AN ANN BASED MOBILE ROBOT CONTROL THROUGH VOICE COMMAND RECOGNITION USING NEPALI LANGUAGE

Neerparaj Rai[1] and Bijay Rai[2]

[1]Department of Electrical and Electronics Engineering, Sikkim Manipal University, Sikkim, India
[2]Department of Electrical and Electronics Engineering, Sikkim Manipal University, Sikkim, India

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). LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modelled signal is called the residue. Back propagation method is used to train the ANN. For each voice signal, LPC method produces 256 data in a frame. Then, these data become the input of the ANN. The ANN was trained by using 120 data training. This data training includes the pronunciation of the six words used as the command, which are created from 20 different people. Experimental results show that the highest recognition rate that can be achieved by this system is 94%.

KEYWORDS

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

1. INTRODUCTION

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 [13]. Regional language like Nepal or other can be used here in the form of normal speech or music [4]. But if collected database is small for training and testing then short utterances speech recognition (SUSR) technique can be used [1]. Authentication issue can be implemented also [16, 3]. 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 [3]. 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 [2].

Back-propagation method is used to train the Artificial Neural Network (ANN) for speech recognition. The features extracted using LPC is used to train the network. The system was built using MATLAB [10] 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.

2. METHODOLOGY

The most important elements used in this paper consist 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 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.

2.1 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.

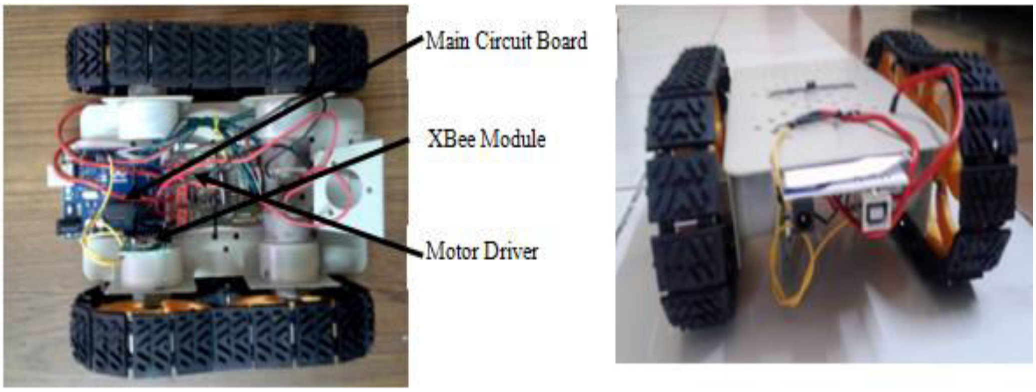

Fig. 1: (a) Bottom View (b) Side View

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).

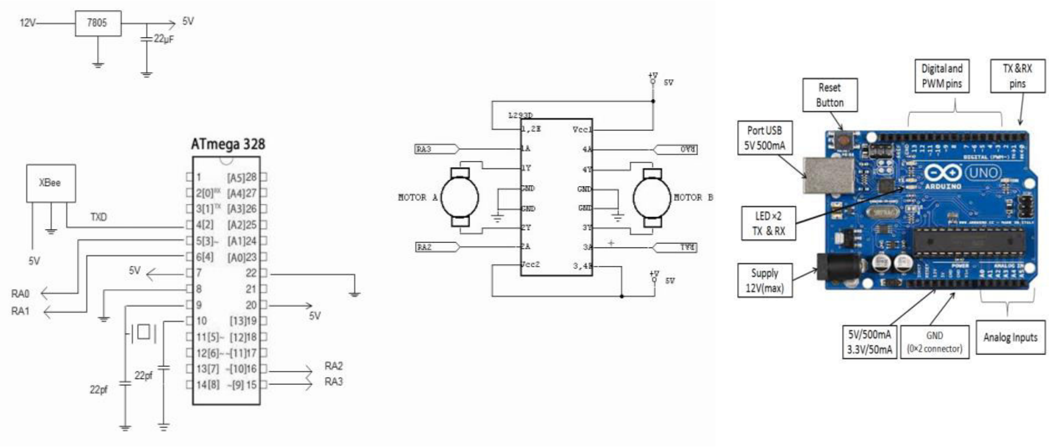

Fig. 2: (a) Main board basic schematic circuit. (b) Main board circuit.

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.

2.2 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 1. 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

2.3 End detection

The implementation for the speaker verification system first addresses the issue of finding the endpoints of speech in a waveform. In the end point detection technique, the average energy level of the signal is calculated to detect the zero crossing of the signal. Fig. 4 shows the energy plot of "Upwards".

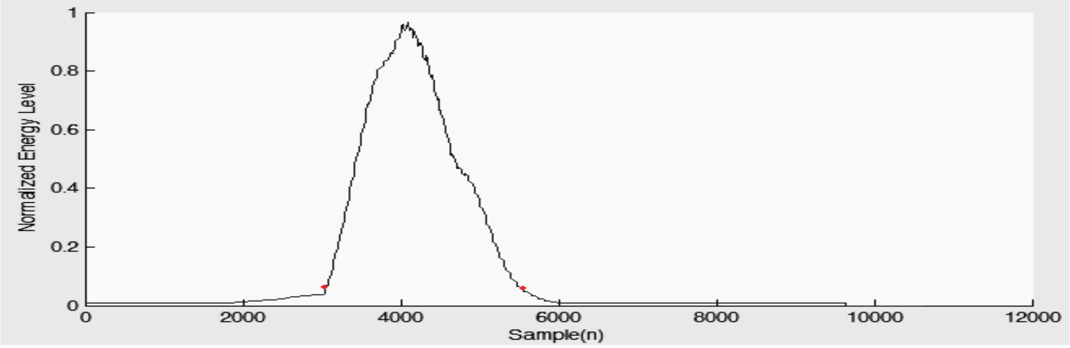

Fig. 3: Original Speech Signal.

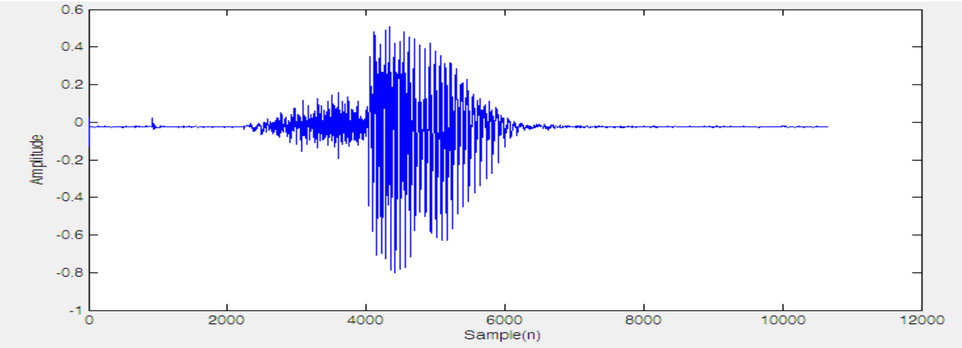

Fig. 4: End point detection using energy level of the speech signal.

Fig. 3 shows the signal of "Upwards" sampled at 8000Hz for 10650 samples or 1.33 seconds. The total frame size is of 10650 samples, but after end point detection voiced part is retained which consists of only 3650 samples. This also improves the processing time and reduces the idle time.

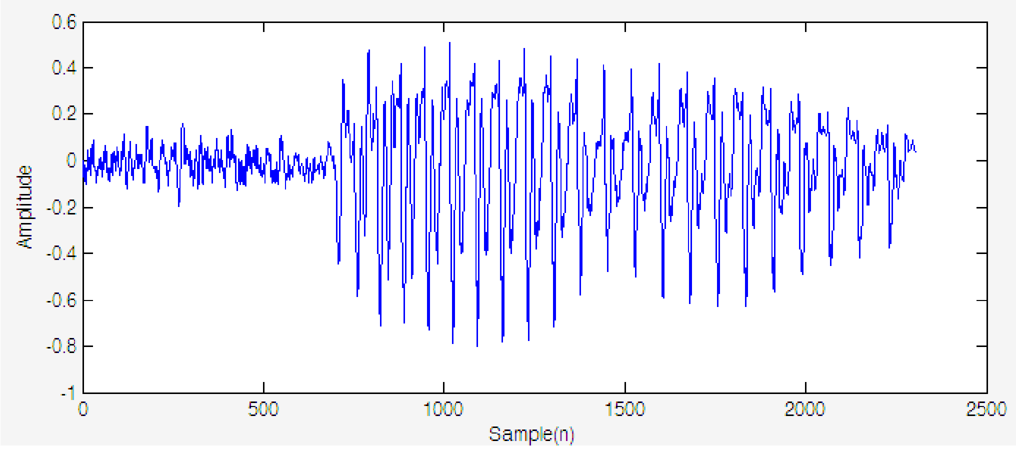

Fig. 5: Cropped Speech Signal.

2.4 Feature extraction

The formants for the speech signal are extracted using inverse filtering so as to minimize the mean squared error. The mean squared error is expressed in equation (4).

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

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

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

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

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

that is,

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

leading to the transfer function:

$$H(z) = \frac{1}{1-\sum_{k=1}^{N} s(n-k) * a_k} = \frac{1}{1-P(z)} \quad (3)$$

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.

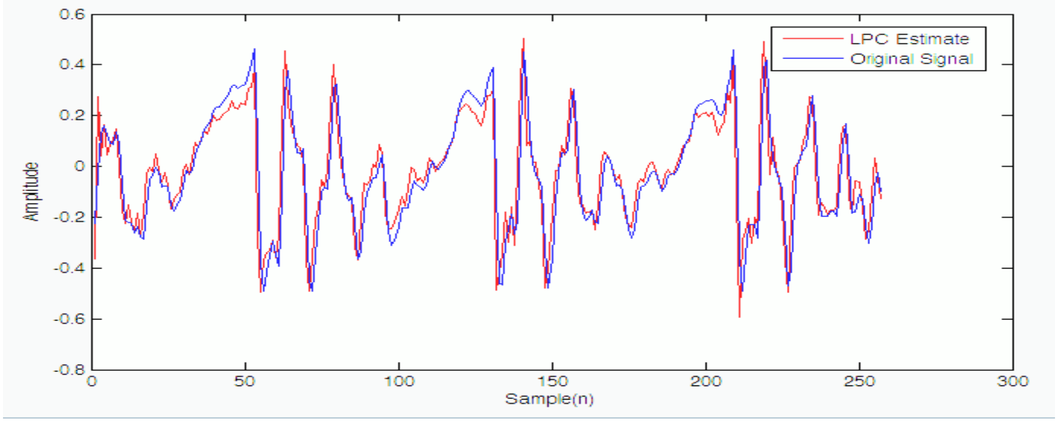

Fig. 6: LPC estimation for a speech signal frame with 256 samples

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

$$E(n) = \sum_n e^2(n) = \sum_n (s(n) - \bar{s}(n))^2 \quad (4)$$

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.

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

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

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

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.

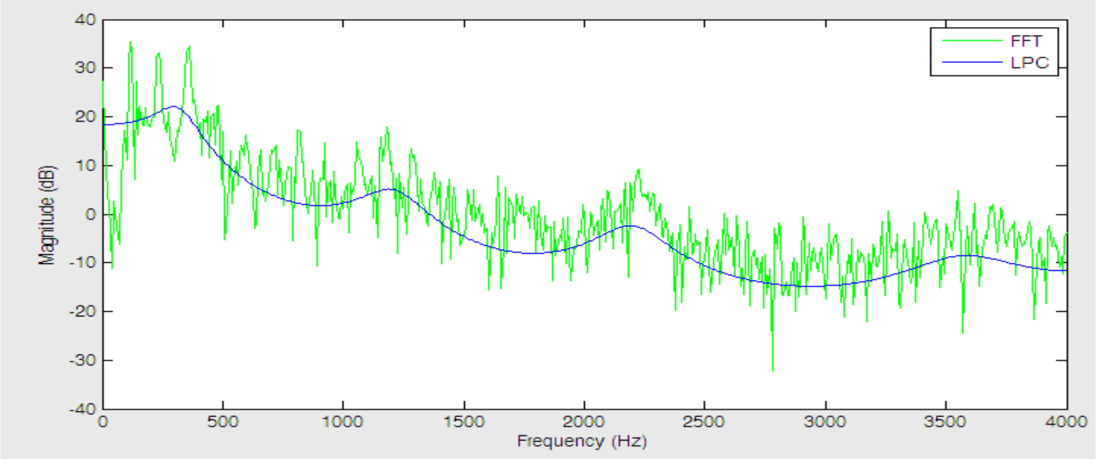

Fig. 7: Spectra for FFT and LPC autocorrelation method for a segment of speech.

Resonance frequency called formants of speech can be noticed in Fig. 7. Equation (3) is used to obtain the formants by find the roots of the denominator.

2.5 Frame selection

In this section, frames of 256 samples are selected to obtain the formants. Three formants are selected for each frame for input to the neural network. A total of 4 frames are selected from each speech signal. Therefore a total of 12 formants are prepared for the neural network input.
Fig 8 shows the selection of four frames starting from the start of the speech at regular intervals.

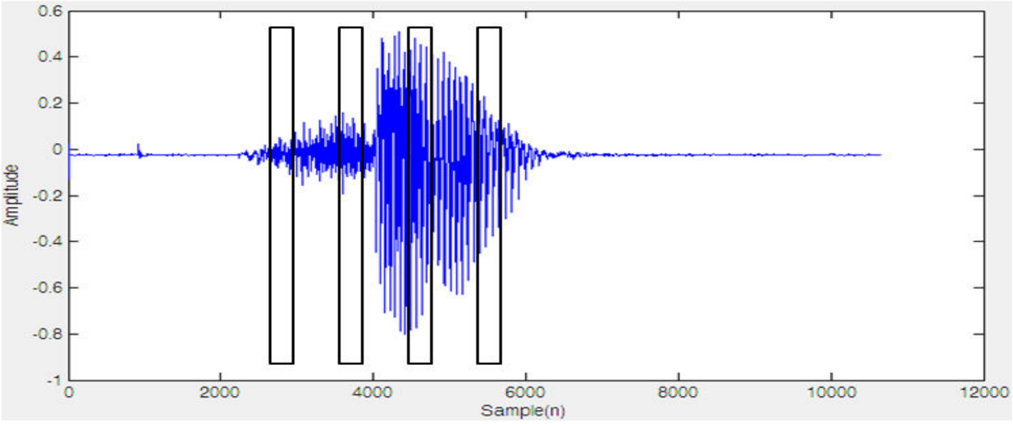

Fig. 8: Selected frames for features extraction

3. NEURAL NETWORKS IN SPEECH RECOGNITION

An Artificial Neural Network is used as recognition method. Architecture of ANN used in this system is a multilayer perceptron neural network. Its basic structure is shown in Fig. 9.

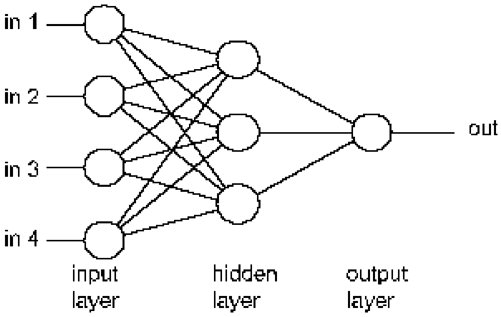

Fig. 9: Structure of a multi-layer perceptron.

Feed-forward networks [12] 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 features can be the LPC coefficients or the first three formants of each frame.

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

About 70% of the database is used for training the multilayer neural network and the rest is used for testing.

There is always a problem to justify the correct number of hidden neurons.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 [15].

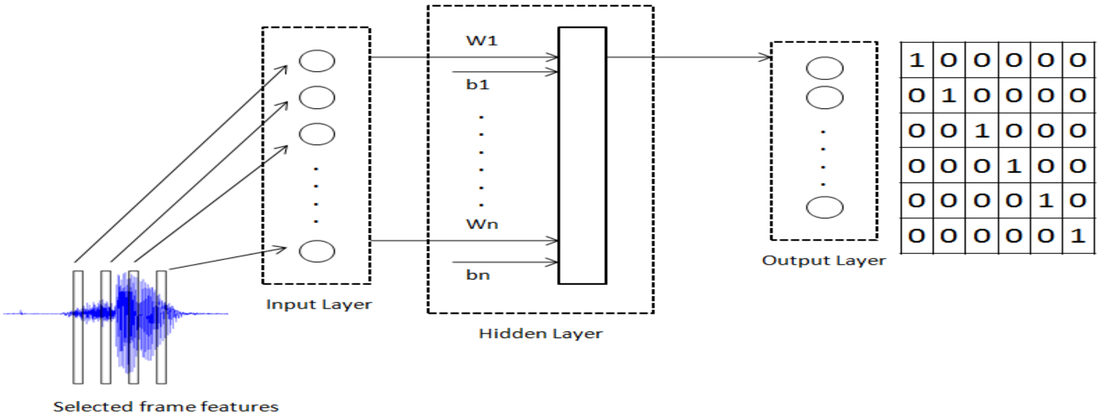

Fig. 10: Neural Network Architecture for speech recognition

4. RESULTS AND DISCUSSION

The neural network consists of 12 input neurons for the 12 formants selected from 4 frames. Only one hidden neuron is used with 10-80 neurons with sigmoid function. For the six voice commands, 6 output neurons are used with linear activation function. The parameters for ANN training are listed in Table 2.

Table 2. 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 3. Performance of neural networks with different numbers of hidden neurons

Hidden Nodes	“SURU”	“ANTA”	“MATTHI”	“MUNNI”	“DAINAY”	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.

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%.

5. CONCLUSION AND FUTURE SCOPE

In this paper, speech recognition using ANN approach 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.

The recognition can further be improved by increasing the number of hidden neurons. Maximum of 94% recognition was obtained which can still further be increased by using other recognition method such as neuro-fuzzy, fuzzy methods. The number of hidden layer can also be altered to notice the behaviour of the recognition system.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

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

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

[16] 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.

[17] 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.

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

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

Authors

Neerparaj Rai was born on the 24th of January1987, in West Bengal, India. In 2009, he completed his four year course in Electrical and Electronics Engineering at Sikkim Manipal University. In July 2013, he received the M.Tech. Degree from Sikkim Manipal University. He joined the Electrical and Electronics Engineering Department as the faculty of Science and Technology at Sikkim Manipal University, where he currently works as Assistant Professor. He is responsible for courses on Basic Engineering, Circuits and Networks, Microprocessors and Computer Architecture. His current research interests focus on data integration in computer vision systems, Robotics and Artificial Intelligence.

Bijay Rai was born on September 29, 1987, in West Bengal, India and has a M.Tech degree on Electrical Engineering at Sikkim Manipal University, specialization in Power Electronics. His research area is Computer Vision and Robotics. He teaches several engineering courses at the Electrical Engineering Department, Sikkim Manipal Institute of Technology, India. He is responsible for courses on Computer Vision, Robotics, Microprocessors and Digital Systems.