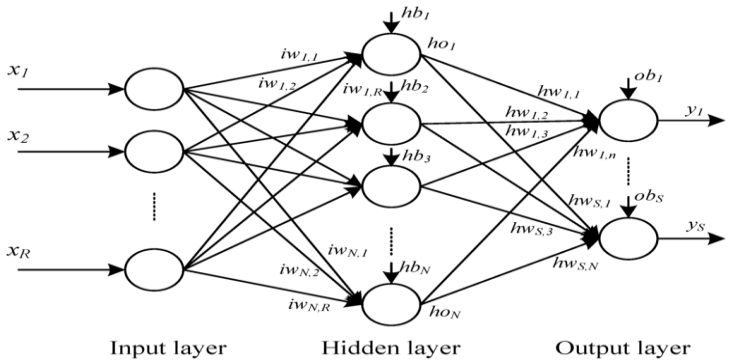


Fig. 4 General Neural Network

The basic Back propagation algorithm consists of 3 steps:

1. Input layer is given by input pattern which are propagated through the network, till those pattern reach the output units. Predicted output pattern will be produced from forward pass.

2. The preferred outputs are fed as a part of training vector. The error signal is produced by subtracting the outputs of actual network from desired outputs.

3. An error signal is origin in back propagation step and those are passed back through the neural networks. The weights of connection are adjusted and neural network has now "learned" from an experience.

For the proposed neural network, total of 228 nodes in an input layer (I), It estimates number of features fed as input to train the network, number of nodes in output layer (O) is 4,it determines a matching of certified speakers command. The number of nodes in hidden layer (n) will be calculated by input and output layer as shown in below equation.

$$n = y + \frac{I + O}{2}$$

From gradient descent method, the algorithm of back propagation provides the minimum of the error function in weight space. The combination of weights which will minimize the error function that is considered as a solution for learning problem. So that this method requires estimation of gradient of the error function at each and every iteration step, it should give the guarantee that continuity and differentiability of error function, the composite function produced by interconnected perceptrons are discontinuous. One of the more popular activation functions for back propagation networks are the sigmoid, a real function Sc: IR (0, 1), it is defined by following expression

$$Sc(x) = 1/1+e^{-cx}$$

Let c will be constant and it is selected arbitrarily and reciprocal of c i.e. 1/c is known as temperature parameter in stochastic neural networks. When the value of c changes then shape of the sigmoid also changes.

D. Proposed Speech Recognition System

In the Proposed speech recognition system, It is preprocesses all the speech samples and extracts the 5700 autocorrelation features (25*228) from each command using LPC coefficients. Further from these unique features, ANN is trained using BPNN.

For each person, total of 22800 features (100 samples * 228 features) obtained and trained using Back propagation neural network. The developed system is tested from speaker by giving input commands to analyze the accuracy result with each person. The following figure 5. Shows the proposed speech recognized system for desktop application.

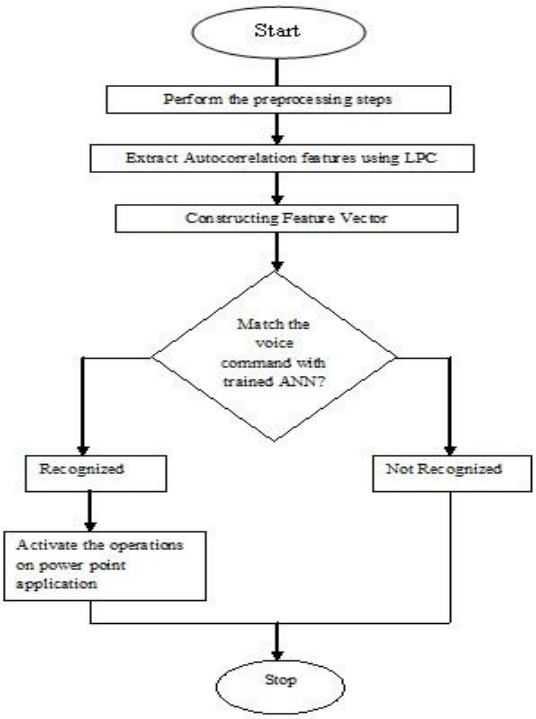


Fig.5 Flowchart for the proposed Voice recognized system for desktop application

III. EXPERIMENTATION

Initially MFCC method has considered developing speech recognition system but there were some issues such as less accuracy, it was difficult to face removal of voice and silence so on. Therefore finally LPC method is selected In the proposed system, there are three speech/voice sample databases, each database is composed of 100 commands from a particular speaker uttering the commands like "ONE','TWO","THREE" and "FOUR", for each command 25 times in English language. When speaker says command "ONE", then power point application will open with named test.pptx (filename),when says command "TWO" then new slide is added in same application, when command "THREE" is said then new image is added and finally when speaker says command "FOUR" then quit the power point application. These commands characterize unique features which then allow the Artificial Neural Network to differentiate individual commands with maximum accuracy.

It is preprocesses all the speech samples and extracts the 17100 autocorrelation features (25*228) from each command using LPC coefficients. Further from these unique features, ANN is trained using BPNN. The test voice sample command is given as input from particular speaker, then follows preprocessing and feature extraction step. Next the extracted features are matched with trained ANN and if match is found, then corresponding operation of power point application will get activated. And finally the speaker can exit from the proposed system.

A.Performance measure of speech recognition system

The performance measure of speech recognition system can be determined by two parameters: The Rate of Correct Recognition (CRR) and Rate of False Acceptance (FAR**).**

The Rate of Correct Recognition (CRR) is the ratio of number of correctly recognized command from the speaker to the total number of authorized speaker's commands.

$$CRR = \frac{\text{Number of correctly recognized speaker commands}}{\text{Total number of authorized speaker commands}} * 100\%$$

The Rate of False Acceptance (FAR) is the ratio of number of incorrectly recognized authorized speaker to total number of speaker's commands.

$$FAR = \frac{\text{Number of incorrectly recognized speakers commands}}{\text{Total number of authorized speaker commands}} * 100\%$$

The experimental results for each command wise accuracy for static speaker using speech recognition system for desktop application are tabulated and depicted below in Table 2.

Table 2 Performance of Speaker in terms of accuracy of each command in static

Sl.No	Name of Speaker	Command Names	Testing Samples	CRR (in %)	FAR (in %)
1	Shivashankar	ONE	25	92.00%	8.00%
		TWO	25	90.00%	15.00%
		THREE	25	95.00%	5.00%
		FOUR	25	90.00%	10.00%

The overall recognition accuracy of the proposed system with static or standard database is about 91.75% when tested with minimum 100 voice sample commands of a particular speaker. But, for real-time system, the accuracy is depending on the environmental situation and it will gives around 85-86% accuracy in worth. Recognition rate of each command from particular speaker is drawn that gives the accuracy of each command and command "ONE" shows 92% accuracy, "TWO" shows 90% accuracy, "THREE" shows 95% and "FOUR" shows 90% accuracy that is shown in figure 6.

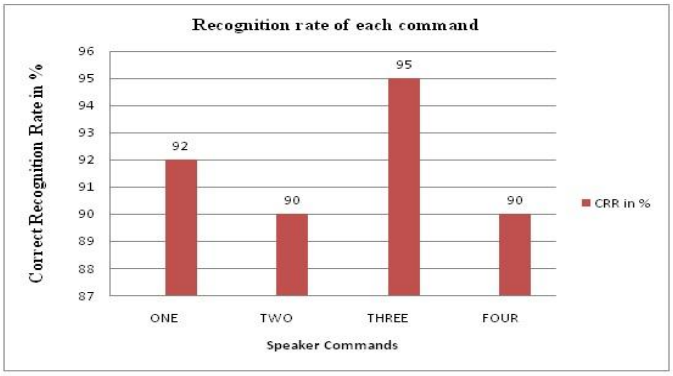


Fig.6 Recognition rate of commands from user.

IV. CONCLUSION

The Speech recognition is the one of the popular and known task for several security related applications. During work with proposed system, variations in speaking rate, style and pitch are commonly faced issues from speech recognition system. In the proposed work, autocorrelation features using LPC coefficients are extracted from each command given by particular speaker. Later Back propagation neural network is used as classifier and for each command corresponding operation of power point application will activated and achieved the recognition accuracy is around 91.75%. In future, operations on any Desktop applications must work with effective and accurate and need to use this speech recognition system for more security, robustness and enabling different operations on real time applications. Since the proposed system work is on focusing only on authorized speaker can access the application and for each speaker, separate database is created for secure purpose using LPC coefficients.

In future, other unique features are extracted from speech sample using different speech analyzer and increase the accuracy and robustness of the system.

ACKNOWLEDGEMENT

The authors of this paper would like to thank to Guide, parents, friends and relatives for their useful suggestions to improve this paper.

References

1. D R Rakesh, K Jayasiman, 2013, “Speaker Recognition and Authentication”, International Journal of Mobile computing and Computer science, Vol 2, Issue. 5, 2013.
2. N Praveen, Tessamma Thomas, 2013, “Text dependent speaker recognition using MFCC features and BPANN” International Journal of computer applications, Vol 74,pp.5,2013.
3. Amitava Das, Tapaswi Makarand, 2010, “Direct modeling of spoken passwords for speaker recognition by compressed time-features representations”, IEEE International Conference on audio, voice &signal processing, pp.4510-4513, 2010.
4. S.S Wali, S M. Hatture and Nandyal S,2014,”MFCC Based Text-dependent Speaker Identification Using Back propagation neural network”, International journal of signal processing systems Vol.3,pp.1-8,June 2014.
5. Vimala C, Dr. V.Radha “A review on speech recognition challenges and approaches”, The world of computer science and Information Technology journal.
6. Nitin Trivedi, Sachin Ahuja, Dr. Vikesh kumar Ramanchand, Sourabh singh,2011, “Speech recognition by Wavelet Analysis”, International journal of computer applications, Vol. 15 No.8.
7. Aju Joseph, Anish Babu K, Karthik Selven,2013,”Speaker recognition system for security applications” ,IEEE Recent advances in intelligent Computational systems, pp.26-30,2013.
8. Shruti Gujrat, Varindar Singh Baidwan,2014,”Comparitive analysis of prosodic features and Linear prediction Coefficients for speaker Recognition using Machine Learning Technique”, IEEE international conference on Devices, Circuits and communications,pp.1-8,2014.
9. M Hanuma Teja, Chaitra N, 2011, “Computationally efficient speaker identification system using AMDF and LPC”, IEEE international conference on Electronics computer Technology,Vol.3,pp.288-289,2011.
10. Md. Fayzur Rahman, Md. Rabiul Islam,2009,”Improvement of Text dependent Speaker identification system using Neuro-Genetic Hybrid algorithm in office Environmental conditions”, IJCSI international journal of computer science Issues,Vol.1,2009.

11. Sonam kumara, Komal saxena,2012,"Controlling of device through voice recognition using Matlab", International Journal of Advance Technology and Engg. Research, Vol. 2, Issue 2, march, 2012.
12. Thiang, suryo wijoyo,2011,"Speech recognition using LPC and Artificial Neural Network for controlling movement of mobile Robot", Indonesia international conference on information and Electronics Engg. IPCSIT Vol.6(2011).
13. Kalid saeed and Mohammad kheir, 2007, "A speech and speaker identification system: Feature extraction", IEEE transactions on industrial electronics, Vol 54,No 2 april 2007-887.
14. Supriya s. and Dr. Y.S Angal, 2012,"Speech recognition using HMM/ANN Hybrid model", International jounal on recognition and Innovation trends in computing and communication, Vol.3 issue 6,2012.

AUTHOR(S) PROFILE

Shivashankar M.Rampur, received the B.E degree in Information Science and Engineering and M.Tech degrees in Computer Science and Engineering from Basaveshwar Engg College Bagalkot, During 2013-2016, Now, currently working as a Lecturer in Malik Sandal Polytechnic, Vijaypur

