Robot-Era Project: Preliminary Results on the System Usability

Roberta Bevilacqua[1(✉)], Elisa Felici[1], Fiorella Marcellini[1],
Sebastian Glende[2], Susann Klemcke[2], Isabel Conrad[2],
Raffaele Esposito[3], Filippo Cavallo[3], and Paolo Dario[3]

[1] Scientific Direction, INRCA, Ancona, Italy
{r.bevilacqua,e.felici,f.marcellini}@inrca.it
[2] YOUSE GmbH, Berlin, Germany
sebastian.glende@youse.de
[3] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pontedera, Italy
{r.esposito,f.cavallo,p.dario}@sssup.it

Abstract. The European project Robot-Era is an ambitious integrated project (FP7-ICT-2011.5.4), which objective is the development of advanced robotic services, integrated in intelligent environments, to provide independent living to older people.

In order to guarantee the matching of the users' need and the demands, two loops of experimentation were conceived, in realistic and real setting.

The aim of the paper is to described the methods applied and the main results coming from the first experimental loop, concerning the degree of usability of the interfaces and provide guidelines for testing socially assistive robots with older people.

Keywords: Usability assessment · Older people · HRI · HCI · Acceptability

1 Introduction

Due to the demographic increase of older people living alone and in need of care [1], a large number of research projects have been focused on the development of robotic services for ageing-well applications [2]. The support offered by the robotic applications can be related to mobility, providing household and safety maintenance and continues monitoring [3, 4].

The complexity of the robotic systems required a careful methodological consideration on how to properly approach the evaluation of the system usability and acceptability, most of all in case of system tailored for older people.

The aim of the paper is to describe the methods applied and the main results coming from the first experimental loop of the Robot-Era project (FP7-ICT-2011.5.4). As first loop of experimentation, the analysis of the usability of the interfaces has deserved a greater importance, in order to understand both the user- and the technology-oriented improvements to be made for the technical improvement. Moreover, the interest of exploring the results obtained has also to be seen in a methodological perspective,

in order to understand which tool can be of support to collect useful data on the usability of complex robotic systems, as the Robot-Era platform.

At this purpose, the aim of the paper is not only to report briefly the first results obtained during the system validation, but to use an extract of the results to critically analyze the methodological choices done and investigate how to implement it in the view of the second experimental loop.

1.1 The Robot-Era Platform

The Robot-Era architecture integrates a multi-robot system able to work in different environment such as outdoor, condominium and indoor [5]. It includes also a domestic Wireless Sensor Network (WSN), constituting an Ambient Intelligence (AmI) infrastructure, that supervises the home and localize the user. Other agents of the system include the elevator and the user interface sub-system. The system is composed by three different robots acting in three different environment.

The Domestic Robot (DORO) is designed to safely navigate in a domestic environment. It is equipped with a robotic arm in order to take small objects. Multicolor LEDs, mounted on the eyes, and speakers provided a feedback to the user. The robot has a removable tablet that user could use for service requests.

The Condominium Robot (CORO) navigates between floors through the elevator. It doesn't have an arm, but it is mounted a roller mechanism in order to be able to exchange goods with the Outdoor Robot.

The Outdoor Robot (ORO), designed on DustCart platform [6], is an autonomous mobile robot for objects transportation in urban environment. It is equipped with a container for the objects, a robotic head and a touch screen used primarily for human-robot interaction and sensors for obstacle detection and localization.

The users interact with the system using a web-based interface (named Graphic User Interface, GUI), from the DORO tablet or with a wearable microphone connected with a speech recognition software module.

The services offered by the Robot-Era platform are shopping delivery, reminding, communication, laundry, food delivery, object transportation and manipulation, garbage collection, surveillance, indoor escort, outdoor walking support. For the purpose of this paper, only the results of shopping delivery, reminding and communication services will be presented. The GUI of the Robot-Era system is described in details in Di Nuovo et al. [7].

2 Methodology

2.1 Recruitment and Quota Plan

The project focuses on a specific target group of elderly people, aged over 65 years old, with moderate health problems and motor and cognitive deficits, living alone or with their relatives but without a devoted caregiver.

Older people have to be able to communicate their use-preferences and opinions on the Robot-Era services, actively participating in the technical development and design

of the technologies. Within the target group of elderly people, different levels of autonomy could be observed: High Level of User Autonomy (HLUA), Middle Level of User Autonomy (MLUA) and Low Level of User Autonomy (LLUA).

A method to receive fundamental statements from smaller samples in qualitative research is to set up a quota sampling [8], that means researchers are choosing the test participants based on theoretical assumptions in condition to the research question [9]. By using quotas as an orientation an arbitrary sample composition based on subjective criteria can be avoided. For the first experimental loop a number of 70 participants was prescribed. For guaranteeing the comparability between the test sides (Italy and Sweden) the sample has be divided into two equal samples (each consisting of 35 participants). The number of women and men was chosen differently because of the different life expectancy of males and females [10]. These facts lead to the assumption, that also a higher proportion of woman would potentially use the robots in the future. The operative inclusion criteria of the first experimental loop were: Older people aged ≥ 65 years old, positive evaluation of mental status at the Short Portable Mental Status Questionnaire [11] (cut off to be enrolled = nr. Errors ≤ 3), autonomy in performing daily activities with domestic tools, evaluated with the Instrumental Activities of Daily Living by Lawton [12] (cut off to be enrolled = score > 2), absence of psychiatric illness, substance abuse and communication impairments.

2.2 Tools for Measuring the Usability

Before developing the test protocol, a detailed analysis of the literature were conducted, with aim of identifying the most appropriate tools and metrics to be applied [13]. The analysis of available theories on usability evaluation has resulted in the decision of adopting a mix-methods approach, based on a quantitative and qualitative data collection.

To evaluate the overall experience and usability of the Robot-Era services, a qualitative usability test with use cases was set up. The participants tried to perform predefined test tasks, which were oriented on use cases. These use cases [14] were based on scenarios that were described in detailed step-by-step descriptions offered by the technical partners. While executing the tasks, the test persons were observed by experts who documented the difficulties. The usability test was combined with the thinking aloud technique (TAL), [15]. The use of the thinking aloud method allows the researcher to investigate in detail the overall user experience because people express their feeling, thoughts and skepticism directly when using the system. Further the Systems Usability Scale (SUS, 1996) [16] was used. This simple and not highly detailed evaluation method uses a standardized form with ten questions to assess the product's usability, on a five-point Likert scale.

In order to deeply understand the quantitative data collected, two detailed phases of video observation were conducted:

the first phase was aimed at evaluating the Human Computer Interaction (HCI) and was mainly oriented to the users' performance with the graphic user interface (GUI), running on a devoted tablet;

the second phase of observation was aimed at evaluation the overall acceptability and Human Robot Interaction (HRI).

Before starting with the analysis, a set of indicators for acceptability and HRI was selected for each service, starting from the definition of acceptability from the Unified Theory of Acceptance and Use of Technology [17] based on 13 core constructs, whose anxiety, attitude and perceived enjoyment were selected to be collected by users' free statements, as well as, the emotional reaction during the task execution, through the non verbal communication. Regarding HRI, it was decided to guide the observation through the analysis of the [18] head orientation (robot head, robot body, robot eyes, experimenter, tablet), gaze (robot, experimenter, default), body orientation (robot, experimenter, straight ahead), proxemics (approach robot, touch robot) as well as free statements of the users on the robot and its features or emotional reaction to the robot presence. The communication analysis for HRI was conducted by analyzing the presence of specific codified events.

2.3   Settings of the Pilot Sites

The tests were performed in two different pilot sites: in Peccioli (Italy) and in Örebro (Sweden). The setting with end-users in Peccioli was composed of indoor and outdoor realistic environments. The indoor environment consisted of a real apartment in a modern structure and the building consisted of two floors, the DomoCasa Lab. The outdoor environment was around the building. The Ängen Research and Innovation Apartments (Ängen-RIA) were situated in the Ängen healthcare complex in Örebro, Sweden, including senior living apartments, a retirement home for elderly in need of day-to-day care, and a local clinic.

3   Procedures

The experimental loop was conducted in Sweden and Italy, involving 67 older users. The results of the three services are available for 35 Italian users and 22 Swedish users. After the preliminary introduction, the interviewers presented the Use Cases to be executed to the participants and explained how to perform the TAL, during the use.

The Use Cases identified for the shopping delivery service were:

– send the robot go shopping
– receive the delivery from the robot;
  for the communication service:

– starting a video call;
– accept an incoming video call;
  for the reminding service:

– create a new appointment/event;
– reminding of taking the medicine.

Once completed all tasks, the SUS was administrated for each service. During the execution of the test, the two phases of observation were conducted, through videos [19].

4 Results

The subparagraphs below reported the results obtained by applying the different methods. Concerning the results of the observation, they are still under elaboration by the partners, so they are shown in the form of narratives, to offer qualitative insights to the developers. Moreover, the evaluation of the communication service was negatively influenced by the numerous technical failures of the systems during the testing phase, not allowing a proper observation of the performance of the participants. Concerning the shopping service, the results of the SUS are available only for the Italian sample. Finally, the results of the video observation on the HRI and acceptability issues are based only on the Italian sample analysis.

4.1   System Usability Scale and Thinking a Loud

In order to get an overall usability index, the SUS score is ranged from 0 to 100. From the literature [20], the following cut off were chosen, to analyze the level of the services: not usable (score < 65), usable ($65 \leq$ score < 85) and excellent (score $\geq$ 85). In addition to the overall evaluation of the services usability, the results of relevant statements were presented, whose scores are ranged from 1 to 5, as described in paragraph 3.2. As regards the overall usability of the shopping delivery service, the results of the Italian sample (n = 35) show that the proposed system is not usable for 6 elderly volunteers (M = 50.83; SD = $\pm$ 7.85), while it is usable for 8 of them (M = 77.81; SD = $\pm$ 6.74) and excellent for 21 participants (M = 90.71; SD = $\pm$ 4.04). In particular, elderly people are neither agree nor disagree about using this service frequently (M = 3.40; SD = $\pm$ 1.35). However the participants report that the GUI for the shopping service is easy to use (M = 4.03; SD = $\pm$ 1.32), and only 3 of them think that they would need the support of a technical person to use it. About communication service, 2 users of the Italian sample (n = 35) reported that the interface was not usable (M = 41.25; SD = $\pm$ 15.91), while it was considered usable for 10 users (M = 77.75; SD = $\pm$ 4.92) and excellent for 23 of them (M = 92.83; SD = $\pm$ 4.90). In addition, the results from the Swedish participants (N = 22) have shown that the communication service is considered not usable for 6 elderly volunteers (M = 50.00; SD = $\pm$ 19.17), while it is usable for 8 of them (M = 71.88; SD = $\pm$ 6.59) and excellent for 8 participants (M = 93.75 SD = $\pm$ 5.89).

Overall, it can be said that all the 57 users would be enough disposed to use the communication service without the support of a technical person, perceiving the task as very easy (IT users M = 4.49; SD = $\pm$ 0.92, SE users M = 3.73; SD = $\pm$ 1.42). Concerning the reminding service, the Italian results (N = 35) have shown that it was considered not usable for 6 users (M = 40.83; SD = $\pm$ 14.72), while it is usable for 10 users (M = 70.25; SD = $\pm$ 3.43) and excellent for 51.52 % of the sample (M = 96.62;

SD = ± 4.41). The reminding service was also tested by 18 Swedish users and it was considered mainly not usable for 55.56 % of the sample (M = 36.75; SD = ± 19.58), while it was considered usable for 6 users (M = 78.33; SD = ± 4.65) and excellent only for 2 of them (M = 87.50; SD = ± 3.54).

In the case of the Robot-Era first experimental loop, the TAL has produced not relevant information, mainly in the case of the Italian sample. Despite the positive results achieved at the SUS, that have suggested a highly degree of usability of the GUI, the qualitative analysis have highlighted some discrepancy, as detailed below.

4.2   Qualitative Usability Evaluation

During the experimentation videos were taken from three different perspectives, which show the test person, the robot and the interaction on the tablet interface. The evaluation of the usability of the Robot-Era System is mainly based on a video-interaction analysis [21] of the experimentation.

All videos taken during the experimental loop were analyzed service by service. The analysis was conducted as an explorative content analysis [22] and was validated through multiple experts in the field of usability studies. After the selection of usability problems was made, the problems were classified and weighted with regard to international usability standards and norms (e.g. EN ISO 9241-110).

To weight a usability problem the following benchmarks were used:

– Frequency of occurrence of the problem
– Influence of the problem on fulfilling the task – Was it possible for the user to complete a task or not?
– Potential of learning from the users side.

The problems then were classified from 1 (no usability problem) to 5 (major usability problem).

Usability problems related to the shopping delivery service are:
– Difficulties in understanding the concept of creating a shopping list
– Difficulties in entering the shopping list
– Difficulties in recognizing the food icons correctly
– Difficulties in finding the items under the correct category
– Purchasing more than one item
– Confusion about the ending of the task

For the communication service, no strong usability problems can be identified during the test, even if the observation was negatively influence by the system failures. Regarding the reminding service, older people failed in changing the title of the reminder, due to the fact that it was presented by a drop-down menu, where the other options are hidden. The standard title of the reminder is "Medicine". The other options "Telephone" and "Generic Alert" are only shown when the person opens it. The person did not expect that there were other options too.

4.3   Overall Acceptability and HRI Evaluation

The results reported are based on the evaluation of the Italian sample video analysis (N = 35). For the shopping service, 12 subjects were focused mainly on the robot with the head, eyes and body orientation. In one case, the user started walking around the robot for looking better at him, moving close around DORO. At the same time, 12 subjects were more oriented to the tablet, suggesting that the use of the GUI would have distracted them from the interaction with DORO. In 5 cases, it was not possible to collect the information, for the technical failure of the system. Only one user touched the fingers of the robot, as she was looking for more interaction with it, as shown also by her statements ["I'd like the robot has a human-like arm, to be adapted to my apartment"; "The robot is very nice"]. From an analysis of the communication, the majority of the users seems to like DORO and look for more speech communication with him ["Is it possible to use a courtesy communication?", "I think that it is better if the robot has the vocal command, because if I have to stay in bed, it is difficult to take the tablet", "I think I like more the robot, while tablet is still a little complicate for me. I need to exercise more"], while no one expressed openly preferences for the tablet interface. Concerning the reminding services, the principal mean of interaction was constituted by the GUI on the tablet. From the observation, it was found that just the users with a higher technological literacy were able to perform autonomously the task – 9 out of 35 –. This has influenced the acceptability of the service, that sometimes was felt as too complicated ["The menu really does not work for me!", "I don't understand this – tablet -","It's too difficult for me!", "You should do too many things for reminding something!"]. It can be understood that the majority of the sample has looked for the step by step support from the experimenter – 14 users -. The more skilled users have given input on what can be done to ameliorate the service satisfaction ["Aesthetically, it should be more similar of the shopping list, so you can choose what to remind trough icons"] and the robot communication capability for this service ["Why DORO does not say more things? For example, he can remind me also who I have to call to"]. Even if they were more concentrated on the tablet, many users have a good HRI, as shown by the adoption of a very friendly communication to DORO ["Many thanks DORO!", "I really need you, DORO!"] and free statements on its features, especially the eyes ["DORO has a really "intelligent" eyes","DORO has a funny eyes"]. In addition, an affective reaction to the robot was detected in one user ["It seems a person, it is a companion!"].

5   Conclusion

The different number of users involved in each method represents the most relevant limit to the validity of the results. Nonetheless, some considerations can be made on the methodological point of view, that can be of usefulness for improving the future assessment.

First of all, it was observed an inconsistency between the results obtained at the shopping delivery service: even if the SUS score shows a high degree of usability of the GUI, the video interaction analysis and the HRI and acceptability observation have

depicted a different situation: users needed the support of the experimenter nearly step by step to complete the task. Concerning the communication service, even if the system failed many times due to technical issues, the users seems to really appreciate it, in contrast with the literature in the field of technology usability and acceptability [23]. Only the results obtained at the reminding service seems to be confirmed both at the quantitative and qualitative evaluation.

Despite the limit of the analysis already expressed at the beginning, it seems that a mix-method approach should be highly recommended, in order to deeply understand the results, avoiding the effect of social desirability on the answers given by the older users, for example [24].

From this first experience, it can be also observed that the choice of the appropriate method to adopt should be also guided by the end-users characteristics, in particular the technological affinity and the cultural predisposition to the task prescribed by the method. For example, the TAL method has produced not relevant results in the case of the Italian sample, maybe due to the unusual double-task for this population and the limited technological literacy that may have affected the overall use [23].

The HRI and acceptability observation has also highlighted the preference of the users in interacting more with the robot: the opportunity of using vocal command for completing the service is highly suggested for improving the overall system usability, switching the attention from the HCI to HRI.

From the statements of the users, in fact, it is interesting to notice the need of more social interaction, maybe mediated by the specific characteristics of the technology. Robots, in fact, may evoke in the users the feeling of empathy and reciprocity, following the theory of mind paradigm [25].

Due to this, a new metrics and benchmarks definition is conceived for the second loop, in which a higher attention is given to the Social Presence dimension [26], through combining different methods.

References

1. Van Den Broek, G., Cavallo, F., Wehrmann, C.: AALIANCE Ambient Assisted Living Roadmap. IOS Press, Amsterdam (2010)
2. Forlizzi, J.: Robotic products to assist the aging population. Interactions **12**(2), 16–18 (2005)
3. Forlizzi, J., DiSalvo, C., Gemperle, F.: Assistive robotics and an ecology of elders living independently in their homes. J. HCI Spec. Issue Hum.-Robot Interac. **9**(1/2), 25–59 (2004)
4. Mynatt, E.D., Essa, I., Rogers, W.: Increasing the opportunities for aging in place. In: Proceedings of the CUU 2000 Conference on Universal Usability, New York (2000)
5. Cavallo, F., Aquilano, M., Carrozza, M.C., Dario, P.: Implementation of 3D services for "Ageing well" applications: Robot-Era project. In: Proceedings of the foritaal2012 Forum, pp. 17−19. Parma, Italy, October 2012
6. Ferri, G., Mondini, A., Manzi, A., Mazzolai, B., Laschi, C., Mattoli, V., Reggente, M., Stoyanov, T., Lilienthal, A., Lettere, M.:Dustcart, a mobile robot for urban environments: experiments of pollution monitoring and mapping during autonomous navigation in urban scenarios. In: Proceedings of ICRA Workshop on Networked and Mobile Robot Olfaction in Natural, Dynamic Environments (2010)

7. Di Nuovo, A., Broz, F., Belpaeme, T., Cangelosi, A., Cavallo, F., Esposito, R., Dario, P.: A web based multi-modal interface for elderly users of the Robot-Era multi-robot services. In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 2186–2191 (2014)

8. Behnke, J., Baur, N., Behnke, N.: Empirische Methoden der Politikwissenschaft. Paderborn, UTB Schöningh (2006)

9. Schnell, R., Hill, P.B., Esser, E.: Methoden der empirischen Sozialforschung. Oldenbourg, München (2005)

10. Eurostat: Life expectancy at age 65 for Female 21,3 years is higher than the life expectancy of men which is 17,8 years at age 65 (2011). http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&plugin=extraction

11. Pfeiffer, E.: A short portable mental status questionnaire for the assessment of Organic brain deficit in elderly patients. J. Am. Geriatr. Soc. 23, 433–441 (1975)

12. Lawton, M.P., Brody, E.M.: Assessment of older people: self-maintaining and instrumental activities of daily living. The Gerontologist 9(3), 179–186 (1969)

13. Bevilacqua, R., Di Rosa, M., Felici, E., Stara, V., Barbabella, F., Rossi, L.: Towards an Impact Assessment Framework for ICT-Based Systems Supporting Older People: Making Evaluation Comprehensive Through Appropriate Concepts and Metrics (2014)

14. Sarodnick, F., Brau, H.: Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung, vol. 1. Huber Verlag, Bern (2006)

15. Ericsson, K., Simon, H.: Protocol Analysis: Verbal Reports as Data, 2nd edn. MIT Press, Boston (1993)

16. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry. Taylor and Francis, London (1996)

17. Venkatesh, V., Morris, M., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. MIS Q. 27(3), 425–428 (2003)

18. Mumm, J., Mutlu, B.: Human-robot proxemics: physical and psychological distancing in human-robot interaction. In: 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 331–338, 8–11 March 2011

19. Nedopil, C., Schauber, C., Glende, S.: Methods of User Integration for AAL Innovations – Toolbox. AAL Association (2013)

20. McLellan, S., Muddimer, A., Peres, S.C.: The effect of experience on system usability scale ratings. Usabil. Stud. 7(2), 56–67 (2012)

21. Knoblauch, H., Schnettler, B., Raab, J., Soeffner, H.G.: Video analysis: methodology and methods Qualitative Audiovisual Data Analysis in Sociology. Lang, Frankfurt am Main (2006)

22. Mayring, P.: Qualitative content analysis. In: Forum: Qualitative social research, 1(2) (2000)

23. Sun, H., Zhang, P.: The role of moderating factors in user technology acceptance. Int. J. Hum. Comput. Stud. 64, 53–78 (2006)

24. Edwards, A.L.: The Social Desirability Variable in Personality Assessment and Research. Dryden Press, New York (1957)

25. Gordon, R.: Folk psychology as simulation. Mind Lang. 1, 158–171 (1986)

26. Biocca, F., Harms, C., Gregg, J.: The networked minds measure of social presence: pilot test of the factor structure and concurrent validity. Paper presented at the Presence, 4th Annual International Workshop, May 21–23, Philadelphia, PA (2001)