19

The Usability of Speech and Eye Gaze as a Multimodal Interface for a Word Processor

T.R. Beelders and P.J. Blignaut
University of the Free State
South Africa

1. Introduction

Communication between humans and computers is considered to be two-way communication between two powerful processors over a narrow bandwidth (Jacobs and Karn, 2003). Most interfaces today utilise more bandwidth with computer-to-user communication than vice versa, leading to a decidedly one-sided use of the available bandwidth (Jacobs and Karn, 2003). An additional communication mode will invariably provide for an improved interface (Jacobs, 1993) and new input devices which use passive measurements to capture data from the user both conveniently and at a high speed are well suited to provide more balance in the bandwidth disparity (Jacobs and Karn, 2003). In order to better utilise the bandwidth between human and computer, more natural communication which concentrates more on parallel and not sequential communication is required (Jacobs, 1993).

Furthermore, the user interface is the connection between the user and the computer and as such plays a vital role in the success or failure of an application. Modern-day interfaces are entirely graphical and require users to visually acquire and manually manipulate objects on screen (Hatfield and Jenkins, 1997) and the current trend of Windows, Icons, Menu and Pointer (WIMP) interfaces has already been around since the 1970s (van Dam, 2001). Unlike their command line counterparts, these graphical user interfaces are not in the least accessible to users with disabilities and it has become essential that viable alternatives to mouse and keyboard input are found (Hatfield and Jenkins, 1997). Specially designed applications which take users with disabilities into consideration are available but these do not necessarily compare with the more popular applications. This chapter therefore aims to investigate various ways to provide alternative means of input which could facilitate use of the mainstream product by disabled users.

These alternative means should also enhance the user experience for novice, intermediate and expert users. Findings from previous studies (Beelders, 2006; Blignaut, Dednam and Beelders, 2007) show that while novice users of word processors experience a number of obstacles in acceptance and usage of the application that are unique to the demographic, alternative pictorial icons, text buttons and translation of the interface into the native language of the user all failed to lessen the learning curve significantly or to increase usability significantly. However, these findings should not discourage researchers but should serve as encouragement to find more innovative and creative means of alleviating the burden on these users. Particularly since these users show remarkable eagerness and enthusiasm to learn, greater effort should be made to accommodate them to become

mainstream users. Although the main focus could be to narrow the gap between novice and expert users, the means to achieve this should not alienate or disrupt the smooth flow of work that an expert user is capable of achieving. Rather, the improvements should serve not only the novice users but also provide an alternative means for experts as a way to improve their interaction with the product. The study that is reported in this chapter therefore proposes to be an extension or continuation of these aforementioned studies, and investigate further ways to improve the interface of a word processor for all user groups.

The eye-tracker has steadily become more robust and reliable and cheaper and therefore, presents itself as a suitable tool for this use (Jacobs and Karn, 2003). However, much research is still needed to determine the most convenient and suitable means of interaction before the eye-tracker can be fully incorporated as a meaningful input device (Jacobs and Karn, 2003). However, the disadvantages associated with eye-tracking as an input device mean that it should be used with caution or as suggested by Istance, Spinner and Howarth (1996), it should ideally be combined with other input modalities which will provide a means to overcome the limitations of eye-tracking, such as speech. As it is, Microsoft Office already comes bundled with an in-built speech engine which makes speech recognition available in all Office packages. There are also a number of affordable alternative speech engines available on the market. Eye-trackers may eventually become cost-effective enough to be a standard feature in future computing devices (Isokoski, 2000). However, given that the hardware and software is available, the task remains to prove that the eye-tracker improves the quality of human-computer interaction as validation for the inclusion in future devices (Isokoski, 2000). Although neither eye-tracking nor speech recognition is new to usability studies or as a potential source of increased usability, few studies have been found that use a combination of the two in a single package as a means of usability improvement.

Therefore, the aim of this study was to determine whether a multimodal interface, using non-traditional input means could be created for a word processing application. In this way, this popular application can cater for a more diverse group of users through a highly customisable interface. The following section will provide some background literature which serves as a foundation on which this study was based.

2. Background

This section will discuss some of the available literature which was used as a foundation for the study.

2.1 Advantages for users

The high incidence of afflictions such as tendonitis, carpal tunnel syndrome and repetitive strain injuries provides ample motivation to reduce typing requirements and device manipulation (Klarlund, 2003). Automatic speech recognition (ASR) offers an interaction means capable of replacing conventional typing.

Moreover, the most sensible way of empowering disabled users is to provide them with a means to be able to use the same software applications as any other computer user, which requires that input devices specifically tailored for these users will have to be developed (Istance, Spinner and Howarth, 1996). Eye movement is ideal for such situations as it requires no additional training, is high-speed and the majority of motor impaired individuals still retain ocular motor abilities (Istance, Spinner and Howarth, 1996).

2.2 Eye-tracking and human-computer interaction

Eye-tracking has been used as an alternative input means in a number of applications (for example Gips and Olivieri, 1996; Hornof, Cavender and Hoselton, 2004; Kumar, 2007). The use of eye-tracking can be facilitated in a number of ways, for example dwell time (Isokoski, 2000), look and shoot (Isokoski, 2000) or eye gestures. The use of dwell time requires the user to look at a target for a certain amount of time before the target is activated. Alternatively, look and shoot requires an additional mechanism to be triggered whilst gazing at the desired target. For example, the user may be required to press a key on the keyboard to activate the target under the eye gaze. Gaze gestures require the users to complete a predefined set of eye movements to activate a command (Drewes and Schmidt, 2007). Gaze gestures have been used to successfully map the entire alphabet, thereby allowing users to type text using only their eye gaze (Wobbrock, Rubinstein, Sawyer and Duchowski, 2008). All of these selection methods will be incorporated into the proposed multimodal interface to allow for maximum customisation of the interface to suit the needs of the user at any given time.

The role of feedback is also vital in the development of eye gaze applications (Hyrskykari, Majarants and Räihä, 2003) and serves to increase the user efficiency and enjoyment (for example, Miniotas, Špako and Evreinov, 2003). Therefore, during this study visual feedback will always be given when eye gaze is used as an interaction technique.

Furthermore, even with advances in technology and continued research, most interfaces which are gaze sensitive are designed with oversized interface elements to facilitate easier acquisition and activation of the element (Ashmore, Duchowski and Shoemaker, 2005). The use of oversize targets impacts negatively on screen real estate as a lot of free space is now occupied by icons, buttons etc. To counteract both the impact on available screen real estate and to exploit the properties of Fitts' Law several target expansion mechanisms have been proposed and implemented for both eye pointing and manual input (Ashmore, Duchowski and Shoemaker, 2005). These include expansion of the target in motor space, expanding or zooming into the entire display uniformly or expanding a portion of the display through the use of a fisheye lens (Ashmore, Duchowski and Shoemaker, 2005). Expansion of the targets can be either visible or invisible when it occurs strictly in motor space, implying the user is not aware of the expansion. The idea behind invisible expansion is to create a larger selection area around the target without visual feedback. This allows room for error and slight displacement of the eye during target selection. Buttons used during this study for text input will be larger than the standard icons in Windows. Even so, invisible expansion of buttons will also be used for the onscreen keyboard. This invisible expansion will be referred to as a gravity well as the actual selectable area of the button will be larger than the physical size of the button. Once the eye gaze is detected within the bounds of the enlarged area of expansion, the button will become selectable, thus creating the impression that the eye gaze is drawn onto the button. Additional visible expansion capabilities, in the form of magnification triggered by the position of the eye gaze, will also be provided.

2.3 Eye-tracking and speech recognition in combination

The limitations created by the lack of accuracy of eye-tracking equipment can be overcome by the simultaneous use of speech recognition (Castellina, Corno and Pellegrino, 2008). Insofar as can be ascertained these particular modalities are often used in isolation. When used in such a manner, these are often ambiguous but when appropriately used in

combination they could result in effective interaction methods (Oviatt, 1999). This would create a multimodal interface, which is an interface that uses several input and output modalities in combination in an effort to assist human-computer communication through utilising natural human communication channels (Pireddu, 2007) such as voice and gaze.

The underlying foundation of this research undertaking is the view that while eye gaze and speech recognition are prone to ambiguity when used in isolation, using them in combination may allow much of the problems to be overcome. User intent can be inferred by providing a means for the user to gaze at certain objects and then issue verbal commands which can then be executed to create a hands-free application (Hatfield and Jenkins, 1997). In this way it is envisaged that the strengths of one interaction technique will be able to compensate for the weaknesses of the other and together speech and vision should provide a better interaction experience than each in isolation. Given the inherent problems associated with target selection via eye gaze, such as accuracy, stability and the Midas touch (everything the user gazes at is selected as the user is not accustomed to an interface which reacts to eye gaze) problem, it seems plausible that an additional modality might make selection easier and more feasible even though to date there have been very few empirical studies conducted to explore this phenomenon. One such study did determine that there is high accuracy of target selection using eye gaze and speech to such an extent that user performance approaches that of manual pointing (Miniotas, Špakov, Tugoy and MacKenzie 2006). Furthermore, integration of voice and speech for a multimodal interaction was shown to be a feasible option and an option that works well with robust eye trackers (Pireddu, 2007).

EyeTalk is a voice and vision integrated application which allows a user to gaze at an object and issue a verbal command which is then captured and merged into a single message and passed to the current application as a mouse click or keyboard event (Hatfield and Jenkins, 1997). EyeTalk is application independent and can therefore be used with a multitude of standard applications. Users are able to fixate on an object, which causes the mouse cursor to move to that position, and then issue a command to execute a mouse click (Hatfield and Jenkins, 1997). Initial results with EyeTalk showed positive feedback and indicated that users were able to operate the system with high efficiency after just a few moments of getting accustomed to the system (Hatfield and Jenkins, 1997). A promising consequence of the EyeTalk application is the indication that a stand-alone application can be developed to interact with any Windows application without any need to re-engineer the entire existing application (Hatfield and Jenkins, 1997).

3. Developed application

The premise of the study that is reported in this chapter - to test the feasibility and usability of a multimodal interface for a word processor – necessitated that an application be developed for these purposes. Since Microsoft Word® enjoys the highest market penetration (Bergin, 2006) and also leads the way as the *de facto* interface standard; it was the focus of the study. Consequently, there were two options available, a complete application could be developed that emulated the look, feel and functionality of Word or the Word application itself could be used with data capturing capabilities being provided.

Since Visual Studio for Office (VSTO) allows .NET developers to customise not only the interface of the Office suite but also to add functionality that is required (Anderson, 2009) it was decided to rather use the tried and tested application and add the required components.

Therefore, VSTO was used to manipulate Microsoft Word to make a multimodal interface within a well-known environment. The integrated development environment (IDE) of Visual Studio 2008 was used for development with C# as the programming language.

The Tobii Studio Software Development Kit (www.tobii.com) was used to add eye gaze functionality to the application and the Microsoft Speech Application Programming Interface (www.microsoft.com) was used to add speech capabilities. MagniGlass Pro® (http://magnifying-glass-pro.softutopia.com) was used for magnification purposes as it was fairly inexpensive and was the only tool that was found to allow interaction on the magnification itself. This means that the user could click on the magnified area and did not first have to close the magnification before being able to click, which defeats the purpose of using magnification for selection of small targets.

Figure 1 shows the tab called "Multimodal Add-Ins" that was added to the ribbon in Word 2007. The magnifier button allows the magnifying capabilities to be toggled on and off. Following this are the buttons to show and hide the onscreen keyboards. An alphabetic or standard QWERTY keyboard layout can be chosen. The onscreen keyboards are used for hands-free text entry using eye gaze and speech recognition. The next button group manages the speech engine. The speech engine can be turned on and off, a trained speech profile can be selected and automatic speech recognition (ASR) can be used for either command or dictation purposes. The final group manages the eye gaze interaction technique. The first step when using eye gaze is to calibrate the eye-tracker. The calibration process has a significant effect on the accuracy of the eye gaze interaction technique. The gaze type can then be set. Dwell time (linked to the sensitivity setting), blinking and look

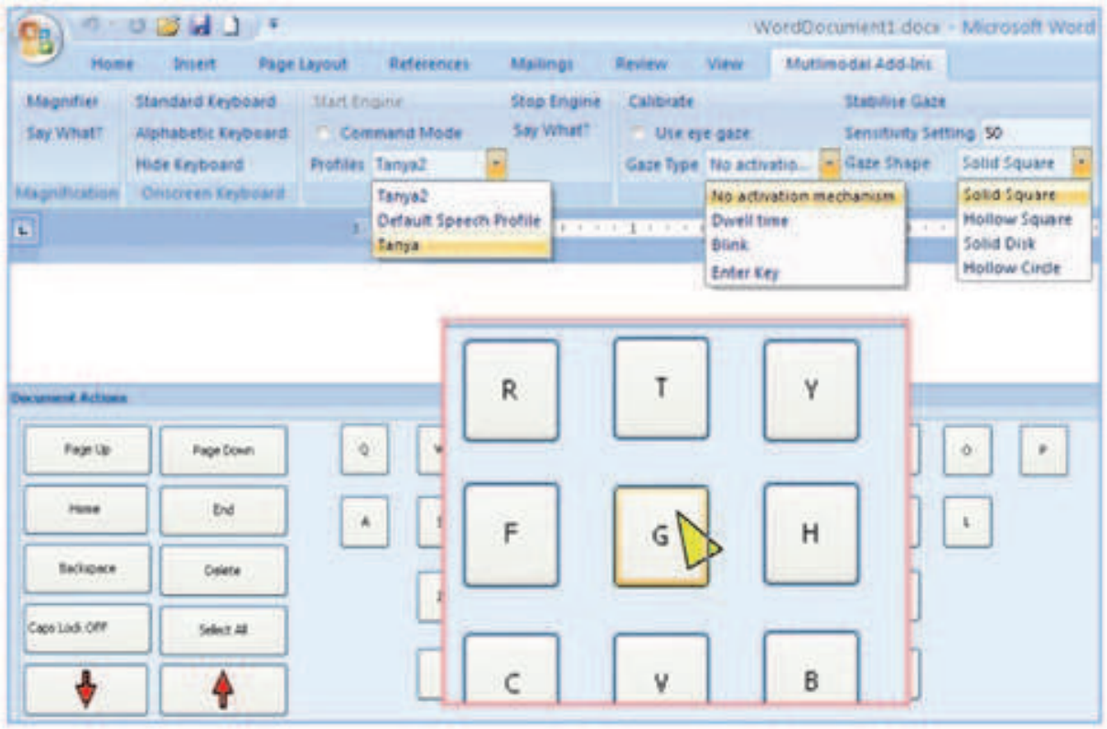

Fig. 1. Multimodal Add-Ins tab in Microsoft Word

and shoot (with the Enter Key) are all available. When the "no activation mechanism" is chosen, then eye gaze can be used in combination with speech recognition. The gaze shape dropdown allows the user to select the shape of the visual feedback cue on the letters of the onscreen keyboard.
The editable region of the document is shown in the figure as a much smaller area than what it was in reality. At the bottom of the screen, the onscreen QWERTY keyboard can be seen with the area directly under the current eye gaze being magnified. The yellow arrow indicates the exact position of the eye gaze.
Speech recognition can be used for both dictation and command purposes. A simple grammar containing common formatting commands (for example bold, italic and underline), cursor movement (for example right, left, up and down) and text selection (for example, select a line, select a word, select whole document) commands was built. In this way it became possible to move around the document or select and manipulate text contained in the document without using either the mouse or the keyboard.
The dwell time can be set by the user to a length of time with which they are comfortable. Blinking requires the user to blink in order to activate the object currently being fixated on. Since blinking is a natural occurrence, the blink required for this activation must be more pronounced. Finally, eye gaze can be used in combination with speech recognition as a text entry method using an onscreen keyboard. When the eye gaze is stable and directed at a certain key, the key is framed with a green square, or the selected shape (see Figure 2). This gives a visual cue/feedback to the user so that they know the key can now be activated. The user can then issue one of several verbal commands in order to type the selected letter to the document at the cursor position. The keys of the onscreen keyboard had a gravity well of 20 pixels on all sides.

Fig. 2. Onscreen keyboard framed in green when selected

By providing all these functions and settings, a highly customisable interface was built within the well-known environment of Word.

4. User testing

The scope of the project did not allow full-scale user testing to be conducted on all the interaction techniques, such as dwell time and blinking. Therefore, the user testing only concentrated on testing the combination of eye gaze and speech when used in a word processor. These interaction techniques could be used for two specific purposes, namely to issue commands in order to perform basic word processing tasks and to enter text within the document. These two types of tasks will be reported on separately within this chapter.
Longitudinal testing was conducted over a ten week period with each participant attending one session per week at the same time and on the same day. During the first session, participants each trained their speech profile using the Microsoft speech training wizard. The participants were then introduced to the multimodal Word that they would be using for the next few weeks and were given a brief tutorial of the speech grammar which was available for use in Word. The participants were then encouraged to interact with the application and to use all the verbal commands as well as attempting to type a full sentence

using the onscreen keyboard and the interaction technique of eye gaze and speech. Every subsequent session followed the same procedure, which was to complete the list of preset task as quickly and correctly as possible.

4.1 User testing of speech commands

The use of speech commands and how their performance compares with that of the mouse and keyboard will be investigated first.

4.1.1 Participants

In total there were 25 participants who participated in the longitudinal study. They were all undergraduate students who were completing their studies at the University of the Free State, South Africa. A pre-requisite for participation in the study was sufficient computer literacy as well as word processor expertise.

There were 17 male participants and 8 female participants with an average age of 21.1 (standard deviation = 1.9). Six participants indicated that English was their first language, 7 Afrikaans and the remainder (12) were African language speakers. Since the University employs a parallel medium tuition policy where classes are offered in either English or Afrikaans, all students are comfortable in either English or Afrikaans. Therefore, each session was conducted in the tuition language of the participant.

4.1.2 Tasks

Participants had to complete 20 tasks, five of which were typing tasks. The majority of the other tasks, for example selection and formatting, had to be completed using the traditional means of a mouse or keyboard. A similar task then had to be repeated using speech recognition. The tasks were set up in such a way that the same types approximately required an equal number of minimum actions to complete it successfully. A summary of the tasks is tabulated below (with typing tasks omitted):

Task Description	Shortened task description	Keyboard	Speech
Select three lines and apply formatting such as bold or italics	Line selection and formatting	1	1
Select all text in the document and remove it by deleting or cutting	Select all text and remove	1	1
Select two words and make them bold	Select words and format	1	1
Paste previously copied text at the current cursor position	Paste	1	1
Undo the previous action	Undo	1	1
Select a single word and copy it	Select word and copy	1	1
Position the cursor at a certain position in the document and paste the previously copied text	Position and paste	1	1

Table 1. Grouped tasks as divided between interaction techniques

4.1.3 Measurements

The measurements that will be analysed are the time taken to complete the task as well as the number of actions that were required to complete the task. The number of errors was also considered as a means to determine how effective the interaction technique is. However, since there are multiple ways to complete a task, it became very difficult to pinpoint exactly what was an erroneous action, particularly where the mouse or keyboard was used. For the speech, the commands that could complete the task could be isolated as an acceptable set of commands for that task and then any command issued that is not a member of that set can be flagged as an error command. However, since there is considerable risk for potentially flagging an action as an error when it might not be, it was decided that the percentage of the task completed correctly were better indicators of the effectiveness of the interaction techniques.

4.1.4 Time to complete a task

The time to complete the task was measured from when the task was started to when the task was considered by the participant to be completed. This time included the time it took the participant to read the description of the task. Since similar tasks had virtually identical wording it was assumed that they would require the same amount of time to read and that, therefore, the time to read would not have an effect on the time required to complete the task.

The charts below (Figures 3-6) plot the least square means for both interaction techniques over all sessions. The least squares means are the means of interest when interpreting significant results of a factorial design (StatSoft, 2010) and will therefore be provided as a visual representation of the descriptive statistics. The vertical bars denote a 95% confidence interval. The blue line plots the completion time for the speech and the red line that of the keyboard.

As can clearly be seen from the graphs above, in some instances the keyboard maintained a faster average completion time and in others the speech interaction technique could surpass the performance of the keyboard.

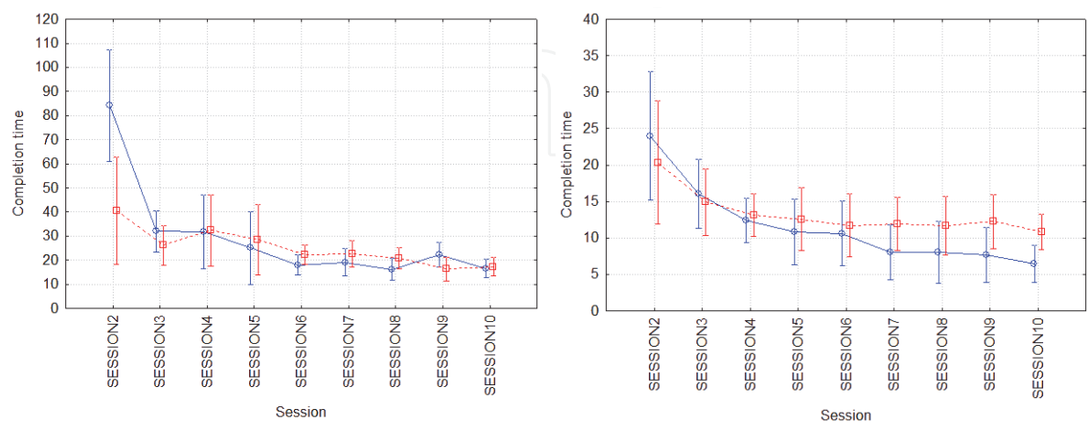

Fig. 3. Average completion times for (a) line selection and formatting and (b) select all and remove

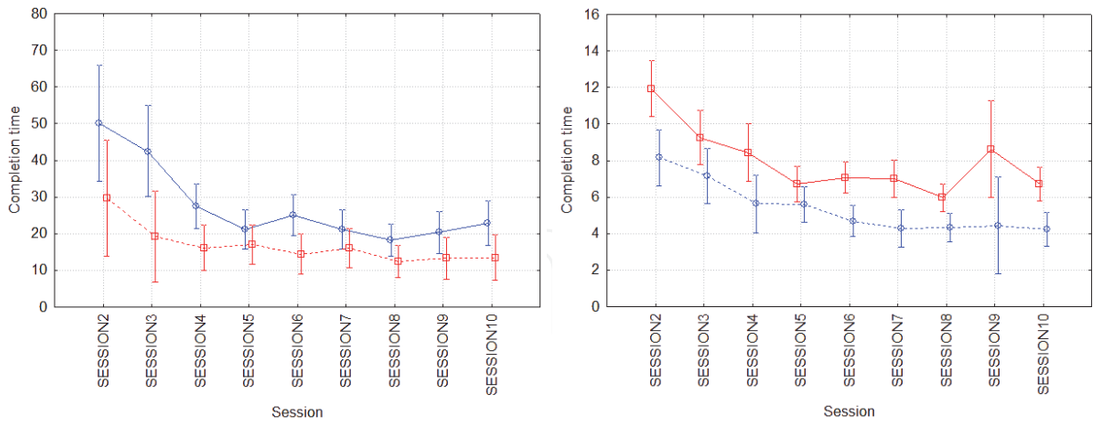

Fig. 4. Average completion times for (a) select words and format and (b) paste

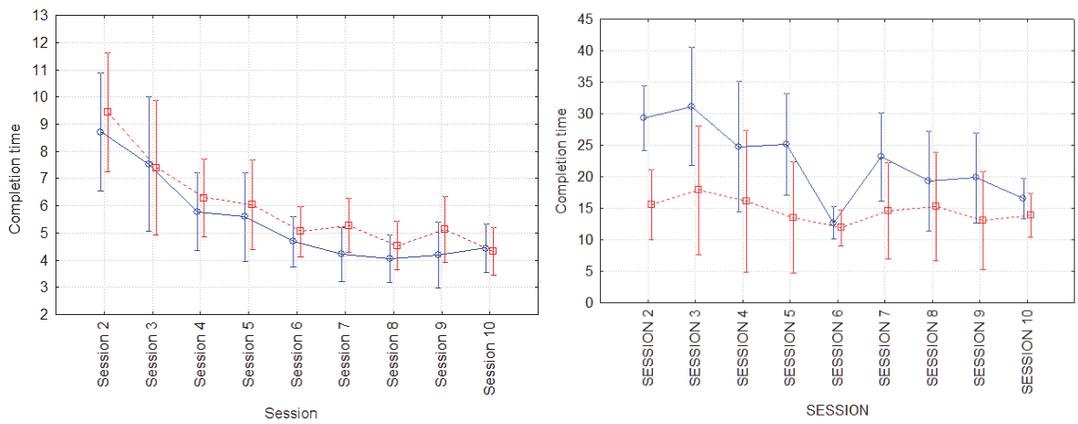

Fig. 5. Average completion times for (a) undo and (b) select word and copy

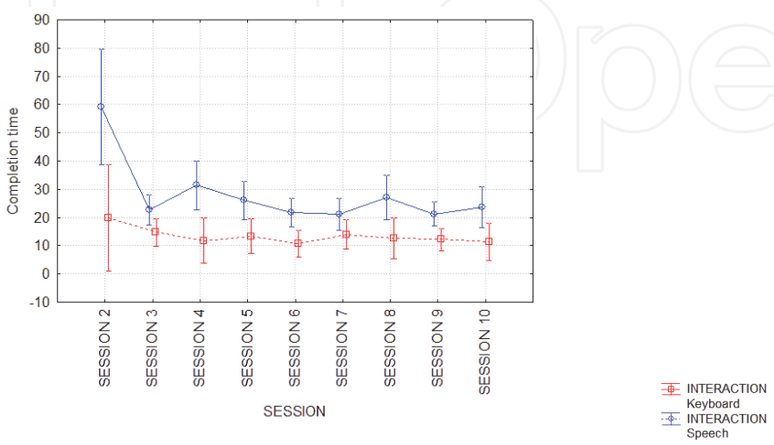

Fig. 6. Average completion times for position and paste

The time measurements were in seconds and there were a vast number of instances in which the normality tests fail for the data. In order to combat this, the time measurement was converted to 1/time.

For each of the tasks, the following hypotheses were formulated:

1. $H_{0,1}$: There is no difference between the time required to complete the tasks when using the mouse and keyboard or speech commands.
2. $H_{0,2}$: Participants did not improve over time with regard to the time taken to complete the tasks.

A repeated-measures within-subjects ANOVA was performed to analyse the aforementioned hypotheses. Where necessary, the adjusted corrections of Geisser-Greenhouse and Huyn-Feldt were applied to the degrees of freedom in the cases where the assumption of sphericity was not met. The table below shows only the results of the original ANOVAs and not, for the sake of brevity, the results of the adjusted corrections. For the Paste task, there was significant interaction between the factors of interaction technique (keyboard and speech) and improvement over time (session) the two hypotheses had to be examined in isolation.

	$H_{0,1}$	$H_{0,2}$
Line selection and formatting	$F(1, 23) = 0.286$, $p > 0.05$	$F(8, 184) = 14.040$, $p < 0.05$
Select all and remove	$F(1, 23) = 4.328$, $p < 0.05$	$F(8, 184) = 15.197$, $p < 0.05$*
Select words and format	$F(1, 26) = 10.447$, $p < 0.05$	$F(8, 208) = 9.487$, $p < 0.05$
Paste		
Undo	$F(1, 24) = 0.001$, $p > 0.05$	$F(8, 192) = 22.148$, $p < 0.05$
Select word and copy	$F(1, 22) = 3.655$, $p > 0.05$	$F(8, 176) = 3.470$, $p < 0.05$
Position and paste	$F(1, 22) = 15.448$, $p < 0.05$	$F(8, 176) = 5.123$, $p < 0.05$

Table 2. Results of ANOVA for time of speech commands

The first null hypothesis could be rejected for the task which required all text to be selected and removed. In this instance, it was the speech commands which averaged a faster completion time. Conversely, the keyboard was significantly faster for the task where words had to be selected and formatted as well as for the position and paste task. This finding could imply that the speech command to select all text was fairly intuitive and easy to learn, which facilitated a faster completion time than using the mouse or keyboard. However, selection of individual words was less intuitive and took longer than when using the keyboard or mouse. It could also mean that participants did not use the keyboard shortcut to select all text as this is the fastest way of selecting all text in a document. Analysis of the number of actions should provide more clarity in this regard.

For those tasks where the second null hypothesis could be rejected, it was under the majority of cases the first few sessions which differed significantly from the last sessions. This provides a very encouraging finding that there is a significant effect of learning which occurs as the amount of exposure to the application is increased.

When a repeated-measures within-subjects ANOVA was performed for the paste task, it was found that there was significant interaction between the two factors of session and interaction technique ($F(8, 192) = 2.356$, $p < 0.05$). Therefore, it was imperative that each factor was isolated and analysed separately to preclude the interaction with the other factor having an effect on the analysis. Firstly, $H_{0,1}$ was evaluated by isolating each session individually and testing for a difference between interaction techniques. For brevity's sake, the actual results of the ANOVA will not be reported here. Suffice it to say that, at an α-level of 0.05, there was a significant difference between the interaction techniques in every session. Therefore, the completion time is significantly better for speech than for the keyboard and mouse throughout all the sessions. Secondly, $H_{0,2}$ was evaluated using a repeated-measures within-subject ANOVA but testing each interaction technique separately. Consequently, it was found that $H_{0,2}$ could be rejected for both the speech interaction technique ($F(8, 96) = 17.727$, $p < 0.05$) and the keyboard and mouse ($F(8, 96) = 6.883$, $p < 0.05$).

4.1.5 Number of actions

The next measurement to be analysed was the number of actions that were performed during task completion. Actions were defined as any mouse click, button press or speech command that was issued during completion of the task. The number of actions were measured per interaction technique and per session for each participant and then, as always, outliers were removed from the data set prior to analysis.

The underlying hypotheses were formulated to analyse the actions for this task:

$H_{0,1}$: The interaction technique does not significantly affect the number of actions required to complete the task.

$H_{0,2}$: Participants did not improve over time with regard to the number of actions required to complete the task.

The charts below (Figures 7-10) plot the number of actions for each interaction technique over all sessions. The red line plots the keyboard and mouse actions, while the blue plots the speech commands.

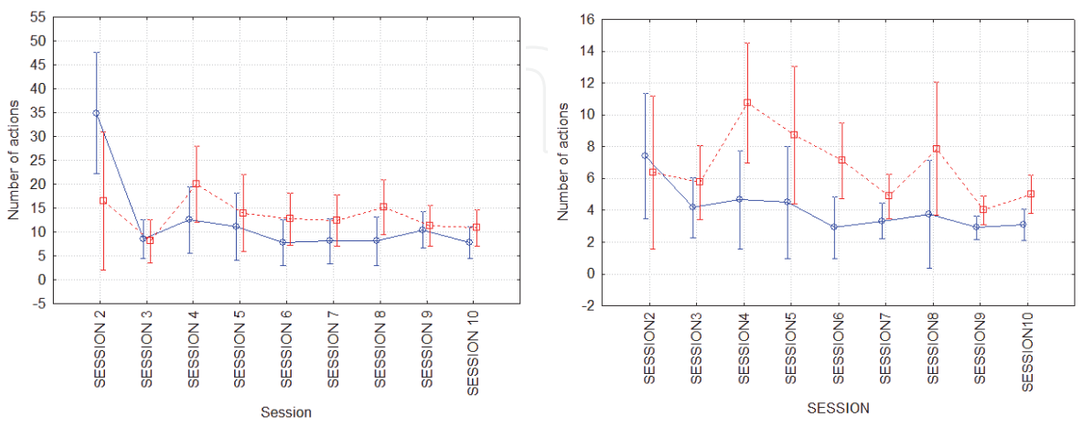

Fig. 7. Average number of actions for (a) line selection and formatting and (b) select all and remove

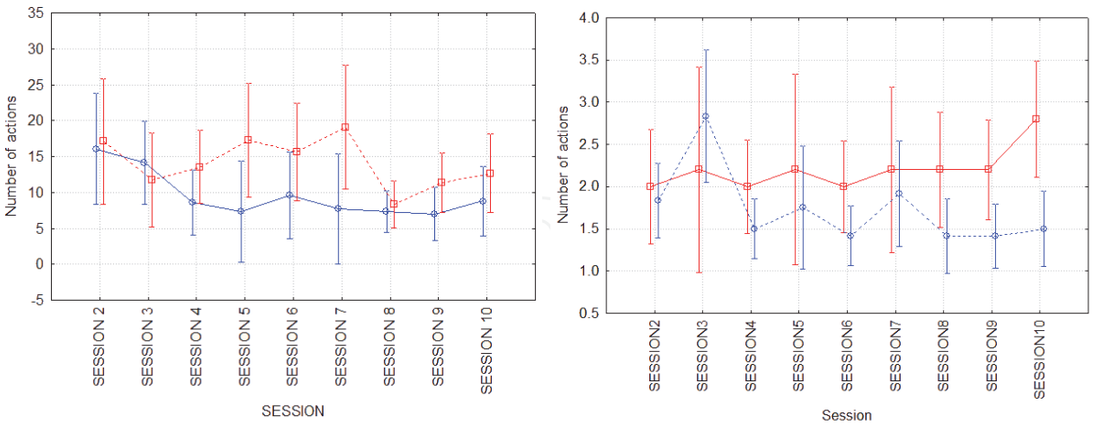

Fig. 8. Average completion times for (a) select words and format and (b) paste

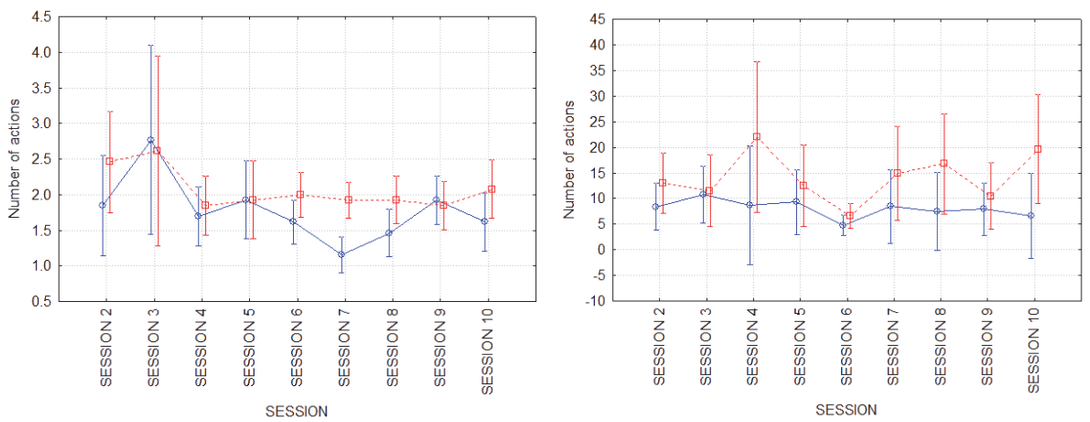

Fig. 9. Average completion times for (a) undo and (b) select word and copy

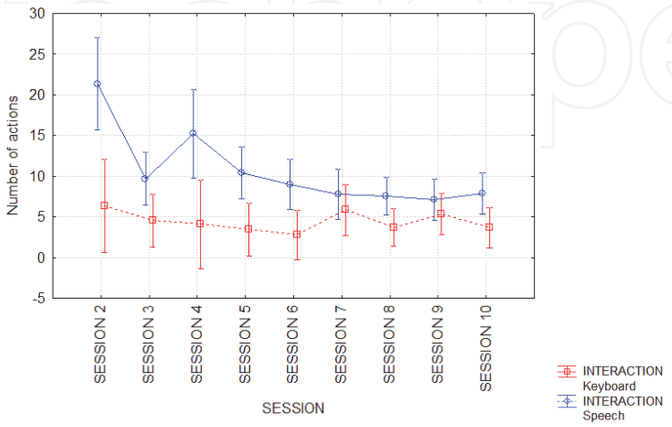

Fig. 10. Average completion times for position and paste

The graphs clearly show that in most instances the use of the keyboard and mouse resulted in more actions being performed. It was only when participants were required to position the cursor and paste previously copied text that the speech commands required more actions. The table below summarises the results of the repeated-measures within-subjects ANOVA for each task.

	$H_{0,1}$	$H_{0,2}$
Line selection and formatting		
Select all and remove	$F(1, 18) = 8.574$, $p < 0.05$	$F(8, 144) = 2.562$, $p < 0.05$
Select words and format	$F(1, 23) = 2.598$, $p > 0.05$	$F(8, 184) = 2.234$, $p < 0.05$
Paste	$F(1, 15) = 6.287$, $p < 0.05$	$F(8, 120) = 1.297$, $p > 0.05$
Undo	$F(1, 24) = 2.294$, $p > 0.05$	$F(8, 192) = 2.934$, $p < 0.05$
Select word and copy	$F(1, 19) = 3.498$, $p > 0.05$	$F(8, 152) = 1.378$, $p > 0.05$
Position and paste		

Table 3. Results of ANOVA for actions of speech commands

In the two instances where there was a significant difference between the interaction techniques, it was the speech commands which required significantly less actions than the keyboard. This result for the selection and removal of all text and the paste task corresponds with the findings that the speech commands were also more efficient, in terms of the time required to complete a task, for these tasks.

For the task which requires that words be selected and formatted, session 2 had a significantly higher number of actions than any other session. During the undo task, session 3 resulted in a significantly larger number of actions than the other sessions.

The two tasks for which there are no results in the above table had significant interaction between the two factors. This meant that individual analyses had to be performed in order to counteract the effect of one factor on another. For the line selection and formatting task, the two interaction techniques differed significantly from one another during the second and eighth session. During the other sessions the number of actions for the two interaction techniques was comparable to one another. The second null hypothesis could be rejected for the keyboard, where a significantly higher number of actions were performed during session 2 than all the other sessions, but not for the speech commands. Closer inspection of the analysis revealed that some participants resorted to using longer methods of text selection when using the keyboard. For example, they would select the text one character at a time instead of using the efficient means which were available. Since it appears that the majority of the participants used the mouse for selection purposes, the fact that there was a minority who employed this very inefficient means was not cause for great concern but cognisance was taken thereof.

For the task where the cursor had to be positioned and text pasted at that specific location, speech required significantly more actions than the keyboard during all the sessions. Even

though the number of actions decreased over the sessions, which indicates learning, the learning did not allow the speech to perform on a comparable level to the keyboard. The higher number of actions for the speech interaction technique could be explained by the types of commands that were issued. Therefore, an analysis was conducted to determine which commands were issued during the completion of this task. This showed a high incidence of the command 'Right' which could be used to move the cursor to the right. This indicated that the participants resorted to moving the cursor to the correct position one character at a time. Obviously very few participants realised that they could use the command 'Select word' and then 'Right' to move the cursor to the right a word at time. Since the keyboard and mouse offers the alternative of simply clicking the mouse pointer at the correct position this could account for the significant difference between the two interaction techniques. This finding could mean that the participants do not seek to find the most efficient method of task completion.

The ANOVA performed to evaluate $H_{0,2}$ for the speech commands showed that there was a significant difference between the sessions ($F(8, 64) = 5.820$, $p < 0.05$*). Post-hoc tests indicated that there was significant improvement between session 2 and the remainder of the sessions.

4.1.5 Discussion

The speech interaction technique performed relatively well when compared with the keyboard and mouse, in some instances even surpassing the performance of the traditional input methods. Clearing of all text in the document and pasting were even faster and completed with less actions than when using the keyboard and mouse. It is only when positioning within the document must occur that the keyboard outperforms the speech interaction technique in terms of both the time that it takes and the number of commands that are issued.

While this finding was very encouraging, the most promising finding was that there was continued improvement in the efficiency with which the task was completed. Even though the improvement between subsequent sessions was not always significant the fact there is continual improvement hints at the possibility that the two interaction techniques could eventually compete on a comparable level for all tasks or that the speech interaction technique could eventually perform better.

Since there are often multiple options available to the user to complete the task when using the traditional means, the most effective method was not always chosen. This was also noticed when using speech to move the cursor. Rather the user chooses the method which results in an intermediate action which is closer to the final result when in reality there is a shorter method that can be used.

The fact that the speech commands resulted in less actions for most of the tasks, may be attributed to the fact that the grammar was fairly simple and provided commands to complete basic operations only. The complexity of the options provided by Word is much higher than accommodated in the grammar. When using Word in the normal capacity there is, more often than not, at least 3 different ways to complete a task which may place an added burden on the user of the application. However, the goal of the study was not to provide a complete alternative to the keyboard and mouse but rather to determine whether common word processing tasks could be achieved using an alternative interaction technique. Therefore, by the very nature of the study, the grammar was required to be simple in composition.

4.1.6 Further research

The tasks that were chosen for this part of the study were chosen as some of the more common tasks that may occur in the word processing application. Therefore, they may be viewed as some of the less complex tasks and other tasks may require less intuitive commands and more complex commands. However, this will parody the nature of any other system which provides access to common tasks "at your fingertips", for example the Home tab in Office while lesser used tasks or more complex tasks require further navigation and perhaps a heavier burden on one's memory. It may be possible to extend the grammar to encompass many more tasks within the word processor application. Another consideration would be to use a default smaller grammar and an optional extended grammar that can be activated on request.

The results of the study indicate that interaction through speech could dramatically increase the efficiency of end-users. However, it remains to be seen if this result holds when the user is free to use the grammar in a normal setting. This would require that the participants would not be given small separate tasks but rather that they would have to compile a document from scratch with pre-defined formatting.

Whether or not an extended grammar is considered, further research will have to be done where the exposure to the application is lengthened in order to determine whether the learning effect can continue to an even greater degree. This study could use a smaller sample as it has already been established that it is possible to use this interaction technique effectively.

4.2 User testing of text input

As previously mentioned, the longitudinal testing also included tasks which required that the participants input text using either the keyboard or eye gaze and speech recognition. This section is a discussion of the comparative study between these two text input methods.

4.2.1 Participants

The participants for this analysis were the same as in the previous section. There were, however, three of the 25 participants who were unable to type using eye gaze and speech for various reasons and they were excluded from the analysis. Fourteen of the remaining participants were male and 8 were female, 6 were English-speaking, 6 Afrikaans-speaking and the remainder (10) had an African language as their first language. The average age of participants was 21.1 (standard deviation = 2.0).

4.2.2 Tasks

In total there were two typing tasks using the keyboard and three using the eye gaze and speech. The tasks required participants to type phrases that were randomly selected from a set of 35 preselected tasks, which were in turn selected from the 500 everyday commonly used phrases as determined by MacKenzie and Soukoreff (2003).

When using eye gaze and speech the size of the buttons was set to 60×60 (≈1.55° visual angle) pixels. Buttons were spaced 60 pixels apart with a gravity well of 20 pixels on all sides of each button. Although there were three typing tasks using these settings, only the last two of each session were included in the analysis. This was due to the fact that the first one was viewed more as a practice typing task to reacclimatise the participants to typing using eye gaze and speech. The participants were not told that the first task would not count towards the analysis and were instructed to complete all tasks to the best of their ability.

In order to investigate the effect of size and spacing between targets, additional typing tasks were added from the fifth session onwards. Within these additional typing tasks, the first one had to be completed using the originally sized and spaced buttons. The next two had to be completed with buttons that were 50×50 (≈1.29° visual angle visual angle at a viewing distance of 600 mm) pixels in size and spaced 70 pixels apart. Following this there were another two tasks which had to be completed using buttons that were also 50×50 pixels in size but were spaced 60 pixels apart. For all typing tasks a gravity well of 20 pixels on all sides of the buttons were employed.

4.2.3 Measurements

Since both input methods (the keyboard and eye gaze and speech recognition) were character based, the measurements that were selected for analysis were the character error rate and the characters typed per second. The character error rate (CER) measures how many insertions, deletions and substitutions have to be done to convert the presented text to the text as entered by the participant (Read, 2005). This measurement is synonymous with the Levenshtein distance between two strings (Levenshtein, 1966) divided by the number of characters that were typed (Read, 2005; MacKenzie and Soukoreff, 2002). This error rate measurement will be used in this section to analyse the effectiveness of the interaction techniques.

For the efficiency of the interaction techniques, the measurement of characters per second (CPS) will be used. This measurement divides the number of characters that were typed by the time taken in seconds. Similar to previous studies (MacKenzie, 2002), the time taken was measured from the time when the first character was typed to the time the last character was typed. This excludes the time required to read the question, including the sentence that must be typed, and the time taken to locate the first character that must be typed. As a consequence, the number of characters becomes n-1.

4.2.4 Results

The initial analysis will only include the data from the original typing tasks using the originally sized buttons.

The leftmost chart below shows the average error rate for input through eye gaze and speech (blue line) and the keyboard (red line). The chart on the right shows the characters per second that were achieved with both interaction techniques and for all sessions. Clearly, the technique of eye gaze and speech results in far more errors than the keyboard when used for text entry while the keyboard facilitates a faster typing speed. Although the error rate of eye gaze and speech declines as exposure increases, the typing speed does not increase significantly. This could indicate that either more practice is required to increase typing speeds or that the typing speed quickly reaches a plateau which cannot be breached. Observation of the participants during their interaction with the system would suggest that more practice is required to increase the efficiency of the text entry.

Using a confidence interval of 95%, it was found that the interaction technique had a significant effect on the number of errors made ($F(1, 21) = 6.516$, $p < 0.05$) but that there was also a significant difference between the sessions ($F(8, 168) = 2.278$, $p < 0.05$). In particular, sessions 9 and 10 differed significantly from sessions 2 and 3. This shows a measure of improvement in the error rate as time went by and would suggest that participants were becoming more accustomed to using eye gaze and speech for text input purposes.

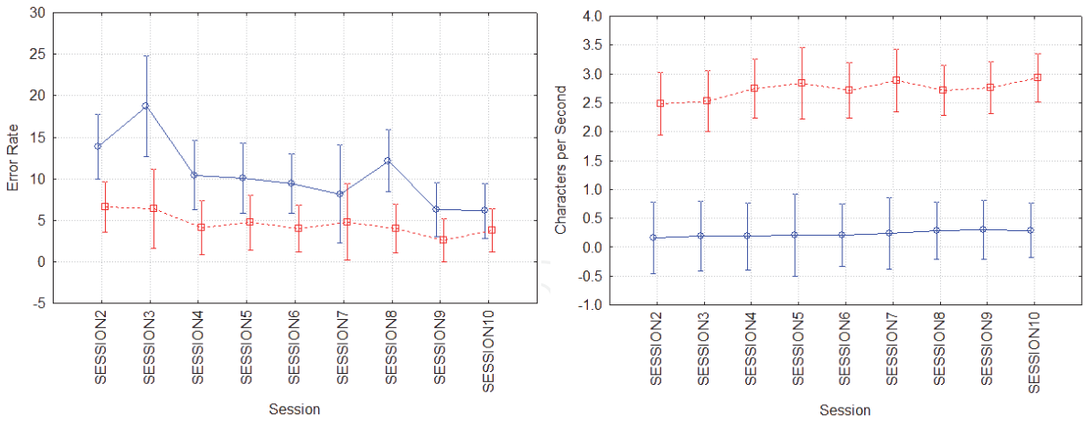

Fig. 11. Least squares mean plot of character error rate and characters per second

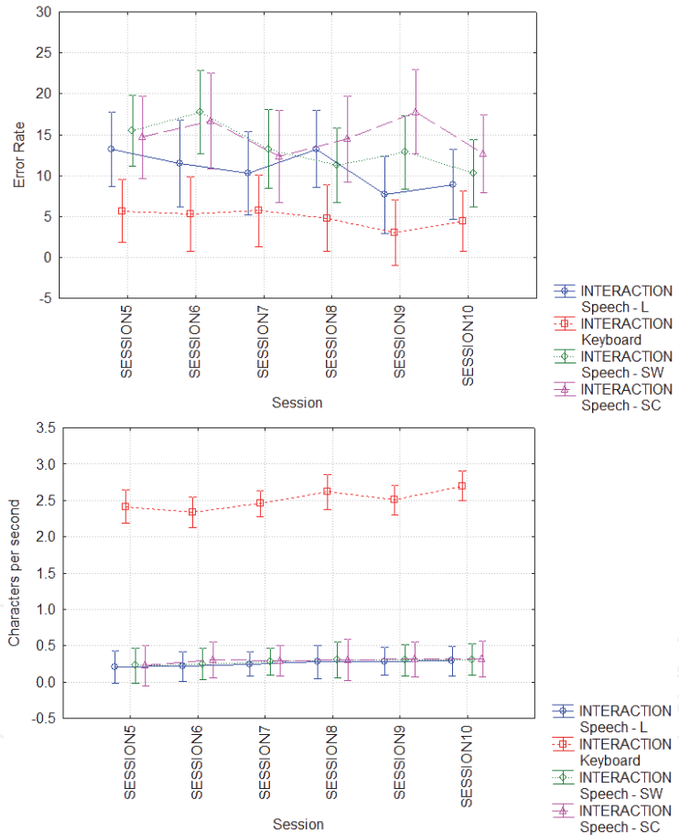

Fig. 12. Least squares mean plot of character error rate and characters per second for all typing tasks

Similarly, the interaction technique ($F(1, 21) = 54.704$, $p < 0.05$) had a significant effect on the characters typed per second but there was no significant difference between the sessions ($F(8, 168) = 1.385$, $p > 0.05$). Therefore, using eye gaze and speech for typing is significantly slower than when typing with the keyboard but there is no significant improvement in typing speed as exposure to the system increases.

The next step was to analyse text input that includes the additional tasks and differently sized and spaced buttons. Since the additional tasks were only completed from session 5 onwards. The analysis was done for these sessions only. In order to distinguish between the different sized buttons, results for the originally sized and spaced buttons will be referred to as speech-L, the smaller widely spaced buttons as speech-SW and the smaller closely spaced buttons as speech-SC.
Figure 12 plot the error rate and characters per second for each of the text entry methods for the sessions during which they were tested.
The keyboard has the lowest error rate of all the interaction techniques and it also has the highest typing speed. Regarding the error rate and typing speed of the eye gaze and speech, the three different methods are virtually indistinguishable from one another.
The interaction technique ($F(3, 44) = 4.100$, $p < 0.05$) causes a significant difference in the error rate but there is no significant difference between the error rates of the various sessions ($F(5, 220) = 1.056$, $p > 0.05$). Post-hoc tests indicate that there is a significant difference between the error rates of the keyboard and those of the speech-SW interaction technique. In terms of typing speed, the interaction technique ($F(3, 44) = 148.369$, $p < 0.05$*) significantly affects this measurement as does the session ($F(5, 15) = 3.002$, $p < 0.05$*). As could be expected the keyboard results in a significantly faster typing speed than all other interaction techniques. The typing speeds in the last session were also significantly faster than the speeds of the first two sessions which indicates some measure of learning.

4.2.5 Discussion

It was found that the eye gaze and speech interaction technique causes a significantly higher error rate than the keyboard. There was no difference between the error rates of speech-L, speech-SW and speech-SC and they all differed from the keyboard at some stage. However, the interaction technique of speech-L did seem to offer the most improved error rate as it did not differ from the keyboard when analysed for the later sessions only. In some instances there was improvement over the sessions, which indicates some measure of learning when using eye gaze and speech. If the learning effect can be maintained, more practice could possibly lead to an effectiveness measurement which is comparable to that of the keyboard.
In terms of efficiency (characters per second), the keyboard outperformed the eye gaze and speech interaction technique. The efficiency of eye gaze and speech also did not improve as exposure increased. This could either indicate that more practice is needed to achieve increased speed or that the typing speed quickly reaches the fastest achievable rate. Neither the size of the buttons nor the spacing between buttons affected the efficiency of the eye gaze and speech.

4.2.6 Further research

Further research can be conducted whereby the participants receive more practice with using eye gaze and speech as a text input mechanism. This will allow more detailed analysis to be performed in order to determine whether a much longer period of exposure would serve to increase the effectiveness and efficiency of the interaction technique. Furthermore, future studies could incorporate the correction of errors so that the character error rate could determine the eventual correctness of the transcribed text in conjunction with the transcribed text before corrections were applied.

Since it was found that neither the size of the buttons nor the spacing between the buttons influenced the usability of the interaction technique, further tests can be conducted to determine whether an increase in the gravity well will impact performance. Although the decrease of physical size and increase of gravity well result in a selectable area with the same size as a large button, the *perceived* accuracy with smaller buttons could serve to boost the confidence, and therefore satisfaction, of end-users.

5. Conclusion

This chapter reported on the results of similar word processing tasks which were compared when they were completed using the mouse and keyboard or when using speech commands. The measurements which were analysed were time to complete the task and the number of actions that were performed during completion of the task. For the majority of the tasks it was found that the interaction techniques could compete on a comparable level, particularly as the participant gained experience. This indicates that the application was indeed learnable. These results indicate that the proposed use of speech commands within a word processor application is viable.

This chapter also reported on the results of the use of eye gaze and speech for text input when compared to a traditional keyboard. Measurements of effectiveness, namely the error rate, and efficiency, namely characters typed per second were analysed. It was found that when using eye gaze and speech for text input, neither the size of the buttons nor the spacing between the buttons affected the performance of the interaction technique. The performance of the keyboard for both these usability measures far outstrips that of the eye gaze and speech. Even with extended exposure to the eye gaze and speech interaction techniques, the effectiveness and efficiency could not reach levels which were equivalent to those achieved by the keyboard.

6. References

Ashmore, M., Duchowski, A.T. & Showmaker, G. (2005). Efficient Eye Pointing with a Fisheye Lens. In *Proceedings of Graphics Interface 2005*

Beelders, T.R. (2006). A comparative study on users' responses to graphics, text and language in a word processor interface. M.Sc dissertation, University of the Free State, Bloemfontein, South Africa

Bergin, T.J. (2006). The Origins of Word Processing Software for Personal Computers: 1976 – 1985. *IEEE Annals of the History of Computing*. 28(4), pp. 32-47

Blignaut, P.J., Dednam, E.H. & Beelders, T.R. (2007). Die opleiding van persone uit benadeelde groepe in rekenaargebruik: Is die agterstand nie té groot om te oorbrug nie? *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie,* 26(3)

Castellina, E., Corno, F., & Pellegrino, P. (2008). Integrated Speech and Gaze Control for Realistic Desktop Environments. In *Proceedings of ETRA 2008*

Drewes, H. & Schmidt, A. (2007). Interacting with the Computer using Gaze Gestures. In *Proceedings of the 11th IFIP TC13 International Conference on Human-Computer Interaction, INTERACT 2007, Rio de Janeiro, Brazil, September 2007*

Gips, J. & Olivieri, P. (1996). EagleEyes: An Eye Control System for Persons with Disabilities. In *Proceedings of The Eleventh International Conference on Technology and Persons with Disabilities*, Los Angeles, March 1996

Hatfield, F. & Jenkins, E.A. (1997). An interface integrating eye gaze and voice recognition for hands-free computer access. In *Proceedings of the CSUN 1997 Conference*

Hornof, A., Cavender, A & Hoselton, R. (2004). EyeDraw: A system for drawing pictures with eye movements. *ASSETS 2004*

Hyrskykari, A., Majaranta, P. & Räihä, K-J. (2003). Proactive response to eye movements. In M. Rauterberg et al. (Eds.), *Human-Computer Interaction -- INTERACT'03*, IOS Press, pp. 129-136

Isokoski, P. (2000).Text input methods for eye trackers using off-screen targets. In *Proceedings of ETRA 2000*

Istance, H.O., Spinner, C. & Howarth, P.A. (1996). Providing motor impaired users with access to standard Graphical User Interface (GUI) software via eye-based interaction. In Proceedings of *1st European Conference on Disability, Virtual Reality and Associated Technology*, Maidenhead, UK

Jacobs, R. J. (1993). Advances in Human-Computer Interaction, Vol. 4. In H.R. Hartson and D. Hix (eds.), *Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces*, pages 151–190. Ablex Publishing Co

Jacob, R.J.K. & Karn, K.S. (2003). "Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary)," in J. Hyona, R. Radach, and H. Deubel (eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pp. 573-605, Amsterdam, Elsevier Science

Klarlund, N. (2003). Editing by Voice and the Role of Sequential Symbol Systems for Improved Human-to-Computer Information Rates. In *Proceedings of ICASSP*

Kumar, M. (2007). Gaze-enhanced user interface design. PhD Thesis, Stanford University.

Miniotas, D., Špakov, O. & Evreinov, G. (2003). *Symbol Creator: An alternative eye-based text entry technique with low demand for screen space*. In M. Rauterberg et al. (Eds.) Human Computer Interaction – INTERACT '03, pp. 137-143

Miniotas, D., Špakov, O., Tugoy, I. & MacKenzie, I.S. (2006). Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 symposium on Eye tracking research and applications (ETRA)*, San Diego, California, pp. 67-72

Oviatt, S. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the ACM SIGCHI 99*, Pittsburgh, Pennsylvania, United States, pp. 576 – 583. New York: ACM Press

Pireddu, A. (2007). Multimodal Interaction: An integrated speech and gaze approach. Thesis submitted at Politecnico di Torino

Van Dam, A. (2001). *Post-Wimp user interfaces: The human connection*. In R. Earnshaw, R. Guedj, A. van Dam and J. Vince (Eds), Frontiers of human-centred computing, online communities and virtual environments (pp. 163-178). London, Great Britain:Springer-Verlag

Wobbrock, J.O., Rubinstein, J., Sawyer, M.W. & Duchowski, A.T. (2008). Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, pp. 11-18

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen