A Semi-autonomous Interactive Robot

B. Schelsinger, M. Mensch, C. Rindosh, J. Votta, Y. Wang

Department of Mechanical Engineering
The College of New Jersey
Ewing, NJ 08534, USA
{ schlesi2, mensch, rindosh, votta, jwang@tcnj.edu}

Abstract

Human-Robot Interaction is a dynamic and expanding research field. This paper presents the creation and concepts of a semi-autonomous interactive robot, TARO. TARO is designed to humanoid in appearance, and can entertain and interact with people through verbal communication and body language.

I. Introduction

One of the best things about modern interactive robotics is the diversity in methods of approaching natural interaction. The main goal of this project was to approach human-robot interaction in such a way that the result would be an unusual first-generation prototype that could serve as a platform for future work.

TARO, the robot developed throughout the course of the project, began as a series of ideas and concepts in the minds of a small group of undergraduate engineering students at The College of New Jersey. Over the course of the eight-month project, TARO was developed from the ground up into a semi-autonomous mobile entity with steadily increasing interactive capabilities.

In addition to ambulatory considerations, topics such as speech recognition, natural language processing, and real-world agent representation were addressed. One long-term goal of the TARO project is to create a robot that is humanoid in both appearance and action.

In its current form, TARO is the result of a union of a several independent programs created in C and PROLOG that are interconnected by interface software written in Visual Basic.

II. Related Work

There have been a large number of projects that included or revolved around human-robot interaction. The Robotics Institute at Carnegie Mellon University has a number of robots that represent advanced contributions to the human-robot interaction community.

GRACE is an excellent example of a mobile robot that uses its humanoid qualities in conjunction with its communicative skills to engage humans in increasingly humanoid interaction. The panning monitor on GRACE's body displays a expressive female face that animates while the robot is speaking. GRACE also possesses a high-quality speech synthesizer and speech recognition capabilities.

Using GRACE's predecessor, Vikia, a study was done on the importance of a robot's expressiveness and attention in face-to-face interaction between humans and robots. [1] From this study, it was concluded that the presence of an expressive face as well as the use of a robot's body to indicate attention makes the robot a more compelling entity for humans to interact with. The results of this study were taken into account when TARO was designed.

Another indicator of the importance of expressiveness beyond speech is the commonly accepted Mehrabian model. In the Mehrabian model, it was noted that only about 7% of the effectiveness of communication is in actual words being spoken. The other 93% of meaning inferred by communication is attributed to body language, facial expressions and tone. [2]

The results of the above studies demonstrate the need for a robot with more interactive capabilities than a microphone and a speaker. For a truly fulfilling interactive experience, a robot needs to give the human interacting with it the impression that they are interacting with a being and not a box.

One of the main differences between the previously mentioned robots and the one developed in the TARO project is that TARO was designed to have more humanoid physical characteristics. Instead of having a cylindrical body, TARO's prototype body has two robotic arms attached to its upper torso, which was designed to roughly emulate the size and shape of the human frame.

III. Hardware and Structure

One thing that makes TARO somewhat different from other robots is its hardware. The first generation prototype TARO uses a relatively small amount of hardware to control all of its functions. However, much of the hardware used was custom-fabricated for the robot. While hardware is not the focus of this paper, it is worth noting that custom hardware and structural designs have contributed to the overall success of the project.

III. Software

Beneath the surface, TARO is composed of a number of interacting pieces of software. Written in C, Visual Basic, and PROLOG, these interdependent programs enable TARO's hardware to function as part of a single entity. Figure 1 is a high-level flow diagram that demonstrates a typical decision cycle. As shown in the diagram, TARO's two different interactive states correspond to two different decision chains that in turn correspond to different parts of its logic software. Described below are the two main pieces of TARO's software system: the Interface Software and the Cognitive Software.

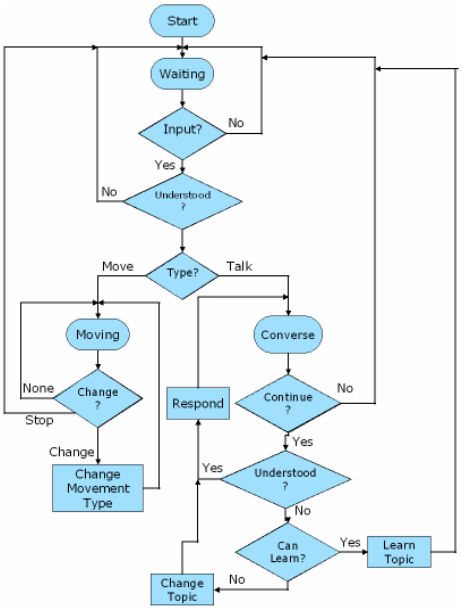

Figure 1: High-level Decision Diagram

4.1 Interface

TARO's interface software is a key part of its identity. The interface software, all of which was written in Visual Basic, allows TARO to take in user utterances as well as generate an appropriate output. The interface software can be split into three subsections: Speech Recognition and Output, Facial Expressions, and Action control.

4.1.1Speech Recognition and Output
TARO's ability to parse utterances and synthesize speech is due largely to the implementation of the Microsoft Speech Software Development Kit. Though other speech recognition packages were considered, the Microsoft Speech SDK proved to be the best for two reasons: it is provided at no cost and it is ready for implementation in Visual Basic.

4.1.1.1 Speech Recognition
Rather than employing simple dictation algorithms for speech recognition which can lead to poor recognition results, a modified command-based system was implemented. By defining structures that limited the number of possible word combinations, the recognition rate was increased dramatically. These structures were dubbed "sub-phrases".

In addition to sub-phrase restrictions placed on the recognition of words, internal state checks were put in place that disallowed some possible word and sub-phrase choices. These state checks monitored the status of the conversation at hand, the state of the robot, as well as the state of the environment. The combination of this with the sub-phrase limitations discussed before was nicknamed "grammar restriction", despite the fact that no formal grammars were defined.

4.1.1.2 Speech Synthesis
The Microsoft Speech SDK provides Text-to-Speech (TTS) engines useful for vocal responses. By passing strings from the logic software to the TTS engine, it was possible to give the robot a voice with which it could interact with other agents in its environment.

4.1.2Facial Expressions
4.1.2.1 Visemes
In addition to enabling Text-to-Speech output, the Microsoft Speech SDK can also return information about the strings passed to it. Two pieces of useful information that can be returned are phoneme and viseme information.
Phonemes are the smallest units of speech in any language that convey meaning. Examples in English are the sounds made by the letters 'R' and 'L'. Phonemes are different from syllables, which divide words at natural pauses. Syllables often contain multiple phonemes.

Visemes are the visual equivalent of phonemes. Whereas phonemes are distinct sounds, visemes are distinct mouth positions. The two main uses of viseme information are for lip reading and animation.

While lip readers learn to associate different mouth positions with sounds, animators use visemes to give the illusion that their characters are creating those sounds.

Disney animators determined that only 13 distinct mouth positions are required to properly animate speech. These are commonly referred to as either the “Disney 12” or the “Disney 13”, depending on whether the closed mouth position is ignored or included.

TARO’s interface software uses viseme information generated by the speech engine to animate the robot’s face. When properly synchronized, the animated face gives the appearance that TARO is speaking the words being generated.

4.1.2.2 Facial Animations
TARO’s face was designed to be simple but effective. Rather than creating a truly humanoid face for the first prototype of the robot, a very simple cartoon-like face was created and used for testing.

Figure 2 shows TARO’s face at rest, meaning that no speech animation is required. TARO’s face is divided into three sections: mouth, left-eye and right-eye. Each of these features can be and often are addressed independently. For example, TARO’s eyes are coded to blink at an appropriate rate to help give the face a more lifelike feel. However, since the two eyes can be addressed separately, a particular situation could cause TARO to wink, closing one eye and leaving the other open.

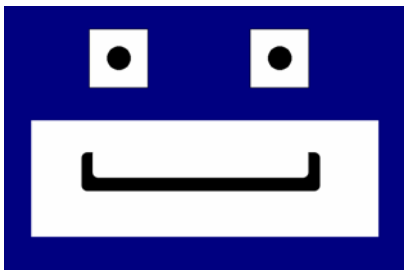

Figure 2: TARO's face at rest

The bulk of the facial animation, however, takes place at the mouth. TARO’s mouth is actually a set of images, each representing one of the “Disney 13” visemes. The mouth remains motionless, held in a small fixed smile until the code signals a change due to speech.

While only 13 images are required to animate speech, the Text-to-Speech engine of the Microsoft Speech SDK can identify 21 distinct visemes. Therefore, it was necessary to map each of the 21 visemes into one of the 13 available images. When a viseme code is generated by the TTS engine, it is first converted into its “Disney 13” equivalent. The new viseme information is then sent to the mouth animation code, which selects the proper image and displays it until new viseme information changes the required image.

4.1.3Action Control

Controlled by the logic software, the Visual Basic action control code is responsible for calling for monitoring all of the movement systems, including robot platform movement and arm control. The software that directly controls the drive motors and the arm motors reside on two microcontrollers. The Action Control code makes use of the COM ports on the computer and transmits movement commands to the two boards as necessary. The microcontrollers send an acknowledgement signal back to the action control code and perform the action dictated by the command.

4.1 Cognitive Software

4.2.1PROLOG
Another thing that differentiates the TARO project from others is the use of PROLOG for the entirety of its decision software, knowledge base, and Natural Language Processing (NLP) capabilities. While the Visual Basic Software described before controls TARO’s input and output functions, it is essentially a body without a brain. All of TARO’s cognitive functions are controlled by its PROLOG code.

Prolog is a prominent logic programming language. Although not as popular in the United States, Prolog is widely used around the world. The decision to use PROLOG was largely based on the fact that PROLOG was originally developed for Natural Language Processing. In future generations of TARO, the NLP capabilities of PROLOG will be used more extensively to create more intelligent NLP software.

4.2.2Implementation Scheme
The logic software was divided into a three categories: Movement, Conversation, and General Logic. Each of these was implemented in a separate logic bank. This way, they could be accessed concurrently. Furthermore, the separation allowed one or more to be ignored as necessary in order to maintain proper operation.

Interoperation with the Visual Basic interface software was a major concern when developing the cognitive software. Therefore, the software was implemented in a way that would allow for easy communication with the VB code. Each of the three logic software categories was implemented as a Dynamic Link Library (DLL). By passing strings to the functions exported by the DLLs and receiving strings corresponding to commands and text to be spoken, the Interface software has access to the all the knowledge needed to run the robot properly. In addition, implementing the cognitive software as DLLs allows for the possibility of implementing the exported functions in any higher level programming language chosen to act as a real-world interface.

4.2.3Movement Control
The movement control DLL indirectly controls both the movement of the robotic arms attached to TARO’s

shoulders as well as the movement of the robot as a whole. Through a set of very simple commands, the movement control code can control motion in a number of ways.

4.2.3.1 Movement of the Robot
Movement of the robot as a whole is controlled either by remote control or by a set of vocal commands. If the robot is in a mode where it is listening for movement commands, a movement request sent from the interface software will cause the robot to move as desired. Table 1 is a list of the one to two-word commands that the robot can listen for when sensitive to movement commands.

Table 1: Vocal Movement Protocol

	COMMAND	DESCRIPTION
Direction	Left	Turn left
	Right	Turn right
	Hard	90° turn
	Forward	Go forward
	Back	Go backward
Speed	Increase	Increase speed
	Decrease	Decrease speed
	Stop	Brings robot to a stop

The protocol listed above includes the command "Hard" which one could prepend to a turn command to signal a 90 degree turn in either direction. Though not in use now, it is expected that the "Hard" command will be utilized in future generations of TARO.

4.2.3.2 Movement of the Arms
The movement of the robotic arms is controlled completely by the logic. The current prototype of TARO offers two options for arm movement. The first is a handshake, which could be called at the beginning or end of a conversation as necessary. In addition, the user can request a handshake during a conversation. The second capability of the arms is waving. Again, as necessary TARO can be directed to wave his left arm. In future generations of TARO, when the robotic arms are redesigned to have more degrees of freedom, this list of commands could be augmented to include more complicated procedures.

4.2.4Conversation

4.2.4.1 General Overview
Conversation was one of the most interesting and challenging aspects of the TARO project. Over the course of the eight-month project, TARO's communicative abilities were developed to a state that has been referred to as "first five minute" conversation.
What is meant by "first five minute" conversation is that TARO can understand and respond to questions and statements that one would encounter in the initial meeting of an individual. TARO's permanent knowledge bank contains information about the robot's origins, design, creators, and future goals. The knowledge bank also contains general information about things such as how to refer to itself or respond to a various greetings. Based on the state of a conversation and the state of the robot, the conversation code can also choose an appropriate response from a group.

4.2.4.2 Natural Language Processing
The trickiest part of conversation is parsing natural language. The solution chosen for the first generation of TARO presented itself in the form of the grammar restriction used in the speech recognition. By training the NLP code to determine meaning of the incoming sub-phrases within strings and routing decisions based on the determined sub-phrase meaning, it was possible to determine the meaning of the entire incoming sentence.
While use of sub-phrase recognition limits TARO's vocabulary to sub-phrases it recognizes, it also ensures accurate meaning determination. In the future, the algorithms governing TARO's Natural Language Processing capabilities will be improved significantly to allow for more intelligent decision making.

4.2.5General Logic
The general logic DLL controls all that is not controlled by either the movement control software or the conversation code. In its current form, the general logic code is perhaps the least used code. However, it does have a few important roles.
The first major task of the general logic DLL is to determine what action state TARO should be in when it is first activated. When activated, TARO comes up in a neutral state where it looks for one of two commands that will determine whether its first task is to move or to interact. After one of those commands is recognized, the rest of the operation is governed by the appropriate DLL.
The other major task of the general logic DLL is to keep track of the robot's internal states. Knowledge of conversation state, environment state, and the robot's emotional state are stored here. At present, the emotional state is only used to determine which greeting and parting phrases to return. However, the use of this emotion information could easily be developed into something more frequently used.

V. Results

Though not the most advanced piece of machinery in its field, the first generation TARO prototype is a promising first step in the development of a state-of-the-art interactive humanoid robot. Figure 3 is a photo of the completed first generation TARO prototype.

Overall, the first TARO prototype was a success for the TARO team. The union of Visual Basic, C, and PROLOG worked exceptionally well as did the Microsoft Speech Software Development Kit which was implemented. The grammar restriction ensured that an phrase recognized by

TARO was an exact match of the uttered phrase nearly 90% of the time.

Most of the problems encountered were hardware related. For example, it was determined late in the project that the arm control motors did not have the torque to consistently lift the arms without slipping. However, none of the problems affected the state of the robot.

The first generation prototype did bring to light several software improvements that can be implemented in coming generations. These improvements are discussed in the Future Work section of this paper.

Figure 3: Finished Prototype

V. Future Work

There are several areas in which the next generation TARO prototype will improve on the current model.

- Physical Body

  First, the physical body of the robot will be redesigned to improve its humanoid characteristics. Increasing the humanoid nature of the robot is one of the main goals of the project.

- Artificial Intelligence Algorithms

  Second, the natural language processing and overall artificial intelligence systems will be augmented and revamped to include more intelligent algorithms and more current topics. A larger array of conversation topics is desired as is a system through which new topics can be learned

- Voice

  A third planned improvement is the creation of a distinct voice for TARO. TARO's current voice is a Microsoft default voice. For the next generation, a new voice will be created to give TARO a recognizable speech sound.

- Vision System

  The final planned improvement is the introduction of a vision system. In its first stages, the system will be used for obstacle avoidance. It is hoped that after some development, the vision system could be used for object recognition and eventually facial recognition.

Acknowledgements

The authors would like to extend their deepest thanks to the rest of the TARO team: Michael Mensch, Christopher Rindosh, and Joe Votta. In addition, our thanks go out to Dr. Miroslav Martinovic, Mr. Jay Ross, and Mr. Alex Michalchuk for there invaluable contributions to the project.

References

[1] A. Bruce, I. Nourbakhsh, R. Simmons, "The Role of Expressiveness and Attention in Human-Robot Interaction," in Proc. ICRA. May, 2002

[2] A. Chapman, "Albert Mehrabian Communications Research". Buisnessballs.com, 2004 – 2006. <http://www.businessballs.com/mehrabiancommunications.htm>