Available online at www.elixirpublishers.com (Elixir International Journal)

Electrical Engineering

Elixir Elec. Engg. 78 (2015) 29484-29488

ISSN: 2229-712X

An ANN based mobile robot control through voice command recognition using Nepali language

Neerparaj Rai and Bijay Rai
Department of Electrical and Electronics, SMIT, India.

ARTICLE INFO

Article history:
Received: 8 August 2014;
Received in revised form:
19 December 2014;
Accepted: 29 December 2014;

Keywords
Mobile Robot,
Feed-forward back-propagation,
Linear predictive coding,
Neural networks,
Speech recognition.

ABSTRACT
The analog speech signal can be used for interacting with machines, computers or robots. In this case, speech algorithm is capable to recognize the voice commands that are given as inputs to a mobile robot through wireless XBee modules. It is actually a form of Word Recognition. In this presented paper, a voice command recognition system is going to be developed by using Artificial Neural Network (ANN). The Commands that used here are all in Nepali Languages (used in North East India). Linear predictive coding (LPC) has been applied to represent speech signal in frames in early stage. Features from the selected frames are used to train multilayer perceptrons (MLP) using back-propagation. The same routine is applied to the speech signal during the recognition stage and unknown test patterns are classified to the nearest patterns. In short, the selected frames represent the local features of the speech signal and all of them contribute to the global similarity for the whole speech signal. The analysis, design and development of the automation system are done in MATLAB, in which an isolated word speaker independent digits recogniser is developed.

Introduction

In the field of speech recognition, a large number of algorithms and methods have been proposed for various purposes. The requirement of different applications drives the researchers to develop new algorithms or improve existing methods to serve the need in different situations. For example, speaker-dependent (SD) systems which accept the speech from specific speakers are usually applied in security systems. On the other hand, speaker independent (SI) recognisers are designed to recognise speech from different speakers such as speech to text engines in word processing programs, as a substitute to a keyboard. Broadly speaking, speech recognition systems are usually built upon three common approaches, namely, the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach [1]. The acoustic-phonetic approach attempts to decide the speech signal in a sequential manner based on the knowledge of the acoustic features and the relations between the acoustic features with phonetic symbols. The pattern recognition approach, on the other hand, classifies the speech patterns without explicit feature determination and segmentation such as in the formal approach. The artificial intelligence approach forms a hybrid system between the acoustic phonetic approach and the pattern-recognition approach.

In Speaker Recognition System, the speaker should be recognized based on collected speech database. Speakers are recognized by their voice according to different voice properties [4]. Regional language like Nepal or other can be used here in the form of normal speech or music [5]. But if collected database is small for training and testing then short utterances speech recognition (SUSR) technique can be used [13]. Authentication issue can be implemented also [2, 7]. Neural network also helps to implement authentication on speaker recognition. For this purpose graphical representation of voice signal is used where signal-images are actually used for extraction of various features [7]. Speaker Recognition further includes two subsequent steps that are Speaker Identification and Speaker Verification. In the case of Speaker Identification, spectral analysis can be done after extraction of features [8]

The artificial intelligence approach becomes the field of interest after seeing the success of this approach in solving problems (especially classification problems) [3]. The application of artificial neural networks is proposed to meet the needs of an accurate speech recogniser. For example, the neural network approach to phoneme recognition [9, 10] is proposed in Japanese vowel recognition. Besides, the combination of neural networks and linear dynamic models is proven in achieving a high level of accuracy in automatic speech recognition systems. Another problem in speech recognition is the increase of error in the presence of noise such as in a typical office environment. Some researchers propose the use of visual information such as the lip movement [11, 12].

We propose the use of a multilayer perceptron (MLP), which is trained using the back-propagation technique to be the engine of an automated speech recognition system. Firstly, the features of the training datasets are extracted automatically using the end-point detection function. The features are then used to train the neural network. The system was built using MATLAB [14] and accuracy greater than 90% was achieved for the unknown patterns. Based on the feature extraction, specific serial ASCII commands are sent to the robot through wireless communication.

Methodology

The most important elements used in this paper consists of a headset microphone which acts as a voice sensor which is connected to a laptop for audio processing. The laptop is also connected to XBee through a USB cable for wireless command communication to the robot which is received at the receiver end

Tele:
E-mail addresses: neerparaj_rai@yahoo.co.in

using another XBee. The XBee acting as a receiver directly sends the received data to Atmega 328 microcontroller which drives the robot motors using L293D motor driver IC.

Robot Design

In this paper, mobile robot was designed in accordance with purposes, mechanism interpretation to control, and final design based on situation. The design of the mobile robot is simple yet convenient for the system. The main board and the XBee module along with the motor driver are placed on the bottom layer as shown in Fig. 2(a) and (b). The mobile robot consists of a chassis mounted on four wheels out of which two are dummy wheels and the other two are attached to 12V gear motors. The complete circuit for the robot operation is placed on the chassis. The gear motors are driven by motor controller driver IC L293D for forward, backward, left and right movements. The chassis also holds XBee module circuit and a 12V battery pack for power supply.

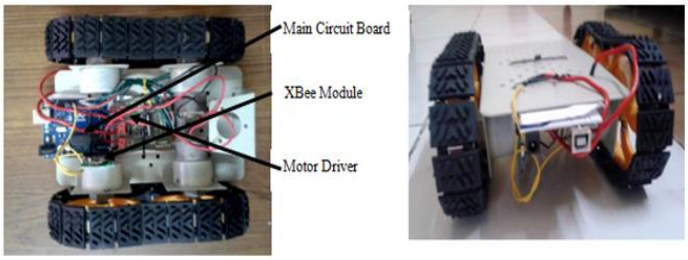

Fig. 1: (a) Bottom View (b) Side View

Atmega 328 is used as the main controller to control the motors and to communicate with the XBee module circuit. Atmega 328 is a 28 pin microcontroller with 14 digital pins and contains program written in C language for its required operation. The schematic for main board circuit is shown in Fig. 3a). Main board is used as the main controller for the control decision of the mobile robot as shown in Fig.3b).

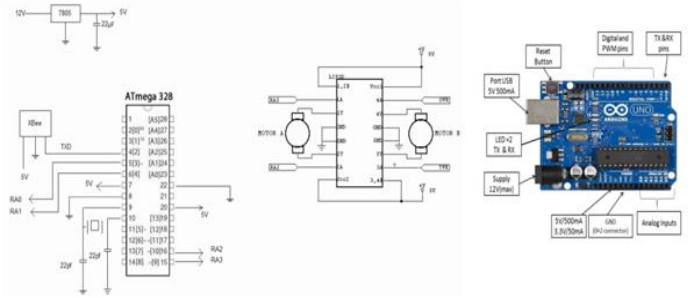

Fig. 2: (a) Main board basic schematic circuit. (b) Main board circuit.

The pwm signals RA0, RA1, RA2 and RA3 from the microcontroller controls the speed as well direction of the robot. The data transmit line of the XBee module is connected to digital pin 2 of the microcontroller working at a clock pulse of 16MHz. The controller then identifies at which pin the PWM signal has to been sent and then it operates the motors of the robot accordingly. The microcontroller is used in 8 bit UART mode with 1 start bit, 8 data and 1 stop bit at 9600 buadrate. Fig. 4 explains the complete block diagram for the operation of the proposed system.

Database Preparation

Our Methodology includes various steps such that Collection of Voice Command data, Pre-processing of voice data, Extraction of various features, Artificial Neural Network Training, Voice Command Recognition. Here speech commands are taken in Nepali language for our database generation. There are total six speech commands that are SURU, ANTA, MATTHI, MUNNI, DAINAY and DEBRAY. The following table is showing the different Nepali speech commands and their corresponding meaning. Here the microphone of personal computer is taken for recording purpose. Recording is done at Mono Channel, 16 bit per sample and 8 KHz sampling rate. After recording (.wav) sound file is generated. Here we are considering 20 different individuals for recording and creating the Speech Command Database.

Table I. Meaning of different Nepali voice commands

SL.No	Voice Command	Meaning in English
1.	SURU	Start
2.	ANTA	Stop
3.	MATTHI	Upwards
4.	MUNNI	Backwards
5.	DAINAY	Right
6.	DEBRAY	Left

End Detection

A speech signal is usually classified into two states. The first state is silence, where no speech is produced. The second state is voices, in which the vocal cords vibrate and produce a quasi-periodic signal. The silence state is usually the unwanted state and has to be removed in order to save the processing time of the speech recognition system as well as to improve the accuracy of the system. In the time domain, the amplitude of the speech signal at each sampling time is plotted over time.

The end point detection technique is applied to extract the region of interest from the speech signal. In other words, it removes the silent region in speech signals. The basic technique of end point detection is to find the energy level of a signal. Signal energy level is calculated in frames, where each frame consists of N samples. The frames usually overlap with the adjacent frames to produce a smooth energy line**.** Fig. 4 shows the energy plot of "Upwards".

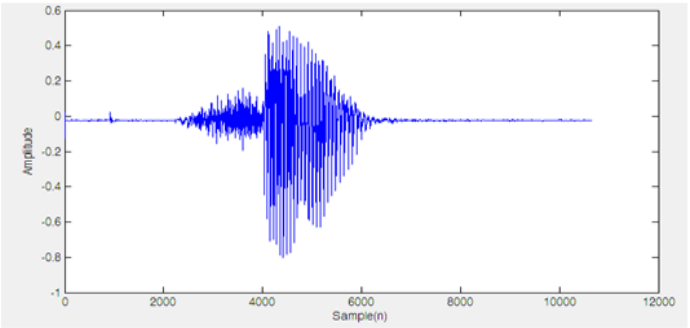

Fig 3. Original Speech Signal

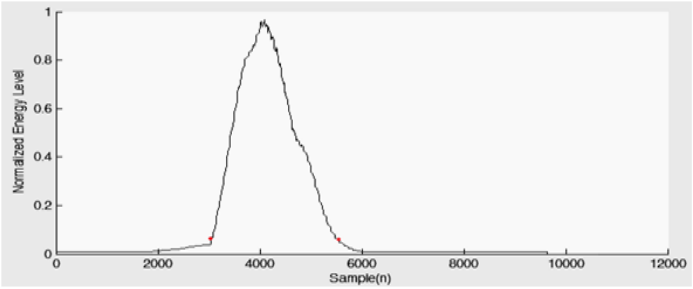

Fig. 4: End point detection using energy level of the speech signal.

Accurate end-point detection is important to reduce processing load and increase the accuracy of a speech recognition system. Basically there are two famous endpoint detection algorithms. The first algorithm uses signal features based on energy level and the second algorithm uses signal features based on the rate of zero crossings. The combination of both gives good results, but increases the complexity of the program and also the processing time. In this project, an end-point detection method that is based on the energy level is applied to reduce the pre-processing time [16].

Fig. 3 shows the signal of "Upwards" sampled at 8000Hz for 10650 samples or 1.33 seconds. Before the speaking begins,

the waveform started as silence for about 2000 samples. After the utterance, the signal remains in a silent state again for about 5000 samples. Throwing the unwanted silence region, the processing time can be improved to 3650/10650 * 100 = 34.3% by assuming all the frames in the region of interest have been processed. Fig. 5 shows the cropped signal, where the silence region has been eliminated, and the remaining regions of interest are used for further processing.

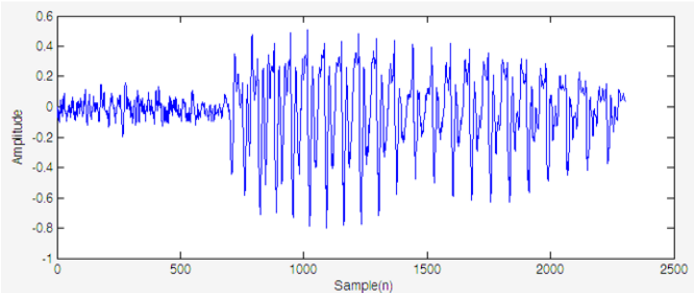

Fig. 5: Cropped Speech Signal.

Feature Extraction

The particular source-filter model used in LPC is known as the Linear Predictive Coding model. It has two key components: analysis or encoding and synthesis or decoding. The analysis part of LPC involves examining the speech signal and breaking it down into segments or blocks. All voice coders tend to model two things: excitation and articulation. Excitation is the type of sound that is passed into the filter or vocal tract and articulation is the transformation of the excitation signal into speech. The filter that is used by the decoder to recreate the original input signal is created based on a set of coefficients. Each speech segment has different filter coefficients or parameters that it uses to recreate the original sound.

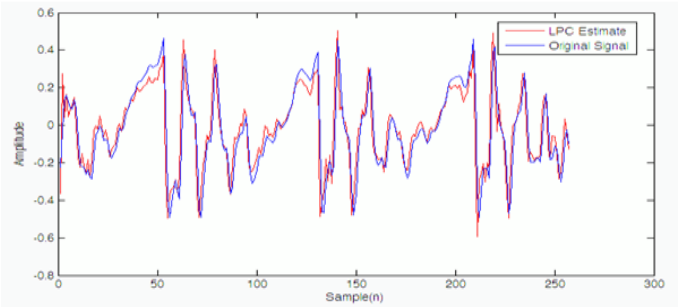

Fig. 6: LPC estimation for a speech signal frame with 256 samples

Suppose we wish to predict the value of the sample $\bar{\mathbf{s}}(n)$ using a linear combination of *N* most recent past samples. The estimate has the form:

$\bar{\mathbf{s}}(n) = a_1*s(n-1)+ a_2*s(n-2)+.....+ a_N*s(n-N)$

$$\bar{\mathbf{s}}(n) = \sum_{k=1}^{N} \mathbf{s(n-k)} * \mathbf{a_k} \qquad (1)$$

The integer *N* is called the prediction order. The estimation error is

$e(n) = s(n) - \bar{\mathbf{s}}(n)$ (2)

that is,

$$e(n) = s(n) - \sum_{k=1}^{N} \mathbf{s(n-k)} * \mathbf{a_k} \qquad (3)$$

leading to the transfer function:

$$H(z) = \frac{1}{1-\sum_{k=1}^{N} \mathbf{s(n-k)}*\mathbf{a_k}} = \frac{1}{1-P(z)} \qquad (4)$$

The LPC spectrum can be obtained by plotting the H(z) as shown in the equation mentioned above. Fig. 6 shows the LPC estimation for a frame of a speech signal with 256 sample.

We denote the average mean squared error as E(n),

$$E(n) = \sum_n \mathbf{e^2(n)} = \sum_n (\mathbf{s(n)} - \bar{\mathbf{s}}(n))^2$$

In order to provide the most accurate coefficients, $\{a_k\}$ is chosen to minimize the average value of E(n) for all samples in the segment.

$$\frac{\partial \mathbf{E(n)}}{\partial \mathbf{a_k}} = 0 \quad ; 1 \leq k \leq \mathbf{N}$$

The optimal value of a_k should be such that the error e(n) is orthogonal to s(n − k), that is,

$$\sum_n \boldsymbol{s(n-k)} * \boldsymbol{e(n)} = 0,\ 1 \leq k \leq \mathbf{N}$$

These equations have been known variously in the literature as normal equations, Yule—Walker Equations. We shall refer to them as normal equations. We can find a unique set of optimal predictor coefficients a_k. Fig. 7 shows the typical signal and the spectra for the LPC autocorrelation method for a segment of speech spoken by a male speaker. The analysis is performed using a p = 7th order LPC analysis over 256 samples at a sampling frequency of 8 KHz.

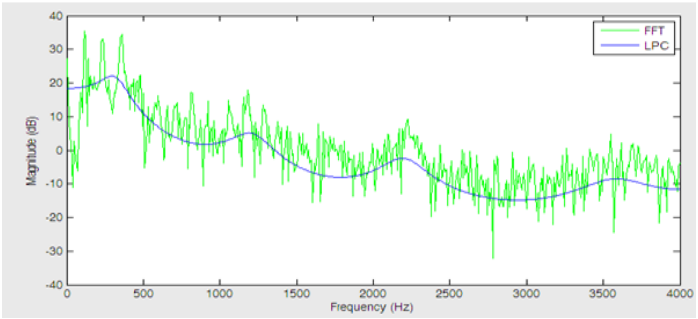

Fig. 7: Spectra for FFT and LPC autocorrelation method for a segment of speech.

In other words, the transfer function of energy from the excitation source to the output can be described in terms of natural frequencies or resonances. Such resonances are called formants of the speech. From the LPC spectral, three resonances of significance can be noticed, and named as F1, F2 and F3 respectively. Mathematically, three formants can be obtained by taking the angle of roots of the denominator in Equation 4.

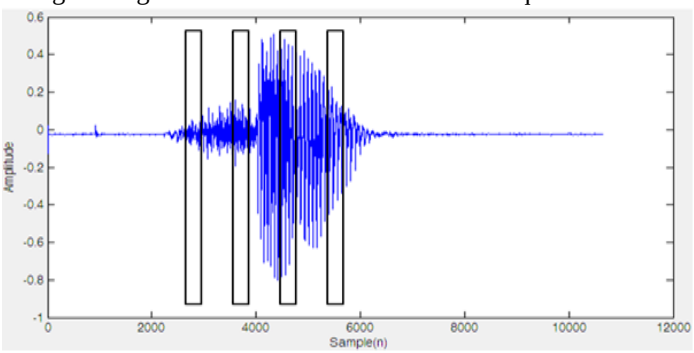

Fig 8: Selected frames for features extraction

Frame Selection

Processing all frames in the region of interest as discussed in the previous section leads to few problems. Firstly, due to the various speaking rates, the number of frames is not equal between signals. Secondly, the processing time for all frames is time consuming. In this paper, specific frames are selected to be presented to the neural network during the training process as well as during the recognition process. The frames are selected in linear distance with reference to the start point and the end point of the signal. Each frame consists of 256 samples of data. Fig. 8 shows four frames that have been selected with reference to the start point and end point. The LPC coefficients of the selected frames are used as the inputs for the neural network which will be discussed in the next section.

Neural Networks In Speech Recognition

Multi-layer perceptrons are one of many different types of existing neural networks. They comprise a number of neurons

connected together to form a network. The "strengths" or "weights" of the links between the neurons is where the functionality of the network resides. Its basic structure is shown in Fig. 9.

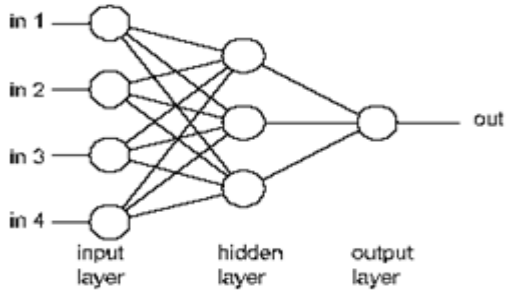

Fig. 9: Structure of a multi-layer perceptron.

The idea behind neural networks stems from studies of the structure and function of the human brain. Neural networks are useful to model the behaviours of real-world phenomena. Being able to model the behaviours of certain phenomena, a neural network is able to subsequently classify the different aspects of those behaviours, recognise what is going on at the moment, diagnose whether this is correct or faulty, predict what it will do next, and if necessary respond to what it will do next.

Feed-forward networks [17] often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. Fig. 10 illustrates the structure of the neural network in this project. The inputs of the network are the features extracted from the selected frames. The features can be the LPC coefficients or the first three formants of each frame.

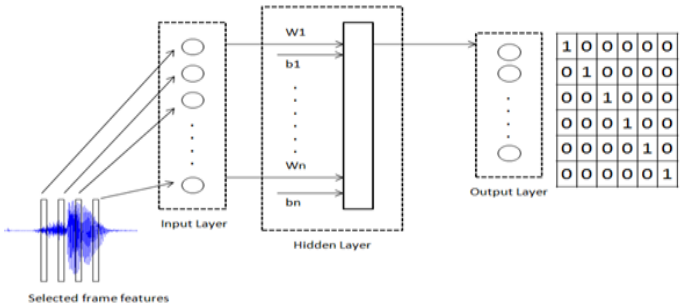

Fig. 10: Neural Network Architecture for speech recognition

The whole database consists of:

i) 20 different speakers speaking at different rates

ii) 6 words, for each speaker

iii) Total number of utterances: 20x6 = 120 utterances

The database is split into two groups, one for training the neural network, the other for testing the performance of the trained neural network. The first group, training database, comprises 70% of the speakers' utterances and the second group, testing database, is comprised of the remaining utterances.

One of the common problems when using Multilayer Perceptrons is how to choose the number of neurons in the hidden layer. There are many suggestions on how to choose the number of hidden neurons in Multilayer Perceptrons. For example, the minimum number of neurons, h, can be:

$$h \geq \frac{p-1}{n+2} \qquad (5)$$

where p is the number of training examples and n is the number of inputs of the network [19].

Results And Discussion

Neural Networks with one hidden layer with sigmoid functions and the output layer with linear functions are used in this paper. There are 6 output neurons for all the networks while the numbers of hidden neurons vary from 10 to 80. The inputs of the network are the features of 4 selected frames with 256 samples per frame i.e. 3 formants * 4 frames = 12 (3 formants per frame). Each frame is represented by either 7 LPC coefficients or the first 3 formants of the signal in the frame. All of the networks share the following common properties.

Table II: Different ANN training parameters and their corresponding value

Sl.No.	ANN parameters	Values
1.	Learning parameter	0.22
2.	Non Linear Activation Function	Tan-sigmoid
3	Maximum Epoch	1,000
4.	Number of Hidden Layer	1
5.	No of nodes in Hidden Layer	10-80
6.	Error goal	0.001
7.	Momentum	0.95
8.	Target Node	6

Table III: Performance of neural networks with different numbers of hidden neurons

Hidden Nodes	"SURU"	"ANTA"	"MATTHI"	"MUNNI"	"DAI NAY"	DEBRAY"
10	84.1	84.1	88.5	81.0	89.4	97.3
20	90.3	92.9	88.5	90.3	88.5	92.0
40	93.5	92.9	93.8	93.8	92.0	90.3
80	93.8	97.3	88.5	80.5	92.9	96.5
Mean=	**90.4**	**91.8**	**89.82**	**86.4**	**90.7**	**94**

Table 3 illustrates the results of comparing the performance of networks with different numbers of hidden layer neurons. The commands "SURU" and "ANTA" are identified more easily than other commands. So, the number of error found in BPA is less compared to other commands. The commands "MATTHI" and "MUNNI" have some spoken similarity. Thus error found here is 2 and 3 respectively out of 20 utterances of each. The left two commands are "DAINAY" and "DEBRAY" where we get 2 and 1 errors respectively out of 20 utterances of each.

On successful completion of each iteration of the audio processing, one byte of data is send to the predefined COM Port to which XBee configured as transmitter is connected. The transmitter then sends them to the receiver side. Depending on the data byte received in the UART buffer, microcontroller operates the robot in specified direction. When there is no byte present in the UART buffer of the microcontroller, it continuously waits for the next data.

The performance of the system also improved with the increasing of the number of neurons in the hidden layer used to train the network. Total input utterance that is measured is 120. Among them the error count is 12. So, the average efficiency given by the system is 90%. Finally, the number of neurons in the hidden layer also affects the performance of the system.

Conclusion And Future Scope

In this paper, the approach of using neural networks for speaker independent isolated word recognition has been studied. We have collected voice commands from 20 persons including both male and female in our databases. But this database is used in both training and testing purpose after pre-processing this speech commands. Finally we get a recognition system with 90 % efficiency.

For large vocabulary systems, this approach can also work together with other models to achieve higher accuracy. For example, it can be modified in order to work with Hidden Markov Models (HMM) to improve the performance of the recognition system [18, 20].

Other training methodology can also be used in place of BPA. Thus we can increase the efficiency of the system. The accuracy can also be improved by increasing the number of speaker for the training.

References

[1] L. R. Rabiner, B. H. Juang, *Fundamental of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[2] Zebulum, R.S., Vellasco, M., Perelmuter, G. and Pacheco, M.A. "A comparison of different spectral analysis models for speech recognition using neural networks.",IEEE , 1996.

[3] P. G. J. Lisboa, *Neural Networks Current Application*. Chapman & Hall, 1992.

[4] Campbell, Joseph P., "Speaker Recognition: A Tutorial", IEEE , VOL. 85, No. 9, pp. 1437-1462, September 1997.

[5] Chakraborty, P., Ahmed F., Kabir Md. Monirul, Shahjahan Md. and Murase Kazuyuki, "An Automatic Speaker Recognition System", Springer-Verlag Berlin Heidelberg, M. Ishikawa et al. (Eds.): ICONIP 2007, Part I, LNCS 4984, pp. 517–526, 2008.

[6]. J. Harrington and S. Cassidy, "Techniques in Speech Acoustics", Kluwer Academic Publishers, ordrecht, 1999.

[7] Shukla, A.,Tiwari R., "A novel approach of speaker authentication by fusion of speech and image features using Artificial Neural Networks", International Journal of Information and Communication Technology, Vol.1,No.2 pp . 159 – 170, 2008.

[8]. Chandra, E. and Sunitha, C., "A review on Speech and Speaker Authentication System using Voice Signal feature selection and extraction", Advance Computing Conference , 2009. IACC 2009. IEEE International, pp 1341 – 1346, March 2009.

[9] B. A. St. George, E. C. Wooten, L, Sellami, "Speech Coding and Phoneme Classification Using MATLAB and Neural Works", in *Education Conference*, North-Holland University, 1997.

[10] M. Nakamura, K. Tsuda, J. Aoe, "A New Approach to Phoneme Recognition by Phoneme Filter Neural Networks". *Information Sciences Elsevier*, Vol. 90, 1996, pp. 109-119.

[11] J. T. Jiang, A. Alwan, P. A. Keating, E. T. Auer L. E. Jr, Bernstein, "On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics". *EURASIP Journal on Applied Signal Processing*, Vol. 11, 2002, pp. 1174-1188.

[12] X. Z. Zhang, C.C. Broun, R. M. Mersereau, M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces". *EURASIP Journal on Applied Signal Processing*, Vol. 11, 2002, pp. 1228-1247.

[13] Fatima, N. and Zheng, T.F. "Short Utterance Speaker Recognition A research Agenda", Systems and Informatics (ICSAI), 2012 International Conference, pp 1746 - 1750, May 2012.

[14] H. Demuth, M. Beale, *Neural Network Toolbox*. The Math Works, Inc., Natick, MA, 2000.

[15] Shukla, A.,Tiwari R., "A novel approach of speaker authentication by fusion of speech and image features using Artificial Neural Networks", International Journal of Information and Communication Technology, Vol.1,No.2 pp . 159 – 170, 2008.

[16] L. R. Rabiner, M. R. Sambur, "An Algorithm for Determining the Endpoints for Isolated Utterances". *The Bell System Technical Journal*, Vol. 54, No. 2, 1975, pp. 297-315.

[17] S. Haykin, *Neural Networks, A Comprehensive Foundation*. Prentice Hall, New Jersey, 1999.

[18] Zhen, B., Wu, X. and Chi, H., "On the Importance of Components of the MFCC in Speech and Speaker Recognition", Center for Information Science, Peking University, China, 2001.

[19] N. K. Kasabov, *Foundations of Neural Network, Fuzzy Systems, and Knowledge Engineering*. The MIT Press Cambridge, London, 1996.

[20] T. F. Li, "Speech Recognition of Mandarin Monosyllables". *The Journal of the Pattern Recognition Society*, 2003.