Speech Controlled Robotics using Artificial Neural Network

Neha Joshi, Anil Kumar, *Student Member, IEEE*, Pavan Chakraborty, *Member, IEEE*, Rahul Kala, *Member, IEEE*
Robotics and Artificial Intelligence Laboratory,
Indian Institute of Information Technology, Allahabad, India
nitti212@gmail.com, anilkumar6079@gmail.com, pavan@iiita.ac.in, rkala001@gmail.com

Citation: N. Joshi, A. Kumar, P. Chakraborty and R. Kala, "Speech controlled robotics using Artificial Neural Network," 2015 Third International Conference on Image Information Processing (ICIIP), Waknaghat, 2015, pp. 526-530.
Full Text Available at: https://ieeexplore.ieee.org/document/7414829

***Abstract*—Humanoid robots nowadays are able to interact with humans in realtime environments. It is always desirable that robots should have similar capacity to humans for their auditory information processing and taking actions based on the same. With this concept, in this paper we proposed a nature inspired algorithm where recognition is done by using Artificial Neural Network and and the recognized word command is used to actuate the HOAP-2 Humanoid robot. In this paper we demonstrate a prototype system to control a humanoid robot using speech. The prototype can be customized for different types of tasks depending upon the utility of the robot.**

***Index Terms*—Speech Recognition, Artificial Neural Network, Deep Learning, HOAP-2.**

I. Introduction

Communication can be visual or aural. Language plays a very important role in the case of human communication. Speech is considered as the best way of interaction with the human being. As we all know speech is one of the most user friendly ways to interact with the human. There are two ways of communication - gesture based communication and speech based communication. Speech based artificial devices enable any layman to access information without any in-depth knowledge of the device configuration or process model[1]. Here we have developed a speech recognition system, initially consisting of five words to give instructions to the robot like âĂIJByeâĂİ, âĂIJRightâĂİ, âĂIJLeftâĂİ, âĂIJNamasteâĂİ (Hello in English) and âĂIJWalkâĂİ.

As a lot of work is already done in the field of speech recognition but the overall accuracy and noise removal always remains the biggest challenge. Using speech recognition techniques to solve real life applications poses more application specific challenges. There are two types of speech recognition systems - speaker dependent systems and speaker independent systems. Speaker independent system generally have low accuracy as compared to speaker dependent systems. Factors such as gender, age, emotions, and speed come into consideration while devising speaker independent systems [2]. Here we are stick to speaker dependent system.

Feature extraction is one of the important step of recognition system. Feature extraction techniques extract useful features from the raw data which helps the classifier to make decisions. Feature extraction techniques help in dimensionality reduction of the data, which is important since most of the classifiers face the problem of curse of dimensionality, wherein the classifier performance erodes dramatically with an increase in dimensions. There are many features of speech which have been used since long like Mel-frequency Cepstral Coefficients (MFCC), Linear predictive Coding (LPC), Human Factor Cepstral Coefficients (HFCC), Discrete Wavelet Transform (DWT), etc. In this paper we have used MFCC features due to their promising results in various kinds of speech recognition tasks. The classification is done by using a Artificial Neural Network. The combination is capable of showing robustness to speaker variation and environment distortion.

Neural networks and their variants have already shown some promising results in the domain of speech and speaker recognition [3][4]. Traditional neural network models and their variants have a stringent limitations of the complexity of data that they can handle. Speech and related signals are highly complicated, which limits the use of traditional neural network models. Deep neural networks use multiple hidden layers trained in an unsupervised manner, wherein each layer processes the data so as to make it easier for the subsequent layers to classify. Interest in Deep Learning is increasing attributed to their capability to handle complex data. As we are using small dataset here. So relatively good accuracy is acquired by ANN. But in future we can use deep neural network for large dataset. In Deep Neural Network a set of neurons activated by the features of input speech.

A larger problem is to use speech as a means to control a humanoid robot. A humanoid can largely be controlled using commands for different types of tasks. Speech recognition can be used as a tool to take command from the human in the form of a speech signal, recognize the command, communicate the command to the robot and then controlling the robot to perform the desired action. In this paper a prototype system has been developed wherein the speech signals are used to control a HOAP-2 humanoid robot. Every command is mapped to a robotic gesture which is independently programmed. The command invokes the desired gesture on the HOAP-2 humanoid.

Performance of the whole system controlled by preprocessing of data, feature extraction and classification[5]. Here preprocessing means dealing with the data i.e changing its size and dimension. In the first stage samples of data is taken i.e digitization of speech is done which is also known as preprocessing of speech. After that spectral analysis is done on it i.e finding features of speech signal like MFCCs. MFCC is used for building a feature vector. Extraction is

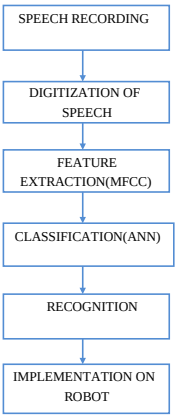

Fig. 1. Proposed Speech based Robot Control System

getting information from the speech signal. Then extracted feature is given to Artificial Neural Network classifier for training purpose. The performance of ANN is appreciable. The algorithm produces good results for increased samples and iterations of training. Artificial neural network perfectly amalgamates with MFCC features.

In the literature numerous techniques have been applied for speech recognition including Hidden Markov Models (HMM), Dynamic Time Warping (DTW), Neural Network, Deep Neural Network[6] [7]. HMM is one of the popular approach of speech recognition system. Already HMM has shown a major success in speech recognition. The drawback of HMM is its low classifying power. Whereas neural network can handle high classification power[8]. Neural network has a history of performing well in incompetency and variability of data. Here supervised learning methods are used. Input given to the system is the acoustic features of the speech signal. Speech is also affected by the emotions of the speakers[2]. A lot of research is also going on Human emotion recognition by speech. Voice authentication with speech processing can eliminate the need of a password which in turn provide high level of security[6].

This paper is organized as follows. Proposed methodology is discuss in *Section II. Section III* specifically deals with Artificial Neural Network used in this work. Finally *Section IV* shows the simulation results on HOAP-2 humanoid robot. Conclusion remarks are given in *Section V*.

II. PROPOSED METHODOLOGY

In this section, we present a general outline of our proposed methodology for speech based robotic control system. The detailed stepwise representation of system for speech recognition system is shown in *Fig* 1. However the general assumption for our training and testing signal having noise between 2-5%.

A. Speech Recording

In this initial step our assumption is that the environment ideally noise free and the signal is recorded at 16 *Khz* frequency for the duration of 3 seconds. All our training and testing speech signals have the same frequency and time frame for each speech input. Although we assume that our recording system is largely noise free, still noise error can be in the range of 2-5%. The sound signal data is written into the disk with *.wav* file format for further processing. All speech signal can represented mathematically using *equation* 1.

$$X(t) = \Sigma_{t=0}^{m} A_p(t) \cos(\omega_i(t)t + \varphi_i(t)) \quad (1)$$

where $A_p(t)$ is an amplitude, $\omega_i(t)$ is frequency, $\varphi_i(t)$ is phase difference.

B. Digitalization of Speech

As natural signal is continuous in nature, this needs to be converted into a digital one for study and experimentation purposes. This is commonly known as sampling of a signal. A matrix of 48000×1 is generated as a result of it.

C. Feature Extraction

After digitization of signal we proceed to feature extraction which is basically a method of retrieving information from the signal. Here we use MFCC as a feature vector[9]. We extract fourteen Mel Frequency cepstral coefficients from the speech. The MFCC can be calculated by using *equation* 2. For the computation of our signal 20 bandpass filters are used. Frequency to mel scale conversion formula can defined using *equation* 3. However the advantage of *equation* 2 is that our original input dimension of the signal was 48000×1 which reduces to 4470×1. Therefore further implementation on ANN will decrease the computation time.

$$M = 1127 \ln\left(\frac{f}{700} + 1\right) \quad (2)$$

$$f = 700\left(e^{\frac{M}{1127}} - 1\right) \quad (3)$$

where f represent the frequency range which will be used in the filter bank and M is a *mel* frequency.

D. Classification

In this step we use Artificial Neural Network (ANN) based machine learning techniques for multilevel pattern classification problem. The technique is nature inspired for better accuracy. The details of ANN are given in *Section III.*

E. Robot Control

The classifier's output is mapped to a robot command which should make the robot perform the desired actions. Performing the specific action is specific to the kind of action, required intelligence for the action, the robot's capability, etc. Typical to the programming of any robot, the basic task is to take the sensory inputs, process the sensory inputs to get a representation of the operational scenario, plan the robot's task based on deliberation or based on the past learning of the robot, and map the same to the robot's actuators which physically performs the task.

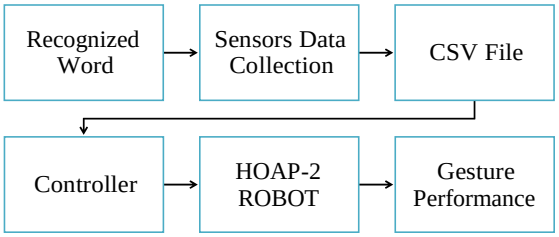

Fig. 2. Process of Gesture Recognition System for HAOP-2

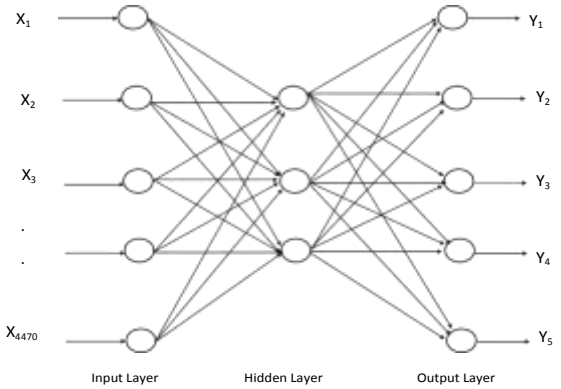

Fig. 3. ANN Network Architecture

Currently a simple system is taken wherein HOAP-2, a humanoid robot, is asked to perform a gesture as asked by the human operator. First the gestures are defined by holding the humanoid and guiding it by hand to perform the gesture. The gestures chosen are currently static and hence only the final positions are important. The same are stored in CSV(Comma-separated value) files for each gesture.

On identifying any gesture, the action specific CSV file is loaded on the robot which gives the desired positions of each joint angles. A controller is used to actuate each joint to the desired joint angle. All joints may not be needed to perform all gestures, in which case only the needed joints are activated and controlled. The overall process is shown in *Fig* 2. And all this performed in a Webot platform.

III. ARTIFICIAL NEURAL NETWORK

Speech recognition is generally a multilevel pattern recognition task, where acoustic signals are structured in sub-word units (e.g. word). ANN with *Back Propagation Neural Network* is one of the most promising approaches for speech recognition[10] and is thus used in this paper. ANNs require a training data which is used to learn the values of the weights and biases, and a testing data over which the network is tested for performance. The data sets consist of an input and an output. The input to the neural network are the MFCC coefficients.

Our speech dataset consists of multiple class so we need to define target class. Here a one-vs-all methodology has been adopted, wherein the number of output neurons are same as the number of target classes. Each output neuron tries to separate the output class from the rest of the classes. For any input, the output of the neuron is a score denoting the possibility of the input belonging to the class that the output neuron represents. The general architecture of the network is shown in *Fig* 3. The class corresponding to any input is taken as the class which gets the highest score. Back Propagation Algorithm (BPA) is used as a training algorithm for ANN. First Feed Forward stage is applied wherein an input is fed to the network and the output is computed at each iteration m. Then the errors are computed and back propagated in the network to adjust the weights and biases. Let the weights of connecting neuron i and neuron j be $W_{ij}(m)$. Also assume the output from neuron j be Y_j. Then error can be calculated from *equation* 4.

$$E_j(m) = D_j(m) - Y_j(m) \tag{4}$$

where E is error, D represent desired output and Y is actual output. By using *equation* 5 weights can updated.

$$W_{ij}(m+1) = W_{ij}(m) + \alpha\delta_i(m)\,Y_j(m) \tag{5}$$

In *equation* 5 α is constant real number having value less then 1 and also called the learning rate. The δ_j can define using *equation* 6 and *equation* 7, where φ represent transfer function.

$$\delta_j(m) = E_j(m)\,\varphi_j'(V_j(m)) \tag{6}$$

$$V_j(m) = \sum_{\forall i} W_{ij}(m)\,Y_j(m) \tag{7}$$

The ANN network was implemented through using MATLAB R2013a toolbox. The ANN was trained with input matrix. The network which close to the target class during the testing of network considered as the "recognized" word if they belong to same class.

IV. EXPERIMENTS AND RESULTS

The testing of the algorithm was done on a self made dataset. The ANN was trained and the trained networks were asked to classify the speech input and accordingly the HOAP-2 robot was made to perform the necessary action. The speech dataset consisted of 5 different classes having 15 samples each. The final recognized command is performed on a simulated humanoid robot HOAP-2 using Webots simulator [11] . WEBOTS is one of the most powerful integrated environments which allows us to model our complex setups with single or several robots in artificial environments. All our experiments are performed on humanoid robot HOAP-2 [12]. HOAP-2 is human-like physical robot with 25 joints or DOF (Degree of

Gesture class	Bye	Right	Left	Namaste	Walk
Bye	100%	0%	0%	0%	0%
Right	0%	92.8	7.2%	0%	0%
Left	0%	0%	93.3%	6.7%	0%
Namaste	0%	6.7%	0%	93.3%	0%
Walk	5.9%	5.9%	0%	0%	88.2%

Fig. 4. Confusion Matrix for ANN with BPA

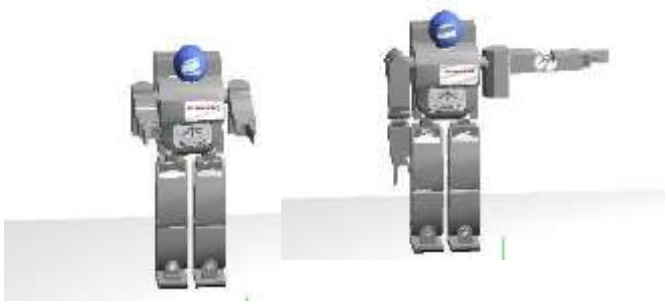

Fig. 5. Normal position Fig. 6. Control command *Right*

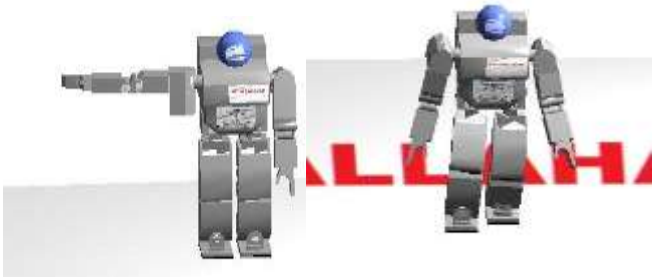

Fig. 7. Control command *Left* Fig. 8. Control command *walk*

Freedom). All training and testing experiments are performed on an Intel i5 processor having 4GB RAM.

The data is recorded for 5 different classes (*right*, *left*, *walk*, *Bye*, *namaste*) and 15 samples for each class. The size of the entire data set is thus 75×4470 which is the input to the speed recognition system. For each class 10 samples are used for training while 5 samples are used for testing. First ANN is applied to the data, on the extracted MFCC coefficients. The ANN architecture is feed forward type consisting of 1 input layer, 1 hidden layer and 1 output layer. There are 35 neurons in the hidden layer. The output layer has 5 neurons for the 5 different classes. The results of the ANN using Back Propagation Algorithm for training are given in *Fig* 4.

ANN with BPA used for classification gave an accuracy of 93.5% for the testing data set. The confusion matrix is shown in *Fig* 4. It can be easily seen the performance of ANN trained with BPA from the confusion matrix. Moreover for test of word based gesture on HOAP-2, process shown in *Fig* 5 – 10 . In *Fig* 5 HOAP-2 is initially on the normal standing mode. However for the test of recognized word using our proposed methodology in *section* 2. We setup several scenario where each gesture independently perform on HOAP-2 robot.

In the first scenario we assume that our control word is recognized *right* and we have to perform same gesture on robot. The gesture is shown in *Fig* 6. Now robot is asked to perform *left* hand up from its own current position which results in gesture shown in *Fig* 7 in second scenario. Similarly in the third scenario robot needs to perform *walk*, then corresponding *walk* gesture will be performed from the current position of the robot. The result is shown in *Fig* 8. In Fourth scenario the robot is asked to perform *Bye* gesture as our testing word which CSV file is given to HOAP-2 then result is shown in *Fig* 9. In the fifth scenario the control gesture is *namaste*. The robot will stand on its current position only and join both hands. The simulation result is shown in *Fig* 10.

V. Conclusion

Innovations in technology call for designing newer ways to interact with the robot and control it to do desirable tasks for the human master. Speech is one of the best modalities to interact with the robot and to command the robot. In this paper we used MFCC coefficients as features to train an ANN with BPA. ANN performed reasonably well to identify words from speech. The word so identified was used as a command to control a humanoid robot.

We have shown here a prototype application wherein the robot was controlled using speech of the human master. Using the application developed, a person could ask the robot to do interesting gestures. The high accuracy of recognition motivate further research in this area. In the future attempts would be to work on continuous speech recognition, making a rich set of user commands, making a vocabulary for complex robot tasks and multi-modal robot control.

References

[1] Sukumar, A.R.; Shah, A.F.; Anto, P.B., "Isolated question words recognition from speech queries by using Artificial Neural Networks," Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on , vol., no., pp.1,4, 29-31 July 2010 doi: 10.1109/ICCCNT.2010.5591733

[2] Dey, N.S.; Mohanty, R.; 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 on , vol., no., pp.311,315.

[3] A. Shukla, R. Tiwari, H. K. Meena, R. Kala (2009) Speaker Identification using Wavelet Analysis and Modular Neural Networks, Journal of Acoustic Society of India, 36(1), 14-19.

[4] Kala, Rahul, et al. "Fusion of Speech and Face by Enhanced Modular Neural Network." Information Systems, Technology and Management. Springer Berlin Heidelberg, 2010. 363-372.

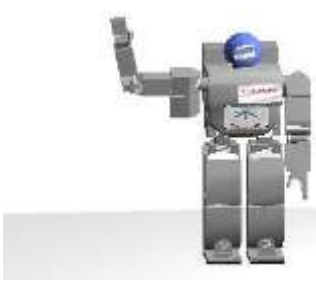

Fig. 9. Control command *bye* Fig. 10. Control command *namaste*

[5] Krishnan, V.R.V.; Jayakumar, A.; Anto, P.B., "Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network," Electronic Design, Test and Applications, 2008. DELTA 2008. 4th IEEE International Symposium on , vol., no., pp.240,243, 23-25 Jan. 2008 doi: 10.1109/DELTA.2008.88

[6] Lim, C.P.; Woo, S.C.; Loh, A.S.; Osman, R., "Speech recognition using artificial neural networks," Web Information Systems Engineering, 2000. Proceedings of the First International Conference on , vol.1, no., pp.419,423 vol.1, 2000 doi: 10.1109/WISE.2000.882421

[7] Kota, R.; Abdelhamied, K.A.; Goshorn, E.L., "Isolated word recognition of deaf speech using artificial neural networks," Biomedical Engineering Conference, 1993., Proceedings of the Twelfth Southern , vol., no., pp.108,110, 1993 doi: 10.1109/SBEC.1993.247339

[8] Botros, N., "Speech recognition using hidden Markov models and artificial neural networks," Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE , vol., no., pp.243,243, 1993 doi: 10.1109/IEMBS.1993.978523

[9] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

[10] Ehab F., M. F. Badran , Hany Selim âĂIJSpeaker Recognition Using Artificial Neural Networks Based on Vowel phonemesâĂİ Electrical Engineeling Department, Assiut University.

[11] WEBOTS website `http://cyberbotics.com`

[12] `http://www.cyberbotics.com/dvd/common/doc/webots/guide/section3.5.html`