Li, Andol X and Bonner, John V H

Improving control panel consistency of wizard of oz design and evaluation studies

Improving Control Panel Consistency of Wizard-of-Oz Design and Evaluation Studies

Andol X. LI and John V. H. BONNER
Live:lab, Dept. Informatics
School of Computing and Engineering, University of Huddersfield
Huddersfield, UK
x.li@hud.ac.uk and j.v.bonner@hud.ac.uk

***Abstract*—This paper investigates how a Wizard of Oz (WoZ) control panel could be developed to improve 'between-subject' consistency. To achieve this we conducted a comparative study of two control panels. Both control panels were used by the experimenter to ostensibly facilitate the design and evaluation of a novel domestic planning application allowing members of a family to coordinate a range of social arrangements and tasks. Based on video analysis and semi-formal interviews, the control panels' effectiveness as a reliable tool was assessed. Results suggested that the component-separated control panel can obviously improve operational effectiveness thus enhancing system consistency.**

Wizard of oz, control panel, operation consistency, ambient intelligence

I. Introduction

This study is a part of our research into consistency problems using Wizard of Oz (WoZ) as a design and evaluation tool for potential future domestic communication applications. Since WoZ was coined by Kelley [1], it has grown in popularity as a user-based design and evaluation tool for unproven future technologies. WoZ can provide valuable interaction, acceptance and usability feedback. However, it is difficult to provide consistency across different user evaluation studies. Furthermore, the experimenter also requires significant training to improve consistency. Studies have attempted to address this problem by controlling system variables such as task complexities and participant preference [2]. Despite this, difficulties still remain in managing situated and unexpected decision making by study participants.

Therefore in this study our aim was to improve the control panel design for the WoZ system so that we could help the experimenter to gather more consistent experimental data. To achieve this we conducted a comparative study which consisted of two different versions of control panel design. The control panels were used by the experimenter to ostensibly facilitate the design and evaluation of the novel domestic planning application which allows members of a family to coordinate a range of social arrangements and tasks. Through video analysis and semi-formal interviews the study accessed control panels' effectiveness in terms of experimenter's responding time, mistake handling and participant engagement.

The paper is structured as follows: section 2 discusses the background to WoZ as a design and evaluation tool, considerations on consistency and control panel concerns with WoZ studies; section 3 discusses assessment criteria and study procedures; section 4 analyses the experimental data; section 5 covers the implications of results in terms of handling 'system mistakes' and response effectiveness, and finally conclusions and thoughts for future work are presented.

II. Background

A. *Wizard of Oz – why using this*

Wizard of Oz (WoZ) is a light-weight HCI methodology which intercepts interaction between users and acts as an 'intelligent' device [3]. The experimenter manipulates input and output behaviour in order to deceive the subject into thinking the device possesses what levels of intelligence. Where in reality the intelligence is modelled and simulated by the experimenter rather than being programmed into the device [3]. This form of simulation provides low cost design and evaluation opportunities for smart products. It allows a wide range of interaction scenarios to be evaluated against acceptance and usability criteria without having to incur heavy development costs. WoZ method has been used for a number of applications, for example, to evaluate the use of a listening typewriter [4], like speech-controlled telephone services [5, 6] or the user of natural body movements [7] or for the selection of the most intuitive movements for vision-based game controls [7] and facial expression [8], and for conversational behaviours in handheld devices with animated characters [8, 9] and environmental sensors for human interruptibility studies [10].

However one disadvantage of WoZ as a design and evaluation tool is that it is difficult for wizards to provide consistent responses. This problem can be found across a multitude of WoZ studies, for example, in simulating speech systems [2], multimodal systems [11] and intelligent agent systems [9]. These studies have put some considerations on addressing the problem by configuring system variables such like communication modalities [2] and participant dynamics [12]. Some other studies looked for solutions from cooperative interaction designs for instance employing multiple experimenters to manipulate multimodal systems [13, 14]. For example in Neimo project the study employed three experimenters to facilitate speeches, face and mouse recognition [11]. There are also some studies which introduce automatic mechanisms for experimenter to facilitate system designs and evaluations, for example, training the wizard at well-defined tasks with 'behaviour instructions [15]' in [16].

B. Adressing inconsistency problems

Solutions proposed by these studies are at high levels that some dynamic factors are concerned such as participant preferences. However, there is a pragmatic way of looking for solutions from control panel designs via which experimenters can simulate ostensible systems. Due to the control panel is the direct tool for experimenter to facilitate the system, improving its design may effectively reduce operation inconsistency.

A few studies have looked into control panel designs for inconsistency problem. Decades ago the control panels were crude due to the system functions were simple [12]. To date with the growth of interactive media which makes multimodal user interfaces popular [11], WoZ is used for designs and evaluations of complex system. Salber and Coutaz (1993) employed multiple interfaces to address multimodal system facilitation. This solution actually introduced more experimenter dynamics into studies, and each experimenter required respective control panels with various complexities, which may deteriorate the inconsistency problem. Balbo, Coutaz and Salber (1993) suggested 'predictive models' for automatic evaluation systems to overcome the inconsistency problem which was caused by human experimenters. They proposed a WoZ platform to allow observations and automatic participant behaviour recordings. However their approach only reduced the inconsistency risks in data collection stage when interacting with multimodal interfaces for experimenter. Mavrikis and Gutierrez-Santos presented a 'tapering' approach to gradually replace a human experimenter as a computer for the design of intelligent learning environments [17]. In their iterative designs the control panel was transferred by computer programs in a gradual manner, although the limitation was that only one facilitating function was considered each time. In some speech-typing systems control panels were equipped with specific filters that inputs were intercepted by the control panel and then outputs were presented in machine-like manners [4, 5, 18]. And some other studies aimed at intuitive graphic interfaces for experimenters to manipulate the system such like gesture recognition systems [12], mixed reality systems [19] and human-robot interaction systems [15]. These WoZ studies mostly emphasised control panel usability rather than the modules and layouts to help experimenters improve the operation effectiveness and consistency.

In this regard, we proposed an empirical study to evaluate control panel designs which, probably, led to consistency improvement. We proposed two different control panel designs for experimenter to facilitate three independent system applications. By asking participants to use the novel system we gave experimenter different control panels to facilitate same experiment tasks. The comparisons between two experiments gave clues on what and why the control panel design could improve operation consistency.

C. Control panels – how should these be designed?

In current WoZ studies end-user interface designs have gained more attentions than control panel designs for experimenter use such as speech telephone service designs [5, 18]. The control panel design should have more considerations due to that it is the tool by which experimenters can facilitate the system. In multimodal systems the control panel is split into sub panels for respective experimenters, and in single modal systems the control panel is first considered on its functions as Klemmer [18] and Whittaker's [20] 'wizard interface' design. Some old-fashion control panel designs were simple, for example, Gould's (1983) listening typewriter had only one textbox for speech transcription, and Höysniemi's (1989) gesture recognition system had one simple interface for direction controls via keyboard. To date control panels are designed with multiple components such as buttons, textboxes and other graphical elements. These components have provided basic accesses to fast responses, although the delay that caused by experimenter is still noticeable as described in [7].

The control panel designs aim to associate the multifaceted and situated relations between system and experimenter [21]. Multiple roles of experimenter exist in WoZ studies, such like controller, moderator and supervisor [22]. These dynamics need to be addressed in control panel designs due to different experimenter roles require different control panels. For example, a supervision control panel has more surveillance functions yet with less control elements and a control panel has more manipulative components. Therefore in our study the control panels are designed on controlling purposes with multiple manipulative elements.

III. METHODOLOGY

Our methodology was conceived as a practical, cyclical progress in which variables were controlled and compared throughout two experiments. Each experiment went through a set of tasks involving three system applications: a domestic calendar, a communication dialogue and a cube-based media manager. The domestic calendar was a speech-recognition application designed for family arrangements planning. The communication dialogue was designed for natural language dialogues and the media manager was a simple multimedia manager. A full experiment cycle contained three action stages - planning, acting and reflecting. The planning stage related to setting up experiments, the acting stage to conducting experiments procedures and the reflecting stage to data analysis.

All experiments were video recorded by a webcam which was set beside the experiment site. Data was manually transcribed into scripts for further analysis which, was the one of most significant part in our study. A semi-formal post-experiment interview was planned. Through interviews we aimed to extract user thoughts about system performances. In this WoZ study participant could not be told the truth until all experiments were completed. However such low-level 'deception' still needed to be explained to participant for data use consents.

A. *Control panel evaluation critera*

Criteria are important for control panel design assessments. Unlike normal usability evaluations, control panels in WoZ studies are invisible to participants. To evaluate these control panels we need to measure how the system performs while being controlled by experimenter. Therefore how participant reacts to the system can affect experimenter control panel operations.

There are two ways of reflecting experimenter operations, one is how 'system mistakes' are handled and the other is how real-time responses are presented. Fraser and Gilbert [2] suggested that the wizard should take account of making some mistakes to keep faithfulness to technologies, and they also suggested a 5 percent mistaken occasions. To fulfil this target the experimenter needs to make some mistakes on purpose. For instance, taking the speech 'thunder' as 'honda' [23] and then presenting a learning progress to convince participants that the system is automatically learning on its own. Some speech misunderstandings may also contribute to mistaken occasions since the experimenter cannot be able to comprehensively understand all speeches which are close to personal lives. By analysing participant reactions toward these mistakes, we can measure how flexible the experimenter-mediated system handles mistakes through control panels.

Responding in real-time is the other important criteria which evaluates how fast the system can respond to participant. In this study these responses are all generated via control panels, thus measuring how fast the system interacts with participants can reflect the extent of operation effectiveness. The response durations can be gained from recorded videos according to time stamps.

B. *Control panels – why united and split designs?*

We designed the application system in the domain of domestic communication due to the home is an ideal site for emerging interactive technologies [24-26]. The three system applications were proposed based on daily routines in the home [27, 28], and these were integrated with mundane rhythms such like organising daily appointments and managing multimedia.

The system was programmed with C++ (MFC) in Visual Studio, therefore it had normal window elements. The system first used intranet-based communication for remote manipulation and surveillance. After several trial tests the system integrated control panels and applications in one computer due to the instable network could not afford simultaneous manipulation and video surveillance.

We proposed two versions of control panels. One was designed with united layouts which contain all modules in one panel (see Fig. 1), and the other design split the control panel into a group of sub-panels (see Fig. 2). This division was made according to system applications that each application had one separate control panel. Both control panels had the same application interfaces. The application interfaces were the same due to this makes participant easy to compare system performance changes (see Fig. 3 and Fig. 4).

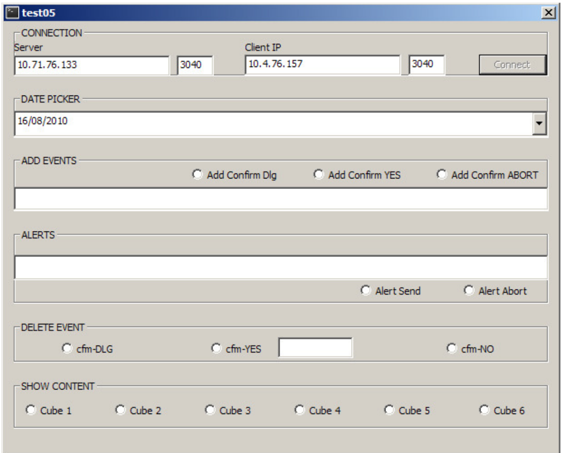

Figure 1. The compact control panel interface

C. *Procedures – how experiments are conducted*

Domestic interactions require experiment spaces which can provide home-like environments. We set up a scenario in laboratory as a domestic communication scene. It consisted of a set of sofa and a coffee table as well as multiple experimental devices such like projectors, webcams and microphones. The site separated some invisible space for experimenter to facilitate the system. Separated by big screens the experimenter could not be seen by participants, while the experimenter could still observe participant via surveillance video. The experimenter manipulated the host computer and monitored application running. The experimenter was located with the computer while application interfaces (Fig.3 and Fig. 4) were separately distributed in front of participant.

A volunteered participant was employed in our study. She had little knowledge of WoZ system but with strong interests in experiencing novel interaction styles. This is due to that experiments may be severely affected if participant becomes aware of experimenter existing. Furthermore the participant had good experiences of mundane affaires. With these conditions the participant was told that she was interacting with an intelligent computer system and her speeches could be recognised and learnt by the system. The participant needed to go through two experiments and she might be asked to compare the system performance changes in interviews.

The experimenter, who designed whole system applications as well as control panels, was playing the role of system facilitator. One advantage of this is that the experimenter does not require extra training to manipulate control panels. Meanwhile the experimenter can handle unexpected errors carefully based on system familiarity. The most important reason is that the experimenter can have first-hand experiences of control panel.

The invisible experimenter could not directly associate with participant. To address that a new role was introduced which was called 'instructor'. Her main responsibility was to deal with the participant face-to-face as an experiment moderator. The instructor could deliver some indications about system applications.

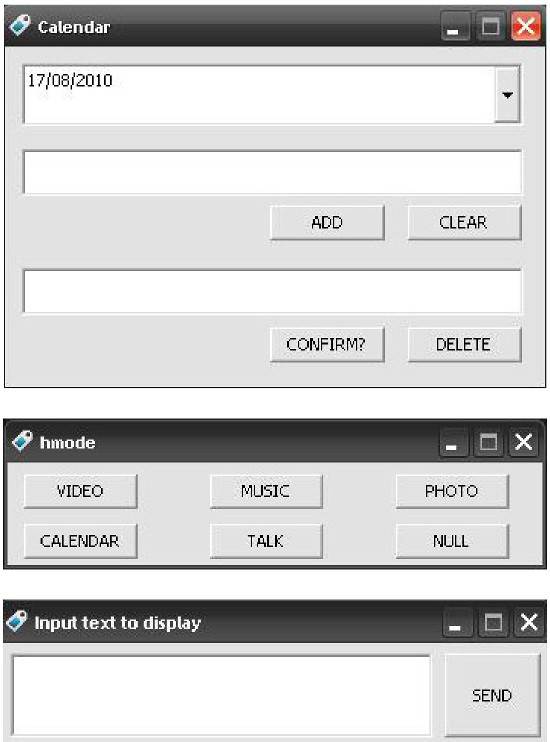

Figure 2. Split control panels: top - calendar, middle - media manager, and bottom - communication dialogue

Each experiment adopted a different control panel as described in Fig. 1 and Fig. 2. The first experiment started from system introduction by the instructor. After that the instructor gave participant a sheet in which command examples and tasks were listed. The instructor then allowed some time for participant to learn system functionalities. This was to make sure that the task completion was based on skilled manipulations, which might low the risks of dealing unfamiliar functionalities and wasting unnecessary time. Once the learning was done, participant was allowed to start tasks. All tasks should be accomplished and these included a) basic calendar operations (viewing / adding / deleting appointments), b) communicating with the system through the dialogue and c) using media manager to play videos. The media manager works with a coloured cube.

System responses were facilitated by the experimenter via control panels. When received incorrect speeches experimenter could use the dialogue to display alert messages and thus communicated with participant. While using the cube the system launched a video and played it on the coffee table. The experimenter also needed to sense the cube movements which were assigned with different operations.

After all tasks the participant was invited to a semi-formal interview which encouraged the participant to express experiences and thoughts about the system. All comments made by participant were logged in videos. These videos were manually transcribed into scripts with time stamps, therefore operation durations could be calculated and analysed.

Figure 3. The domestic calendar application interface

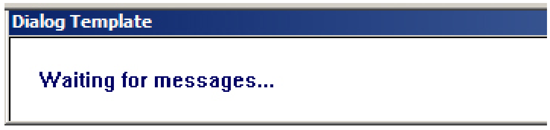

Figure 4. The communication dialogue application interface

The participant was told about the simulation system after last experiment's interview had accomplished. Then participant was encouraged to express the thoughts about experimenter facilitation. And finally participant's consents were required for further data analysis.

IV. Experiment Analysis

The experimental materials were collected from two aspects: observations and video analysis. These materials provided quantitative data in terms of response durations and mistake numbers, and also provided qualitative data which reflects how facilitation interacted with participant.

Videos were transcribed into texts by the experimenter who could recall the motivations of facilitation. These were complementary to understand participant reactions. Below is an example of scripts that demonstrates how the participant used calendar.

[00:05] Subject: today [*Speaking without hesitation and starting to wait for system responses*]

System: showing today's events

[00:08] Subject: create an event. [*Looking at the table and giving the speech when saw today's events*]

[00:10] System: popping up an input window for event contents

[*a short pause due to the participant was thinking about an appointment to add*]

[00:19] Subject: event for tomorrow and 2 o'clock

[*speaking naturally, still looking at the table and waiting for next response*]

[*A short pause due to the participant forgot to use the 'confirm' to finish the input*]

[00:22]System: popping up a message 'Confirm?'

[00:23] Subject: ye, confirm [*suddenly realising the system would not respond without 'confirm' , then giving the right command and waiting for system responses*]

As two experiments followed same procedures the analysis compared some phenomenons in terms of real-time responses and 'system mistake' handling.

In the first united control panel experiment, the average response time was approximate 5 seconds, and the average speech typing duration was about 6 seconds. The speech typing was observed as the most time-consuming part in this experiment. The observation also showed that participant became inpatient in some occasions for example system appeared out of response. When participant added a new appointment 'Go Shopping' the experimenter typed the speeches without responding to other inputs. In this case participant might repeat the speeches. However using the colour cube to manipulate videos did not have participant repeat speeches, and the communication dialogue neither.

In this experiment the experimenter facilitated some deliberate system mistakes. For example, in adding an appointment the speech 'Go Shopping' was mistaken as 'Go For Shopping'; and in another case the system popped a prompt of 'Speech not recognised, please repeat slowly' when participant spoke, after a slower speech repetition the system again popped the reminding. Participant was observed to follow the system prompts.

In the second split control panel experiment the overall response time was reduced to about 3 seconds, and the average speech typing durations were still as high as 5 seconds. The split control panel was not observed to improve the typing speed while it noticeably reduced overall response time. Throughout the statistic we noticed that the massive time reduction of click operations contributed to the overall response improvement.

In this experiment same mistakes were avoided and some new mistakes were proposed. For example the system prompted '13th July' instead of participant's speech '30th July'. In this case participant repeated correct speeches until the system recognised them.

In both experiments there were response differences across system applications. It was observed that the colour cube manipulation had fastest responses and the domestic calendar had slowest responses. Meanwhile participant's attitudes towards these applications were also different that communication dialogue gained most tolerances of slowness and the cube manipulation gained least. Through interviews the participant validated these observations.

V. IMPLICATIONS AND DISCUSSIONS

Experiment comparisons explicitly illustrate the effectiveness differences between two control panels. Due to all experiment facilities were the same, the effectiveness improvement is deemed as a result of control panel design changes.

The reason of the split control panel having higher response effectiveness may be twofold, first the control panel can be rearranged according to participant preferences, and second the split may make control panel more intuitive. The rearrangement reduces unnecessary operations, for example moving cursor from left top to right bottom over the control panel, and also it allows participant to create an environment that fits better. The intuitive control panel may lowen time waste on looking for functions. A complicated layout requires extra endeavours to locate a facilitating function. This may be even difficult when facing a control panel with massive boxes and buttons, such like Whittaker's (2002) wizard interface designs.

However there is no strong evidence showing which control panel handles system mistakes better. The durations of generating and responding to system mistakes are similar in two experiments. One of reasons may be the same control panel elements through which experimenter can only facilitate limited mistakes. The only discovery is drawn upon system application differences in terms of user tolerance. Speech typing consumed most time while clickable elements (such like buttons and checkboxes) showed great potential to shorten response time in the second experiment. Therefore introducing clicking-style elements into control panel may help to handle system mistakes more consistently due to these elements provide fixed operations in fast response.

VI. CONCLUSIONS

This paper presented a comparative study between two control panel designs to investigate how they should be developed for inconsistency problems in WoZ studies. The study built on control panel designs, and assessed operation inconsistency changes in terms of real-time responses and system mistake handling.

Based on video analysis and interviews, the study suggests that the control panel design can intuitively affect experimenter and thus affect system operation inconsistency. To address this issue, designs should consider the types of elements and manners of layouts. From this study it also shows that clicking-style elements are suitable for system mistake handling due to these provide limited yet fast responses, and separately grouped control panels are helpful for system responses as these help participant locate facilitating functions efficiently.

Despite this, we need further considerations on facilitating personal information-related intelligent systems in the future. This study merely considers control panels which receive participants' data; future control panels should be designed capably to collect user information autonomously. Based on this, future studies will investigate some other principles for control panel designs with which experimenters can facilitate autonomous applications for domestic communication studies.

REFERENCES

[1] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," 1, ACM, 1984, pp. 26-41.

[2] N. M. Fraser, and G. N. Gilbert, "Simulating speech systems," Computer Speech and Language, vol. 5, pp. 81-99, 1991.

[3] A. X. Li, and J. Bonner, "Developing smart domestic applications using a wizard of oz methodology," in 2nd International Workshop on Human-Centric Interfaces for Ambient Intelligence, Nottingham, 2011.

[4] J. D. Gould, J. Conti, and T. Hovanyecz, "Composing letters with a simulated listening typewriter," 4, ACM, 1983, pp. 295-308.

[5] I. Bretan, A.-L. Ereback, C. MacDermid *et al.*, "Simulation-based dialogue design for speech-controlled telephone services," in Conference companion on Human factors in computing systems, Denver, Colorado, United States, 1995.

[6] N. Dahlback, A. Jonsson, and L. Ahrenberg, "Wizard of Oz studies: why and how," in Proceedings of the 1st international conference on Intelligent user interfaces, Orlando, Florida, United States, 1993.

[7] J. Hoysniemi, P. Hamalainen, and L. Turkki, "Wizard of Oz prototyping of computer vision based action games for children," in Proceedings of the 2004 conference on Interaction design and children: building a community, Maryland, 2004.

[8] T. Bickmore, "Towards the design of multimodal interfaces for handheld conversational characters," in CHI '02 extended abstracts on Human factors in computing systems, Minneapolis, Minnesota, USA, 2002.

[9] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through Wizard of Oz," in Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, Amsterdam, The Netherlands, 1993.

[10] S. Hudson, J. Fogarty, C. Atkeson *et al.*, "Predicting human interruptibility with sensors: a Wizard of Oz feasibility study," in Proceedings of the SIGCHI conference on Human factors in computing systems, Ft. Lauderdale, Florida, USA, 2003.

[11] D. Salber, and J. Coutaz, "Applying the Wizard of Oz technique to the study of multimodal systems," Human-Computer Interaction, pp. 219-230, Berlin/Heidelberg: Springer Berlin/Heidelberg, 1993.

[12] A. G. Hauptmann, "Speech and gestures for graphic image manipulation," in Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind, 1989.

[13] L. Molin, "Wizard-of-Oz prototyping for co-operative interaction design of graphical user interfaces," in Proceedings of the third Nordic conference on Human-computer interaction, Tampere, Finland, 2004.

[14] S. Balbo, J. Coutaz, and D. Salber, "Towards automatic evaluation of multimodal user interfaces," in Proceedings of the 1st international conference on Intelligent user interfaces, Orlando, Florida, United States, 1993.

[15] Y. XU, S. TAKEDA, and T. NISHIDA, "A WOZ Environment for Studying Mutual Adaptive Behaviors in Gesture-based Human-robot Interaction." pp. 40-46. in AAAI 2007, Human Implications of Human-Robot Interaction Workshop, Vancouver, British Columbia, Canada, 2007.

[16] Y. Xu, K. Ueda, T. Komatsu *et al.*, "WOZ experiments for understanding mutual adaptation," AI & Society, vol. 23, no. 2, pp. 201-212, 2009.

[17] M. Mavrikis, and S. Gutierrez-Santos, "Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering," Computers & Education, vol. 54, no. 3, pp. 641-651, April 2010, 2010.

[18] S. R. Klemmer, A. K. Sinha, J. Chen *et al.*, "Suede: a Wizard of Oz prototyping tool for speech user interfaces," in Proceedings of the 13th annual ACM symposium on User interface software and technology, San Diego, California, United States, 2000.

[19] S. Dow, J. Lee, C. Oezbek *et al.*, "Wizard of oz interfaces for mixed reality applications," in CHI 2005, 2005.

[20] S. Whittaker, M. Walker, and J. Moore, "Fish or fowl: A wizard of oz evaluation of dialogue strategies in the restaurant domain," in Language Resources and Evaluation Conference, 2002.

[21] A. Voss, M. Hartswood, R. Procter *et al.*, "Introduction: configuring user-designer relations: Interdisciplinary perspectives," Configuring user-designer relations, A. Voss, M. Hartswood, R. Procter *et al.*, eds.: Springer, 2009.

[22] S. Dow, B. MacIntyre, J. Lee *et al.*, "Wizard of Oz Support throughout an Iterative Design Process," 4, IEEE Educational Activities Department, 2005, pp. 18-26.

[23] A. X. Li, and J. Bonner, "Enhancing social relationships through human-like intelligences." in The 9th International Workshop on Social intelligence Design (SID2010), Egham, UK, 2010.

[24] A. F. Blackwell, J. A. Rode, and E. F. Toye, "How do we program the home? Gender, attention investment, and the psychology of programming at home," International Journal of Human-Computer studies, vol. 10, no. 1016, 2008.09.11, 2008.

[25] J. Hughes, J. O'Brein, and T. Rodden, "Understanding Technology in Domestic Environments: Lessons for Cooperative Buildings " Cooperative Buildings: Integrating Information, Organization, and Architecture, pp. 248-261: Springer Berlin / Heidelberg, 2007.

[26] R. Harper, *Inside the smart home*, New York: Springer-Verlag, 2003.

[27] A. S. Taylor, R. Harper, L. Swan *et al.*, "Homes that make us smart," Pers Ubiquit Comput, 2007.

[28] A. Crabtree, and T. Rodden, "Domestic routines and design for the home," Computer Supported Cooperative Work, vol. 13, pp. 191-220, 2004.