R3D3: the Rolling Receptionist Robot with Double Dutch Dialogue

Jeroen Linssen
University of Twente
PO Box 217 7500 AE
Enschede, the Netherlands
j.m.linssen@utwente.nl

Mariët Theune
University of Twente
PO Box 217 7500 AE
Enschede, the Netherlands
m.theune@utwente.nl

ABSTRACT

We discuss the design of R3D3, a rolling receptionist robot with the ability to conduct 'double Dutch dialogues': dialogues (in Dutch) that involve, besides a human user, both a robot and a virtual human. R3D3 is intended to assist people when they visit shops, museums, or other establishments by acting as a host or receptionist. In the R3D3 project, we investigate how users can interact with the robot through natural language and nonverbal behaviour. R3D3 uses computer vision to determine user characteristics and combines this with automated speech recognition to analyse users' intentions. The robot and virtual human complement each other's affordances for verbal and nonverbal behaviour. Our preliminary studies have shown that 'double Dutch dialogues' require careful attention management through verbal and nonverbal behaviours by the robot and the virtual human, and that a receptionist robot should take the initiative in conversations to fulfil its intentions.

1. INTRODUCTION

Current human-robot interactions rarely involve the use of natural language for communication [3]. In the R3D3 project, we strive to create a robot with a specific focus on this modality for interaction with human visitors of museum, shops, and governmental buildings. R3D3 is a mobile (rolling) robot with receptionist functionality, that conducts 'double Dutch dialogues' with visitors through a duo consisting of a robot and a virtual human. The robot is capable of limited verbal and nonverbal behaviour, while the virtual human has richer conversational behaviour.

R3D3 should understand what people are saying (*content*) and why they are saying it (*intention*). This is a challenge in the fields of both human-robot interaction and human-virtual agent interaction [5]. In R3D3 these fields come together. We investigate the three research areas involved in such interactions: the sensing of (audiovisual) information, the processing of this information for decision-making, and the realization of (embodied) behaviour.

Mavridis discusses ten desiderata for conversational robots [3], of which we specifically address the following three in our research. (1) R3D3 should be able to understand sentences and intentions that are more complex than clear directives (breaking the 'simple commands only' barrier). (2) R3D3 should be able to take initiative in conversations, but should also allow users to take initiative (mixed-initiative dialogue). (3) R3D3 should express itself by synchronizing its verbal and nonverbal behaviour, for example, gazing at users it addresses verbally (motor correlates of speech and non-verbal communication). We believe that these desiderata can be met by combining a physical robot and a virtual human, which can complement each other's strengths and weaknesses [2]. The robot's physicality allows it to approach and guide people; the virtual human's extended range of expressions and verbal behaviour allow for more elaborate dialogues. However, this approach implies new challenges, such as more complex turn-taking behaviour than in bilateral interactions [5]. In the following section, we address how we intend to overcome these challenges.

2. DESIGN OF R3D3

R3D3 combines a virtual human (*Leeloo*) and a social robot (*EyePi*) in one mobile manifestation. Figure 1a shows a conceptual sketch of R3D3; Figure 1b shows the current prototype. The two entities each have their own role in the dialogues. EyePi's physical embodiment allows it to attract the attention of users and physically guide them to points of interest. Because EyePi's nonverbal and verbal behaviour are limited, R3D3 uses Leeloo's richer verbal and nonverbal behaviour to complement EyePi's behaviour.

The software architecture of R3D3 is shown in Figure 1c. It is implemented in ASAP, a platform for social agents that is suitable for controlling the behaviour of both robots and virtual humans [4]. R3D3 accepts audiovisual input through modules for automated speech recognition and computer vision. Language models for the speech recognition software are trained through deep learning, using *Kaldi*.[1] The computer vision module can detect users' demographics and emotions.[2] The dialogue manager uses these two information streams to store user models for all recognized users. Using these models, it determines whether someone is speaking and whether R3D3 itself should take initiative to speak. The dialogue manager also estimates the intentions of users, for example, whether they want directions to a point of interest or information about the current exhibition in a museum.

HRI '17 Companion March 06-09, 2017, Vienna, Austria
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4885-0/17/03.
DOI: http://dx.doi.org/10.1145/3029798.3038372

[1] See http://kaldi-asr.org.

[2] See http://www.vicarvision.nl/facereader/.

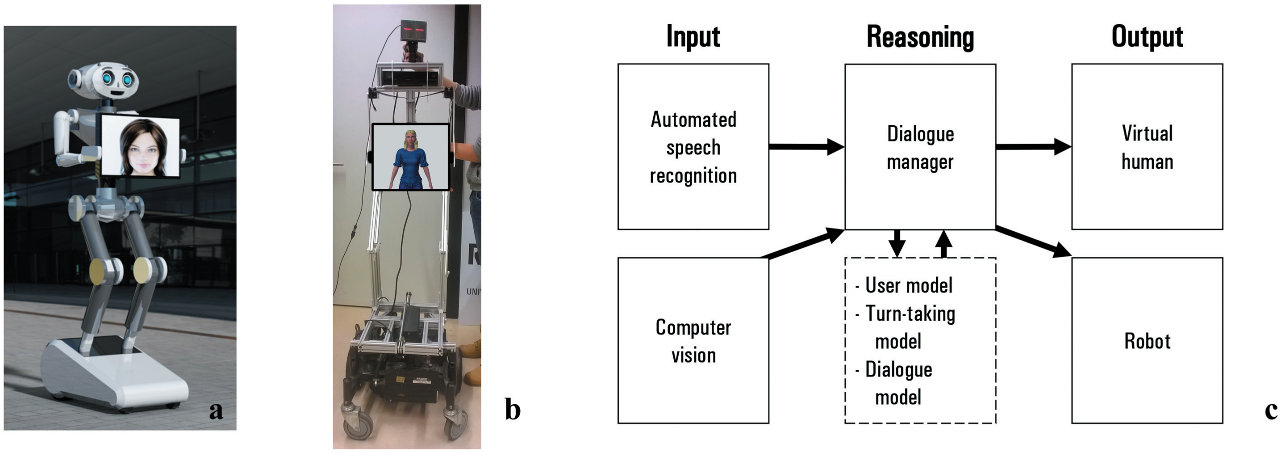

Figure 1: (a) A conceptual sketch of R3D3. (b) The current prototype, including a sketch of the envisioned placement of the tablet. (c) The architecture connecting all the software modules of R3D3.

Additionally, the dialogue manager keeps track of the phase of the dialogue, such as introduction, establishing intentions and navigation. Based on the current dialogue phase, the dialogue state (who is speaking), and the user's intention, the dialogue manager decides which tasks should be assigned to either EyePi or Leeloo. Certain tasks can be carried out in tandem. For example, when Leeloo explains to the user where a certain object is located in the physical environment, she can ask EyePi to point it out through gaze. By addressing the other entity, Leeloo attempts to shift a user's attention to EyePi, ensuring that he or she is not confused by the dialogue with two conversational partners.

3. FIRST STUDIES

We have investigated user interactions with several early prototypes of R3D3 at different venues. At a language festival, we used a first prototype of R3D3, consisting of Leeloo and EyePi's movable head, to investigate how people would interact with the two entities. EyePi supported Leeloo by nodding in agreement when she explained certain things, or by gazing at a point of attention Leeloo pointed at. The robot head attracted much attention, but its behaviour was limited to tilting and emoting. This caused people to pay attention mostly to Leeloo, whom they could actually talk to. From this, we conclude that there needs to be more explicit cooperation between the two synthetic entities to make sure the users' attention is appropriately divided between them.

At the Dutch Police Academy, we demonstrated a second prototype with the fully assembled robot. People could ask R3D3 questions in natural language, which were translated and matched to answers in R3D3's database. We constructed the possible dialogues in a way that Leeloo explicitly asked about which topic someone would want to hear, listing several options. This guidance assisted users in choosing what to say, resulting in fairly fluent conversations, though still limited in length and depth. EyePi assisted Leeloo by expressing matching emotions. When someone's utterance wasn't completely understood, Leeloo asked the user to repeat herself. On some occasions, the user would simply repeat the keyword central to her request. This is an indication that people are aware of R3D3's limited understanding and are willing to cope with it.

4. NEXT STEPS

Our pilot studies confirm that attention management in multi-party interactions is an important challenge. We have already taken the first steps toward modelling engagement in multi-party interactions, similar to [1]. We created a model to divide a robot's attention between multiple users by assigning priorities to each of them, based on their behaviour. We will investigate this approach in upcoming user studies.

The following step in this project is expanding the possible behaviours for both the robot and the virtual human, improving their synchronicity. Furthermore, we aim at widening the range of recognizable user intentions, matching them to each of the contexts R3D3 will be placed in. Finally, R3D3 should be adaptable to even more to unexpected situations it may encounter. We will investigate this in upcoming evaluations. Automatically keeping track of successful and failed interactions will provide valuable feedback to further iterate R3D3's design and expand its behaviour.

5. ACKNOWLEDGMENTS

This publication was supported by the Dutch national program COMMIT.

6. REFERENCES

[1] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *SIGDIAL '09*, pages 225–234, 2009.

[2] J. Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015.

[3] N. Mavridis. A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, 63(P1):22–35, 2015.

[4] H. van Welbergen, R. Yaghoubzadeh, and S. Kopp. AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs. In *IVA '14*, pages 449–462, 2014.

[5] Z. Yumak, J. Ren, N. M. Thalmann, and J. Yuan. Tracking and fusion for multiparty interaction with a virtual character and a social robot. In *SIGGRAPH ASIA '14*, 2014.