R3D3 in the Wild: Using A Robot for Turn Management in Multi-Party Interaction with a Virtual Human

Mariët Theune, Daan Wiltenburg, Max Bode, and Jeroen Linssen
Human Media Interaction
University of Twente
Enschede, The Netherlands
Email: m.theune@utwente.nl

***Abstract*—R3D3 is a combination of a virtual human with a non-speaking robot capable of head gestures and emotive gaze behaviour. We use the robot to implement various turn management functions for use in multi-party interaction with R3D3, and present the results of a field study investigating their effects on interactions with groups of children.**

I. Introduction

The Rolling Receptionist Robot with Double Dutch Dialogue (R3D3) is a social robot consisting of two agents: a robot (EyePi) and a virtual human (Leeloo), carried by the robot on a tablet as shown in Figure 1. R3D3 is intended to serve as an assistant to visitors of public places, interacting with people using natural language and nonverbal behaviour. While Leeloo is capable of simple spoken conversations, EyePi can only show nonverbal behaviour by moving its head and showing emotions with its eyes. Robot gaze and emotion have been shown to be effective for attracting people and engaging them in a conversation [2]. However, our first tests in public venues with an early prototype of R3D3 showed that, after being initially attracted to EyePi, during the subsequent conversation with Leelo users focused almost exclusively on the virtual human [5].

In previous research we experimented with giving EyePi a turn allocation role in the conversation, by selecting the addressee of a question with its gaze [4]. Although successful, this previous study was limited to a controlled 'Wizard of Oz' experiment involving a small number of subjects. In the current study, we extended the robot's turn management role with a number of additional behaviours. It now encompasses bystander acknowledgement, turn confirmation, turn allocation, and interruption management. Furthermore, instead of a controlled experiment we performed a field test with groups of children at the NEMO Science Museum in Amsterdam, with R3D3 functioning autonomously instead of being controlled by a wizard.

Below, first we briefly discuss related work that formed an inspiration for this research. Then we describe R3D3 and the turn management behaviours we designed for it. We present our findings from a pilot test and from the field study, ending with conclusions and future work.

Fig. 1. R3D3: Rolling Receptionist Robot with Double Dutch Dialogue.

II. Related Work

Turn-taking, in interactions between humans [9] and more recently, between humans and both virtual characters [1] and social robots [3], is seen as an important factor in managing fluent interactions. Engaging people to make them assume certain roles can be done with the right cues [1]. A key behaviour in doing so is the intentional direction of gaze [8], which, in human-robot interaction, has been shown to be a highly effective turn-taking mechanism [6]. Gaze is of even higher importance when interacting with multiple people, as group dynamics play a role there. Especially in crowded places, interactions between robots and multiple users can benefit from turn management through gaze, both for enabling

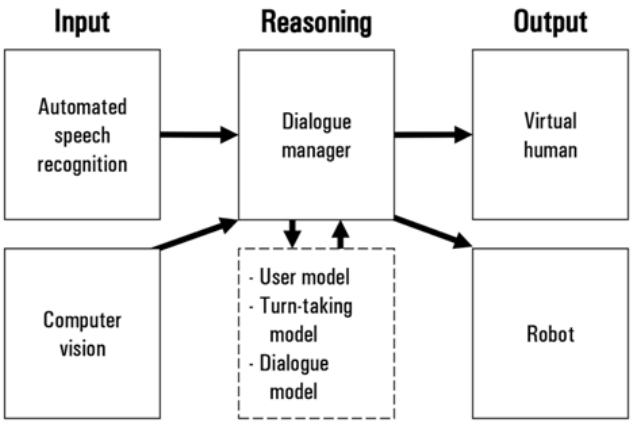

Fig. 2. R3D3 architecture.

a robot to express its intentions and for controlling users' attention [3], [10]. Previous work has also looked at the dynamics of interaction with multiple virtual humans [7] and multiple users [11]. In the current paper, we investigate multi-party interaction with two synthetic entities, a virtual human and a social robot. Such a setup has been previously explored in [12], who studied how such a pair could, together, engage multiple users.

III. R3D3 Architecture

In our previous studies, R3D3 was controlled manually with a Wizard of Oz method [4]. In contrast, in our current research R3D3 was designed to interact with humans autonomously. R3D3 includes several sensory and communication modules that enable autonomous interaction; see Figure 2. The computer vision module can detect humans in its field of view. It can make an estimation of the interlocutors' age, gender, emotion, and detects when he or she is speaking. The R3D3 speaker detector is implemented by a linear support vector machine, which takes as input features derived from the facial action units classified by a deep network. The microphones of R3D3's KINECT camera are used to detect the direction of the speaker. Furthermore, R3D3 has a speech recognition module for Dutch based on deep learning, using Kaldi.[1] The recognized speech is processed by the dialog manager module, which matches the user's input to the most similar utterance in a list of utterance-response pairs. The corresponding response is returned and spoken by the Virtual Human using a text-to-speech module.[2]

IV. Turn Management in R3D3

For our field study, we added the following turn management functions to be carried out by the robot, EyePi.

Acknowledgement (happy gaze). When the presence of a person is detected, an acknowledgement sequence is initiated. EyePi briefly directs a happy gaze (see Figure 3) to the new user, and Leeloo introduces R3D3. If another person comes into view, EyePi acknowledges this user with a happy gaze as well. Currently, the system can detect and remember up to three people; additional newcomers are ignored.

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

[2]http://www.fluency.nl

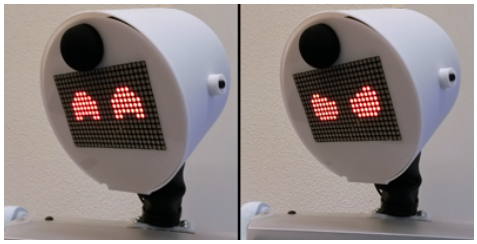

Fig. 3. EyePi expressing happiness (left) and anger (right).

Turn confirmation (neutral gaze). The first user who starts talking is stored (temporarily) as the first speaker, and regarded as the 'owner' of the current turn. When the first speaker is detected, the robot gazes toward this user to confirm his or her turn ownership. The turn owner keeps the turn until he or she finishes speaking. Then, after a between-speaker interval of 1 second (to avoid accidental interruptions), the virtual human responds and the robot turns its gaze towards the virtual human as the new turn owner. Whenever a new person enters the field of view while the turn owner (either a user or the virtual human) is speaking, the robot briefly acknowledges them as described above, assigning them the role of bystander [6].

Turn allocation (neutral gaze). After the virtual human finishes talking, the robot gazes toward a randomly picked person within its field of vision, to give this user the next turn. This form of turn allocation had been tested earlier in a small-scale lab experiment under controlled conditions [4] and was shown to be most effective after the robot had been explicitly addressed by the virtual human.

Interruption management (angry gaze). When several users are speaking, only the first speaker is regarded as having the turn. If another user is detected speaking at the same time as the current turn owner, the interruption management function is initiated and the robot briefly directs an angry gaze towards the interrupting user (see Figure 3). After that, it returns its gaze to the current turn owner.

V. Pilot test

To test R3D3's new turn management functions, a pilot test was conducted at a day care centre. The goal of the test was to see how the new autonomous behaviour of R3D3 functioned with groups of children, and how the children would react to R3D3. Only turn allocation was not tested, as it had not yet been implemented at the time of the pilot test.

In the pilot, five groups of three children, aged between 5-10 years old, were invited to interact with R3D3. When any of the children approached R3D3, Leeloo would introduce herself and EyePi as R3D3, and ask if the children wanted to know something about robots. Then, the children were free to ask anything to R3D3.

Our main finding was that the speech recognition module, trained on data from adults, did not work at all for children. It could not recognize any words spoken by the children, which made a true conversation impossible. The computer vision, similarly trained on adults only, also had some difficulties with the children. In some cases it could not detect all the children in view and the estimated age was off every time. Still, it was able to pick up enough to detect when a child was talking and what the children's locations were.

In spite of the occasional computer vision problems, the acknowledgement behaviour worked moderately well. EyePi acknowledged at least two children in each session. The turn confirmation function seemed to work correctly as well. However, a clear effect could not be noticed because surprisingly, most children were very quiet throughout the session. If they did speak, it often was a short utterance of one to three words. Due to the lack of speech recognition, Leeloo did not speak much either. This caused long moments of silence and a lack of actual conversation. Nevertheless, interrupters were occasionally recognized, in which case EyePi gazed angrily at this person. This made a big impression on the children, but the reason for EyePi's anger was not clear to them.

Based on the pilot, we redesigned the dialogues for R3D3 to function without speech recognition. This was done by having Leelo take the initiative in the dialogue. When the first speaker finishes speaking, Leeloo responds with a randomly selected preprogrammed utterance. This can be an informative statement (e.g., some general information about robots), or a question ("Do you know...?"). After a question, Leeloo waits for a few seconds to give the users a chance to answer, before giving the corresponding follow-up explanation. The robot, EyePi, gazes at Leelo while she is speaking and randomly allocates the turn to one of the users after she is finished. The following is an example (translated from Dutch):

Question: For what would you like to use a robot?
(User response)
Follow-up explanation: There are robots of many different types. You have them in factories, but also for use at home. Do you know for example the automatic lawn mower or vacuum cleaner? Those are robots too.

VI. Field test in the Science Museum

With the final version of R3D3 as described above, we carried out a field test in the NEMO Science Museum in Amsterdam. The main target group of NEMO are children between 6 to 16 years old. The visitors were allowed to approach R3D3 and interact with it as they pleased. The interactions were video recorded and two observers were present (behind R3D3). The observers interfered in the interactions only when errors occurred or when people touched R3D3.

During this 'in the wild' experiment we tested three different conditions, in which different functions of the robot's gaze behavior were turned on:

1) Turn confirmation, turn allocation, acknowledgement and interruption management (all functions)
2) Turn confirmation and turn allocation
3) Turn confirmation, turn allocation and acknowledgement

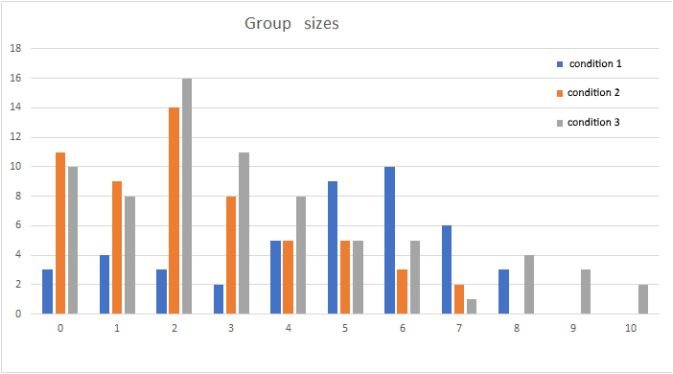

Fig. 4. Group sizes (x axis) and number of occurrences of each group size (y axis) per condition.

Due to R3D3's ability to respond to the user's speech (although without actual understanding), this time, in contrast to the pilot test, 'real' conversations occurred often. Below we briefly report on group sizes and interaction times, and provide some general observations concerning the functioning of R3D3's turn management.

A. Participants

Figure 4 shows the sizes of the groups interacting with R3D3 in the three different conditions. Intervals when no one was interacting with R3D3 are counted as occurrences of group size zero. Groups of two were overall the most frequent; this was also the case in conditions 2 and 3. In condition 1, groups of six were the most frequent. In condition 3, the biggest group interacted with R3D3: in this condition, R3D3 interacted with a group of ten people twice. In condition 1, the biggest group consisted of eight people. In condition 2, the biggest group consisted of seven people. Group sizes were not fixed, as children could join or leave the conversation while R3D3 was interacting with a group.

Although the age of the participants was not measured, their estimated age was between eight and fourteen years old.

B. Turn Management

The robot's turn confirmation behaviour worked well for the virtual human; the children often followed the robot's gaze toward Leelo when she was talking. Most children were quiet or stopped talking during Leelo's turn. However, for human users, R3D3 had problems detecting which person was talking. One reason was that the computer vision does not work very well for children. Another reason was the overall noise in the NEMO environment. The system could detect if someone was speaking, but with sound coming from different directions it could not distinguish correctly who was speaking. Since the function of gazing towards the first speaker relied on this specific data, the function simply did not work.

Fortunately, the turn allocation function could be carried out whenever a person was in view. Nevertheless, it was only

successful (in the sense that the intended user took the turn) in 27% of the cases. This is considerably lower than the results we obtained with adults in a lab environment [4]. This may be partly due to the fact that other than in our earlier experiment, in the current design of the dialogues the virtual human did not explicitly draw attention to the robot. As a consequence, participants may have missed the robot's turn allocation cues. The relatively large group sizes in the field test are another factor. With a success ratio of 33%, turn allocation was most successful with groups consisting of two persons. Then, with each increase in the number of interlocutors, the success/fail ratio decreased, with a lowest success rate of 17% with seven interlocutors. (With only three attempts in total, we did not consider the data of groups with eight or higher to be representative, so these are disregarded here.)

As with the turn confirmation function, interruption management suffered from the faulty speaker detection. Only seven interruption behaviours were performed without error, that is, when an actual interruption occurred. This is not enough to draw any conclusions on the effectiveness of the function when performed correctly. Still, the function had great effect on the users. Each time the robot directed an angry gaze at someone (which happened many times due to an error), the children reacted surprised, disapproving or even shocked. It seemed that each time the user in question immediately understood that he or she did something wrong according to the robot, but not what it was exactly. Furthermore, it seemed as if the function made the interaction with R3D3 more interesting. This is supported by the interaction times reported below.

C. Interaction times

In condition 1, the mean interaction time per person was 2 minutes and 11 seconds (SD: 01:19.9). In condition 2, the mean was 00:51.2 (SD: 00:50.1), while in condition 3 the mean was 00:46.1 (SD: 00:19.5). These results suggest that R3D3 was most interesting for the visitors in condition 1, and least interesting in condition 2. In the latter condition, six of the interactions lasted less than 30 seconds.

It seems that the robot's behaviour played an important role in the overall experience when interacting with R3D3. In condition 1, the robot's functions were all activated, meaning that the robot was very active and showed a range of emotions. The interaction time is longer both in mean and in total in this condition. In contrast, in condition 2 all functions except turn confirmation and allocation were turned off, meaning that the robot was not very active and showed no emotions. These findings are in line with [2], who found that "having an expressive face and indicating attention with movement both make a robot more compelling to interact with" ([2], p. 4142).

VII. Conclusion

The aim of this study was to create an autonomous turn allocation system for R3D3 for interaction with groups of children. A pilot test showed that R3D3's speech recognition is not (yet) usable for group conversations with children. In a subsequent field test in a science museum three behavioural conditions were compared. Although not all behaviours functioned without errors or were understood by the interlocutors, the results suggest that having EyePi show varying gaze behaviour and emotions made R3D3 more interesting as a whole. When EyePi looked angry because an interruption was detected, most visitors showed a strong reaction, even though they did not understand the reason behind the robot's anger.

During the experiment, groups of different sizes interacted with R3D3. While R3D3 was interacting with a group, other children could join or leave the conversation. The rapid alternation of group sizes made it difficult for the computer vision to keep track of interlocutors. In future work a more effective way to keep track of interlocutors should be developed.

During the field study a large amount of video data has been gathered. In future work, this data should be analysed more thoroughly to investigate the effects of turn management.

Acknowledgements

This publication was supported by the Dutch national program COMMIT. We thank NEMO for hosting our field test. We also thank Marten den Uyl for being the initiator of the R3D3 project and for coming up with the R3D3 concept. Sadly, he passed away during the project.

References

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

[2] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2002, pp. 4138–4142.

[3] I. Leite, H. Hajishirzi, S. Andrist, and J. Lehman, "Managing chaos: Models of turn-taking in character-multichild interactions," *Proceedings of ICMI '13*, pp. 43–50, 2013.

[4] J. Linssen, M. Berkhoff, M. Bode, E. Rens, M. Theune, and D. Wiltenburg, "You can leave your head on - attention management and turn-taking in multi-party interaction with a virtual human/robot duo," in *Proceedings of Intelligent Virtual Agents (IVA 2017)*, 2017.

[5] J. Linssen and M. Theune, "R3D3: the Rolling Receptionist Robot with Double Dutch Dialogue," in *Proceedings of the Companion of HRI '17*, 2017, pp. 189–190.

[6] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 1–33, 2012.

[7] R. Nishimura, Y. Todo, K. Yamamoto, and S. Nakagawa, "Chat-like spoken dialog system for a multi-party dialog incorporating two agents and a user," in *Proceedings of iHAI '13*, 2013.

[8] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception," *Computer Graphics Forum*, vol. 34, no. 6, pp. 299–326, 2015.

[9] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[10] M. Shiomi, T. Kanda, S. Koizumi, H. Ishiguro, and N. Hagita, "Group attention control for communication robots with wizard of oz approach," in *Proceedings of HRI '07*, 2007, p. 121.

[11] G. Skantze, "Predicting and regulating participation equality in human-robot conversations," in *Proceedings of HRI '17*, 2017, pp. 196–204.

[12] 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 *Proceedings of SIGGRAPH Asia '14*, 2014, pp. 1–7.