Evaluating a social multi-user interaction model using a Nao robot

Simon Keizer[1], Pantelis Kastoris[1], Mary Ellen Foster[1], Amol Deshmukh[1] and Oliver Lemon[1]

Abstract— **This paper presents results from a user evaluation of a robot bartender system, which supports social engagement and interaction with multiple customers. The system is a Nao-based alternative version of an existing robot bartender developed in the JAMES project [1]. The Nao-based version has given us a local experimentation platform, allowing us to focus on social multi-user interaction rather than the robot technology of object manipulation. We will describe the design of the Nao-based system and discuss the differences with the original JAMES system. In a recent evaluation of the JAMES system with real users, a trained and a hand-coded version of the action selection policy were compared [2]. Here we present results from a similar comparative user evaluation on the Nao-based system, which confirm the conclusions of the previous experiment and provide further evidence in favour of the trained action selection mechanism. Task success was found to be almost 20% higher with the trained policy, with interaction times being about 10% shorter. Participants also rated the trained system as significantly more natural, more understanding, and better at providing appropriate attention.**

I. Introduction

The use of service robots in the home as well as in public spaces has become increasingly viable over the last decade. The development of effective and robust models for social multi-user human robot interaction is continuing to be vital to this development. This paper builds on previous work on using machine learning techniques in this area, applied to the example of a robot bartender. This bartender should not only be *task effective*, i.e., taking orders from customers and serving the drinks they ordered, but also exhibit *socially appropriate* behaviour, e.g., serving multiple customers in the appropriate order, and following other social conventions such as greeting and responding to a customer's "thank you" with "you're welcome", making interactions with the robot more acceptable and pleasant for customers.

With the purpose of testing and evaluating such models for social interaction locally on a regular basis, the bartender system developed in the JAMES project[1] was ported to a modified robot platform comprising a Nao torso robot and a single Microsoft Kinect for vision. This modified robot cannot track the full range of the users' non-verbal behaviour and is not physically able to serve drinks; however, it provides a useful platform for experiments specifically addressing aspects of social multi-user interaction.

Using this Nao-based robot bartender, we carried out a user evaluation similar to the one described in [2], focusing on comparing a hand-coded and trained version of the action selection component of the system. Based on the current social state, which contains relevant higher level information such as which customers are present, whether a customer is seeking attention, or what they want to order, this action selection component decides what action (communicative and/or non-communicative) the robot should produce next.

[1]Interaction Lab, Heriot-Watt University, Edinburgh, UK s.keizer@hw.ac.uk

[1]EU FP7 project JAMES: james-project.eu

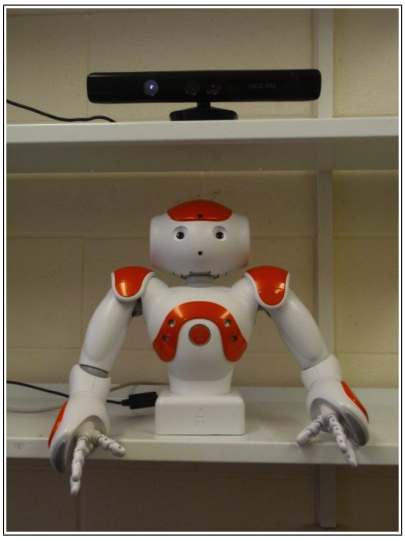

Fig. 1. Nao/Kinect robot bartender

The paper is organised as follows. In Section II we describe the Nao-based robot bartender system and contrast it with the original JAMES system. In Section III we describe in more detail the action selection component on which the evaluation is focused, called the Social Skills Executor (SSE). The evaluation itself is described in Section IV, followed in Section V by the results and discussion. The paper is concluded in Section VII.

II. Robot bartender system

The JAMES robot bartender [1] is equipped with modules for vision and speech processing, along with modules controlling the robot behaviour. The robot behaviour is realised in the form of both speech and head/arm gestures. Based on observations about the users in the scene, the system maintains a model of the social context, and decides on effective and socially appropriate responses in that context. The system thus aims to engage in, maintain, and close interactions with users, take a user's order via spoken conversation, and serve their drinks. For the Nao-based bartender, we implemented versions of the vision system and the robot behaviour controller making use of a Nao torso robot and a single Microsoft Kinect as shown in Figure 1.

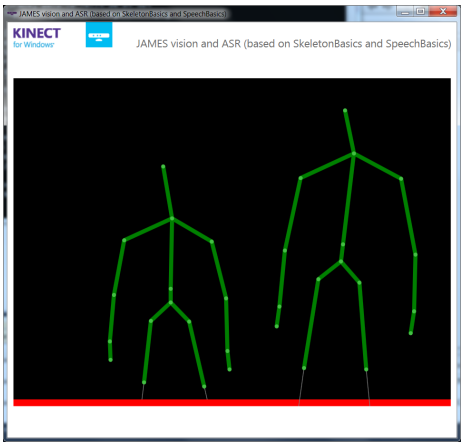

Fig. 2. Kinect-based vision system

A. Vision module

The full JAMES computer vision system [3] tracks the location, facial expressions, gaze behaviour, and body language of all people in the scene in real time, integrating signals from multiple sensors including two calibrated stereo cameras and a Microsoft Kinect depth sensor. For the current study, we have developed a vision system that uses only a single Kinect sensor to track the location and torso orientation of all customers, using the built-in skeleton tracking provided by the Kinect for Windows SDK [4] as shown in Figure 2. Although the features tracked by the Kinect-based vision system are a subset of those handled by the full system, the information it provides is still sufficient for the Social State Recogniser (SSR) [5] to estimate the social state of all customers for use in selecting response actions as described in Section III.

B. Realisation of robot actions

As shown in Figure 1, the Nao robot hardware consists of two 5-degrees-of-freedom arms with hands, along with a head with 2-degrees-of-freedom for pan and tilt. The robot head is equipped with multi colour LED lights, along with two speakers in its ears for producing synthesised speech, and a camera which can capture 30 images/second. The Nao torso has an embedded computer provided with an API developed in Python [6] for programming different gestures and for data acquisition from sensors. A set of Nao robot behaviours was developed for the bartender domain, based on those supported by the full JAMES robot. These behaviours included both verbal actions and gestures. The verbal actions were realised using the Nao's built-in text-to-speech (TTS) facility, while the gestures were realised using motion from hands and the robot head. Table I lists the behaviours supported by the Nao in this domain; note that, for practical reasons, the Nao serves drinks "symbolically" through the *ServeDrink* gesture in which the robot grasps and hands over an imaginary drink.

III. Social Skills Execution

The Social Skills Executor (SSE) controls the behaviour of the robot bartender by selecting both communicative and non-communicative robot actions, based on the social state updates it receives from the SSR. In the bartender domain, the non-communicative actions typically involve serving a specific drink to a specific user, whereas the communicative actions have the form of dialogue acts [7], directed at a specific user, e.g., `setQuestion(drink)` ("What would you like to drink?") or `initialGreeting()` ("Hello"). Full details of the SSE are presented in [2]; we summarise the main points here.

In our design of the SSE, the decision making process that leads to the actions for the robot to realise (or the decision to do nothing) consists of three stages: 1) ***social multi-user coordination***: managing the system's engagement with the users present in the scene (e.g., accepting a user's bid for attention, or proceeding with an already engaged user), 2) ***single-user interaction***: if proceeding with an already engaged user in the social multi-user coordination stage, generating a high-level response to that user, in the form of a communicative act or physical action (e.g., greeting the user or serving him a drink), and 3) ***multi-modal fission***: selecting a combination of modalities for realising a response selected in the single-user interaction stage (e.g., a greeting can be realised through speech and/or a nodding gesture), using the Nao behaviours listed in Table I.

For the multi-user coordination and single-user interaction stages, the decision making process happens through a combination of two Markov Decision Process (MDP) models, which can be trained using reinforcement learning in interaction with a Multi-User Simulated Environment (MUSE) developed for this purpose [2]. MUSE allows for rapidly exploring the large space of possible states in which the SSE must select actions. A reward function incorporating individual rewards from all simulated users in the environment is used to encode preferred system behaviour in a principled way. A simulated user assigns a reward if they are served the correct drink, and gives penalties corresponding to their waiting time and various other forms of undesired system behaviour.

The architecture for interactions in simulation using MUSE includes both SSE and SSR. MUSE produces inputs for the SSR: a) a vision input stream containing information about the visible users, including their location and gaze direction; b) speech events in the form of user dialogue acts; c) rewards provided by the simulated users; and d) feedback about the execution of robot actions. MUSE also processes the output of the SSE to simulate action execution: the start of the action is signalled to the SSR, and when an action is completed, it is made available for processing by the simulated users. In this way, robot actions can be given a duration in the simulated environment (in terms of a number of simulated time frames), and MUSE can thus produce an input stream for the SSR, whereas the SSE processes input and generates output on the basis of events.

TABLE I

NAO BEHAVIOURS FOR THE BARTENDER DOMAIN

Behaviour	Description
Say	Nao speaks the given text with default TTS
LookAt	Nao looks at a given location position in space
Nod	Nao nods or shakes its head
GreetExpression	Nao greets the customer by waving a hand
ServeDrink	Nao makes a grabbing action and brings the hand to a serving position
SmileExpression	Nao's eyes change colour, and its head and body move to show joy

IV. USER EVALUATION

For the user evaluation of the Nao based robot bartender, 48 subjects—university students with varying backgrounds—were recruited and asked to interact with the system, resulting in a total of 96 two-user interactions. Each pair of subjects interacted four times with the system, two times whilst running the hand-coded SSE (labelled SSE-HDC), and two times whilst running the trained SSE (SSE-TRA). To cancel out any bias due to learning effects, the order of these four interactions was varied between subject pairs. Before the four interactions, both users of each pair determined which of the three possible drink types (coke, blue lemonade, or green lemonade) they were going to order. Note that the SSE policies compared in this study were identical to those used in the previous study described in [2].

A. Godspeed evaluation

As a way to evaluate the overall impression of the Nao-based system, the subjects were asked to fill out a questionnaire based on the *Godspeed* questionnaire series [8], both before and after being exposed to the system. The pre-experiment questionnaire was to give us an insight into the users' expectations about the system, and to what extent these expectations were met. The full Godspeed questionnaire consists of five sets of questions, but in the interest of time we limited that to the two categories that we considered to be the most relevant for our purposes: *likeability* and *perceived intelligence* (Figure 3).

Godspeed III: Likeability

Please rate your impression of the robot on these scales:

Dislike	1	2	3	4	5	*Like*
Unfriendly	1	2	3	4	5	*Friendly*
Unkind	1	2	3	4	5	*Kind*
Unpleasant	1	2	3	4	5	*Pleasant*
Awful	1	2	3	4	5	*Nice*

Godspeed IV: Perceived Intelligence

Please rate your impression of the robot on these scales:

Incompetent	1	2	3	4	5	*Competent*
Ignorant	1	2	3	4	5	*Knowledgeable*
Irresponsible	1	2	3	4	5	*Responsible*
Unintelligent	1	2	3	4	5	*Intelligent*
Foolish	1	2	3	4	5	*Sensible*

Fig. 3. Godspeed questionnaire sections used for evaluation. Note that for the pre-experiment test the questions were formulated to ask about the users' expectation about the robot, rather than their impression.

Q1: What did you try to order? [coke/blue lemonade/green lemonade]

Q2: Did you successfully order a drink from the bartender? [Y/N]

Please state your opinion on the following statements:
[1:strongly disagree; 2:disagree; 3:slightly disagree; 4:slightly agree; 5:agree; 6:strongly agree]

Q3: It was easy to attract the bartender's attention [1–6]

Q4: The bartender understood me well [1–6]

Q5: The interaction with the bartender felt natural [1–6]

Q6: Overall, I was happy about the interaction [1–6]

Fig. 4. User questionnaire after each interaction

B. Subjective evaluation metrics

In order to compare the two versions of the system in terms of subjective performance, every subject filled out a questionnaire after each of the four interactions, as shown in Figure 4. The questions were designed to measure perceived system performance in terms of task success, ease of seeking the robot's attention, ease of making the robot understand a drink order, and naturalness of the interaction.

C. Objective evaluation metrics

Besides the subjective evaluation we also analysed the system logs, resulting in a number of objective evaluation metrics. These metrics are averages for the following values for each user in each interaction:

- *Attention seeking time*: the time in seconds between the moment the (vision) system has detected the user and the moment the user is recognised as seeking attention by the social state recogniser;
- *Interaction time*: the time in seconds between the moment of detection by the vision system and either the moment a drink has been served to that user, or, if the user was not served, the moment the user leaves the scene (i.e., is no longer visible by the system);
- *Serving time*: the time in seconds between the moment of the user recognised as seeking attention and the moment the user has been served a drink (assuming the user has been served at all);
- *Number of SSE level 1 decisions*: the number of times the SSE multi-user coordination policy was triggered to make a decision;
- *Number of SSE level 2 decisions*: the number of times

the SSE single user interaction policy was triggered to make a decision;

- *Number of speech input events*: the number of times the Kinect speech processing module detected speech input;
- *Number of speaker identification failures*: the number of times the social state recogniser could not assign speech input to a known customer; and
- *Number of ASR failures*: the number of times speech input was discarded because the confidence score (provided by the Kinect speech recogniser) was below the threshold of 0.5.

V. Results and discussion

A. Godspeed questionnaire

The results from the Godspeed questionnaire pre and post test are shown in Figure 5. The responses on all questions decreased from the pre test to the post test: on the likeability questions, only a marginal decrease was observed, whereas the decrease in perceived intelligence was significant at $p < 0.05$ on a Wilcoxon signed rank test. It is likely that one reason for the decrease in perceived intelligence is the rather limited task domain currently supported by the system; on the other hand, the behaviour of the bartender system did not have as much effect on the users' impression in terms of likeability. Also, recall that this questionnaire is aimed to evaluate an overall impression of the robot system, whereas our focus in developing this Nao-based version of the robot was on approximating the basic capabilities of the original JAMES robot and on replicating the previous comparison between the two SSE versions. Note that a decrease in Godspeed scores was also found in two recent user evaluations of the full JAMES robot system [9], [10].

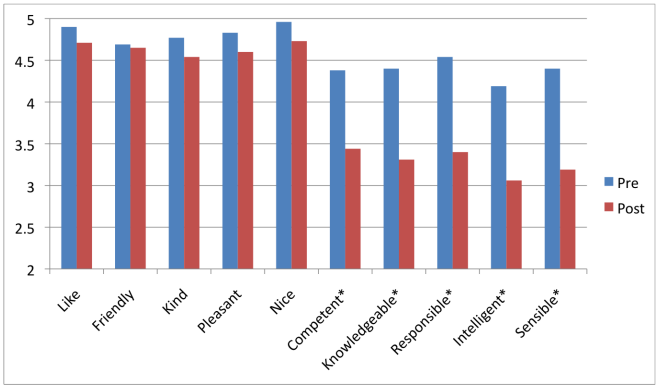

Fig. 5. Godspeed results (significant differences indicated with an asterisk)

B. Subjective evaluation

The results from the questionnaire in Figure 4 are given in Table II, in the form of a percentage (success rate) for question Q2, and average scores for questions Q3 to Q6. As indicated by the asterisks in the table, the system that used the trained version of the SSE (SSE-TRA) significantly outperformed the hand coded version (SSE-HDC) on all scores ($p < 0.05$ on a Wilcoxon signed rank test).

TABLE II

Perceived performance results for the hand-coded and trained SSE

	SSE-HDC	**SSE-TRA**
Q2 (task success) *	55.21%	75.00%
Q3 (attention) *	4.36	4.92
Q4 (understanding) *	3.53	4.52
Q5 (naturalness) *	3.35	4.04
Q6 (overall) *	4.13	4.64

TABLE III

Objective measures on the system logs of the collected interactions

	SSE-HDC	**SSE-TRA**
AvgAttTime	0.519	0.617
AvgIntTime *	45.119	40.492
AvgServTime	30.166	27.172
AvgNumDecs1	14.56	14.21
AvgNumDecs2	5.56	6.00
AvgNumDecs2a	2.83	3.79
SpeakerIdFailRate	44.63%	46.78%
AsrConfRejRate	56.69%	50.62%

C. Objective evaluation

The results of the objective analysis of the experiments are summarised in Table III, again with an asterisk indicating differences that were significant on a Wilcoxon signed-rank test ($p < 0.05$). The average time between detecting a user and recognising them to seek attention (AvgAttTime) was very short, and only marginally longer for the interactions with the system that used the trained SSE policy (SSE-TRA). This is according to expectations since subjects were asked to enter the scene and (immediately) approach the robot to order a drink, and the SSE component does not play a direct role in recognising users seeking attention. The average time of interactions (AvgIntTime) with the SSE-TRA system was significantly shorter than those with the SSE-HDC system, suggesting that the trained policy resulted in more efficient interactions. The average time it took to serve a user (AvgServTime) was also somewhat shorter for the SSE-TRA system. Part of this difference can be explained by the performance of the speech processing component: in the interactions with the SSE-HDC system, the speech rejection rate (i.e., rejections due to failed speaker identification, denoted by SpeakerIdFailRate, or due to the ASR confidence score being too low, denoted by AsrConfRejRate) was 6% higher, which could affect the overall performance. However, this difference is not statistically significant.

D. Discussion

The results from the human user evaluation presented here indicate that in terms of subjective measures as well as objective interaction time, the trained SSE outperforms the hand-coded version. This result confirms the findings in [2], where the trained SSE also obtained better subjective scores overall, and a significantly better score for perceived success.

One of the main differences between the two SSE versions is that in the initial state of a single user interaction, the hand-coded SSE decides randomly between asking the user for their order and doing nothing, i.e., waiting for the user to order on their own initiative, whereas the trained SSE always asks the user immediately for their order. This difference is also reflected in the average number of decisions made by the single user interaction policies in Table III. Especially when no-action decisions are excluded, this number (AvgNumDecs2a) is higher for the SSE-TRA policy because it asks the user for their order more often. Although in real bar situations, it seems perfectly reasonable to assume that a customer can order without the bartender explicitly asking, in this more artificial human-robot interaction setting, this strategy might have been too confusing, resulting in the lower scores presented above.

VI. Related work

The relative affordability of the Nao robot has allowed for an increase in human robot interaction research, especially in domains that do not require advanced object manipulation or mobility. For example, [11] describes a multi-user robot application in the restaurant domain that was created with the Nao torso built-in development environment 'Choregraphe'. The Nao platform has also been used in experiments in the context of a multi-user quiz game [12], and there has also been research on using reinforcement learning for single-user human robot interaction [13]. Other work on multi-user engagement and interaction using a virtual agent has been reported in [14], [15].

VII. Conclusion

In this paper, we have presented a socially intelligent robot bartender, ported to the Nao torso robot platform. This new robot bartender is not capable of actually serving drinks, but uses a gesture for this action. However, sufficiently realistic interactions can be supported for useful multi-user human robot interaction experiments.

Using the new Nao robot bartender system we carried out a user evaluation, focused on comparing a trained and hand-coded version of the Social Skills Executor (SSE), the action selection component of the system. The results confirmed the results from a similar, recent evaluation on the original robot bartender system [2], and provided even further evidence in favour of the trained version of the SSE, which received significantly higher subjective scores than the hand-coded version, and objectively also resulted in more efficient interactions. Task success was found to be almost 20% higher with the trained policy, with interaction times being about 10% shorter. Participants also rated the trained system as being significantly more natural, more understanding, and better at providing appropriate attention.

To date, we have extended the functionality of the JAMES bartender system, aiming at more realistic and natural interactions. For example, users can now order multiple drinks, instead of a single drink of given type only, and also additional actions have been included for handling natural opening and closing of interactions (e.g., by generating and responding to greetings and goodbyes) and clarification questions (e.g., "Did you say 'blue lemonade'?"). Details of this revised system are available in [10].

For future work we plan to adopt these extensions in the Nao-based robot bartender as well, and carry out experiments for both system evaluation and data collection purposes. We aim to use the collected data to train new interaction models for this version of the system, either from scratch or by adapting the existing models that were based on data collected with the original JAMES robot bartender.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems, `http://james-project.eu/`.

References

[1] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. A. Petrick, "Two people walk into a bar: Dynamic multi-party social interaction with a robot agent," in *Proceedings ICMI*, Santa Monica, CA, 2012.

[2] S. Keizer, M. E. Foster, O. Lemon, A. Gaschler, and M. Giuliani, "Training and evaluation of an MDP model for social multi-user human-robot interaction," in *Proceedings SIGdial*, Metz, France, 2013.

[3] M. Pateraki, M. Sigalas, G. Chliveros, and P. Trahanias, "Visual human-robot communication in social settings," in *Proceedings ICRA Workshop on Semantics, Identification and Control of Robot-Human-Environment Interaction*, Karlsruhe, Germany, 2013.

[4] Microsoft Corporation, "Kinect for Windows," http://www.microsoft.com/en-us/kinectforwindows/.

[5] M. E. Foster, A. Gaschler, and M. Giuliani, "How can I help you? Comparing engagement classification strategies for a robot bartender," in *Proceedings ICMI*, Sydney, Australia, December 2013.

[6] Aldebaran Robotics, "Python SDK," http://www.aldebaran-robotics.com/documentation/ref/python-api.html.

[7] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum, "Towards an ISO standard for dialogue act annotation," in *Proceedings LREC*, Valletta, Malta, 2010.

[8] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[9] M. Giuliani, R. P. A. Petrick, M. E. Foster, A. Gaschler, A. Isard, M. Pateraki, and M. Sigalas, "Comparing task-based and socially intelligent behaviour in a robot bartender," in *Proceedings ICMI*, Sydney, Australia, December 2013.

[10] S. Keizer, M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, and O. Lemon, "Handling uncertain input in multi-user human-robot interaction," in *Proceedings RO-MAN 2014*, Edinburgh, Aug. 2014.

[11] P. Paranthaman, "Designing a restaurant assistance robot - involving engagement of multiple parties in conversation," MSc in Artificial Intelligence, School of Mathematics and Computer Science, Heriot-Watt University, Edinburgh, Scotland, United Kingdom, 2012.

[12] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *Proceedings SIGdial*, Portland, OR, 2011.

[13] H. Cuayáhuitl and I. Kruijff-Korbayová, "An interactive humanoid robot exhibiting flexible sub-dialogues," in *Proceedings NAACL HLT*, Montreal, Canada, 2012.

[14] D. Bohus and E. Horvitz, "Learning to predict engagement with a spoken dialog system in open-world settings," in *Proceedings SIGdial*, London, UK, 2009.

[15] ——, "Models for multiparty engagement in open-world dialog," in *Proceedings SIGdial*, London, UK, 2009.