Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab

Andreea Niculescu[1], Betsy van Dijk[1], Anton Nijholt[1], Dilip Kumar Limbu[2], Swee Lan See[2], and Alvin Hong Yee Wong[2]

[1] HMI Group, University of Twente, The Netherlands
[2] Institute for Infocomm Research, (I^2R), A*STAR, Singapore
{niculescuai,bvdijk,anijholt}@ewi.utwente.nl,
{dklimbu,slsee,hli,hyawong}@i2r.a-star.edu.sg

Abstract. In this paper we present the evaluation results of an exploratory study performed in an open environment with the robot receptionist Olivia. The main focus of the study was to analyze relationships between the robot's social skills and the perceived overall interaction quality, as well as to determine additional important interaction quality features with potential general validity. Our results show positive correlations between the investigated factors, as the ability to socialize with humans achieved the second highest correlation with the perceived interaction quality. One of the most relevant functional aspects for the interaction quality was found to be the ability to respond fast. Performance abilities, such as speech or object recognition were, surprisingly, considered less important. The voice pleasantness was regarded as one of the most important non-functional aspects being ranked higher than a nice physical appearance.

Keywords: social robots, quantitative evaluation, non-laboratory conditions.

1 Introduction

Since technology advances in engineering and computer science of the last decade brought the use of robots outside their traditional industrial 'playground' there has been a growing interest in designing socially competent robots for entertainment [1], educational purposes [2], healthcare assistance [3], as museum tour-guides [4], or receptionists [5,6].

As more and more social robots become available to the general public, there is an increasing trend to carry out experiments and evaluation studies in real-environment settings where the robots are meant to function. A number of social robotic projects have performed such studies with different research goals: for example Robovie [7], an interactive humanoid robot was used to explore friendship relationships between the robot and elementary school children; the robot receptionist Valerie [5] and its updated version Tank [6] were deployed to investigate long-term relationships with humans. Studies performed with the robot Pearl [3] investigated how the robot's social skills helped to improve task performance in the elderly. Minerva [4], a museum tour guide, was used to explore short-term spontaneous interactions with crowds of people.

S.S. Ge et al. (Eds.): ICSR 2010, LNAI 6414, pp. 50–62, 2010.

In this paper we present the evaluation results of an exploratory study performed with the robot receptionist Olivia during the annual two-day exhibition *TechFest,* organized in October 2009 at I^2R, Fusionopolis (Singapore). Olivia is the 4th service robot model developed by the A*STAR robotic team from I^2R (Singapore). The robot's role was not only to act as a receptionist but also to represent the institution as a kind of mascot.

Unlike the robots presented in the other studies, Olivia has the embodiment and behavior of a child. Olivia uses her childish charm to draw adults' attention so that they will interact with her. Since it has been proven that humans often treat artificial entities as though they were real [8] we hoped that Olivia's cute behavior would induce the sympathy people usually feel for young children and consequently, her overall abilities would be more positively assessed. Thus, the main goal of this research paper is, besides exploring the social relationship between humans and Olivia, to determine how the robot's social skills relate to the overall interaction quality as perceived by users. Furthermore, we investigate how people assign different ranked priorities to several functional and non-functional conversation aspects according to their importance for the interaction quality with the robot. With our study we expect to gain insights not only into interaction with Olivia, but also into the general quality features relevant to human-robot interactions.

2 Methods

2.1 Experimental Set-Up

Many human-robot evaluations presented in the social robotic literature were carried out under controlled lab conditions, where the human social 'landscape' was artificially re-constructed. These studies are especially useful for experiments aiming to determine the effects achieved through different variable manipulations. On the other hand, such experiments do not provide insights on how people would interact with the robot in spontaneous real-life situations, nor are the testing conditions comparable, i.e. systems that work well in the lab are often less successful in noisier field environments. Hence, it is necessary to evaluate human-robot interactions as socio-culturally constituted activities outside the laboratory [9].

In our study the robot receptionist Olivia interacted with human visitors in an uncontrolled real-life environment. The robot's tasks were to inform and entertain the *Techfest* visitors by presenting information about building amenities, daily horoscopes and by playing a simple game consisting of recognizing and tracking different objects.

Attached to the robot was a touch screen where additional information cues were displayed (see figure 1). Visitors could communicate with Olivia using speech or the touch screen. The topics and the games were randomly initiated by the robot: being equipped with visual-recognition capabilities Olivia was able to detect a person standing in front of her and accordingly, could initiate the conversation naturally. A conversation with Olivia typically lasted around 3-4 minutes. Olivia was accompanied by a human assistant standing at 2-3 meters distance. Visitors were free to talk with the assistant and ask questions about the robot, if they wished to. After interacting with Olivia visitors were kindly asked to fill in an evaluation questionnaire.

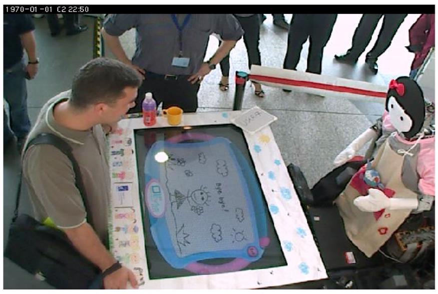

Fig. 1. Visitor interacting with the robot

2.2 Designing a Robot with Human Social Behavior

Engaging socially in verbal interactions, as simple as it might appear for humans, is in fact a highly complex process, requiring a synchronized interplay of affective, conversational and personality related behavioral cues. Fong et al. [10] translated these cues into a list of design characteristics that robots aiming to exhibit human social behavior should possess. These characteristics are: 1. express and/or perceive emotions; 2. communicate with high-levels dialogues; 3. learn/recognize models of other agents; 4. use natural cues (gaze, gesture, etc); 5. exhibit a distinctive personality or character. All these characteristics – except for 3rd - were implemented in Olivia's behavior design.

One of Olivia's most distinctive features is her role: she represents a robot mascot that looks and talks like a child; dressed up in a cute pink skirt and wearing a red 'hair' ribbon Olivia speaks to visitors with the typical charm of a very young person. Since her 'job' as receptionist and entertainer requires interacting with many people Olivia's personality was designed to be extrovert - as shown in the literature [11], extroverted individuals have enhanced social skills that allow them to communicate easily with others.

Olivia's personality profile was derived from Eysenck's [12] personality extrovert model. The model contains 7 traits such as being outgoing, talkative, lively, carefree (e.g. cheerful), responsive, easygoing (e.g. cooperative), leadership (e.g. dominant). These personality traits were implemented as follows:

1) Outgoing: Olivia's outgoingness manifests in a very friendly way of approaching people: always ready to engage in a conversation the robot usually makes the first 'move', greeting people passing by and asking them to spend time with her (see table 1).

2) Talkative: Olivia loves to talk and often adds a very personal touch to her discourse: visitors are informed not only about building amenities or horoscopes but also about Olivia's family members living in the building, about her preference for *kaya* toast or her passion for swimming. Because talkative people often use gestures to communicate, Olivia's statements are accompanied by head, arm and body movements meant to emphasize the intended message: for example, the robot uses her arm to point at relevant information cues on the screen or to show direction, shakes her head to express dizziness, waves her hand to greet people or rotates her arms to demonstrate how she swims.

3) Lively and cheerful: Olivia's speech and gestures unveil highly emotional features that leave the impression of a cheerful character with a highly animated personality: using a colorful intonation and many interjections Olivia shows surprise ("wow"), when a visitor's horoscope sign matches the one of her 'mommy', fear ("oh") when she meets a dangerous Scorpio person or joy ("hey") when she comes across a Cancer man as she likes the "yummy" taste of "chili crabs"; she yells for joy when she finds an object during the visual recognition game, 'yawns' to show boredom or complains of getting 'dizzy' when the tracking game lasts too long; in the end she gives visitors a cute onomatopoeic good-bye kiss.

4) Responsive and cooperative: Olivia's responsiveness and cooperativeness is expressed at three different levels: at the dialogue structure level, semantic level and gestural level. Through an implicit feed-back strategy ("Oh I see, you are a Leo!") the visitors are directly addressed by the robot and confronted with the internal processing state of their inquiry. On the semantic level the robot shows her interests in people's horoscopes revealing often positive characteristics: Taurus are flattered for being big and strong and Virgos for being intelligent people. Also at the gestural level Olivia shows her readiness to help by leaning her upper body and head towards the touch screen in an attempt to look for corresponding answer images that she can point at.

5) Dominant: Since our experiment was carried out in uncontrolled environment settings there is a need to guide the visitors in order to maintain a smooth interaction. Thus, this personality trait derived from the leadership characteristics was added to Olivia's personality model. Olivia's dominance is expressed on the dialogue structure level - the conversations is initiated, lead and ended by the robot - and on the semantic level - the robot uses the first person to refer to herself often displaying an assertive verbal behavior (e.g. "I like *kaya* toast!").

Table 1. Excerpt from two conversations with Olivia

Amenities dialogue	Game dialogue
Olivia: Hi *(waving hand)*! I am Olivia! Nice to meet you! **User**: Nice to meet you too! **Olivia**: Would you spend some time with me? **User**: Sure! **Olivia**: Hmhm (*clearing her 'throat'*)… I know a lot about amenities here (*makes a round movement with the arm showing the amenities depicted on the screen*). Tell me, which one you like to know more? **User**: I would like to know more about "Fitness First." **Olivia**: Hm … my daddy works out at "Fitness First", located at level 23 (*points on the screen where the fitness center details are displayed).*	**Olivia**: Now, let's play a game! **User**: Ok, what game? **Olivia**: Hmhm *(clearing her 'throat')...* pick up my toy *(points to her book left on the screen and looks up at the visitor)* and move it slowly in front of me, as I follow the motion. **User**: (*moves the book too fast*) **Olivia**: Hey! It's too fast! I can't catch up with you! **User**: (*moves the book slowly*) **Olivia**: It is fun! (*moves her head following the book; after sometime starts 'yawing' and pushes her upper-body closer to the book; after sometime starts shaking her head and brings back her upper-body*) I am getting dizzy! Let's stop here! **User**: (*puts the book down*) **Olivia**: Muac (*kiss sound*)! Thank you for playing with me and have a nice day (*waving hand*)!

2.3 Technical Features

Olivia is approximately 152 kg and 1.6 m tall. The robot has 13 degrees of freedom in total: head (3 degrees), body (2 degrees) and hands (2x4 degrees). It is built on a PowerBot base mobile platform and equipped with several hardware/mechanical components, including actuators (servomotor, harmonic gear system, drive unit and harmonic drive servo actuators) a laser (Hokoyu URG-04LX), cameras (Bumblebee®2 and DVN1501 mono camera), microphones and speakers. The robot has several independent software modules for controlling and executing several functions: a motion control (MC), a dialog management system (DMS) and a vision understanding (VU). The MC module employs advance motion control algorithms, such as nonlinear task space control and joint space control to control the robot's movements. The DMS module utilizes the Loquendo 7.52 text-to-speech (TTS) software to generate a female, child like voice with an American English accent (timbre & pitch=70, speech rate=30, volume=50). The TTS enables the use of several emotion cues, such as hesitation sounds, coughs, yawning, etc. To increase the speech recognition accuracy the DMS' acoustic model was trained with 13.5 hours of read speech data, collected from 40 English non-native speaker subjects (mostly male). For the data collection a 200 word vocabulary was used; additional word entries related to the two main conversation topics (building amenities and horoscopes) were included (50 words per topic). The VU module deploys a multi-model fusion maximum likelihood method by integrating four different approaches: stereo-based human detection, HOG- based human detection, color-based tracking, and motion estimation for human detection and tracking. All software modules run two PC boards: one Intel Corei7 (2.8 GHz) and one Atom processor (1.2GHz).

2.4 Questionnaire Design

Since we are interested in the relationship between the robot's social skills and the perceived overall interaction quality, it is important to find adequate ways to measure them. Additionally, we are interested in finding the most relevant conversational aspects contributing to a better interaction assessment.

A tool widely used in behavioral research for social skills evaluation was developed by Gresham and Elliott [13]. The tool is meant to assess human social skills along five categories: cooperation, assertion, empathy, self-control and responsibility. These categories were found to match social abilities aspects involved in human-robot interaction [14] being related to the design characteristics presented by Fong et. al [10]. Translated to Olivia, these abilities, partly overlap with her extrovert personality characteristics and are expressed in the following way: cooperation manifests in her readiness to help others by sharing information in a highly sociable manner, referring to her 'own' experiences and using gestures to enhance explanations. Assertiveness relates to Olivia's extrovert personality, as she initiates the conversation, introduces herself and shows openly her preferences and dislikes. Olivia expresses empathy through emotional, verbal interjection. For self-control and responsibility no direct related aspects were found. Since many authors [15] suggested that humor has an important role in interpersonal relationships, being a social skill in itself, we included it in our investigations. Olivia's humor is, however expressed only through a personalization effect: the robot often refers to itself as it would be human, creating a hilarious impression (see table 1). Consequently, we built up a ***social skills*** subscale with 5 items: the ability to socialize (i.e. ability to be friendly), to use natural gestures, to express emotion, personality and humor.

To evaluate the interaction quality we used the SASSI [16] questionnaire as inspiration. The questionnaire was developed to evaluate the usability of uni-modal speech-based interfaces and it addresses five different dimensions: response accuracy, likeability, cognitive demand, annoyance, habitability and speed. Because evaluating the interaction quality with a multimodal interface differs somewhat from assessing the usability (fit-for-use) of a uni-modal system we needed to modify the questionnaire to suit our purpose. Accordingly, we retained only items corresponding to the interaction features and their effects on users' mood; additionally we replaced the accuracy dimension with a more precise category referring to the robot's multimodal performance and we semantically re-grouped the items in two factor subscales: ***interaction features*** and ***user feelings***. The interaction features subscale contained 8 items: interaction easiness, level of concentration, response speed, usefulness, flexibility, speech/object recognition and object tracking. The user feelings subscale includes only 3 items: enjoyment, calm and comfort; the comfort was not listed in the SASSI questionnaire, but it is often mentioned in the literature along with user enjoyment as contributing to the overall interaction quality perception [17].

According to Hassenzahl et al. [18], the user evaluation of a system is influenced by its pragmatic and hedonic quality. Applied to conversational interactions, pragmatic quality would refer to ***functional aspects*** determining how well a certain communicative goal is achieved, while hedonic quality would relate to ***non-functional aspects*** indicating how much the user enjoyed the interaction. Thus, we selected from the SASSI questionnaire and other relevant evaluation studies concerning multimodal conversational interaction [17] a total of 16 items, applicable to Olivia and social robots in general. 7 items were related to functional aspects, such as interaction speed, content relevance, clarity of answers, speech/visual recognition accuracy, system transparency and easy recovery from errors. The other 9 items were concerned with non-functional aspects, such as voice and appearance pleasantness, friendliness, politeness, humor, emotion display, gestures and mimic, display of human-like physical characteristic (gender and age).

The questionnaire was divided in two parts. In the first part visitors scored the subscale items using a 5-point Likert scale with 'strongly agree/disagree' as endpoints. In the second part they ranked the functional and non-functional aspects according to their importance for the interaction quality. A 7-point scale with 'not important at all/extremely important' as endpoints was used for the ranking in order to ensure more differentiated results. The questionnaire ended with a general question about the perceived ***overall interaction quality***.

3 Results and Discussions

From 121 visitors who interacted with Olivia 88 filled in the questionnaire. 67.8% were male and 32.2% were female. 73.3% were of Chinese origin, 14.4% Indian, 12.1% other nationalities. The majority (71.1%) had an IT & engineering background, the rest sharing a background in business (13.3%), arts & humanities (5.6%) and other areas (9.9%). 66.7% were aged between 26-40 years, 20% between 18-25 years and 13.30% were above 41 years. More than half of the visitors (54.5%) were Master or Ph.D. holders; 34% had a Bachelor degree and 11.1% held other diploma degrees. Probably, due to a technical educational background a relatively high percentage (47.2%) had seen or read about robots and some visitors (24.5%) had even interacted with them; also other few (6.7%) had expertise in robot design & development. 24.5% had no knowledge of robots. A lower percentage of visitors (38%) had used speech recognition devices – mostly as input modality for mobile phones, video games, cameras, dictation systems; a very small number of visitors (3%) used the Microsoft SDK tool to build speech recognition applications; 61.10% of the visitors had no knowledge about speech recognition devices.

Next, we checked the internal consistency of the proposed subscales, as well as the cumulated negative ('disagree' + 'strongly disagree'), positive ('agree' + 'strongly agree') and neutral scores achieved by each subscale item[1]. The reason behind listing the cumulated values lies in understanding the general item evaluation tendency. Subscales with α >.600, item total correlations r >.300 and a reduced number of items (<10) are generally considered as acceptable [19].

Table 2. Robot social abilities subscale and item cumulated score values

Robot's social skills	Items	Item-total correlation r	C_{neg}	Neutral	C_{pos}
N of items: 5 Cronbach α =.789	Socialize	.524	5.70%	42.00%	52.30%
	Nat. gestures	.536	15.90%	43.20%	40.90%
	Personality	.646	17.00%	48.90%	34.10%
	Emotions	.621	26.10%	39.80%	34.10%
	Humor	.499	15.90%	52.30%	31.80%

The analysis of the robot's social skills subscale (see table 2) revealed a high internal consistency (α = .789). According to a Friedman test, significant differences between the items were found ($\chi 2(4)=26.671$, p=.000). Thus, we conducted a post-hoc analysis with a Wilcoxon Signed-Rank test (W_+) applying a Bonferroni correction (BC) for multiple comparisons; a new p-value was set at p<.012. The test showed that the ability to socialize was significantly higher scored than all others subscale items (p=.000), except for the ability to express natural gestures (p=.016).The lowest rated item seems to be the ability to express emotion, however no significant difference with respect to the other items was found. All subscale items, except for the ability to socialize show high frequency distributions in the neutral category. This means that Olivia's social skills are acceptable, as she can 'socialize' but most of the features need improvements. Especially the ability to express emotions - a key item with the second highest subscale correlation (r=.621), but also highest negative ratings (26.10%) - should be given special attention in the future. The lack of mimicry on Olivia's face, most probably might have lowered the rating, as humans typically expect emotion expression to appear synchronized at both voice and face level.

[1] Since the data is not normally distributed we do not report the mean.

Table 3. User feelings subscale and item cumulated score values

User feelings	Items	Item-total correlation r	C_{neg}	Neutral	C_{pos}
N of items: 3 Cronbach α =.696	Comfort	.433	10.20%	36.40%	53.40%
	Enjoyment	.551	6.80%	29.50%	63.60%
	Calm	.574	2.30%	21.60%	76.10%

The reliability analysis performed on the user feelings subscale proved an internal consistency of α=.696 (see table 3). Since all items showed relatively good scores we assume the majority of the visitors felt comfortable and calm while interacting with the robot, enjoying the conversation.

The interaction features subscale showed an internal consistency of α=0.645 (see table 4). Two items -attention level required and interaction flexibility- were removed because of low correlations with all scale items and with the overall interaction quality (r=.087, and respectively r= -.174, *p*=n.s.).

Table 4. Interaction feature subscale and item cumulated score values

Interaction features	Items	Item-total correlation r	C_{neg}	Neutral	C_{pos}
N of items: 6 Cronbach α =.645	Easiness	.438	13.60%	40.90%	45.50%
	Interaction speed	.331	27.20%	44.30%	28.40%
	Usefulness	.361	1.10%	34.10%	64.80%
	Speech recognition	.390	1.10%	22.70%	76.20%
	Object recognition	.385	4.50%	23.90%	71.60%
	Object tracking	.329	5.70%	37.50%	56.80%
Removed parameters	**Items**	**Item-total Correlation r**	C_{neg}	**Neutral**	C_{pos}
N of items: 2 Cronbach α (if included) =.485	Attention level required	.091	4.50%	15.90%	79.50%
	Flexibility	-.198	30.60%	37.50%	31.80%

A post-hoc analysis ($W_{+,}$ BC, new p<.012) performed after the Friedman test ($\chi^2(5) = 86.336$, p=.000) showed significant differences between the interaction speed and easiness on one side and all the other items, on the other side (p=.000). Also, significant differences were found between speech recognition and object tracking capabilities (p=.000).

The response slowness was mostly caused by speech[2] and visual recognition difficulties in respectively, 54.40% and 28% of the cases; since no feed-back or error recovery strategies were implemented, i.e. no reaction came when the recognition score was below a certain threshold, the visitors were left with the impression the robot's response was slow. Table 5 presents an overview of response latencies[3] for speech and object recognition, as well as for visitors' tolerance level[4] to speech response latencies. The robot's response latency in speech recognition error-free cases

[2] The speech recognition problems were caused by a noisy environment (92.80%), wrong pronunciations (4.8%) and other technical issues (2.4%).

[3] Response latency refers to the time elapsed between last user input and robot's response.

[4] The tolerance level refers to the time elapsed until a user re-prompts her input when no response is given.

was on average 2.51sec, a value still far behind that found in human face-to-face conversation - 0.97 sec, [20].

In cases with speech recognition errors response latencies were higher, lasting on average 3.18 sec; in such cases, total response time[5] could achieve extreme (but luckily infrequent) values of even 45 sec. Compared with human face-to-face conversations, where a delay of more than 2-3 seconds in providing a response was found to cause discomfort [21] it becomes clear that such response latencies are unacceptable. Interestingly, the delay value mentioned above corresponds roughly to the tolerance level of 3.74 sec measured in our study. Thus, we would expect high correlations between speech recognition errors and the speed scores. But surprisingly, this was not the case: the speech recognition and the interaction speed have a correlation coefficient of only r=.279. A further detailed analysis revealed the speech recognition performance correlates with the speed scores in only 30% of the cases: many visitors (44.30%) scored the interaction speed as being neutral - neither fast, nor slow, even by high response latencies (23 up to 45 sec.) or by relatively low latencies (1.5-2 sec.). This leads to the following three remarks. Firstly, an average response latency of 2.51 sec is too high. Secondly, the question referring to speech recognition abilities ("The robot was able to recognize my speech") should have been formulated more accurately (e.g. "The robot's ability to recognize my speech was very good"); since Olivia always provided a response, even after long response delays, it means she was 'able' to recognize speech; therefore, the question might have generated misleading responses. Thirdly, many visitors tended to avoid negative scores choosing instead neutral ratings; this tendency of scoring more positively in order to please the interviewer or to be helpful was also observed by other studies [22].

Table 5. Robot's speech and visual average response latencies in seconds

Response latency	Mean	Median	Modus	Min.	Maxi.
ASR error free	2.51	2	2	0.75	7
ASR with errors	3.18	3	2	1	15.50
Total time until response-ASR errors	14.96	11.25	6	4	45
Visual recog. error free	1.77	1	1	0.25	11
Visual recog. with errors	13.52	12	11	4	27
Tolerance level	3.74	3.31	3	2	9.33

Table 6. Overall interaction quality scale with cumulated score values

Scale	$C_{neg.}$	Neutral	C_{pos}
Overall interaction quality	2.00%	30.7%	63.6%

Similar rating behaviors were observed between speech recognition performance and the scores obtain in overall interaction quality (see table 6): the presence or absence of speech recognition errors corresponds to a negative/positive overall quality assessment in only 34.66% of the cases; visitors gave more neutral scores, even if the robot obviously failed to recognize their speech and her response had long delays. We

[5] The total response time refers to time measured between first user input and robot's response.

also compared the ratings for enjoyment in cases with speech recognition error and found that in 60% of the cases nevertheless, the visitors gave high ratings. Looking at the enjoyment correlations outside its own subscale we found the highest correlations with interaction easiness (r=.442), ability to socialize (r=.436) and overall quality (r=.418). Thus, the visitors' tendency towards more positive ratings as observed in [22] might have an additional, complementary explanation: people might have rated the interaction features and overall quality better because they experienced an enjoyable (and not particularly difficult) interaction with a sociable robot.

Finally, we analyzed the correlation between the subscales and the perceived overall quality (see table 7). All subscales correlate significantly with the overall interaction quality, whereas the interaction features have the highest correlation coefficient (r=.600). The robot's social skills have a lower correlation coefficient (r=.444) but to some extent higher as compared with the user feelings. Additionally, we checked the correlations between the overall interaction quality and each subscale item to detect the highest correlations; we found that the interaction easiness (r=.490), the ability to socialize (r=.435), the enjoyment (r=.418) and usefulness (r=.409) had the highest correlations (>.400) with the overall interaction quality. On the other hand, the interaction features were significantly better evaluated than the robot's social skills (W_+, p=.000).

Table 7. Overall quality scale with cumulated ratings

Subscales / Overall Quality	**Min.**	**Max.**	**Mean**	**Std. deviation**	**Correlation with overall quality r**
Robot social skills	2.00	5.00	3.2409	.55183	.444**
User feelings	2.00	5.00	3.6629	.55936	.435**
Interaction features	2.67	4.83	3.5568	.43468	.600**
Overall scale	2.64	4.79	3.4156	.37021	.589**
Overall quality	**2.33**	**5.00**	**3.5083**	**.58880**	1.000

** p<0.01

Next, we analyzed the priorities ranks visitors assigned to different functional and non-functional aspects that might be involved in the face-to-face conversation with a robot (see table 8). Despite a non-normal data distribution we chose the mean as sole option to build a differentiation order. However, the rank order can be validated by only applying a non-parametric significance test. The mean scores show a demarcation line between the functional (1-7) and non-functional (8-16) aspects. The functional aspects were on average significantly higher ranked than the non-functional (W_+, p=.000). This finding is not surprising, since functional aspects are, from a pragmatic point of view, more important than non-functional aspects, e.g. the robot's nice appearance would not replace its poorly working speech recognition. Nevertheless, this does not mean non-functional aspects are unimportant. In fact, many studies proved the benefits of non-functional aspects such as emotion displaying, gesture and mimicry for the robot's social acceptance or human-like skills [10].

The result of the post-hoc analysis (W_+, BC, new p<.008) performed after the Friedman test ($\chi^2(6)$=18.38, p=.005) revealed that the interaction speed was statistically higher ranked than the error-free speech/object recognition (p=.006/.007); this means that users could be more tolerant to errors, but less understanding if they have to wait too. No significant rank differences were found between the other aspects, except for the system transparency whose mean was significantly lower than those of the interaction speed (p=.005).

Among the non-functional aspects statistically significant differences could be found ($\chi^2(8)$=175.696, p=.005). The post-hoc test (W_+, BC, new p <.006) showed that the pleasant voice, friendly behavior and politeness were significantly higher ranked than the humor and the gender/age displaying. In fact, both gender/age displaying were included on the aspects list because of their relative importance in verbal addressing in Asian cultures; however, they achieved the lowest statistically significant ranking of all aspects (p=.000). Interestingly, a pleasant voice achieved a statistically significant higher mean than a nice physical appearance (W_+, p=.005). This result could be explained as follows: even if the visual impression of the robot would impact visitors in the first place, its voice might play a more important role in the interaction, since it conveys the required information.

Table 8. Mean scores and significance levels for functional and non-functional face-to-face conversational aspects

No.	Category	Mean	Significance level relative to the other item rank[6]	
1	Interaction speed	5.83	*=5,6,7; ***=8-16	ns= 2-4
2	Easy recovery from errors	5.80	***=8-16;	ns=1,3,4-7
3	Clarity of answers	5.77	***=8-16;	ns=1,2,4-7
4	Delivering relevant information	5.72	***=8-16;	ns= 1-3, 5-7
5	Error free speech recognition	5.67	*=1; ***=9-16;	ns=2-4,6-8
6	Error free object recognition	5.61	*=1; ***=9-16	ns=2-5, 7,8
7	System transparency	5.52	*=1; ***=10-16;	ns=2-6, 8, 9
TOTAL	**FUNCTIONAL ASPECTS**	**5.70**	***** TOTAL NON FUNCTIONAL**	
8	Pleasant voice	5.26	**=1-4,12, 14-16;	ns=5-7,9-11,13
9	Friendly behavior	5.22	**=1-6,14-16;	ns=7,8,10-13
10	Gestures and mimic	5.01	**=1-7,15,16;	ns=8, 9,11-14
11	Polite way of talking	5.01	**=1-7,14-16;	ns=8-10,12,13
12	Nice physical appearance	4.95	**=1-8,15,16;	ns=9-11,13,14
13	Emotion displaying	4.92	**=1-7,15,16;	ns=8-12,14
14	Humorous way of talking	4.72	**=1-9,11,15,16;	ns=10,12,13
15	Gender displaying	4.10	**= 1-14, 16	ns= none
16	Age displaying	3.66	**=1-16	ns= none
TOTAL	**NON-FUNCTIONAL ASPECTS**	**4.80**	***** TOTAL FUNCTIONAL**	

4 Conclusions

In this study we analyzed relationships between a robot's social skills, interaction features and user feelings on one side, and the perceived overall interaction quality, on the other side. Our results showed significant correlations between these three factors and the perceived overall quality, the interaction features showing the highest correlation. We would have expected a stronger relationship between the robot's social skills and the perceived overall interaction quality. Nevertheless, the ability to socialize seems to play an important role, being the second most correlated item with both enjoyment and overall interaction quality.

[6] *** significant at: p<.005 (9 comparisons), ** at p<.006 (8 comparisons), * at p<.008 (6 comparisons); '**ns**' stands for not significant.

The robot's speech recognition performance was better ranked than the error logs and total competition time would have predicted. This might be explained, partly by the question formulation bias, partly by a general human tendency to give more positive ratings and partly because the visitors enjoyed the interaction despite errors and long response delays.

The conversational aspects ranking brought us important information that can be used to improve the robot design and set priority decisions: for example, it seems that visitors were more tolerant to errors than to long response latencies. A pleasant voice seems to be more important than a nice physical appearance while humor and gender/age displaying appear to be less important conversational aspects for the interaction quality.

Both, aspects ranking and correlations obtained from the items' evaluation suggest that the overall interaction quality relates more to the robot's ability to lead the interaction (response speed, clarity of answers, interaction easiness) and to appear agreeable (friendly, i.e. sociable, having a pleasant voice), than to its performance accuracy, in terms of speech and object recognition/tracking.

In the future Olivia's dialogue design would incorporate an error-handling strategy to reduce the robot's response perceptions as being slow. Also, help options and a better system transparency would be integrated to enhance the interaction easiness. Further, adding mimicry to Olivia's face to show emotions, improving her gesture to becoming more natural and making the dialogue script more amusing might increase the robot's perceived ability to socialize. Our results are originated from an exploratory study and therefore, cannot prove causal relationships between the analyzed items. Nevertheless, the study revealed significant item correlations that can be used to improve the current robot design. Their significant impact could be examined in future contrastive laboratory conditions to find statistical evidence.

Acknowledgments

We are grateful to A*STAR Robotics team for their excellent development work on Olivia 4.0 service-robot model. A special thank to Alexandre Thebaud and Lynn Packwood for careful proof-reading and useful suggestions. This work has been supported by the EU's 7th Framework Program (FP7/2007-2013) under grant agreement no. 231868 (SERA).

References

1. Fujita, M.: AIBO: Toward the era of digital creatures. Int. Journal of Robotic Research 20 (10), 781–794 (2001)
2. Kanda, T., Hirano, T., Eaton, D.: Interactive robots as social partners and peer tutors for children: A field trial. In: HCI, vol. 19, pp. 61–84 (2004)
3. Pollack, M.E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J.T., Dunbar-Jacobs, J., McCarthy, C., Thrun, S., Montemerlo, M., Pineau, J., Roy, N.: Pearl: A mobile robotic assistant to the elderly. In: AAAI Workshop 2002 notes, Edmonton, AB (2002)

4. Schulte, J., Rosenberg, C., Thrun, S.: Spontaneous, short-term interaction with mobile robots. In: Proc. of IEEE Int. Conference on Robotics and Automation, pp. 658–663 (1999)
5. Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A.C., Wang, J.: Designing robots for long term social interaction. In: IEEE/RSJ Int. Conference on Intelligent Robots and Systems, pp. 2199–2204 (2005)
6. Lee, M.K.: How do people talk with a Robot? An analysis of human-robot dialogues in the real world. In: Proc. of CHI 2009, Boston, pp. 3768–3774 (2009)
7. Kanda, T., Sato, R., Saiwaki, N., Ishiguro, H.: A two-month field trial in an elementary school for long-term human-robot interaction. IEEE Transactions on Robotics 23(5), 962–971 (2007)
8. Reeves, B., Nass, C.: The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, NY (1997)
9. Sabanovic, S., Michalowski, M.P., Simmons, R.: Robots in the wild: observing human-robot social interaction outside the lab. In: Proc. of AMC 2006, Istanbul (2006)
10. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42(3-4), 143–166 (2003)
11. McCrae, R.R., Costa, P.T.: A five-factor theory of personality. In: Pervin, L., John, O.P. (eds.) Handbook of personality: Theory and research, 2nd edn., Guilford, NY, pp. 139–153 (1999)
12. Eysenck, H.J., Eysenck, S.B.G.: Manual of the Eysenck personality inventory. University of London Press, London (1964)
13. Gresham, F.M., Elliot, S.N.: Social skills rating system manual. American Guidance Service (1990)
14. Heerink, M., Kröse, B.J.A., Wielinga, B.J., Evers, V.: Studying the acceptance of a robotic agent by elderly users. Int. Journal of Assistive Robotics and Mechatronics 7(3), 33–34 (2006)
15. Yip, J.A., Martin, R.A.: Sense of humor, emotional intelligence, and social competence. Journal of Research in Personality 40(6), 1202–1208 (2006)
16. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering 6(3-4), 287–303 (2000)
17. Möller, S., Engelbrecht, K., Kühnel, C., Wechsung, I., Weiss, B.: A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In: Proc. QoMEX, pp. 7–12 (2009)
18. Hassenzahl, M., Platz, A., Burmester, M., Lehner, K.: Hedonic and ergonomic quality aspects determine software's appeal. In: Proc. CHI 2000, Den Haag, pp. 201–208 (2000)
19. Tabachnick, B.G., Fidell, L.S.: Using multivariate statistics. Allyn and Bacon, Boston (2001)
20. Jaffe, J., Feldstein, S.: Rhythms of Dialogue. Academic Press, New York (1970)
21. McLaughlin, M.L., Cody, M.J.: Awkward silences: behavioral antecedents and consequences of the conversational lapse. Human Communication Research 8(1), 299–316 (1982)
22. Worcester, R.M., Burns, T.R.: A statistical examination of the relative precision of verbal scales. Journal of Market Research Society 17(3), 181–197 (1975)