technical transactions

No. 2021/002

mechanics

sciendo

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

Adwait Naik

adwaitnaik2@gmail.com | http://orcid.org/0000-0001-9655-7564

University of Mumbai, India

Scientific Editor: Jacek Pietraszek, Cracow University of Technology
Technical Editor: Aleksandra Urzędowska, Cracow University of Technology Press
Language Editor: Tim Churcher, Big Picture
Typesetting: Małgorzata Murat-Drożyńska, Cracow University of Technology Press

Received: June 1, 2020
Accepted: February 5, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing interests: The authors have declared that no competing interests exist.

Citation: Naik, A. (2021). HMM-based phoneme speech recognition system for the control and command of industrial robots. Technical. *Technical Transactions: e2021002*. https://doi.org/10.37705/TechTrans/e2021002

Abstract

In recent years, the integration of human-robot interaction with speech recognition has gained a lot of pace in the manufacturing industries. Conventional methods to control the robots include semi-autonomous, fully-autonomous, and wired methods. Operating through a teaching pendant or a joystick is easy to implement but is not effective when the robot is deployed to perform complex repetitive tasks. Speech and touch are natural ways of communicating for humans and speech recognition, being the best option, is a 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 investigates the potential of the linear predictive coding technique to develop a stable and robust HMM-based phoneme speech recognition system for applications in robotics. Our system is divided into three segments: a microphone array, a voice module, and a robotic arm with three degrees of freedom (DOF). To validate our approach, we performed experiments with simple and complex sentences for various robotic activities such as manipulating a cube and pick and place tasks. Moreover, we also analyzed the test results to rectify problems including accuracy and recognition score.

Keywords: Speech recognition; phoneme; robotics; human-robot interaction (HRI); linear predictive coding (LPC); hidden Markov model (HMM)

1. Introduction

Today the technologies centred on artificial intelligence (AI) make our lives easier. Speech recognition is one such AI technology that enriches our lives. Speech recognition has successfully managed to pave its way into our lives and changes are also foreseeable in the field of robotics. Robots are destined to perform repetitive tasks with razor-sharp precision and accuracy.

Speech recognition is the ability of the computer to analyse, understand, and interpret the human voice. Generating speech is an easy, intuitive, and fast process. The processing and storage of speech signals is less time and space consuming. Modern speech recognition systems have recognition rates but lack the much-required accuracy to accomplish crucial tasks. Eliminating noise by cleansing the speech signal is one of the major challenges in building a robust and stable speech recognition system and this is our motivation for researching 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, such as the mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), and LPC. Based on the results, we preferred LPC because our system had successfully recognised the commands with the highest level of accuracy (96.64%).

Our target application is the human-robot collaboration domain, in particular social robotics in which 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. A stable speech recognition system is a crucial part of this interaction.

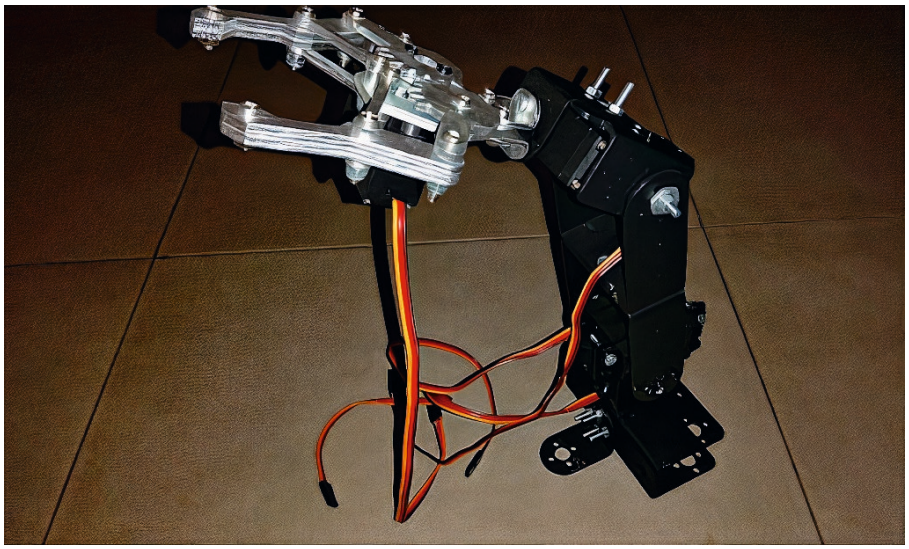

Fig. 1. A 3-DOF robotic arm (photo by author)

The key objectives of our study are:

- achieving wireless control over the robotic arm using voice commands to perform the pick and place operation;
- performing tasks with greater accuracy, higher precision, and better compliance;
- eliminating challenges like background noise by ensuring a smooth interaction between the robot and the operator.

The remainder of this paper is organised 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

https://doi.org/10.37705/TechTrans/e2021002

stage. Section 6 summarises the robot actuation process. Section 7 describes the experimental results followed by Section 8 which summarises the observations. Section 9 concludes the manuscript and enumerates the future scope.

1.1. Related work

Previous research has employed speech recognition to search for casualties in disaster-prone areas (Alifani, Purboyo, Setianingsih, 2019). The system was installed in an unmanned aircraft (drone) with the help of a micro-computer that was programmed to capture a variety of sounds and commands that are generally spoken by casualties 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. The 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 reflect the fact that the system is capable of recognising the sounds/commands at a maximum distance of six metres with 70% accuracy. Also, one of the major drawbacks of the system was the lack of optimal filtering to eliminate noise.

Other research proposed an intelligent method called (THHMB-STOI) to eliminate noise from a densely contaminated speech signal (Alifani, Purboyo, Setianingsih, 2019). The Twin-HMM model is used for extracting features from the signal and for the enhancement of the speech signal. This method is an intelligible prediction framework that measures short-time objective intelligibility (STOI) to produce clean and accurate speech. Its usage in voice recognition systems has drastically improved accuracy by improving the speech transmission index (STI) and articulation index (AI). 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 of the methods mentioned above are non-intrusive and use MFCC for feature extraction. In other words, the methods can be implemented without the requirement for specific hardware. Additionally, the accuracy of these methods is low compared to the acoustic modelling technique proposed by the authors in previous research (Palaz, Magimai-Doss, Collobert, 2019) for the HMM-based speech recognition system. Using convolutional neural networks (CNN) to train the HMM model is a peculiarity of this technique. The ability to generate intermediate feature representations with a subsequent filtering approach using CNN is the main advantage of this technique. Moreover, the system improves its accuracy in recognising commands by a large margin; however, there is considerable loss of information or aliasing witnessed in the method mentioned above. This drawback is overcome by a technique that is mentioned in literature (Zhou, Schlüter, Ney, 2020). 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.

The author of another paper (Ande, Kuchibotla, Adavi, 2020) proposes an interesting application of HMM, namely audio-visual speech recognition. In this study, visual features are combined with audio features using the early integration method followed by the classification of speech using the hidden Markov model. The Gammatone frequency cepstral coefficient (GFCC) and optical flow (OF) are integrated seamlessly 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 was performed on a Hindi language database.

It comes as no surprise that speech recognition has successfully managed to pave its way in the field of robotics to perform difficult tasks varying from the control of a robotic car (Bahar, 2020) to commanding a PR2 robot (Novoa, 2018)

to perform clandestine tasks. The papers introduced a robust speech recognition system by combining a deep neural network (DNN) with the hidden Markov model (HMM). In other research (Bahar, 2020; Ting, 2019; Zhou, Schlüter, Ney, 2020). the authors reviewed a speech-controlled automation system (SCAS) which is also closely related to our application as well. The importance of speech recognition in Human-Machine Interaction is briefly discussed in three papers (Abdelaziz, Kolossa, 2016; Kennedy, 2017; Ting, 2019) in which the authors propose various applications integrating the principles of speech recognition, robotics and machine learning to solve many problems in the field of social robotics. Apart from industrial applications, speech recognition has also witnessed rapid growth in the field of social robotics (Naik, 2020; Bendel, 2020; Bongomin et al., 2020). The application of voice recognition controlled computers (VRCC) is another conventional assistive technique usually used to control most ABB and KUKA robots (Naik, 2020; Abdelaziz, Kolossa, 2016). It comes with a typical human-robot interface which ensures that the commands given as input by the end-user are recognised by the robot. This technique has been used for torch welding and trajectory planning purposes (Alifani, Purboyo, Setianingsih, 2019; Bendel, 2020; Charles, Vishwas, Ruixi, 2020).

Cooperation and collaboration is another important application of robotics (Naik, 2020; Palaz, Magimai-Doss, Collobert, 2019; Sharma, 2019]. 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 such as care homes, rehabilitation centres, etc. According to the author, the *cobot*, as it's referred to in the paper, was successfully deployed into sensitive areas like hospitals and rehabilitation facilities taking into consideration the safety and reliability of the patients surrounded.

1.2. Motivation for using hmm

In this section, we briefly discuss the merits and demerits of using the HMM-based algorithm as opposed to the-state-of-the-art artificial neural network (ANN) based algorithms.

Hidden Markov models (HMM) are generative models based on the Markovian assumption that the current state S' depends only on the previous state S. On the other hand, recurrent neural networks, a variant of ANN widely used for modelling sequenced data such as voice signals, find spurious patterns in the data and overfit. This makes the prediction process difficult and the returned sequences (output) of the RNN becomes irregular with errors. HMM is simple and robust compared to RNN when the dataset used is not complex and smaller in size. By contrast, for a large-scale dataset, RNN is a perfect candidate.

The aspect of non-linearity greatly affects the applicability of RNN for the application of speech recognition. Linear models like HMM. Here, **linearity** is used in terms of the way HMM models function – the weight matrices of the HMM model are normalized and add up to unity. By contrast, RNN's weight matrices are normalised based on a different criteria governed by the activation function.

For our application, we required an algorithm that uses the part-of-speech (POS) technique for mapping the words in commands. POS tagging is based on tagging of the nouns, verbs, and adjectives in a command or sentence. RNN, on the other hand, adopts a completely different procedure based on **one hot encoding** the input sequence. This justifies HMM being a better candidate for our application than RNN.

1.3. Motivation for using lpc

In this section, we present the motivation for adopting the linear prediction technique for encoding our speech signals. From a theoretical viewpoint, linear prediction is a maths-intensive technique which requires the thorough

https://doi.org/10.37705/TechTrans/e2021002

calculation of many variables based on input data such as window length, frequency of the input sequence at each time step, spectral density etc. On the other hand, implementing this algorithm on hardware is, comparatively speaking, a simpler task that requires no special equipment. This makes it cheaper and computationally efficient in comparison with other techniques. However, adopting techniques based on neural networks is computationally intensive and costly.

From a practical viewpoint, our application of primary interest is encoding speech into a digital sequence with a relatively low bit rate for the ease of communication between the speaker and the robot. This makes it easy for the robot to recognise the speech. However, minor fluctuations in accuracy are observed in an environment highly-perturbed with noise, which is a disadvantage of this algorithm. The main advantage of LPC is that it produces an output sequence with correct spectral properties.

2. Methodology

This section gives a brief overview of the preparation of the speech dataset, training the hidden Markov model (Charles, Vishwas, Ruixi, 2020; Bongomin, et al. 2020; Abdelaziz, Kolossa, 2016; Kennedy, 2017) and the underlying principle of the LPC technique used for extracting important data from the audio (speech) input. Moreover, we briefly discuss the advantages and limitations of the LPC technique (Becker, 2016).

2.1. Training stage

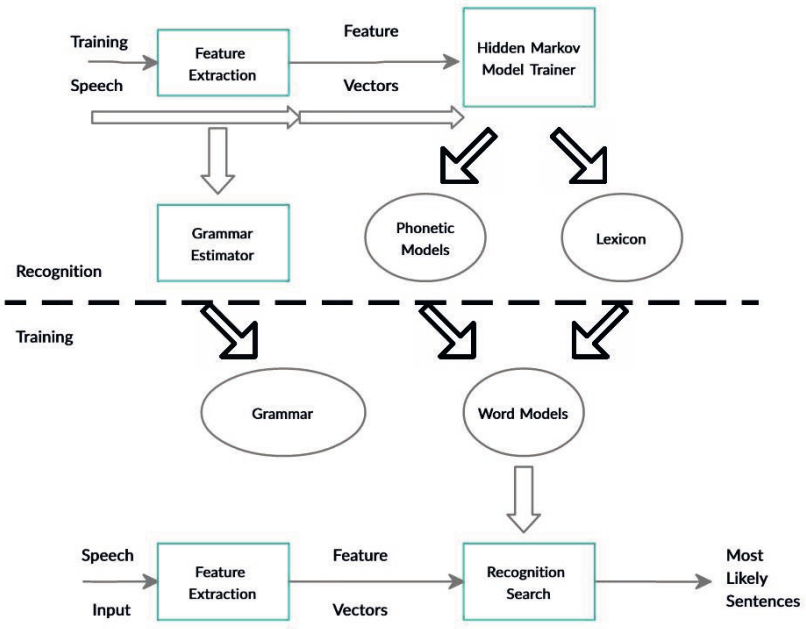

Fig. 2. Automatic speech-recognition system (based on http://www.creately.com)

The training procedure is divided into two parts; the training stage and the recognition stage (Fig. 2).

2.2. Dataset preparation and structure

Datasets have always been instrumental for the advancement of speech recognition technologies. In this section, we present a small-scale audio dataset of human speech – **ROBSPEECH**. It contains over 1000 utterances of 150 different commands from 20 speakers, both male and female. Some

commands are listed in Table 1 below. The commands are stored under two categories – male speaker and female speaker. Vital information related to the speaker's audio range, frequency, audio wavelength, and pitch is also included in the dataset.

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 a clockwise direction
4	**Rotate the shoulder**	This command rotates the servo motor attached to the shoulder by 180 degrees in a 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 then transferred to the Arduino microcontroller to control the servo motors.

2.3. Recording the commands

The text corpus was used for recording the commands using a close-talking microphone at a dual-frequency channel of 44.1 kHz at a -6db peak and a 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 with four commands to build the model using the recorded commands in .wav format. The Hidden Markov Model Toolkit (HTK) software was used.

2.4. Feature extraction

Feature extraction is an essential step in the process of speech recognition as it segregates a particular voice sample 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, such as mel frequency cepstral coefficients (MFCC), linear predictive coding (LPA), revised perceptual linear prediction (RPLP), Bark frequency cepstral coefficients (BFCC). We prefered to use the linear predictive coding (LPC) technique, as shown in Fig. 3 below, for the following advantages:

- high computation speed and robustness;
- lower bit rate requirement for transmission.

https://doi.org/10.37705/TechTrans/e2021002

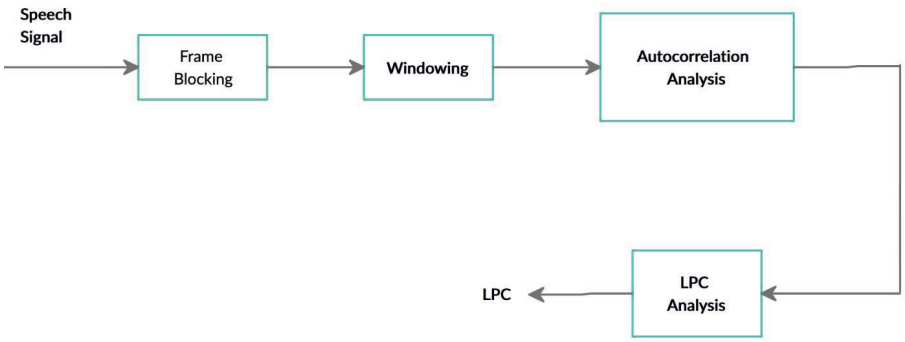

Fig. 3. Block diagram for the LPC technique (based on http://www.creately.com)

2.4.1. Linear predictive coding

Linear prediction coding, in its operation, resembles the human vocal tract. Also known as a formant estimation technique, it is used to estimate the formants and reduce their effects on the signal. Here, the formants are peaks or local maxima 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 the 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:

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

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

The prediction error is given (Charles, Vishwas, Ruixi, 2020):

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

After the speech signal is pre-processed, it is passed for frame blocking as shown in the above block diagram (Fig. 4). Each frame is autocorrelated and the highest autocorrelation value is chosen for linear predictive analysis (Lakomkin, 2018; Ninh, 2019). In linear predictive analysis, the coefficients are calculated, which are given by (Charles, Vishwas, Ruixi, 2020; Bongomin, et al. 2020):

$$a_m=\left(\log\left[1-k_m\right]/\log\left[1+k_m\right]\right) \tag{3}$$

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 is thus a very effective technique employed in the tonal analysis of stringed instruments like the violin and the guitar.

3. 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).

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

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 (ak)	16
7	linear prediction cepstral coefficients	28

3.1. Model training and deployment

As shown in Fig. 4, HTK, **HRest**, **HInit**, and **HERest** programs are used to make the acoustic model. The parameters are initialised 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.

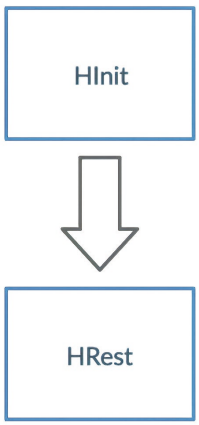

Fig. 4. HTK programs for model creation (based on http://www.creately.com)

4. 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.

4.1. Likelihood with the forward algorithm

Given that we have successfully modelled 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:

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

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

4.2. Hardware

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

The voice module (Fig. 5) 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 SC57X series digital signal processor 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.

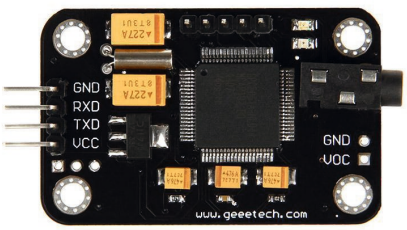

Fig. 5. Speech recognition module (source: http://geetech.com)

https://doi.org/10.37705/TechTrans/e2021002

5. Robot actuation and recognition

speech recognition processes are shown in Fig. 6,. 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, the servo motor is attached to the particular joint viz. The shoulder, elbow, or the gripper actuates and as a result of which, the movement occurs.

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, a decision is made by the comparison stage. In the case of a mismatch, background noise, or the wrong pronunciation of a command, the speech recognition system fails to identify the phonemes in the command, and as a result, the robot is not actuated.

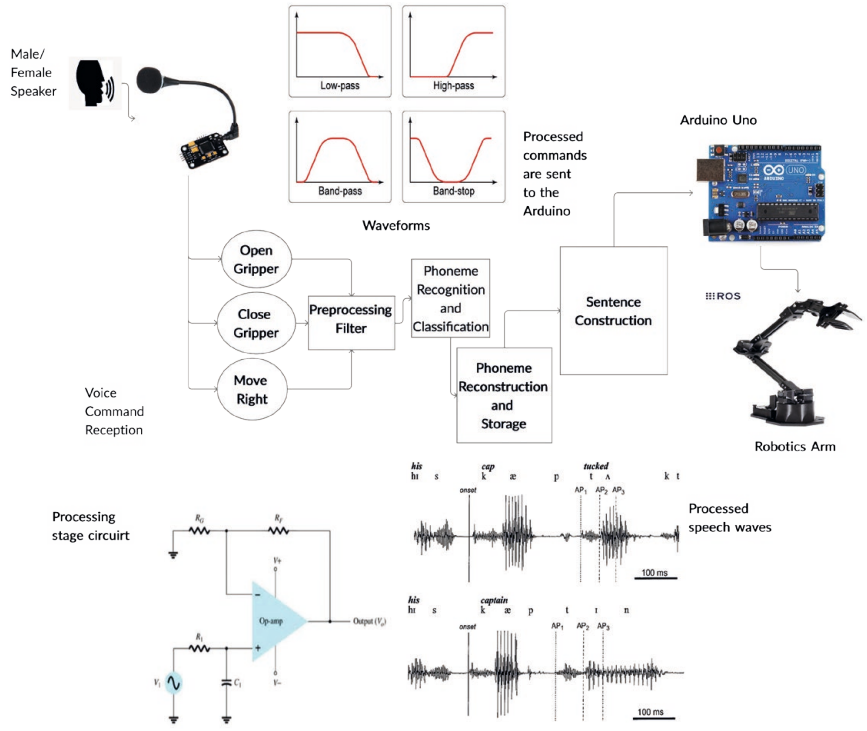

Fig. 6. Speech recognition to control the robot (based on http://www.creately.com)

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

6. Experimental results

The training and testing for the speech recognition system were performed using the recorded commands. 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 is used in double differentiation mode. Here, we have used the HTK toolkit to train the acoustic model and deploy it.

In the experiments, we evaluated word error rate, recognition score, and word accuracy rate – these are shown in Table 3 below.

Word Error Rate:
Word error rate (WER) is the measure of the difference between the recognised word sequence and the input word sequence. It is the most commonly used performance metric for speech recognition systems. Its computation is based on the Levenshtein distance. WER is calculated at the phoneme level. WER is given by

$$\mathrm{WER} = (S + D + I)/N \tag{4}$$

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 percentage word accuracy and is defined as %WAcc = 100 - %WER. It should be noted that the word accuracy can be negative. WAcc is given by

$$\mathrm{WAcc} = (N - S - D - I)/N = (H - I)/N \tag{5}$$

Where H = number of words that are correctly recognised

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	**96.64****
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

7. Observations and conclusion

To validate the performance of our speech recognition system, we conducted an experiment involving eight speaker volunteers (4 male speakers and 4 female speakers) for testing accuracy. Recognition scores are shown below in Fig. 7(a-c).

In Table 3, we observe that Female Speaker 2 was able to attend a remarkable recognition score **of 96.64%** which is the highest of all volunteers. Additionally, the word accuracy rate maximum for the same candidate was **94.04%** with a word error rate of **5.96%**, which was the lowest of all volunteers. This signifies that the recognition system, although not biased towards a particular voice pattern, is significantly affected by quality, pitch, and other wavelengths of the speech signal. By contrast, Male Speaker 1 has the maximum word error rate

https://doi.org/10.37705/TechTrans/e2021002

and the minimum word accuracy rate. With these observations, we conclude that the female voice tends to rise and fall more dramatically than 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 (Fig. 8(a)) due to which the wavelets (small portions of the waves) overlap, which results in a loss of information. The solution to overcome this problem is the use of an anti-aliasing filter which avoids the waves from folding and thus prevents the loss of information.

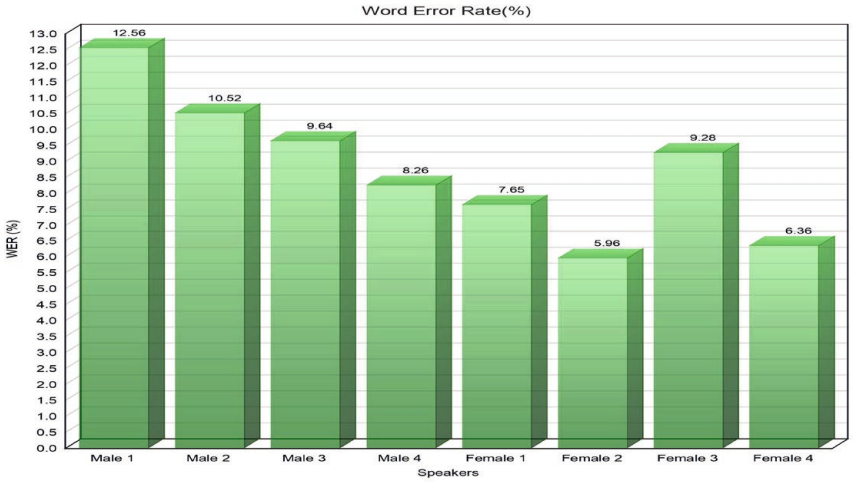

Fig. 7. a) Word error rate graph (by author, based on: http://threegraphs.com)

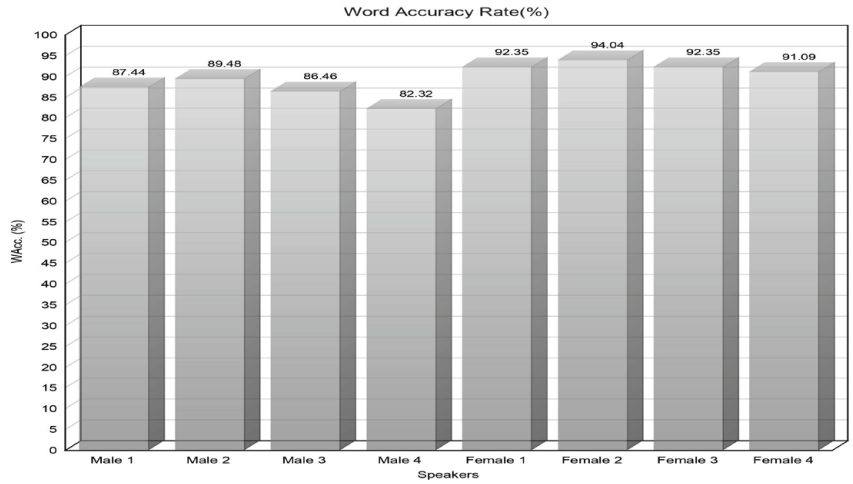

Fig. 7. b) Word accuracy percentage graph (by author, based on: http://threegraphs.com)

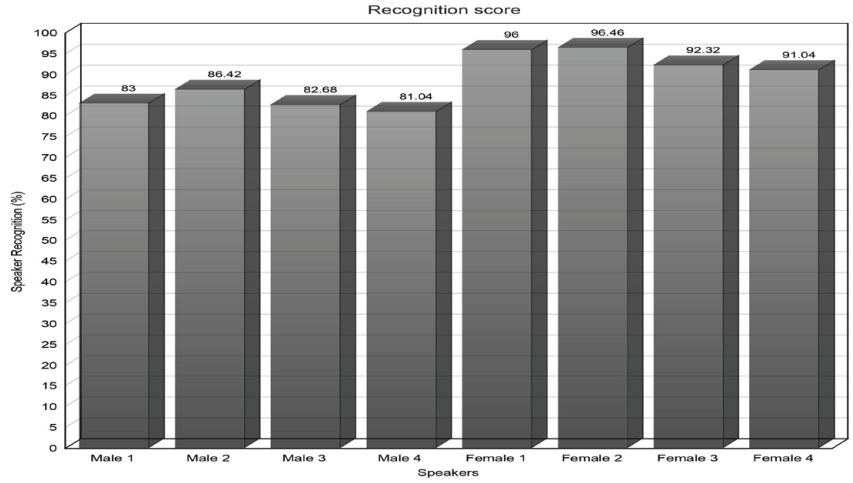

Fig. 7. c) Speaker's recognition score (by author, based on: http://threegraphs.com)

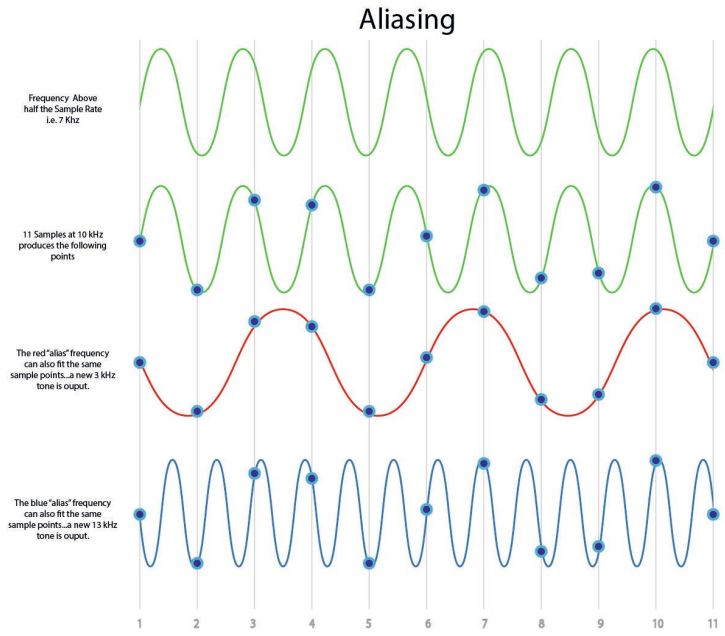

Fig. 8. a) Aliasing or fold over (source: https://www.realhd-audio.com/wp-content/uploads/2014/05/140528_aliasing_illustration.jpg)

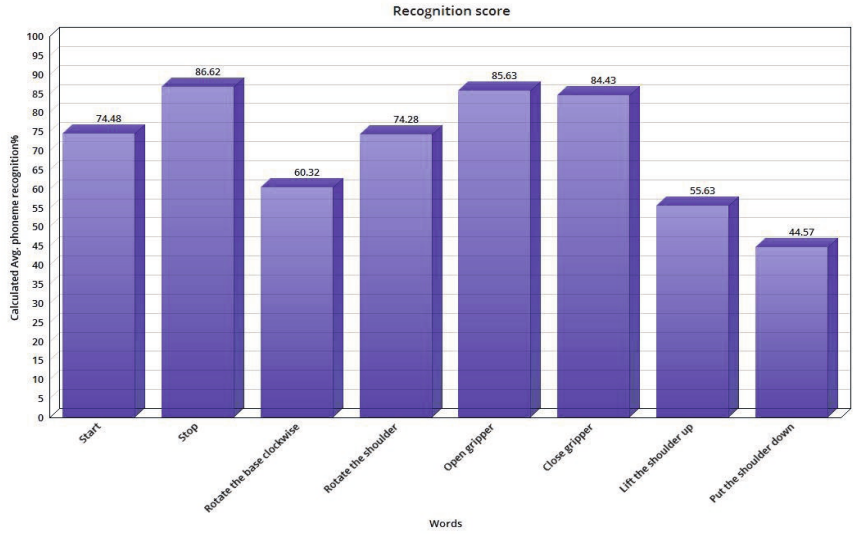

Fig. 8. b) Calculated recognition score graph (by author, based on: http://threegraphs.com)

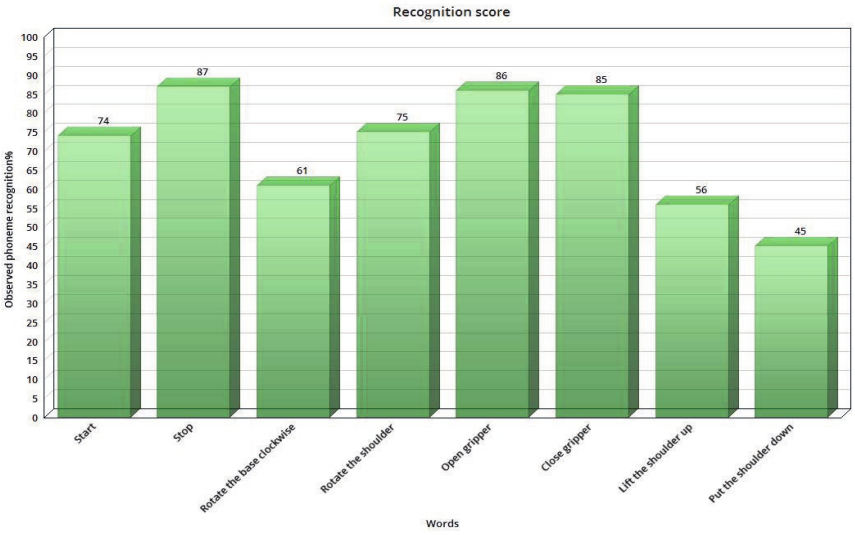

Fig. 9. Observed recognition score graph (by author, based on: http://threegraphs.com)

https://doi.org/10.37705/TechTrans/e2021002

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

8. 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. From the experiments performed so far, we are able to conclude that our system is highly capable of recognising and classifying each phoneme accurately (Fig. 8(b) & Fig. 9); however, we have identified the following limitations:

- There is a reduction in the recognition score when long sentences were used by the speaker (see Fig. 9). There is a significant difference in the recognition score of the **Start** command and **Put the shoulder down command.**
- There is a 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 m).
- The utterance of a word by two 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 briefly discuss them 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** that resembles a tree or a graph with each node having the acoustic properties of a male or a female speaker.
2. **Audio-visual recognition:** This is concerned with combining the optical flow (OF), which is defined as a pattern formation upon the random motion of objects, surfaces, etc. caused by relative motion between the observer and the camera; with our approach, this would allow the system to recognise the object that generates the sound.
3. **The pepper social robot:** Deploying social robots in 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 (Zhou, Schlüter, Ney, 2020); however, their performance is unmatchable for tasks like socialising. Our approach, if scaled properly, can be very effective and, 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 four methods for feature, namely Bark frequency cepstral coefficients (BFCC), perceptual linear prediction (PLP), linear predictive coding (LPA) and Mel-frequency cepstral coefficients (MFCC) for different signal to noise ratios (SNR). Please note that ROBSPEECH is the database we 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 based on noise sources (like the ceiling fan, wall-mounted fan, etc.) present in the laboratory. With this study, we can conclude that our technique outperforms all other techniques.

The article is an extended version of the publication: Adwait Naik: HMM-based phoneme speech recognition system to control and command industrial robots. Authorea. April 20, 2020. DOI: 10.22541/au.158739988.85564998
The authors of the manuscript would like to express their gratitude to Prof. Annu Abraham and Dr. J H Nirmal for their guidance and valuable suggestions.

References

Alifani, F., Purboyo, T.W., Setianingsih, C. (2019). Implementation of Voice Recognition in Disaster Victim Detection Using Hidden Markov Model (HMM) Method. *International Seminar on Intelligent Technology and Its Applications (ISITIA)*.

Alim, S.A., Rashid, N.K. (2018). Some Commonly Used Speech Feature Extraction Algorithms.

Ande, S. K., Kuchibotla, M. R., Adavi, B. K. (2020). Robot acquisition, control and interfacing using multimodal feedback. *Journal of Ambient Intelligence and Humanized Computing*, 1–11.

Bahar, P., Makarov, N., Zeyer, A., Schlüter, R., Ney, H. (2020). Exploring A Zero-Order Direct Hmm Based on Latent Attention for Automatic Speech Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7854–7858.

Baranwal, N., Singh, A. K., & Hellstrom, T. (2019). Fusion of Gesture and Speech for Increased Accuracy in Human Robot Interaction. *24th International Conference on Methods and Models in Automation and Robotics (MMAR).*

Becker, K. (2016). Identifying the Gender of a Voice using Machine Learning. Retrieved from http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning (access: 29/05/2020).

Bendel, O. (2020). Co-Robots as Care Robots. Preprint arXiv arXiv:2004.04374.

Bongomin, O., Yemane, A., Kembabazi, B., Malanda, C., Mwape, M. C., Mpofu, N. S., Tigalana, D. (2020). The Hype and Disruptive Technologies of Industry 4.0 in Major Industrial Sectors: A State of the Art.
M., Abdelaziz, A. H., & Kolossa, D. (2016). Twin-HMM-based non-intrusive speech intelligibility prediction. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

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

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., Belpaeme, T. (2017). Child Speech Recognition in Human-Robot Interaction. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. HRI '17. *ACM/IEEE International Conference on Human-Robot Interaction.*

Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., Wermter, S. (2018). On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

Naik, A. *HMM-based phoneme speech recognition system for control and command of industrial robots.* Preprint arXiv:2000.01222, 1–23.

Ninh, D. K. (2019). A Speaker-Adaptive HMM-based Vietnamese Text-to-Speech System. 2019 11th International Conference on Knowledge and Systems Engineering (KSE). *11th International Conference on Knowledge and Systems Engineering (KSE).*

Novoa, J., Wuth, J., Escudero, J. P., Fredes, J., Mahu, R., Yoma, N. B. (2018). DNN-HMM based automatic speech recognition for HRI scenarios. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 150-159).

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

Sharma, U., Maheshkar, S., Mishra, A. N., Kaushik, R. (2019). Visual Speech Recognition Using Optical Flow and Hidden Markov Model. *Wireless Personal Communications*, *106*(4), 2129–2147.

Ting, W. (2019). An Acoustic Recognition Model for English Speech Based on Improved HMM Algorithm. In *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 729–732.

Zhou, W., Schlüter, R., Ney, H. (2020). Full-Sum Decoding for Hybrid HMM based Speech Recognition using LSTM Language Model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7834–7838.
http://geetech.com (access: 29/05/2020).
http://threegraphs.com (access: 29/05/2020).
http://www.creately.com (access: 29/05/2020)

https://doi.org/10.37705/TechTrans/e2021002