# HMM-based phoneme speech recognition system for control and command of industrial robots

Adwait Naik
(K J Somaiya College of Engineering, Mumbai, India)
adwaitnaik2@gmail.com

***Abstract-*** In recent years integration of Human-Robot interaction with speech recognition has gained a lot of pace in the manufacturing industries. Undoubtedly, this bridges the large gap created between the operator and robot by communication's point of view. Although there are numerous ways in which communication can be established between a human operator and the robot-like, controlling with a teaching pendant, or a joystick. Currently, the robots are controlled semi-autonomously by means of a computer. However, speech and touch [16] are natural ways of communication in humans, where speech recognition, being the best, is heavily researched technology. In this study, we aim at developing a stable and robust speech recognition system to allow humans to communicate with machines (Robotic-arm) in a seamless manner. This paper intends to investigate the potential of the Linear Predictive coding technique to develop a stable and robust HMM-based phoneme speech recognition system for robotics applications. Our system is divided into three segments: a microphone array, a voice module, and a 3-DOF robotic arm (Figure 1). To validate our approach, we have performed tests with simple and complex sentences for various robotics activities like manipulating a cube and pick and place tasks. Moreover, we also analyzed the test results to rectify problems like accuracy, recognition score, etc. Also the paper briefly enumerates the future prospects and applications of our approach.

Keywords - Speech recognition, phoneme, robotics, Human-Robot interaction (HRI), Linear Predictive Coding (LPC), Hidden Markov Model (HMM)

## 1. INTRODUCTION

Today the technologies centered on artificial intelligence (AI) are making our lives easier. Speech recognition is one such technology empowered by AI to enrich our lives. Speech recognition has successfully managed to pave its way in our lives and the changes are foreseeable in the field of robotics as well. Robots are destined to do repetitive tasks with high precision and consistency. But the future requirements are complex and dictate the requirement to integrate speech recognition and robotics to yield conceivable solutions.

Speech recognition is the ability of the robot to analyze and understand the human speech signal (commands) to perform a particular task. Modern speech recognition systems have recognition rates, but lack the much-required accuracy to accomplish the crucial tasks. Eliminating the noise by cleansing the speech signal is one of the major challenges to build a robust and stable speech recognition system and our motivation for

carrying out research in this field. In this paper, we propose a technique that employs Linear Predictive Coding (LPC), an approach based on maximum likelihood using automatic phoneme discrimination. Phoneme extraction is the key element of this process. We conducted experiments with a variety of techniques like Mel-frequency cepstral coefficients (MFCC), Perceptual Linear Prediction (PLP), and LPC. Based on the results, we preferred LPC because our system had successfully recognized the commands with the highest accuracy of 96.64 %.

Our target application is the human-robot collaboration domain, in particular social robotics where the robots are involved in conducting social experiments to study communication with humans. Interacting with the operator often requires something more than a co-operative, precise and highly accurate robot to facilitate flexibility in accomplishing the most tedious tasks. And a stable speech recognition system is a crucial part of this.

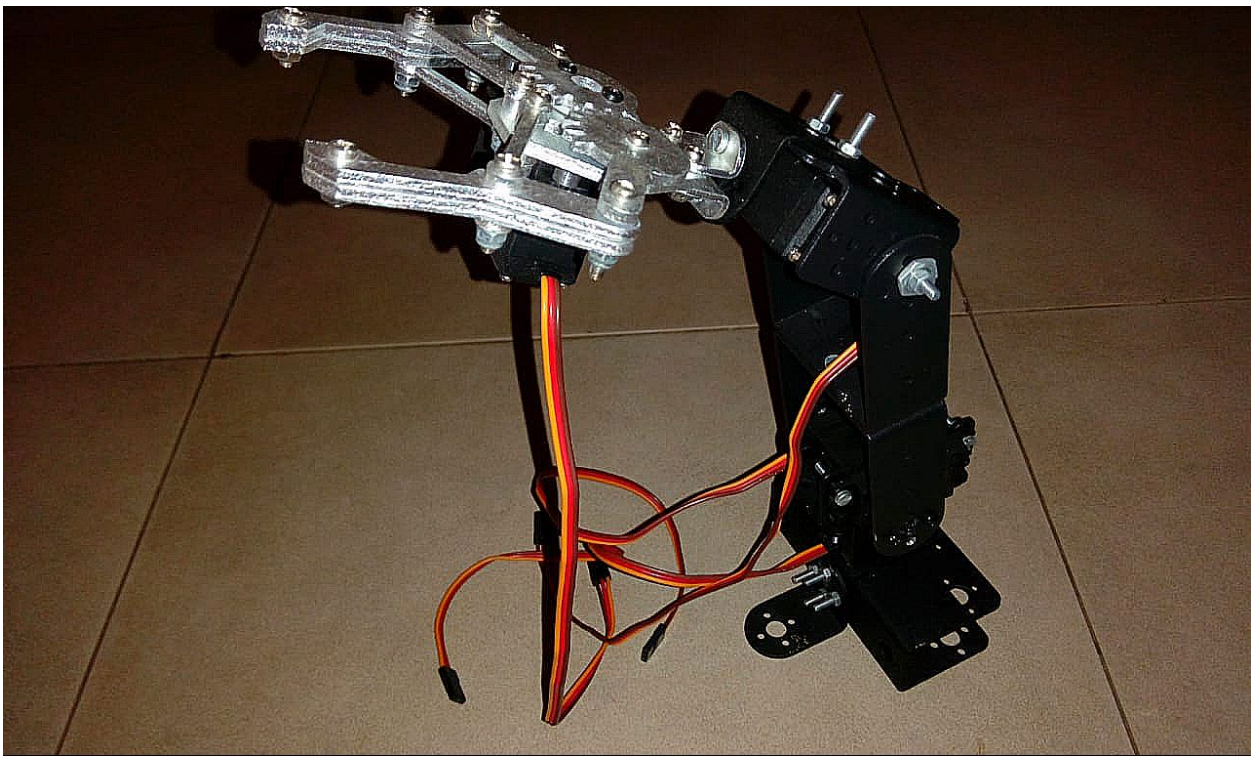

Figure 1: A 3-DOF robotic arm

This key objectives of our study are:

- Achieving wireless control over the robotic arm using voice commands to perform the pick and place operation.

- Perform tasks with greater accuracy, higher precision, and better compliance.

- Eliminate challenges like background noise by ensuring smooth interaction between the robot and the operator.

The remainder of this paper is organized as follows. Section 2 discusses the prior work followed by section 3 which provides a brief overview of the system design and Phoneme extraction process. In section 4, the training and model deployment is outlined. Section 5 describes the testing and hardware deployment stage. Section 6 summarises the robot actuation process. Section 7 describes the experimental results followed by section 8 which summarizes the observation. Section 9 concludes the manuscript and enumerates the future scope.

## 2. RELATED WORK

In [1], the authors have employed speech recognition to search for the victims in disaster-prone areas. The system was installed in an unmanned aircraft (drone) with the help of a micro-computer that is programmed to capture a variety of sounds or commands that are generally spoken by the victims in need of help. The paper briefly discusses the Hidden Markov Model with the Gaussian Mixture emissions (HMM-GMM) technique used to identify the words from audio files. Mel-frequency Cepstral Coefficient technique was used to extract the features from the voice commands to generate a dataset. Moreover, the experiments which were conducted to validate the study reflects that the system is capable of recognizing the sound/ commands at a maximum distance of 6 meters with 70% accuracy. Also, one of the major drawbacks of the system is the lack of optimal filtering to eliminate the noise.

The authors propose an intelligent method called (THHMB-STOI) in [2] to eliminate noise completely from a densely contaminated speech signal. The Twin-HMM model is used for extracting features from the signal and enhancement of the speech signal. An intelligible prediction framework that measures the short-time objective intelligibility (STOI) to produce clean and accurate speech. Its usage in the voice recognition system has drastically improved the accuracy by improving the speech transmission index (STI) and articulation index (AI). Moreover, the experiments conducted to examine the validity of this approach show a high correlation with human speech recognition results. This technique is mainly used in robotics, automatic speech recognition (ASR), etc.

All the methods mentioned above are non-intrusive and use MFCC for feature extraction. In other words, the methods can be implemented without the requirement of specific hardware. Also, the accuracy of these methods is low compared to the acoustic modeling technique proposed by the authors in [3] for the HMM-based speech recognition system. Using Convolutional Neural Network (CNN) to train the HMM model is the peculiarity of this technique. The ability to generate intermediate feature representations by subsequent filtering approach using CNN is the main advantage of this technique. Moreover, the system improves its accuracy of recognizing commands by a large margin but there is considerable loss of information or aliasing witnessed in the method mentioned above. This drawback is overcome by the technique that is mentioned in [5]. In this study, the author

investigates the Full-sum decoding method applied over the HMM-state sequences instead of the famous Viterbi algorithm. Full-sum decoding is tested both on the Librispeech and Switchboard corpora. Additionally, the paper briefly discusses the possible ways by which the tuning effort, efficiency can be improved to eliminate extra cost.

In [4], the author proposes an interesting application of the HMM, Audio-Visual speech recognition. In this study, the visual features are combined with the audio features using the early integration method followed by the classification of speech using the hidden Markov model. Gammatone frequency cepstral coefficient (GFCC) and Optical flow (OF) are integrated in a seamless manner to enhance the accuracy of the system. Moreover, the OF analysis gave a significant improvement in the Signal to noise ratio (SNR). The speech recognition is performed on a Hindi language database.

It's not a surprise that speech recognition has successfully managed to pave its way into the field of robotics to perform difficult tasks varying from Robotics Car control [7] to commanding the PR2 robot [6] to perform clandestine tasks. The papers introduced a robust speech recognition system by combining the Deep Neural Network (DNN) with the hidden Markov model (HMM). In [7,8,9], the author reviewed a speech controlled automation system (SCAS) which is closely related to our application as well. The importance of speech recognition in Human-Machine Interaction is briefly discussed in [9,10,11] where the author proposes various applications integrating the principles of speech recognition, robotics, and machine learning to solve many problems in the field of social robotics. Apart from industries, speech recognition has witnessed rapid growth in the field of Social Robotics as well [14,15]. Voice Recognition Controlled Computer (VRCC) is another conventional assistive technique usually used to control most of the ABB and KUKA robots. It comes with a typical Human-Robot interface which ensures that the commands given as input by the end-user are recognized by the robot. The author in [15] uses this technique mainly for torch welding and trajectory planning purpose.

Co-operation and collaboration is another important application of robotics [16, 19]. The paper focuses on another dimension of robotics application - speech recognition for care and support. *Lio*, a robotic arm by F&P robotics, unlike other robots is programmed to interact with humans in a completely different environment like old age homes, rehabilitation centers, etc. According to the author, the *cobot*, as it's addressed in the paper was successfully deployed into sensitive areas like hospitals, rehabs taking into consideration the safety and reliability of the patients surrounded.

## 3. SYSTEM DESIGN

This section gives a brief overview of the preparation of the speech dataset, training the hidden Markov model [7-10], and the underlying principle of the LPC technique used for extracting important data from the audio (speech) input. Moreover, we have briefly discussed the advantages and limitations of the LPC technique [18].

### 3.1 TRAINING STAGE

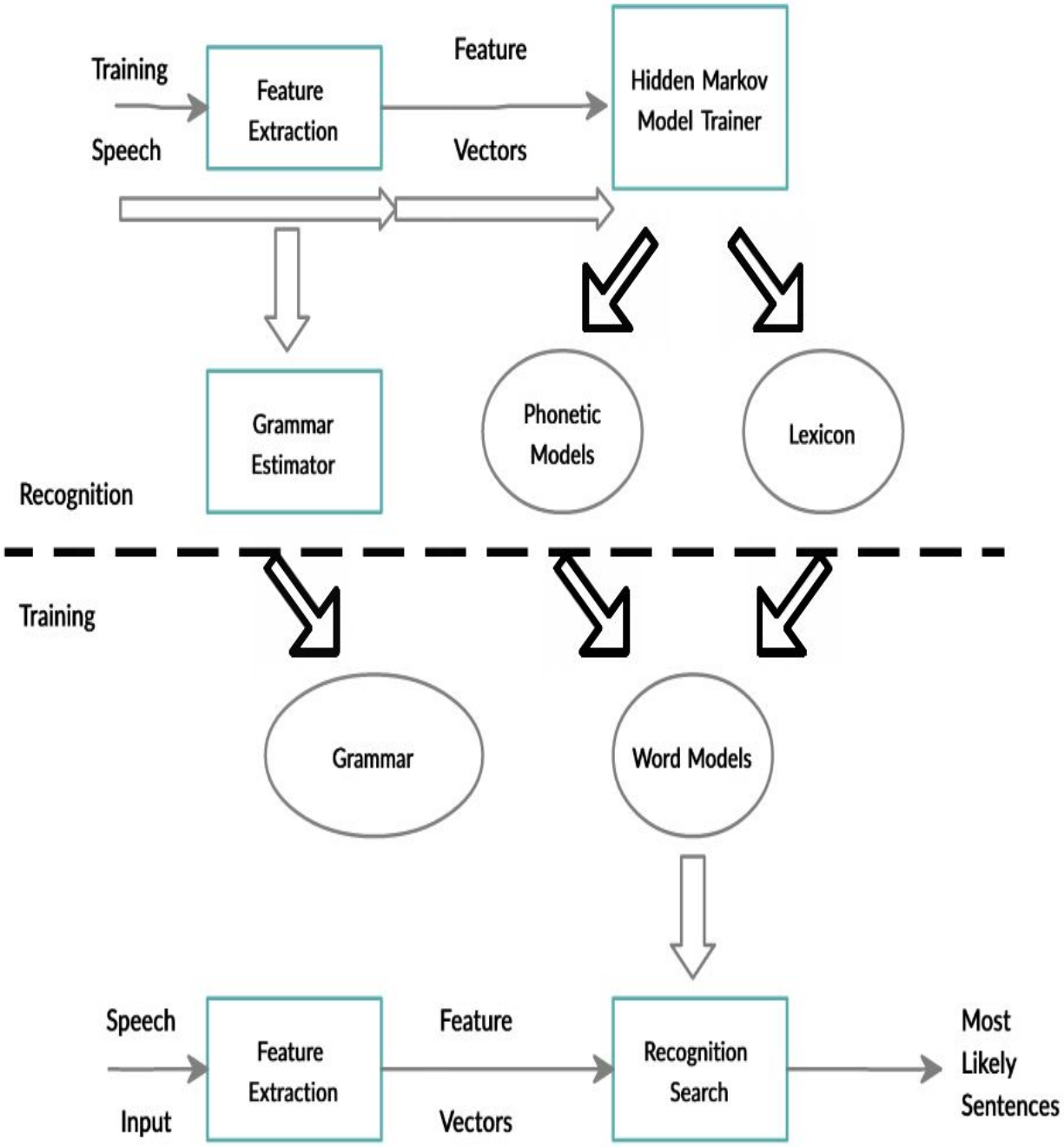


Figure 2: Automatic-Speech recognition system (Makhoul et al. 1995)

The training procedure is divided into two parts; the Training stage and the Recognition stage (Figure 2).

### 3.2 DATA PREPARATION

The corpus is a dataset of commands recorded to train the (HMM) trainer. Commands used during training are shown in Table. 1 below.

**Table 1** List of commands

| | **Primitive Commands for robot control** | **Description** |
|---|---|---|
| 1 | **Start** | This command initiates the movement in the servo motor attached to the base. |
| 2 | **Stop** | This command halts the movement of the servo motor attached to the base |
| 3 | **Rotate the base clockwise** | This command rotates the servo motor attached to the base by 180 degrees in the clockwise direction |
| 4 | **Rotate the shoulder** | This command rotates the servo motor attached to the shoulder by 180 degrees in the clockwise direction |
| 5 | **Open gripper** | This command rotates the servo motor attached to the gripper in the clockwise direction thereby opening the gripper. |
| 6 | **Close gripper** | This command rotates the servo motor attached to the gripper in an anticlockwise direction thereby closing the gripper. |
| 7 | **Lift the shoulder up** | This command rotates the servo motor attached to the shoulder by 90 degrees upwards. |
| 8 | **Put the shoulder down** | This command rotates the servo motor attached to the shoulder by 90 degrees downwards. |

The text corpus is a collection of eight primitive commands used to control the robot. For each command, a separate program had to be written in C++ which was fed to the Arduino manager to be transferred to the Arduino microcontroller to control the servo motors.

### 3.3 RECORDING THE COMMANDS

The text corpus was used for recording the commands using 2 speakers at a dual-frequency channel of 44.1 kHz at -6db peak and 256 Kbps data rate using the Audacity® software which is an open-source application used for digital audio editing, mixing, and recording. Commands were recorded in three groups, each having four commands. The recording was exported and stored in .wav format as shown in Figure. 3 below.

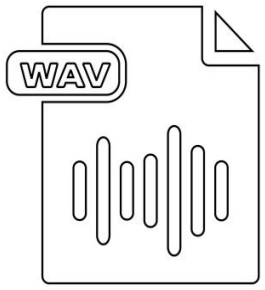


Figure 3: recorded command in .wav format

To build models using the recorded commands in .wav format Hidden Markov Model Toolkit (HTK) software was used.

### 3.4 FEATURE EXTRACTION

Feature extraction is an essential step in the process of speech recognition as it segregates the speaker's voice from all other voices and generates observational vectors. It reduces the magnitude of the speech signal responsible for causing damage to the power of the speech signal. In this case, the input given is an audio signal. Various techniques for feature extraction like MFCC, Linear Predictive coding, FFT, and RASTA. We preferably used the Linear Predictive Coding (LPC) technique as shown in Figure. 4 below, for the following advantages:

- High computation speed and robustness.
- Bit rate requirement is less for transmission.

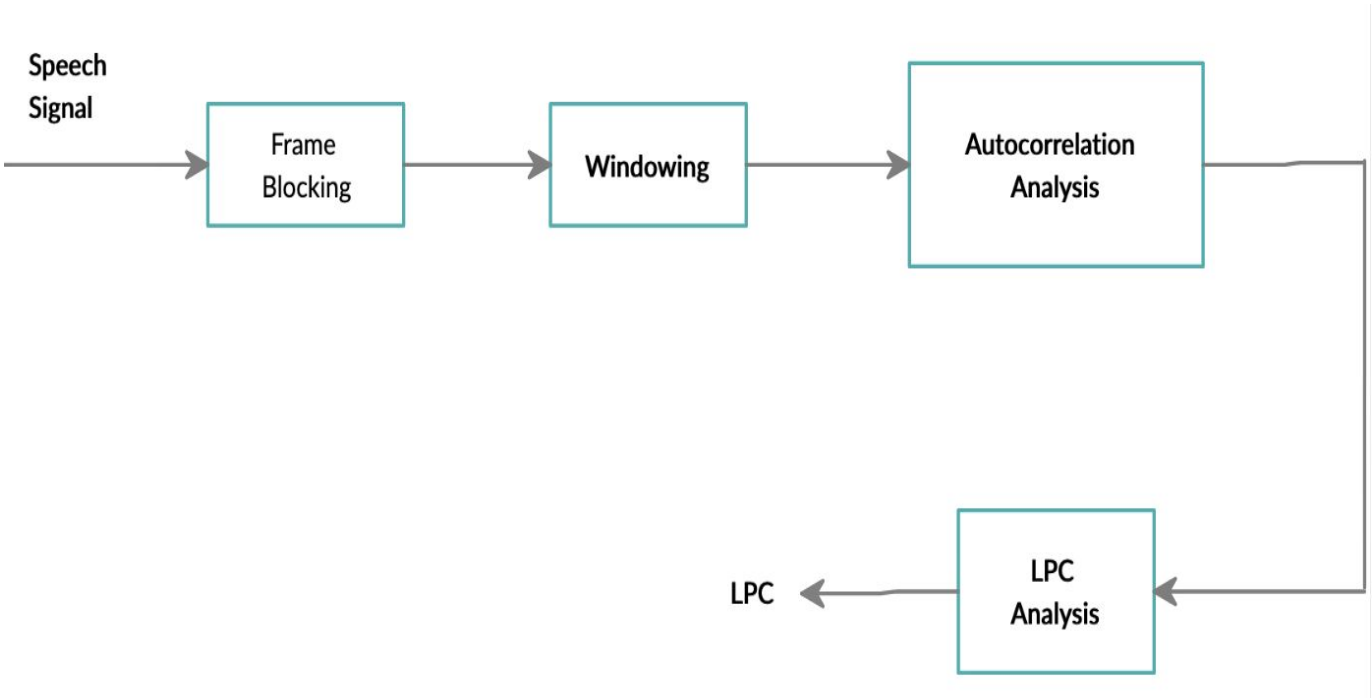


Figure 4: Block diagram for the LPC technique (Sabur Ajibola Alim et al. 2018)

### 3.4.1 LINEAR PREDICTIVE CODING

Linear prediction coding, in its operation, resembles the human vocal tract. Also known as a format estimation technique, it is used to estimate the formants and reduce their effects on the signal. Here, the formants are peaks or local maximum occurring in the spectrum as a result of resonance. The frequencies where the formants appear are defined as formant frequencies. The location of the formats in the spectrum can be deduced by calculating the linear predictive coefficients.

LPC is based on the principle of reducing the mean square error (shown in Equation 2) between the input speech and estimated speech. The speech sample at any time interval is expressed as a linear weighted aggregation of preceding samples. The linear predictive model of speech creation is given as:

$$\hat{S}(n) = \sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

where $\hat{S}$ is the predicted sample, s is the input speech sample, and p is the predictor coefficients.

The prediction error is given as [7]:

$$e(n) = S(n) - \hat{S}(n) \quad (2)$$

After the speech signal is pre-processed, it is passed for frame blocking as shown in the block diagram in Figure. 4 above. Each frame is autocorrelated and the highest autocorrelation value is chosen for the linear predictive analysis [11, 12]. In linear predictive analysis, the coefficients are calculated which are given by [7, 8]:

$$a_m = ( log[1-k_m] / log[1+k_m] ) \quad (3)$$

where $a_m$ is the linear prediction coefficient and $k_m$ is the reflection coefficient.

Furthermore, LPC is used to very accurately estimate the vocal tract properties from the speech signal and hence is a very effective technique employed in the tonal analysis of string instruments like violin and guitar.

## 4. TRAINING HMMs

In this stage, to create the final trained model, we developed the speech recognition system using the Hidden Markov Model-based toolkit HTK Version 3.2 in the Mac OS (Official site of HTK toolkit, htk.eng.cam.ac.uk).

Table 2 Parameters for feature extraction using LPC

| S. No | Parameters | Value of parameters |
|---|---|---|
| 1 | Features Extracted | LPC |
| 2 | Window used | Hamming |
| 3 | Window length | 15 ms |
| 4 | Frame count | 12 |
| 5 | Pre-emphasis (Pre-processing) | 0.67 |
| 6 | Number of coefficients ($a_k$) | 16 |
| 7 | linear prediction cepstral coefficients | 28 |

The parameters listed in Table 2 were calculated during the feature extraction from the speech signal using the LPC process.

### 4.1 MODEL TRAINING AND DEPLOYMENT

In HTK, **HRest**, **HInit**, and **HERest** programs as shown in Figure 5 are used to make the acoustic model. The parameters are initialized by the HInit program using the Viterbi Extraction algorithm. HRest estimates the parameters shown in Table 2 above, using the Baum-Welch algorithm. On comparing the performance HRest is outperformed by HERest in a noisy environment.

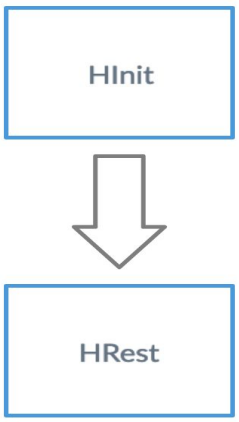


Figure 5: HTK programs for model creation

## 5. TESTING

The **HVite** program was used for testing the recorded commands. It uses the Token passing algorithm to perform offline testing using the recorded database. HVITE takes as input a network describing the allowable word sequences, a dictionary defining how each word is pronounced and a set of HMMs.

### 5.1 LIKELIHOOD WITH THE FORWARD ALGORITHM

Given that we have successfully modeled the HMM, to calculate the likelihood for our observations, we use the forward algorithm which is based on summing the probabilities (equation 4) of all the states in all possible state sequences:

Mathematically, the probability is calculated using

$$p(X) = \sum_{s} p(X, S) = \sum_{s} p(X, S)\, p(S) \qquad (4)$$

Where X denotes the observed events, $\sum$ s denotes sum over all possible time sequences of internal states, p(X, S) is the emission probability and p(S) is transition probability.

### 5.2 HARDWARE

After training the Hidden Markov Model, it was deployed on the Geetech speech recognition module as shown in Figure 5 (Official site Geetech module).

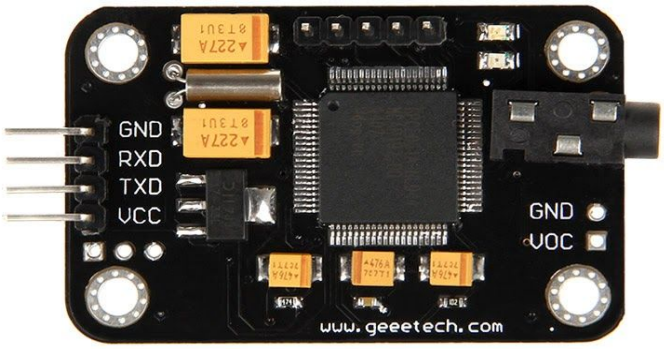


Figure 6: speech recognition module

The voice module (Figure 6) is one of the key components of this system. It works on the principle of serial data transfer when connected to the Arduino board. Equipped with a Digital Signal Processor of SC57X series based on SHARC (Super Harvard Architecture Single-Chip Computer) architecture. It comes with the ARM® Cortex-A5 system control capability, which provides high performance for complex applications demanding the latest advanced algorithms.

## 6. ROBOT ACTUATION AND RECOGNITION

In Figure 7, speech recognition processes are shown. When the speaker gives a command, for example - “Open Gripper”, every phoneme in this command is isolated and matched with the commands used to train the Hidden Markov Model. If the command matches then the servo motor attached to the particular joint viz. Shoulder, elbow, or the gripper actuates as a result of which the movement takes place.

The comparison stage contains the trained model that is responsible for carefully matching the commands given by the speaker. Based on the parameters calculated in Table 2, the decision by comparison stage is made. In case of mismatch, background noise, and wrong pronunciation of command the speech recognition system fails to identify the phonemes in the command, as a result, the robot is not actuated.

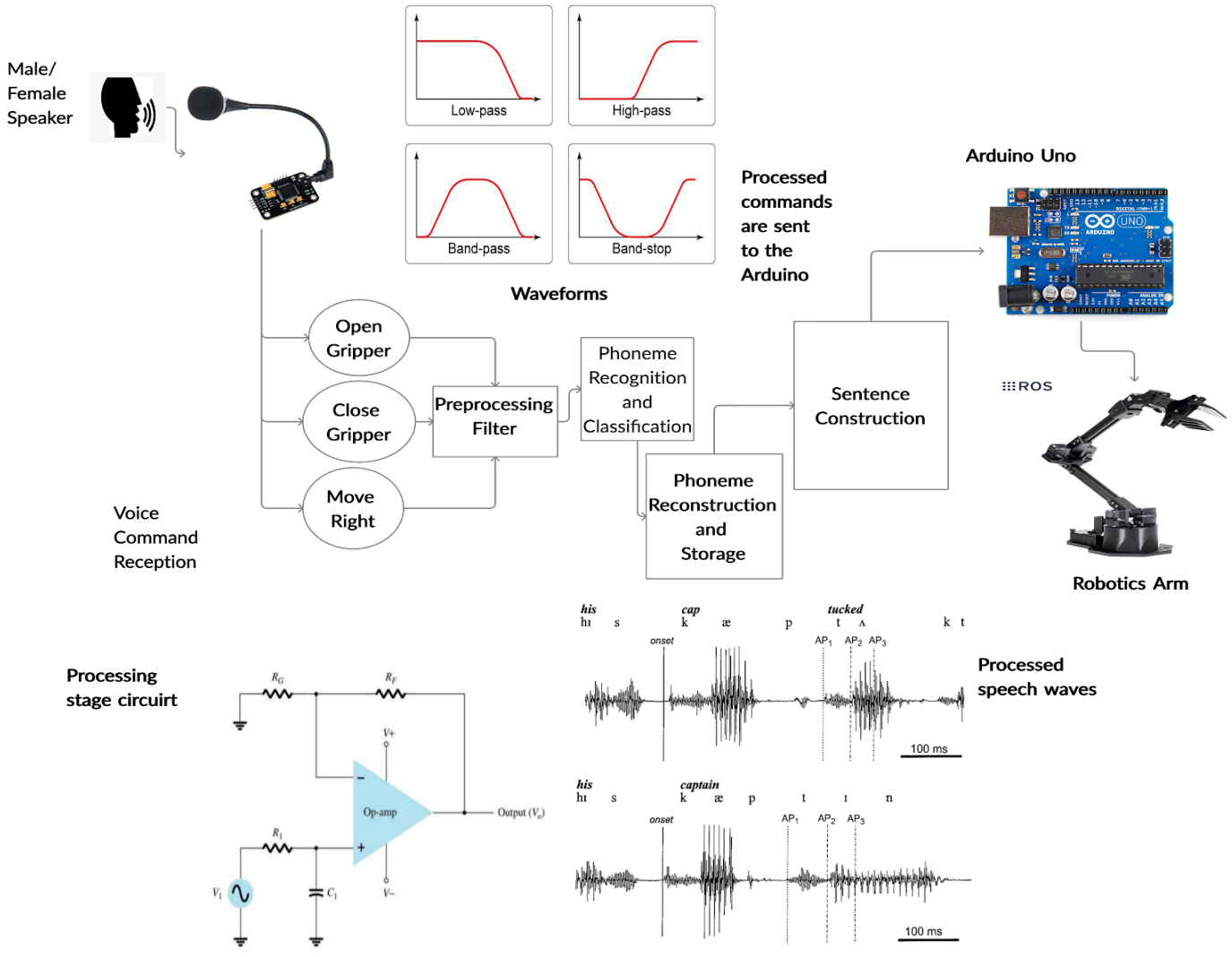


Figure 7: Speech recognition to control the robot

As mentioned above, a program in C++ was designed for each command listed in Table 1 above. The commands are used to form functions, to which the parameters like time, angle, and the number of steps are passed; so when the program is executed, the speech signal actuates the servo motor attached to the respective joint. Here the commands are allotted transmission channels on the Arduino board, for example- Channel 1 is designated for "Start" command, Channel 2 is designated for "Stop" command, etc.

## 7. EXPERIMENTAL RESULTS

The training and testing for the speech recognition system were done using the commands recorded. The database used for testing the speech recognition system consists of 200 speech samples from different speakers involved in the testing process. As mentioned above, the Linear Predictive Coding (LPC) technique used in double differentiation mode. Here, we have used the HTK toolkit to train the acoustic model and deploy it.

In the experiments, we have evaluated Word Error Rate, Recognition Score, and Word Accuracy Rate which are shown in Table 3 below.

***Word Error Rate:***

Word Error Rate (WER) is the measure of the difference between the recognized word sequence and input word sequence. It is the most commonly used performance metrics for speech recognition systems. Its computation is based on the Levenshtein distance. WER is calculated on the Phoneme level. WER is given by

$$WER = (S + D + I) / N \quad (4)$$

Where *S* is the number of substitutions,

- *D* is the number of deletions,
- *I* is the number of insertions,
- *C* is the number of correct words,
- *N* is the number of words in the reference (N=S+D+C)

***Word Accuracy Rate:***

Word Accuracy Rate (WAcc) is the percent word accuracy is defined as %WAcc = 100 - %WER. It should be noted that the word accuracy can be negative. WAcc is given by

$$WAcc = (N - S - D - I) / N = (H - I) / N \quad (5)$$

Where H = No. of words that are correctly recognized

Table 3 WER and WAcc calculation

| **Technique Used** | | **Rate in %** | |
|---|---|---|---|
| **Linear Prediction Coding** | ***Word Error Rate*** | ***Word Accuracy Rate*** | ***Recognition Score**** |
| **Male 1** | **12.56** | 87.44 | 83 |
| **Male 2** | 10.52 | 89.48 | 86.42 |
| **Male 3** | 9.64 | 86.46 | 82.68 |
| **Male 4** | 8.26 | 82.32 | 81.04 |
| **Female 1** | 7.65 | 92.35 | 96 |
| **Female 2** | 5.96 | 94.04 | **96.46**** |
| **Female 3** | 9.28 | 92.35 | 92.32 |
| **Female 4** | 6.36 | 91.09 | 91.04 |

*The recognition score is calculated programmatically. ** Highest recognition score achieved

## 8. OBSERVATIONS

To validate the performance of our speech recognition system we conducted an experiment involving 8 speakers (4 Male speakers and 4 Female speakers) volunteers for testing the accuracy, recognition scores which are shown in Figure 8 (a), Figure 8 (b), and Figure. 9 below.

In Table 3 we observe that Female speaker 2 was able to attend a remarkable recognition score **of 96.46 %** which is highest compared to other volunteers. Also the Word Accuracy Rate maximum for the same candidate **94.04 %** with the Word Error Rate being lowest **5.96 %**. This signifies that the recognition system, though not being biased towards a particular voice pattern, is significantly affected by quality, pitch, and other wavelengths of the speech signal. On the contrary, Male speaker 1 has the maximum Word Error Rate and minimum Word Accuracy Rate. With these observations we conclude that the female voice tends to rise and fall more dramatically when compared with the male voices. We also observed that the background noise significantly affects the accuracy of the speech recognition system. The noise perturbations tend to degrade the quality of the speech input by aliasing (Figure 9(b) ) due to which the wavelets (small portions of the waves) overlap resulting in loss of information. The solution to overcome this problem is using an Anti-aliasing filter which avoids the waves from folding and thus prevents loss of information.

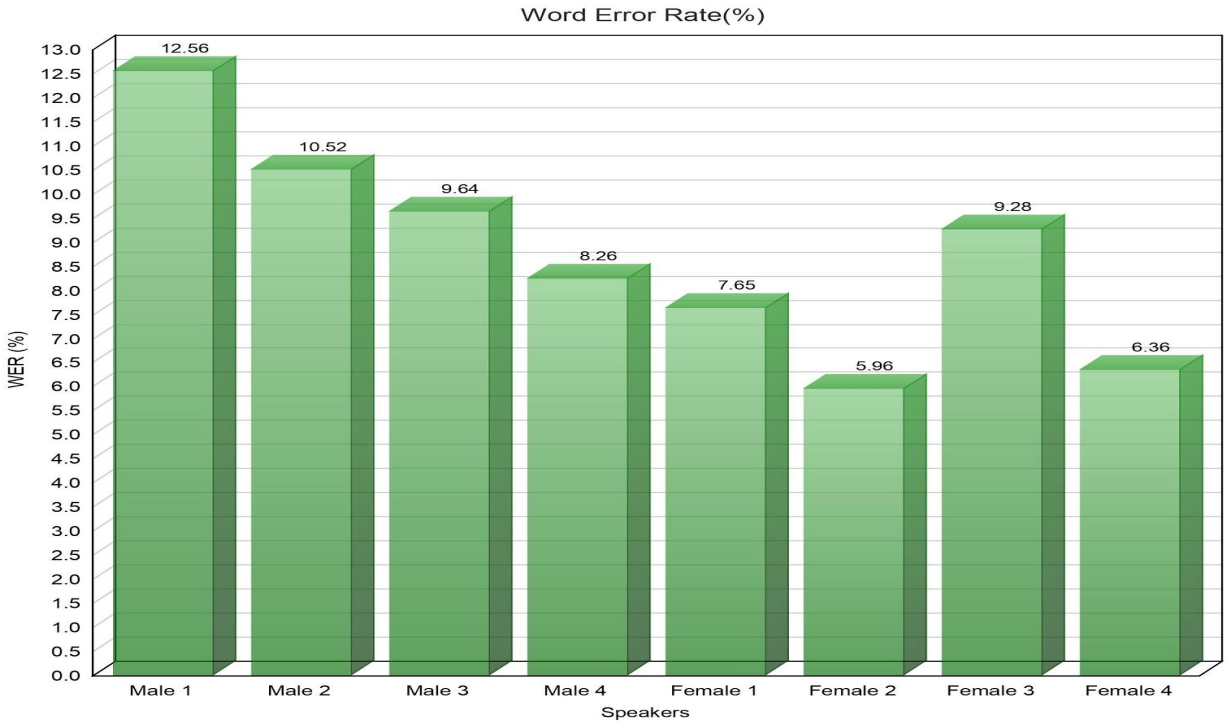


Figure 8: (a) Word Error Rate graph

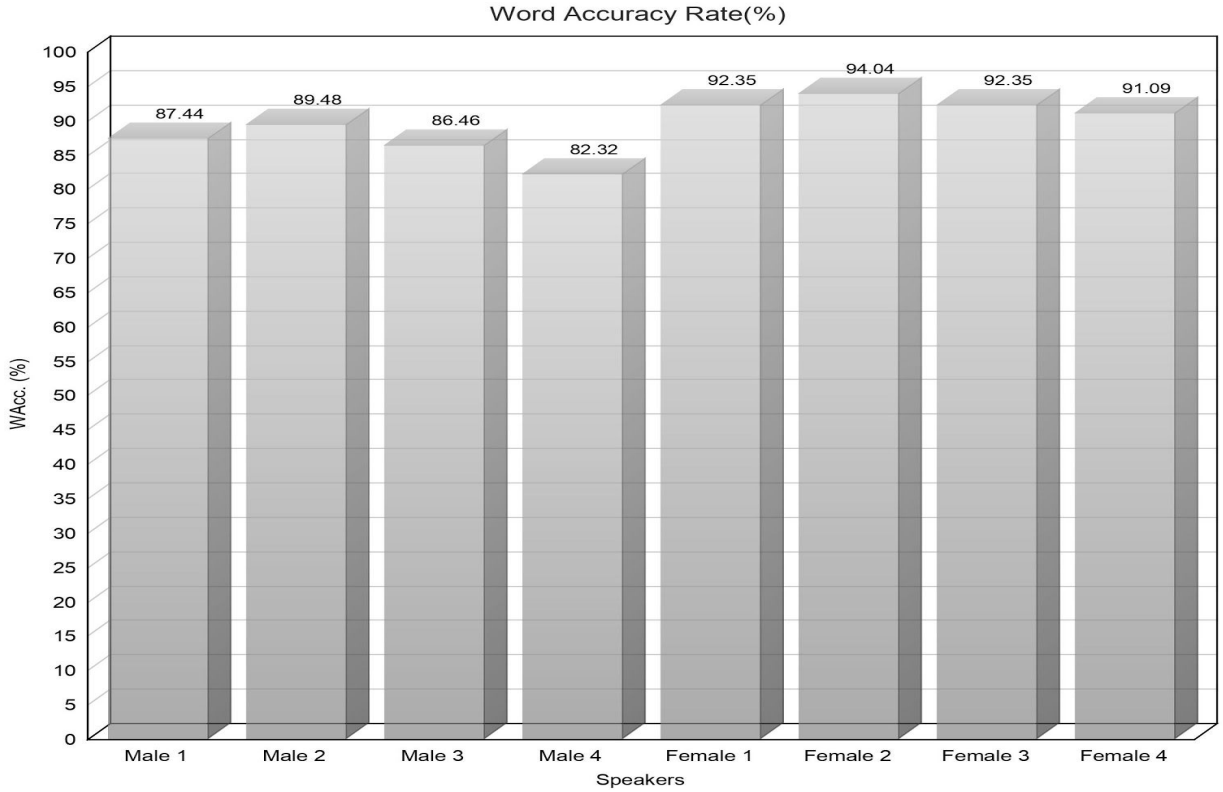


Figure 8: (b) Word Accuracy percentage graph

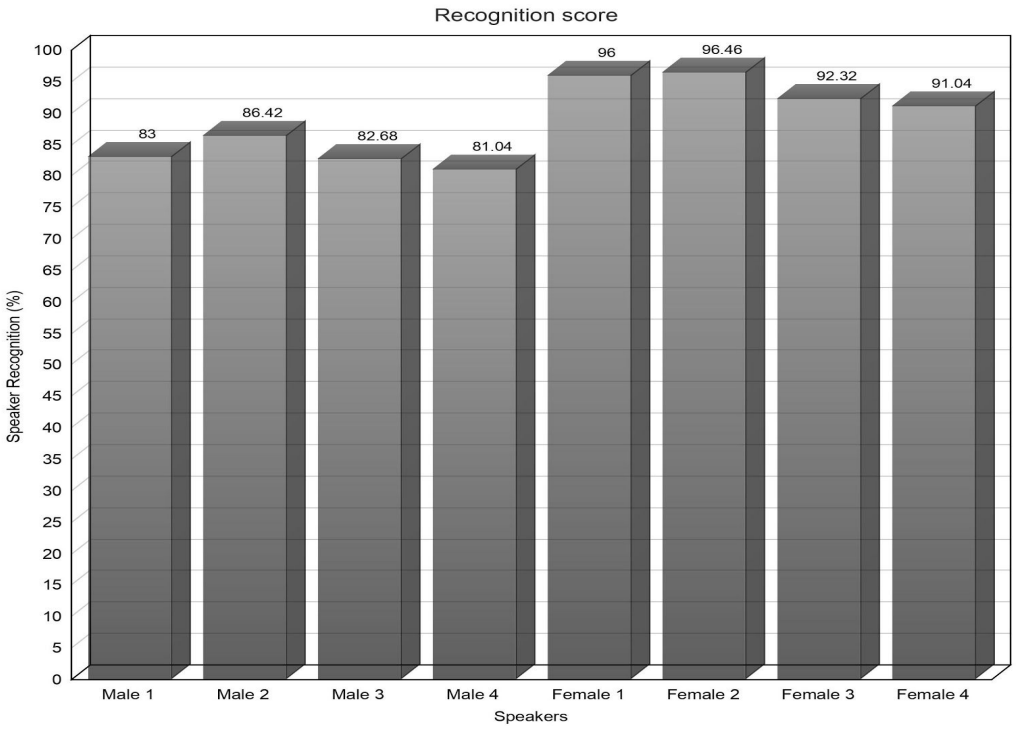


Figure 9 (a): Speaker's recognition score

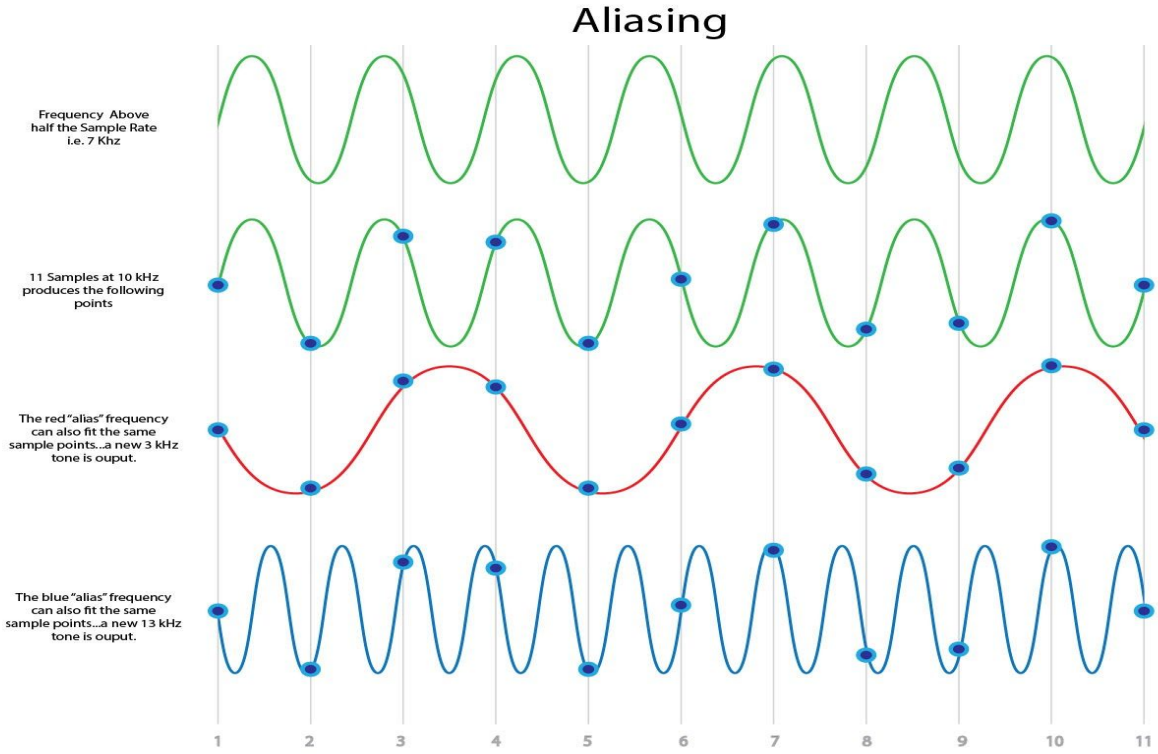


Figure 9 (b): Aliasing or Foldover (*source cited in references*)

Table 4 Recognition scores

| **Words** | **Phonetic pronunciation** | **Calculated Avg. phoneme recognition%** | **Observed Recognition score** |
|---|---|---|---|
| Start | st a r t | 74.48 | 74 |
| Stop | st a w p | 86.62 | 87 |
| Rotate the base clockwise | row teit the beis klawk vaiz | 60.32 | 61 |
| Rotate the shoulder | row teit the showl duh | 74.28 | 75 |
| Open gripper | ow pn grip er | 85.63 | 86 |
| Close gripper | kle oz grip er | 84.43 | 85 |
| Lift the shoulder up | lift the showl duh up | 55.63 | 56 |
| Put the shoulder down | put the show duh da un | 44.57 | 45 |

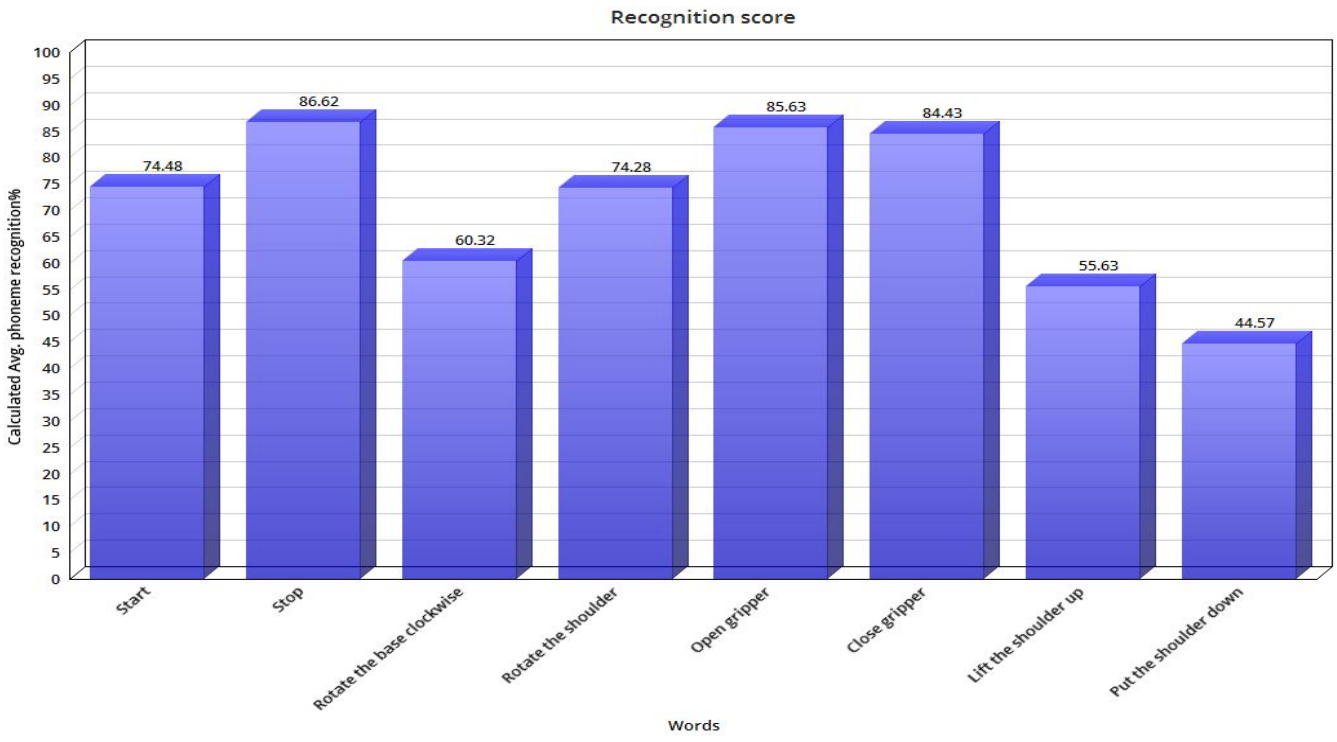


Figure 10: Calculated recognition score graph

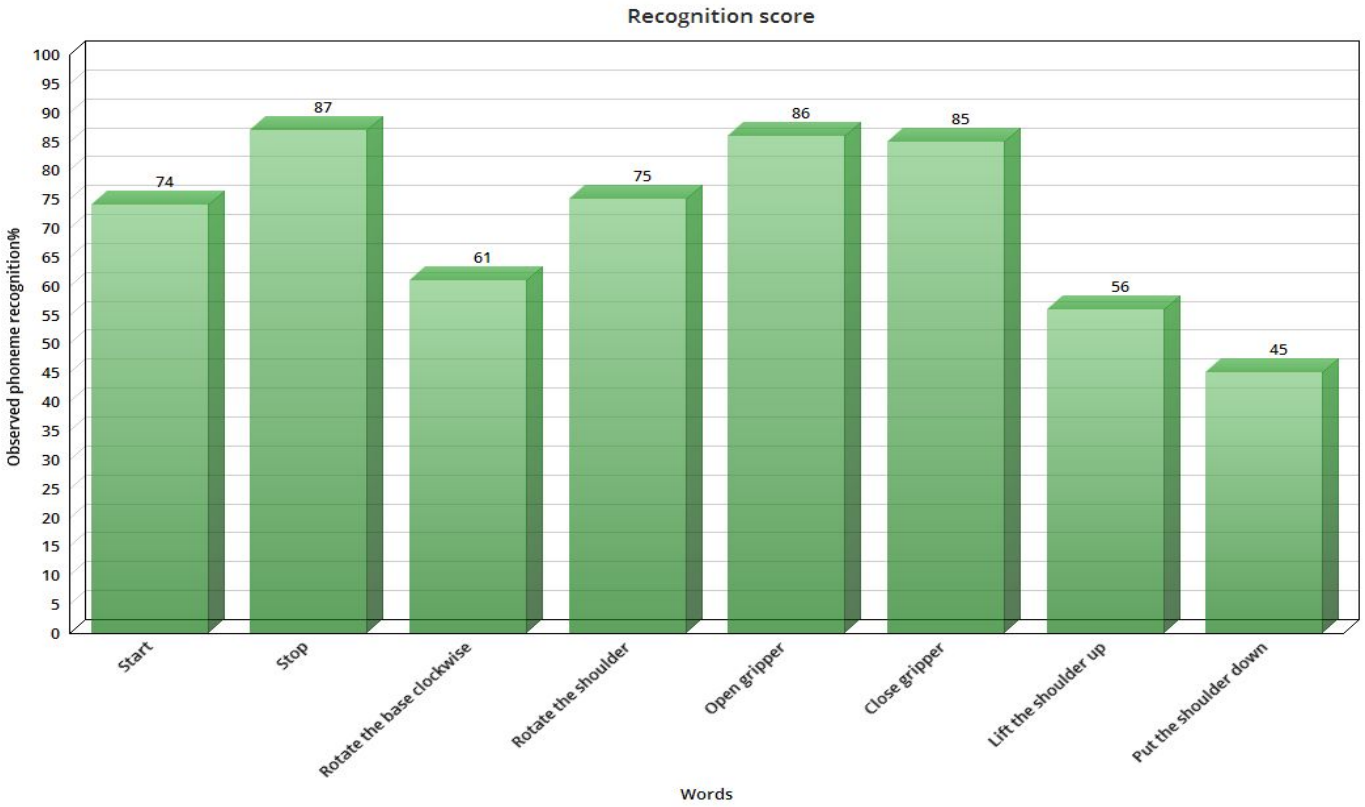


Figure 11: Observed recognition score graph

## 9. CONCLUSION AND FUTURE SCOPE

In this paper, we present the Hidden Markov Model-based technique to develop a robust and stable speech recognition system. We investigate the Linear Predictive Coding (LPC) method for feature extraction from the speech signals to build the dataset. While from the experiments carried so far, we are able to conclude that our system is highly capable of recognizing and classifying each phoneme accurately (Figure. 10 & Figure. 11) at the same time there are limitations which we have pinpointed like;

- Considerable reduction in the recognition score when long sentences were used by the speaker (see Figure. 11). There is a significant difference in the recognition score of **Start** command and **Put the shoulder down command.**
- Considerable reduction in the accuracy and recognition rate when the distance between the source (speaker) and the robot is increased. (Please note that all the readings (Table 4) are taken from a distance of 3 meters)
- The utterance of a word by 2 or more speakers at the same time significantly affects the recognition process.

The proposed approach paves the path for further research and development. We enumerate and discuss them briefly below.

1) **Gender Recognition from voice sample**: This application focuses on identifying the gender based on the voice samples provided. It incorporates algorithms like logistic regression, decision tree, Support Vector Machines (SVM), etc**.** The Harvard-Haskins Database of Regularly-Timed Speech dataset can be used to train an **Acoustic Voice Model** [20] that resembles a tree or a graph with each node having the acoustic properties of a male or a female speaker.

2) **Audio-Visual Speech recognition:** Combining the Optical Flow (OF), which is defined as pattern formation upon random motion of objects, surfaces, etc caused by relative motion between the observer and the camera, with our approach will allow the system to recognize the object that generates the sound.

3) **The Pepper Social Robot:** Deploying social robots at public places is one of the biggest challenges for developing a robust Automatic Speech recognition system (ASR). Conventional ASR engines from Google, Microsoft, Nuance, etc. can be very expensive [19]. However, their performance is unmatchable for tasks like socializing. Our approach, if scaled properly, can be very effective at the same time cost-efficient.

## COMPARATIVE ANALYSIS

Table 5 Analysis for SNR-5 dB

| Sr. No. | Feature Extraction Technique | Language | Database | Average Recognition Score (%) | Application |
|---|---|---|---|---|---|
| 1 | BFCC+CFIF+CFSE | English | ROBSPEECH | 67.13% | Robotics Manipulation |
| 2 | PLP | English | ROBSPEECH | 77.53% | Robotics Manipulation |
| 3 | LPA | English | ROBSPEECH | 94.08% | Robotics Manipulation |
| 4 | MFCC | English | ROBSPEECH | 91.02% | Robotics Manipulation |

Table 6 Analysis for SNR-10 dB

| Sr. No. | Feature Extraction Technique | Language | Database | Average Recognition Score (%) | Application |
|---|---|---|---|---|---|
| 1 | BFCC+CFIF+CFSE | English | ROBSPEECH | 62.193% | Robotics Manipulation |
| 2 | PLP | English | ROBSPEECH | 74.263% | Robotics Manipulation |
| 3 | LPA | English | ROBSPEECH | 90.08% | Robotics Manipulation |
| 4 | MFCC | English | ROBSPEECH | 85.02% | Robotics Manipulation |

Table 7 Analysis for SNR-15 dB

| Sr. No. | Feature Extraction Technique | Language | Database | Average Recognition Score (%) | Application |
|---|---|---|---|---|---|
| 1 | BFCC+CFIF+CFSE | English | ROBSPEECH | 58.13% | Robotics Manipulation |
| 2 | PLP | English | ROBSPEECH | 70.73% | Robotics Manipulation |
| 3 | LPA | English | ROBSPEECH | App. 88.92% | Robotics Manipulation |
| 4 | MFCC | English | ROBSPEECH | 80.34% | Robotics Manipulation |

In this section, we have compared 4 methods for feature namely; BFCC, PLP, LPA and, MFCC for different Signal to Noise ratios (SNR). Please note ROBSPEECH is the database we have created for validating our approach. This study justifies that our approach to adopt the LPA technique witnesses the least reduction with an increase in the SNR value. Please note that the values calculated here (table 5, table 6, and table 7) are

approximated on the basis of the noise sources (like the ceiling Fan, wall-mounted fan, etc) present in the laboratory. With this study, we are able to conclude that our technique outperforms all other techniques.

## REFERENCES


[1] F. Alifani, T. W. Purboyo and C. Setianingsih, "Implementation of Voice Recognition in Disaster Victim Detection Using Hidden Markov Model (HMM) Method," *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Surabaya, Indonesia, 2019, pp. 445-450, doi: 10.1109/ISITIA.2019.8937290.

[2] M. Karbasi, A. H. Abdelaziz and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 624-628, doi: 10.1109/ICASSP.2016.7471750.

[3] Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech Commun., 108*, 15-32.

[4] Sharma, U., Maheshkar, S., Mishra, A.N. *et al.* Visual Speech Recognition Using Optical Flow and Hidden Markov Model. *Wireless Pers Commun* **106,** 2129–2147 (2019). https://doi.org/10.1007/s11277-018-5930-z

[5] Zhou, Wei & Schluter, Ralf & Ney, Hermann. (2020). Full-Sum Decoding for Hybrid Hmm Based Speech Recognition Using LSTM Language Model. 7834-7838. 10.1109/ICASSP40776.2020.9053010.

[6] José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, and Néstor Becerra Yoma. 2018. DNN-HMM based Automatic Speech Recognition for HRI Scenarios. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18). Association for Computing Machinery, New York, NY, USA, 150–159. DOI:https://doi.org/10.1145/3171221.3171280

[7] Shafiq, Ayesha & Alvi, Fareed & Tariq, Humera & Amjad, Usman. (2019). Voice Recognition System Design Aspects for Robotic Car Control. IJCSNS International Journal of Computer Science and Network Security, VOL.19 No.1, January 2019.

[8] N. Baranwal, A. K. Singh and T. Hellström, "Fusion of Gesture and Speech for Increased Accuracy in Human Robot Interaction," 2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR), Międzyzdroje, Poland, 2019, pp. 139-144, doi: 10.1109/MMAR.2019.8864671.

[9] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg and S. Wermter, "On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 854-860, doi: 10.1109/IROS.2018.8593571.

[10] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg and S. Wermter, "On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 854-860, doi: 10.1109/IROS.2018.8593571.

[11] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and

Recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17). Association for Computing Machinery, New York, NY, USA, 82–90. DOI:https://doi.org/10.1145/2909824.3020229

[12] W. Ting, "An Acoustic Recognition Model for English Speech Based on Improved HMM Algorithm," 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Qiqihar, China, 2019, pp. 729-732, doi: 10.1109/ICMTMA.2019.00167.

[13] D. K. Ninh, "A Speaker-Adaptive HMM-based Vietnamese Text-to-Speech System," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 2019, pp. 1-5, doi: 10.1109/KSE.2019.8919326.

[14] Ande, S.K., Kuchibotla, M.R. & Adavi, B.K. Robot acquisition, control and interfacing using multimodal feedback. *J Ambient Intell Human Comput* (2020). https://doi.org/10.1007/s12652-020-01738-0

[15] Gao, Z., Wanyama, T., Singh, I., Gadhrri, A., & Schmidt, R. (2020). From Industry 4.0 to Robotics 4.0-A Conceptual Framework for Collaborative and Intelligent Robotic Systems. *Procedia Manufacturing*, *46*, 591-599.

[16] Bendel, O. (2020). Co-Robots as Care Robots. *arXiv preprint arXiv:2004.04374*.

[17] Ande, Stanly Kumar, Mallikarjuna Rao Kuchibotla, and Bala Krishna Adavi. "Robot acquisition, control and interfacing using multimodal feedback." *Journal of Ambient Intelligence and Humanized Computing* (2020): 1-11.

[18] Sabur Ajibola Alim and Nahrul Khair Alang Rashid (December 12th 2018). Some Commonly Used Speech Feature Extraction Algorithms, From Natural to Artificial Intelligence - Algorithms and Applications, Ricardo Lopez-Ruiz, IntechOpen, DOI: 10.5772/intechopen.80419. Available from: https://www.intechopen.com/books/from-natural-to-artificial-intelligence-algorithms-and-applications/some-commonly-used-speech-feature-extraction-algorithms

[19] Charles J., Vishwas M., Ruixi L. (2020). Improved Robust ASR for Social Robots in Public Spaces. arXiv preprint arXiv:2001.0.04619.

[20] Kori B, "Identifying the Gender of a Voice using Machine Learning", Link:http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/

[21] Figure 9 (b), "Aliasing", http://www.realhd-audio.com/wp-content/uploads/2014/05/140528_aliasing_illustration.jpg

## ACKNOWLEDGEMENTS

The authors of the paper would like to express their gratitude to **Prof. Annu Abraham** and **Dr. J H Nirmal** for their guidance and moral support.

## CONFLICT OF INTEREST

The authors have no conflicts of interest to declare regarding or related to the contents of the manuscript.

## ETHICAL DECLARATIONS

The authors of the manuscript declare that no animals were involved in the experiments performed during the study.


## DISCLOSURE OF FUNDING

The authors of this paper would like to mention that no specific funding was received for this project.


## Author's biography

Adwait P Naik, born in 1997, graduated from K. J. Somaiya College of Engineering, affiliated to the University of Mumbai with B.tech (Hons) in Electronics Engineering in 2019. Currently, working as a research intern at HTIC, IIT-Madras. The author's research interest spans over various fields including Robotics, Artificial Intelligence, Speech Recognition, and Machine learning.