Robotica: (2021) volume 39, pp. 1779–1790. © Iran University of Science and Technology, 2021. Published by Cambridge University Press.
doi:10.1017/S0263574720001496

Design and Implementation of the Voice Command Recognition and the Sound Source Localization System for Human–Robot Interaction

M. H. Korayem†*, S. Azargoshasb†, A. H. Korayem‡ and Sh. Tabibian¶

†*Robotic Research Laboratory, Center of Excellence in Experimental Solid Mechanics and Dynamics, School of Mechanical Engineering, Iran University of Science and Technology, Tehran 1684613114, Iran*
E-mail: azargoshasb71@gmail.com
‡*Mechanical and Mechatronics Engineering Department, University of Waterloo, ON N2L 3G1, Canada*
E-mail: amin.korayem@uwaterloo.ca
¶*Department of Mechanical and Mechatronics Engineering, Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran*
E-mail: sh_tabibian@sbu.ac.ir

(Accepted December 17, 2020. First published online: March 15, 2021)

SUMMARY
Human–robot interaction (HRI) is becoming more important nowadays. In this paper, a low-cost communication system for HRI is designed and implemented on the Scout robot and a robotic face. A hidden Markov model-based voice command detection system is proposed and a non-native database has been collected by Persian speakers, which contains 10 desired English commands. The experimental results confirm that the proposed system is capable to recognize the voice commands, and properly performs the task or expresses the right answer. Comparing the system with a trained system on the Julius native database shows a better true detection (about 10%).

KEYWORDS: Voice command; Human–robot interaction; Hidden Markov model; Social robotic; Robotic face.

1. Introduction

Nowadays, robots are becoming more socialized and they have been used in human daily lives more than before. In terms of human–robot interaction, robots should have communication capabilities including speaking, listening, sensing, and interaction skills. Although there are different ways to communicate with the robots, such as keyboards, mouse, touchscreen, joystick, sign language for deaf people,[1] vision-based interaction,[2] and human intention,[3] the ability to communicate via speech is more suitable for increasing social interaction with robots.

Speech recognition and especially voice command detection is an important technology in human–robot social interaction. Kennedy et al. evaluated some automatic speech recognition (ASR) engines under a variety of conditions, inspired by real-world social HRI conditions with a focus on interactions with children.[4] Ivaldi et al have assessed the influence of two factors, namely extroversion

* Corresponding author. E-mail: hkorayem@iust.ac.ir

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

and negative attitude toward robots, on speech and gaze during a cooperative task.[5] Also, human–robot interaction (HRI) can be used as a non-invasive screening tool for the assessment of cognitive decline.[6] Grigore et al. have considered how two basic input communication modalities (speech recognition and touch-based selection) can shape users' perceptions of the agent.[7]

Speech recognition systems have been implemented on different robots using different algorithms and tools. Fardana et al. have presented a real-time system for controlling a small mobile robot via speech and gesture using a Kinect sensor. The commands were recognized using a hidden Markov model (HMM)-based approach with a limited vocabulary. The overall recognition rate of 85% and the average execution time of 2.2 s were reported for the speech recognition system.[8] Gomez et al. have proposed an interface that is controlled based on voice and lip-reading. Voice command detection system was modeled using an HMM-based isolated word recognition approach. The detection rate was 100% for the system with eight voice commands.[9] Mustofa[10] used the MFCC features and ANFIS methods for the recognition of sound patterns. The results were tested on the mobile robot. The runtime and recognition rates were not mentioned.[10] Zinchenko et al.[11] presented a voice command detection system for the control of the three degrees of freedom of a robotic endoscope holder. The system was implemented on Ubuntu, using CMU Sphinx. The average recognition rate was 90%, with a standard deviation of 5%.[11] Burger et al. described a flexible multimodal interface to control a mobile robot named Jido. The speech recognition system was developed based on the open-source Julius engine and reached an accuracy of 84.71%. The achieved accuracy was acceptable considering that near half of the speakers were non-native.[12] Abdullahi et al. developed a voice recognition system using a development methodology known as the waterfall model to control the robot's motion using the voice. The average rate of voice recognition was 80%.[13] Abed et al. used the Dynamic Time Warping (DTW)-based speech recognition for five isolated Arabic words. The results have shown a good recognition rate with a moderate time.[14]

According to the purpose of HRI and type of the robots, the interaction system can be designed in different ways. The HRI has been considered in various areas from giving single-word commands to the robots to more complicated interaction with social robots. In this paper, a verbal communication system has been designed and implemented that can be used for different purposes in HRI applications. Accuracy, robustness, runtime, and cost for hardware implementation are important aspects that have been considered for developing a practical system. Due to the hardware requirements, number, and microphone configuration, a low-cost communication system is proposed, which is the main contribution of the paper as similar tasks have been done with a more complicated system in the literature.[8] The voice command detection system is designed based on the hidden Markov model, which has shown acceptable results in previous works.[8,9] Moreover, since the Persian speakers are supposed to use the system, a non-native speech database for 10 English commands using Persian (non-native) speakers has been collected/recorded for training the algorithm to increase the performance of the system. The execution time for recognition of the input command with an average duration of 3 s is 0.9 s, which is improved when compared with ref. [8]. This enables the proposed system to practically be implemented on the robotic hardware. Moreover, we can integrate some other tools including the voice localization package of the HARK (robot audition system)[15] and the voice synthesis module into our proposed system using the ROS middleware to improve the social interaction level. A comparison between the speech recognition package using the native database and the proposed HMM-based system using the non-native database is provided. Furthermore, it has been experimentally shown that the Scout and facial robots are capable to detect the speaker's direction, turn toward the speaker, and express the appropriate response in addition to recognizing the voice commands based on the designed algorithm.

The remaining of this paper is as follows. The HMM-based system and the non-native database are introduced in Section 2. In Section 3, the experimental setup and the hardware implementations are explained. Finally, the experimental tests and results evaluations are presented in Sections 4 and 5.

2. Methodology

2.1. HMM-based voice command detection system

The purpose of this paper is to detect single-word commands based on the hidden Markov models. The HMMs are the most popular acoustic models for modeling the time-varying nature of speech

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

signals. The parameters of the HMM are the number of the states, the transition probability matrix between the states, the initial state presence probability, and the observation matrix, which determines the distribution of observations in each state of the model. The main voice command detection system consists of features extraction, acoustic modeling, and decoding phases.

Various feature extraction methods are used in speech processing techniques. Some of the most famous methods are Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive (PLP), Coefficients, Mel-Frequency Cepstral Coefficients (MFCC), and Wavelet-based feature extraction methods.[16] In the feature extraction phase of our proposed voice command detection system, the MFCC features are extracted from the input voice command. They have been used due to their acceptable performance in most of the speech recognition and voice command detection systems. Each voice command has been segmented to 25 ms frames using the Hamming window and with 50% overlap between frames. Thirty-nine MFCC coefficients, which include 12 main coefficients, 1 energy coefficient plus the first and second derivatives of these 13 coefficients are extracted from each segment of the input voice command. In order to compensate for the effects of channel noise, we normalized the 39 MFCC coefficients using the cepstral mean and variance normalization method (CMVN).[17]

The acoustic models can be trained based on different phonetic units such as words, phonemes, triphones, diphones, syllabus. In this paper, the phonetic unit is the word. The proposed speech recognition system has been trained based on a word unit using the HMM toolkit (HTK).[18] It means that a hidden Markov model is trained for each word and also the probable silence at the start, end, or between the words of the input command. Finally, 11 word-based HMM models have been trained for silence and 10 control words. Each model includes 8 states (input and output states plus 6 internal states) with 16 Gaussian mixtures in each state to model the observation vector of that state.

The decoding phase is the final phase of the voice command detection system. Each voice command could consist of one or more words. The arrangement of the words in the voice commands has been determined using a dictionary. The Viterbi decoding algorithm[18] has been used to detect a whole voice command based on its constituent words using the mentioned dictionary. In this phase, we used single-word commands. However, the proposed voice command detection system can be used for commands with more than one word.

Speech databases play an important role in the performance of speech recognition and voice command detection systems. Although there are many native datasets containing voice commands for controlling the robots, there are not any available non-native dataset of the mentioned voice commands. It is clear that *in the face of non-native speakers*, the voice command detection system trained on the non-native dataset has a better performance compared with the voice command detection system trained on a native dataset. Therefore, in this paper, a database consisting of non-native speeches has been recorded to train the system. The mentioned database is introduced in the following sub-section.

2.2. Preparing the non-native database

According to the application of the speech recognition or voice command detection system, the database can be small (includes 1–99 words), medium (includes 100–999 words), and large (includes more than 1000 words). In this paper, a small database with 10 control words including the words "ready", "forward", "backward", "left", "right", "stop", "move", "faster", "slower", and "go" has been used. These words are used as a single-word voice command for the robot.

The database consists of 18 speaker's speeches and each word is spoken 3 times by each person. Thus, the whole database includes 540 wave files (30 files for each speaker). Moreover, the data are recorded in the .wav format, single band, with a sampling frequency rate of 16,000 Hz and 32-bit size. In practice, it is better to record speech in the conditions (such as microphones, environments, noisy situations) similar to the conditions in which the robot is used.

A graphical user interface (GUI) is designed for recording different commands as shown in Fig. 1. In this GUI, the audio files can be recorded individually by entering each person's ID and gender of the speakers.

After recording the dataset, we labeled the recorded commands at the word level. The labeling can be done manually or automatically. Fully automated labeling systems are not accurate and should be used in combination with manual methods. In this paper, the dataset is labeled manually.

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

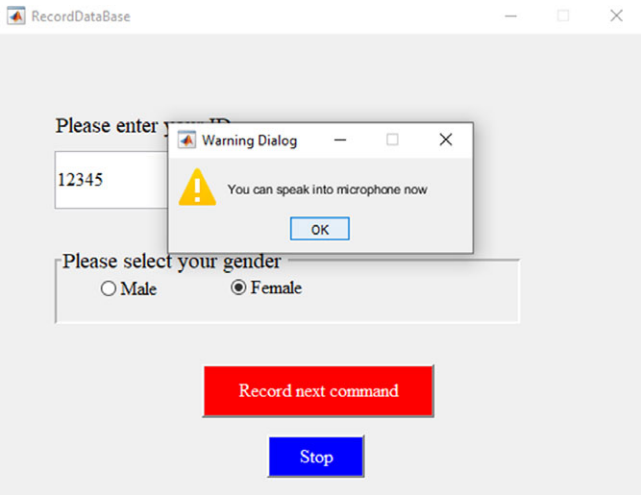


Fig. 1. The designed GUI for recording commands.

3. Experimental Setup

3.1. Sociable robotic face

The ability to perform different facial expressions is an important factor for HRI. Meghdari et al. have indicated that the use of robots that can perform different facial expressions is effective in helping to improve facial expression understanding disorders in children with autism.[19–21] These facial expressions can be generated by moving various facial features such as eyebrows, lips, and eyelids. Each of these components plays an important role in communicating and transferring facial expressions in human communication.

In this paper, the movement of the lips, eyebrows, and neck of the robotic face, plus eye color changes are considered. According to surveys done,[22–25] a robotic face is designed to be able to express happiness, anger, and sadness by moving the mentioned facial elements. Accordingly, a robotic face with five degrees of freedom is designed and built. The robotic face has two degrees of freedom in the eyebrows. Each of the eyebrows rotates independently around a parallel axis with a sagittal axis. Independent movement of eyebrows empowers the robot to express other emotions such as surprise, disgust, or suspect.

The semantic role of the lips in different facial expressions, as well as the high range of lips movement in different facial expressions, add to the complexity of the design because the length of the lips varies in different facial expressions. Finally, using two degrees of freedom for each corner of the lips, the various forms of the lips are well achieved in all facial expressions. Two servo motors are used for generating rotational motion to the corners of the lips.

Studies have shown that eyes are of great importance in human social communication. Due to the simplifications made in the design of the face, the LEDs are embedded inside the eyes to enhance the desired emotion by the face, therefore, they could have a more effective impact on the audience by changing the eyes' color in different modes. Correspondingly, red is used for anger (Fig. 4(c)), white for happiness (Fig. 4(b)), and blue for sadness (Fig. 4(a)).

Also, the robot is capable to turn toward the sound source using a servo motor as a neck to show attention to the speaker. The designed robotic face and the various facial expressions are shown in Figs. 2–4.

3.2. The Scout robot

The Scout robot has two mechanical arms placed on the mobile body. The base of the robot has two main wheels. Each arm has five degrees of freedom, which performs pitch and roll operations by DC motors. To control the robot with voice commands, each command after detection is converted to a robot's usable command and applied to the robot using the kinematic model of the robot. A view of the Scout robot is shown in Fig. 5(a).

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

Table I. Denavit–Hartenberg parameters of Scout robot.[27]

Axis	θ	α(deg)	αm	d(m)
1	θ_1	$-\phi/2$	0	$d_1 = 0.06$
2	θ_2	0	$a_2 = 0.1$	0
3	θ_3	0	$a_3 = 0.21$	0

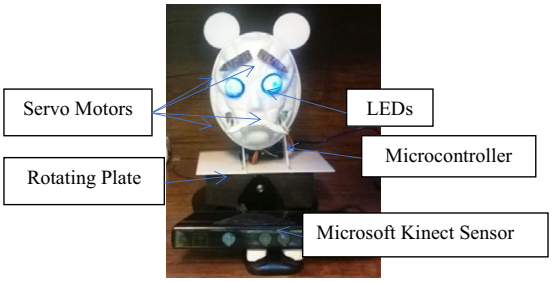


Fig. 2. General schema of the hardware system.

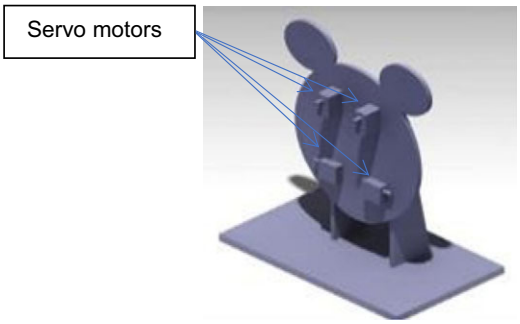


Fig. 3. 3D model in CATIA with SG90 servo motors.

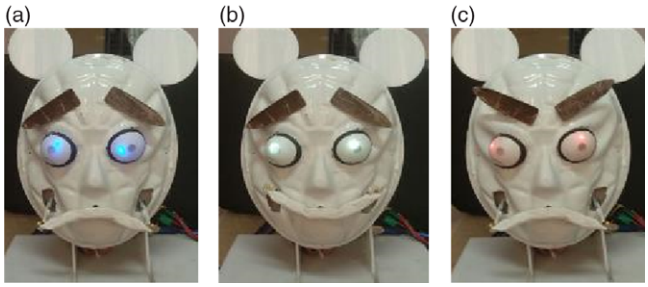


Fig. 4. Facial expressions of robotic face (a) Sad. (b) Happy. (c) Angry.

To transfer the voice command to the robot's movement, the equation of the motion of the robot is needed. The kinematic and kinetic of the Scout robot are derived based on the Denavit–Hartenberg (D-H) method as shown in Table I. The dynamic of the robot is introduced in ref. [26] and based on the D-H parameters (Fig. 5(c)), and the dynamic equation of the motion of the system is presented in ref. [27].

The inverse kinematic of the Scout robot is introduced in ref. [28]. It requires to transfer the end-effector desired path to the joint space. Therefore, for each command, the joint space desired values ($q_{i,des}$, $i \in [1, 2, 3]$) is obtained as the desired path (X_{des}, Y_{des}, Z_{des}) is provided. For instance, if the "forward" command detects by the robot, the desired path would be ($X_{des} = 0$, $Y_{des} = 0 : 0.1$, $Z_{des} = 0$), which means the robot starts to move forward and it continues until the next command will be detected.

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

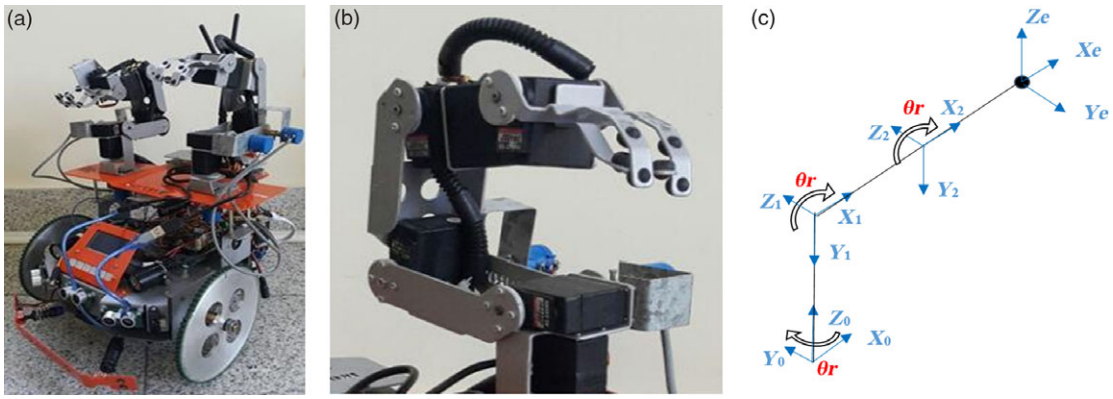


Fig. 5. Scout robot. (a) Scout robot front view. (b) Scout's arm. (c) Arm coordinates

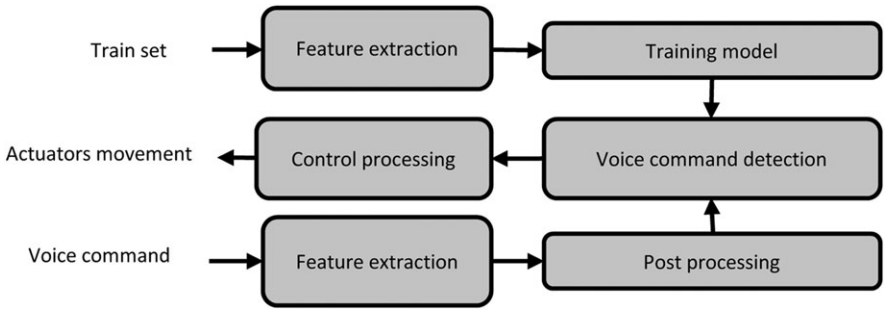


Fig. 6. The overall process of the HMM-based voice command detection system.

3.3. Implementation of the HMM-based voice command detection system on the robot

In this section, the steps for implementing the proposed communication system on the robot are investigated. In the proposed system, each voice command is converted to a motion command and applied to the robot. ARM 32-bit core microcontroller is used to control the actuators of the robot. The robot is waiting for voice commands by simultaneously executing speech detection and the control unit. The user can give the voice command to the robot by pressing a button from the keyboard or a GUI-based pushbutton at any time. The control unit then provides the actuation signal as it receives the analyzed voice command. Eventually, the robot does the desired action as the user asked by the voice command. The overall procedure of the designed system is shown in Fig. 6.

To communicate remotely and get rid of wire connections between the PC and robot, which may interrupt the human–robot interaction, the HC-05 Bluetooth module is used. This module receives and sends the data at specified intervals using the Bluetooth protocol. The HC-05 module communicates with the electronic interface board via the serial protocol. Finally, the voice is received and processed using a PC, and then the result is sent via the Bluetooth module to the robot control board and the robot performs the control command. The schematic of the overall system hardware is shown in Fig. 7.

There is also a sound source localization package using HARK and a voice synthesis package and control packages to increase the level of robot interactions. Therefore, for communication between different modules and packages, ROS middleware has been used and the processing result of each package is published through ROS messages and topics to be available for other packages. This software is capable to run different packages at the same time. In order to control the robot using the voice command, communication between the MATLAB, HARK, ROS, and microcontroller has been established. The steps are as follows: the voice command is received and processed by the MATLAB and speech recognition is done. On the other hand, the HARK performs the sound source localization using the Kinect microphone array. Finally, the results of speech recognition and sound source localization, after the required processing, are published through ROS topics and received through the ROS serial by the ARM controller. Depending on the control system designed in the

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press



Fig. 7. The hardware schematic of the HMM-based voice command detection system.

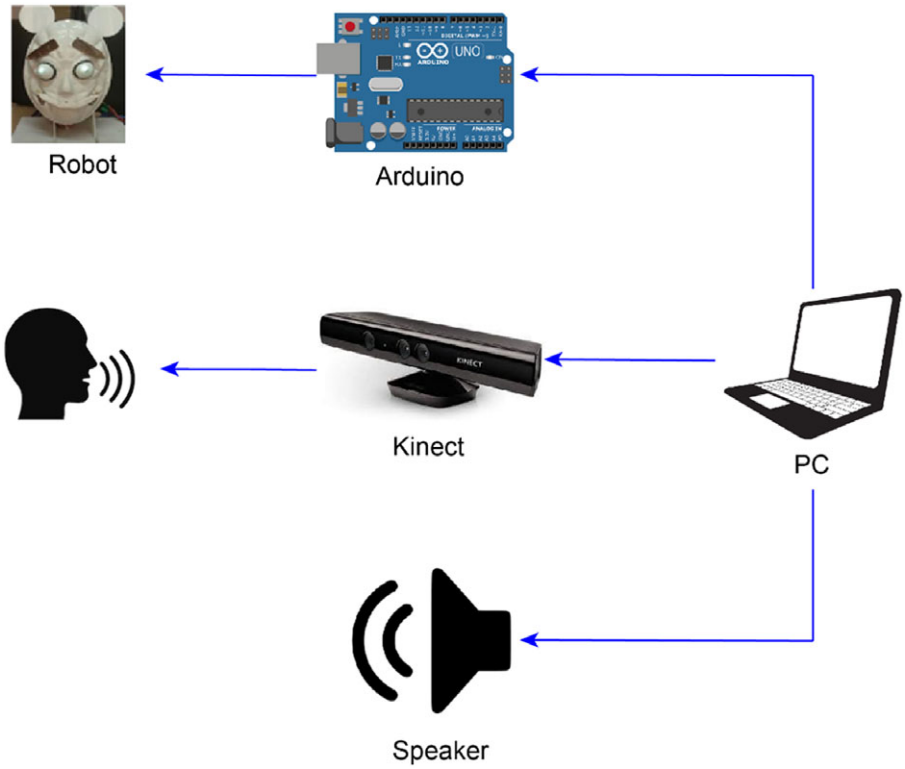


Fig. 8. Hardware schematic of the whole communication system.

ARM controller, the motors move and the corresponding tasks are performed. The responses are also determined and expressed by the voice synthesis package.

In the final designed system, the connection between the various hardware devices needs to be established. The system includes different hardware devices such as Microsoft Kinect sensor, Processor unit, ARM controller, robot, and a speaker. The hardware schematic of the designed system is shown in Fig. 8. The system is designed to receive sound by the Microsoft Kinect sensor's microphone array and send it to the PC. The result – after processing in different modules – is sent to the ARM controller and the movement control of the robot is carried out. The result of the sound synthesis package is also played through the speaker.

4. Results

In this section, the results of the designed model are presented.

4.1. The results of HMM-based voice command detection system

In this paper, a database of 10 English words is collected using 18 different speakers. Then, the data are labeled and used to train the model. The dataset is divided into test and train sets in such a

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

Table II. The test results of the HMM-based voice command detection system.

	Stop	**Go**	**Left**	**Right**	**For ward**	**Back ward**	**Slower**	**Faster**	**Ready**	**Move**
Stop	100	0	0	0	0	0	0	0	0	0
Go	0	100	0	0	0	0	0	0	0	0
Left	0	0	100	0	0	0	0	0	0	0
Right	0	0	0	100	0	0	0	0	0	0
Forward	0	0	0	0	100	0	0	0	0	0
Backward	0	0	0	0	0	100	0	0	0	0
Slower	0	0	0	0	0	0	100	0	0	0
Faster	0	0	0	0	0	0	0	100	0	0
Ready	0	0	0	0	0	0	0	0	100	0
Move	0	0	0	0	0	0	0	0	0	100

way that the speakers of the test set are not necessarily the same as the speakers in the trainset. The average execution time for the detection of a single trained command with an average duration of 3 s is 0.9 s. The test results of the voice command detection system are presented in Table II.

As it is clear from the confusion matrix presented in Table II, the word error rate was 0%.

4.2. Experimental test result of the system on the Scout robot

The proposed voice command detection system is implemented on the Scout robot. To do so, the Scout robot actuators (two actuators for the base platform and three actuators for each arm) are controlled to follow the desired robot action provided by the voice command system. Although it is possible to control the robot/machine by more complicated voice commands, like asking the robot to move with the desired distance specified by the user command with multiple words, the single command movement is opted to make the Scout robot's movement simple. Additionally, because of our small database (we collected our non-native database for Persian speakers), the performance of the speech recognition system would be decreased by combining, for instance, two words for one command. Thus, the main focus of this paper is to propose a voice-based interaction unit that can easily implement on the robot and follows single-word commands. The robot can correctly recognize the voice command and perform the required movements (e.g., start/stop moving or turn right/left by 90°). A combination of the system with the kinematic equation has been done in refs. [28, 29] focusing on the navigation of the Scout robot by receiving speech from its surroundings and sending the processed speech to the central processor. They analyzed the processed speech to estimate the speaker's location (distance and direction with respect to the robot's current position). Then, the estimated speaker's location is sent to the robot via the communication system, and the robot moves toward the speaker by taking the advantage of robot kinematic equations[28,29] as shown in Fig. 9 (please see refs. [28, 29] for more similar test results).

In this study, the voice command system is integrated with the kinematics of the robot for real-time application. Thus, the robot could easily, for example, be stopped at the specified distance by kinematics equations. The start of the actuator movements is shown in Figs. 10 and 11 focusing on the real-time application of the system.

The commands are planned in such a way that the robot starts to move with the "go" command and stops with the "stop" command. The robot turns to the right or left 90° with the "right" and "left" command, the speed of the robot is reduced or increased by a certain amount, with "slower" and "faster" commands. The robot moves forward or backward for 10 cm, once the "forward" and "backward" commands are applied. Using the "move" command, the robot moves sinusoidally; and with the "ready" command, the robot arms (Fig. 5(b)) are positioned in the home position of the robot shown in Table I. The speed of the actuators is controlled using Pulse Width Modulation (PWM).

To address the accuracy of the designed method concerning implementation time delay and the speed of voice detection on the Scout robot, the results of the right wheel motor of the Scout robot are considered. The right wheel of the Scout robot is rotated by a DC motor which has an encoder on its rotation shaft.

The test result of the Scout right wheel motor is shown in Fig. 10. In this figure, the change in the Scout right wheel motor's angle and the variable corresponding to the recognized word is shown over time to indicate the synchronization of the programs. As it is shown in the figure, the Scout

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

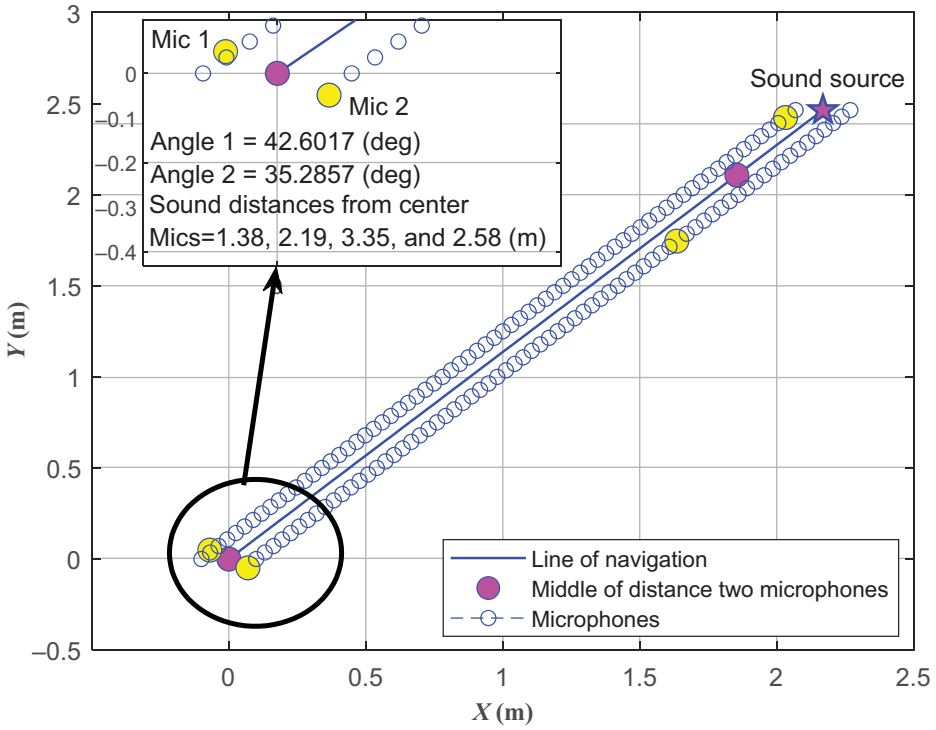


Fig. 9. The Scout robot path toward the speaker by analyzing the speaker's speech for estimating its location.[28]

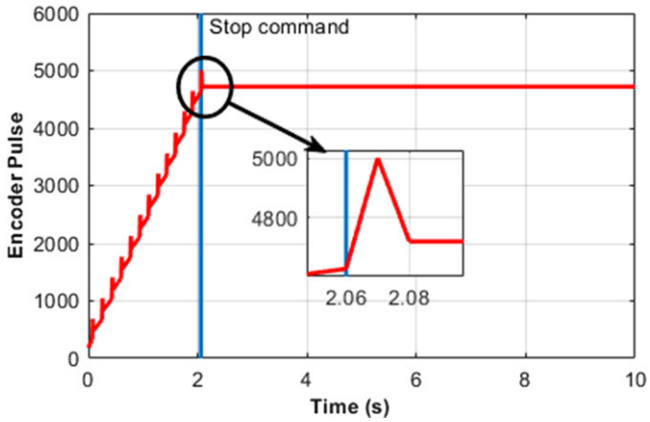


Fig. 10. The test result of STOP voice command.

right wheel motor is stopped with a delay of 0.02 s by receiving a voice command and changing the corresponding command (CommandID).

The ability to receive different commands is tested by applying the 10 voice commands on the Scout robot. The result for the Scout right wheel motor is presented. As soon as a voice command is received, the Scout right wheel motor rotates according to the recognized command. In this test, the Scout right wheel motor starts moving clockwise using voice commands "move", "go", "right", "forward", and "faster" moves counterclockwise using "left", "backward", and "slower" commands and stop moving using "stop" and "ready" commands. If the system does not recognize a command, the Scout right wheel motor will continue to move. The Scout right wheel motor position is recorded by its encoder. It is shown along with the waveform of the voice commands over time in Fig. 11.

As can be seen in Fig. 10, by receiving each voice command, the system recognizes it and the motor moves according to the recognized command.

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

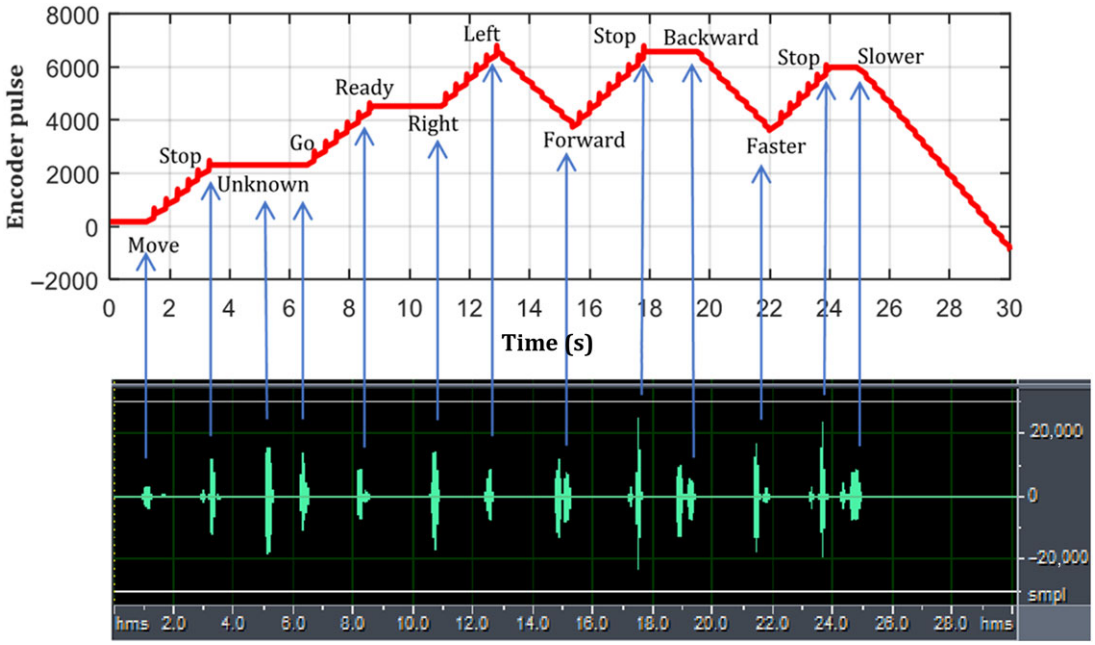


Fig. 11. Experimental test results of 10 voice commands.

4.3. The test results of the system on a robotic face

The sound processing is performed in MATLAB and the detected result is sent by the Bluetooth module to the robot control board. As a result, the robot moves according to the control command and performs the desired facial expression. The movement of different actuators is shown in Fig. 12.

As it is clear from Fig. 12, the facial robot motors are started/stopped with a delay of 0.03 s by receiving a voice command and changing the corresponding command (CommandID), which means that there is a 0.03-second time delay between the facial expression detection and implementation on the facial robot with the designed method.

5. Discussion and Validation

In this paper, a communication system is designed and implemented for human–robot interaction. According to the experimental test results, when the Persian (non-native) speakers use a system that is trained under native conditions, system performance decreases. Therefore, the HMM-based voice command detection system, which is trained with Persian speakers' data has high accuracy (100% word recognition rate). To compare with a native database, the Julius software, and the acoustic models of English have been used and tested for the same 10 English commands. The results have naturally shown lower accuracy in the test with the non-English speakers (85.16% word recognition rate).

The execution time for the detection of a command with an average duration of 3 s is 0.9 s, which is improved when compared with ref. [8]. The experimental tests on a robot show that the system is capable to successfully detect the command and perform the right task. Given the acceptable results, it can be concluded that the HMM-based robot command recognition system can be used in systems that require a precise recognition of a user's command. For example, commanding robots that perform certain tasks such as guiding a wheelchair by voice commands or using a robot as a teaching assistant in an English class.

In this paper, a communication system is designed using low-cost hardware. The test results show that the word recognition rate is higher in the systems designed in this paper compared with other researches such as in refs. [8,12] and [13] in the human–robot interaction area. Also, the experimental results show that the designed systems are capable to control the movement of the robots successfully with acceptable execution time.

6. Conclusion and Future Work

A voice command detection system based on the hidden Markov model was designed and implemented. This system has shown an acceptable recognition rate. Also, integrating available packages

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

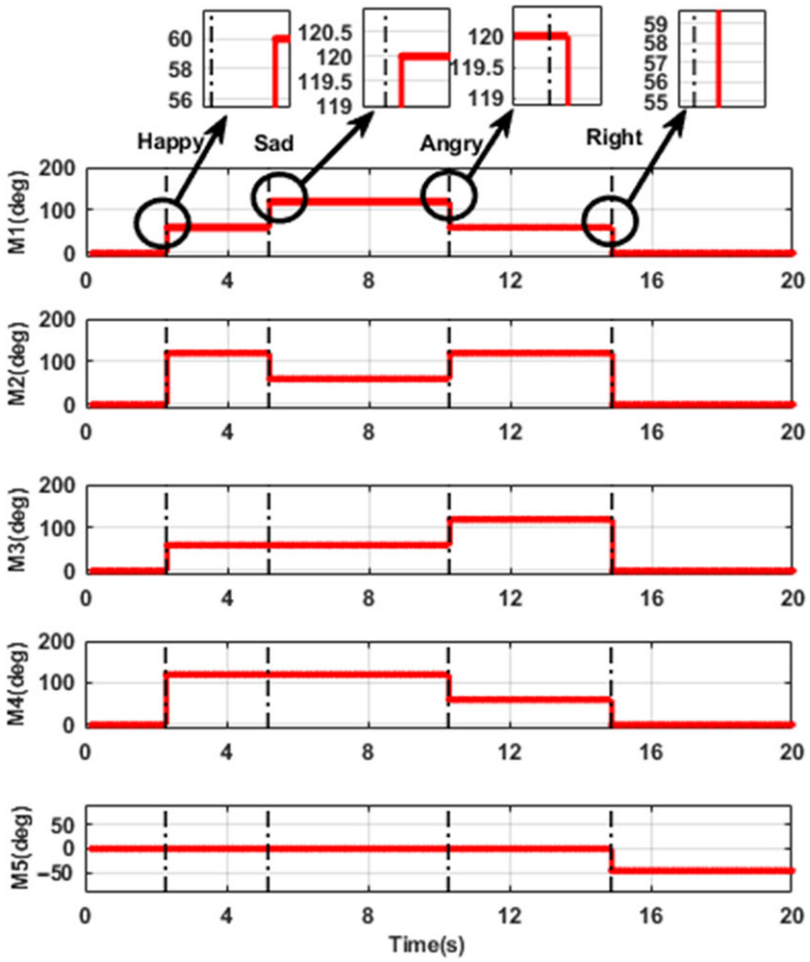


Fig. 12. The robot motors position using the HMM-based voice command detection system.

such as sound source localization and voice synthesis module using ROS middleware has led to improve the social interaction level. Therefore, the presented model is capable to detect the speaker's direction, turn toward the speaker, and express the appropriate response in addition to recognizing and performing the voice commands.

Improving the goals of voice command detection can be achieved by extending the dictionary, acoustic, and language models of the system. However, this situation will lead to a reduction in the recognition rate of the voice command detection system. Therefore, to achieve his improvement, more stability is needed. Although the current system has achieved the goals of this paper, more sophisticated methods can be considered to overcome constraints of recognizing more words, more complicated commands and achieving more realistic relationships between the human and robots. However, it provides a framework that is not only suitable for use as an English language teaching assistant, but also any other robotic platform in human–robot interactions.

References

1. S.-Y. Lo and H.-P. Huang, "Realization of sign language motion using a dual-arm/hand humanoid robot," *Intell. Serv. Rob*. **9**(4), 333–345 (2016).
2. A. Halawani, S. Ur Réhman, H. Li and A. Anani, "Active vision for controlling an electric wheelchair," *Intell. Service Rob*. **5**(2), 89–98 (2012).
3. P. Leica, F. Roberti, M. Monllor, J. M. Toibero and R. Carelli, "Control of bidirectional physical human–robot interaction based on the human intention," *Intell. Serv. Rob.* **10**(1), 31–40 (2017).
4. J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations," *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (2017).

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press

5. S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi and E. Zibetti, "Towards engagement models that consider individual factors in HRI: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *Int. J. Soc. Rob.* **9**(1), 63–86 (2017).
6. S. Varrasi, A. Lucas, A. Soranzo, J. McNamara and A. Di Nuovo, "IBM Cloud Services Enhance Automatic Cognitive Assessment via Human-Robot Interaction," **In:** *New Trends in Medical and Service Robotics* (Springer, 2019) pp. 169–176.
7. E. C. Grigore, A. Pereira, I. Zhou, D. Wang and B. Scassellati, "Talk to Me: Verbal Communication Improves Perceptions of Friendship and Social Presence in Human-Robot Interaction," *International Conference on Intelligent Virtual Agents* (Springer, 2016).
8. A. Fardana, S. Jain, I. Jovančević, Y. Suri, C. Morand and N. Robertson, "Controlling a Mobile Robot with Natural Commands based on Voice and Gesture," *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (2013).
9. J.-B. Gómez, A. Ceballos, F. Prieto and T. Redarce, "Mouth Gesture and Voice Command Based Robot Command Interface," *International Conference on Robotics and Automation* (2009).
10. T. Mustofa, "Implementation speech recognition for robot control using MFCC and ANFIS," *J. Telematics Inf. (JTI)* **5**(2), 47–55 (2017).
11. K. Zinchenko, C.-Y. Wu and K.-T. Song, "A study on speech recognition control for a surgical robot," *IEEE Trans. Ind. Inf.* **13**(2), 607–615 (2016).
12. B. Burger, I. Ferrané, F. Lerasle and G. Infantes, "Two-handed gesture recognition and fusion with speech to command a robot," *Auto. Robots*. **32**(2), 129–147 (2012).
13. Z. H. Abdullahi, N. A. Muhammad, J. S. Kazaure and F. A. Amuda, "Mobile robot voice recognition in control movements," *Int. J. Comput. Sci. Electron. Eng. (IJCSEE)*. **3**(1), 11–16 (2015).
14. A. A. Abed and A. A. Jasim, "Design and implementation of wireless voice controlled mobile robot," *Al-Qadisiyah J. Eng. Sci.* **9**(2), 135–147 (2016).
15. H. G. Okuno, K. Nakadai, T. Takahashi, R. Takeda, K. Nakamura, T. Mizumoto, T. Yoshida, A. Lim, T. Otsuka, K. Nagira and T. Itohara, HARK Document Version 2.0. 0.(Revision: 6357).
16. S. Tabibian, "A voice command detection system for aerospace applications," *Int. J. Speech Technol.* **20**(4), 1049–1061 (2017).
17. S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Cepstral Gain Normalization for Noise Robust Speech Recognition," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (2004).
18. S. J. Young and S. Young, *The HTK Hidden Markov Model Toolkit: Design and Philosophy* (University of Cambridge, Department of Engineering Cambridge, UK, 1993).
19. M. T. Elahi, A. H. Korayem, A. Shariati, A. Meghdari, M. Alemi, E. Ahmadi, A. Taheri and R. Heidari, "'Xylotism': A Tablet-Based Application to Teach Music to Children with Autism," *International Conference on Social Robotics* (Springer, 2017).
20. A. G. Pour, A. Taheri, M. Alemi and A. Meghdari, "Human–robot facial expression reciprocal interaction platform: Case studies on children with autism," *Int. J. Soc. Rob*. **10**(2), 179–198 (2018).
21. A. Taheri, A. Meghdari, M. Alemi and H. Pouretemad, "Human–robot interaction in autism treatment: Acase study on three pairs of autistic children as twins, siblings, and classmates," *Int. J. Soc. Rob.* **10**(1), 93–113 (2018).
22. C. Bennett and S. Šabanovic, "Perceptions of Affective Expression in a Minimalist Robotic Face," *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2013).
23. C. C. Bennett and S. Šabanoviæ, "Deriving minimal features for human-like facial expressions in robotic faces," *Int. J. Soc. Rob*. **6**(3), 367–381 (2014).
24. M. Blow, K. Dautenhahn, A. Appleby, C. L. Nehaniv and D. Lee, "The Art of Designing Robot Faces: Dimensions for Human-Robot Interaction," *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (2006).
25. C. F. DiSalvo, F. Gemperle, J. Forlizzi and S. Kiesler, "All Robots are not Created Equal: The Design and Perception of Humanoid Robot Heads," *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (2002).
26. A. Korayem, S. Nekoo and M. Korayem, "Sliding mode control design based on the state-dependent Riccati equation: Theoretical and experimental implementation," *Int. J. Control* **92**(9), 2136–2149 (2019).
27. A. H. Korayem, S. R. Nekoo and M. H. Korayem, "Optimal sliding mode control design based on the state-dependent Riccati equation for cooperative manipulators to increase dynamic load carrying capacity," *Robotica*, **37**(2), 321–337 (2019).
28. A. Davouddoust, Upgrading the Scout Mobile Robot, by Creating Social Interactions with Humans, Using Audio and Video Channels in a Dynamic Environment *Master's Thesis* (Islamic Azad University of Tehran Research Sciences, 2018).
29. S. Tanha, S. Dehkordi and A. Korayem, "Control a Mobile Robot in Social Environments by Considering Human as a Moving Obstacle," *2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM)* (2018).

https://doi.org/10.1017/S0263574720001496 Published online by Cambridge University Press