Assisted Robot Navigation based on Speech Recognition and Synthesis

Silas F. R. Alves, Ivan N. Silva
São Carlos Engineering School
University of São Paulo
São Carlos, Brazil
{salves, insilva}@sc.usp.br

Caetano M. Ranieri
Institute of Mathematical and Computer Sciences
University of São Paulo
São Carlos, Brazil
cmranieri@yahoo.com.br

Humberto Ferasoli Filho
Faculty of Sciences
University of São Paulo State
Bauru, Brazil
ferasoli@fc.unesp.br

***Abstract*—Interactive robots can help people with or without disabilities. In this sense, research has been made in order to help children with motor disabilities to explore the world around them, which is important for their cognitive development. However, most of these initiatives lack on natural and intuitive interfaces, or are prohibitively expensive to be adopted in a larger scale. This paper describes an experimental environment to use speech recognition and synthesis to improve human-robot interaction (HRI) with children. The proposed system main goal is to perform activities with physically disabled children, however it can be used with other children. Thus, robots that are attractive, small-sized and relatively low-cost are used to implement such environment. The system recognizes a set of simple speech commands, which allows human-assisted navigation.**

Keywords — human-robot interaction, speech interaction, assistive technologies, social robots, mobile robotics.

I. INTRODUCTION

The development of socially interactive robots have different motivations. These robots can aid patients with certain kinds of impairments, providing means of therapy for autism [1] or severe motor disabilities [2]. On the other hand, they can also help people without disabilities to perform tasks. In the latter, the literature points to different applications according to the age of the audience. For children, these robots may serve as tutors or interactive toys. For adults, as personal assistants. For elderly people, as company or tools to help dealing with daily activities [3].

One important issue for research on the field of social robotics is the development of Human-Robot Interaction (HRI) based on natural interfaces, such as speech interaction [4] or social gaze [5]. To improve the interface, one approach is to provide the systems with emotions, allowing humans to identify intentionality on the robot [6]. Besides, robots with emotions may serve as a platform for experiments in studies about living beings' emotions and neurobiology [7].

Some projects consist on the implementation and further validation of socially interactive robots, dealing with different aspects or applications [3] [4] [6] [8]. The system presented in this paper aims to provide an interface to enable a human user to partially control a mobile robot endowed with social aspects. The presented robot may serve as a tool for educative or ludic activities for children with and without disabilities. Future projects may apply this system in a larger robot to help elderly people.

II. MOTIVATION

The inclusion of children with disabilities on daily activities presents several challenges. They need specialized tools and trained personnel, which is associated high costs. Still, this is a steady growing issue in Brazil, where the number of special students in public education grew from 752,305 in 2011 to 840,433 in 2012 [Brasil, INEP 2013]. In the case of children with severe motor impairments, the usage of robotics have enabled them to interact with the world, which important for their cognitive development [9] [10].

These children are usually unable to use traditional user interfaces (UI), such as the computer keyboard and mouse, or a joystick. They need different UIs, which do not rely on precise manipulation and are intuitive. A common communication channel for humans, but still is a challenge for robots, is speech. Even though speech recognition and natural language processing present several open problems, it may provide an easy way of interaction under some limitations.

There is also research suggesting that children can regard robots as living beings [11] – even though such kind of robot is still out of reach with current technology – which may help to create interesting interaction scenarios between children and robots to foster education. However, the high cost of robots sometimes forbids their use in the classroom.

In this sense, the use of smartphones may be a promising approach. Originally created to serve as advanced personal assistants, smartphones have increased its relevancy on the market, expanded its functionalities, improved its connectivity and energetic efficiency, and had its cost reduced. In general, smartphones converged to a minimum set of devices, which are available within its hardware architectures, such as cameras, accelerometers, touchscreen, Wi-Fi and Bluetooth communication, and audio capture and reproduction. Each of these features provides a useful capability for interactive robots. This set of devices can improve the robot hardware by providing new sensors and actuators, besides providing a versatile processing unit. Thus, we propose its application in an entertainment mobile robot.

III. SYSTEM OVERVIEW

An Android operated smartphone and a mobile robot are the main devices of proposed environment. As described in section IV, the robot has a mechanical support for a smartphone, and a Bluetooth link connects both devices. The robot firmware does not have control software, which is solely performed by the smartphone. Therefore, the whole controller works on a smartphone application, responsible for both the robot control functions and the human-robot interface. Google recognizer is responsible by speech recognition in the described system, and eSpeakTTS is responsible for speech synthesis in Brazilian Portuguese.

Once a user starts the application, the system operates as represented in Fig. 1. The robot wanders around, avoiding obstacles and periodically changing the smartphone's tilt angle, until the system detects a face on the image captured by the smartphone's camera. When a face is recognized, the robot stops all kinds of movements, and the smartphone synthetizes a speech statement in Portuguese, through its speakers. In such statement, the system asks for a speech command. To do so, the human user might speak one between a few available commands, all of them in Portuguese.

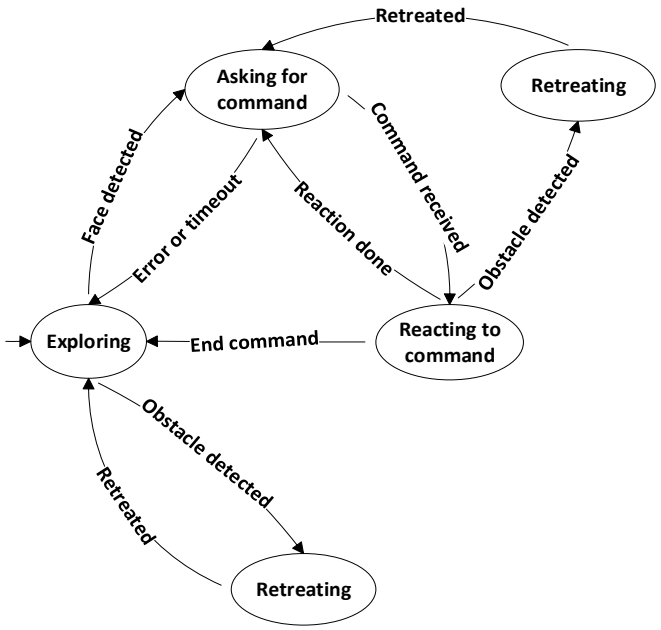


Fig. 1. State machine relative to the implemented system

The smartphone utilizes its own microphone or an external headset to capture any sound long and loud enough to be a recognizable command. If the speech recognizer detects a sentence, it will return a set of up to 10 possibilities and their confidence. If any of these sentences match one of the available commands, such command is executed. Otherwise, the robot asks the user to repeat the command. While executing a command, if the robot encounters an obstacle, it retreats, just as it does when it detects an obstacle while it is wandering around.

The available commands refer to simple discrete actions, which can be used to provide an assisted navigation for the robot. For better comprehension, Table I lists the possible commands in English. The actions related to each command are also described.

TABLE I. AVAILABLE COMMANDS FOR SPEECH INTERACTION.

Command	Action
You can go	Finalizes speech interaction.
Circle	Move in a circumference-like trajectory.
Flip back	Turn 180º in clockwise fashion.
Forward	Move forward, at 6.0 cm/s, for 1.5 s.
Clockwise	Turn 30º clockwise.
Counterclockwise	Turn 30º counterclockwise.

A. Control Architecture

The subsumption architecture, proposed by Rodney Brooks, is the basis for the control architecture implemented on the described system. This architecture consists on the organization of the system in hierarchical behavioral modules, named levels of competence, each comprising a control subsystem. The more elevated the level of competence, the more specific the subsystem defined by it. Each level of competence may suppress inputs or inhibit outputs of inferior levels' behaviors.

As presented in Fig. 2, four levels of competence compose the developed control architecture. From the lowest to the highest, they are: *avoid obstacles*, *wander around*, *seek faces* and *speech interaction*. In *avoid obstacles* level, the proximity sensors are continuously checked. When the system detects an obstacle, the robot retreats. In *wander around* level, the robot walks randomly around the environment. In *seek faces* level, the smartphone's inclination angle is changed periodically, aiming to increase the probability of detecting a human face. Finally, on *speech interaction* level, the robot stars, after detecting a face, a verbal interaction, through which the user can control the robot by giving speech commands. Once such interaction begins, *seek faces* and *wander around* outputs are inhibited.

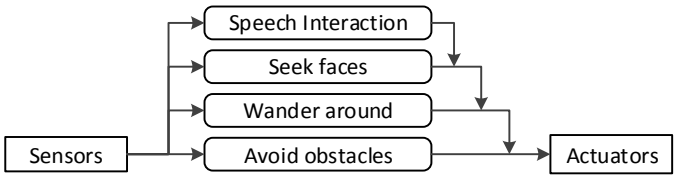


Fig. 2. Implemented control architecture.

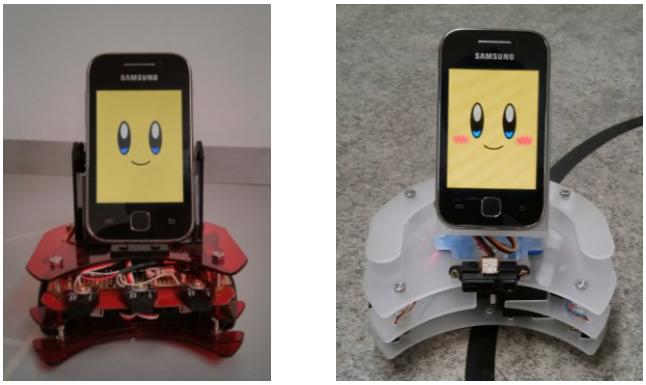
As sensors, the system uses infrared light refection proximity sensors to detect obstacles, the smartphone's camera to capture video and the smartphone's microphone or an external headset to capture audio. As actuators, the system uses the robot motors, whose activation result in a robot locomotion or on a smartphone's tilt change.

IV. ROBOTS

Some of the difficulties faced by assistive robotics are the cost and the experience necessary to work with robots [12]. Robots adapted to HRI are expensive and require specialized personnel, and many researchers, especially in the areas of cognitive science and human-computer interfaces, do not have access to them [13].

In order to provide a robotic platform with low cost, we developed two 100 USD robots that are easy to replicate and provides high flexibility, Roburguer, Fig. 3 (a), and its latest version, Pomodoro, Fig. 3 (b). These robots were created to serve as entertainment robots for children, thus they use vivid colors to draw children's attention, and the body is designed with rounded shapes to not harm the user. The body structure uses flat acrylic parts that can be easily replicated.

Work is in progress to provide these robots with emotional skills, which may explore the exhibition of faces describing an emotional state, as shown in Fig. 3. Currently, these faces are not yet used, and are shown in the picture to illustrate the possibility of implementing such feature.



(a) (b)

Fig. 3. Roburguer (a) and Pomodoro (b) robots.

The robots also have sensors and actuators that allow safe navigation, avoiding falling and collisions, as presented on Table II. Additionally, a smartphone was adopted as the embedded computer.

TABLE II. DEVICES AVAILABLE IN POMODORO.

Device	Qty.	Function
Floor sensor	5	Avoid falling
Floor line sensor	2	Line following
Proximity sensor	3	Avoid collision
Distance sensor	1	Object detection (Pomodoro only)
DC motor	2	Differential drive
Servo-motors	2	Move (pan-tilt) smartphone

The hardware consists of two separate modules: the embedded board, with the circuit to activate the motors and read sensors presented by Table II, and a smartphone, which will implement the control architecture. The embedded board implements a supervisor system that is not autonomous, but receives commands from a computer through a Bluetooth connection. In this case, the computer is a smartphone that will be attached to the robot.

The smartphone was adopted due its relative low-cost, connectivity, sensors, and processing power. In this project, smartphone was used to capture the user's speech, search for the user's face within the camera image, and reproduce a synthetized speech.

V. EXPERIMENTS AND RESULTS

To validate the developed system, some experiments were made at the Integration of Systems and Intelligent Devices Laboratory (LISDI) of São Paulo State University (UNESP), campus Bauru. Although the system was experimented only by adults without signs of impairments, it was possible to evaluate the technical aspects of the system and its appliance as a human-robot interaction tool.

In the experiments, the robot was placed on a flat surface, measuring 80 cm x 60 cm. A wall with 5 cm height surrounds this platform, shown in Fig. 4. The smartphone, a Samsung Galaxy Y with a back camera and no frontal camera, was placed on the robot with its back camera pointing forward. Thus, the display could not be seen by the user.

These experiments consisted in individual tests of each of the four levels of competence, verifying their adequate functionality. To register the results, four videos were produced, each related to a level of competence. These videos are available at http://www2.fc.unesp.br/gisdi/speech/. The procedures to test each level of competence and the respective results are described below.

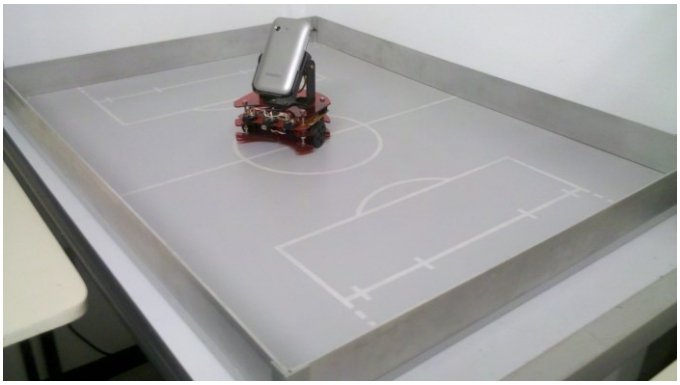


Fig. 4. Platform used for experiments, with Roburguer robot.

- **Avoid obstacles:** the robot was left on the platform, standing still until an obstacle was positioned in front of its proximity sensors. This way, it was possible to conduct the robot do different places on the platform, by successively positioning the obstacle next to convenient sensors.
- **Wander around:** an obstacle was left on the platform. The robot, moved in different directions, following the *wander around* level principles. It is shown that the robot can avoid the platform's borders. When the obstacle set on the platform is detectable by the robot's proximity sensors, the system reacts fast enough to avoid collisions. Though obstacles are not always detectable before the collisions happen, the robot's integrity is assured, because the robot moves slow enough to avoid damages during collisions.
- **Seek faces:** at this level, the robot's behavior is similar to the behavior seen at *wander around* level. *Seek faces* adds periodical changes at the smartphone's inclination, which does not affect the behaviors defined by lower levels.

- **Speech interaction:** comprises the system as a whole. The test was done while there was a stable Wi-Fi connection to the Internet, requisite for this level to work properly. While there was no face to be detected, the system worked similarly to *seek faces* level. Using the proximity sensors, it was possible to conduct the robot to a position in which the user's face could be detected. When a face was detected, all the robot's movements stopped, and speech interaction began. All the implemented commands were experimented, having worked as expected. During the execution of "forward" command, the user put his hand in front of a proximity sensor, which detected it as an obstacle. The robot retreated and, through synthetized speech, asked for another command.

VI. CONCLUSION

This paper presented the use of control architectures receptive to human commands and supported by the use of smartphones. Search and detection of faces was effective and interesting within the premise of the robot behaving like a pet. Speech interaction, during the current stage of development, merely interpret and execute commands. In this sense, the interactive system showed results that encourage the exploration of this area, adapting and expanding it in order to meet actual demands. It is expected that the behavioral actions of robots arrest the attention of children, giving the idea of interacting with a small pet.

The latency between recognizing a command and executing a correspondent action does not compromise the interaction. A richer verbal interaction between the user and the robot, which would approach the idea of an artificial living being, will be addressed in further works with use of natural language processing – another challenging research field. Also in this context, a more incisive use of computer vision can also increase the possibilities of interaction with human and robot with the external environment, which fosters interest in increasing the use of this technology. For example, emotion analysis based on facial expressions may enhance human-robot interaction, as well as object recognition, scene understanding and situation awareness may improve the collaboration between robots and humans or other robots.

In this work, we used a smartphone Samsung Galaxy Y, which has no frontal camera. In future work, using a smartphone with front camera, it would be possible to introduce facial expressions by exhibiting them on the smartphone display. It would encourage experiments involving emotions.

The need for sophisticated algorithms, requiring computational power to the robotic system in timely responses, may be satiated by modularity provided by the architecture, which allows the use of Wi- Fi and a remote computational basis. With the fast development of mobile devices' hardware, however, it is expected that the processing power of smartphones will grow, fitting most of the requirements for such projects.

ACKNOWLEDGEMENT

We thank the Intelligent Automation Laboratory (LAI), of the Department of Electrical Engineering, of the Center for Technology, Federal University of Espirito Santo, for their support.

REFERENCES

[1] I. Werry, K. Dautenhahn, B. Ogden and W. Harwin, "Can social interaction skills be taught by a social agent? The role of a robotic mediator in autism therapy," in *Cognitive Technology: Instruments of Mind*, Springer, 2001, pp. 57-74.

[2] C. M. Ranieri, S. F. d. R. Alves, H. F. Filho, M. A. C. Caldeira and R. Pegoraro, "An Envoronment Endowed with a Behavior-Based Control Architecture to Allow Physically Disabled Children to Control Mobile Robots," in *Robocontrol - 5th Workshop in Applied Robotics and Automation*, Bauru, 2012.

[3] M. Malfaz and M. A. Salichs, "A new architecture for autonomous robots based on emotions," in *5th IFAC Symposium on Intelligent Autonomous Vehicles*, Lisboa, Portugal, 2004.

[4] K. Aoyama and H. Shimomura, "Real world speech interaction with a humanoid robot on a layered robot behavior control architecture," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, 2005.

[5] N. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews,* vol. 24, pp. 581-604, 2000.

[6] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Kyongju, 1999.

[7] M. A. Arbib and J.-M. Fellous, "Emotions: from brain to robot," *TRENDS in Cognitive Sciences,* vol. 8, no. 12, pp. 554-561, Dezembro 2004.

[8] G. A. Hollinger, Y. Georgiev, A. Manfredi, B. A. Maxwell, Z. A. Pezzementi and B. Mitchell, "Design of a social mobile robot using emotion-based decision mechanisms," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Pequim, China, 2006.

[9] J. K. Tsotsos, G. Verghese, S. Dickinson, M. Jenkin, A. Jepson, E. Milios, F. Nuflo, S. Stevenson, M. Black, D. Metaxas, S. Culhane, Y. Ye and R. Mann, "PLAYBOT A visually-guided robot for physically disabled children," *Image and Vision Computing,* pp. 275-292, 4 March 1998.

[10] L. Alvarez, A. M. Rios, K. Adams, P. Encarnação and A. M. Cook, "From Infancy to Early Childhood: The Role of Augmentative Manipulation Robotic Tools in Cognitive and Social Development for Children with Motor Disabilities," *Converging Clinical and Engineering Research on Neurorehabilitation,* pp. 905-909, 2013.

[11] S. Turkle, "Authenticity in the age of digital companions," *Interaction Studies,* pp. 501-517, 2007.

[12] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction,* pp. 203-275, 2007.

[13] J. L. Burke, R. R. Murphy, E. Rogers, V. J. Lumelsky and J. Scholtz, "Final Report for the DARPA/NSF Interdisciplinary Study on Human-Robot Interaction," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews),* pp. 103-112, 05 2004.