Integrating Robots with Speech Recognition to Perform Tasks

https://github.com/dennyboby/social_robot_navigation

Denny Boby
Robotics Engineering
Worcester Polytechnic Institute
Worcester, USA
dboby@wpi.edu

Nihal Suneel Navale
Robotics Engineering
Worcester Polytechnic Institute
Worcester, USA
nsuneelnavale@wpi.edu

***Abstract*—The speech recognition system is a comfortable and quicker way to control the machines in today's life and is a hot topic in the development of social assistive robots. In this paper, the speech recognition system is integrated with the robotic operating system (ROS) and implemented onto a mobile robot called "FREIGHT". This paper also describes the implementation of speech recognition system with robot's movement according to human's spoken commands. The system will first convert the audio streams into text through Automatic Speech Recognizer (ASR) using Picovoice API, Natural Language Processing (NLP) is also done on Picovoice framework to retrieve the navigation task, Human-Robot Interaction associated model and use this model information to plan the route. This method is used to provide an efficient approach to translate natural language tasks to machine understandable format.**

***Keywords*—Speech Recognition, Social Assistive Robots, Navigation, Automatic Speech Recognizer, Natural Language Processing.**

I. INTRODUCTION

In recent years dialog based systems have become very popular among Human Machine Interfaces. This voice/speech-based interaction is highly valuable when the user is performing critical tasks. One such scenario is a nurse/doctor who is attending a covid patient and is using our robot to transport medicines from the pharmacy to the ward. In this scenario every second of a patient matters. There are moments where the patients are un-supervised and these moments may end-up in life threatening scenarios. Designing a social robot which can recognize and understand human speech and perform predefined tasks will assist the nurses/doctors in their duties.

The applications of Speech Integrated robots are immense. Such robots will be more trusted among humans and will be able to bridge the gap between human and robot communication. This technology will help integrate robots into human lives better, the robots can then be placed to assist professionals in the Hospitality Industry, Education Industry, Medical Industry and can also become a part of the family and help with early-child education, therapy, assistance of the elderly and also provide home security.

II. RELATED WORK

In the work by Gordon Briggs, Tom Williams and Matthias Scheutz (2017)[1], experimental evidence is provided that humans tend to phrase their directives to robots indirectly, especially in a social context. They introduced a dialog based mechanism to infer intended meanings from such indirect speech and demonstrated that these mechanisms can handle all indirect speech acts in their experiment as well as in other common forms of request. In another paper by Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili and Daniele Nardi (2017)[2], they present an approach to the design of natural language interfaces for human robot interaction, to translate spoken commands into computational structures that enable the robot to execute the intended request. The proposed solution is achieved by combining a general theory of language semantics with the state-of-the-art methods for robust spoken language understanding. Their results show that their processing chain, trained with generic resources, provides a solid baseline for command understanding in a service robot domain. A paper by Tada, Y., Hagiwara, Y., Tanaka, H., & Taniguchi, T. (2020)[3], describes a new method that enables a service robot to understand spoken commands in a robust manner using off-the-shelf automatic speech recognition (ASR) systems and an encoder-decoder neural network (sequence to sequence) with noise injection. The noise is injected into phoneme sequences during the train phase of encoder-decoder neural network based semantic parsing systems They demonstrate that the use of neural networks with a noise injection can mitigate the negative effects of speech recognition errors in understanding robot-directed speech commands. In this paper Withanage, P., Liyanage, T., Deeyakaduwe, N., Dias, E., & Thelijjagoda, S. (2018)[4], gives insights about using ASR and natural language processing for road navigation. They propose a user centric roadmap navigation application called "Direct Me". The approach of generating the user preferred route, the system will first convert the audio streams into texts through ASR using Pocket Sphinx Library, followed by NLP utilizing Stanford CoreNLP framework to retrieve the navigation-associated information and process the route in the Map using Google Map API. In the paper by Sharan, S., Nguyen, T. Q., Nauth, P., & Araujo, R. (2019)[5], a voice control software system is integrated with the SLAM algorithm available in robotic operating system (ROS) and implemented to a mobile system of a robot "ROSWITHA". This paper also expresses the different ways of controlling the robot by developing a graphical user interface (GUI) with different controlling tabs. The experiment and test is performed real time in environments with different speakers to analyze the accuracy of the system.

III. PROPOSED METHOD

We can divide the method of approach into 4 points:

- **Speech Recognition**- In this task we are required to detect audio/voice from the user and convert the audio/voice signals into text using Automatic Speech Recognizer (ASR). The ASR must be fast and reliable. We have tested 3 ASRs namely Google ASR and IBM Watson and Picovoice. We are using Picovoice because of its high accuracy of 97.6% in a noisy environment. IBM Watson-86.7% and Google-76.4% as shown in Fig1. and Fig 2. shows the Accuracy vs. Noise graph of Picovoice.

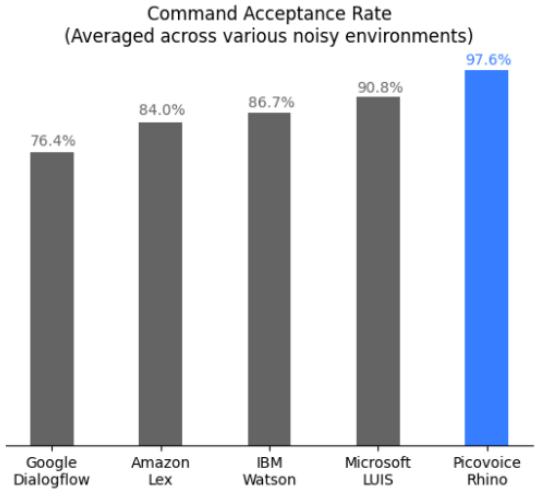


Fig. 1. Accuracy vs. Noise (dB)

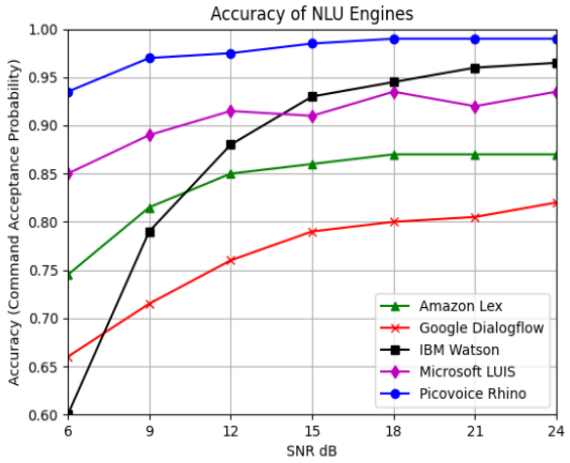


Fig. 2. Accuracy vs. Noise (dB)

- **Semantic Analysis**- The ASR is then inputted into a Semantic Analyzer to pick out the key/trigger words and map these words to the command library which is integrated with the robot through ROS.
- **Speech Synthesis**- In this task we develop a conversational based model which gives the robot the functionality to have a conversation, this task uses Speech Recognition and Semantic Analysis to understand human speech. The Speech Synthesis model is developed on Picovoice Rhino Console.
- **Navigation**- In this task the recognised navigation command from human speech is used to give a point A to B task or point A to B then C task to the robot. This task uses move_base library to execute such commands.

IV. EXPERIMENT

A. *Task*

- Understand human intent when the speech is command-like and conversational depending upon the perceived intent the robot must perform accordingly.
- When a speech is provided containing ambiguity of intent the robot must communicate and ask for clear commands or ask further questions to clear ambiguity.

B. *User Study*

To understand how complex human speech can be, we conducted a user study/survey where we ask the participants to write down different ways someone can ask a robot to perform a task.
The tasks included navigation between three rooms. We found that humans are able to convey the same meaning using a myriad of sentence formations either by using different words or even by omission of a few words but the overall meaning of the spoken dialogue remains intact. The survey can be accessed on Github.

C. *Evaluation Metrics*

For the Evaluation, we are checking for the following:

- Accuracy of the robot to understand different accents of humans regardless of age and gender.
- Statistics of the number of times a person has to repeat speech to execute the right command.
- Accuracy of intent
- User rating on collaboration with the robot and the effort in communication.
- Performance analysis of Speech Recognition, Semantic Analysis & Speech synthesis.

V. RESULTS

A. *Simulation*

For this project we are using Unity simulation with Nursing Robot setup. Inorder to get familiarized with the Unity installation and simulator, the Pick and Place tutorial from Unity-Robotics-Hub was executed as shown in Fig 3 and Fig 4.

Fig. 3. Niryo One picking the object.

Fig. 4. Niryo One placing the object onto the marked location.

Next the Nursing Robot simulation for the project was set up with a Freight robot in it. The nursing robot simulation is shown in Fig 5 and the Freight robot is shown Fig 6.

Fig. 5. Nursing robot simulation in Unity.

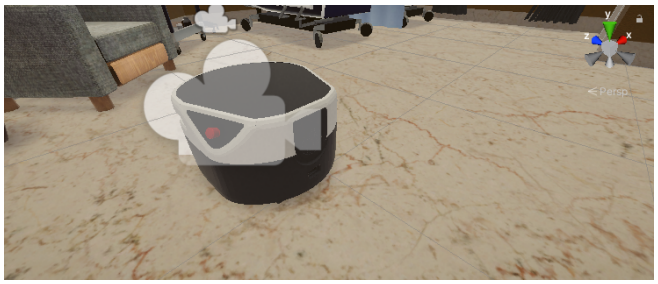


Fig. 6. Freight robot (Differential drive robot) in Unity.

Plugins like Joint State Publisher (publishes the current state of both the wheel joints), Pose Stamped Publisher (publishes the current pose of robot w.r.t unity world frame) and Laser Scan Publisher (publishes the laser scan data from unity to a ROS topic) were added to the robot in Unity so that in future we could implement a better navigation stack with obstacle avoidance. Apart from simulating the robot in unity the freight robot was also visualized in Rviz (visualization tool from ROS) showing RobotModel, TF and LaserScan as shown in Fig 7.

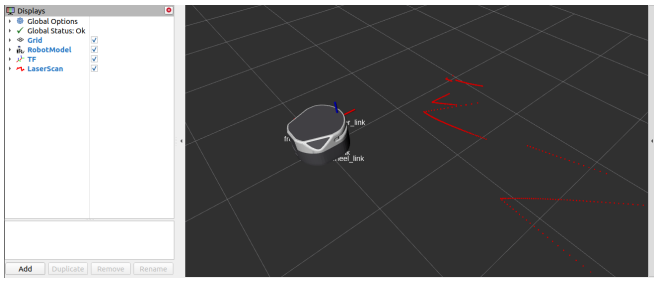


Fig. 7. Freight robot visualized in Rviz with RobotModel, TF and LaserScan

B. *Speech*

For the development of Speech recognition we have tested Google ASR, IBM Watson and Picovoice to detect speech for Semantic Analysis. Among the Google, IBM and Picovoice ASRs, the Picovoice ASR is the fastest and most reliable with a command acceptance rate of 97.6% in a noisy environment of 12dB. The Picovoice ASR also performs significantly better at noise of 24dB as shown in Fig. 1. The time taken by the ASR depends on the size of the conversation/speech.

For the development of Intent detection Speech Analysis model we conducted a survey to understand how humans speak commands and form various sentences which convey the same meaning, we are using the Picovoice Rhino Console, we are using this console because of its integration with the Picovoice ASR, its user-friendly interface for developing the model and its accuracy in output.

For example, the speech is spoken as “Please go to room 1”. For this spoken speech we get the Intent model output as shown in Fig. 8.

```
{
  intent: "move",
  slots: {
    dst1: "1"
  }
}
```

Fig. 8. Example of NLP model output

The intent is detected as “move” and the destination dst1 is “1”. Now, the robot moves to room 1.

We have also trained the model to understand commands like “Please go to the cardiology lab and orthopedics lab”. For this spoken command we get the Intent model output as shown in Fig. 9.

```
{
  intent: "move",
  slots: {
    dst1:
"cardiology lab"
    dst2:
"orthopedics lab"
  }
}
```

Fig. 9. Example of NLP model output

Here, the robot will first move to the room mapped to “cardiology lab” and then move to the room mapped to “orthopedics lab”. Our model also has the ability to understand other intents such as “exit”, “convo” and “help”. We are currently developing a voice talkback script for better Human Robot Interaction which will use the “convo” and “help” intents to the fullest.

A intent_recognition.py python script runs this model and when a wake word “Jarvis” is said the script starts recording the navigation command untils a pause or silence is there. Then this audio is processed using the trained model and the intent goal is identified. The intent is further processed to generate a navigation goal message which is published via the topic “/navigation_goal”. We have also integrated a talk back feature which allows the robot to clear ambiguity in the speech recognition by asking the user to repeat

themselves, this makes the Human-Robot Interaction more intuitive and human-like. Our NLP model also has the capability to understand other speech commands like, Can you please take me to where Dr. Mellissa is currently stationed, here the robot will guide/move to the room where the doctor is specified for. We can also ask the robot to say a joke, the date, time as well as ask where the pre-defined specific staff (doctor/nurse) are stationed. These features make the Human-Robot Interaction more friendly and easier to integrate with the daily lives of humans.

C. *Navigation*

In Order to control the Freight robot in Unity via ROS, the first step is to control the Freight robot via the Keyboard Wheel Controller plugin present in the Unity simulation. Once that is done the next step is to connect Unity with ROS and control the robot via /cmd_vel topic. In Order to achieve this Twist Subscriber plugin was used which listens to the /cmd_vel topic and provides the linear and angular velocity in the message to the wheel controller in the simulation. Using the terminal, velocity commands were published to the /cmd_vel topic. The output of the work is shown in Fig 10.

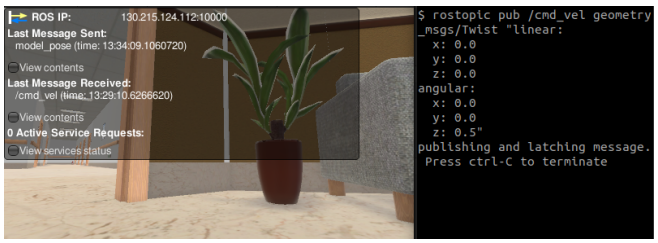


Fig. 10. Freight robot controlled via messages published to /cmd_vel topic using terminal.

Once that was done a navigation.py script was made which listens to the topic "/navigation_goal". When a navigation goal is received the script is responsible for finding the path to the goal location from a set of predefined paths. Once it finds the path the script makes the robot move by publishing appropriate velocities to the "/cmd_vel" topic. The path is followed using two components. One component makes the robot rotate to the desired orientation so that the robot can just follow a straight line path. The other component makes the robot follow the straight line path using a PID controller which computes correction angular velocity based on the feedback from "/model_pose" topic that publishes the current pose of the robot. Fig 11 shows how different nodes are communicating with each other via topics.

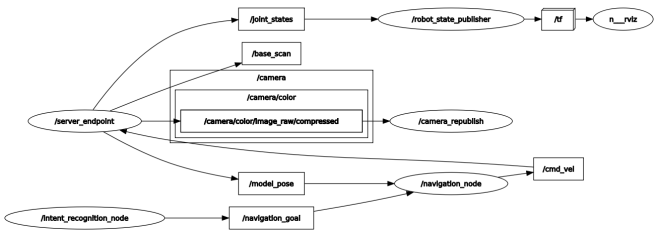


Fig. 11. Different nodes communicate with each other through topics.

Fig. 12 shows the output of the speech "Jarvis, go to room 2". The bottom terminal for Fig. 12 runs the intent recognition script and the script prints out the recognised intent on the terminal. Once the navigation script receives the command it prints out "Navigation request received" and moves the robot to room 2. Once the robot reaches room 2 the script prints out "Reached room 2".

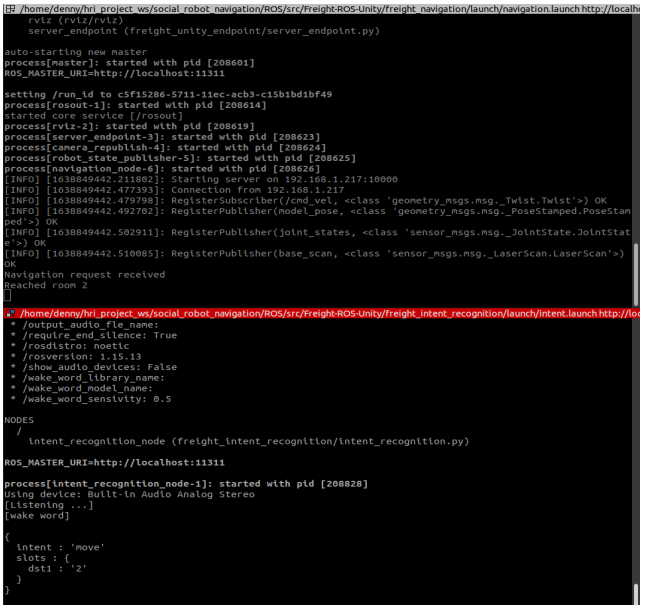
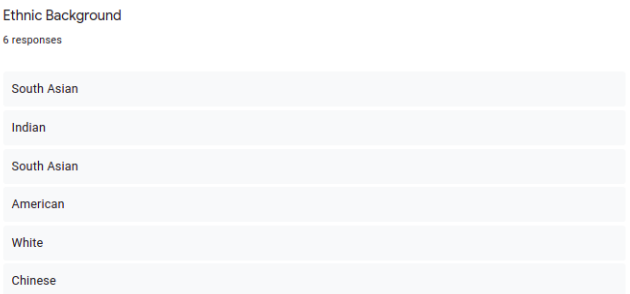


Fig. 12. Terminal output of navigation script (top) and intent_recognition script (bottom).

D. *User Study*

We conducted a user survey (the survey asks short answers and rating from 1-10) to collect data on how good our project's speech recognition is for people of different accents/ethnicity as shown in Fig. 13, the number of times the user had to repeat themselves, if the users were able to understand the talk-back function and more, the user survey can be accessed through the link https://drive.google.com/drive/folders/14saLGhgnFf94iduA1Lh-W2hBCulZIk1z?usp=sharing. For our survey we asked 5 males and 1 female of different ethnicity/accent to have a conversation with our robot, from the survey we are able to



give certain conclusions,

Fig. 13. Ethnic/Accent Diversity in the conducted user study.

1. **How good was the voice recognition ?** : Overall the users were satisfied with the voice recognition capability and scores were all above 7.

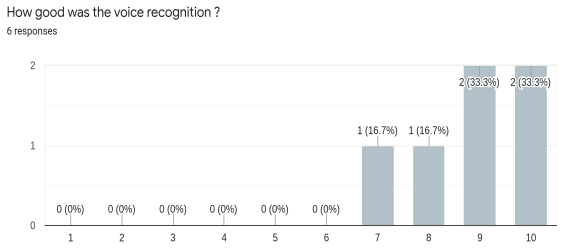


Fig. 14. Survey response to the associated questionnaire.

2. **How well do you think the robot understood your command or intent ?** : Command Intention detection is high, with scores above 8.

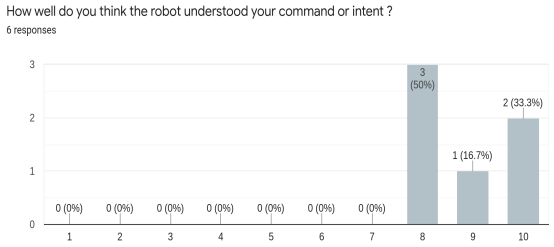


Fig. 15. Survey response to the associated questionnaire.

3. **How well were you able to understand the robot?** : Scores of 10 justify the talk-back capability is good and that users had no problem understanding the robot.

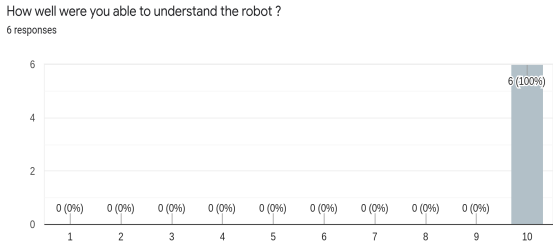


Fig. 16. Survey response to the associated questionnaire.

4. **How many times did you repeat yourself ?** : The results for these are very spread out, with User repeating as high as 6 times and low as 1. From this we can see that our project has a few shortcomings which can be resolved by developing a better speech model.

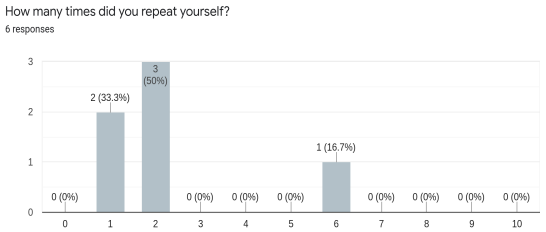


Fig. 17. Survey response to the associated questionnaire.

VI. CONCLUSION

In this paper, propose an approach to the design and implementation of a natural language model with the aim of enabling a robot with the ability of understanding natural speech commands for a set of navigation tasks.

The proposed method is characterized by three main components. First, we take the user's spoken speech and convert it into text based format, this is achieved by ASRs. Second, we develop a NLP model using Picovoice Rhino console, the NLP model is specifically designed to recognize different possibilities of predefined commands. Since we are using Picovoice as our ASR and NLP model trainer, we can skip converting the speech into a text file, the voice and NLP modules are integrated within the Picovoice Console. The trained NLP model gives us the intent of the spoken command and the task to be performed. Third, send the detected intent to the navigation script which achieves the desired goal by moving the robot in simulation. We have successfully achieved performing the task using the user's spoken command.

We have also faced a few setbacks and we recognize these setbacks as potential limitations, these limitations are regarding the trained NLP model, currently the NLP model is trained for specific navigation tasks, so any command given unrelated to or out of scope of the model's capability would not work. Since we are working with voice or speech, the NLP model is highly dependent on the voice, noise in the surrounding and the spoken language, our project is currently limited to the English language so our model will not work with other languages.

VII. SCHEDULE

TABLE 1; shows the schedule planned for the entire HRI project for the semester(August - December 2021):

TABLE I. SCHEDULE

WEEKS	DENNY BOBY	NIHAL NAVALE
Week 1 - 2	• Literature Review • Project Proposal PPT • Setup Simulation	• Literature Review • Project Proposal PPT • Familiarize with ROS • Setup Simulation
Week 3 - 4	• Navigation code for point A to B in simulation.	• Automatic speech recognition code.
Week 5 - 7	• Implement static parser for the detected speech • Implement semantic analysis module for post processing	
Week 8	• Combine speech command with navigation	
Week 9	• Testing the entire system	
Week 10	• Metrics collection	
Week 11	• Buffer	
Week 12	• Achieve extended goal	
Week 13	• Report	• PPT
Week 14	• Presentation	

REFERENCES

[1] Briggs, Gordon, T. Williams and Matthias Scheutz. “Enabling robots to understand indirect speech acts in task-based interactions.” Journal of Human-Robot Interaction 6 (2017): 64 - 94.

[2] Bastianelli, Emanuele, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. “Structured Learning for Spoken Language Understanding in Human-Robot Interaction.” The International Journal of Robotics Research 36, no. 5–7 (June 2017): 660–83.

[3] Tada, Yuuki, Yoshinobu Hagiwara, Hiroki Tanaka and Tadahiro Taniguchi. “Robust Understanding of Robot-Directed Speech Commands Using Sequence to Sequence With Noise Injection.” Frontiers in Robotics and AI 6 (2019): n. pag.

[4] P. Withanage, T. Liyanage, N. Deeyakaduwe, E. Dias and S. Thelijjagoda, "Road Navigation System Using Automatic Speech Recognition (ASR) And Natural Language Processing (NLP)," 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2018, pp. 1-6, doi: 10.1109/R10-HTC.2018.8629859.

[5] S. Sharan, T. Q. Nguyen, P. Nauth and R. Araujo, "Implementation and Testing of Voice Control in a Mobile Robot for Navigation," 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 2019, pp. 145-150, doi: 10.1109/AIM.2019.8868892

[6] https://github.com/Picovoice/speech-to-intent-benchmark

[7] https://picovoice.ai/docs/tips/syntax-cheat-sheet/