Pers Ubiquit Comput (2006) 10: 269–283
DOI 10.1007/s00779-005-0048-1

ORIGINAL ARTICLE

Eamonn O'Neill · Manasawee Kaenampornpan
Vassilis Kostakos · Andrew Warr · Dawn Woodgate

Can we do without GUIs? Gesture and speech interaction with a patient information system

Received: 26 January 2004 / Accepted: 6 January 2005 / Published online: 10 November 2005

Abstract We have developed a gesture input system that provides a common interaction technique across mobile, wearable and ubiquitous computing devices of diverse form factors. In this paper, we combine our gestural input technique with speech output and test whether or not the absence of a visual display impairs usability in this kind of multimodal interaction. This is of particular relevance to mobile, wearable and ubiquitous systems where visual displays may be restricted or unavailable. We conducted the evaluation using a prototype for a system combining gesture input and speech output to provide information to patients in a hospital Accident and Emergency Department. A group of participants was instructed to access various services using gestural inputs. The services were delivered by automated speech output. Throughout their tasks, these participants could see a visual display on which a GUI presented the available services and their corresponding gestures. Another group of participants performed the same tasks but without this visual display. It was predicted that the participants without the visual display would make more incorrect gestures and take longer to perform correct gestures than the participants with the visual display. We found no significant difference in the number of incorrect gestures made. We also found that participants with the visual display took longer than participants without it. It was suggested that for a small set of semantically distinct services with memorable and distinct gestures, the absence of a GUI visual display does not impair the usability of a system with gesture input and speech output.

E. O'Neill (✉) · M. Kaenampornpan · V. Kostakos · A. Warr · D. Woodgate
Human-Computer Interaction Group, Department of Computer Science, University of Bath, BA2 7AY Bath, UK
E-mail: eamonn@cs.bath.ac.uk
E-mail: mk@cs.bath.ac.uk
E-mail: vk@cs.bath.ac.uk
E-mail: aw@cs.bath.ac.uk
E-mail: daw@cs.bath.ac.uk

Keywords Multimodal interaction · Gesture input Speech output · GUI · Mobile · Ubiquitous

1 Introduction

Input and output devices and techniques that work (at least some of the time!) with deskbound computers are often inappropriate for interaction away from the desktop. Simplistic solutions seen with several current mobile and wearable devices, such as making the keyboard much smaller, create their own usability problems. Physically shrinking everything including the input and output devices does not create a usable mobile computer. Instead, we need radical changes in our interaction techniques, comparable to the revolution in the 1980s from command line to graphical user interfaces. For example, Brewster and colleagues have investigated overcoming the limitations of tiny screens on mobile devices by utilising sound and gesture to augment or to replace conventional mobile device interfaces [1, 2].

A persistent problem with the usability of mobile computing has been the conflation of the physical characteristics of the device with the characteristics of the interface between the user and the computing services that the device delivers. For example, as mobile and wearable devices become ever smaller, their display areas, which typically serve as both input and output devices, become ever smaller and less usable. The legitimate desire to make mobile and wearable devices more mobile and easier to wear through miniaturisation will render these devices less and less usable so long as the interface and its associated interaction techniques continue to be conflated with the physical characteristics of the device itself. In attempting to resolve this dilemma, we have been exploring ways of decoupling the interaction techniques from the physical characteristics of the devices.

In our current research, we assume that there will be an increasing convergence between mobile/wearable

computing and ubiquitous computing. For many applications, the user may want to use, say, the wall display in the hospital waiting room or café with the high bandwidth connection, rather than the tiny display on her PDA with its relatively poor connectivity. For other applications, the user may prefer to take advantage of the characteristics of her mobile device. Indeed, some applications may be most usable through simultaneous use of a combination of ubiquitous and mobile computing power. In the context of converging mobile and ubiquitous technologies, this implies developing input and output techniques that will work with devices ranging from the smallest wearable computer or smart ring with no visual display to a wall-sized display driven by a powerful fixed-location computer in a shop or street or hospital. Again, this motivates us to decouple the interaction technique from the particular devices. Ideally, we should have a range of common, usable interaction techniques that operate across the gamut of desktop, mobile, wearable and ubiquitous devices.

In our recent work [3], we have developed a gesture-based input technique that attempts to achieve this goal. Clearly, however, we also need to consider output and in the work reported here, we have gone on to combine this gestural input technique with speech output. We propose that this combination of gestural input and speech output will satisfy our goal of decoupling interaction technique from device, providing a common, usable interface. To test this proposal, we implemented these interaction techniques in a prototype system developed from our field studies in a hospital Accident and Emergency (A&E) Department [4]. This paper reports an experimental evaluation of this prototype, investigating the effect of the presence or absence of a graphical user interface (GUI).

Multimodal interaction is likely to become increasingly important as a wide range of different people use a wide range of mobile, wearable and ubiquitous devices in a wide range of different situations, in many of which a visual display may not be effective or available at all. In addition to the difficulties noted above of producing a usable visual display for mobile and wearable devices, ubiquitous systems have their own problems with visual interaction. The most fundamental of these is that wireless technologies of various kinds, from Bluetooth to 802.11 to UMTS, enable the delivery of information and services in many, many more locations than one can expect to find visual displays through which to interact with these services. For example, a single 802.11 base station might give potential access to services over a radius of, say, 50 metres but it is clearly unreasonable to expect all of that area to be covered in visual displays. In this case, while the wireless connectivity is ubiquitous, the visual display based access to that connectivity is not. The absence of an effective, or indeed any, visual display motivates the use of other interaction modalities, individually and in combination. In addition, the standard desktop GUI model can be problematic for users with disabilities [5] and so alternative models may offer advantages. Attempts have been made simultaneously to ameliorate the problems of restricted visual displays on mobile devices and users' disabilities by combining, for example, auditory and tactile interaction techniques [6].

But the GUI paradigm introduced substantial advantages that drove the revolution from the previously ubiquitous command line interfaces. We must be careful in driving the shift to mobile and ubiquitous interaction that we do not lose the advantages that the GUI brought and so hamper, rather than improve, usability. In particular, a major advantage that standard GUIs brought is that users can see what services are currently available and how to invoke them. For example, in a word processor, the user can see that she may use a function to check spelling and grammar. She can also see that in addition to clicking on the Spelling and Grammar menu option, she may use the F7 key to invoke the same functionality. In addition, she can see that there is a Copy function but it is currently unavailable. At a traditional command line interface, there are no such visual cues and the cognitive burden on the user is correspondingly greater. In moving to a world where mobile and wearable devices are too small to have an effective visual display and ubiquitous systems can offer wirelessly networked services and communications coverage much more widely and pervasively than they can provide visual displays through which to interact, shall our novel interaction techniques prove to be less usable because we have lost the cognitive support gained in the move from command line interfaces to visual display based GUIs?

In tackling this question, the study reported here evaluated our prototype system combining gestural input and speech output in the presence or absence of a visual display of the available services and the gestures that invoked them. The main contribution of this paper is an experimental evaluation of the effect of having no visual display on the usability of such multimodal interaction. In our study we had one group of users who were able to see a menu of services and their corresponding gestures displayed on a GUI while they were asked to use the services by making the appropriate gestures. The results of invoking each service were presented as speech output via audio speakers. Our other group of users were also asked to use the same services from the same set by making the same gestures. However, during their trials, they could not see the menu of services and their corresponding gestures, and therefore had to remember them. Since they needed to see them in the first place in order to be able to remember them, all participants were given a training period practising the gestures while able to see the GUI menu of services and gestures. In order to explore the training effect, half of the users in each of the GUI and no-GUI conditions were given 5 min training and half were given 10 min training.

The primary experimental hypothesis (H1) predicted that users who could see the Services and Gestures visual display would (i) perform gestures more quickly and (ii)

perform fewer incorrect gestures than users who could not see the visual display. Taking these two measures as indicators of usability, a significant result confirming this experimental hypothesis would support the argument that usability would suffer in a paradigm of gesture and audio based interaction that lost the GUI paradigm's visual presentation to the user of the available services and means of accessing them. Our secondary experimental hypothesis (H2) predicted that users who had 10 min training on using the gestures would (i) perform gestures more quickly and (ii) perform fewer incorrect gestures than users who had 5 min training.

2 Information requirements for Accident and Emergency patients

In addition to our theoretical concerns, we have a desire to conduct our research in real world domains and challenges. Our recent work has included studying the complex, mobile, collaborative activities in the A&E department of a busy hospital, with the goal of identifying opportunities for technological support of these activities. The use of information displays by staff in healthcare settings has been shown to provide important support for patient care [7, 8], for example in organizing and locating clinical information, and coordinating and managing patient care. Despite initially entering the domain with a focus on the collaborative activities of the clinical staff, our fieldwork in the A&E department has also identified valuable opportunities for the exploitation of information technology by the patients themselves.

Patients were frequently observed to show signs of annoyance, stress and exasperation. Our field studies and previous research [e.g. 9] suggest that a major contributing factor is long waiting times with no explanation or information. In addition to causing stress for the patients, continual requests for information caused stress to the staff. The frequent need to respond to these requests was often distracting, interrupting their ongoing work. Such interruptions at times had the unfortunate effect of increasing the patients' waiting times still further. Previous work has shown that urgent care patients who were told the expected waiting time for treatment and were kept busy while waiting, had higher satisfaction perceptions of their treatment [9]. Maister [10] suggested that customers who were given information about how long they would have to wait are less likely to be anxious about the wait. Dansky and Miles [11] found that telling patients in an urgent care department how long they would have to wait was positively related to their satisfaction with the treatment.

This research suggests that the provision of information of this type might be a useful tool not only for reducing stress, but also in influencing patients' perceptions of satisfaction with their visit. In the A&E waiting area under study, some information was on display, though nothing that related to likely waiting times. There was clearly a requirement for this information since staff were continually asked by patients both for general information about the average waiting times that day and for specific information about their personal wait. This kind of information would enable patients to make transport arrangements, and to let anxious family members know roughly how long they would be at the hospital. It would also help to reassure them that they had not been forgotten.

Our prototype design for such a system included a range of services we identified as potentially useful to the patients in this setting. The combination of the patients' requirements and the requirements of the physical setting in the hospital suggested a system that offered a mixture of ubiquitous and mobile functionality using a variety of modalities and devices. Hence, it provided a useful example domain for our experimental evaluation of the effect of the presence or absence of a GUI on participants' use of our combined gesture and speech interaction techniques.

3 Input and output techniques for mobile and ubiquitous systems

Given the inadequacies of traditional desktop input techniques in a ubiquitous computing environment and even more so with mobile and wearable computing, there has been considerable research investigating alternative techniques [e.g. 12, 13]. Prominent amongst these is gesture or stroke based input [14]. Furthermore, speech output has been considered as an alternative to visual output, and advances in text-to-speech technology have made the use of speech output more realistic [15].

3.1 An input technique for mobile, wearable and ubiquitous systems

Gesture input has formed the basis for many of the input techniques used with PDAs, whether in the form of touchscreen strokes to perform commands or in the form of alphabets, such as Graffiti on the Palm range of PDAs. Its range of uses over many years illustrates a key feature of stroke recognition as an input technique: it is not tightly bound to a particular device. Pursuing our goal of decoupling interaction techniques from the physical characteristics of particular devices, we have exploited this feature to develop an input technique that can be used seamlessly across a wide range of devices in a mobile-populated, ubiquitous computing world

Designing for independence from the diverse characteristics of such devices, and potential future devices, imposes key requirements on such an interaction technique. At one end of the scale, the user may wish to interact with a device as limited in processing power and surface area as a smart ring or credit card, perhaps using

a stylus to make the gestures. At the other end of the scale, the user may wish to interact with a wall-size display, perhaps using the smart ring itself, or indeed using just the user's hand, to make the gestures in the air. Igarashi et al. [16] present a framework for using wall-sized displays using pen input. They describe how defining different application behaviours can provide a means of dealing with input strokes in different ways.

In our recent work [3], we have developed a technique for recognizing input strokes which can be used successfully on a wide range of devices right across this scale. Previously, we have demonstrated the technique with mouse input on a desktop computer, stylus and touch screen input on a wearable computer and hand movement input using real-time video capture. We have termed our technique Directional Stroke Recognition (DSR). As its name implies, it uses strokes as a means of accepting input and commands from the user. In this section we give a brief synopsis of how our technique works and in which situations it can be utilized. A fuller description of the technique is available in [3].

The technique is based exclusively on the direction of strokes and discards other characteristics such as the position of a stroke or the relative positions of many strokes. The algorithm is given an ordered set of coordinates (x, y) that describes the path of the performed stroke. These coordinates may be generated in a number of different ways, including conventional pointing devices such as mice and touch screens, but also smart cards, smart rings, and visual object tracking. The coordinates are then translated into a 'signature' which is a symbolic representation of the stroke. For instance, an L-shaped stroke could have a signature of 'South, East'. This signature can then be looked up against a table of pre-defined commands, much as a mouse button double-click has a different result in different contexts. An advantage of using only the direction of the strokes is that a complex stroke may be broken down into a series of simpler strokes that can be performed in situations with very limited input space (Fig. 1).

The flexibility of our method allows switching between input devices and methods with no need to learn a new interaction technique. For example, someone may at one moment wish to interact with her PDA using a common set of gestures and in the next moment move seamlessly to interacting with a wall display using the same set of gestures. At one moment the PDA provides the interaction area on which the gestures are made using a stylus; in the next moment the PDA itself becomes the 'stylus' as it is waved in the air, during the interaction with the wall display. Any object or device that can provide a meaningful way of generating coordinates and directions can provide input to the gesture recognition algorithm (Fig. 2).

Some important characteristics of this technique include the ability for users to choose the scale and nature of the interaction space they create [17, 18], thus influencing the privacy of their interaction and others' awareness of it. In addition, the physical manifestation

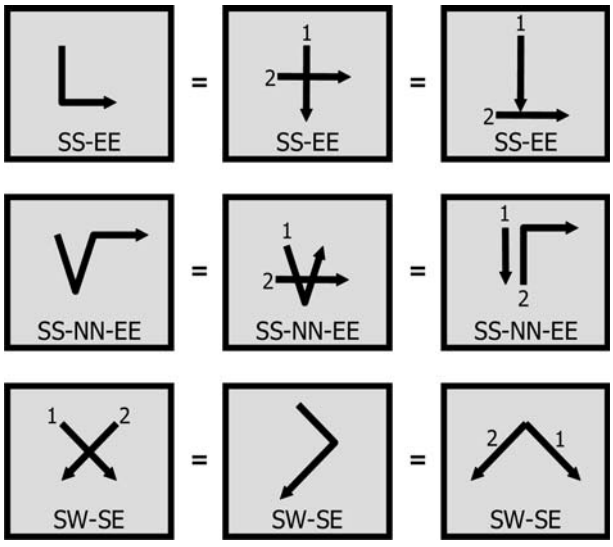

Fig. 1 The recognition algorithm allows a signature to be accessed via different strokes

of our interaction technique can be tailored according to the situation's requirements. As a result, the technique also allows for easy access, literally just walking up to a system and using it, with no need for special equipment on the part of the users. This makes the technique very suitable for use in domains such as a hospital.

The Directional Stroke Recognition technique is flexible enough to accommodate a range of technologies (and their physical forms) yet provide the same functionality wherever used. Thus, issues concerning physical form may be addressed independently. In contrast, standard GUI based interaction techniques are closely tied to physical form: mouse, keyboard and monitor. The technique we have described goes a long way towards the separation of physical form and interaction technique.

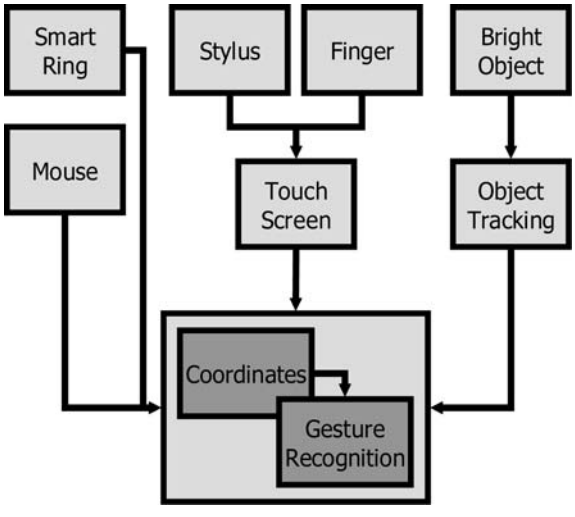

Fig. 2 Using various techniques with the stroke recognition engine

3.2 Speech as an output technique for mobile and ubiquitous systems

For the purposes of our experiment, we decided to use solely speech to present output to the users of our system. In exploring alternatives to the visual display based GUI paradigm, gesture is a primary candidate as an input technique, while speech is maturing as a viable output technique. Much of the research on speech output has been done in the area of assistive technologies, largely due to its relevance to visually impaired users [e.g. 6]. However, speech output has also been proposed for mobile and ubiquitous systems targeted at users who are not disabled [e.g. 19]. The motivations for this research largely reflect the issues raised in Sect. 1 around the usability of GUIs in such systems. As described above, we have chosen to use speech as a more appropriate output technique in mobile and ubiquitous systems for users who are not visually impaired. However, our experimental condition of not having a visual display available for users is quite similar to studying visually impaired users, for whom a visual display is not available. In a study of speech-augmented ATM machines [20], Manzke suggests that speech output should be provided in short sequences and at an appropriate pace. Furthermore, Ross and Blasch [21] identified timing as a key issue with speech output used in a system for blind users. In our study, short sequences of speech output were presented at a pace determined by the user's progress through a set of simple tasks. Each discrete speech output was triggered by a user's input gesture, so the timing was tightly bound to the user's own actions.

We are currently witnessing major advances in speech output technology. Industry standards, such as the Microsoft Speech API and .Net Speech, have helped the widespread use of speech for output purposes. Furthermore, a lot of research is directed at improving the quality of speech output itself, and the results are quite impressive. A range of companies now offers text-to-speech dictionaries and voices, such as AT&T's Natural Voices, ScanSoft's RealSpeak and similar packages by IBM, as well as the Microsoft Windows XP built in text-to-speech engine.

For our experiment, we used the Delphi development environment, and interfaced the Microsoft Speech API (version 5.1) to generate speech on a Windows XP machine. Our design included a simple program that listened to a network port for a text string. Upon receiving a string, it simply played it back through the speakers using the text to speech engine. This set up allowed us to decouple the output issues from other concerns simply by adopting this simplistic network communication protocol.

4 Evaluating the effect of a GUI on gesture and speech interaction

Our field studies identified requirements for the provision of information to patients (and their relatives) waiting for treatment in the A&E department. Our related work suggests providing this information via integrated mobile, wearable and ubiquitous technologies. Our concerns, outlined above, to test the usability of interaction techniques in the absence of visual displays led us to develop a prototype system for providing information to A&E patients through a combination of gesture input and speech output. We used our DSR technique for the gesture input and speech synthesis for the output. We ran an experimental evaluation of this prototype system. The main question addressed by the evaluation was: if we move away from the standard desktop GUI paradigm and its focus on the visual display, do we decrease usability by losing a major benefit

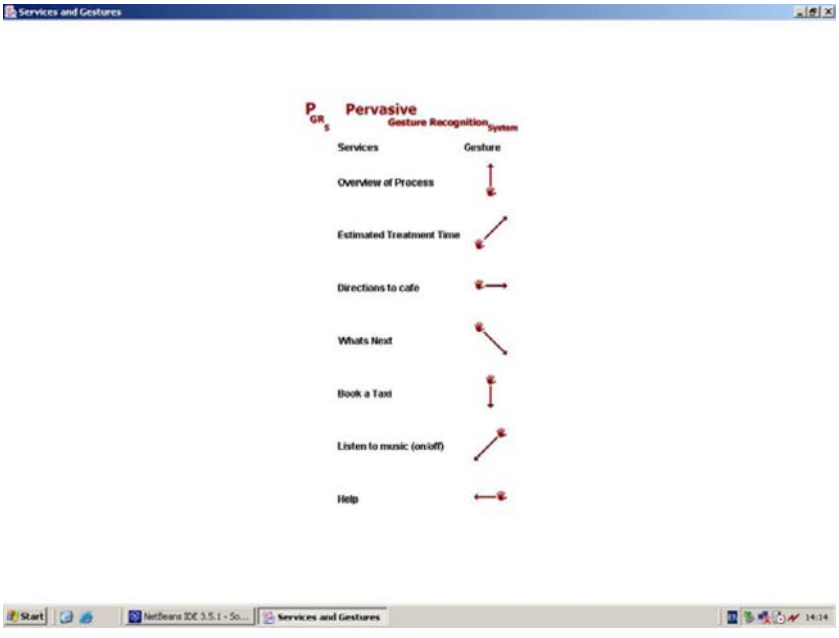

Fig. 3 The 'Services and Gestures' visual display

that the GUI brought, i.e. being able to see the currently available functionality and how to invoke it?

4.1 Method

4.1.1 Design

The experiment had a between participants design. There were two factors, each of which had two levels. The independent variables were: (i) whether the 'Services and Gestures' screen was visible to the participant or not; and (ii) whether the participant had 5 or 10 min training time. The dependent variables were: (i) the time it took a participant to perform a gesture; this was the time in seconds from the end of an instruction being given, to the correct gesture being performed; and (ii) the number of incorrect gestures performed by the participant. It was predicted that participants who could see the 'Services and Gestures' screen (see Fig. 3) would: (i) perform the gestures more quickly; and (ii) perform fewer incorrect gestures than those participants who could not see the screen. It was also predicted that participants who had 10 min training time would: (i) perform the gestures more quickly; and (ii) perform fewer incorrect gestures than those participants who had 5 min training time.

4.1.2 Participants

A total of thirty two participants took part in this experiment, eight per condition. 18 of the participants were male. The participants varied in age within the range 21 to 50, with a mean age of 28.4 years. The participants were all from the University of Bath, varying in occupation and attached to various departments (although the majority of the participants were from the Department of Computer Science). The participants were mainly recruited personally by the experimenters from the Department of Computer Science, with others responding to an e-mail sent to various mailing lists within the University asking for volunteers. None of the participants was given financial incentives to take part in the experiment.

4.1.3 Apparatus

The set up of the apparatus can be seen in Figs. 4 and 6. In line with our goal of decoupling interaction technique from device, our gesture system was designed ab initio to be used with diverse interaction devices, from a standard desktop combination of mouse and monitor to real time tracking of the user's hand. Our current implementation using camera based tracking is rather slow due to our inefficient image processing so, for the purposes of this study, we used a standard mouse and monitor set up. It was not our intention here to evaluate whether or not the mouse is the most appropriate input device. Our emphasis on decoupling interaction technique from device renders the mouse as good (or bad) as any other device for making our gestures and the particular input device used in this study has no bearing on the study's hypotheses, independent variables or dependent variables. If, however, we were to move from a prototype system in a controlled experimental setting to a deliverable system in a setting closer to the application domain, we should need to investigate the relative merits of different devices in the hospital setting.

Each participant sat in front of a standard 15'' monitor with a Microsoft IR mouse positioned on a small table directly in front of the monitor (see Fig. 4). The monitor and mouse were powered by a computer enabled with an 802.11b wireless connection. This computer also ran our Gesture Client software which allowed the participant to input gestures using the

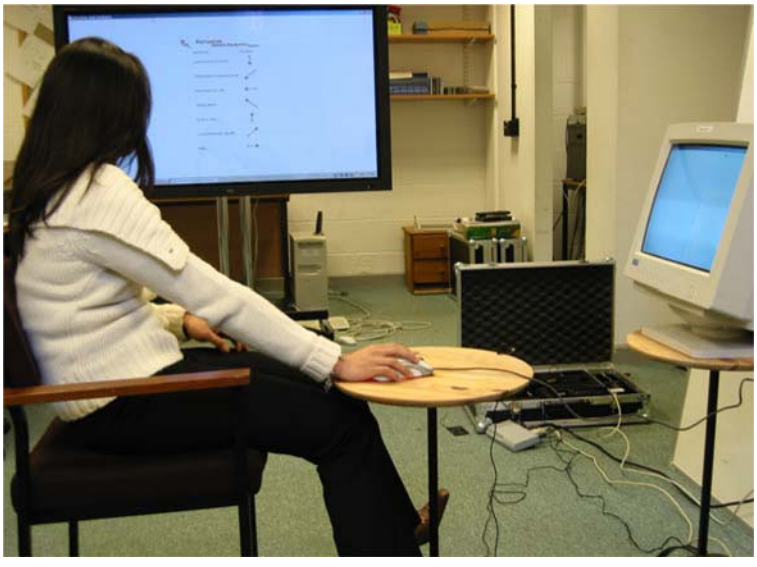

Fig. 4 A participant with the mouse and the Gesture Client displayed on a 15'' monitor, with the 'Services and Gestures' visual display on the 61'' plasma screen to the participant's left

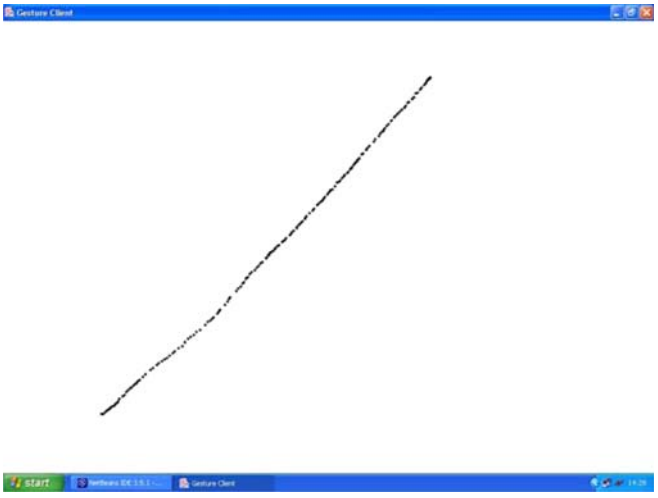

Fig. 5 The Gesture Client screen showing a gesture made by the user

mouse. The Gesture Client displays a blank, white canvas to the participant allowing her to perform a gesture by pressing the left mouse button, moving the mouse in the desired direction, and finally releasing the left mouse button. A trail of black dots appears on the canvas as a gesture is performed (see Fig. 5). This gave the participant visual feedback of the gesture being made. All participants, regardless of condition, were provided with this feedback to try to ensure that the dependent variables were influenced only by whether or not the participants could see the 'Services and Gestures' presented on another screen. When the left mouse button is released, a string is sent via an 802.11b connection to a program running on the evaluators' PC, informing it of the gesture performed.

Directly to the left of the participant was a 61'' NEC plasma screen which displayed to the participant the 'Services and Gestures' visual display (see Fig. 4). This was a simple GUI that presented the services available to the participants and the gestures that invoked each service. Its placement allowed us easily to observe when the participant was looking at the visual display. The services provided were identified from our fieldwork in the hospital as commonly desired by patients or their relatives [4]. There were seven services available:

- Overview of process (which provided information about the process – triage etc—that a patient would go through upon arriving at the A&E department, thus helping to explain the sometimes very long waiting times);
- Estimated treatment time;
- Directions to café;
- What's next (which provided real-time updated information about which stage in the process a patient – or more likely a patient's file – had reached and what the patient could expect as the next step);
- Book a taxi;
- Listen to music (on/off); and
- Help (which reminded the participant—via speech output—of the available services and corresponding gestures).The seven gestures that invoked these services were, respectively, North; Northeast; East; Southeast; South; Southwest; and West. (Note that, for example, Northeast is a single stroke, whereas North, East is a combination of two strokes, North followed by East. Details of combining the strokes are reported in [3].) The Services and Gestures program was run on a stand-alone desktop PC. A screen shot of the Services and Gestures display is reproduced in Fig. 3 above.

Directly to the right of the participant was the evaluators' table (see Fig. 6). On this table was the evaluators' computer, a standard desktop PC enabled with an 802.11b wireless connection. This PC was used to run the software that received the gesture signature sent from the participant's computer via the 802.11b wireless connection. This allowed the program to determine which gesture had been input by the participant and to give the appropriate audio feedback generated using a Delphi Chat Server and the Microsoft Speech SAPI 5.1 through the PC's stereo speakers.

Also on the evaluators' table was a Sony Vaio notebook computer with desktop stereo speakers, which ran Windows Media Player to play the scenario and instructions for the experiment (see Appendix). The

Fig. 6 The evaluators' table with a desktop PC providing the speech output responses to the participants' gestures and a notebook PC used to play instructions and music

scenario and instructions could simply have been read out by the evaluators. However, we chose to have the computer play them from pre-recorded files in order to ensure consistency of content, tone and speed of delivery across all the trials. Winamp was also used on this computer to play a music audio file whenever a participant performed the gesture for music to be played during the evaluation scenario.

A video camera recorded the participant's activities. A clip-on microphone was attached to the participant. A scan converter captured the gestures from the 15" monitor in front of the participant. The signals from the video camera and the scan converter were fed to a Picture-in-Picture unit. The resulting combined video signal and the audio signal from the microphone were then fed to an MPEG converter. The latter recorded the resulting MPEG file to a hard disk for later analysis.

4.1.4 Procedure

Participants were run individually in a controlled laboratory. Upon entering the lab, the participant sat in front of a 15" monitor displaying the Gesture Client, the microphone was attached to the participant's clothes and the video camera was adjusted if necessary. One experimenter (Evaluator 1) was positioned to the right of the participant at the evaluators' desk. Another experimenter (Evaluator 2) was positioned to the left of the participant to control the video recording. The equipment and software were then described to the participant, and the participant was informed of the experimental procedure.

Before the actual experiment started, the participant was played an audio file that presented a background scenario to set the scene and informed her about the experiment (see Appendix). The participant was then given 5 or 10 min training time (depending on the condition) in which she could use the Gesture Client, study the services available, practise making the associated gestures, and ask the evaluators questions. Once the training time was over the 'Services and Gestures' screen was turned off or remained showing (depending on the condition) and any final questions were dealt with. Once both the experimenters and participant were satisfied, the participant was informed that the actual experiment was about to commence.

Evaluator 2 started recording and Evaluator 1 began the audio files that talked the participant through the scenario and asked the participant to perform a gesture when required (see Appendix). Whenever an instruction was given to the participant, the audio file was paused while the participant performed the gesture. The participant used the mouse to perform a gesture by holding the left mouse button down, dragging the mouse in the appropriate direction and releasing the left mouse button. The gesture was then accepted by the system and the corresponding speech output response was automatically presented to the participant by the software.

Depending on the gesture performed, there were three possible results:

(i) the correct speech output response was played because the correct gesture was made;
(ii) an incorrect speech output response was played due to a wrong gesture being performed; and
(iii) an 'unrecognised gesture' message was played because the user's gesture could not be recognised by the system.In principle, there is a fourth possibility that the user could attempt to make one gesture but the system recognises it as a different gesture. However, we had the facility to calibrate our DSR system to minimise misrecognition of gestures and, in practice, there were no misrecognitions in our experimental trials.

For example, if the user was instructed to book a taxi and she made the correct gesture 'South', the speech output response would be: 'Your taxi has been booked'. On the other hand, if the user instead made the gesture East, an incorrect response would be output (in this case, directions to the café). If an incorrect speech output response was played or the 'unrecognised gesture' message was played, the participant would perform another gesture until she performed the correct gesture. When the correct gesture was performed by the participant, the scenario continued. This continued until the scenario was completed.

4.2 Results

The number of incorrect gestures and the processing times for producing correct gestures were calculated for each participant.

4.2.1 Incorrect gestures performed

The number of incorrect gestures from both the visual display and no visual display conditions were calculated for each participant. 'Incorrect gestures' did not include gestures that were unrecognised by the system and, as noted in Sect. 4.1.4, there were no misrecognised gestures. The mean number of incorrect gestures produced in response to each instruction is shown in Table 1, for all four conditions. Standard deviations are given in parentheses for each.

The data were analysed using a 2-way unrelated ANOVA. There was no significant main effect of screen presence ($F_{1,444} = 1.526$, $P = 0.217$). There was also no significant main effect of training time ($F_{1,444} = 0.549$, $P = 0.459$).

However, the screen presence × training time interaction was significant ($F_{1,444} = 4.943$, $P = 0.027$). For participants who could not see the Services and Gestures visual display, additional training time decreased the number of incorrect gestures. Intriguingly, for participants who could see the Services and Gestures visual

Table 1 Mean (and SD) incorrect gestures per instruction (n = 112 for each condition)

Screen presence	Training time	
	5 min	10 min
With screen	0.01 (0.09)	0.04 (0.19)
Without screen	0.07 (0.29)	0.02 (0.13)

display, additional training time increased the number of incorrect gestures.

4.2.2 *Processing time*

The processing time (in seconds) from both screen and no screen conditions were calculated for each participant. We defined processing time as the time taken from the end of an instruction until the start of the automated audio feedback for a correct gesture, disregarding the time it took to perform unrecognised gestures. In Table 2, we show the mean processing time per instruction, i.e. how long it took participants, on average, to complete each instruction given to them. Again, standard deviations are given in parentheses.

The data were analysed using a 2-way unrelated ANOVA. There was no significant main effect of training time ($F_{1,444} = 0.80$, $P = 0.372$). However, there was a significant main effect of screen presence ($F_{1,444} = 7.593$, $P = 0.006$). As may be seen in Fig. 8, mean processing time was longer in the presence of the Services and Gestures visual display than in its absence.

The screen presence × training time interaction was not significant ($F_{1,444} = 0.001$, $P = 0.979$).

The relatively large standard deviations in some cells of Tables 1 and 2 may be the result of a few outliers. This is likely given that we were not dealing with particularly large sample sizes.

5 Discussion

Users' performance was assessed in terms of (i) time taken to make gestures in response to instructions given by the experimenters, and (ii) the number of incorrect gestures made. Only one of the main effects reached statistical significance, and this was in the opposite direction to the one-tailed experimental hypothesis (H1) which predicted that users who could see the Services and Gestures visual display would perform gestures more quickly than users who could not see the visual display. There was no significant interaction effect on instruction processing time between visual display presence and training time. There was a significant interaction effect between these two factors on the number of incorrect gestures performed.

Table 2 Mean processing time in seconds (and SD) per instruction (n = 112 for each condition)

Screen presence	Training time	
	5 min	10 min
With screen	4.67 (4.85)	4.36 (3.42)
Without screen	3.72 (3.31)	3.42 (2.57)

5.1 Visual display *versus* no visual display

Significant main effects in the direction of Hypothesis 1 would have reinforced our concern that forsaking the standard GUI paradigm and its associated visual display will impair usability. Instead, we found that participants who could see the Services and Gestures visual display made just as many incorrect gestures as participants who could not see this display, but took significantly longer to do so. Hence, the question arises: why should participants with the visual display take longer to achieve the same gesture performance level as participants without the display? The required tasks were the same in both conditions, so it is not the case that participants with the visual display had more difficult tasks and therefore took longer to perform them. Even if that were the case, it assumes that participants would conscientiously take longer over the more difficult tasks in order to get them right. In fact, we would expect some of that effect but also some effect of participants simply getting the more difficult tasks wrong, resulting in higher incorrect gesture rates. This was not found.

Another explanation for the longer processing times for participants with the Services and Gestures visual display is that those who had the display chose to spend time looking at it (even though it did not improve their performance). This is borne out by the data: on average, participants with the Services and Gestures visual display spent 58 sec looking at it. Obviously, those without the display spent no time looking at it.

The question then becomes: why did those without the screen not make more incorrect gestures? Quite simply, they remembered correctly and therefore were not disadvantaged by the lack of a visual display to remind them of the services and corresponding gestures. So what explains their good memory for the services and gestures? In the setting of our study, we had a small set of available services, a small set of simple gestures in a memorable pattern, a highly constrained user context and semantically very distinct services. All of these features may have contributed to assisting the users in the absence of a visual display of the available services and gestures.

There were only seven services available: Overview of process; Estimated treatment time; Directions to café; What's next; Book a taxi; Listen to music (on/off); and Help. It is generally accepted that most people have little difficulty in working with this number of distinct items [22]. The gestures themselves were deliberately simple, consisting of a single stroke. These gestures were related in a memorable pattern, i.e. the points of the compass,

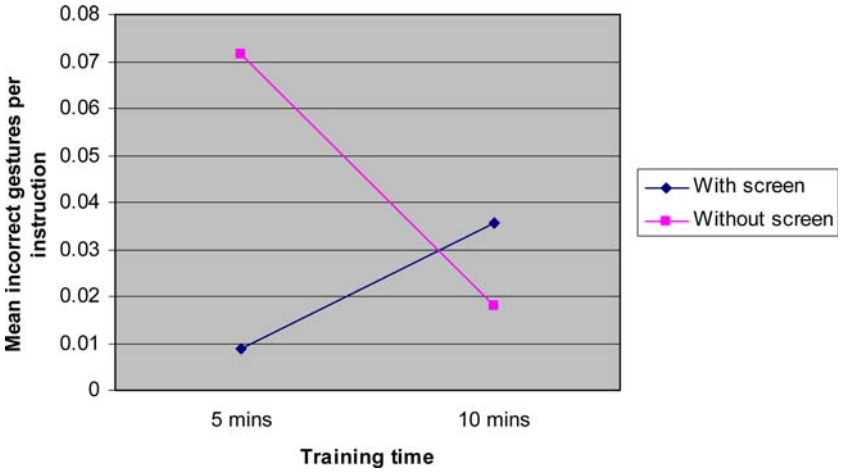

Fig. 7 Interaction of factors on number of incorrect gestures

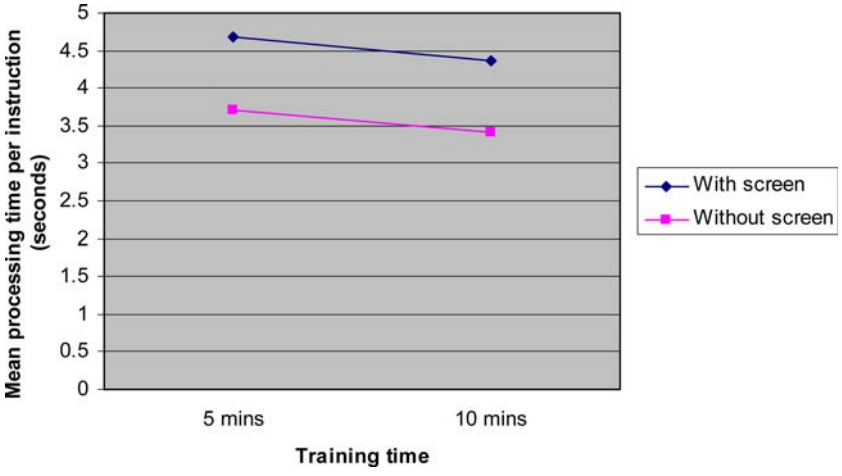

Fig. 8 Interaction of factors on instruction processing time

and while services had been arbitrarily assigned to particular gestures, some participants reported imposing their own semantic groupings or relationships between gestures and services, in a form of 'method of loci', a memorisation technique with a very long reported history [23].

In addition, the services we offered were semantically very distinct from each other. For example, asking for directions to the café is very distinct from booking a taxi. This dissimilarity amongst the available services should help users to remember and choose between them, compared to a situation with a number of semantically similar options.

The user context in our study was patients attending a hospital A&E department. This greatly constrains the range of services that users might wish to use and even more tightly constrains those that they typically would expect to find. The things one currently can do in a hospital while waiting for treatment are fairly limited. Hence, a participant might well expect that 'Estimated treatment time' would be one of the options but might dismiss the notion that 'Trade my personal share portfolio' might be available. Note that this is not to suggest that we should not be offering options such as the latter, simply that currently most people in the A&E setting would not expect it to be available. We take up this issue of wider ranging tasks in Sect. 6.

There is another difficulty in interpreting our data that is caused by the small number of gestures and limited range of contextually suggested services. Having decided, for example, that 'Estimated treatment time' is a possible option in the context of the A&E domain, with only seven gestures available there is a 14% chance of making the correct gesture randomly. Unfortunately, we cannot tell to what extent any of our participants were making random gestures in response to the instructions they were receiving.

All of these features could serve to create a 'ceiling' effect on both measures of users' performance. That is, in the absence of the visual display of services and gestures, users would be helped to remember them, resulting in fewer incorrect gestures and reduced processing time. These features could well have contributed to nullifying the advantage claimed for the GUI paradigm:

that of providing a visual cue of the available services and means of accessing them. The very low number of incorrect gestures recorded is suggestive of a ceiling effect. It is possible that scaling up the complexity of the interaction would generate more incorrect gestures, and indeed misrecognised gestures. This issue is taken up in Sect. 6. But the point remains that within the constraints discussed here, the presence of a GUI does not reduce the occurrence of incorrect gestures and actually leads to longer processing times.

5.2 Training time

A significant main effect in the direction of Hypothesis 2 would have reinforced the conventional assumption that longer training time will improve users' performance. It is worth asking what may have prevented the main effect from reaching significance. A possible explanation is that the difference between 5 and 10 min was not enough, especially since many of the participants in the 10 min training condition were observed to 'waste' the training time, seemingly having become bored or decided they had learned all they could after a few minutes. There was, therefore, inadequate control to ensure that the training effect of the 10 min was actually double that of the 5 min. Further investigation of a training effect should require greater differences in both training time and the nature of the training itself, with pre- and post-tests, to be more confident that one group of participants were reliably being trained more effectively than the other.

The results are interesting, however, from another perspective. The kind of public information system that we discuss in Sect. 2 could be expected to be usable on a 'walk up and use' basis, without the requirement for user training. If such a system were highly usable on that basis, one should expect no significant effect of training time, at least over the periods used in this study.

5.3 Interaction effects

There was no significant interaction effect on instruction processing time between visual display presence and training time. Intriguingly, there was a significant interaction effect between these two factors on the number of incorrect gestures performed. For participants who could not see the Services and Gestures visual display, the additional training time (10 min rather than 5) decreased the number of incorrect gestures. This is not an unexpected result and may be explained by the additional training time compensating for the adverse effect of not being able to see the visual display, an effect that would be in line with the original hypotheses.

However, for participants who could see the Services and Gestures visual display, additional training time increased the number of incorrect gestures. There is no clear explanation for this effect and it may well be a statistical anomaly. Unfortunately, the putative ceiling effect, leading to the low overall numbers of incorrect gestures, prevents further meaningful analysis.

5.4 Generalisability of findings

In considering the generalisability of our findings, one aspect to consider is the ecological validity of the study, i.e. the extent to which findings may be generalised from experimental settings to 'real world' situations. Clearly, our study was not one of people using mobile and ubiquitous systems in a real hospital environment. Rather, our participants were made to sit in a usability laboratory, alone but for the experimenters, and using a conventional mouse to make their gestures.

As noted in Sect. 4.1.3, the particular input device used in this study has no bearing on the study's hypotheses, independent variables or dependent variables. Moreover, in related work, we have shown degradation in performance when participants used 'unconventional' I/O devices, so not using a mouse here could have introduced another possible confounding factor. In so far as we expect users not to use a conventional mouse in many mobile and ubiquitous computing situations, the desire for experimental control conflicts with the desire for ecological validity. This is a perennial issue with such experimental studies and the experimenter must decide where the balance lies in designing a study [24].

Similarly in the experimental setting, participants could see immediate feedback of their gesture on the 15'' monitor in front of them (see Figs. 4 and 5). It is possible that this feedback may have improved the participants' performance of the gestures. In a 'real world' setting of mobile and ubiquitous systems without displays, users would not have this feedback. In the experimental setting, we could have prevented the participants from seeing the trace of their gesture simply by not providing them with the monitor. However, if we had one group with it and one group without it, there would have been an obvious confounding risk. If we had both groups without it, there would have been the risk that participants' performance would degrade *because* they had no real-time feedback of their gesture. Since all conditions might degrade to the same extent (though we cannot assume even this), this should not give the same kind of confusion as the first case but could give a falsely large degradation in performance across the board for the hypotheses we were tackling.

Hence, we could have increased the ecological validity of the study by not using a mouse for input or by giving the participants no visual feedback of their gestures, and added potential confounding factors in the process, but this kind of lab-based experimental study has real worth only if it sticks to its methodology of teasing out and testing one factor at a time. Our knowledge and understanding of non-GUI interaction in a mobile and ubiquitous computing world is still

immature and so we need, amongst other approaches, systematic lab-based experiments to answer some of the basic questions. There is little point in increasing ecological validity at the expense of confidence in our results since ecological validity then becomes largely irrelevant.

We are concerned that it is difficult to scale up the kind of interaction investigated in this study. In addition to the problem of adding to what the user has to remember and so increasing the likelihood of incorrect gestures, increasing the number of available services and corresponding gestures also increases the likelihood of misrecognised gestures. The absence in this study of gesture misrecognitions by the system is in part due to our DSR approach of dealing with simple directional input. However, this was aided by the restriction here to a small number of highly distinct gestures. This made for straightforward and effective calibration of the system and inputs from the user that were relatively easily distinguished. But taking up to eight points on a compass rose as distinct directional gestures gives 45° between each direction. Adding another option between existing gestures would reduce this to 22.5° and so increase the likelihood of misrecognition.

Our initial concern remains for the development of non-visual interaction techniques for *general* use in a mobile and ubiquitous computing world. Our DSR technique for gestural input can handle arbitrarily complex gestures comprised of multiple strokes. There is no requirement for it to be confined to simple single strokes to compass points. Its potential for much richer syntax coincides with the requirement for much richer semantics in general purpose mobile devices. The user may wish to trade her personal share portfolio, and check her email, and book cinema tickets, etc, etc, while she endures the long wait in A&E. A user may build a repertoire of gestures, some simple, some complex, that she uses frequently. She will not want to leave them behind at the door of the hospital.

6 Conclusion and future work

The results of the evaluation reported here may be interpreted as good news for those developers of multimodal interaction who want to mitigate our reliance on the increasingly unsuitable visual displays of mobile and wearable devices and ubiquitous systems. We found no significant evidence that usability suffered in the absence of one of the major benefits of the GUI paradigm: a visual display of available services and how to access them. Indeed, we found that spending time looking at this visual display took longer but did not lead to better performance by users of a gesture and speech based system. However, we must sound a note of caution. The results of this study should not lead us to conclude that in general we may happily develop interaction techniques involving non-visual modalities, safe in the knowledge that the loss of the visual display based GUI will not impair usability. Our study suggests only that *with particular constraints*, the effects of losing the cognitive support provided by a standard GUI visual display are mitigated. These constraints include having a small set of available functions, a small set of simple input gestures in a memorable pattern (e.g. the points of the compass), a tightly constrained user context and semantically very distinct functions.

A potential strategy for generalising the multimodal interaction method demonstrated here is to develop a range of local gesture-service sets. Thus, in the tightly constrained user context of the A&E department, the user could have access to a small set of simple, distinct services invoked by a small set of simple, distinct gestures, as in this study. When the user then goes to the bank or the supermarket or the garage, she could, in each local context, be offered a similar arrangement. However, in the absence of visual displays, how would the user know which services and gestures were available in the local context? A *de facto* standard might develop out of a basic set of gestures, such as the compass points. This would mitigate the problem of knowing what gestures were available in a new setting. But the problem remains of mapping particular services to the gestures. The same small set of simple services is unlikely to be sufficiently central to the activities in, for example, a hospital's A&E department and a bank for both local settings to map the same services to the same few simple gestures.

In addition to tackling these practical problems, further work is required to evaluate the use of gesture and voice based interaction in more general use. Specifically, we are planning further research to investigate whether or not there was a ceiling effect in this study, i.e. whether the predicted poor usability in the absence of a visual display failed to reach significant levels due to the influences discussed in Sect. 5. To test the range of issues raised by this paper, further studies could be conducted with the same hypotheses but with a greater range of services available, less semantically distinct services, a larger set of gestures, more complex gestures, a less tightly constrained context of use, and more complex and longer speech outputs. Indeed, a greater range of services and similarities amongst the services offered in itself will require further research on the number and complexity of gestures that could be usable, with or without a GUI's visual display.

It is also worth conducting a further study in which the participants do not have visual feedback of their gestures (see Fig. 5). Such feedback was delivered in this study via the 15'' display in front of the participant (see Fig. 4) to try to ensure that the participant's performance was influenced only by whether or not she was presented with the GUI's hypothetical advantage of visually cueing the services available and how to invoke them. However, it is possible that the real time visual feedback of the gesture being made helps the user, possibly with self-calibration of her gestures. In this case, one would expect performance to deteriorate in the absence of this feedback. This would need to be inves-

tigated independently of the effects studied here in order to avoid confounding.

We are currently planning a follow-up study to investigate a further aspect of the paradigm shift from command line to GUI that has not been explored here. We have described an advantage that the GUI paradigm brought in presenting the user with visual cues to available services and how to invoke them. This study tested the proposal that usability would suffer if we lost this advantage in moving to a mobile and ubiquitous computing world in which GUIs are restricted or absent. There is, however, another advantage that GUIs brought: they provide visual cues to the system's *state*. Once again, we might predict that removing the GUI would remove this advantage and so render gesture and speech based systems less usable. This hypothesis was not tested in this study as the patient information system was effectively stateless. Once again, this potential effect needs to be tested independently and will require the development of a different testbed application.

Clearly there are a large number of factors with potential influence on the usability of multimodal interaction in the presence or absence of GUIs. It is necessary to tease these out in a series of independent but related studies, as suggested in this section. The study reported here offers the first intriguing results in this series and suggests that in the right conditions we can do without GUIs.

Acknowledgements The research reported here is part of a UK EPSRC-funded research project 'Designing for common ground in mobile distributed collaborative systems', award number GR/R24562/01. We thank Hilary Johnson and Leon Watts for their advice and insightful comments.

7 Appendix

Transcript of the speech output file played to participants in the experimental evaluation. The text here was prerecorded. Speech output responses to participants' gestures were generated in real time based on which gestures were made.

7.1 Gesture Interaction in a Hospital Waiting Room

7.1.1 Background

You have just come to the hospital with a very painful wrist and you want to have it checked over by a doctor. You have never really had the need to go to the hospital before and therefore are unsure of the process of being seen by a doctor.

Upon entering the hospital the first thing you see is a Reception desk. You are greeted by a nurse at the desk who takes your personal details and asks you to wait. You sit down in the waiting area. On the wall of the waiting area is a large computer display, listing services that are available to you and the gesture that invokes each service. When you make a gesture, the service is delivered to you via loudspeakers.

7.1.2 Evaluation

We are running a test of the gesture based system to see how it performs and how usable people find it. Please note that we are evaluating the performance of the computer system, not of you. During the evaluation you will be asked to perform specific gestures, which will provide you with particular services during your time at the hospital. Please note that, in addition to these requested gestures, you are able to perform the 'help' gesture at any time to hear what services are available to you and their associated gestures.

You now have < 5/10 min > training time to familiarise yourself with the functionality and their associated gestures. If you have any questions please ask one of the evaluators before the experiment starts.

7.1.3 The Tasks

As you are unfamiliar with the hospital process, you wish to hear an overview of what is involved.

1. Perform a gesture to hear an overview of the hospital process.

   When you gave your details at Reception, you were asked to take a seat in the waiting room and told a nurse will see you as soon as possible. You are impatient and would like information on how long you will have to wait before treatment.

2. Perform a gesture to hear the estimated treatment time.

   The estimated treatment time is quite long, so you decide to go and get a drink from the café.

3. Perform a gesture to get directions to the hospital café.

   After getting your drink you return to the waiting room and continue to wait. While you are waiting you decide to listen to some music.

4. Perform a gesture to listen to some music.

   You are now called by the nurse for a pre-examination of your injury to see how serious it is. You don't want to be distracted by the music while you are talking with the nurse.

5. Perform a gesture to turn the music off.

While your injury seems to be not too serious, the nurse would like you to be seen by a doctor in case of a possible fracture. The nurse also suggests that you may need to be x-rayed. The nurse asks you to take a seat in the waiting room and tells you a doctor will see you as soon as possible.

While in the waiting room, you wish to pass the time and therefore decide to listen to some music again.

6. Perform a gesture to listen to some music.

After waiting for some time, you wonder if you are meant to do anything before seeing the doctor, such as going for the x-rays suggested by the nurse.

7. Perform a gesture to hear the next step in the process.

The doctor comes to the waiting room and calls you to be seen.

8. Perform a gesture to turn the music off.

You follow the doctor to a treatment cubicle. After examining you, the doctor decides that your wrist must be x-rayed. You are told to wait in the radiology waiting room.

9. Perform a gesture to listen to some music.

After a short wait there, you are called by a radiographer.

10. Perform a gesture to turn the music off.

Your wrist is x-rayed and you are again asked to wait in the radiology waiting room. You are now unsure what you should be doing next.

11. Perform a gesture to hear the next step in the process.

The radiographer returns and gives you your x-ray plates. You return to the treatment cubicle. The doctor is satisfied that your wrist is not broken but is just slightly sprained. The doctor leaves you with a nurse who bandages your wrist and is then called away.

You are now unsure if it is all right for you to leave, or if you have to make a follow up appointment.

12. Perform a gesture to hear the next step in the process.

You now know that you have been discharged and are therefore able to leave the hospital and go home.

13. Perform a gesture to book a taxi.

While you are waiting for your taxi you decide to go to the hospital café and get some breakfast. However you have forgotten where the café is.

14. Perform a gesture to get directions to the hospital café.

You get some breakfast from the café and your taxi arrives to take you home.

References

1. Brewster SA (2002) Overcoming the lack of screen space on mobile computers. Personal and Ubiquitous Computing 6(3):188–205
2. Brewster SA, Lumsden J, Bell M, Hall M, Tasker S (2003) Multi-modal 'eyes free' interaction techniques for wearable devices. Proc. CHI'03 Conference on Human Factors in Computing Systems, CHI Letters, ACM Press, 5(1):473–480
3. Kostakos V, O'Neill E (2003) A directional stroke recognition technique for mobile interaction in a pervasive computing world. In People and Computer XVII, Proc. HCI 2003: Designing for Society. Bath, UK, 197–206
4. O'Neill E, Woodgate D, Kostakos V (2004) Easing the wait in the Emergency Room: building a theory of public information systems. Proc. DIS'04 Conference on Designing Interactive Systems, ACM Press, pp 17–25
5. James F, Roelands J (2002) Voice over Workplace (VoWP): voice navigation in a complex business GUI. Proc. Fifth International Conference on Assistive Technologies, ACM Press, pp 197–204
6. Amar R, Dow S, Gordon R, Hamid MR, Sellers C (2003) Mobile ADVICE:an accessible device for visually impaired capability enhancement. Extended Abstracts CHI'03 Conference on Human Factors in Computing Systems, CHI Letters, ACM Press 5(1):918–919
7. Xiao Y, Lasome C, Moss, J, Mackenzie C, Faraj S (2001) Cognitive properties of a whiteboard: a case study in a trauma centre. Proc. ECSCW 2001 Seventh European Conference on Computer Supported Cooperative Work Kluwer, pp 259–278
8. Clarke, K., Hughes, J. and Rouncefield, M (2002) When a bed is not a bed: the situated display of knowledge on a hospital ward. In Workshop on Public, Community and Situated Displays, at CSCW 2002, ACM Conference on Computer Supported Cooperative Work. Available at http://www.appliancestudio.com/cscw/papers.htm. Last accessed December 2004
9. Naumann S, Miles JA, (2001) Managing waiting patients' perceptions the role of process control. Journal of Management in Medicine 15(5):376–386
10. Maister DH (1988) The psychology of waiting lines. In: Lovelock J (ed) Managing services: marketing, operations and human resources. Prentice-Hall, pp 176–183
11. Dansky KH, Miles JA (1997) Patient satisfaction with ambulatory healthcare services: waiting time and filling time. Hospital and Health Services Administration 42:165–177
12. Goldstein M, Book R, Alsio G, Tessa S (1999) Non-keyboard QWERTY touch typing: a portable input interface for the mobile user. Proc. CHI'99 Conference on Human Factors in Computing Systems, ACM Press, pp 32–39
13. Wigdor D, Balakrishnan R (2003) TiltText:using tilt for text input to mobile phones. Proc. 16th Annual ACM Symposium on User Interface Software and Technology, ACM press, pp 81–90
14. Pirhonen A, Brewster S, Holguin C (2002) Gestural and audio metaphors as a means of control for mobile devices. Proc. CHI'99 Conference on Human Factors in Computing Systems, ACM Press, pp 291–298

15. Lines L, Hone KS (2002) Older adults' evaluations of speech output. Proc. Fifth International Conference on Assistive Technologies, ACM Press, pp 170–177
16. Igarashi T, Edwards WK, LaMarca A, Mynatt ED, (2000) An architecture for pen-based interaction on electronic whiteboards. Proc. Working Conference on Advanced Visual Interfaces, ACM Press, pp 68–75
17. Kostakos V, O'Neill E (2004) Pervasive computing in emergency situations, Proc. Thirty-Seventh Annual Hawaii International Conference on System Sciences. IEEE Computer Society Press, p.30081b
18. Kostakos V (2004) A design framework for pervasive information systems. PhD thesis. Department of Computer Science, University of Bath. Technical Report CSBU-2005-02, Technical Report Series ISSN-1740-9497
19. Sawhney N, Schmandt C (2000) Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. ACM Transactions on Human-Computer Interaction 7(3):353–383
20. Manzke JM (1998) Adaptation of a cash dispenser to the needs of blind and visually impaired people. Proc. Third International Conference on Assistive Technologies, ACM Press, pp 116–123
21. Ross DA, Blasch BB (2002) Development of a wearable computer orientation system. Personal and Ubiquitous Computing 6(1):49–63
22. Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. The Psychological Review 63:81–97
23. Cicero MT, De Oratore, II lxxxvi. Translated by H. Rackham, pp 350–353
24. Hoc J-M (2001) Towards ecological validity of research in cognitive ergonomics. Theoretical Issues in Ergonomics Science 2(3):278–288