An Efficient ANN based Human-Machine Interaction through Voice Command Recognition using Bengali Language

Tiyasa Chakraborty[1], Sanjay Kumar Singh[2], Prateek Agrawal[3] and Saruchi[4]
Department of Computer Science Engineering[1 2 3]
Department of Computer Application[4]
Lovely Professional University, Jalandhar, INDIA
Email: tiyasachakraborty@gmail.com, Email: sanjayksingh.012@gmail.com,
Email: prateek061186@gmail.com, & Email: ganpati.saruchi@gmail.com

***Abstract*- The analog speech signal can be used for interacting with machines or computers. In this case, machine should be capable to recognize the voice commands that are given as inputs to that machine. 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 Bengali Languages. In our paper, four speech features i.e. LPC, LPCC, MFCC and FORMANTS are used. These features are further applied to Artificial Neural Network (ANN) model. This model is trained by using Back Propagation Algorithm. Six different words are taken from each individual as a form of commands in Bengali Language to get best voice command recognition accuracy. The accuracy of recognizing these voice commands, when tested on neural network, is about 90.83%.

***Keywords*- Formants Frequencies (FORMANTS), Linear Prediction Cepestral Coefficient (LPCC)Linear Prediction Coefficient (LPC), Mel-Frequency Cepestral Coefficient (MFCC),and Multilayer Perceptron (MLP).**

I. Introduction

Voice is the main communication medium for human being. Similarly, it also can be used for user friendly communication between human and computing system. In this case, machine should understand voice signal. Here, voice signal understanding is done in the form of Voice Command Recognition. To make this user-friendly interface option more efficient machine should understand the voice commands in regional languages also. So here, we are taking voice command in one of the Indian regional language i.e. Bengali. Automatic Voice Recognition System can be developed using Neural Network where small, medium or large vocabulary can be recognized [3][10].To develop such an interface for traditional computing system both voice processing and voice recognition system are required.

A. Speech Recognition

Speech Recognition is the subfield of Artificial Intelligence in modern computing studies [2]. It can be divided into two subcategories that are Text Dependent Speech Recognition System and Text Independent Speech Recognition System.

In Text Dependent Speech Recognition System, same dataset is used for both training and testing purpose. It provides maximum accuracy as unknown situation is not handled in this case. System only implemented according to various speech samples taken in the database [9].

In Text Independent Speech Recognition System, it is able to deal with unknown situation also as training and testing samples are different here. One approach is to use PCA for feature detection following by training through NN model [12]. Speech Recognition System can be used for different purpose like Speaker Recognition, Gender Recognition, Language Recognition and Word Recognition. Here, voice command recognition is done as a form of word recognition.

1) Speaker Recognition

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 Bengali 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 [1, 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]. Here, Pitch should be detected also [15]. Actually the speaker is recognized in this phase. In Speaker verification step, this particular recognized speaker is verified or authenticated. One approach can be the use of Auto Associative Neural Network (AANN) model for training and that subsequently can be designed to implement the verification system [14].

ISBN: 978-81-923777-9-7

2) Gender Identification

Gender Identification is another application of speech recognition system. Here, System is implemented based on a binary decision that is whether the speaker is male or female. Different types of classifier can be used for this purpose like Vector Quantization, Multi-Layer Perceptron, Gaussian Mixture Model [4, 16].

3) *Language Recognition*

In Language Recognition System should able to recognize one particular language or multiple languages in the case of multilingual system. Different algorithm can be used for this purpose to improve the accuracy of the recognition system [17].

4) Word Identification

Now in this paper the Word Identification is explored. In word recognition system, sub word units can be generated using dictionary based table [6].Word can be recognized from handwritten text also where word images are used for extraction of feature vectors [18]. But this paper is an attempt to recognize words as a form of speech commands where database is also taken in the form of speech samples. In the case of pattern classification in signal processing, Vector Quantization Technique can be used also where extracted features are stored in codebook [11, 15]. This codebook is used to make the cluster of training vector during training period and then matched with test vectors during testing period [5].

Voice command recognition is thus a form of Word Identification. These words that are chosen for the recognition purpose can be of any languages.

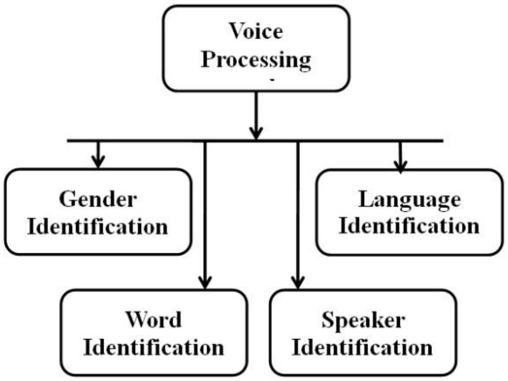

Fig. 1: Voice signal Processing

B. *Artificial Neural Network*

Artificial Neural Network plays very important role in the Implementation of Speech Recognition System. Artificial Neural Network provides a model that gives the output based on some knowledge. This knowledge is acquired by this model previously through learning. Our system can be able to deal with a completely new situation also by a systematic adaptation and generalization process. In speech recognition system, we use complex artificial neural network model where hidden layers act as main computing element.

In this paper, we will use Back Propagation Algorithm for learning purpose. After learning through BPA algorithm the system can deal with unknown data also. In speech recognition automatic machine learning approach can be implemented through spectral analysis, feature extraction and training by BPA [19]. The back propagation algorithm actually finds the minimum amount of the error in terms of error function. To find this minimum amount of error the gradient descent method is used. The summation of present weights is actually used to reduce the error function. Thus the learning problem is considered to be solved by this reduction method. Since at each iteration step, this learning process actually requires calculation of the gradient of the particular error function.

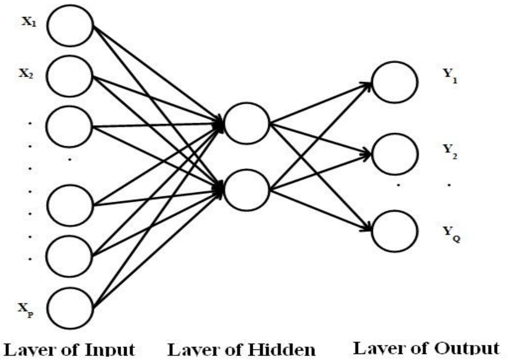

Fig. 2: Non Linear Feed Forward Neural Network Architecture

II. Methodology

A. *Diagrammatic representation of implementation methodology*

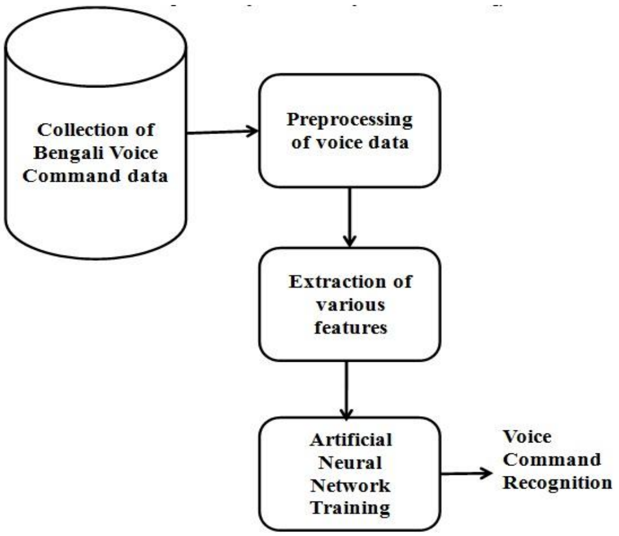
Collection of Bengali Voice Command data
Preprocessing of voice data
Extraction of various features
Artificial Neural Network Training
Voice Command Recognition

Fig. 3: Steps for Implementing Voice Command Recognition System Using ANN

Our Methodology includes various steps such that Collection of Voice Command data, Preprocessing of voice data, Extraction of various features, Artificial Neural Network Training, Voice Command Recognition.

ISBN: 978-81-923777-9-7

B. *Data Base Preparation*

Here speech commands are taken in Bengali language for our database generation. There are total six speech commands that are SURU, SESH, DAAN, BAAM, UPORE and NICHE. The following table is showing the different Bengali 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 44.100 KHz sampling rate. After recording (.wav) sound file is generated. Here we are considering 40 (20 male+20 female) different individuals for recording and creating the Speech Command Database.

TABLE I: MEANING OF DIFFERENT BENGALI VOICE COMMANDS

SL No.	**Voice Command**	**Meaning in English**
1	Suru	Start
2	Shesh	End
3	Daan	Right
4	Baam	Left
5	Upore	Top
6	Niche	Bottom

C. *Preprocessing*

In this step the recorded data is actually preprocessed for accurate and efficient speech command recognition. Preprocessing is done for mainly two purposes. One is noise removal and another one is blank or unnecessary section removal from each speech command sample that is recorded. It is done because our objective is to create approximate noise free database for our recognition system.

D. *Feature Extraction*

In normal speech recognition system as well as in audio-visual speech recognition system feature extraction is an important factor. There are various types of features available but here only 4 different features are used. They are Linear Predictive Coefficient (LPC), Formant Frequencies (FORMANTS), Mel-Frequency Cepestral Coefficients (MFCC) and Linear Predictive Cepestral Coefficient (LPCC). Among those, LPC & LPCC are commonly known as LPC and LPC derived features. Features can be applied on VQ, SVM or ANN model for training of the system [20].

1) *LPC and LPC derived features*

This technique is very famous for feature extraction in the field of modern digital signal processing. By using this technique we can easily find various features form our voice commands according to peach, word structure etc.

Each sample of speech signal is actually the weighted sum of q previous voice command sample and an excitation [21]. Here, the prediction error is reduced for calculating LPC coefficient by using least squares sense. In time domain, the LPC speech production can be written as [21, 22];

$$Sp[m] \approx \sum_{i=1}^{q} a[i]Sp[m-i] \qquad (1)$$

Where *Sp[m]* denotes sample of voice signal, predictor coefficients is denoted by $a[i]$ and the order of predictor is represented by q. Now, the weighted sum of squared prediction error can be represented by;

$$Er = \sum_{m} \left(Sp[m] - \sum_{i=1}^{q} a[i]Sp[m-i]\right)^{2} \qquad (2)$$

Here, for each speech frame the coefficient $a[i]$ is calculated so that error should be reduced. For this purpose, we can take the partial derivation of eq. (ii) w. r to $a[i]$ tends to zero. In frequency domain, LPC can be represented by;

$$Hz(z) = \frac{1}{1 - \sum_{i=1}^{q} a[i]z^{-i}} \qquad (3)$$

2) *MFCC*

Mel-Frequency Cepestral Coefficient is an efficient method for extraction of speech feature vector from given speech commands. It is a benchmarking method for speech and speaker recognition system. MFCC can be calculated based on Hamming Window, Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT). If the phase information is combined with traditional MFCC, it can provide much more efficiency in speaker identification and verification process in leads to text independent speech recognition system [23].

3) *Formats*

Formants are defined by the peaks that can be determined by the spectral analysis of the voice signal spectrum. Formant can also be used for the acoustic resonance in phonetic that is actually the resonance of vocal tract of normal human being. It can be calculated by the peak in the frequency spectrum of the voice command, using a particular spectrum analyzer [22].

E. *Artificial Neural Network*

We are using Artificial Neural Network to train our System. The feature vectors that are calculated in previous Feature Extraction step, are organized in a proper format in matrix form. This organized data is used to train the model by using Back Propagation Algorithm which is the main training algorithm of Artificial Neural Network. Our Neural Network Model consists of 1 hidden layer with 24 hidden neurons. Different learning parameters are shown in the following table:

TABLE II: DIFFERENT ANN TRAINING PARAMETERS AND THEIR CORRESPONDING VALUE

SL. No.	ANN Parameters	Values
1	Learning Parameter(α)	0.22
2	Non Linear activation Function	Tan-sigmoid
3	Maximum Epoch	1,000
4	Number of total hidden layer	1
5	No of Nodes in hidden layer	24
6	Error goal(δ)	0.001
7	Momentum	0.95
8	Target Node	6

III. Results and Discussion

Our whole system can be trained through Artificial Neural Network by using BPA Algorithm. Here we have 40 different people consist of both male and female categories. To implement the database each of six commands is spoken by each individual. Thus we can get 40 input utterances for each of our voice command and Total 240 input utterances. In this recorded command, the commands ‘NICHE’ and ‘UPORE’ are identified more easily than other commands. So, the number of error found in BPA is less compared to other commands. The commands ‘DAAN’ and ‘BAAM’ have some spoken similarity. Thus error found here is 6 and 5 respectively out of 40 utterances of each. The left two commands are ‘SURU’ and ‘SHESH’ where we get 5 and 3 errors respectively out of 40 utterances of each.

TABLE III: TABULAR REPRESENTATION OF ERROR COMPUTED BY BPA AND CORRESPONDING EFFICIENCY CALCULATION

SL. No. &Word	No. of Input Utterances	No of Errors (BPA)	Efficiency (%) (BPA)
1. (Suru)	40	5	87.5
2. (Shesh)	40	3	92.5
3. (Daan)	40	6	85.0
4. (Baam)	40	5	87.5
5. (Upore)	40	1	97.5
6. (Niche)	40	2	95.0
Total	∑= 240	∑ =22	∑=90.83

Finally, Total input utterance that is measured is 240. Among them the error count is 22. So, the average efficiency given by the system is 90.83%. Here the System is fully Text Dependent as here the training and testing data is not Different.

IV. Conclusion And Future Scope

In this paper, we are introducing a word recognition system that in Text dependent as testing and training samples are not different. We are taking voice commands from 40 persons including both male and female in our databases. But this database is used in both training and testing purpose after preprocessing this speech commands. Finally we get a Text Dependent System with 90.83 % efficiency.

In future we should work on a text independent system by using different testing and training sample. Other training methodology can be used also 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] 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.

[2]. J. Harrington and S. Cassidy, “Techniques in Speech Acoustics”, Kluwer Academic Publishers, ordrecht, 1999.

[3] Love, Brian J., Vining J. and Sun X., “Automatic Speaker Recognition Using Neural Networks”, EE371D Intro. To Neural Networks, Electrical and Computer Engineering Department, The University of Texas at Aus, Spring 2004.

[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] Singh, R., Bhiksha, R., and Richard, M., “Automatic Generation of Sub word Units for Speech Recognition Systems”, IEEE Transactions on Speech and Audio Processing, VOL. 10, NO. 2, February 2002.

[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]. Firoz, S.A., Raji, S.A. and Babu A.P.,“Speaker and Text Dependent automatic emotion recognition from female speech by using artificial neural networks”, Nature and Biologically Inspired Computing, 2009. NaBIC 2009.pp 1411 – 1413, December 2009.

[10] 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.

[11] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," Proc. IEEE International Conference, Acoustics, Speech Signal Processing, S6.3, pp. 377-380,1991.

[12] Lu Xiao-chun, Yin Jun-xun and Hu Wai-ping,"A text-independent speaker recognition system based on Probabilistic Principle Component Analysis", System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2012 3rd International Conference, Vol.1 pp 255 – 260, 2012.

[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] Garimella, S., Mallidi, S.H. and Hermansky, H., "Regularized Auto-Associative Neural Networks for Speaker Verification", Signal Processing Letters, IEEE, Vol. 19 Issue: 12, pp 841 – 844, December 2012.

[15] Chougule, S. and Rege, P., "Language Independent Speaker Identification", Industrial Technology, 2006. ICIT 2006. IEEE International Conference, pp 364-368, December 2006.

[16] Djemili, R., Bourouba, H. and Korba M.C.A, "A speech signal based gender identification system using four classifiers", Multimedia Computing and Systems (ICMCS), 2012 International Conferences, pp 184 – 187, May 2012.

[17] Hategan, A., Barliga, B. and Tabus, I., "Language identification of individual words in a multilingual automatic speech recognition system", Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, IEEE International Conference, pp 4357 – 4360, April 2009.

[18] Bhowmik, T.K., Roy, U. and Parui, S.K., "Lexicon Reduction Technique for Bangla Handwritten Word Recognition", Document Analysis Systems (DAS), 2012 10th IAPR International Workshop, pp 195-199, March 2012.

[19] Dey, N.S., Mohanty, R. and Chugh, K.L., "Speech and Speaker Recognition System Using Artificial Neural Networks and Hidden Markov Model", Communication Systems and Network Technologies (CSNT), 2012 International Conference, pp 311-315,May 2012.

[20] Goyani, M., Dave, N. and Patel, N. M., "Performance Analysis of Lip Synchronization Using LPC, MFCC and PLP Speech Parameters", Computational Intelligence and Communication Networks (CICN), 2010 International Conference, pp 582-587, November 2010.

[21] Ranjan, R., Singh, S. K.,Shukla, A. and Tiwari, R."Text-Dependent Multilingual Speaker Identification for Indian Languages using Artificial Neural Network", ICETET-IEEE, pp.632-635, November 2010.

[22] Agrawal, P., Shukla, A. and Tiwari, R., "Multilingual Speaker Recognition using Artificial Neural Networks", Advances in Intelligent System on Springer, pp. 1-9, 2009.

[23] Nakagawa, S., Longbiao Wang and Ohtsuka, S., "Speaker Identification and Verification by Combining MFCC and Phase Information", IEEE Transaction on Audio, Speech, and Language Processing, Vol. 20, Issue: 4, pp 1085-1095, May 2012.