



## Conversational agent or direct manipulation in human–system interaction

Els den Os <sup>a</sup>, Lou Boves <sup>b,\*</sup>, Stéphane Rossignol <sup>c</sup>,  
Louis ten Bosch <sup>b</sup>, Louis Vuurpijl <sup>c</sup>

<sup>a</sup> *Max Planck Institute for Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands*

<sup>b</sup> *Centre for Language and Speech Technology, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands*

<sup>c</sup> *Nijmegen Institute for Cognition and Information, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands*

Received 10 December 2004; received in revised form 1 April 2005; accepted 5 April 2005

---

### Abstract

In this paper we investigate the usability of speech-centric multimodal interaction by comparing two systems that support the same unfamiliar task, viz. bathroom design. One version implements a conversational agent (CA) metaphor, while the alternative one is based on direct manipulation (DM). Twenty subjects, 10 males and 10 females, none of whom had recent experience with bathroom (re-)design completed the same task with both systems. After each task we collected objective measures (task completion time, task completion rate, number of actions performed, speech and pen recognition errors) and subjective measures in the form of Likert Scale ratings.

We found that the task completion rate for the CA system is higher than for the DM system. Nevertheless, subjects did not agree on their preference for one of the systems: those subjects who were able to use the DM system effectively preferred that system, mainly because it was faster for them, and they felt more in control.

We conclude that for multimodal CA systems to become widely accepted substantial improvements in system architecture and in the performance of almost all individual modules are needed.

© 2005 Elsevier B.V. All rights reserved.

*Keywords:* Multimodal interaction; Usability; Conversational agent; Direct manipulation

---

### 1. Introduction

It has long been believed that speech would bring a breakthrough in human–system interaction, but it has become increasingly clear that the fact that speech is the most natural medium for

---

\* Corresponding author. Tel.: +31 24 361 2902; fax: +31 24 361 2907.

E-mail address: [l.boves@let.ru.nl](mailto:l.boves@let.ru.nl) (L. Boves).

human–human communication does not necessarily imply that it is also the best medium for human–system interaction (Boves and den Os, 1999). During the last couple of years the focus of speech-based interaction research has shifted towards multimodal interaction (Oviatt, 2003). However, it is all but certain that making applications multimodal and speech centric is the best way for solving interaction problems.

One of the open issues in human–system 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 (Shneiderman and Maes, 1997), while others provide evidence in favour of the conversational agent metaphor (McGlashan, 1995). Proponents of direct manipulation (DM) emphasise the importance that users attach to the feeling that they are always in control. Proponents of the conversational agent (CA) metaphor object that it is not clear how users could feel in control if they do not fully understand the application they are trying to use. There is increasing evidence that people accept their own mistakes, as well as mistakes of a system, as long as they understand why the system misinterpreted their request (for example because it took the wrong sense of a polysemic word). Therefore, it is quite likely that the users' preference for the interaction metaphor depends strongly on their knowledge of the application domain (which determines whether they may need help) and the functionality of the interface (especially its intelligence in dealing with lack of knowledge and uncertainty at the user side). For example, in (Sturm et al., [accepted for publication](#)) it is shown that users do not appreciate the guidance of a conversational agent in completing the query form for a timetable information system, most probably because they do not need any help, while spoken instructions slow down the interaction. 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, in a domain where they lack detailed technical and procedural knowledge.

Until now, few studies have been reported in which non-expert subjects try to use multimodal

applications. The main reason is that there are still very few operational multimodal systems in existence. Although the lack of fully operational systems can be sidestepped by reverting to Wizard\_of\_Oz techniques, we have argued that this approach may lead to conclusions that cannot very well be generalized (den Os and Boves, 2005). Most multimodal systems are research systems that demonstrate that the technology can be useful if the user knows the system sufficiently well to stay within the confines of the functionality. It takes a lot of work and effort to tune a research system for doing user experiments. The little usability research with non-expert users in multimodal interaction that has been carried out in the past is based on relatively simple applications, such as route finding or time table information. In (Kvale et al., 2003) a more complex multimodal map-based application is evaluated with naïve users. From these evaluations it becomes clear that if the application is well known or relatively simple and therefore fast to learn, the putative benefits of multimodal conversational agents are not appreciated or completely understood by non-expert users. Rather, one must expect to see a bias in favour of DM, because subjects are unlikely to need support in using the application. Thus, there are very few, if any, results for a comparison of DM and CA interaction styles for non-expert subjects who use a semi-professional service in a field that is most probably not familiar to them. Our research is meant to fill this void. As an example of a task that naïve users perform seldom, but of which they still have a global mental model, we have chosen architectural design, instantiated in the form of a bathroom design application. Most people buy a new bathroom only once or twice in their life, so it is unlikely that randomly chosen subjects have fresh experience with software to support the task. 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. At the same time it holds that virtually all subjects have a global knowledge of how bathrooms look like, and what they like and dislike. In principle, a task such as bathroom design can be implemented both in the

form of direct manipulation and a conversational agent. In the FP5 project COMIC (<http://www.hcrc.ed.ac.uk/comic/>) we are working on the implementation of a conversational agent system for bathroom design (den Os and Boves, 2003). At the same time we see emerging on-line services for bathroom design based on the direct manipulation approach.

In this user study we compare a conversational agent system and a direct manipulation system for bathroom design on a number of usability issues for non-expert users. We are not aiming at optimising the application per se. Rather, the results of this study should lead to the formulation of guidelines for improving the design and the implementation conversational agent systems in general. The first system is the conversational agent system that is developed in the COMIC project (called the CA system from now on), the other system is a direct manipulation system developed by one of the partners in the COMIC project, viz. the SME ViSoft, which is available on the web for ViSoft customers (called the DM system in the remainder of this paper).

From previous studies we have learned that there is a very large difference between subjects in the way they use multimodal systems. So far, it has been difficult to relate differences between individual subjects to group characteristics, although recently it has been suggested that the sex of the subjects may play a significant role (Buisine et al., 2004). Therefore, we decided to make subject sex a factor in the design of our experiment.

In Section 2 of this paper we will explain the design of the experiment in more detail. To that end, we first describe the characteristics of the two systems that are most important for our usability evaluation. We also describe the subjective and objective measures that we obtained, and we explain why we focus on subjective measures. In Section 3 we present the actual data that we collected in the experiment, and the results of statistical analyses of these data. In Section 4 we present a discussion on the results and we formulate guidelines for developing improved CA systems.

## 2. Method

### 2.1. The systems

The first step in bathroom (re-)design is to input the shape and dimensions of the room, and the location and dimensions of doors and windows. Especially for the doors it must also be indicated how they open, because that determines how the floor space adjacent to the doors can be used. The input results in a machine readable blueprint of the room, adorned with some additional annotation (for example for the height of window sills). In existing commercial software packages (all of which implement DM interfaces) this information must be entered by means of drawing and drag and drop actions, combined with keyboard input.

#### 2.1.1. The CA system

In the COMIC project we have designed and implemented a multimodal system for bathroom design that is suitable for usability evaluations with non-expert users. The version of the system that was used in this study is definitely not the final one. In the present version we paid much attention to robustness of speech and pen input recognition, while the interaction design and user interface are only in the first cycles of iterative design and optimisation. In fact, one of the goals of the experiment was to obtain guidelines for improving the interaction design and the interface.

The complete bathroom design task in the COMIC system consists of four phases. In the first phase, users enter the shape and dimensions of the bathroom, including the position of the doors and windows (if any). In the second phase they can decide what sanitary ware goes where in the room. In the third phase they select tiles and decoration, while the fourth phase consists of a 3D tour of the newly designed and furnished room. The complete system is described in (Herzog et al., 2003). The underlying architecture and implementation platform (called MULTIPLATFORM system) is adapted from the German VerbMobil and SmartKom projects (Wahlster, 2003). The user study presented in this paper concentrates on the evaluation of phase one.

In phase one, users can use pen and speech to input the requested items (walls, measures, doors, windows). A naturalistic talking head gives instructions and some back channel information (e.g. thinking, agreement); the ‘agent’ also asks for clarification if the speech and/or pen input could not be interpreted. Last but not least, the gaze and facial expression of the ‘agent’ were meant to provide additional cues for turn taking. If the system could make sense of the speech and pen input, the recognised values were verified by displaying them on the tablet screen in the form of ‘beautified’ renderings: printed characters for text and measures, straight lines for walls, etc. The dialogue is in English and it is system driven, i.e., the system gives detailed instructions for what to do next. In interactions with the type of system under development, two kinds of errors can be distinguished, viz. mistakes made by the users, and recognition errors committed by the system. Users were told that they could correct errors, either by saying “erase this”, or by using the pen (pressing a button on the pen and tapping on the item one wants to erase) each time the system displayed a recognition result on the tablet screen. However, the erase function could only be applied to the last item that was entered. The functionality of the system was also limited in another respect, that was relevant for the present study: it is not possible to indicate the exact position of doors and windows relative to the corners of the room. However, it was possible to indicate exactly how the door opens.

The system has been tuned for the user evaluation by running a large number of pilot tests with naïve subjects. By doing this we were able to repair a large number of system bugs, and we tuned the speech and pen input recogniser. See Fig. 1 for the configuration of the CA system.

### 2.1.2. Turn taking

Multimodal turn taking is an essential aspect of multimodal conversational agent systems. Today, the most elaborate computational model of multimodal turn taking is probably the Ymir Turn Taking Model (YTTM) (Thórisson, 2002). YTTM implements perception-action loops on several parallel layers. The lowest layer, which essentially only



Fig. 1. The COMIC system for bathroom design.

senses action without attempting a semantic interpretation, makes it possible for a system to show immediately that it has noticed some user action and that it is now paying attention. Higher layers, which do involve semantic interpretation and reasoning before an action command can be issued, still impose some delay. However, the YTTM architecture is not compatible with MULTIPLATFORM architecture on which the COMIC system is built. Therefore, we were obliged to define and implement a more conventional turn taking protocol, similar to what has previously been used in multimodal interaction projects like SmartKom. In these systems multimodal turn taking relies on the synchronization of information entering through the parallel input signals in a Fusion module. For the experiment reported here, we implemented a straightforward timing-based end-of-turn detection protocol, which is illustrated in Fig. 2. An End\_of\_Speech input event can be triggered by three conditions: if the user does not start speaking within 2 s after the end of the system prompt, if the ASR system detects a silence exceeding 500 ms, or if the

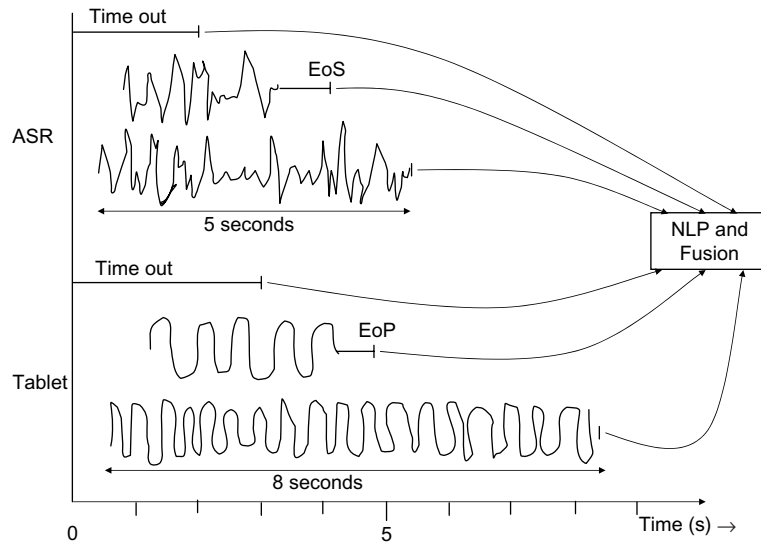


Fig. 2. Definition of multimodal turns and end-of-turn synchronization. Time is on the  $X$ -axis.  $T = 0$  represents the end of a system prompt; microphone and pen tablet open. ASR: Automatic speech recognition. Tablet: Pen input device; EoS: End\_of\_Speech Detector fires; EoP: End\_of\_PenInput Detector fires. The NLP and Fusion module combines and synchronizes the end\_of\_input signals.

input speech exceeds a duration of 5 s. The event is issued by the condition that comes first. End\_of\_Pen input events are determined in the same manner, with the only difference that a user is given a time interval of 3 s until the first pen action, and 8 s to complete an input.

Chopping up potentially continuous input into ‘turns’ enables the Natural Language Processing and Fusion module to make a definitive interpretation of the inputs. Moreover, strict synchronisation, as implemented in COMIC as well as other speech-plus-pen based multimodal interaction systems (e.g. Kvale et al., 2003), has the advantage that input processing and interpretation become manageable with limited computational power. 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.

One inevitable but undesirable side effect of the system architecture that COMIC inherited from SmartKom and from the software architecture of the individual modules is that it is bound to cause considerable response latencies. Although the MULTIPLATFORM system supports asynchronous parallel processing, except for the ASR

module, no other input module was capable of incremental processing. The Pen Input Recogniser, Natural Language Processing and Fusion could only start processing after a complete turn was available. Thus, NLP could only start processing after both pen and speech inputs had detected an end\_of\_input. Reducing the latencies that accumulate in a system where modules cannot do incremental processing is a major Human Factors issue. However, we are convinced that the system that we had available was good enough for the main objective of this study, i.e. to compare a DM and CA system for the task at hand and to derive guidelines for improving the conversational behaviour of the CA system.

### 2.1.3. The DM system

The direct manipulation system is a web system that ViSoft offers to its customers, who are dealers of tiles and sanitary ware. It is not available for the general public. The intended users are experts in design applications who should not need specific instructions how to deal with the DM system. Since our subjects are non-expert users, we decided to provide some introduction on how to deal with

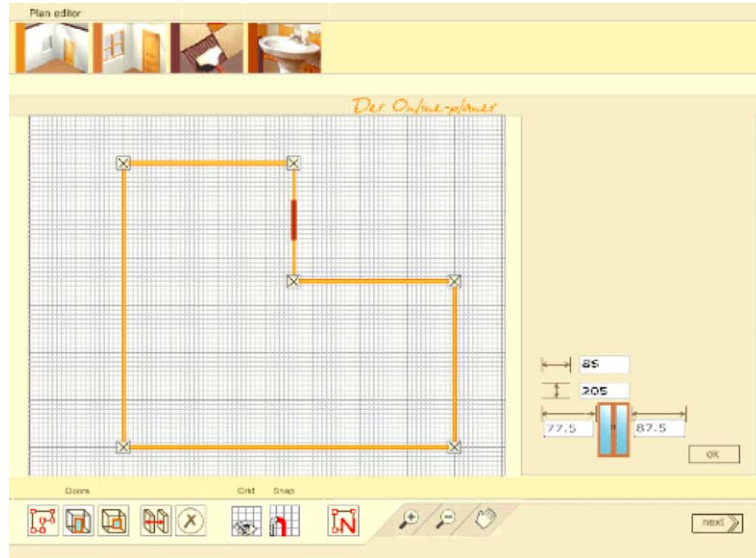


Fig. 3. Snapshot of the direct manipulation tool.

this application. Subjects were told that they first had to get the shape and dimensions right, and that only then they could place the door and the window. This should be done by first tapping on the relevant icon, followed by adjusting the measures, and finally by placing the object at the correct location in a wall. Fig. 3 shows a snapshot of a screen as it may appear in the DM system. Users had to find out that the measures of the room were presented in a menu window at the same time as one was drawing the walls. This menu window appeared next to the grid, at the right hand side of the screen. The functionality of this system has another limitation than the CA system: it is not possible to indicate the way the door opens.

After the shape and dimensions of the room have been entered, the user can proceed to the second phase, in which sanitary ware can be selected and positioned. However, in this experiment we only tested phase one.

## 2.2. Design of the evaluation

### 2.2.1. Non-expert users

Ten male and ten female non-native English non-expert users participated in this study. Their

ages range between 22 and 59 years (mean age was 33 years). The users were not paid for their participation. Test sessions lasted between 35 and 55 min. The educational level of all subjects is high (academic level) and their knowledge of English is very good. We opted for a within-subject design, in which all subjects tested both systems. Half of the female and male users started with the CA system, the other half with the DM system. Before the test started, the test leader checked whether the subjects understood task specific words like “window sill”. All subjects spend more than four hours on a computer every day. They consider their computer experience as advanced or expert, and their reported programming experience ranges from beginner to expert.

### 2.2.2. Task

Subjects had to imagine that they were in the process of re-designing their bathroom, and that they were visiting a busy bathroom store in which two systems were available that could help them in the design process while waiting for a salesperson to become available. The users were asked to use both systems for *exactly* copying the same blueprint of what was supposed to be their bathroom.

The blueprint consists of a rectangular room of 2.5 by 3 m. Exactly in the middle of one wall of 3 m is a door that opens to the inside and which is 85 cm wide; in the middle of the other wall of 3 m is a window that is 100 cm wide, 75 cm high, and with a window sill 120 cm from the floor. It was explained that both systems have more or less the same (but not identical) functionality, but that the way to use them is rather different. Given the fact that the functionality of both systems is not *exactly* the same, users were told not to worry if they could not find the way to input certain data exactly in the way they were specified in the example blueprint, and just to stop when they thought something is not possible.

Since we wanted to investigate the situation in which users were confronted for the first time with these systems (which is a realistic “real-life” situation), we did not offer any practice time. In this study we were not interested in any learning effects. Rather, we wanted to investigate whether the two systems could be used effectively ‘out of the box’. The impact of omitting a training session on the two systems is difficult to predict. The CA system may suffer from the fact that none of the subjects had prior experience with fully multimodal pen-speech systems, whereas some subjects did know DM-style web design applications. At the same time it may turn out that the CA has an advantage because it guides inexperienced users through the task, while they are left on their own with the DM system. Because of the help and guidance provided by the prompts of the CA system, persistent recognition errors are the most likely cause for failing to complete the task. In case of recognition errors, users could erase these by voice or pen and they could switch between input modes. The major cause why users might fail to complete the task with the DM system is that they did not manage to find out how the system worked. Subjects were told that when they felt frustrated because they were unable to find out how things worked, they could stop. For both systems these working conditions have obvious ecological validity: customers in a shop will only make so many attempts to correct recognition errors, or to try different ways to use a DM interface for an unfamiliar task.

After the users had finished with a system, the test leader discussed with them the result. For the CA system, the result was still visible on the tablet. For the DM system, the experimenter played back the recorded mouse and keyboard events. After this discussion the users were asked to fill in a questionnaire.

### 2.2.3. Questionnaire

One questionnaire was designed that was used for both systems, so that a clear comparison between both systems could be made. For the larger part of the questionnaire Likert scales were used. This scaling instrument was adapted from [Sturm and Boves \(submitted for publication\)](#) and more indirectly from [Love et al. \(1994\)](#) and [Larsen \(2003\)](#). Subjects had to indicate whether they completely disagreed (1), disagreed (2), were neutral (3), agreed (4), or completely agreed (5) with 34 statements. These statements concerned the working of the system, the ease of use, the controllability, and the general acceptance and appreciation. Next to these Likert scales, six open questions were asked that addressed the experienced duration, the easiest, the hardest, the unexpected things, the possible improvements, and general comments. One final question asked for the preference for a system.

### 2.2.4. Objective analysis

All data were logged by both systems. For the COMIC system, we used the built-in logging facilities; for the Web system we used Camtasia, a programme that records all mouse movements and mouse clicks, as well as the keyboard input. Given the big differences between both systems that make that durational measures for sub tasks rather meaningless, we decided to restrict the objective measures to six yes/no observations (walls present?, wall measures OK?, door present?, door features OK?, window present?, window measures OK?) and durations of the complete session. A task completion measure of 6 points means that the session ended in fully correct items. If a user was not able to get anything right, zero points were given.

### 3. Results

#### 3.1. Order and sex

First we performed an independent *t*-test (2-tailed), and Anova to find out whether an order effect and/or sex effect could be observed. It turned out that neither of these independent variables had a significant effect. Therefore, we will concentrate on the differences between both systems in the rest of this paper.

Due to the large differences between the two systems, we had not expected a significant order effect: there was very little, if anything, that subjects could learn from the first system that would facilitate their using the second.

We did not find the gender effect reported by Buisine et al. (2004) either. It is not completely clear why that is so. Perhaps, gender effects in human–system interaction are dependent on the details of the task. In addition, a small gender effect would have been obscured by the large differences between the individual subjects.

#### 3.2. Objective measures

In Table 1, mean durations of complete interactions (time to task completion) are presented. The users were also asked to estimate durations of an interaction. Mean *estimated* durations are also presented in Table 1. To see whether users systematically over- or underestimate interaction durations, we also calculated the difference between

actual and estimated duration. The mean difference is also given in Table 1. Finally, a measure of task completion rate is given; 6 points means that all information items (walls, door, window) are present and that also all measures or features are correct. A paired *t*-test showed that neither of these measures differed significantly between the two systems. This leads us to the conclusion that for the blueprint used in this study neither time to task completion nor task completion rate differed between the CA and the DM system. Again, the lack of significance may be attributed at least to some extent to the large differences between the subjects.

It can be observed from the mean values in Table 1 that in general it took much longer than the minimum duration to input the information. This is especially true for the DM system, where the minimum duration for an expert user is about 0.8 min. Here we also observe relatively large standard deviations, indicating large differences between users. The fastest user needed 1.2 min, the slowest one needed 14.7 min to input the blueprint. Also for the CA system, users needed more time than the minimum of 3.8 min. The mean number of turns amounts to 21.7 (with a range between 15 and 31). The minimum number of turns (if no corrections are necessary) is 15.

It is interesting to note that—while our subjects needed more time to complete the task than what would be theoretically possible—the excess time was larger for the DM system than for the CA counterpart. This is due to the time it took our subjects to figure out how to use the DM system without initial guidance and training. Apparently, DM systems are not fully transparent, especially when it comes to unfamiliar tasks. This can be taken as a strong argument in favour of CA systems, especially in the case of applications that users need only occasionally, so that it becomes difficult to remember what one learned in the previous confrontation. Excess time in the CA system was mainly due to correcting recognition errors.

We observed that a large proportion of the errors in pen input were due to the fact that subjects were given a fixed amount of time to complete their input. This window of 8 s turned out to be too short for some of the subjects. This is a direct

Table 1  
Objective duration (min), estimated duration (min), difference between both durations (min), and task completion (min = 0, max = 6) ( $N = 20$ )

	System	Mean duration (min)	SD
Objective duration	CA	6.5	1.8
	DM	7.1	3.4
Estimated duration	CA	8.2	3.6
	DM	8.4	5.7
Difference between both durations	CA	−1.7	3
	DM	−1.3	3.5
Task Completion	CA	5.1	.8
	DM	4.3	1.8

(but unexpected) consequence of the turn taking protocol explained in Fig. 2. It was clear that some users did not understand that gestures produced after the end of the window started at the detection of the first pen activity were not processed. This problem was aggravated by the way the electronic ink on the tablet screen was handled: all electronic ink remained visible on the screen. We did not encounter many speech recognition errors because of utterances that were truncated at the beginning or the end. This is, at least in part, due to the turn taking cues provided by the talking head. Also, the application did not elicit speech utterances long enough to be truncated at the end. Most recognition errors were due to out-of-grammar utterances, hesitations and disfluencies.

From the difference between duration and estimated duration we can learn that users tend to overestimate the duration of an interaction session. This is true for both systems, which is somewhat surprising. In (Sturm and Boves, submitted for publication) we found that subjects overestimate the time it takes to complete the query form in a tap-to-talk multimodal train timetable application, but not the time needed to complete that form in a graphical user interface (GUI). This difference was explained by assuming that subjects were misled by the need to wait for the result of the Automatic Speech Recognition module after each spoken input, while in the GUI they were responsible themselves for the way they spent their time. The most likely explanation for the fact that subjects also overestimated the time to task completion in the bathroom design experiment is that they weighed the time spent in trial and error attempts to understand the DM interface too heavily. Again, this suggests that there is a usability advantage for CA interaction in unfamiliar applications.

The mean figures of task completion show that not all users were able to input all items correctly. This also holds for both systems. For the CA system most wrong results are due to recognition errors that were not corrected by the users. For the DM system most mistakes are due to the fact that users were not able to get the measures of the room right. They did not notice that the measures appeared in the menu window left of the grid.

### 3.2.1. Mode switch

We analysed whether users switch to the alternative mode in case of recognition errors. Four of the 20 users completed their input in less than 18 turns. For these users no mode switches were observed. This is not surprising, since almost invariably mode switches are triggered by persistent recognition errors (Xiao et al., 2003; den Os and Boves, 2005). Of the remaining 16 users, 8 users did switch modality and 8 users did not. The mean number of turns was the same (23) for the switchers and the non-switchers.

Our results show that mode switching behaviour is more complex, and especially more user-specific than one would have expected on the basis of experiments in which subjects have to cope with repeated errors in Wizard\_of\_Oz settings. In our experiment, where subjects were left free to select the input mode, it appeared that some subjects very early on decided that either speech or pen was not going to work for them. These subjects naturally did not switch mode in case of recognition errors. The other half of the subjects who encountered a sufficient number of errors to allow for a meaningful interpretation did switch modes. Therefore, the contribution that multimodal input can make to error correction is difficult to predict. It seems to depend on subject idiosyncrasies; it may also depend on details of the interaction design (den Os and Boves, 2005).

### 3.3. Subjective measures

In Table 2, we present the mean scale values and standard deviations for the 34 Likert scale questions. Also here paired *t*-tests (2-tailed) were performed, with the threshold for significance  $p < 0.05$ . With 34 independent *t*-tests this threshold is very lenient, so that significant differences are only interesting if they can be explained by means of reference to mental and cognitive models of the task, or if all differences point in the same direction.

Given the fact that both systems were not exactly the same, question 12 had to be stated slightly differently for both systems. The results show that for about half of the statements the users did not differ significantly in their opinions

Table 2

Likert scale values (I completely disagree = 1 and I completely agree = 5); means, SD, and sig. level ( $N = 20$ , values with  $p < 0.05$  are marked with an \*)

Question	Mean scale value and SD		Sig.
	CA system	DM system	
1. I was able to use the system successfully	3.80 (.951)	3.05 (1.432)	.061
2. I was able to input the walls of the bathroom	4.40 (.681)	4.10 (.968)	.301
3. I was able to input the <i>measures</i> of the walls of the bathroom	3.70 (1.081)	3.10 (1.714)	.186
4. I was able to place the door	4.15 (.745)	4.00 (1.414)	.679
5. I was able to place the window	4.30 (.923)	4.80 (1.436)	.096
6. It was clear what I had to do to input the walls	4.45 (.999)	3.50 (1.051)	.016*
7. It was clear what I had to do to input the door	4.25 (.910)	3.45 (1.276)	.049*
8. It was clear what I had to do to input the window	4.40 (.598)	3.50 (1.192)	.004*
9. It was clear what I had to do when something went wrong with the walls	3.85 (.933)	2.05 (.933)	.000*
10. It was clear what I had to do when something went wrong with the window	3.40 (.995)	2.50 (1.000)	.007*
11. It was clear what I had to do when something went wrong with the door	3.20 (.894)	2.55 (1.234)	.120
12. CA: It became clear to me that I could not input the measures of the wall next to the door and the window DM: It became clear to me that I could not input the way the door opens	3.10 (1.071)	3.15 (1.387)	.895
13. It was easy to use the system	3.35 (1.089)	2.75 (1.118)	.124
14. The system worked the way I expected it to	3.35 (.933)	2.45 (1.146)	.009*
15. I knew what I could do at each point	3.70 (1.081)	2.40 (1.142)	.001*
16. The system responded quickly to my requests	2.85 (1.040)	3.85 (1.040)	.008*
17. It was clear what to do when the system misunderstood me	3.80 (.951)	2.25 (.786)	.000*
18. I found the system to be cooperative	3.45 (.999)	2.65 (1.226)	.035*
19. I felt confused when using the system	2.50 (1.000)	3.30 (1.174)	.025*
20. I found the system to be flexible	2.75 (.910)	2.55 (.686)	.464
21. I felt in control when using the system	2.70 (.979)	2.50 (1.147)	.585
22. I found the system complicated to use	2.20 (.834)	2.80 (1.240)	.124
23. I felt frustrated when using the system	2.50 (1.000)	3.25 (1.020)	.044*
24. I found the system engaging	3.40 (.598)	2.80 (.894)	.030*
25. I found it exciting to interact with the system	3.60 (.940)	2.80 (.894)	.014*
26. I was so involved in the interaction that I lost track of time	2.70 (1.031)	2.85 (1.089)	.673
27. I felt tense when using the system	2.65 (1.040)	2.50 (1.000)	.651
28. I really had to concentrate on the system	2.75 (1.118)	3.05 (1.146)	.437
29. I found the system boring	2.25 (.910)	2.75 (.639)	.066
30. I liked using the system	3.40 (.883)	2.80 (.894)	.055
31. I found the system to be friendly	3.50 (1.051)	2.50 (.761)	.010*
32. I found the system to be knowledgeable	2.85 (.988)	2.65 (.813)	.447
33. The system appeared to be intelligent	3.10 (1.165)	2.45 (.945)	.033*
34. I would like to use the system again	3.40 (1.046)	3.05 (1.146)	.320

about both systems. However, a comparison of the mean scale values shows that overall the CA is evaluated as more positively than the DM system (note that higher scale values indicate more positive opinions for most questions. For the questions 19, 22, 23, 27, 28, and 29 a lower measure indicates a more positive opinion). For 16 statements users

have a significantly different opinion about the two systems. For all of these, except statement 16, users rate the CA as more positively. For statement 16 (“The system responded quickly to my requests”) the CA system was rated less positively than the DM system, a results that is easy to explain as a result of the latencies caused by the turn

taking protocol and the pipelined information flow, due to which the processing latencies of all modules accumulate. When we look at the statements that are rated more positively for the CA system, it becomes clear that almost all statements refer to the same underlying phenomenon, namely the transparency of the working of the system. The CA system gives a clearer idea of how to input items and what to do (statements 6, 7, 8, and 15), of how to handle when something goes wrong (statements 9, 10, and 17), and the CA system is more predictable (statement 14). The CA system is also rated as more friendly and more intelligent and people found it more exciting to use it than the DM system. However, when we look at the mean values we have to conclude that the users are not so positive about the CA in an absolute sense: e.g. they think that the system is not so knowledgeable (statement 31), flexible (statement 20), or controllable (statement 21); here the scale values are below 3.

We also asked the users for their preference for a system by means of the statement: “I preferred the pen-speech system above the web system”. The mean scale value amounts to 3.3, indicating that overall the subjects were neutral: some users preferred the CA system, others the DM system, and some had no preference at all. Female users preferred the CA more (mean scale value 3.6) than male users (mean scale value 2.9), but this difference is not significant. In Table 3 we present an overview of the preference for a system, and the task completion measure for each system. It turned out the correlation between task completion rate and preference for a system was significant ( $p < 0.01$ ). Thus, it appears that preference is related to the effectiveness of the applications.

### 3.4. Open questions

We asked the users what they found hardest and easiest to accomplish, and what they thought was unexpected behaviour of the systems. We also asked what type of improvements should be made for each system. Here we concentrate on the most salient responses related to the CA system. For the CA system nine users explicitly stated that the system should be made faster than it is now. The

Table 3

For 20 users we present gender, preference for the pen-speech system above the web system (5 = I fully agree), and task completion measures for both systems (6 = all items correct, 1 = only 1 item correct)

User number	Gender	Preference CA above DM	Task completion CA	Task completion DM
1	F	5	4	0
2	M	5	6	3
3	F	4	6	5
4	F	4	6	3
5	F	4	5	5
6	M	4	6	4
7	F	5	4	5
8	F	5	6	1
9	M	2	4	6
10	F	2	6	5
11	F	5	5	2
12	M	4	5	5
13	M	3	4	6
14	M	2	5	6
15	M	2	6	6
16	M	2	5	6
17	F	1	5	6
18	M	2	5	5
19	M	3	4	2
20	F	1	4	5

slowness is not only due to latencies in the system, but also to the relatively large number of turns, caused by the fact that many atomic information elements are prompted for independently. Other problems that were repeatedly mentioned were recognition errors (in pen as well as speech recognition), and the difficulty to repair misunderstandings. Users also complained about the fact that the electronic ink remained on the screen. This caused confusion with pen input: users were misled into thinking that they could complete inputs that they did not finish writing in the previous turn. Obviously, this is a design error in the graphical part of the multimodal interface of the CA system.

From the comments of the subjects it appeared that the talking face was not always observed, despite the fact that it was in the same visual field as the tablet (cf. Fig. 1). Some users never looked at the face, others did pay attention and they reported that they used facial behaviour for getting feedback on whether the system was busy or not and on whether input was understood (then the

face nods). This is in accordance with results of other research into human–human interaction: if humans are talking about some object that is visibly present and must be manipulated, they seldom have gaze contact, and the moments when they do look at each other cannot be predicted from the type and contents of the speech acts (Argyle and Cook, 1976). Nevertheless, in human–system interaction (possibly exaggerated) gaze behaviour of an avatar can help to smooth turn taking.

The main reason why some subjects preferred the DM system was that they felt more in control when using this system. The main reason for preferring the CA system was that users thought it more natural, and also error correction was clearer than in the DM system.

#### **4. Discussion and guidelines for future research in conversational interaction**

The effectiveness in dealing with a system correlates with the preference for a system. This is in line with the findings in (Walker et al., 2000). The background of the users (their experience in using certain interfaces and interactions) determines to a large extent how effective a system is for this person, and therewith their preference. For the near future this means that it will not be easy to develop good user tests for evaluating multimodal conversational systems. As long as there are no “commercial” multimodal systems around, users will not get acquainted to this type of interaction and will find it difficult (and perhaps even less ‘natural’) to deal with these systems (although they are relatively transparent). A solution would be to run user tests over a longer period of time to see whether users will change their behaviour and their opinions. However, this would defeat what we think is the most appealing advantage of CA interaction, viz. that CA systems can be used without instruction and training.

The overall picture we get is that especially users who have no experience in using drawing software programmes have a more positive opinion about the CA system than about the DM system. We base this conclusion on the ratings of the statements related to transparency (i.e. the ease of

inputting data and correcting errors). However, given the mean scale values the users do not have clearly positive opinions about either system. For the DM system the less than positive assessment is probably due to the fact that it was not designed for use without at least a minimal amount of prior instruction. At the same time, this shows that the DM interaction metaphor per se is not sufficient to overcome conceptual problems that users may have with an application or an interface. The mediocre (at best) ratings of the CA system can be explained by technical shortcomings, some of which were known a priori (the latencies, the fact that neither speech nor pen recognition performed at close to 100% accuracy, combined with restricted error handling functionality) while others (especially the way in which the electronic ink was treated, and the unexpected end-of-turn truncation problem in pen input) became apparent during the user test. The latter problems occurred despite intensive testing with uninformed subjects prior to the experiment.

Our results suggest that multimodal conversational systems certainly have potential for the future. Unfortunately, the design of our experiment makes it impossible to determine to what extent the promise is held by the ‘conversational’ or by the ‘multimodal’ part of the system. In applications like architectural design, where any interaction design that would not offer the combination of pen and speech would feel unnatural, the ‘multimodal’ bit is probably essential. But it is also likely that few persons would be able to complete a complex design task without the assistance of a conversational agent. As we have pointed out above, our subjects tended to prefer the CA system over its DM counterpart, mainly because it is more transparent to the user and because it provides valuable assistance with an unfamiliar task. However, many things need to be improved and to be further developed, especially for the type of complex multimodal interaction addressed in the bathroom design application. When we concentrate on the CA system and especially on the issues that need urgent improvement, we see that the system is not fast enough, that too many pen and speech recognition errors occur, that it is not always clear to users that a turn has ended, and that the way in

which electronic ink was handled is confusing for users.

#### 4.1. *Improvement of speed*

The system can be made faster in two different ways: Firstly, the modules themselves and the interaction between the modules may be made faster. However, the accumulation of latencies in a multi-modular system can only be avoided if all modules are able to process their input incrementally. This requires a change in attitude and approach in most sub-fields of Natural Language Processing and Dialogue Systems. The requirement of incremental processing also holds for the modules that are responsible for generating output. To accomplish really transparent multimodal interaction it is necessary to move towards an architecture similar to the Ymir Turn Taking Method (Thórisson, 2002). However, it is questionable whether such an architecture can be implemented with currently available inter-module communication platforms like the MULTIPLAT-FORM system.

Secondly, the number of turns may be reduced by offering the possibility to combine elements in the input (drawing a wall and providing measures at the same time) and by making the system more intelligent (“this room is rectangular and it is 3 by 4 m”). In this way the user will have more freedom to input the data, will need fewer turns (gaining time through a smaller number of system prompts) and will get a feeling of being more in control.

#### 4.2. *Improvement of recognition performance*

It is clear that recognition performance for speech as well as pen input need to be improved beyond the performance in the system that we tested. In our user study we ‘helped’ the recognisers, because we knew what could be inputted (the blueprint was known). This was necessary for us to be able to focus on multimodal interaction, instead of multimodal error correction. If the system must be made usable for naive users who may input their own bathroom, much effort will have to be paid to collecting domain specific data that are necessary for tuning the application.

#### 4.3. *Better turn-taking protocol*

In the system tested in this study we implemented a rather simple turn taking protocol. The effect of this turn taking protocol was that some users did not understand that the system stopped processing their input after the end of a fixed time window. Some users were rather slow in drawing and/or writing, probably because they had no experience in using an electronic pen. They were still writing on the tablet, although it was already closed. They did not notice that their turn was over, and that the system had to cope with unfinished input. Often the system came back with a message that was unclear to the users. In a later version of the system a more intelligent end-of-turn detection protocol was implemented, but that was still based on the assumptions underlying half-duplex interaction, to avoid the need for incremental processing throughout the system. Therefore, it improved the confusion about turn taking only slightly. A more fundamental solution can only be expected from a leap towards full-duplex interaction. The Ymir turn taking method is a promising step in that direction.

### 5. Conclusion

In this paper we reported on a comparison of conversational agent and direct manipulation as interface metaphors in an unfamiliar task. We have shown that if subjects need help with completing their task, there is an advantage for the CA interaction style. At the same time we have pointed out a number of technology issues that need substantial improvement if CA interfaces are ever going to be successful for services and applications that are used on a regular basis.

### Acknowledgements

The research reported in this paper was conducted as part of the FP5 project COMIC, contract number IST-2001-32311. Special thanks are due to ViSoft GmbH who made their experimental Web solution available for the research reported in this deliverable, and to Jan Peter de Ruiter of the

Max Planck Institute for Psycholinguistics for assistance with the statistical analyses.

## References

- Argyle, M., Cook, M., 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge.
- Boves, L., den Os, E., 1999. Applications of speech technology: designing for usability. In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Co, 11–15 December 1999.
- Buisine, S., Abrilian, S., Martin, J.C., 2004. Evaluation of multimodal behaviour of Embodied Agents. In: Ruttkey, Zs., Pelachaud, C. (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers.
- den Os, E., Boves, L., 2003. Towards ambient intelligence: multimodal computers that understand our intentions. In: *Proceedings Challenges 2003*.
- den Os, E., Boves, L., 2005. User behaviour in multimodal interaction. In: *Proceedings of HCI International 2005*.
- Herzog, G., Kirchmann, H., Merten, S., Ndiaye, A., Poller, P., Becker, T., 2003. MULTIPLATFORM Testbed: An integration platform for multimodal dialog systems. In: *Proceedings, HLT-NAACL 2003 Workshop: Software Engineering and Architecture of Language Technology Systems (SEALTS)*, Edmonton, Alberta.
- Kvale, K., Rugelbak, J., Amdal, I., 2003. How do non-expert users exploit simultaneous inputs in multimodal interaction? *Proc. Internat. Symposium on Human Factors in Telecommunication*, Berlin, 1–4 December 2003.
- Larsen, B.L., 2003. Assessment of spoken dialogue system usability: what are we really measuring? In: *Proc. European Conf. on Speech Communication and Technology (Eurospeech'03)*.
- Love, S., Dutton, R.T., Foster, J.C., Jack, M., Stentiford, F.W.M., 1994. Identifying salient usability attributes for automated telephone services. In: *Proc. Internat. Conf. on Spoken Language Processing (ICSLP94)*, pp. 1307–1310.
- McGlashan, S., 1995. Speech interfaces to virtual reality. In: *Proc. 2nd Internat. Workshop on Military Applications of Synthetic Environments and Virtual Reality*.
- Oviatt, S.L., 2003. Multimodal interfaces. In: Jacko, J., Sears, A. (Eds.), *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence, Mahwah, NJ, pp. 286–304.
- Shneiderman, B., Maes, P., 1997. Direct manipulation vs. interface agents, excerpts from the debates at IUI 97 and CHI 97. *Interact. ACM* 4 (6), 42–61.
- Sturm, J., Boves, L., submitted for publication. Mobile access to information services: pen, speech or both? *Internat. J. Speech Technol.*
- Sturm, J., Boves, L., Cranen, B., Terken, J., accepted for publication. Direct manipulation or conversational agent: what is the best metaphor for multimodal form-filling interfaces? *Human–Computer Interact. J.*
- Thórisson, K., 2002. Natural turn-taking needs no manual: computational theory and model from perception to action. In: Granström, B., House, D., Karlsson, I. (Eds.), *Multimodality in Language and Speech Systems*. Kluwer Academic, Dordrecht, pp. 173–207.
- Wahlster, W., 2003. SmartKom: symmetric multimodality in an adaptive and reusable dialogue shell. In: Krahl, R., Günther, D. (Eds.), *Proceedings of the Human Computer Interaction Status Conference 2003*. DLR, Berlin (Germany), pp. 47–62.
- Walker, M.A., Kamm, C., Litman, D.J. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue systems*.
- Xiao, B., Lunsford, R., Coulston, R., Wesson, M., Oviatt, S.L., 2003. Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In: *Proceedings of the International Conference on Multimodal Interfaces*. ACM Press, Vancouver, BC, pp. 265–272.