Natural multimodal interaction for design applications

Els DEN OS[1], Lou BOVES[2]
[1]*Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, 65xx YY, Netherlands*
Tel: +31 24 352 1333, Fax: + 31 24 352 1XXX, Email: E.denOs@mpi.nl
[2]*Radboud University Nijmegen, Erasmusplein 1, Nijmegen, 6525 HT, Netherlands*
Tel: +31 24 361 2902, Fax: + 31 24 361 2907, Email: L.Boves@let.kun.nl

Abstract:

One of the open issues in multimodal interaction is the antagonism between proponents of the direct manipulation and the communication agent metaphors. In this paper we describe an experiment in which we tested a system based on the communication agent metaphor for a bathroom design application where direct manipulation is possible in principle, but perhaps not very attractive in terms of usability.

We found that non-expert users have substantial difficulties interacting with the direct manipulation systems, due to the fact that they failed to discover the mental model applied by the application designers. This problem did not occur in the communication agent application, but here we observed the need for substantial improvements of the recognition technology needed to build conversational interfaces.

1. Introduction

One of the important open issues in multimodal interaction is the question whether the direct manipulation or the communication agent metaphor should be preferred. Some authors argue that direct manipulation is always best [1], while others provide evidence in favour of the conversational agent metaphor [2]. However, it is quite likely that the users' preference for the interaction metaphor depends strongly on their knowledge of the application domain and the functionality of the interface. In [3] it is shown that users do not appreciate the guidance of a conversational agent in completing the query form for a timetable information system. However, the authors suggest that the help and guidance that an agent can offer will be appreciated if users need to accomplish a task that they perform seldom, and that addresses a domain where they lack detailed technical and procedural knowledge. Architectural design –instantiated here in the form of bathroom design application- is probably a good example. Most customers buy a new bathroom only once or twice in their lifetime. Yet, designing a new bathroom requires substantial knowledge about existing options for tiles and sanitary ware, as well as of guidelines for how to arrange sanitary and select designs that go together well.

In principle, a task such as bathroom design can be implemented both in the form of direct manipulation and a conversational agent. In the COMIC project[1] we are working on the implementation of a conversational agent system for bathroom design [4]. Some companies have launched competing solutions based on the direct manipulation approach.

[1] http://www.comic.mpi.nl

In this paper we focus on our experiences with the design, implementation and test of the conversational agent approach to bathroom design. In doing so, we will address a number of technical issues that have a large impact on the users' appreciation. In addition, we will report on our observations of a number of subjects who used direct manipulation applications for the same task, as a stepping stone for a larger comparative analysis of the two approaches that is presently being carried out.

The information provided in this paper should be of immediate interest for scientists in the area of multimodal interaction for *e*Commerce and *e*Work services. In addition, it should help managers and investors to better assess the opportunities and risks involved in developing multimodal applications.

2. Objectives

The first step in bathroom (re-)decoration is to input the shape and dimensions of the room, and the location and dimensions of doors and windows. The best way to do this is by specifying a blueprint of the room, adorned with some annotation (for example for the height of window sills). In existing commercial software packages this information must be entered by means of drag and drop actions, combined with keyboard input. Figure 1 shows a snapshot of a screen as it may appear in such an application.

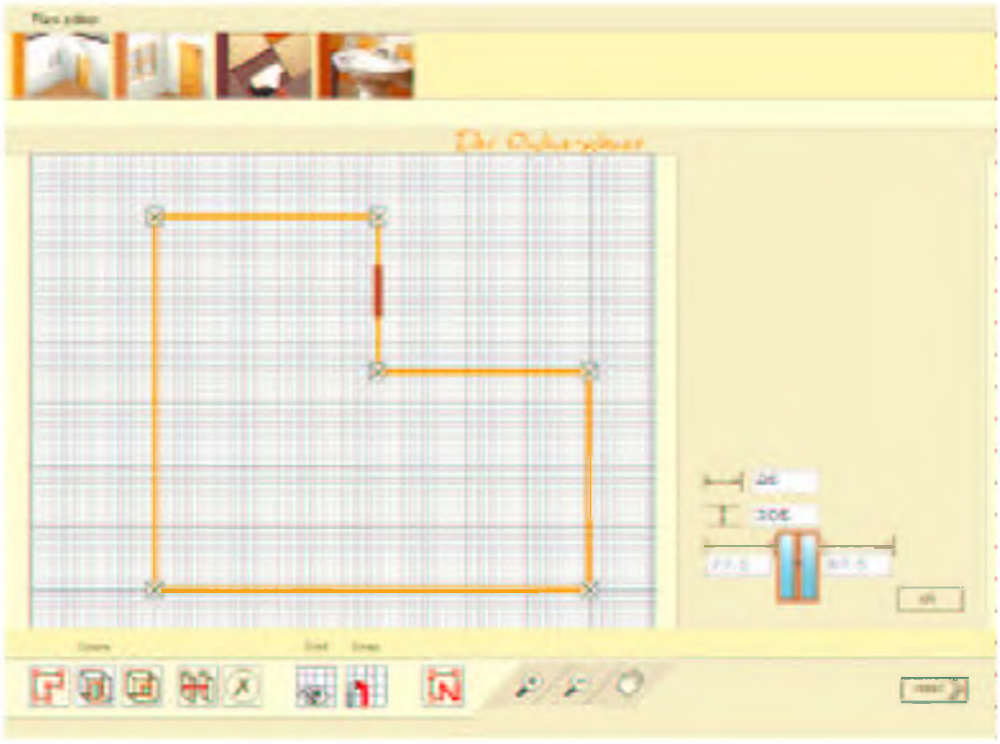

Figure 1: Snapshot of the direct manipulation tool. The outline of a room is shown, together with the location and the dimensions of a door. Icons at the bottom indicate objects. The digit fields at the right must be used to indicate dimensions.

Although expert sales persons can learn to operate this type of software packages in a training course, one must doubt whether non-expert customers are able to use this kind of applications, for example as a self-help tool to get a global impression of the possibilities for (re-)decorating their bathroom within given budget constraints. To investigate this, we observed six highly educated subjects while they attempted to enter a pre-specified

blueprint with the ViSoft Websolution, a prototype GUI application for bathroom design and the Saniweb tool[2], an application that can be accessed by the general public.

The primary goal of this paper is to report on the results of an experiment that was designed to investigate the task completion and user satisfaction of the conversational system for bathroom design that is under development in COMIC. In doing so, we need to pay special attention to the performance of some of the modules that are responsible for the recognition and interpretation of the users' input, since it has been shown repeatedly that this performance has a major impact on user satisfaction [5].

3. Methodology

The usability evaluation of the ViSoft Websolution was performed in the form of participating observation. The authors observed six colleagues while they tried to input a pre-specified blueprint with the two GUI applications. During and after the interaction the beliefs and opinions of the users were recorded in the form of unstructured question-answer conversations. The results of these interviews are presented in section 5.

The conversational agent system that we are building in the COMIC project uses pen and speech for input, and two graphical screens and a loudspeaker for output. One screen is used to display an avatar that interacts with the user by means of synthetic speech and by head gestures (thinking, listening) as back channel information while the user is speaking or drawing. The second screen is actually a tablet computer, which renders the ink produced by the subjects, as well as the interpretation that the system made of the user's input. Figure 2 shows a user interacting with the system, wearing a headset and head mounted microphone.

Figure 2: The COMIC system is shown on the right. The left screen shows the flow of active modules for demonstration purposes.

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

To evaluate the user appreciation of our system, we designed an experiment in which 28 subjects (20 females and 8 males) entered three blueprints, that of their private bathroom, that of their parents' and that of a friend or relative. The age of the subjects ranged between 20 and 37. All subjects were students or scientists associated with the University of Nijmegen or the Max Planck Institute for Psycholinguistics in Nijmegen. Thus, all subjects belong to the group the highest possible education level. After completing the three tasks, subjects filled out a questionnaire comprising 26 Likert scales, designed to measure their appreciation of the system. Examples of statements that we used are "It was easy to use the system", "I always understood what the system expected me to do", "I would be happy to use this system a next time", etc. Subjects had to indicate whether they totally disagreed (scale value 1), disagreed (scale value 2), neither agreed nor disagreed (scale value 3), agreed (scale value 4), or totally agreed (scale value 5) with the statements.

During the experiment all data pertaining to the interaction between the subjects and the system were logged and stored in a database. After each experiment the interaction data were annotated by hand. In this manner we added information about the recognition performance of the pen and speech recognizer, the Natural Language Processing and the Media Fusion modules to the database. Using the combined subjective and objective data of the 28 subjects, we conducted statistical analyses to search for interesting relations between the behaviour of subjects in response to the behaviour of the system.

4. Technology Description

The bathroom design system tested in this experiment has essentially the same functionality as the GUI system that it will eventually supersede: users can draw the outline of the room, specify the lengths of the walls with centimetre precision, and indicate the position and dimensions of the doors and windows. A subsequent version will also allow users to add sanitary ware to the blueprint, but this was not implemented in the tested version.

Pen input was acquired by means of a Wacom CintiQ 15X graphical LCD tablet. The pen input recognition software was developed at the NICI [6]. The recogniser is able to automatically classify pen input as deictic gestures (point-and-click), sketching or handwriting. This interpretation is helped by the expectation that is generated by the Dialog and Action Manager: after a request from the system to draw the shape of the room, it is reasonable to expect drawing, while pen input is much more likely to be handwriting after a system prompt to specify the dimensions of the room.

Speech input used a head mounted noise cancelling microphone. The speech recognition system is based on the open source Hidden Markov Toolkit (HTK). Both pen and speech input were fed into a Natural Language Processing module that is responsible for the semantic interpretation of the individual input channels. A Fusion module combined information from the two parallel input channels. During the experiment the Dialog Manager was partly simulated in the form of a Wizard-of-Oz design, but at the time of this writing is has been replaced by a fully implemented software module. Feedback of the system to the user was in the form of stylised graphical output. Finally, system guidance was provided by means of spoken utterances.

4.1 Multimodal Turn Taking

One of the important issues that came up during the implementation of the COMIC system was the concept of multimodal turn taking. The concept of turn taking originates from the theory of conversations between two or more persons. In an 'ideal' conversation at any one time only one person is speaking. It is also assumes that speakers convey clear signals to inform their interlocutors that a turn has ended, so that another person can take the floor.

Moreover, it is assumed that the meaning of what a speaker just said when (s)he stops talking can be determined from the verbal contents of the message. In face-to-face interaction, where visual signals are exchanged in addition to verbal information, the simple concept of dialog turns does not hold. Visual and verbal information do not need to be synchronised. For example, a speaker can assume a posture that indicates an imminent end of turn well before the speech message is complete. But if two persons are talking about a bathroom design, and support the information exchange by sketching and pointing, it is quite possible that sketching continues after a spoken utterance is completed, and that the sketch is essential for interpreting the meaning of the verbal information [7].

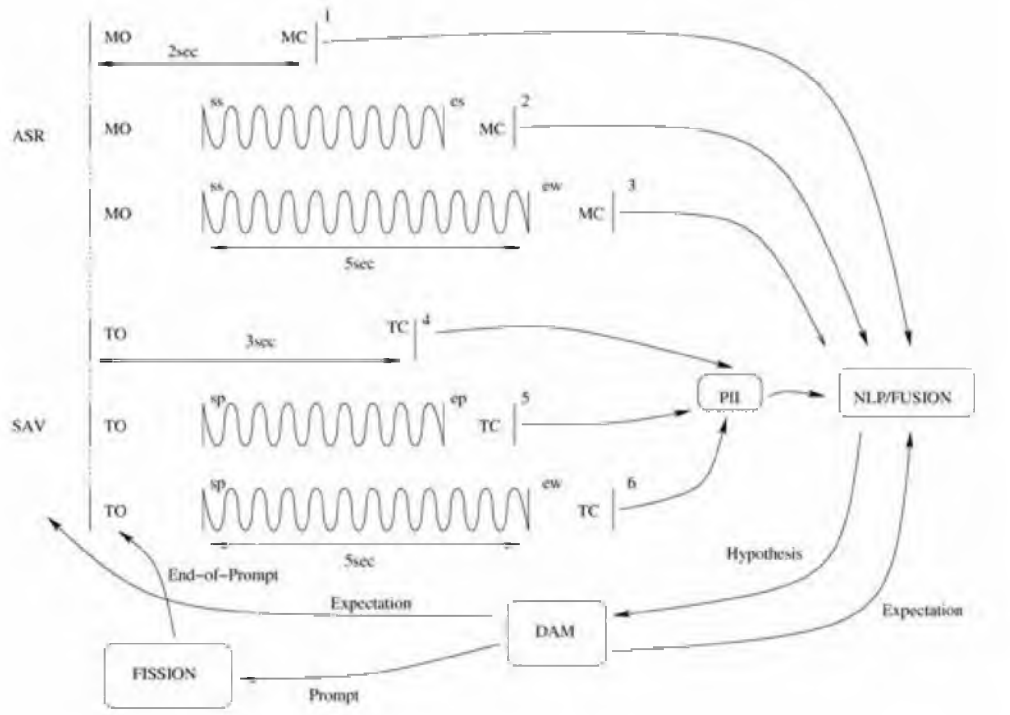

Figure 3. Definition of multimodal turns and end-of-turn synchronisation. Time is running from left to right. The leftmost edge represents the end of a system prompt. MO: Microphone Open; MC: Microphone Closed TO: Tablet input Open; TC Tablet input Closed

Presently, there is no computational theory of multimodal turn taking. But to build a multimodal interaction system a rudimentary version of such a theory must be formulated and implemented. In the COMIC system under test in this paper we opted for a definition of turns in which the verbal and gestural input channels are synchronised (cf. Figure 3). This enables the system to determine the end-of-turn and to make a definitive interpretation of the inputs. Strict synchronisation, as implemented in COMIC, has the advantage that it turns input processing and interpretation into a process that is manageable. However, it has the disadvantage that users must learn that speech or gestures produced after the end-of-turn detected by the system will get lost. Presently, we are working towards a less rigid definition of multimodal turn taking, in which the system still starts processing the input received up to a moment when both pen and speech contain a relatively long pause, but where speech and pen input produced after the pause are captured and stored for processing in the next user turn. However, the system's reaction to the first turn will not reflect the information provided after the pause that triggered the interpretation process. We are

planning experiments to investigate under what conditions this behaviour that is different from what we expect in human-human interaction may confuse users.

5. Results

5.1 The GUI applications

The main observation we made form observing six users trying to enter a blueprint is that they all started with a different mental model of how the application expected them to input the data (walls with measures, doors with measures, and windows with measures). These mental models never matched the "default mental model" of the system. This caused a large number of repair actions that sometimes resulted in a perfect copy of the blueprint but most often did not. One user started drawing the room without looking at the right measures (she thought that she could adjust these later, which turned out to be rather complicated). Another user started inputting measures of the walls without first drawing them, and again another user was able to input all data, but was not able to position the window at the correctly. Only one subject discovered how she could change the width and height of a window in the easiest way. None of the users read the instructions; all just started to input data and some of the subjects tried to recover the instructions when they got stuck. This is a well known phenomenon for almost all applications.

It also appeared that the experience gained with one GUI application had a negative effect on the other, because they were based on different choices for the interaction protocol: one used the 'select object, then indicate action' protocol, while the other used 'select action, then select object' order. On the other hand, once the system had been explained to the users, they learned very fast how to use this system in an efficient way.

It was also evident that users found it very difficult, if not impossible, to get a good idea of the design options that were available for sanitary ware and tiles. This is due to the fact that browsing on the basis of thumbnail pictures is a very difficult task for users who do not know how to interpret tiny pictures. Adding short explanations to the thumbnails in pop-up windows does not solve this problem, because more often than not users are not able to understand the meaning of crucial terms in the explanations.

5.2 The Conversational Agent application

Twenty-eight subjects participated in the user test with the COMIC system. This may be considered as a medium-scale user test for multimodal systems. The data logs obtained in multimodal user tests are usually very large (Mbytes) and difficult to analyse automatically. This is due to the fact that not all events that happen during a dialog can be classified and annotated automatically. For example, the performance of the speech and pen input recognisers can only be established by having a person compare the acoustic and visual input with the recognisers' output. All user data have been analysed by an analysis tool that was developed for this project [8]. In addition, the answers to the questionnaire have been analysed.

Log data analysis showed that, similar to the direct manipulation application, there was a learning effect: users were faster when inputting the third bathroom than when inputting the first bathroom. Mean turn duration decreased from 12.3s for the first bathroom to 11.5s for the third bathroom. The total number of turns (35) did not change. From the log analysis it became clear that recognition performance for handwriting and speech was far from perfect. Handwriting recognition was 83% correct, and speech recognition was correct in only 56% of all utterances. This is due to the fact that users have no way to know the types

of input the system can handle (cf. [5]). The lack of application specific training data, especially for speech recognition, is responsible for the low recognition performance. Our user test provided valuable data that will be used to train the recogniser for this specific task.

We observed that most users showed the tendency to continue using the modality they started with, and only switched to the alternative modality after repeated recognition errors. Thus, when users used speech for inputting the measures of the walls, they often only switched to handwriting after three or more misrecognitions. Similar observations have also been reported by [9]. Apparently, changing modality requires conscious cognitive activity, which users are not able or willing to perform easily. This means that also for multimodal systems speech and gesture recognition performance should be high, and that one should not capitalise on the hope that one modality naturally acts as a back-up if another fails. It is also evident that users do not easily discover the most efficient and affective manner of using a mix of modalities.

The answers to the questionnaire made clear that the users found it easy to form a mental model of the system. The instruction was clear (mean scale value 4.2) and most subjects understood what they were supposed to do while using the system (mean scale value 3.5). They liked that they could combine pen and speech (mean scale value 4.1), and the feedback by the system was very clear to the users (mean scale value 4.3). However, they were not happy with the recognition performance and did not consider the system to be efficient (mean scale value2.5). There was a high correlation between the objective performance of the system and most subjective judgements.

From the answers to the questionnaires it can be deduced that the turn taking protocol that we implemented did not confuse the users, despite its difference from multimodal turn taking in human-human interaction. At the same time, however, an analysis of the timing of user and system turns suggests that this protocol has contributed to the impression that the system is quite rigid, and needs improvement.

6. Business Benefits

This user study suggests that at least for this type of design applications, an interface using a multimodal communication agent metaphor will provide user-friendly and intuitive user interfaces. Since users can more easily form a mental model of the application, this type of interface solves the problems that occur with the direct manipulation application in which users have to struggle to find out how to use mouse and keyboard. Moreover, the conversational agent metaphor makes it much more natural and easy to provide guidance and explanation to non-expert users during an interaction, compared to consulting on-line help systems when problems occur in interacting with a direct manipulation interface to an application that assumes that users have a considerable amount of domain knowledge. Intuitive interfaces are of course very important for offereing eInclusion to the highest possible proportion of European citizens. However, this study also made clear that the basic technologies, like pen and gesture recognition, need to be improved and should be made more robust against all types of unexpected user behaviour.

7. Conclusions

Until now, few user studies with multimodal systems have been reported. The most important result of our experiments is that non-expert subjects are able to use the COMIC system to accomplish tasks that could not easily be completed with web-based direct manipulation applications. The multimodal communication agent metaphor makes it easier to form an adequate mental modal of the application. In addition, users like to combine pen

and speech, because for them this is the most intuitive way to communicate design issues. In addition, the results of the user studies provide clear guidelines for further improvements of the recognition technology. Last but not least, the analysis of the interactions between users and the COMIC system suggests that the concept of turn taking in multimodal interaction needs to be revised and redefined.

The need for intuitive user interfaces is huge, given the aim of Europe to include as many citizens as possible into the world of electronic data exchange. However, many more user studies have to be performed with multimodal communication agent applications to fully understand how these interfaces can best be designed and deployed. In the final year of the COMIC IST project, we will perform additional user studies that should give more insights into the benefits and needs of this type of interface.

Acknowledgement

We thank all members the COMIC team for their contributions to the work reported in this paper. The COMIC is funded by the EU, under contract IST-2001-32311.
Special thanks are due to ViSoft GmbH, one of the partners in COMIC, who made their experimental Websolution available for the research reported in this paper.

References

[1] Shneiderman, B. and Maes, P. (1997), "Direct Manipulation vs. Interface Agents, excerpts from the debates at IUI 97 and CHI 97" In: *Interactions of the ACM*, 4 (6), pp. 42-61.

[2] McGlashan, S. (1995), "Speech Interfaces to Virtual Reality" In: *Proc. of 2nd International Workshop on Military Applications of Synthetic Environments and Virtual Reality.*

[3] Sturm, J., Boves, L., Cranen, B. and Terken, J. (submitted) Direct manipulation or conversational agent: what is the best metaphor for multimodal form-filling interfaces? *Human Computer Interaction Journal.*

[4] den Os, E. and Boves, L. (2003) Towards Ambient Intelligence: Multimodal Computers that Understand Our Intentions. *Proceedings eChallenges 2003.*

[5] Kvale, K., Rugelbak, J. and Amdal, I. (2003) How do non-expert users exploit simultaneous inputs in multimodal interaction?, *Proc. International Symposium on Human Factors in Telecommunication*, Berlin, 1-4. December 2003.

[6] Van Erp, M., Vuurpijl, L.G., & Schomaker, L.R.B. (2002). An overview and comparison of voting methods for pattern recognition. *Proc. of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR.8)* pp. 195-200.

[7] Lee, J. (1999) Words and pictures -- Goodman revisited. In: R. Paton and I. Neilson (Eds.) *Visual Representations and Interpretations*, London: Springer Verlag, pp. 21-31.

[8] Vuurpijl, L et al. (2004) Evaluation of multimodal dialog systems. Workshop on MMCorpora, in press

[9] Sturm, J. and L. Boves (submitted) Effective error recovery strategies for multimodal form-filling applications. *Speech Communication.*