Effective Spoken interfaces to service robots: Open problems.

Guido Bugmann
School of Computing, Communications & Electronics
University of Plymouth, Plymouth PL4 8AA, United Kingdom
gbugmann@plymouth.ac.uk

Abstract

This paper discusses some of the open problems in spoken man-machine interfaces that were highlighted during the development of a human-robot interaction (HRI) system enabling humans to give route instruction to a robot. i) Naïve users only know how to explain tasks to other humans, using task decomposition consistent with human execution capabilities. Robots able to understand such instructions need similar (high-level) execution capabilities. Consequently, the current lack of knowledge in some areas of artificial perception, motor control, etc. is a limiting factor in HRI and in the development of service robot applications. ii) Human language is full of inaccuracies and errors, yet communication is effective because of the use of error-repair strategies. Future HRI systems may need human-like repair mechanisms. iii) At the sensory level, the inability to deal with noisy environments limits the range of possible applications. It is suggested that the analysis, not of the user's needs, but of the user's ways of expressing these needs should drive research in robotics and HRI.

1 Introduction

This paper discusses a number of hard (yet unsolved) problems encountered during the development of a NL interface for the instruction of service robots. The development of such interfaces is based on following working assumptions:

1. A service robot cannot be pre-programmed by the manufacturer for all possible tasks and users will need to give some form of "instruction" to their new robot. E.g. specifying which pieces of furniture can be moved during cleaning, and which ones should not be moved, or how to prepare a given variety of tea.
2. To give instruction is in general a multimodal process including verbal description of rules, pointing movements to define objects and demonstration of complex movement sequences. Such process, although akin to programming, is not tractable with conventional programming tools.
3. Human instructors are familiar with methods for instructing other humans, but are unskilled in the art of robot programming. Few have the ability or inclination to learn formal programming languages.

It appears therefore that service robots need to be programmed in a novel way, by users who only know how to explain a task to another human. A solution to that problem is to give robots the ability to understand human-to-human instructions.

Such a solution was explored in a project on Instruction-Based Learning (IBL) which focused on verbal instructions in a direction-giving task in a miniature town. In one way, this project achieved its objectives in that it demonstrated an effective method for generating robust robot programs from spoken instructions (Bugmann et al., 2004). For that purpose, a corpus-based method was developed for building into the linguistic and functional domain of competence of the robot all expressions and action concepts natural to unskilled users, through the analysis of a corpus of utterances representative of the domain of application. However, the work also revealed a number of hard problems that need to be solved before effective commercial robot instruction systems can be developed. Interestingly, these are never pure robotics or natural language processing problems, but involve both areas to various degrees. Following sections detail these problems.

2 Spoken interfaces constrain the robot’s design

In the area of computer software development, it is a recognized practice to specify the user interface early in the design process and then to design the software around the interface. In robotics, this is a new concept, as spoken interfaces were very much seen as the last component to be added to a robot. This traditional approach then automatically requires the user to learn the specific language and keywords prepared by the robot designer. However, if one expects the robot to understand unconstrained spoken language, then the question of interface needs to be considered prior to robot design.

To illustrate this, let us assume that a user of a domestic robot-cook needs to give an instruction involving the expression “a pinch of salt”. This will clearly exert constraints on how the robot’s manipulators are to be designed. Similarly, if a mobile robot needs to understand the command “turn right at the blue sign”, it will need to be provided with colour vision.

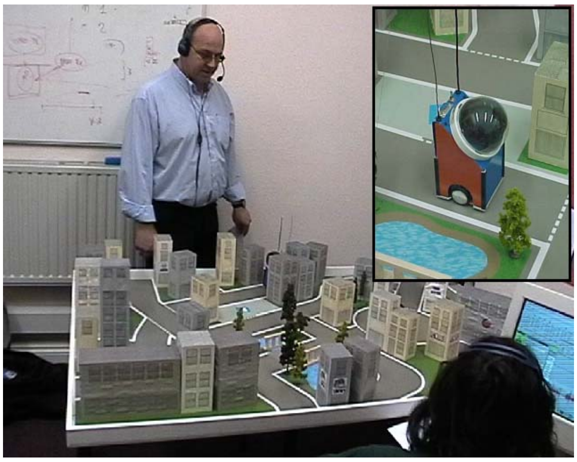

Figure 1. A subject instructing the robot during corpus collection. Inset: remote-brained mini-robot.

In the IBL project, work started by collecting a corpus of route instructions from subjects explaining to human how to drive a robot in a miniature town towards a destination (fig 1). Their analysis revealed 13 primitives functions, some which where navigation procedures such as “take the nth turn right/left” some where just informative statements such as “you pass the post-office to your left” (for more details see Bugmann et al., 2004). Only after this analysis did work start on designing the vision and control system, to build all robot functions required by HRI (Kyriacou et al., 2005).

Note that a command such as “turn right” is highly under-specified for a robot, with no details on what actuators must do. Hence service robots must gather missing information from the environment and make autonomous decisions, e.g. recognize the layout and plan a trajectory. To understand natural language, a robot needs a high level of functionality. In fact, utterance like “clean this window” or “hang up the washing” make demands on robot design and control that are beyond current knowledge. Given that these are expressions that future users are likely to use, it is of concern that relatively little research is devoted to the corresponding robot skills.

There are also examples where particularities of human language, e.g. references to combinations of procedures, exert more subtle constraints on various aspects of robot design, such as its computational architecture (see e.g. next section).

3 Spatial-language-specific problems.

Hereafter are examples of difficulties that natural language in the domain of route instruction creates in both the NLP and robotics domains.

D*etecting references to previous routes.* During corpus collection, subjects were encouraged to refer to previously taught routes whenever possible, rather than re-describing every step of a route. It turned out that such references are very difficult to detect in instructions. In one third of the cases, subjects referred to previous route implicitly, e.g. via a landmark that was part of a previous route. For instance, when a subject said “go to the roundabout”, it was impossible to tell if this referred to a roundabout that is just in front of the robot or a roundabout further away that can be reached using parts of a route previously instructed. In two third of the cases, the destination of a previous route was explicitly mentioned “start as if you were going to the post-office” but in half of these cases, the sentences had structures that could not be properly translated by our NLP system.

Interestingly, experiments with human subjects listening to the same instructions showed that only 55% of references to previous routes were detected in the instruction. Only when subjects started to drive the robot (by remote control) did they notice that there was a problem.

Using references to previous routes when creating program codes. Almost all references to previous routes required only a partial use of the instruction sequence: e.g. “take the route to the station, but after the bridge turn left”. One of the problems is that the bridge may not even be mentioned in the instruction of the route to the station. No definite solution has been found to that problem. One proposal was to implement a multi-threaded concurrent processing scheme where the robot would “follow the road to the station" and at the same time "try to find the left turn after the bridge". The second process would remain the sole active as soon as the turn is found (Lauria et al., 2002). It remains to be seen if this solution is general enough, but it is interesting to note that the way users express themselves could end up dictating the computational architecture of the robot controller.

Programming the final instruction. The final instruction of a route instruction is often a statement like "and you will see it there on your left". The final instruction is especially interesting as it is the one requiring the most autonomy from the robot. It is highly under-specified and the robot needs to visually locate the destination and then plan a path towards it. In our miniature town, we have not undertaken the difficult task of detecting the building, identifying it from its sign and locating its entrance. Instead, a coloured strip was placed at the foot of the building to signal its position. In a real urban environment the final instruction would pose vision and control challenges that are at the limits of current technical capabilities.

4 Handling misunderstandings

Robots are designed with a limited vocabulary corresponding to their action capabilities. In principle this simplifies the design of NLP components and improves the performance of speech recognition. However, users do not know the limits of the domain of competence of the robot and often address the robot with utterances that it cannot understand.

The standard approach to solving this problem is increasing the grammar, e.g. by collecting a larger corpus of utterances natural to users in that domain, then tuning the grammar to that corpus. However, this approach improves the grammar only modestly for a large effort in corpus collection (Bugmann et al., 2001). Another approach is to adding to the grammar a sample of potential out-of corpus expressions (Hockey et al., 2003). However, no matter how large the coverage of the grammar, a robot always has a limited domain of linguistic and functional competence. When the user steps out of this domain, communication brakes down.

Another approach is to accept the domain limitation and work with it. Somehow, the robot should be able to help the user naturally discover its domain of competence. An impractical alternative would be to ask the user to undergo long and detailed training sessions on the robot's capabilities. Both approaches can also be seen as two stages of dialogue system development (fig. 2)

A dialogue system that informs the user about the robot's competences is not straightforward to design. First, it requires that the out-of-domain error is detected. Unfortunately, speech recognition systems are not good at detecting that their interpretation of the speech sounds is incorrect. They tend to generate a translation that is consistent with the grammar of the domain of competence. For instance, if the user asks the robot "go to Tescos", and the word "Tesco" is unknown to the system, the translation may be "go to the school", a perfectly legal request. It is possible that the user detects the error at some point in the dialogue and attempts to engage in a correction dialogue with the robot. However, research on such dialogues is still in its infancy. If the system detects the error, e.g. through a low speech recognition confidence score, how will it inform the user that it does not know the word "Tesco" if it is not in its vocabulary? Some errors may not be speech recognition errors, but requests incompatible with the robot's capabilities. In general, to generate a helpful message, the speech generation sub-system must be aware of what the robot can and cannot do. However, this is a manufacturer-specific knowledge. How should it be represented? In conversations between humans, all these comprehension problems also occur but, after a few clarifications, speakers usually manage to align their utterance to the domain of common ground (Garrod and Pickering, 2004), or learn new concepts if necessary. This is an area where findings from life sciences could help develop more effective human-robot dialogue systems. We are currently planning work in this area.

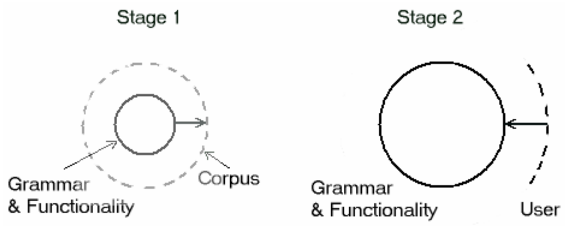

Figure 2. Two stages of spoken user-interface development. In stage one, the system is adapted to the user represented by the corpus. In stage 2, the user is informed of the capabilities of the robot.

5 Speech recognition in noisy backgrounds

Speech recognition has made significant progress in recent years, as evidenced by a number of effective commercial software packages, and does not constitute anymore the bottleneck in natural language interfaces. However, this has given other problems more prominence. In the IBL project, we used a microphone placed near the mouth and switched it on only for the duration of the speech. This enabled effective speech recognition even in a noisy background such as an exhibition. However, if the microphone were always on, the system would start interpreting the background noise. A possible solution could be to establish a directional window using an array of microphones (e.g. eight microphones used on the JiJO-2 office robot (Asoh et al. 2001) or two “ears” used on the SIG active head by Nakadai et al. (2003)). How much of the problem is solved by such systems remains to be seen. Biological systems are also able to track an individual voice from its features, and ultimately hold the solution to noisy speech recognition. Until then, speech-enabled devices will require the user to wear a microphone. In practice, this eliminates all applications where an unknown user addresses a machine in a noisy environment.

6 Multimodal integration

This section is a brief reminder that verbal communication alone is insufficient for HRI. Natural language is a powerful tool for expression rules and sequences of operations. However, it is less expressive for shapes, locations and movements. Natural spoken communication is usually supported by gestures such as pointing to an object or a direction. Many tasks cannot be explained and are best demonstrated. This has long been recognized and research in speech interfaces must be considered as a part of the wider area of multi-modal communication. Some good examples are the GRAVIS system developed in Bielefeld (Steil et al., 2004), and systems developed by Imai et al. (1999) and Ono et al., (2001).

Given the functional consequences of accepting unconstrained spoken input (noted above), it may be interesting to investigate a corpus-based approach to unconstrained multimodal input. This should be done in the context of the instruction of tasks relevant for future users. It is possible that new aspects of verbal communication and its interaction with other forms of communication would then be highlighted.

7 Conclusion

For a robot to understand everyday language, it also needs to be able to execute tasks referred to in everyday language. At present, the problem of designing smart sensory-motor functions is much more difficult than speech recognition. How to recognize a dirty window, a wet piece of cloth? Realizing such difficult tasks could benefit from biological inspiration, especially in the area of vision.

Dialogues are full of misunderstandings and the ability to overcome these makes human-human communication so effective. In this respect, human-robot communication is very poor. A large number of problems remain to be solved, such as error detection, error repair, learning new words and actions, informative dialogues, etc. Such research is very much guided by findings and methods in psychology.

The human auditory system shows capabilities of filtering out background noise and can adapt to the speakers pitch and accent. Speech recognition systems do not process effectively voices in noisy environments or with unusual characteristics. Here, findings in the area of neuroscience of sensory systems could accelerate the solution of these problems.

Overall, speech interfaces require a high level of functional competence from the robot, as humans refer to high-level functions in their everyday language. What these functions should be is still speculative for most applications. The handling of misunderstandings requires from robots a high level of cognitive competence, mimicking many characteristics of human listeners.

Acknowledgements

The author is grateful for comments and references provided by anonymous referees.

References

Asoh H., Motmura Y., Asano F., Hara I., Hayamizu S., Itou K., Kurita T., Matsui T., Vlassis N., Bunschoten R., Krose B. (2001) Jijo-2: An office robot that communicates and learns. *IEEE Intelligent Systems*, 16:5, 46-55.

Bugmann G., Stanislao Lauria, Theocharis Kyriacou, Ewan Klein, Johan Bos, Kenny Coventry (2001) Using Verbal Instructions for Route Learning: Instruction Analysis . *Proc. TIMR 01 – Towards Intelligent Mobile Robots, Manchester 2001. Technical Report Series, Department of Computer Science, Manchester University*, ISSN 1361 – 6161. Report number UMC-01-4-1

Bugmann G., Klein E., Lauria S. and Kyriacou T. (2004) Corpus-Based Robotics: A Route Instruction Example. *Proceedings of IAS-8*, 10-13 March 2004, Amsterdam, pp. 96-103.

Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends in Cognitive Sciences. 8 (1): 8-11.

Hockey B.A., Lemon O., Campana E., Hiatt L., Aist G., Hieronymus J., Gruenstein A. and Dowding J. (2003) Targeted help for spoken dialogue systems: Intelligent feedback improves naïve user's performance. *Proc. 10th Conf. of the European Chapter of the Association for Computational Linguistic* (EACL'03), Budapest, Hungary.

Imai M., Kazuo Hiraki, Tsutomu Miyasato. Physical Constraints on Human Roboto Interaction, *Proceedings of 16th International Joint Conference on Artificial Intelligence* (IJCAI99), PP.1124--1130 (1999).

Kyriacou T., Bugmann G. and E., Lauria S. (2005) Vision-Based Urban Navigation Procedures for Verbally Instructed Robots. To appear in *Robotics and Autonomous Systems*

Lauria S., Kyriacou T. Bugmann G., Bos J and Klein E. (2002) Converting Natural Language Route Instructions into Robot-Executable Procedures. *Proceedings of the 2002 IEEE Int. Workshop on Robot and Human Interactive Communication* (Roman'02), Berlin, Germany, pp. 223-228.

Nakadai K, Hiroshi G. Okuno, Hiroaki Kitano: Robot Recognizes Three Simultaneous Speech by Active Audition. *Proceedings of IEEE-RAS International Conference on Robots and Automation* (ICRA-2003), 398-403, IEEE, Sep. 2003

Ono T., Michita Imai, Hiroshi Ishiguro (2001). A Model of Embodied Communications with Gestures between Humans and Robots. *Proceedings of Twenty-third Annual Meeting of the Cognitive Science Society* (CogSci2001), pp. 732--737.

Steil J.J, Rothling F., Haschke R, and Ritter H. (2004) Situated Robot Learning for Multi-modal Instruction and Imitation of Grasping. Robotics and Autonomous Systems, 47, 129-141.