Voice Enabled Interface for Interactive Tour-Guide Robots

Plamen J. Prodanov[1], Andrzej Drygajlo[2], Guy Ramel[1], Mathieu Meisser[1], Roland Siegwart[1]

[1]*Autonomous Systems Lab,* [2]*Signal Processing Institute*
Swiss Federal Institute of Technology, Lausanne, Switzerland, plamen.prodanov@epfl.ch

Abstract

This paper considers design methodologies in order to develop voice-enabled interfaces for tour-guide robots to be deployed at the Robotics Exposition of the Swiss National Exhibition (Expo.02). Human-robot voice communication presents new challenges for design of fully autonomous mobile robots, in that interactivity must be robot-initiated in conversation and within a dynamic adverse environment. We approached these general problems for a voice enabled interface, tailored to limited computational resources of one on-board processor, when integrating smart speech signal acquisition, automatic speech recognition and synthesis, as well as dialogue system into the multi-modal, multi-sensor interface for the expo tour-guide robot. We also focus on particular issues that need to be addressed in voice-based interaction when planning specific tasks and research experiments for Expo.02 where tour-guide robots will interact with hundred of thousands of visitors during six months, seven days a week, ten hours per day.

1. Introduction

Mobile tour-guide robots intended for large-scale exhibitions pose a unique challenge to trans-disciplinary research. They require the integration of sensing, acting, planning and communicating within a single system. These are all difficult problems that must be solved if we are to have truly autonomous, intelligent systems.

An important aspect of the tour-guide robot is its interactive component [12], [14]. Human-robot interfaces are of great importance for robots that are to interact with ordinary people. In the setting of the national exhibition, where people typically do not spend extensive amounts of time with a single robot, two criteria are considered most important: ease of use, and interestingness. The human-robot interfaces must be intuitive, so that untrained and non-technical visitors of the exhibition can operate the system without prior instruction. Interestingness is an important factor in capturing people's attention.

Natural spoken communication is the most user-friendly means of interacting with machines, and from the human standpoint spoken interactions are easier than others, given that the human is not required to learn additional interactions, but can rely on "natural" ways of communication [4]. Although human-robot voice enabled interfaces are still in their infancy, they have the potential to revolutionize the way that people interact with tour-guide robots. As a result, there have been several attempts to build tour-guide robots with spoken language interaction capabilities [2], [1], [13], [14].

In an exhibition environment, the tour-guide robot often interacts with individual visitors as well as crowds of people. The type of interaction faced by such a robot is spontaneous and short-term; since visitors typically have no prior exposure to robotics technology, and can not be instructed beforehand as to how to operate the robot. This type of interaction differs significantly from the majority of interactive modes studied in the field, which typically assume long-term interaction with a robot [8]. It is important that the expo robot takes the initiative and appeals to the "intuitions" of visitors. Thus, a primary component of a successful tour-guide is the ability to be aware of the presence of people, and to engage in a meaningful conversation in an appealing way.

The major issues in implementing human-robot voice enabled interfaces are: speech output (loud-speakers) and input (microphones), speech synthesis for voice output, speech recognition and understanding for voice input, dialogue and usability factors related to how humans interact with tour-guide robots [11]. They function by recognizing words, interpreting them to obtain a meaning in terms of an application, performing some action based on the meaning of what was said, and providing an appropriate spoken feedback to the user. Whether such a system is successful depends on the difficulty of each of these four steps for the particular application, as well as the technical limitations of the system. Robustness is an important requirement for successful deployment of such a technology (in particular speech acquisition and speech recognition) in

real-life applications. For example, automatic speech recognition systems have to be robust to various types of ambient noise and out-of-vocabulary words. Automatic speech synthesis should not only sound naturally but also be adapted to an adverse acoustical environment. Lack of robustness in any of these dimensions makes such systems unsuitable for real-life applications.

In this paper, we report about the joint efforts of the robotics and speech processing research groups of the Swiss Federal Institute of Technology Lausanne in providing a voice enabled interface to the interactive tour-guide robot RoboX, which has been recently developed for Expo.02 at the Autonomous Systems Lab [5].

2. Design Philosophy Background

The first specificity for the Swiss National Exhibition Expo.02 is that tour-guide robots to be deployed in the robotic exposition should be capable to interact with visitors using four official languages: French, German, Italian and English. They have to attract people's attention, to show them the way to exhibits and to supply information about these exhibits. Studying other specificities of autonomous, mobile tour-guide robots led us to the following observations.

First, even without voice enabled interfaces, tour-guide robots are very complex, involving several subsystems (e.g. navigation, people tracking using laser scanner, vision) that need to communicate efficiently in real time. This calls for speech interaction techniques that are easy to specify and maintain, and that lead to robust and fast speech processing.

Second, the tasks that most tour-guide robots are expected to perform typically require only a limited amount of information [11] from the visitors. These points argue in favor of a very limited but meaningful speech recognition vocabulary and for a simple dialogue management approach. The solution adopted is based on yes/no questions initiated by the robot where visitors' responses can be in the four official languages of the Expo.02 (oui/non, ja/nein, si/no, yes/no). This approach lets us simplify the voice enabled interface by eliminating the specific speech understanding module and allows only eight words as multi-lingual universal commands. The meaning of these commands depends on the context of the questions asked by the robot.

A third observation is that expo tour-guide robots have to operate in very noisy environments, where they need to interact with many casual persons (visitors). This calls for speaker independent speech recognition and for robustness against noise.

The basic philosophy of the design methodology proposed in this paper is to develop voice enabled interfaces that are adapted to the nature of autonomous, mobile tour-guide robots, behavioral requirements on the side of visitors and real-world noisy environments. The automatic speech recognition and synthesis systems have to cope with this.

3. Architectural Overview

A functional architecture model for voice-enabled interface of RoboX is shown in Figure 1.

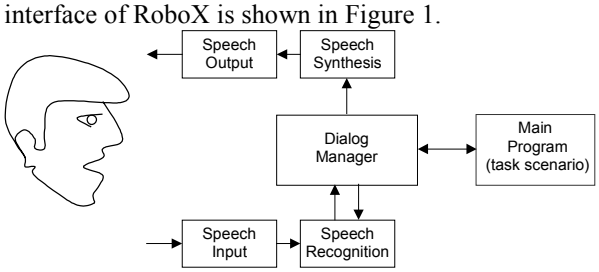

Figure 1 *Voice-enabled interface*

It consists of speech output (loud-speakers) and speech input (microphones), speech synthesis for voice output, speech recognition for voice input and dialogue management that controls the sequence of verbal information exchange between the visitor and the robot given the state of the other modalities and a pre-defined sequence (task scenario) of events (scenario objects) [5].

Speech is one of the input/output modalities within the multi-modal, multi-sensor interface of the robot and should naturally fit into functional layers of the whole system. On the other hand, from a functional and conceptual point of view, the addition of a voice enabled interface does not affect the overall system organization, implementation should take some specific constraints into account.

3.1 Hardware Architecture

Figure 2 presents the hardware architecture of RoboX. It consists of three layers: input/output (I/O) layer and two (low- and high-level) processing layers.

Multiple sensors and other input/output devices of the I/O layer are used by the robot to communicate with the external world, in particular with users. In this set of multi-modalities, loud-speakers and a microphone array (Andrea Electronics DA-400 2.0) represent the output and input of the voice enabled interface. They are

installed at half the height of the robot, which is a convenient position for both children and adults.

Among input devices that have to cooperate closely with this interface, when verifying the presence of visitors, are two SICK laser scanners mounted at knee height and one color camera placed in the left eye of the robot. The blinking buttons help in choosing one of the four languages, and the robot's face, which consists of two eyes and two eyebrows can make the speech of the robot more expressive and comprehensive. Finally, a LED matrix display in the right eye of robot may suggest the "right" answer to the robot's questions [5].

The low-level processing layer contains hardware modules responsible for pre-processing of signals dedicated to input and output devices. The voice pre-processing is represented in this layer by the digital signal processor of the microphone array and the audio amplifier for the loud-speakers.

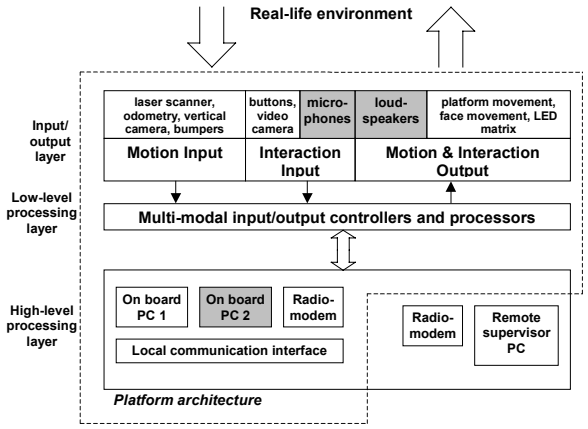

Figure 2 *Hardware architecture*

The high-level processing layer consists of two on-board computers: Pentium III (700M Hz, 128MB, 30GB HDD, Windows 2000) dedicated for all interaction tasks, including speech synthesis, speech recognition and dialogue management, and PowerPC 750 (400 MHz) for navigation. Both computers can communicate with each other via local Ethernet and with external monitoring computer via wireless modems.

3.2 Software Architecture

In the case of RoboX, the principal robot operations are controlled by one main program called sequencer, which executes a predefined sequence (task scenario) of events (scenario objects). The overall architecture of the sequencer including speech synthesis and recognition objects and dialogue sequence management is depicted in Figure 3. The sequencer program is implemented in SOUL (Scenario Object Utility Language) designed at ASL to meet the requirements of the autonomous, interactive, mobile tour-guide robot. The main program is defined as a graph like scenario where the execution of the sequence of events corresponding to a predefined task is strictly linear [5]. The events generated by the sequencer should be treated as logical events. Therefore, each of the scenario objects have a finite number of possible outcomes, which reflect the different states of this object after its execution. For example the speech recognition object has three possible outcomes, corresponding to yes and no answers, and maximal execution time flag (time-out). Several scenario objects may be running in parallel, e.g. speech synthesis and face movement objects.

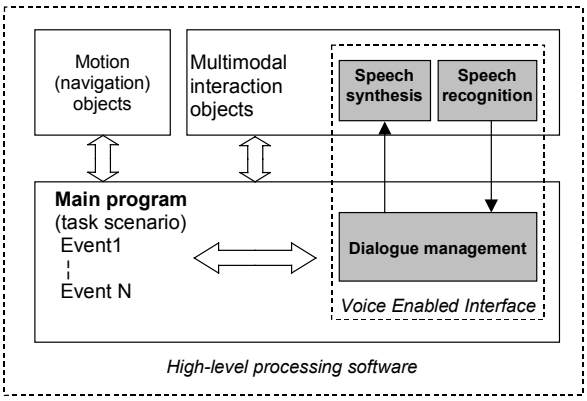

Figure 3 *Software architecture*

The main task scenario of Robox is to guide the visitors of the exhibition in accordance with the predefined tour plan and visitor's expectations, when coordinating the various robot activities related to sensing, motion and visitor-robot interaction. A dialogue scenario has to fulfill these required properties of the main task scenario by appropriate verbal expressions, explanations and questions of the robot and the visitors' confirmations. In the main program, the dialogue scenario in the form of the sequence of dialogues named Introduction, Exhibit 1, Exhibit 2, … , Next Guide, as presented in Figure 4, is embedded in the task scenario.

1. *Introduction.*
2. *Exhibit 1*
    a. *Description.*
    b. *Do you know that…?*
    c. *Do you want to …?*
3. *Exhibit 2*
4. *…*
5. *Next guide*

Figure 4 *Dialogue scenario*

Some examples of dialogue sequences are presented in section 5. Concepts of speech synthesis and speech recognition objects and the corresponding programs are presented in section 4.

4. Voice-Enabled Interface

To start interacting with people, a method for detecting them is needed. We have found that in the noisy and dynamically changing conditions of the robotics "exposition", a technique based on motion tracking using laser scanners, and on face detection with a color video camera gives the best results [5]. When RoboX finds people in the distance smaller than 1.5 meters it should greet people and inform them of its intentions and goals. The most natural and appealing way to do this is by speaking. In the context of the national exhibition (four official languages) and having the possibility of rapid prototyping of complex interaction scenarios when using the voice enabled interface, speech becomes one of the most important output modalities to be used for communicating with visitors.

4.1 Speech Synthesis

In the noisy environment of the exposition, the automatic speech synthesis system should generate speech signals that are highly intelligible and of an easily recognizable style; if possible, this style should correspond to the style of an excellent human guide. On the other hand, and to preserve the robot's specificity, the quality of its speech should not mimic perfectly the human speech, but such speech has to sound natural. Two main criteria that we have used to choose an appropriate method for automatic speech synthesis were intelligibility and naturalness.

Therefore, a solution adopted for the speech synthesis event is a text-to-speech (TTS) system based on concatenation of diphones (phonetic units that begin in the middle of the stable state of a phoneme and end in the middle of the following one) [3]. The actual task of the synthesizer is to produce, in real time, an adequate sequence of concatenated segments, extracted from its parametric segment database and the prosodic parameters of pitch pattern and segmental duration adjusted from their stored values, to the one imposed by the natural language processing (NLP) module. The intelligibility and naturalness of the synthesized speech highly depends on the quality of the segment database, grapheme-to-phoneme-translation and a prosodic driver for pitch and duration modification.

During the experimentation phase with RoboX, the best results, e.g. for French, were achieved for the combination of LAIPTTS (NLP) [7], Mbrola reproduction tools and a Mbrola parametric segment database. For all four application languages (French, German, Italian and English) the structure of the speech synthesis system is the same, and the system can be limited to Mbrola phonetic files generated off-line by the NLP module, Mbrola synthesis engine and parametric segment databases for different languages.

When RoboX needs a yes/no response from the visitor, the speech synthesis event is directly followed by the speech recognition event in the task scenario.

4.2 Speech Recognition

The first task of the speech recognition event is the acquisition of the useful part of the speech signal that avoids unnecessary overload for the recognition system. The adoption of limited in time (3 seconds) acquisition is motivated by the average length of yes/no answers. During this time the original acoustic signal is processed by the microphone array. The mobility of the tour-guide robot is very useful for this task since the robot, when using the people tracking system, can position his front in the direction of the closest visitor and this way can direct the microphone array. The pre-processing of signals of the array includes spatial filtering, de-reverberation and noise canceling. This pre-processing does not eliminate all the noise and out-of-vocabulary (other than yes/no) words. It provides sufficient quality and non-excessive quantity of data for further processing.

Recognition should be speaker independent and multi-lingual performing equally well on native speakers and on speakers who are not native of the target language. The system is intended to recognize the limited vocabulary of eight words (oui, non, …) but can accept an unlimited vocabulary input. In such a system, we are not only interested in a low error rate, but also in rejection of irrelevant words

At the heart of automatic speech recognition system of the robot lies a set of the state-of-the-art algorithms for training statistical models of words and then using them for the recognition task [10]. In speech recognition event the signal from the microphone array is processed using a Continuous Density Hidden Markov Model (CDHMM) technique where feature extraction and recognition using the Viterbi algorithm are adapted to a real-time execution. The approach selected to model eight key words (oui, non, ja, nein, si, no, yes, no) is the speaker independent flexible vocabulary approach. It

offers the potential to build word models for any speaker using one of the four official languages of Expo.02 and for any vocabulary from a single set of trained phonetic sub-word units. The major problem of a phonetic-based approach is the need for a large database to train, initially, a set of speaker-independent and vocabulary independent phoneme models. This problem was solved using standard European and American databases available from our speech processing laboratory, as well as specific databases with the eight key-words as recorded during experiments. The CDHMM toolkit (HTK) [15] based on the Baum-Welch algorithm was used for the training.

Out-of-vocabulary words and spontaneous speech phenomena like breath, coughs and all other sounds that could cause a wrong interpretation of visitor's input have also to be detected and excluded. For this reason a word spotting algorithm with garbage models have been added to the recognition system. Finally, the basic version of the system is capable to recognize yes/no words in four languages, the name of the robot (RoboX) and speech acoustic segments (undefined speech input) associated to the garbage models.

5. Dialogue Management

A particular problem, when designing a dialogue system is to describe its functional structure in a compact, precise and readable way. Some graphical state-based formalism has been adapted to represent the sequences of dialogues. Some of the possible sequences are presented in Figures 5 – 7. They include not only speech events but also some non-speech events, e.g. move event, motion tracking event, behavior event, etc.

6. Planning of the Expo.02 Experiments

During a five-month period from May 15 to October 21, 2002, ten RoboX systems will interact with the visitors of Expo.02. An important aspect of the tour-guide robot interactive component (voice-enabled interface) is the robot's physical reaction to people and vice versa. During this period we plan to monitor the performance of the voice enabled interface in adverse environment conditions and to experiment with different scenarios. For this purpose, a database including visitors' responses and other data will be recorded. These data will be used for optimizing the existing scenario and for developing new ones. They will also be used to optimize the parameters controlling the two main events of the voice-enabled interface: speech synthesis and speech recognition.

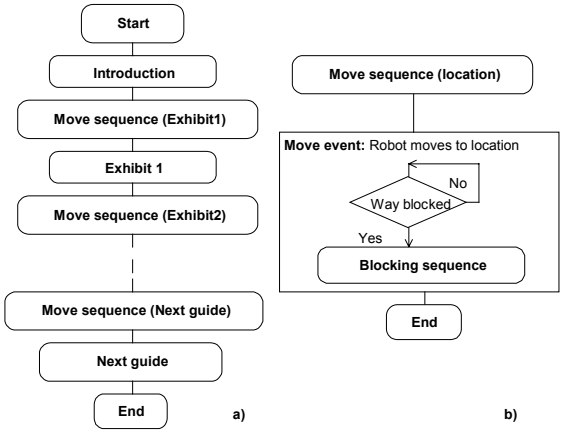

Figure 5 *a) Main sequence, b) Move sequence*

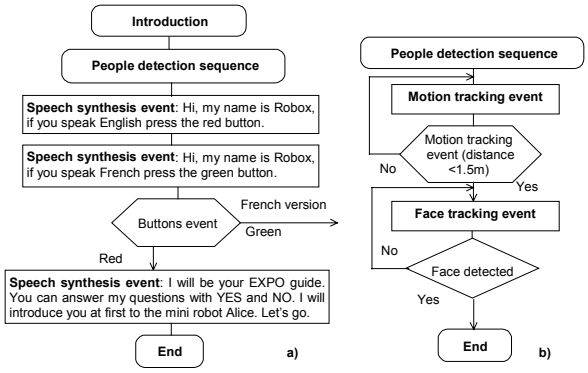

Figure 6 *a) Introduction sequence, b) People detection sequence*

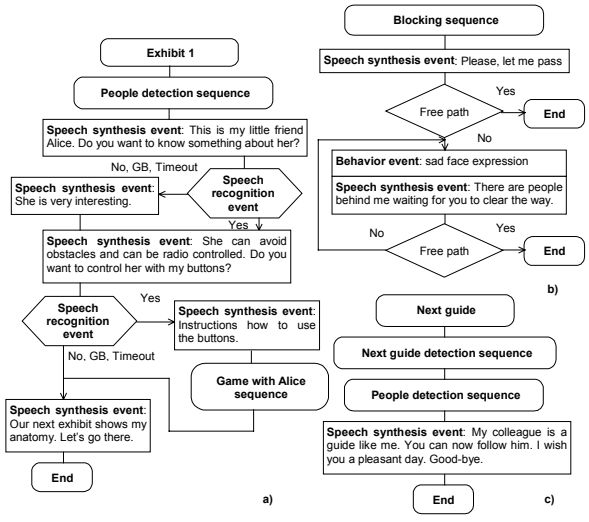

Figure 7 *a) Exhibit 1 sequence, b) Blocking sequence, c) Next Guide sequence*

7. Conclusions

In this paper, the complete methodological concept for designing and implementation of voice-enabled interface for a tour-guide robot to be deployed at the Swiss National Exhibition (Expo.02) was presented. While the issue of navigation of mobile robots is relatively well understood, the issue of voice enabled interaction in adverse environment remains largely open; despite the fact that interaction is a key ingredient of any successful application.

The design methodology proposed in this paper is conceived for developing voice enabled interfaces that are adapted to the nature of autonomous, mobile tour-guide robots with all their constraints, behavioral requirements of visitors and real-world noisy environments that the automatic speech recognition and synthesis systems have to cope with. In the approach presented, the development was focused on the potential user, from the very beginning of the design process through to the complete system.

In the paper, it is demonstrated that this methodology allows for developing a variety of tour-guide scenarios, involving real-time state-based dialogue management on a single on-board computer responsible for the overall interaction of the robot and related modalities. We lack however a convincing methodology for "intuitive" human robot interaction planning.

Acknowledgements

This research work was supported partly by Expo.02 and the Swiss National Science Foundation project "Intelligent Voice Enabled Interfaces for Human Guided Mobile Robots". We acknowledge the help provided by Prof. Keller from the Laboratory for Computational Analysis of Speech (LAIP) at the University of Lausanne, in the form of software tools (LAIPTTS – SpeechMill) for high quality French and German speech synthesis.

8. References

[1] W. Burgard et al., "Experiences with an interactive museum tour-guide robot", Artificial Intelligence, 114(1-2), 2000, pp. 1-53.

[2] A.C. Domínguez Brito et al., "Eldi: An Agent Based Museum Robot", *European Workshop on Servive & Humanoid Robots (ServiceRob 2001)*, Santorini, Greece, June, 2001.

[3] Th. Dutoit. *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht Hardbound, ISBN 0-7923-4498-7, April 1997.

[4] X. Huang, Al. Acero, Hs. Hon, *Spoken Language Processing*, Prentice Hall, New Jersey, 2001.

[5] B. Jensen, G. Froidevaux, X. Greppin, A. Lorette, M. Messier, G. Ramel, R. Siegward, "The Interactive Autonomous Mobile System RoboX", submitted to *IROS 2002.*

[6] K. Kawamura et al., "Design Philosophy for Service Robots", *Robotics and Autonomous Systems*, Vol. 18, Nos. 1-2, July 1996, pp. 109-116.

[7] E. Keller, S. Werner, "Automatic Intonation Extraction and Generation for French". *14th CALICO Annual Symposium*, ISBN 1-890127-01-9, West Point.NY, June 1997.

[8] T. Matsui et al., "Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services", *AAAI-99*, Orlando, FL, July 1999.

[9] R. de Mori (Ed.), *Spoken Dialogues with Computers*, Academic Press, London, 1998.

[10] Ph. Renevey, A. Drygajlo, "Securized Flexible Vocabulary Voice Messaging System on UNIX Workstation with ISDN Connection", *European Conference on Speech Communication and Technology, Eurospeech 99*, Rhodes, Greece, 1997, pp 1615-1619.

[11] D. Spiliotopoulos, I. Androutsopoulos and C.D. Spyropoulos, "Human-Robot Interaction Based on Spoken Natural Language Dialogue". *European Workshop on Service and Humanoid Robots (ServiceRob 2001)*, Santorini, Greece, 2001.

[12] S. Thrun et al., "Experiences with two deployed interactive tour-guide robots", *International Conference on Field and Service Robotics (FSR'99)*, Pittsburgh, PA, August, 1999.

[13] S. Thrun et al., "Minerva: A second generation museum tour-guide robot". *IEEE International Conference on Robotics and Automation (ICRA'99)*, Detroit, Michigan, May 1999.

[14] T. Willeke, C. Kunz, I. Nourbakhsh, "The History of the Mobot Museum Robot Series: An Evolutionary Study", *FLAIRS 2001*, May, 2001.

[15] S.Young, J. Odell, D. Ollason, P. Woodland. *The HTK Book. Version 3.0*, 2000.