

Multimodal Interaction in Architectural Design Applications

Lou Boves¹, Andre Neumann¹, Louis Vuurpijl¹, Louis ten Bosch¹, Stéphane Ros-signal¹, Ralf Engel², Norbert Pflieger²

¹NICI, Nijmegen, the Netherlands

²DFKI, Saarbrücken, Germany

Abstract. In this paper we report on ongoing experiments with an advanced multimodal system for applications in architectural design. The system supports uninformed users in entering the relevant data about a bathroom that must be refurbished, and is tested with 28 subjects. First, we describe the IST project COMIC, which is the context of the research. We explain how the work in COMIC goes beyond previous research in multimodal interaction for *eWork* and *eCommerce* applications that combine speech and pen input with speech and graphics output: in design applications one cannot assume that uninformed users know what they must do to satisfy the system's expectations. Consequently, substantial system guidance is necessary, which in its turn creates the need to design a system architecture and an interaction strategy that allow the system to control and guide the interaction. The results of the user tests show that the appreciation of the system is mainly determined by the accuracy of the pen and speech input recognisers. In addition, the turn taking protocol needs to be improved.

1. Introduction

Research in multimodal interaction tends to divide into two categories that have little in common. One field focuses on applications where users interact with some kind of map, or complete some kind of form using a combination of speech and pen for input. More often than not, the pen can only be used as a pointing device. For entering alphanumeric input with the pen, a soft keyboard must be used, or the user must write isolated characters in a dedicated field on the screen. Examples of projects in this category are SmartKom [1] and MUST [2]. The other category addresses virtual reality applications, where the user can move around freely, while the system interprets all speech and gestures that are relevant for the completion of a specific task [3]. In the ongoing IST project COMIC¹ [4] we intend to narrow the gap between the two categories, by extending the input and output capabilities of an application in the first

¹ <http://www.hcrc.ed.ac.uk/comic/>

category. At the output side the COMIC system features a talking head, displaying naturalistic turn taking behaviour, expressed by means of speech prosody, eye contact and gaze. At the input side the COMIC system supports pen input processing, in addition to automatic speech recognition.

Projects in multimodal interaction differ in yet another aspect. Many projects aim at a fundamental investigation of how several input and output modalities can be combined in human-system interaction. Here, the focus is on experiments with procedures to interpret multimodal input, and methods for rendering information in parallel output channels. Another category of projects aims at developing operational multimodal services, often in digital telecommunication networks, but also desktop applications for non-expert users. Projects in this category focus –by necessity –on developing interfaces that can be implemented and maintained cost-effectively, yet are easy to use for a broad range of customers. It is well known that there is a large difference between customers who pay for using a service and subjects who are paid for participating in experiments. Perhaps it is less well appreciated that the difference between computer scientists who have developed their own multimodal interfaces and uninformed users (be they subjects or customers) is at least as large.

In COMIC, we move one step beyond the conventional map and form filling applications, by addressing an architectural design task, instantiated in the form of bathroom design. In this paper we first introduce the COMIC project in more detail. In section 3 we explain the fundamental problems that must be solved to enable natural human-system interaction in architectural design. Section 4 describes the system that we built for entering a blueprint of a bathroom, and section 5 reports on an experiment in which uninformed subjects tried to use the system. Section 6 completes the paper with conclusions and recommendations.

2. The COMIC project

COMIC [4] is an FP5 project in Key Action 2, in the area of Long Term, High Risk Research. COMIC combines software and system development with experiments in human-human and human-computer interaction in language-centric multimodal environments. The experiments are based on a scenario that can be controlled experimentally, but that at the same time is relevant for *eCommerce* and *eWork* applications. The bathroom design application has speech and pen input recognition at the input side (cf. Fig. 1). In addition, users can point at objects on the screen, such as bathtubs, basins, faucets, etc., and ask the system to shown alternative designs. The system can explain advantages and disadvantages of specific designs. In doing so, it takes into account a dynamically evolving model of the preferences, likes and dislikes of the user. In addition to the tablet screen, where designs and drawings can be shown, the system features a second screen that displays a highly realistic talking head. To enhance the naturalness of the interaction this ‘avatar’ is able to express the moods and attitudes that a customer expects from an expert sales consultant (but the automatic system will always stay polite and will never show irritation). A schematic image of the layout of the application during the phase when the shape and dimensions of the room is being entered is shown in Fig. 1. The avatar guides the user

through the application by explaining what it is expecting and by asking questions if the input is ambiguous. The user can simultaneously draw or write and speak.



Fig. 1. Overview of the bathroom design application. The tablet is used for pen input to enter size and dimension of the bathroom.

The interaction starts with the user entering the blue print of the room, including the position of the door(s) and windows, the opening direction of the door and the height and width of the windows, since these determine feasible layouts of sanitary ware and additional bathroom furniture. After the ground plan of the room is entered, it can be decorated with tiles and sanitary equipment. Subsequently, the user can move through a 3D image of the design, and discuss possible changes. However, the present paper only addresses the process of entering the shape and size of the room.

3. Issues in multimodal interaction in design applications

In order to get an impression of how naive subjects go about entering the shape and dimensions of a room into a computer system with human-like capabilities, we conducted an experiment in which we asked several people to perform the task. They were told that they could draw, write and speak freely. The experimenter provided backchannel feedback to encourage the subjects to speak as if they were addressing a person, and asked clarification questions if he did not understand the information. In addition, the experimenter prompted the subjects to provide all the information that they were instructed to give. Figure 2 shows a representative example of the resulting pen input [see also 5]. The problems that the experimenter experienced in interpreting the sketches and the verbal explanations given by the subjects are very similar to the issues addressed in [11], where it is shown that there is no fixed and predictable relation between sketches and speech: in some cases verbal expressions can only be interpreted with the support of a sketch, while in other cases sketches can only be interpreted with the help of verbal explanations.

From Fig. 2 it is clear that unconstrained pen and speech input pose recognition problems that are insurmountable with existing technology. In addition, it appeared

that all subjects needed substantial guidance and help from the experimenter to complete the task of specifying a complete bathroom. Virtually all subjects needed help in devising ways for expressing the opening direction of a door and the height of a window and a window sill. In Fig. 2 it can be seen that this subject tried to solve the latter problem by drawing a side view of the wall containing the windows. To avoid insurmountable problems for subjects trying to interact with an automatic system, we decided to design a much more structured interaction strategy. To simplify the task for on-line pen and speech input recognition as much as possible, we opted for a system driven interaction style, in which the system prompts the user to enter individual information elements, such as the position and the length of the walls, the position and opening direction of the doors, and the position, height and widths of the windows.

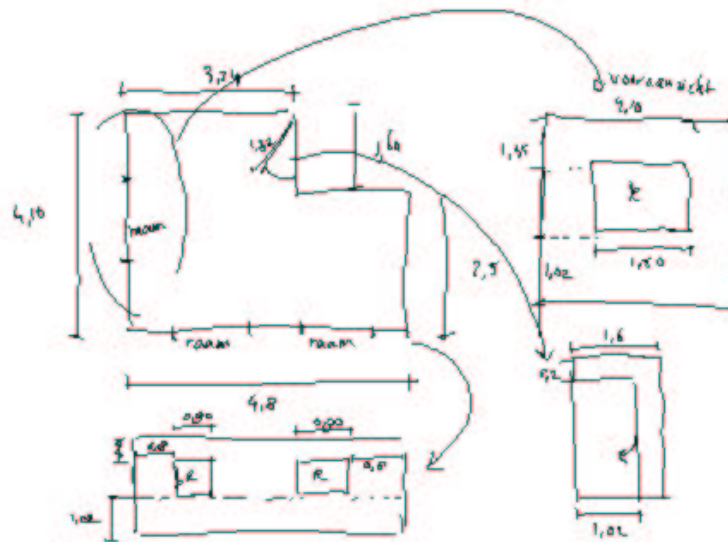


Fig. 2. Example of pen input of the blueprint of a bathroom.

4. The COMIC system for entering blueprints

Fig. 3 shows the architecture of the system that we built for conducting a Human Factors experiment to investigate whether uninformed subjects are able to enter the blueprint of a bathroom using pen and speech as input channels. The system is built using the MultiPlatform environment for implementing multimodal applications that was developed in the Verbmobil and SmartKom projects [6] and that is now publicly available². The present implementation of the system is a simplified version of the eventual COMIC system in that it does not yet include the Dialogue and Action Management (DAM), Fission and Output modules that are described in [4]. The task of

²² <http://sourceforge.net/projects/multiplatform/>

the DAM is taken over by a Wizard, who essentially decides whether or not a user input can be interpreted, and triggers the appropriate system response. System outputs consist of spoken prompts requesting the user to enter an information element and feedback about the interpretation of the user input in the form of graphical output on the Wacom Cintiq LCD Tablet. The recognition of walls, doors and windows is echoed by ‘beautifying’ the user’s pen input: it is overlaid by straight lines for the walls, and standardised graphics for doors and windows. Lengths and measures are echoed as printed characters on the tablet. Users can erase wrongly recognized input by means of spoken utterances (“*No, I meant three meters and thirty centimetres*”), or by erasing the system output with the upper end of the pen (that doubles as an eraser).

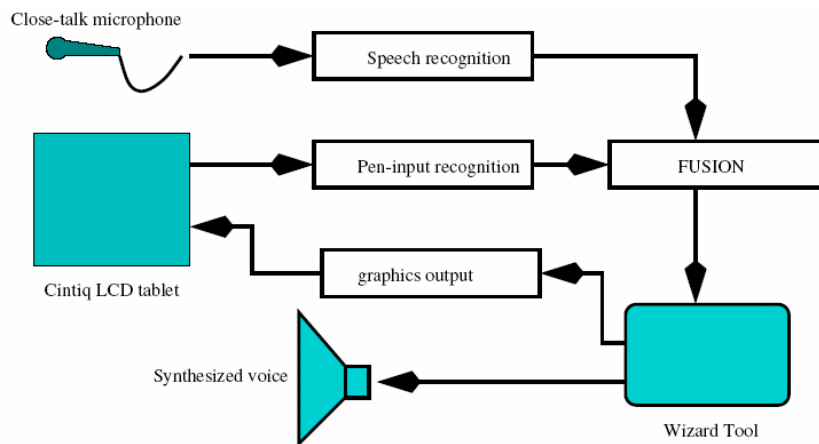


Fig. 3. Schematic representation of the system for entering ground floor plans.

Speech recognition is implemented with the HTK toolkit, adapted for interactive usage [7]. Context dependent phone models were trained using the German SpeechDat database [8]. The language model was inferred from recordings in pilot experiments, and extended with the intuitions of the experimenters about plausible types of expressions. Pen input recognition is implemented with algorithms developed in the NICI. Fusion is implemented on the basis of the procedures and software developed at DFKI in the framework of the SmartKom project. The Dialog Management protocol followed by the wizard is presently being automated, taking due account of the experience gained in the Human factors experiments reported in this paper. The data recorded in this experiment were processed using a tool developed in the NICI [9].

The first fully operational version of the system depicted in Fig. 3 did not implement a strict definition of ‘turn’ and turn taking, allowing for fully asynchronous, full-duplex interactions. This asynchrony caused severe misunderstandings between subjects and wizard, and the lack of a mutually agreed communication protocol caused the partners to run out-of-sync. Therefore, we were obliged to define a strict turn taking protocol that boils down to half duplex communication in which speech and pen input were confined to a fixed time window following the end of a system prompt. The moment when the subject could start writing and speaking was indicated by a

green square that appeared in the left upper corner of the tablet. At the end-of-turn that square turned red, after which input would not be processed.

5. Experiment and results

We have conducted a large scale experiment in which 28 subjects (8 male, 20 female) aged between 20 and 37 (median 23) have used the system to enter ground floor plans of three bathrooms. All subjects had a university level education, but no computer science background; more importantly, none of the subjects was familiar with the research project. However, most reported to have substantial computer experience, but very little experience using speech recognition systems and even less using pen tablets. At the start of the experiment subjects were given a short explanation of what was going to happen. Next, they were requested to specify three bathrooms, their own, that in their parents' house, and a third one of their choice. During this phase they could freely speak and draw and become acquainted with writing on the tablet and using the head mounted microphone. The experimenter would ensure that all data were given (including doors and windows) and suggest ways for expressing specific information elements when needed. Hardcopies of the ground plans were then made and given to the subject to serve as a mnemonic during the next part, in which they had to copy the information into the 'automatic' system, one room at a time. Before starting with the first room subjects saw a short instruction video that explained how the system would show what it had recognised, and how they could correct recognition errors. After completing the third room subjects were asked to fill out a questionnaire comprising 26 Likert scales; a score of 1 corresponded with 'I disagree completely', whereas a score of 5 meant 'I fully agree'. Below we present the results of the analysis of the objective interaction data that were logged during the experiments, the subjective scores on the Likert scales, and interesting correlations between the objective and subjective measures.

From the scores on the Likert scales it appears that subjects had no difficulty understanding the task: mean score is 4.09. It was less clear what to do while using the system (3.43), although the prompts were clear (4.36), and it was easy to understand the way in which the system showed its recognition result (4.14). In general, subjects knew what to do to correct recognition errors (3.59). Despite the fact that they did understand the task, subjects said that they found it rather difficult to use the system (2.68) and that it was not very efficient (2.59). As a consequence, they said that they needed to concentrate hard (3.82), and that it took long to enter all information (3.91). Also, they did not find themselves in control (2.18), the system was not seen as very reliable (2.14), and it definitely needs improvement (4.18). Several subjects found it difficult to wait for the green square to appear on the tablet, and to react within the time window. Also, subjects did not always understand how they had to correct recognition errors in numbers and dimensions. If subjects said or wrote the equivalent of 3.25 m, and the system recognised 2.25 m (i.e., substituting the '3' in the number by a '2'), they could only erase the complete string (both the number and the dimension) and they had to re-enter both. Quite a number of subjects wanted to erase or to re-enter only the digit that was misrecognised. Although on average subjects disagreed

with the statement that the system was too fast (mean score 2.64), we observed that a substantial proportion of the input utterances were truncated because they exceeded the maximum allotted time window. With respect to the input modalities subjects reported that they found the pen easy to use (3.55); the use of the eraser was even simpler (4.27). The combination of pen and speech was easy (3.5), the naturalness of the interaction was assessed as almost neutral (3.09). Only a small proportion of user utterances contained simultaneous and related pen and speech input. However, as is apparent from the Likert scales, subjects appreciated the possibility to choose between pen and speech (4.05).

Although the performance of the pen recogniser was substantially higher than that of ASR, subjects tended to first try and speak the answer to prompts about sizes and dimensions. Only after repeated misrecognitions they switched to writing. However, subjects for whom ASR performed worst changed their behaviour during the course of the experiment: especially while entering the third room they tended to avoid speech and used the pen exclusively. Natural Language Processing and Fusion could do little to improve recognition accuracy, since subjects hardly ever combined pen and speech to enter size and dimension.

While the major cause of the ASR errors is mostly related to robustness of ASR against out-of-grammar utterances, most of the errors in handwriting recognition can be traced back to the fact that many subjects used a comma as the 'decimal point', whereas the recogniser was trained with a bias towards the Anglo-Saxon use of the full stop for that purpose. The mediocre recognition performance is the major explanation of the finding that the subjects were not very happy with the system. Objective data about recognition performance explain more than 50% of the variance in the (negative) scores on the Likert scales. However, several subjects said that the system would have been easy to use if the recognition performance had been better.

6. Conclusions and recommendations

The positive scores on the Likert scales addressing the transparency of the task show that the overall design of our system is sound from a Human Factors point of view. However, it is also evident that substantial technical development and improvement is needed before uninformed subjects are able to use the system in an easy and transparent manner. Both input recognisers must be improved substantially, to enable them to handle the behavior of subjects who are task oriented, instead of focusing on human-system interaction per se. In addition, we have found that –although the system driven interaction strategy did not frustrate our subjects- the turn taking protocol needs to be improved. Subjects' inputs should not be constrained to fixed duration time windows, the start of which is determined by the end of the system prompt.

Our data confirm previous results that show that subjects tend to stick to a given input mode, despite the fact that this may not be the most effective one [10]. Moreover, our results suggest that the subjects' preferred mode is heavily influenced by the mode used by the system to address its user: in our design all system prompts are spoken, eliciting spoken replies whenever that is feasible.

Multimodal interaction combining pen and speech input in a system driven interaction can support non-experts in performing a complex task that would be very difficult to perform without substantial guidance of the system. Yet, the turn taking paradigm should be made more flexible than was the case in our system. Most importantly, the accuracy of the input recognisers needs to be improved. It is important to investigate methods for error correction that allow subjects to repair only those parts of a complex expression that were recognised incorrectly, without having to re-enter the parts that were correctly recognised in the first place.

Acknowledgement

This research is partially funded by the European Commission, under the 5th Framework Programme, project number IST-2001-32311.

References

- [1] W. Wahlster, "SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions". *Proc. First International Workshop on Man-Machine Symbiotic Systems*, Kyoto, Japan, 2002, pp. 213-225.
- [2] L. Almeida et al., "User-friendly Multimodal Services - A MUST for UMTS. Going the Multimodal route: making and evaluating a multimodal tourist guide service". *Proc. EUESCOM Summit*, 2001.
- [3] T. W. Bickmore, and J. Cassell, "Relational Agents: A Model and Implementation of Building User Trust". *CHI 2001*, Seattle, WA.
- [4] E. den Os and L. Boves, "Towards Ambient Intelligence: Multimodal computers that understand our intentions". *Proc. eChallenges*, Bologna, October 2003.
- [5] S. Rossignol, L. ten Bosch, L. Vuurpijl, A. Neumann, L. Boves, E. den Os, and J.P. de Ruiter, "Human Factors issues in multi-modal interaction in complex design tasks". *Proceedings HCI International 2003*.
- [6] G. Herzog, H. Kirchmann and P. Poller, "MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems". *HLT-NAACL'03 Workshop Software Engineering and Architecture of Language Technology Systems*, 2003.
- [7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University, Cambridge, UK, 1997.
- [8] H. Hoege, C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders & H.S. Tروف, "Speechdat multilingual speech databases for teleservices: across the finish line". *Proc. EUROSPEECH'99*, Budapest, Hungary, 5-9 Sep. 1999, pp. 2699-2702
- [9] L. Vuurpijl, L. ten Bosch, S. Rossignol, A. Neumann, N. Pflieger and R. Engel, Evaluation of multimodal dialog systems, *LREC Workshop Multimodal Corpora and Evaluation*, Lisbon 2004.
- [10] Oviatt, S. and VanGent, R., "Error resolution during multimodal human-computer interaction". *Proc ICSLP 1996*, pp. 204-207.
- [11] Lee, J. "Words and pictures -- Goodman revisited". In: R. Paton and I. Neilson, *Visual Representations and Interpretations*, London: Springer-Verlag, 1999, pp. 21-31