VOICE USER INTERFACE IN ROBOTICS - COMMON ISSUES AND PROBLEMS

Shafkat Kibria, Thomas Hellström

Lecturer, Dept. of Computer Science & Engineering, Sylhet International University, Bangladesh, Associate Professor, Dept. of Computer Science, Umeå University, Sweden
E-mail: shafkat80@yahoo.com, thomash@cs.umu.se

***Abstract* - The area of human machine interaction and machine intelligence in a social context is interesting and challenging to the Artificial Intelligence and Robotics community. The issue of developing humanoid robots with human-like perception is, for example, described in Brooks [13] research. Humans have five classical senses and sensors – vision, hearing, touch, smell, and taste, by which they perceive the surrounding world. This paper presents experiences resulting from experiments in introducing the "hearing" perception, using Speech Recognition technology in Robotics control [1].**

I. INTRODUCTION

Humans are used to interact through Natural Language (NL) in the social context. This idea underlies the Roboticists' inclination to create NL interfaces through Speech for the Human-Robot Interaction (HRI) part of the robots. Several Speech Recognition (SR) interfaces in robotic systems have been presented [1,2,6,12,8]. Most of the systems focus on Mobile Robots, which nowadays are becoming popular as service robots both indoors and outdoors [3]. The goal of a service robot is to help people in their daily tasks in a social context. With mobile robots used for service, it is important to communicate with the users. Speech Recognition is one of the easiest ways of communication with humans, making it possible for novice users to interact with robots without a need for special training.

Introducing Voice User Interface (VUI) for communicating with robots is one of the challenging areas in HRI. A project has been conducted to add SR capabilities to a Mobile Robot, and investigate the use of a Natural Language (NL) (English) as a VUI as HRI. SR technology is an apt aid for introducing a VUI to control a machine (computer, robot, etc.). This paper presents the experience of finding a suitable SR tool, SR system performance, and the limitations in the public place or social environments, as well as the challenge of introducing NL interfaces for novice users to control a robot.

The first objective of the project is to find a suitable SR tool to introduce VUI for Robotic control. To achieve this objective, both an SR Hardware Module (SRHM) and PC-based SR Software Program (SRSP) have been tried. Another important goal of the project is to introduce and evaluate VUI for novice users. This goal has been achieved through introducing simple and complex sentences/dialogs (in English) using an existing SR system to control the system (robot) within a specific domain, and with some specific robotic activities. Examples of some spoken sentences/dialogs that the robot is able to understand and perform:

Move
Move <10> centimeters
Turn left
Turn right
Follow the wall
Move <10> centimeters and then turn left.

Note: The tagged words are variables. For example, in "Move <10> cm" any number can replace the tagged number.

The simple and complex sentences used to control the robot have been tried with both the SR tools (SRHM & SRSP) to determine their performance and limitations. Tests and analyses of the test results have also been performed to find out the problems and limitations.

In the following sections we learn more about the issues and problems arising while introducing SR systems to Robotics for interaction purposes. We start with a literature review about SR and VUI systems (Section 2). Then we describe the issues arising in the project's implementation phase and relevant findings (Section 3). A discussion about conclusions, problems, limitations, and future work is presented in the concluding part (Section 4).

II. LITTERATURE REVIEW

SR systems represent a prominent technology, by which machines can recognize human speech. UNECE's press release about world Robotics survey stated that there are "*Over 600,000 household robots in use - several millions in the next few years"* [3]. This indicates that household robots are becoming popular and robots no longer are limited to industrial use, but can also serve in various ways in social contexts. The important thing for a service robot is to communicate with its users (humans) in an easy way, e.g. by NL. Using an SR system as a VUI to interact with humans enables the robot to behave more similarly to humans (at least from the perception-response point of view). *The term "robot" generally connotes some*

anthropomorphic (human-like) appearance [4]. The promise of robot workers, intelligent enough to replace humans in certain tedious or dangerous tasks, is starting to become a reality.

A. Speech Recognition (SR) system

SR is the process of converting an acoustic signal, captured by microphone or telephone, to a set of words [5]. This technology will help us to change the way of interacting with machines in the near future. Recognition of a series of sounds and identification of words from the sounds are the two important parts of Speech Recognition. Many parameters are involved in SR techniques; namely, Speaking Mode and Style, Speaker Enrolment, Size of Vocabulary, Language Model, Perplexity, Transducer, etc. [5]. Depending on the mode of speech, the SR system can be divided into word speech (one word at a time) and continuous speech (one or more sentences at a time). The speaker enrolment factor also divides the SR technique to two types: one is Speaker-dependent systems, in which the system has to enroll the speaker in a training session to be able to recognize the speech. The other one is Speaker-independent systems, which can identify any speaker's speech without training. Two other important factors in a Speech Recognition technique are Vocabulary Size and Language Model. Vocabulary size in an SR system should be carefully thought out, since large vocabularies or many similar sounding words make recognition difficult for the system. Another factor, the Language model (also called Artificial Grammars) is used to confine the possible word combinations to a series of words or sounds. Several techniques are used in SR systems: Hidden Markov Models (HMM), Artificial Neural Network (ANN), Fast Fourier Transform (FFT), and Learning Vector Quantization (LVQ) [7]. Among them, HMM has been the most popular and dominating technique for the last two decades. Thanks to these techniques, both Speech Recognition Hardware Module (SRHM) and Speech Recognition Software Program (SRSP) are now available in the market. See Table 1 for a list of available SRSPs for developers and their vendors/open sources. Table 2 shows some of the available SRHMs in the market.

SR software programs (SRSP) for developer	Vendors/Open Sources
IBM Via Voice	IBM
Dragon Naturally Speaking 8 SDK	Nuance
Voxit (Swedish)	www.voxit.se
VOICEBOX: Speech Processing Toolbox for MATLAB	www.ee.ic.ac.uk/hp/staff/dmp/voicebox/voicebox.html
Java Speech API	Sun Microsystems, Inc
The CMU Sphinx Group Open Source SR Engines	http://cmusphinx.Sourceforge.net
SpeechStudio Suite	SpeechStudio Inc.

TABLE 1: LIST OF SOME AVAILABLE SRSPS FOR DEVELOPERS AND THEIR VENDORS/OPEN SOURCES [1].

For the current project, SpeechStudio Suite for PC-based VUI and Voice Extreme™ Module for stand-alone-embedded VUI have been employed for Robotics control.

B. Voice User Interface (VUI) in Robotics

Interaction is used general method to guide or control something or somebody. In the case of man-machine interaction, the machine should be equipped with a User Interface (UI) that can be handled by the human user, possibly a novice, to conduct this interaction. In the past decades, GUI (Graphical User Interface), Keyboard, Keypad, and Joystick have been the dominating tools for interaction with machines. Using spoken language to instruct or guide a task is an inherent property of humans in a social context. This suggests using Voice User Interface (VUI) for Robotics control. The last years' development in SR technology has made this possible, even for usage in commercial products.

SR Hardware Module	Manufacturer
Voice Extreme™ Module	Sensory,Inc.
VR Stamp™ Module	Sensory,Inc.
HM2007 – SR Chip	HUALON Micro-electronic Crop. USA
OKI VRP6679 – Voice Recognition Processor	OKI Semiconductor and OKI Distributors Corporate, CA
Speech Commander – Ver-bex Voice Systems	Verbex Voice Systems, USA
Voice Control Systems	Voice Control Systems
VCS 2060 Voice Dialer	Voice Control Systems, USA

TABLE 2: LIST OF SOME AVAILABLE SRHMS AND THEIR MANUFACTURERS [1].

Design of the UI is often given first priority at the designing stage in software development. The primary reason for this is that the UI has an impact on the other designing stages. The UI design for any system (machine) depends on the available input devices. In the Robotics area, UI design depends on the robot's sensors, which are the input devices for the robot. The VUI concept is completely new in Robotics. If we consider the social context, people expect a robot (machine) to understand unconstrained spoken language, so the question of interface requires to be considered prior to the robot design [6]. E.g. if a service robot has to understand the command "Follow the wall", it has to be provided with sensors, by which it can detect the wall. Another important factor is the robot's structure or shape, because instructions introduced in VUI should be related to this factor. E.g. if the robot has wheels like a car, the "move" instruction to the robot should correspond to the car driving environment [1]. Human users have normally adapted the scenarios of giving instruction from the social context. E.g. when they see a robot with wheels like a car, they interact with the robot as if they were navigating a car. *The instruction design for the robot should focus not only on the individual user, but also on other members of the environment, who can be regarded as "secondary users" or "bystanders", who tend to relate to the robot actively in various ways* [8]. Object identification in the environment is another important factor in robot navigation. Instructions are normally given with respect to social and natural language contexts. So the natural description of an object in the social context, like "my room" or "at office", should be interpreted by a robot (machine) that knows about the surrounding environment [2].

One of the most important components of an SR system is the microphone. The system receives the speech from the

environment, through the microphone. A very challenging issue is how to handle the noisy data.

The microphone hears everything, but the system should obey only the valid instructions. E.g. if there is a wall behind the robot and it hears the instruction "back", then it could inform the user instead of driving into the wall. In case of Spoken Natural Language interaction, this feedback is normally given by robot speaking, which means speech generation. This is achieved by a *Speech Synthesizer*, which synthesizes or produces speech, and presents the output through a loudspeaker.

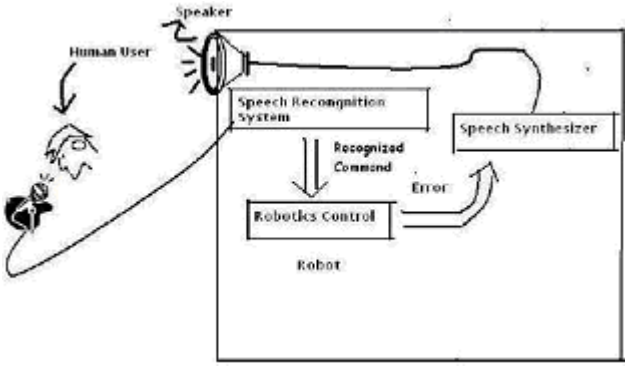

Figure 1: Typical Spoken Natural Language Interface in Robotics [1].

This is normally an efficient way to give feedback to the user. The process of producing speech/sound through a machine is called *Speech Synthesis*. The latest technology of Speech Synthesis creates voice/speech from text, which is called Text-to-Speech Synthesis (TTS). For information about the current situation, the robot (machine) can use a Speech Synthesizer and a loudspeaker, to give feedback to the user through speech or dialog – like "I don't understand". Figure 1 shows a scheme of this example using VUI with TTS, for Robotic Control.

III. Implementations And Tests

Some requirements for the Spoken Natural Language Interface have been identified: Spoken Language should be in English, the Robot should understand the task from the dialog, the system should be speaker-independent, the Robot should have some user feedback and the Robot should understand the dialogs. At the outset of the project, some robot activities have been chosen for implementation, and the dialogs are defined accordingly. Both simple and complex sentences/dialogs have been thus introduced to the system [1].

A. Issues in the Implementation Stage

The implementation has been organized in two phases. In Phase I, a Speech Recognition Hardware Module (SRHM) has been introduced to implement the VUI. In Phase II, a Speech Recognition Software Program (SRSP) has been introduced for the same role. Several hardware and software components have been employed in these implementation phases. Some of the important components are: A miniature mobile robot named Khepera [9], Voice Extreme™ Toolkit (for SRHM), SpeechStudio Suite (for SRSP), LEDs, Microphone, Loudspeaker, MATLAB 7.0.4, and Visual Basic 6.0. The software and hardware components besides the Mobile Robot (Khepera) have been chosen according to the Robot's specifications.

The challenging parts in the implementation phases are implementing the robot's intelligence and then connecting the identified commands with the robot's intelligence. The *Hybrid deliberative/reactive paradigm* [4, 11] has been followed to implement the robot's intelligence. According to this paradigm some behaviors have been employed to introduce intelligence to the robot. These behaviors are: *Move* – straight movement, *Turn* – for turning, *Avoid-Obstacle* – for obstacle avoidance, *Follow-wall* – follow the wall to the left or to right, *Move-to-goal* – find out and follow the goal heading, *Obstruction* – obstruction detection, *At-goal* – identify the goal position [1]. There are also a number of corresponding *commands/dialogues*, which are recognized by the SR tool and then connected to the behaviors. For example, the command "*Move 10 cm"* is connected to the *Move* behavior. Another important part is to organize the behaviors for the specific robotic paradigm, here it is Hybrid paradigm. In general the Hybrid paradigm has five components – Sequencer, Resource manager, Cartographer, Mission planner, Performance monitoring and problem solving. Some modules are developed to emergent the behaviors through the above five components. See the Hybrid architecture in Figure 2.

In Phase I, a challenging part is implementing interaction between the Voice Extreme™ (VE) Module and the Khepera Robot – how can these two devices intercommunicate? A packet transfer communication protocol has been employed to solve this problem. The developers also have to take care of the Language model and the Lexicon issued in the SRHM (in our case the VE module). Therefore, the Language Model and the Lexicon structure have been designed in this phase. The LEDs are used for user feedback. An overview of the implemented system in Phase I is shown the Figure 3.

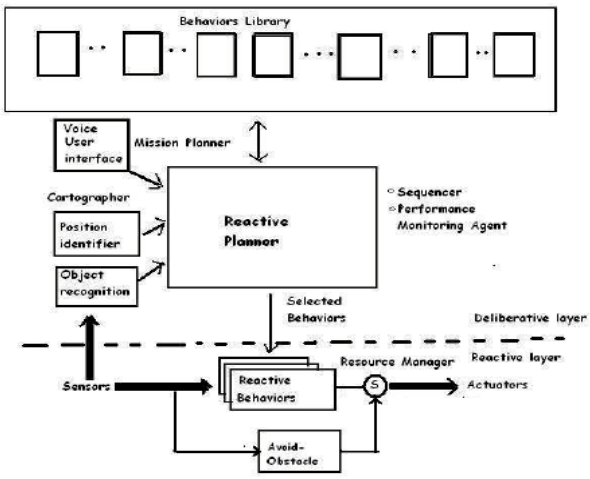

Figure 2: Hybrid architecture for the implemented prototype [1].

In Phase II the Robotic Control and SR program are both integrated in the PC; and a serial protocol [9] is used to control the robot's movements from the PC through a serial cable. The artificial grammar, or language model, has been written in a *grammar file*, which is an XML file [10]. XML is a general language for exchanging

information in a structured way. Here SRSR (SpeechStudio) takes care of both the Language model and the Lexicon. In this Phase TTS technology has been used for Speech Synthesis for feedback to the user: a loudspeaker has been employed to conduct the output of the Synthesize Speech. An overview of the implemented system in Phase II and its domain is shown in the Figure 4.

B. *Test results*

Both implemented systems have been tested according to a *test plan* [1]. Two approaches for testing have been applied. In the first approach the systems have been tested with simple dialogs, and in the second approach the test has been done with the complex dialogs.
The test results show that the software approach (SRSP) with SpeechStudio is more mature than the hardware approach (SRHM) using VE Module.

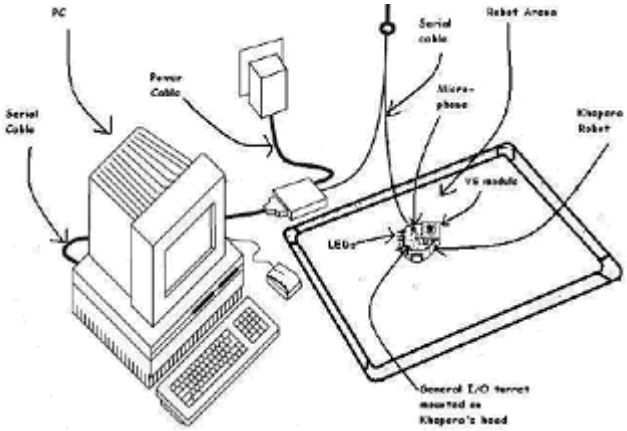

Figure 3: Overview of the implemented hardware-based system in Phase I [1].

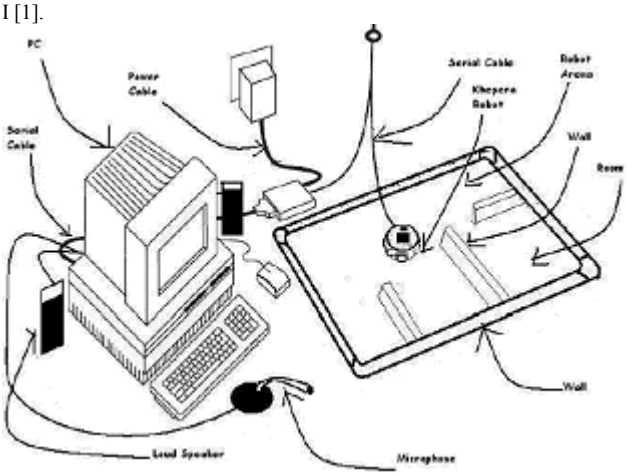

Figure 4: Overview of the implemented software-based system in Phase II [1].

The VE Module is extremely sensitive to noise, which makes it difficult to implement speaker independence. Sentences with three words are not always recognized, because of the user's varying tone. LEDs for issuing user feedback are not suitable, since they demand too high an attention of the user, and there is high chance of missing the feedback [1].
Some limitations have also been found in SRSP with SpeechStudio: the user has to keep the noise at a low level when he or she uses the system. When the user is not planning to communicate with the system, the microphone has to be muted or switched off, to avoid malfunctioning caused by the surrounding noise. For such malfunctioning, the mobile robot needs the Avoid-obstacle behavior, to avoid getting hurt or crashing against the wall. In some cases the system cannot react to the user's speech. The reasons are most often noise, the user's speech not being clear, or the user saying something the robot is not designed to respond to.

The system was tested at the Stockholm International fair 2005, where the visitors could try the SPSP system themselves, and also comment in a questionnaire. The results show that almost every age group found the system easy to control.

IV. Conclusions

Designers of mobile service robots have to focus on the UI, not the least for novice users in a social context. SR technology helps the researcher to implement NL communication with the robot.
SRHM (VE Module) has to be more mature, and because of its memory limitation the developers can only introduce a limited number of instructions or dialogs. SRHM does not support the complex grammar sentences. For user feedback, LEDs are not a suitable interface. To get better performance with the SRSP, the surrounding noise should be kept at a minimal level. The user should mute the microphone to avoid spurious malfunctioning. The conducted user tests indicate that it is relatively easy and quick to learn to control a robot by a Natural Spoken Language interface.

A. *Future work*

The future work will focus on combining voice with gestures into a multi-modal communication [6] interface. This enables more complex instructions and activities. Adding recognition of non-speech sounds [7], like footsteps (close) and footsteps (distant) is another way to extend and improve the communication between the robot and the user.

References

[1] Kibria Shafkat, "Speech Recognition for Robotic Control," Master Thesis Report, Dept. of Computer Science, Umeå University, Umeå, Sweden, Dec. 2005.

[2] Christian Theobalt, Johan Bos, Tim Chapman, Arturo Espinosa-Romero, Mark Fraser, Gillian Hayes, Ewan Klein, Tetsushi Oka and Richard Reeve, "Talking to gobot: Dialogue with a mobile robot," in Proc. IEEE/RSJ International Conference on Intelligent Robots and System 2002, Scotland, UK, pp. 1338–1343, 2002.

[3] UNECE: United Nation Economic Commission for Europe. Press Release ECE/STAT/04/P01, Geneva, 20 Oct. 2004, http://www.unece.org/press/pr2004/04stat_ p01e.pdf (visited 2005-08-25)

[4] Robin R. Murphy, Introduction to AI ROBOTICS, MIT press, UK, 2000.

[5] "Survey of the state of the art in human language technology." Cambridge University Press ISBN 0-521-59277-1, 1996. Sponsored by the National Science Foundation and European Union, Additional support was provided by: Center for Spoken Language Understanding, Oregon Graduate Institute, USA and University of Pisa, Italy.

[6] Guido Bugmann, "Effective spoken interfaces to service robots: open problems," in Proc. AISB 2005, Hatfield, UK, pp. 18–22, April 2005.

[7] Michael Cowling and Renate Sitte, “Analysis of speech recognition techniques for use in a non-speech sound recognition system.,” in Proc. ATS 1997, Akita, Japan, pp. 101–104, Dec. 2002.

[8] Helge Hűttenrauch, Anders Green, Michael Norman, Lars Oestreicher, and Kerstin Severinson Eklund, “Involving users in the design of a mobile office robot,” Systems, Man and Cybernetics, Part C, IEEE Transactions on, 34, Issue:2:113 –124, May 2004.

[9] K-Team Corporation, Rue Galile 9 - Y-Parc, 1400 Yverdon, SWITZERLAND, Khepara User Manual. http://www.kteam.com/download/khepera.html (visited 2005-11-13).

[10] SpeechStudio Inc., 3104 NW 123rd Place Portland, OR97229. SpeechStudio Tutorial for VB6.0 Introduction. http://www.speechstudio.com/. (visited 2006-08-21)

[11] Ronald C. Arkin. BEHAVIOR-BASED ROBOTICS, The MIT press, Cambridge,Massachusetts, London,UK, 1998.

[12] Pierre Nugues, Mathias Haage, and Susanne Schötz. “A prototype robot speech interface with multimodal feedback.” in Proc. of the 2002 IEEE- Int. Workshop Robot and Human Interactive Communication, Berlin Germany, pp. 247–252, September 2005.

[13] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson, “The cog project: Building a humanoid robot.”, Lecture Notes in Computer Science, 1562:52–87, 1999. citeseer.ist.psu.edu/brooks99cog.html