Enhancing a WIMP based interface with Speech, Gaze tracking and Agents

Lau Bakman, Mads Blidegn, Martin Wittrup, Lars Bo Larsen, Thomas B. Moeslund

{bakman,blidegn,wit,lbl,tbm}@kom.auc.dk
Center for PersonKommunikation,
Aalborg University
Aalborg, DK-9220, Denmark

ABSTRACT

This paper describes an attempt to enhance a windows based (WIMP - Windows Icon Menu Pointer) environment. The goal is to establish whether user interaction on the common desktop PC can be augmented by adding new modalities to the WIMP interface, thus bridging the gap between todays interaction patterns and future interfaces comprising e.g. advanced conversational capabilities, VR technology, etc.

A user survey was carried out to establish the trouble spots of the WIMP interface on the most common desktop work station, the Windows 95 PC. On the basis of this, a number of new modalities were considered. Spoken in- and output and gaze tracking were selected together with the concept of an interface agent for further investigation.

A system was developed to control the interaction of the in- and output modalities, and set of five scenarios were constructed to test the proposed ideas. In these, a number of test subjects used the existing and added modalities in various configurations.

1. INTRODUCTION

The paper presents the results of a survey and an experiment with an enhanced WIMP user interface [1]. The purpose of this work is to investigate an intermediate stage between the current and the next generation computer, under the assumption that the next generation is characterised by new interaction methods. This intermediate stage involves combining the advantages of the current generation with some of the new interaction methods expected to be part of the of the next generation computer interface.

More specifically, the work presented here examines which new modalities could easily and naturally be integrated into todays standard computer interface. An investigation of the current state-of-the-art, with regard to research and commercial availability is carried out, after which a system called Herbert is designed and implemented. This is tested in a usability test, in order to evaluate the proposed ideas.

Section 2 discusses the survey. Based on the conclusions from the user survey, an enhanced interface is proposed in section 3. A software system, capable of integrating the added modalities into the standard Windows 95 interface is implemented and described in section 4. A set of scenarios are constructed and a user test is carried out to establish to what degree users will accept the proposed interface. The scenarios and results of the user test are discussed in section 5. Finally, the results of the usability experiment are presented, and conclusions are drawn trying to give an assessment of future interfaces on the standard work station.

2. USER SURVEY

A survey carried out in 1997 [2] established that Windows 95 is the dominating operating system on personal computers, and consequently also defines the predominant user interface. The survey showed that more than 85% used either Win 95 or 3.1. Therefore, the Win 95 user interface were chosen as the subject for this work.

It was decided to identify common problems in the human-computer interface by doing a user survey, asking users to fill out a questionnaire. The problems of interest were tasks or actions that a user may find tedious, annoying or inconvenient during normal use of a computer. The questionnaire was then distributed throughout Aalborg University via Usenet news. Although convenient, this of course limited the demographics of the of the survey to mostly students and employees within certain age groups. However, this was also deemed an advantage, as some of the questions concerned new interaction modes, such as eye-gaze tracking, presupposing a certain knowledge.

The results of the survey were used as guidelines to which problems are generally encountered in Windows 95. The survey could have been conducted in other ways than using a questionnaire, for example as a think aloud test, interviews, etc., but the questionnaire was found to be the easiest method of collecting information from a large number of people. Approximately 200 persons participated in the survey.

2.1 The questionnaire

Only the most important results of the survey are reported in the paper. See [1] for a full description.

One part of the questionnaire concerns the users identification of interaction problems in the current WIMP interface. The most common problems identified are:

“Remember short-cut keys”, *“Find menu item”*, and *“Change from keyboard to mouse”*. Of these three problems, one is particular interesting: The respondents does not see the mouse in general as a problem. It is when changing between keyboard and mouse that a problem may arise. This points towards finding a solution which minimises the number of situations where it is necessary to change between mouse and keyboard., rather than completely trying to substitute the mouse.

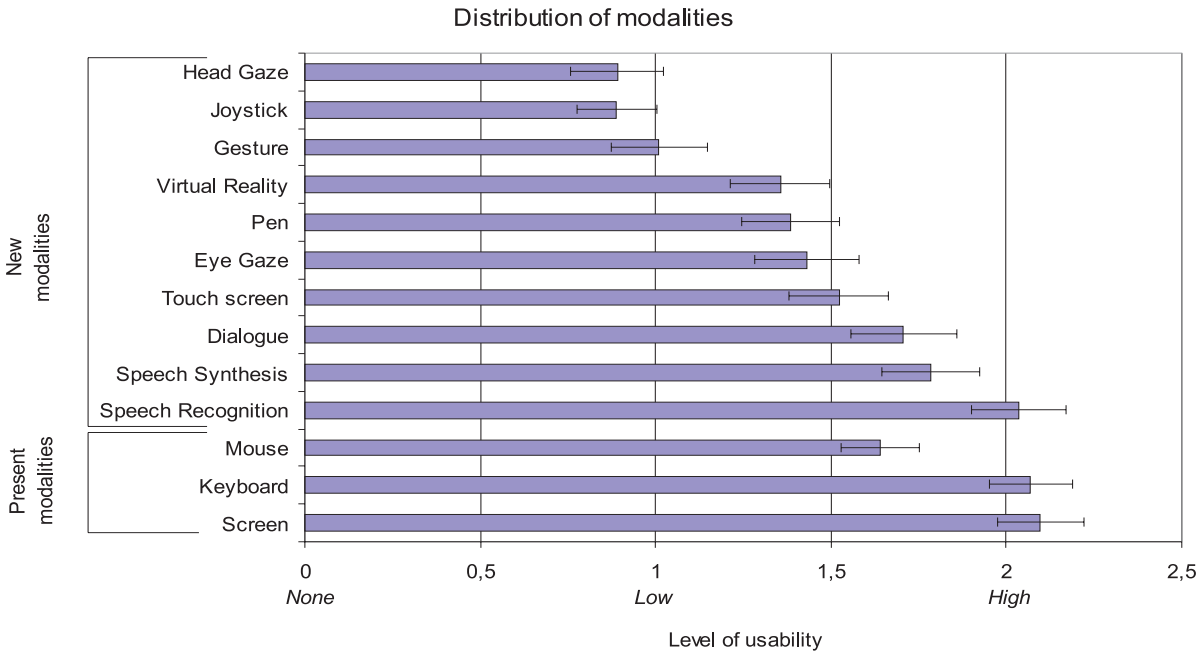

Figure 1: Usability of modalities with 95% confidence intervals, grouped by the interface most commonly used (denoted "present" and non-traditional or future (denoted "new"). From [1].

Another interesting result is that a large number of respondents agrees that it is difficult to remember short-cut keys. A solution to this could be to implement a more natural way of executing short-cuts, e.g. by using speech. This problem also includes finding menu items, as short-cut keys mostly are introduced to make the menus easier to access. Therefore the two problems are closely connected.

The usability distribution on figure 1 shows how the respondents assessed the usability of traditional and new ways of interacting with the computer. It is grouped by present and new modalities, where the most popular of the present modalities is the monitor, followed by the keyboard. Among the new modalities speech recognition is the most interesting, followed closely by speech synthesis and "dialogue". It is interesting to see that the mouse actually is down on a sixth place on the overall usability level.

About 15% of the respondents were dissatisfied with the present interface, 50% were satisfied, but would like new possibilities, and 28% were very interested in new possibilities. Less that 10% were satisfied with the present interface and did not want new ways of interaction.

2.2 Results of the Survey

When combining the investigation of the current problems with the users' rating of traditional and new interaction modes, it becomes apparent that the time is not (yet) for replacing the monitor and keyboard. On the other hand, the survey identifies some problems with issues concerning the mouse. The task of finding menu items and the switch from keyboard to mouse are considered problems.

Figure 1 points directly towards the inclusion of speech understanding and -generating capabilities as the most likely candidates for new modalities. Consequently, speech communicating as well as (at least a partial) substitute for the common mouse pointing device will be considered in the following.

3. SELECTION OF MODALITIES

This section discusses how to introduce speech and an alternative pointer control mechanism into the existing interface. It is important to keep in mind that the resulting interface should be an addition to the existing WIMP interface, as stated in section 1. Therefore, a complete redesign is not considered, although the authors are well aware that this is the ultimate solution to some of the problems identified above. In [3] this fact is demonstrated for a spread-sheet task, where errors decreased significantly with a deeper integration of speech into the application.

Introduction of speech based communication is not without problems. The speech recogniser must achieve a certain level of performance in order to be acceptable to users. However, this is rapidly becoming a minor problem for normal operation conditions. Likewise, the synthesised voice must have an acceptable quality. Good performance is not enough, however, to make users automatically switch to spoken communication. Many people feel uncomfortable or awkward when speaking "to the screen". One concept, "Interface Agents" ([4] [5] [6]) has emerged to amend this problem and act as a "personification" of the computer.

Interface agents

Interface agents are programs that can manipulate objects in a direct-manipulation interface, but without explicit instructions from the user [6]. Often interface agents are represented on the screen by a face or a small humanoid figure. The agent can be used in several ways, by observation or by request. The agent might observe a pattern in what the user is doing and offers to perform this task automatically, for example a search and

replace operation.

Microsoft has developed a software package implementing an interface agent[9]. Figure 2 shows an example of one of the characters' (Merlin) appearance. The agent can perform a number of small animations, enabling the agent to perform gestures such as pointing, showing attention, listening, moving, etc., when the corresponding commands are sent to it. Furthermore, it has a direct interface to speech recogniser and -synthesiser, enabling the character to "listen" and "talk" in synchrony with the actual processing.

Figure 2: The interface agent "Merlin"

A survey of commercially available speech recognisers and synthesisers have been conducted. The main conclusions from the survey are:

Speech recognition:

Many commercially available speech recognisers exist and the amount of research in the field is huge. Experiments with a number of products have been conducted and Microsofts "Whisper" speech recogniser [7] was chosen.

Speech synthesis

Nearly as many speech synthesis products are available. Different products were tested and the performance of these was more or less the same. Lernout & Hauspie's TruVoice [8], which uses the formant technique was chosen because of its high intelligibility.

Apart from the fact that the two products chosen above fulfil the requirements, another issue has been important for the task at hand, namely that both are capable of directly interacting with Microsoft Agents [9].

Pointing Device

As discussed in the previous section, a solution to the "keyboard - mouse problem" must be found. Tracking the direction of the users eye-gaze is an obvious candidate for consideration. However, the product survey showed that although very accurate devices are available [10], they have a number of drawbacks making the technology unlikely to appear on the consumer market in the near future. Most of them require the user to wear special, uncomfortable equipment like helmets or electrodes. Furthermore, they are very expensive (in the range of 10 - 50.000 USD), and are clearly intended for experimental use only. Apart from this, no definitive solution to the "Midas Touch Problem"[1][11],[12] has been proposed yet. Until good solutions to this are found, eye-gaze appears to be problematic for general purpose applications, like the present.

Instead of eye-gaze it was chosen to investigate the usability of head-gaze, i.e. to control the pointer by positioning of the head. A product intended for the handicapped, the Madenta Tracker [13], was chosen. It is roughly twice the size of an ordinary mouse and is placed on top of the monitor. It is capable of tracking the position of a small reflective dot, which the user places on his forehead or glasses. The Tracker costs less that 1.500 USD and is shown in figure 3.

Figure 3: The Madenta Tracker

4. IMPLEMENTATION

A system capable of controlling the devices and interface agent has been designed and implemented. It is denoted "Herbert". The overall architecture is shown in figure 4. Herbert is an application, that receives information from the input devices, which are the top-most objects in figure 4. The figure should be interpreted as Herbert being the client application, which coordinates information from the input devices, which in turn act as servers or services. The coordinated information is then passed on to the appropriate output devices shown at the bottom of the figure. Besides the output modalities Herbert uses the operating system for performing some actions. It should be noted that the monitor is not directly controlled by Herbert, but Herbert produces additional output on the monitor by using the interface agent. For further details, see [1].

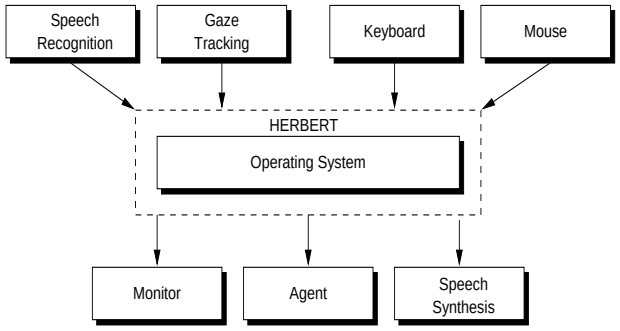

Figure 4: Architecture of the implemented system

5. EXPERIMENTS

An experiment was set up with five cases, which together investigates the new modalities introduced into the Win 95 user interface.

Case 1) Program Management: This case presents three new methods for starting applications in Windows 95, and one for terminating them. The methods involve all the new modalities except gaze control.

Case 2) Window Management: This case mainly focuses on the problem of managing windows present on the monitor by means of the gaze tracker. The case uses two modalities, speech recognition and gaze tracking. By combining speech

[1] i.e. that everything you look at is affected in some way, as your gaze interacts with the objects on the screen.

input with gaze tracking, it is e.g. possible to point out inactive windows on the screen and thereby reducing the need for the ordinary mouse.

Case 3) Common Functions: In this case the focus is on menu navigation and the problems involved when using different functions. The case should reduce the problems of navigating complex menu structures, as the user should be able to activate functions by speech in terms of embedded keywords describing the functionality.

Case 4) Hints and Help: This case focuses on helping the user in situations where the user performs monotonous repeated actions. This is done by monitoring the acts of the user and presenting alternative ways of performing the repeated actions. Also new interaction methods for user initiated help requests are investigated.

Case 5) Gaze tracking: The purpose of this case is to reduce the number of situations in which the user is forced to use the mouse and to investigate whether gaze tracking is a viable supplement to the mouse for manipulating small objects like menus and icons.

The experiment was carried out with 18 test subjects completing the 5 cases, both with the standard Win 95 user interface and the enhanced one. Quantitative as well as qualitative data were collected.

6. RESULTS AND CONCLUSIONS

The results showed that the standard interaction methods were approximately 60% faster on average for performing the tasks of the test. This was expected, as the test subjects all had a long experience with the standard interface. The only improvement in time was achieved when navigating complex menus which was the task in case 3. For these tasks the enhanced interface were on average 33% faster than the standard one. However, when using the gaze tracker, the standard interface were on average between 2 and 3 times faster than the enhanced one.

The results showed that it was possible to perform the tasks with 40% less mouse to keyboard switches when using the enhanced interface. On a scale from ‘useless’ to ‘very useful’ all the new modalities (except the gaze tracker) were rated to be between ‘useful’ and ‘very useful’ on both their present level of development and on a future, ideal level. The Madenta Tracker was rated between bad and useless on its present level, which corresponded with the large time differences in the cases where gaze tracking was used as described above. When asked to rank the new modalities, the order of preference matched the results obtained in the first survey. Speech recognition was rated most useful, and also the modality appealing to the widest group of people from beginners to professionals. The use of agents was considered most appealing to beginners and ordinary users. The presented combination of modalities was recommended by all test subjects without exception. The enhanced interface were also rated as being between ‘useful’ and ‘very useful’ on a general level whereas the usability of the standard interface were rated to be between ‘medium’ and ‘useful’.

In the cases concerning the execution of common functions and methods for getting hints and help, the enhanced interface were rated higher than normal methods. Except for the methods that involved using the gaze tracker (at its current performance), the new interaction methods were all considered easier and more natural than the traditional methods.

The work presented here clearly indicates that spoken interaction can be expected to be widely accepted very quickly, when applications including this begin to appear. More investigation is needed to determine whether gaze control is viable. This work indicates that head-gaze is not viable (at least in it’s present level of development), and it remains to be shown whether eye-gaze will be accepted, and whether technology suitable for mass production will appear.

7. REFERENCES

[1] Bakman, L., Blidegn, M., Wittrup, M. “Improving Human-Computer Interaction by adding Speech, Gaze-Tracking, and Agents to a WIMP-based Environment”. Master Thesis, the IMM programme, Aalborg University, June 1998.

[2] The Réseau Interordinateurs Scientifique Québécois. Third RISQ survey March 97, 1997. http://www.risq.qc.ca/survey/3/installation/inst_ordi.html.

[3] Jim Hugunin and Victor Zue.”On the Design of Effective Speech-Based Interfaces for Desktop Applications.” In Proc. Eurospeech’97, Volume 3, pages 1335–1338, Rhodes, Greece, September 1997.

[4] Tomoko Koda and Pattie Maes. Agents with Faces: “The Effects of Personification of Agents”. In Proceedings of HCI’96, London, UK, August 1996. http://pattie.www.media.mit.edu/people/pattie/cv.html.

[5] Brenda Laurel. “Interface Agents: Metaphors with Character.” In The Art of Human-Computer Interface Design, pages 355–365. Addison-Wesley Publishing Company, 1994.

[6] Henry Lieberman. “Autonomous Interface Agents.” ACM Association of Computing Machinery, 1997. Available from: http://www.acm.org/sigchi/chi97/proceedings/paper/hl.htm.

[7] Xuedong Huang et. al. “Microsoft Windows Highly Intelligent Speech Recognizer: Whisper”. IEEE International Conference on Acoustics, Speech and Signal Processing, USA, 1995. http://www.research.microsoft.com/research/srg/whisper.htm.

[8] Lernout & Hauspie Speech Products. 52 Third Avenue, Burlington, MA 01803, USA. http://www.lhs.com.

[9] Microsoft Corp. “Microsoft Agents.” Technical report, Microsoft Corporation, December 1997. http://www.microsoft.com/workshop/prog/agent/

[10] Theo Engell-Nielsen, Arne John Glenstrup. “Eye Controlled Media: Present and Future State”. Technical report, University of Copenhagen, June 1995. http://www.diku.dk/panic/eyegaze/

[11] Robert J. K. Jacob. “The Use of Eye Movements in Human-Computer Interaction Techniques: “What You Look At is What You Get””. ACM Transactions on Information Systems, 9(3):152–169, April 1991.

[12] Robert J. K. Jacob. “Eye Tracking in Advanced Interface Design”, pages 258–288. Oxford University Press, New York, 1995. http://www.eecs.tufts.edu/jacob/papers.html.

[13] Madenta Inc. Madenta. Madenta Inc. 3022 Calgary Trail South Edmonton, Alberta T6J 6V4, Canada, http://www.madenta.com.