Enhancing the user experience for a word processor application through vision and voice

By

Tanya René Beelders

Submitted in fulfilment of the requirements for the degree

PHILOSOPHIAE DOCTOR

In the Faculty of Natural and Agricultural Sciences
Department of Computer Science and Informatics
University of the Free State
Bloemfontein
South Africa

2011

Promotor:

Prof. P.J. Blignaut
Department of Computer Science and Informatics

Two roads diverged in a wood and I - I took the one less travelled by,
and that has made all the difference.

~ Robert Frost ~

Acknowledgements

I would like to express my utmost thanks and gratitude to the following:

- Professor Pieter Blignaut, my promoter for his guidance, assistance and patience throughout this undertaking.
- The staff of the Computer Science and Informatics Department at the University of the Free State for their moral support and friendship.
- My friends and family for their support, understanding and patience.

PREFACE

The study contained within this thesis has, to date, yielded a number of publications. Most recently, a submitted manuscript has been accepted for publication as a chapter in an upcoming book on speech technologies. The book is currently in press. The following is a list of articles which have been published from this work (the publications are reproduced in Appendix I).

1. Beelders, T.R. and Blignaut, P.J. (2009). A multi-modal interface for a popular word processor. *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns Studentesimposium 2009*, Bloemfontein, South Africa.
2. Beelders, T.R. and Blignaut, P.J. (2010). Using vision and voice to create a multimodal interface for Microsoft Word 2007. *Proceedings of the Symposium on Eye-Tracking Research and Applications (ETRA)*, Austin, Texas, United States of America, 173-176.
3. Beelders, T.R., Blignaut, P.J. and Greeff, F. (2010). Eye-tracking and speech recognition instead of a computer mouse. *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns Studentesimposium 2010*, Pretoria, South Africa.
4. Beelders, T.R. and Blignaut, P.J. (2011). The Usability of Speech and Eye Gaze as a Multimodal Interface for a Word Processor. In I. Ipšić (Ed), *Speech Technologies* (pp. 385-404). ISBN: 978-953-307-996-7.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF CHARTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

A word processor is a software application which allows for composition, editing and formatting of a printable document (wordiQ, 2010). The word processor has become a very popular tool in the everyday use of a computer and has displayed a remarkable ability to evolve and incorporate emerging technologies. The original word processor was developed by IBM in 1969 (Eisenberg, 1992) and since then it has evolved constantly, exploiting the advances in technology.

As an integral part of everyday life for many people a word processor should cater for a very diverse group of users and it offers a unique environment which is rich in potential for improvement of the user experience. However, it may be highly unlikely that only one such complex application would be able to offer the best possible experience to all users. The word processor and the improvement of the usability thereof are the main focus areas of this research study.

1.2 Aim

The aim of the study is to investigate various means to increase the usability of a word processor for use by a diverse group of users, including users of different expertise levels, ages and abilities. Specifically, it will be to determine (i) whether it is feasible[1] to incorporate a truly multimodal interface into a popular existing word processor application through the use of non-traditional input methods and (ii) how usable such an interface will be[2].

1.3 Motivation

Communication between humans and computers is considered to be two-way communication between two powerful processors over a narrow bandwidth (Jacob & 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 (Jacob & Karn, 2003). An additional communication mode will invariably provide for an improved interface (Jacob, 1993a) and new input devices which capture data from the user both conveniently and at a high speed are well suited to provide more balance in the bandwidth disparity (Jacob & Karn, 2003). In order to better utilise the bandwidth between human and computer, more natural communication which concentrates on parallel rather than sequential communication, is required (Jacob, 1993a). The eye-tracker is one possibility which meets the criteria for such an input device. Eye-trackers have steadily become more robust, reliable and cheaper and therefore, present themselves as a suitable tool for this use (Jacob & 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 (Jacob & Karn, 2003).

[1] A feasibility test is aimed at determining whether the proposed interface is viable and whether it could offer a potentially usable interface to any users. Therefore, contrary to a more formal usability study, it does not require that objective measurements be captured and analysed statistically.

[2] This aim will require more formal usability measurements to be captured and analysed.

Furthermore, the user interface is the conduit 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 & Jenkins, 1997) and the current trend of Windows, Icons, Menu and Pointer (WIMP) interfaces have been around since the 1970s (Van Dam, 2001). These graphical user interfaces may pose difficulties to users with disabilities and it has become essential that viable alternatives to mouse and keyboard input should be found (Hatfield & 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. Disabled users should be accommodated in the same software applications as any other computer user, which will naturally necessitate new input devices (Istance, Spinner & Howarth, 1996) or the redevelopment of the user interface. Eye movement is well-suited to these needs as the majority of motor impaired individuals still retain oculomotor abilities (Istance et al., 1996). However, in order to disambiguate user intention and interaction, eye movement may have to be combined with another means of interaction such as speech. This study 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. Previous studies (Beelders, 2009; Blignaut, Dednam & Beelders, 2007) show that novice users of word processors experience a number of obstacles in acceptance and usage of a word processor that are unique to their particular 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 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. This study therefore proposes to be an extension or continuation of these aforementioned studies, and to investigate further ways to improve the interface of a word processor for all user groups. Eye-tracking, which was identified as a possible means of interaction to increase bandwidth use and meet the needs of disabled users, also provides a possible means of achieving this for these users.

The technologies chosen to improve the usability of the word processor are speech recognition and eye-tracking. As it is, Microsoft Office already comes bundled with an in-built speech engine which makes speech recognition available in all Office packages. Speech recognition offers an interaction means capable of replacing conventional typing and alleviating strain which may be caused by using an onscreen keyboard. Eye-trackers may eventually become affordable enough to be a standard feature in future computing devices (Isokoski, 2000). As it is, fairly inexpensive eye-tracking solutions have successfully been developed and used within gaze-based solutions (cf. Corno, Farinetti & Signorile, 2002; Haro, Essa & Flickner, 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). The underlying foundation of this research undertaking is the view that while eye gaze and speech recognition may be prone to ambiguity when used in isolation, using them in combination may allow many 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 & 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 problem (Chapter 2), it seems plausible that an additional modality might make selection easier and more feasible. Additionally, the actions required within a

word processor can all be facilitated through the combined use of eye gaze and speech as interaction techniques (He & Kaufman, 1993).

The goal of this study is therefore to determine whether the combination of eye gaze and speech can effectively be used as an interaction technique to replace the use of the traditional mouse and keyboard.

1.4 Problem statement

The research problem of the study is twofold: firstly to determine whether a multimodal interface using eye gaze and speech as interaction techniques is possible and feasible for a word processor; and secondly, as a feasible application does not necessarily imply a usable application, to establish the usability of such an application by comparing it to standard or traditional interaction techniques currently in use in a word processor.

1.5 Research questions

The research study will be conducted in a series of linear phases, each of which will have its own research question. The underlying proposal of the study is to determine whether the combination of eye gaze and speech as an interaction technique is a viable solution for a multimodal interface for a word processor. Therefore, it will first have to be established whether an existing word processor can be changed or emulated to incorporate a multimodal interface. Once this has been achieved, feasibility of this multimodal interface will have to be established.

Following this, the usability of the multimodal interface will have to be tested through extensive user testing. For this purpose, three main features which an interaction technique must facilitate within a word processor were identified. The user must be able:

1. to type text into the document;
2. to use the interaction technique as a pointing device in order to click on icons within the ribbon and menu of the application;
3. to achieve common word processing tasks such as formatting, document manipulation and navigation through a document without having to click on an icon or menu option.

There will therefore be three primary research questions in this study, namely:

1. Can a customisable multimodal interface be developed and successfully incorporated into a mainstream word processor with the aim of providing an all-inclusive application to a diverse group of users?
2. How feasible is such an interface and in which context is it feasible?
3. How usable is the multimodal interface compared to the traditional interaction techniques?

Based on the identification of the word processing features above, research question 3 could be further subdivided into the following secondary questions:

a. How usable is the combination of eye gaze and speech when used to simulate a pointing device?
b. How usable are speech commands for performing common word processing tasks?
c. How usable is the combination of eye gaze and speech when used for text entry?

Both the first and second research questions are exploratory in nature while the third question is a causal question as the effect that the proposed interaction techniques have on the usability of a word processor will be examined.

1.6 Scope

The possibilities presented by the proposed research study are vast and wide-ranging. Therefore, the scope of the study must be clearly defined at the outset to avoid scope creep occurring.

Since the multimodal interface is only now being proposed, this study will include both the development and the testing of the feasibility of the proposed interface. By testing the feasibility, it will allow a more learned sample to evaluate the potential, both short- and long-term, that the interface offers.

Thereafter, the usability of the interface must be investigated through objective, measurable usability metrics. Since the user base of a word processor is very diverse and the interface proposes to extend this base even further, the population which will be concentrated on must be clearly defined. Since the interface has not yet been tested, the scope of the study will include testing on proficient able-bodied users only. This will determine whether the interface is usable for the context in which it will be used.

Although the study has identified three main features of a word processor that will be concentrated upon, it is not possible to include testing on all the functionality that a word processor offers. Therefore, the tasks that will be included in the testing will represent only a subset of the functionality, but will be chosen based on the consideration that they are the most commonly used functions in a word processor environment.

1.7 Limitations of the study

Keates and Trewin (2005) state that in order to provide interfaces which compensate for disabilities, it is necessary first to fully understand the difficulties of the users. This implies that each disability will present its own challenges and require unique compensatory actions to be taken. This viewpoint is further supported by Gajos, Wobbrock and Weld (2008), who evaluated systems which automatically generated adaptable interfaces based on individual motor capabilities of users with motor impairments. Since the proposed interface may be an ideal solution for disabled users it would have to be tested using disabled users. Unfortunately, the scope of the study will not allow for these tests to be conducted, specifically not in the order that they will be required. Therefore, a limitation of the study is that only able-bodied users will be tested.

The initial motivation of the study was to provide an interface which is suitable for both novice and more experienced users. However, the nature of a longitudinal study, especially within the context of the hardware which is required for this study, together with time and budget constraints, was not conducive to the use of a large sample. Therefore, only experienced users will be tested as these will not require additional training on a word processor. Other target groups will not be tested and will have to be tested in the future in order to determine whether the proposed interface provides a viable solution to all users.

Dwell time, look-and-shoot and blinking will also be added as interaction techniques for use within the developed application. However, although these functionalities will be provided, they cannot all be tested during the formal usability testing. Therefore, only the proposed solution of eye gaze and speech for text entry will be tested and compared to the traditional means of keyboard and mouse. Furthermore, a limited grammar for speech input will be tested which implies that it will not be possible to complete all word processing tasks

through speech commands. Although this is undoubtedly a limitation of the study, it was felt that within the scope of the study it was sufficient to provide speech commands for only the common word processor tasks.

1.8 Methodology

The thesis is based on the premise of testing the principle behind the inclusion of both speech recognition and eye-tracking in a word processor application. To this end, the five research questions (section 1.5) were posed. Each of these research questions will be answered in turn using its own specific methodology, each of which will be discussed further in this section.

Research question 1: Can a customisable multimodal interface be developed and successfully incorporated into a mainstream word processor with the aim of providing an all-inclusive application to a diverse group of users?

In order to make user interaction with the test system as natural as possible, the system must emulate the real-world application as closely as possible. Therefore, a popular word processor application will be chosen as the application which must be emulated or changed to incorporate the multimodal interface (Chapter 3). Since Microsoft Word® is the most popular word processor in the current market, it was chosen as the application on which the study would focus. Moreover, Visual Studio Tools for Office (VSTO) allows programmers to add additional functionality and change the interface of applications within the Office Suite. Therefore, using these and other tools and software development kits (SDKs) which are available, eye gaze and speech functionality will be added to Word. By providing a number of means through which additional modalities can be used, the interface can be customised to suit the needs of a particular user at any given time.

This study will make use of surveys and experiments to resolve the empirical research questions, namely the second and third research questions. Surveys, both in the form of questionnaires and interviews, will be used. Questionnaires will be used in a number of capacities such as to capture user demographics, to measure user opinion of as well as user satisfaction with the proposed interface (Appendices A, C - H). Interviews will also be conducted with test participants in order to gauge their satisfaction, general impressions and comfort level with the application. Interviews will allow more open-ended questions to be posed to participants than would be the case with questionnaires. Questionnaires will contain some open-ended questions but for the most part the questionnaire will follow a structured approach.

Research question 2: How feasible is such an interface and in which context is it feasible?

In order to answer this research question, a feasibility study with a carefully selected sample will be conducted (Chapter 4). The sample will be a convenience sample and will be drawn exclusively from a population which is familiar with the human-computer interaction field. Since the study will be more qualitative in nature a sample size of 5 will be sufficient (Nielsen, 2000). The sole data collection method for this feasibility study will be a questionnaire with both closed- and open-ended questions.

This feasibility review will require participants to give an unbiased opinion of a system as their experience should allow them to accurately judge the long-term possibilities of a system, should there be no immediate short term benefits. This will allow the viability of the chosen interaction techniques to be determined without concentrating on usability measures *per se*. The aim of the feasibility review is to establish a more subjective view about whether the interface which is suggested has long-term usage potential and whether it can offer a solution that meets the needs of users.

Research question 3: How usable is the multimodal interface compared to the traditional interaction techniques?

Experiments will be used to answer all three secondary research questions. Usability experiments in human-computer interaction (HCI) generally take the form of user testing which requires that representative users must perform representative tasks on the application (Al-Qaimari & McRostie, 2001; Dillon, 2001; Preece et al., 1994; Shneiderman, 1998). Therefore, for each of the secondary questions suitable tests will have to be designed which will allow the usability of that particular word processing function to be measured (these tests will be discussed in Chapter 3). The International Standards Organisation (ISO) stresses that in order to test the usability of a product both the performance and satisfaction of the end-users must be measured in some way (ISO, 1998). In order to do this, effectiveness, efficiency and satisfaction must be defined in terms of measurable attributes (ISO, 1998; Bevan & Macleod, 1994; Scholtz, 2004). Ultimately, this research study has adopted the viewpoint that it is obligatory to select at least one measurement for each of the usability components of effectiveness, efficiency and satisfaction. The actual objective measurements which will be used will be discussed in Chapter 3. Objective measurements will be complemented by questionnaires designed to elicit subjective measurements of usability (Appendices E, G and H). Each of the user tests will make use of a convenience sample as the participants will be sourced from the university at which the study is being conducted. For the purposes of the user testing an endeavour will be made to maintain a minimum sample size of 20 (Nielsen, 2006).

Research question 3a: How usable is the combination of eye gaze and speech when used to simulate a pointing device?

The accepted means of testing and comparing pointing devices is through the use of the International Standards Organisation (ISO) standard 9241-9 (Chapter 5). This test will be used to test how best to increase the usability of eye gaze and speech as a pointing device to such an extent that it may be comparable to the performance when using the traditional mouse. The literature review (Chapter 2) will identify possible means through which usability can be increased. These will be tested and compared to the use of a mouse as a pointing device.

Research question 3b: How usable are speech commands for performing common word processing tasks?

User testing will be conducted to compare the use of traditional methods to achieve common word processor tasks and the use of speech commands (Chapter 6). These common word processor tasks will include such functions as selecting text, formatting of text, navigating through a document and manipulating the text in the document (for example, cutting and pasting). These tasks will be of such a nature that they can be completed without having to click on an icon or menu option in the application. Speech commands will be provided for these tasks so that they can be completed without the use of either a mouse or keyboard. A preset list of tasks will require study participants to complete tasks using either a mouse or keyboard and then to complete an equivalent task using speech commands. Since it may require some time for participants to become accustomed to the speech commands a longitudinal study will be undertaken. This will therefore be a repeated-measures within-subjects study. Efficiency measurements, such as time to complete a task, and effectiveness measurements, such as the level of correctness with which the task can be completed, will be measured and analysed. Furthermore, questionnaires will be used to analyse the subjective measurement of user satisfaction.

Research question 3c: How usable is the combination of eye gaze and speech when used for text entry?

The final research question will be answered using the same method as for the previous research question. Within the task list for the longitudinal testing, there will be a number of tasks which will require the participant to type random phrases using either the keyboard or eye gaze and speech (Chapter 7). Efficiency

and effectiveness measurements will be analysed. Once again, questionnaires will be used to test the subjective measurement of satisfaction.

To round off the exploration of the third research question, subjective satisfaction will be measured using established questionnaires (Chapter 8).

Data analysis will be conducted in order to make insightful conclusions from the data that has been collected. For these purposes, descriptive as well as inferential statistical analysis (section 3.5), which will be dependent on the data that is collected, will be conducted.

1.9 Outline of the thesis

This thesis will proceed according to the following outline. Chapter 2 will provide a discussion of the some of the available literature. Motivation will also be provided for the study which was undertaken. This will include discussions on the technologies which were chosen for inclusion in the study, with their associated disadvantages and how these could possibly be overcome.

Thereafter, Chapter 3 will focus on the experimental methodology and design of the study. Specific details will be given of all instruments which will be used or developed in order to explore the research questions. This will include the questionnaires which will be used as well as an in-depth discussion of the application which will be developed in order to answer the posed research questions.

Chapter 4 will discuss the results of the feasibility study which was conducted in order to establish the viability of the developed multimodal interface. Chapter 5 will report on the user testing which was conducted in order to determine how usable the proposed interaction techniques are when used to replace a pointing device.

The following two chapters (Chapters 6 and 7) will report on the results of the longitudinal user testing which was designed to evaluate objective usability measurements for the multimodal interface. This will include the comparative analysis with the more traditional means of interaction currently available for a word processor. Chapter 8 will then discuss the subjective feelings of the test participants towards the proposed multimodal interface. A number of anecdotal observations will also be reported on.

The final chapter (Chapter 9) will provide a summary of the results found as well as make some recommendations for use and further research.

1.10 Summary

This chapter provided a brief introduction to the study which was undertaken. The motivation for undertaking the study stemmed from a number of sources and provided an opportunity for a wide-reaching study with broad scope. The scope was, however, narrowed down to a manageable size which sufficed for the purposes of the thesis. A number of limitations were identified which have to be considered during the course of the study. Finally, the methodology which will be used to answer the research questions was presented and briefly discussed.

The following chapter will provide a more in-depth discussion of some of the available literature which provided the basis and motivation for the research study.

CHAPTER 2

THEORETICAL BACKGROUND

2.1 Introduction

The previous chapter gave an overview of the objectives, motivation and methodology which will be used to answer the research questions that were posed. This chapter will discuss some of the relevant literature which formed the foundation for this study. Various concepts pertinent to the study will be defined and their use explained. These include discussions on concepts such as usability, user interfaces in general and computer users.

Previous studies which are related to the current study will be reported on. In particular, the focus will be on the modalities of speech and eye gaze. In order to facilitate this discussion, the human physiology behind these technologies must be discussed. Following this, the specific technologies of speech recognition and eye-tracking will be discussed with reference to relevant studies that have used them as interaction techniques. Thereafter, the combination of the two within a multimodal interface will be reported on with specific reference to how it can be used for text entry and as a pointing device.

2.2 Word processors

> "Word processing, a concept that combines the dictating and typing functions into a centralized system, is replacing the one-man, one-secretary, one-typewriter idea in a growing number of firms. By organizing the flow of office correspondence on a more efficient basis, word processing is becoming to typing what Henry Ford's assembly line was to the original methods used for automobile making." (Administrative Management Article, December 1970 as cited in Haigh, 2006, p. 8)

Word processing is a system which allows for the flexible composition, editing, formatting, storage and printing of digital documents (Daintith & Wright, 2008) and is often regarded as the first step towards office automation (Freedman, 1998). A word processor is therefore, the software that provides these capabilities on a computer (Freedman, 1998).

The word processor application has evolved substantially since its initial inception. The original word processor - in the true sense of the word - was developed by IBM in 1969 and was known as the Magnetic Tape Selectric Typewriter or MT/ST (Eisenberg, 1992). In this model, keystrokes were recorded on a 16 mm magnetic tape and, while the MT/ST was capable of distinguishing between words, lines and paragraphs, the division of the full text into pages and the numbering of pages still had to be manually completed by a human operator (Eisenberg, 1992). Since then the word processor has undergone a virtual metamorphosis to achieve the capabilities that are available in these applications today. The introduction of MS-DOS yielded great improvement in the capabilities of word processors with the inclusion of features such as endnotes, footnotes and the ability to edit more than one document by utilising the provision of increased memory and disk space (Eisenberg, 1992). The introduction of WordStar in 1979 saw the first release of a "what you see is what you get" (WYSIWYG) word processor (Bergin, 2006a). Its developers touted WordStar as being the first word processor that was capable of showing onscreen page breaks, that had in-line help, was keystroke sensitive, had automatic word wrap and allowed users to set the left and right margins (Bergin, 2006a). When Microsoft Windows replaced MS-DOS, Microsoft Word became the word processor of choice (Bergin, 2006a; Bergin 2006b).

Two trends in the widespread adoption of the word processor are notable. Firstly, when word processing became synonymous with a computerised application, this niche in the technology field became the fastest growing and most competitive of the field (Haigh, 2006). Secondly, the falling cost associated with such technology facilitated the widespread adoption of these tools in business arenas that otherwise might not have been possible (Haigh, 2006).

One drawback of the current word processors in circulation is that they depend heavily on the user's ability to read without impediments and to be able to remember and execute a sequence of actions to perform a desired action (Dickinson, Gregor & Dickinson, 2003) – traits which not all word processor users possess in equal measures. For example, SeeWord was developed for use by dyslexic users and found to be far more suitable for this audience than a WYSIWYG word processor (Dickinson et al., 2003). Since it is evident that the word processor is constantly evolving to adapt to the needs of users and to exploit the increased capabilities offered by the newer technologies, it offers a unique environment and one rich in potential for improvement of the user experience, particularly since the current word processor may assume that its users possess certain abilities. Furthermore, the adoption of the word processor by a large group of users as it becomes affordable bodes well for the adoption of other technologies which may currently be beyond the budget of mainstream users (such as eye-tracking). Should such a feat be emulated and a new application become known for its usability and customisability it may enjoy such widespread adoption as the traditional word processor originally did. For all these reasons, the word processor and the improvement of its usability were the main focus areas of this research study. In particular, the possibility of developing a multimodal interface for a mainstream word processor and establishing the usability thereof is the aim of the study.

2.3 Usability and user experience

There are many definitions available for usability (cf. Shackel, 1991; Shneiderman, 1998; Wixon & Wilson, 1997). The International Standards Organisation (ISO) formalised the definition of usability in the 9126-1 standard as "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions". This definition is further expanded upon in ISO 9241-11 where usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO, 1998).

These ISO definitions were considered appropriate for the purposes of the current study and were combined into a single definition which encompasses all the salient parts, namely:

> *The **usability** of the system can be measured as the extent to which a software product can be used to achieve specified goals with effectiveness, efficiency and satisfaction as well as the extent to which it exhibits the capability to be learned and understood by the user*.

The definitions allow measureable components to be extracted in order to determine the usability of a product by requiring users to complete certain tasks on the system in question. The four identified components can be defined as follows:

- Effectiveness is how well the user is able to achieve that which must be done by using the system (ISO, 1998) and can be measured in terms of accuracy and completeness (Cato, 2001).
- Efficiency is the amount of resources required to complete the desired task (ISO, 1998), such as time, money or mental effort (Bevan & Macleod, 1994).
- Satisfaction is a subjective feeling and relates to the attitude of the user towards the system (ISO, 1998).
- Learnability measures not only the time taken for a user to become familiarised with the system but also how well the user is able to remember system functionality (Cato, 2001).

Each of these components can be measured in some way depending on the task at hand and which part of the system is being tested. For the purposes of determining usability in the current study, certain measurements, all conforming to the above-mentioned components of usability, will be recorded and analysed. Since usability encompasses all of these components, where possible a measure of each of these will be used to provide a representative view of the usability of the inspected interface feature. The actual measurements which will be analysed for each part of the study will be discussed under the relevant sections in Chapter 3.

In recent years, there has been a movement towards evaluating the user experience and not simply evaluating usability. Similar to usability, user experience has a number of definitions ranging from being synonymous to usability, to encompassing beauty, affective or experiential aspects of using technology (Hassenzahl & Tractinsky, 2006). Some texts consider user experience to be a broader field than usability and represent it as the convergence of usability, branding, functionality and concept (Rubinoff, nd) or "the creation and synchronisation of the elements that affect users' experience with a particular company, with the intent of influencing their perceptions and behaviour" (Unger, 2009).

Essentially, the user experience can be summarised as "a consequence of a user's internal state, ... the characteristics of the designed system ... and the context within which the interaction occurs" (Hassenzahl & Tractinsky, 2006) or "the characterisation of what a user feels while using a product" (Paluch, 2009). ISO 9241-210 defines the user experience as a very subjective concept in terms of the person's perceptions and responses while using the system (ISO, 2010).

From all these definitions it is clear that the user experience can be interpreted as being much broader than the definition of usability which has been accepted for use in this study. This sentiment was echoed in the definitive guide on user experience, where Tullis & Albert (2008) reiterate that usability and the user experience are two separate concepts with user experience including aspects such as the thoughts, feelings and perceptions that result from interaction with the product. They use the term usability and user experience interchangeably and advocate the use of usability metrics which measure some aspect of the user experience. To this end, efficiency, effectiveness and satisfaction must be measured together with expectations, ease-of-use, awareness and behavioural and physiological metrics such as eye-tracking, facial expressions and measures of stress (Tullis & Albert, 2008).

While this study will test the usability of the proposed application it will inspect aspects beyond the formal usability metrics which have been defined. Interviews will be conducted in an attempt to elicit other self-reported responses to the application which do not generally fall within the confines of usability. Additionally, a smaller study will be conducted to test the feasibility and user reaction to the proposed system. A large portion of the following chapter will contain a discussion of the characteristics of the system, how it was developed and how it was endeavoured to create a full-scale highly customisable application which caters for a large, diverse group of users. Therefore, the term user experience will be used to refer to the all-encompassing study which includes all the aforementioned aspects.

2.4 User interfaces

The user interface is the conduit between the user and the computer and as such plays a vital role in the success or failure of an application. A graphical user interface (GUI) is an interface that makes use of input devices other than the keyboard (Daintith & Wright, 2008). GUIs usually make use of windows, icons, menus and pointing devices, and are therefore often referred to as WIMP interfaces. Currently, the user interface finds itself in the dubious situation of being in a design rut, with the current trend of interfaces having been around since the 1970s (Van Dam, 2001). Some designers feel that the time has arrived to concentrate on discovering innovative post-WIMP interfaces which do not rely solely on menus and icons (Van Dam, 2001). A natural consequence of this may be that the mouse and keyboard will no longer be suitable input methods.

Instead, indications are that future computer interfaces should foster natural and intuitive communication that emulates human-human communication (Wachs, Kölsch, Stern & Edan, 2011). Since users are often notorious for their unwillingness to accept changes in their interaction methods, these interfaces must be accessible and not require long periods of learning (Wachs et al., 2011). One of the major challenges facing the HCI community at present is the development of these alternative means of input which move away from the traditional manual inputs of mouse and keyboard (Miniotas, Špakov, Tugoy & MacKenzie, 2006). This may well be the area in which perceptual, attentive, non-command, brain-computer and multimodal user interfaces find themselves on the forefront of the technology wave.

2.4.1 Perceptual, attentive and non-command user interfaces

Perceptual user interfaces coordinate perception using multimodal input and multimedia output modelled after natural human-computer interactions (Maglio, Matlock, Campbell, Zhai & Smith, 2000), aimed at allowing users to interact with technology in much the same way as they interact with each other (Turk, 2001). Perceptual interfaces are interactive and utilise the senses to provide interactions which cannot be accomplished through traditional input devices (Turk & Kölsch, 2004). An example of a perceptual interface is the Kid's Room, a narrative play-space for children with two walls constructed from video projection screens to allow the room to be transformed into a magical play area (Bobick et al., 1999). For example, a river world is created where children are encouraged to use the bed in the room as a boat and to row it down the river. Should any person exit the boat a splashing sound is heard and the person is encouraged to climb back into the boat.

Attentive user interfaces go a step further than perceptual interfaces since they must not only perceive, but also anticipate the user's next action (Maglio et al., 2000). For example, in a multi-monitor set-up, the attentive interface developed by Ashdown and Sato (2005) automatically moves the mouse cursor to the monitor the user was looking at and causes the topmost window on that monitor to receive focus. This would undoubtedly be an attractive solution to any user of multiple screens and windows.

Jacob Nielsen (1993) coined the term non-command interfaces to describe the future generation of user interfaces which would infer meaning without having to receive explicit commands from the user. These interfaces seem to be very similar to attentive user interfaces but also lean heavily on the theory of ubiquitous computing which allows the interfaces to be embedded into the user's physical environment. Virtual reality is an example of a non-command interface as it allows users to immerse themselves in a simulated world and move about and interact with the virtual world in the same way as they would in the real world.

The fact that these three types of interfaces utilise human senses or strive to react to human attention means that current input devices will not be sufficient (cf. Dirican & Göktürk, 2009; Turk & Kölsch, 2004). Instead, technologies which allow human senses to be mimicked or reacted to (Turk & Kölsch, 2004), such as speech recognition or olfactory devices are required. Furthermore, eye gaze position (or the point of regard) is possibly the simplest means by which to infer the focus of users' attention (Just & Carpenter, 1976). For these purposes eye-tracking, which will be discussed in a later section, can be used. While perceptual, attentive and non-command user interfaces are an exciting and promising new field of research they are beyond the scope of the current study. The technologies used in this study are, however, ideal for these interfaces and cognisance will be taken of this throughout the study as the potential of the proposed interface stretches far beyond that which will be investigated.

2.4.2 Brain-computer user interfaces

A brain-computer user interface (BCI) is a computer interface that is able to respond to human thoughts and intentions (Nijholt & Tan, 2008). Although the main focus of BCIs is to enable disabled users, BCIs may become equally acceptable for use within the able-bodied community, in the form of gaming interaction or map navigation (Nijholt & Tan, 2008). Although many may shy away from the use of such a seemingly complex system, the fact remains that it presents certain advantages above and beyond traditional interaction devices – particularly in environments where the user's hands are busy and additional interaction devices are required to increase productivity, or in systems where bandwidth is insufficient, such as some gaming environments (Nijholt & Tan, 2008). Similar to the way in which BCIs can be justified for use within the mainstream computer population, so too can eye-tracking and speech recognition.

BCIs are beyond the scope of the current study but remain an exciting prospect for the future of user interfaces.

2.4.3 Multimodal user interfaces

In order to define a multimodal interface it is necessary first to define a modality. A modality is a means of communication using one of the five human senses or type of computer devices that is equivalent to human senses (Jaimes & Sebe, 2005) or the way an action is performed (Coutaz & Caelen, 1991). Similar to usability, there are a number of definitions available for multimodal interfaces, some of which are listed below:

- A multimodal interface uses a combination of communication means using the human senses or equivalents thereof, thereby responding to multiple input channels (Jaimes & Sebe, 2005).
- A computer system is said to be multimodal if "it supports human modalities such as gesture, written or spoken natural language" (Coutaz & Caelen, 1991).
- Multimodal interfaces "process two or more combined user input modes – such as speech, pen, touch, manual gestures, gaze and head and body movements – in a coordinated manner with multimedia system output" (Oviatt, 1999).
- A multimodal interface is one in which several input and output modalities are combined in an effort to assist human-computer communication through utilising natural human communication channels (Pireddu, 2007).
- The aim of a multimodal interface is to make a computer behave in a fashion similar to human communication which should facilitate easier learning and use (Kaukènas, Navickas & Telksnys, 2006).

There are common elements to these definitions, for example that the purpose is to make interaction more natural through the emulation of human-human communications. For the purposes of this study, a combination of the definitions will be used to define a multimodal interface. The following definition therefore applies:

> *A **multimodal interface** uses several human modalities which are combined in an effort to make human-computer interaction easier to use and learn by using characteristics of human-human communication.*

Multimodal interfaces themselves date back to 1980, when Richard Bolt, in his seminal work entitled "Put that here" (Bolt, 1980), combined speech and gestures to select and manipulate objects. Using a projected image of a workspace, a media room was used to create the impression of a virtual workspace as opposed to simply working on a computer. Speech recognition provided a grammar through which commands could be issued to create, move and change onscreen elements. Gestures were interpreted so that elements could be positioned and moved from one location to another. Commands such as "Create a blue triangle here" would cause a blue

triangle to be drawn where the user was pointing. Similarly, saying “Move this there” would move an element from its position, indicated by pointing at it, to the new location the user was pointing at.

The following year, using the same media room with a projected workspace, Bolt was responsible for the first gaze-controlled interface when he tested an application called World of Windows (WOW) (Bolt, 1981). WOW was essentially a gaze-controlled interface which allowed multiple windows to be displayed simultaneously to a user. Based on the user’s gaze, a window could be “zoomed into” by concentrating on it long enough (analogous to dwell time) or through an additional action such as a spoken command. The currently zoomed window could be reset by either looking away from the workspace or gazing at another window in the display.

A distinct advantage of multimodal interfaces is that they offer the possibility of making interaction more natural (Bernhaupt, Palanque, Winkler & Navarre, 2007). Furthermore, a multimodal interface has the potential to span across a diverse user group, including varying skill levels, different age groups as well as increasing accessibility for disabled users whilst still providing a natural, intuitive and pleasant experience for able-bodied users (Oviatt & Cohen, 2000). This statement played a significant role in the motivation to undertake this study. Current interfaces are not capable of living up to these expectations and a multimodal interface that meets these needs must still be discovered.

Therefore, this study will propose a multimodal interface for a word processor application which makes use of natural human modalities in an attempt to provide a hands-free, intuitive, easy to use and learn interface which can cater for a diverse group of users. In this regard, speech offers an intuitive means of communication which requires very little to no training in its application within a computer interface. Coupled with speech, humans often make use of their hands, body language and eye gaze to infer meaning and intent with the spoken words. These human qualities offer a wide variety of possibilities in this new generation of interfaces and this provides ample motivation for the investigation into a multimodal interface using speech and eye gaze, which forms the basis of the suggested multimodal interface in this study.

2.4.4 Interaction techniques

Using a physical input device in order to communicate or perform a task in human-computer dialogue is called an interaction technique (Foley, Van Dam, Feiner & Hughes, 1990 as cited in Jacob, 1995a). However, for the purposes of this study, the definition will be modified and used in the following context:

> *An **interaction technique** is the use of* any *means of communication in a human-computer dialogue to issue instructions or infer meaning.*

The proposed multimodal interface will provide a number of interaction techniques to heighten the customisability it offers to the user. In this way, it may be possible to cater for a very diverse group of users as previously stated. Interaction techniques may oftentimes be a single modality but the term will also be used to refer to a number of modalities which are combined into a single interaction technique.

2.5 Computer users

Users are those people who will eventually interact with the product or application (ISO, 1998). The profile of computer users has changed from the inventors themselves (first generation), to the technocrats and computer professionals (third generation), to include everybody in the current and future generation of user interfaces (Nielsen, 1993). This has a direct consequence in that applications must cater for a large and diverse user base in order to ensure the continued use of the application. One of the primary tasks which must be completed in any graphical user interface is to position a cursor over an object and select that object (Keates &

Trewin, 2005). This action can prove difficult for older users and users with disabilities (Keates & Trewin, 2005). Some of the different types of computer users, including aged and disabled users, will be concisely discussed in the following sections.

2.5.1 Types of users

Any computer user, regardless of physical abilities or age, can be classified according to level of expertise. The level of expertise is measured in terms of experience with both the task and the interface domain (Shneiderman, 1998). Novice users have little knowledge of either the task or the interface concepts while first-time users have sufficient knowledge of the task concepts but limited knowledge of the interface concepts (Shneiderman, 1998). Knowledgeable intermittent users have knowledge of both the task and the interface domain but their infrequent use of the interface prevents adequate retention of the interface components (Shneiderman, 1998). The final category of user is the expert user who displays extreme competence in both domains (Shneiderman, 1998).

Originally, the aim of the study was to test as many of the user categories as possible on the multimodal interface using eye gaze and speech. However, it quickly became clear that the scope would be far too large; therefore the target group was limited to expert word processor users so that no training on word processor use would be needed. Since the proposed multimodal interface was a new concept, there could be no users of any expertise levels for the interface. There could, however, be users of the individual modalities of eye gaze and speech. However, the exclusivity of the modalities and the cost of high quality equipment needed for the modalities, especially eye-tracking, reduced the chances of finding a large enough sample size. Therefore, it was decided to rather focus on first-time and novice users of both of these modalities.

2.5.2 Aged users

The aging IT population could have severe consequences for the design of user interfaces. There are several factors unique to the aging demographic that must be accounted for in user interface design. These users experience a reduction in light sensitivity, colour perception, dynamic and static visual acuity and contrast sensitivity (Murata, 2006). Furthermore, there is a decrease in sensory and motor function (Thomas, Basson & Gardner-Bonneau, 2008) which impacts on the use of the mouse by increasing the pointing time (Murata, 2006). Aging users will play a pivotal role in the development of assistive technologies as indicated by the projected statistics on the age of computer users. In the United States, 1 in 5 workers in 2020 will be over the age of 55 – an increase of over 50% from the year 2000, when only 13% of the workforce fell in that category (Thomas et al., 2008). The same phenomenon is evident in Asia Pacific and Europe where 20% of the Japanese and Italian population were 65 or older in 2006 (Thomas et al., 2008).

A representative sample in terms of age and gender showed that when positioning the cursor, older users take longer and pause more often than younger users do (Keates & Trewin, 2005). Murata (2006) conducted an experiment to determine the impact of aging on mouse handling amongst three age groups, namely, young, middle-aged and older (Murata, 2006). All groups consisted exclusively of men, all of whom had experience with mouse handling but none of whom had ever used an eye-tracker. Participants were required to point to a predefined target on screen first using a standard mouse and then using the eye-tracker. In terms of just mouse pointing time, it was found that older adults required more time to point with a mouse than younger adults, which indicates a lengthening of manual input time required as age increases. All age groups performed better when using eye gaze to select targets with a marked decrease in the difference between the groups when using eye gaze. Results show that eye gaze pointing can be mastered and performed at high speed regardless of the user's age, and is a recommended pointing device for the older computer generation. Younger adults preferred pointing with a mouse while middle-aged and older users found the eye gaze

pointing to be very easy. A limitation of Murata's study is, of course, that only one gender was tested and the results were not verified with females as well. Nevertheless, it is acknowledged that aging users have characteristics that must be considered and which are unique to this user demographic but which can possibly be overcome through the use of non-manual interaction techniques.

While aged users will not be tested in the current study, this discussion is relevant in the development of a multimodal interface. While it has been theorised that eye gaze might not be suitable for aging users due to the natural effects of aging on the eye, results from previous studies fail to verify this theory. Therefore, eye gaze remains a possible means of interaction for older users and may still be a more usable interaction technique than the mouse. Similarly, results for speech-enhanced user interfaces with older users have yielded mixed results (Basson, Fairweather & Hanson, 2007) and further investigation is required. The combination of the two modalities has, as far as can be ascertained, not been tested with this demographic. Therefore, this group will be taken into account when designing a multimodal interface and the importance of testing these users on the proposed interface is recognized and suggested for future research based on the findings of the current study.

2.5.3 Disabled users

A computer can play a vital role in the everyday lives of users with functional impairments (Keates, Hwang, Langdon, Clarkson & Robinson, 2002). However, the effective use of a physical input device is heavily dependent on the ability of the user to feel, or have the knowledge that they are touching or holding the device as well as being able to move, manipulate and operate the device (Bates, 2002). Regrettably, traditional input devices, such as the mouse and keyboard, are designed with able-bodied users in mind. Consequently, disabled users cannot always manage to use these devices to their full potential or even use them at all (Su, Su & Chen, 2005). The most sensible way of empowering disabled users is to provide them with 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 et al., 1996). Therefore, it is imperative that alternative means of interaction be found for these users as this will allow them to use technology on a level comparable with able-bodied users.

Although no official statistics are available for South Africa, it is estimated that between 400 and 500 spinal cord injuries are sustained annually (Quadriplegic Association of South Africa, nd). In the United States of America, this figure escalates dramatically with 11 000 sustained injuries every year with a total of 250 000 spinal cord injured Americans (www.sci-info-pages.com). Eye movement is ideal for such situations as learning time may be reduced through the use of a "natural" means of pointing (Istance et al., 1996). Furthermore eye movement is high-speed (Istance et al., 1996) and the majority of motor impaired individuals still retain oculomotor abilities (Hornof, Cavender & Hoselton, 2004). However, the disadvantages associated with eye-tracking as an input device mean that it should be used with caution or, as suggested by Istance et al. (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. For example, when using eye gaze as a pointing device, speech can be used as a triggering event instead of just using eye gaze. Speech can also offer a means of interaction to these users in terms of text input, as in the case of Speech Dasher (section 2.9.4). However, there may be instances where the vocabulary of the user is limited and dictation is not possible. This study proposes a solution to the problem by allowing text input using eye gaze and speech. This combination may prove to be more usable than using eye gaze in isolation and may offer a means of text input for those users who are incapable of using speech recognition to its fullest potential.

It would be very beneficial to test the proposed interface using disabled users since they could very well be the user group which will benefit the most from such an application. However, the fact that disabilities can be wide-ranging and unique to an individual makes it very difficult to infer findings back to a general population

and analysis will have to be performed carefully on such small samples. Therefore, in order to increase the applicability of the findings, disabled users were not included in the current study but remain a prospect for future research. The possible needs of these users will be considered throughout and the decisions made will reflect this. For example, the fact that these users may have restricted mobility or limited vocabularies will be a consideration throughout.

2.6 Human modalities

As established in a previous section (section 1.3), the current study will focus on the development and testing of a multimodal interface which uses eye gaze and speech. In order to understand the technologies which make the use of these human modalities possible, it is necessary first to understand the human physiology behind the human vocal and vision system. The subsequent sections will briefly discuss these.

2.61. Human vocal system

When humans speak (Figure 2.1), air is forced from their lungs through their mouths and nasal cavities and then changed by the lips and tongue (Forsberg, 2003). The air exhaled from the lungs during speech causes oscillations of the vocal chords which are situated in the larynx (Fitch, 2000). Acoustic energy is produced and filtered (Fitch, 2000) to eventually create discernible sounds.

Since speech is the most common form of communication between humans its incorporation into user interfaces offers the possibility of a more natural human-computer interaction. When humans speak, they do so expressively, which means they use their eyes, hands and body to convey more meaning than with just simple spoken words. Therefore, the combination of speech with one of these expressions would seem to be a most natural means of communication. For the purposes of this study, speech was coupled with eye gaze.

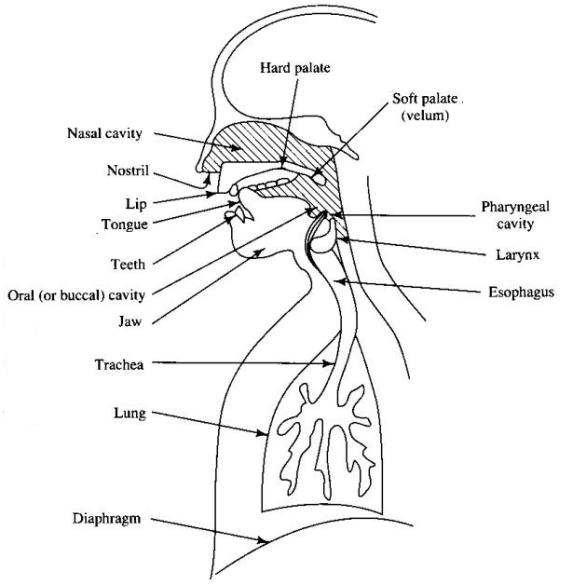


Figure 2.1: Cross-section view of human vocal system

Source: www.msu.edu

2.6.2 Human vision system

2.6.2.1 Physiology of the eye

The eye (Figure 2.2) is an organ which is responsible for collecting light and sending it to the brain to be processed into images (Yale Medical Group, nd). The outer layer of the eye consists of the anterior transparent cornea and the posterior sclera which is a dense, opaque, fibrous tissue (Atchinson & Smith, 2000). The cornea is responsible for the most refraction of light while the lens is responsible for accommodation, which is achieved by changing the shape of the lens as required (Gregory, 1966). The iris is situated on the middle layer of the eye (Atchinson & Smith, 2000). Pigmentation is present in the iris, giving humans the colouring of their eyes (Gregory, 1966). The inner layer of the eye is the retina which is connected to the brain via the optic nerve (Atchinson & Smith, 2000). The pupil is a tiny hole formed by the iris through which light passes to reach the lens and then fall onto the retina to form an image (Gregory, 1966).

The visual field can be divided into three areas, namely the foveal, parafoveal and peripheral areas (Rayner, 1998). The central region of the retina, called the fovea, is very densely packed with receptors (Gregory, 1966) and in order to see an object clearly, the eye is moved, using the six oculomotor muscles, so that the fovea, which has the highest visual acuity, is placed on the object of interest (Rayner, 1998). Objects that are large enough can be accurately identified in the peripheral vision (Rayner, 1998).

2.6.2.2 Eye movements

There are various types of eye movements in a human, some of which will briefly be discussed in this section.

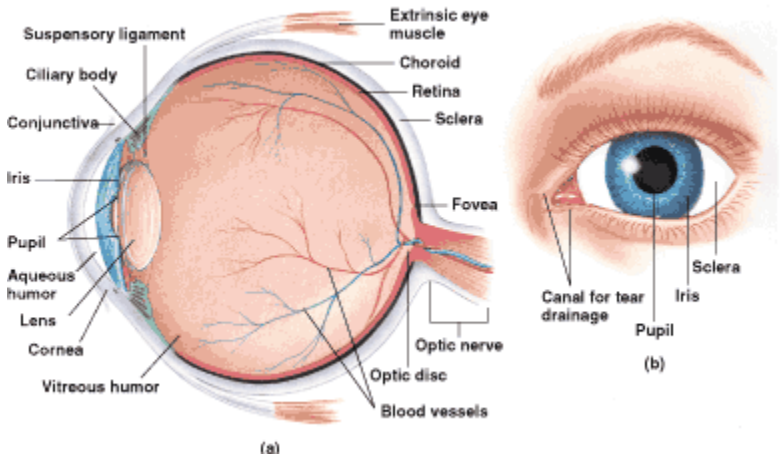


Figure 2.2: Physiology of the eye

Source: Yale Medical Group (nd)

Eye movements are required to locate a stationary object and are essentially a series of rapid jerks, known as saccades (Gregory, 1966). Visual sensitivity is reduced during saccades (Rayner, 1998). Apart from saccadic movement, there are three different eye movements which can be identified. They are (i) pursuit, which occurs when the eye follows a moving target, (ii) vergence, which occurs when the eyes are both moved inward in order to fixate upon an object and finally (iii) vestibular movements are compensatory movements to maintain visual direction and are made in response to head and body movements (Rayner, 1998). Since

these are all rapid eye movements, they are not of significance in this study and therefore, will not be discussed any further or taken into account during the experimental design or analysis of the collected data.

Between saccades, the eyes experience a period in which they remain relatively still (Rayner, 1998). These periods of stability are called fixations and generally last between 200 and 300 milliseconds (Rayner, 1998). A fixation occurs when an individual attempts to maintain their eye gaze on a stationary point (Ditchburn & Ginsborg, 1953), which can be regarded as focusing of attention on a specific object. During a fixation three different eye movements are present namely, tremor, drifts and microsaccades (Martinez-Conde & Macknik, 2008). These are collectively referred to as fixational eye movements.

The purpose of tremor, also known as nystagmus, is unclear but it may be responsible for assisting the nerve cells of the retina to keep firing in order to ensure perceptual acuity (Rayner, 1998). Drifts occur simultaneously with tremor and are slow motions of the eye, possibly used in the absence of microsaccades to maintain accurate visual fixation (Martinez-Conde, Macknik & Hubel, 2004). Due to imperfect control of the oculomotor system, the eyes sporadically experience small drifting movements away from the fixated target and then a microsaccade occurs to compensate for this drift and to move the eyes back to where they were (Rayner, 1998).

There are also three types of overshoot in saccadic eye movement, namely dynamic overshoot, glissadic overshoot and static overshoot (Bahill & Clark, 1975). A static overshoot is corrected using a corrective saccade but when the eye stops short of the intended target it tends to drift to its final position (Bahill & Clark, 1975). These slow drifts are called glissades (Bahill & Clark, 1975).

The focus of this study will be to use eye gaze, specifically in the capacity of fixations to indicate intention, as an interaction technique within a multimodal interface. Therefore, it will be necessary to determine how fixations can be used. These methods will be discussed in section 2.8.3. When using eye gaze for interactive purposes, fixational eye movements play a role in the stability of the eye gaze pointer. Since eye gaze will be used as a pointer in the current study, it is necessary to determine how the accuracy and stability of eye pointing can be improved. A more in-depth discussion of these techniques will follow in section 2.8.4.2.2.

2.6.3 Temporal relationship between eye gaze and speech

When engaged with objects, the eyes tend to look directly at the objects but the fixation which provides the information required to interact with the object occurs prior to the action (Land & Tatler, 2009). Psycholinguistic studies have also shown that there is a temporal relationship between eye gaze and speech (cf. Just & Carpenter, 1976; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995), often referred to as the eye-voice span. The eyes move to an object before the object is mentioned (Griffin & Bock, 2000) with an approximate interval of 500 milliseconds between the eye movement and speech (Velichkovsky, Springer & Pomplin, 1997 as cited in Kammerer, Scheiter & Beinhauer, 2008). However, recently it has been shown that these fixations on objects of interest could occur anywhere from the start of a verbal reference to 1500 milliseconds prior to the reference (Prasov, Chai & Jeong, 2007). While the relationship between eye gaze and speech could be confirmed in a separate study, a large variance in the temporal difference between a fixation and a spoken reference to an object was also found (Liu, Chai & Jin, 2007) which could explain the various temporal differences reported on in different texts. In a situation where multiple objects must be referred to in a single verbal utterance, the next object is already being fixated upon while speech is being produced for a certain object (Griffin, 2001).

Eye gaze has been successful in resolving ambiguities when using speech input (Tanaka, 1999). However, when implementing systems which use both eye gaze and speech, it is important to respond to the input channels by correctly identifying how to synchronise the two. It has been found that for the majority of verbal requests,

users were looking at the object of interest when the command was issued. For the remainder of the instances, users tended to look at the object more before the request was issued than after the request was issued (Maglio et al., 2000). More specifically, where eye gaze and speech were combined in an interface it was found that input events will generally occur within a range of 60-100 milliseconds of one another (Kaur, et al., 2003).

Since eye gaze and speech will be used as an interaction technique in the current study, the temporal relationship between the two modalities is of relevance as it may impact the usability of the interaction technique. In order to maximise the disambiguation of both modalities, the user will be expected to maintain eye gaze on the desired object whilst issuing the verbal command to interact with that object.

2.7 Speech recognition

Automatic speech recognition (ASR) is the process whereby human speech is interpreted in a computer (Forsberg, 2003) through the process of mapping the acoustic signals generated by the human vocal system to words (Jurafsky, 2000). Speech that is captured is first digitised, confirmed against a dictionary and then converted and, if required, displayed as typed text (Freedman, 1998).

The first foray into speech recognition produced a toy dog dubbed Radio Rex in the 1920s which recognised its name and emerged from its doghouse when called (Russel & Norvig, 2009). However, the first electronic speech synthesiser was only developed in 1936 by AT&T Bell Labs (Russel & Norvig, 2009). In the early 1970s the Defense Advanced Research Projects Agency (DARPA) took an interest in speech recognition and funded four projects to develop high performance speech recognition systems (Russel & Norvig, 2009). It was however only in the 1980s that speech recognition became commercially available (Dragon Naturally Speaking, nd). With the first release of speech recognition engines, the technology was expensive and therefore not suitable for the mass market. Since then, advances in technology as well as in the fields of digital signal processing, pattern matching and classification algorithms have made speech recognition commercially attainable even for personal computer environments (Karl, Pettey & Shneiderman, 1993). This has allowed speech recognition to become widely available to the extent that it is now a standard feature of current computers. Since the technology is now readily available to the general user population, ways in which it can effectively be utilised must be investigated.

2.7.1 How speech recognition works

Vocal communication between humans and computers can either be in the form of text-to-speech (TTS) synthesis or ASR, also known as speech-to-text conversion. Algorithms designed for synthesis have been more successful than those designed for recognition, due to the complexity of interpreting speech (O'Shaughnessy, 1995).

Essentially, ASR is a pattern recognition task which requires that the received speech signal be matched to corresponding text (O'Shaughnessy, 1995). Pattern recognition tasks generally have a training phase which is followed by a recognition phase (O'Shaughnessy, 1995). The training involves the creation of a reference memory or a dictionary of speech patterns (O'Shaughnessy, 1995). Recognition involves a number of steps, namely (1) normalisation, (2) parameterisation, (3) feature extraction, (4) similarity comparison and (5) a decision. Normalisation involves the removal of variability in the input signal as a consequence of the environment, after which the signal is divided into parameters and features. Parameters constitute the outputs from standard speech analysis while features are the outputs of further analysis (O'Shaughnessy, 1995). Recognition is then attempted by comparison of the input signal with reference templates obtained

during training (O'Shaughnessy, 1995). A decision as to what text must be output is then made, based on the template with the closest match to the received signal (O'Shaughnessy, 1995). However, if the match is too poor the decision must be postponed (O'Shaughnessy, 1995).

In order to successfully recognise boundaries between words, some technologies require speakers to pause briefly between word utterances (O'Shaughnessy, 1995), a situation which hardly supports natural communication between human and computer. These types of isolated word recognition engines can only process speech at a rate of 20 to 100 words per minute (O'Shaughnessy, 1995). However, connected systems allow sequences of concatenated words while continuous recognition affords the speaker the naturalness of speaking fluently (Nusbaum, De Groot & Lee, 1995) and are typically capable of processing between 150 and 250 words per minute. This does, however, increase the intricacy of recognition substantially (O'Shaughnessy, 1995).

The grammar which will be used for the current study consisted of commands comprising only a few words at a time. Therefore, any one of the afore-mentioned speech engines would be suitable. However, since the interface has to exhibit potential for a multitude of interaction techniques, a continuous speech recognition engine will be used to provide for more natural speech interaction in the future. All participants in the study will be expected to complete training to create a speech profile for themselves. This will increase the recognition accuracy and improve system reaction to the grammar.

2.7.2 Functions of speech recognition

Speech recognition can fulfil two types of functions, namely dictation and command and control (Pireddu, 2007). Dictation is primarily used to transcribe documents into a digital form (Pireddu, 2007). Command recognition requires a spoken word to be verified against a grammar and then the system responds to that command. Therefore, speech recognition and generation is ideal for use in situations where the environment is hands-busy, eyes-busy, mobility-required or hostile and is very promising for use in telephonic services (Shneiderman, 2000).

For the purposes of the study, the naturalness and theoretical high performance levels which can be achieved with continuous speech recognition is ideal. Within the environment of a word processor, dictation may be of the utmost importance as the use of dictation will allow the user to speak aloud and have the spoken text automatically transcribed to the current document. Memorised text can be spoken 5 times faster than it can be written, however since the composition of the desired text consumes the majority of the time taken, dictation may only increase a writer's speed by 20 to 65 percent (Schmandt, Ackerman & Hindus, 1990). Even so, this may be a significant improvement over typing speeds. Although a fully functional dictation engine will be provided, the testing and comparison thereof to standard text input is beyond the scope of the study as the focus of the study is on determining the usability of different means of text input. Therefore, participants will be expected to input text using an onscreen keyboard one letter at a time. The question which is now posed is whether this will result in increased typing speeds or whether the requirement of concentrating on the onscreen keyboard will affect the composition skills of the participant. Theoretically, the onscreen typing will be similar to typing using a keyboard but the effort required to use the eyes as a control device may cause additional strain and negatively affect the compositional speed of the user. Free-writing text will not be tested in this study.

The second way in which speech will be incorporated is through the compilation of a grammar containing common word processor commands. When these commands are issued and recognised by the speech engine, their word processing counterparts will be executed. This will alleviate the need to use either the mouse or keyboard to edit or format word processing documents. In addition, the speech grammar will also allow text input via an onscreen keyboard when used in conjunction with eye gaze to establish user attention.

2.7.3 Considerations and factors influencing speech recognition

Some tasks simply interfere with one another as they draw on the same cognitive resources and as such cannot be performed to their full potential when executed in parallel (Suhm, 2008). For example, most humans can type and think simultaneously but find it much harder to speak and think at the same time (Shneiderman, 2000). Hand-eye coordination is accomplished in a different part of the brain which allows it to be performed in parallel with problem solving (Shneiderman, 2000). It was found that when issuing commands such as “bold” and “page up” there was a marked increase in the speed with which tasks could be accomplished. However, when coupled with a memorisation task followed by a “page down” command users found it difficult to complete the task and had to repeatedly scroll back to the symbols that had to be memorised (Shneiderman, 2000).

Moreover, users voiced concern over the commands that had to be memorised to complete tasks in a word processor environment (Karl, Pettey & Shneiderman, 1993). Memorisation and the level of difficulty required by the user to remain within boundaries of the stipulated sublanguage belonging to the domain are important factors when designing a speech recognition system (Forlines, Schmidt-Nielsen, Raj, Wittenburg & Wolf, 2005). However, in this instance it is somewhat surprising that participants were concerned over the amount of memorisation required for command execution. Within the confines of the word processor environment there is a grammar somewhat unique to its environment and learning this terminology is part of the learning curve of a word processor. Therefore, the use of these within a spoken grammar should not require additional memorisation on the part of the user and they should comfortably be able to stay within the confines of the provided grammar as it would closely resemble the names and descriptions already used within the application. The grammar used in the study by Karl et al. (1993) consisted of only 18 words, the majority of which were well-known word processing terms which the users should have been familiar with. It is envisaged that it will only be the less commonly used tasks which will require some effort to retrieve from memory but since this is the case even when the icons or menu are used it can hardly be considered a drawback of speech recognition.

Nevertheless, in order to alleviate the memorisation required by the user, an application can also be tasked with the responsibility of learning a vocabulary which gives the user the freedom to provide speech commands that are understandable and intuitive to them. For example, after the Java-based jfig drawing application was adapted by Gorniak and Roy (2003) to include speech, users could train the application to associate their chosen speech commands to actions in the application. Additionally, as a consideration to ensure that the user stays within the confines of the permitted sub-language, it is possible to design a context sensitive menu, in the same vein as the Things To Say (TTSay) menu used in a dialogue system for thermostat control (Freudenthal, Keyson, DeKoven & De Hoogh, 2001). When the system determines that specific commands are applicable to the current context in which the system finds itself, a menu displaying these acceptable commands can be shown to the user to restrict them to a limited number of commands. In such a way it is analogous to a context-sensitive menu or enabling of allowable menu items. Furthermore, it is comparable to the pop-up context toolbar and the context-sensitive ribbon (replacement for the menu and standard toolbar) that has been in use from Word 2007 onwards. Regardless of the strain that memorisation of commands places on the user, using speech for non-dictation purposes is faster than using the mouse and keyboard as confirmed by Cohen, McGee and Clow (2000) although dictation was found to be slower than text input using the keyboard (Sears et al., 2001).

The underlying fact that all participants in the current study will be experts in word processing was the influencing factor in deciding not to place the burden of learning the grammar on the application. The commands included in the grammar will be tailored according to the terminology of a word processor and should not be beyond the abilities of the user. Therefore, it is theorised that the bulk of the grammar will quickly be adopted by the users and those isolated incidents where there is no verbal counterpart for the command, such as with the keyboard navigation, will easily be learnt by the users. Furthermore, providing a

grammar with visual examples of the effect the command will have, may actually serve to ease the learning curve of novice and first-time users, particularly for users attempting to teach themselves the intricacies of word processing. Participants in the study will be closely monitored to determine the difficulty experienced in memorising the command list provided.

An additional consideration of speech recognition is the switching between voice commands and typing, which can be "disruptive" (Morrison, Green, Shaw & Payne, 1984 as cited in Karl et al., 1993) but unfortunately, all systems which use ASR in some way face the challenge of correctly distinguishing between dictation and commands. A rocker switch on the microphone has been used to toggle between dictation and command issuing (Oviatt et al., 2000) but this eliminates the possibility of a truly hands-free environment. For the purposes of the current study toggling between command and dictation is irrelevant since dictation will not be tested, although solutions should still be considered. Since eye gaze will be included, dwell time (section 2.8.3.1) can be used to toggle between command and dictation mode by simply glancing at the ribbon. This may however be considered just as disruptive as the mechanisms for toggling between the states. Alternatively, a voice command can be used to toggle between dictation and commands which would logically seem to be the fastest method of switching states.

Other factors which influence the performance of speech recognition in a given situation include fatigue, effort and stress (Nusbaum, De Groot & Lee, 1995). The human ability for resilience and flexibility can, however, counteract these effects and humans can be taught to use a speech recognition system effectively through learning to overcome these shortcomings (Nusbaum et al., 1995). Users of speech recognition must, however, be vigilant that they do not resort to "hyperarticulation", which occurs when a speaker attempts to speak more clearly in the event that the system repeatedly fails to recognise the word (Oviatt et al., 1998). The recognition system is also heavily dependent on the environment in which it is used as this environment is subject to ambient noise, as well as possible conversations that are not directed at the recognition system but which may be interpreted by the system as being so (Suhm, 2008). Consequently, the system should ideally be trained under the same conditions under which it will eventually be used (Nusbaum et al., 1995).

For this reason, training of the speech profiles will take place in the same venue, ideally under the same conditions under which the tests will be conducted. Participants will be observed during their test sessions to establish whether they learn to adapt and master the use of speech recognition.

2.7.4 Speech-enhanced user interfaces

> "How do people want to talk to their computers – and do they want to talk to them at all?" (Berg, Gröber & Weicht, 2010, p .19)

The proliferation of computers in everyday life has ensured that most people use a computer on a daily basis and for this they depend heavily on the standard keyboard for text input (Feng & Sears, 2004). Speech technology is an exciting concept which provides an ideal input device for significantly reducing the large amount of typing that must be performed. Although typing is an integral part of computer use, it requires high levels of practice and many users are not able to achieve high typing speeds even with prolonged use of a keyboard (Feng & Sears, 2004). However, speaking is an innate ability and most individuals are capable of a high average rate of spoken words per minute. The high incidence of such afflictions like tendinitis, carpal tunnel syndrome and repetitive strain injuries also provide ample motivation to reduce typing requirements and device manipulation (Klarlund, 2003). Considering the possibility of high speed recognition when using continuous speech recognition provides ample motivation to incorporate ASR as a standard input method in typing-intensive environments, such as a word processor or text editor. Taking into account that an experienced typist can reach average speeds of 68 words per minute (Logan & Crump, 2009; Liu, Crump & Logan, 2010), ASR provides a means to considerably increase efficiency.

However, a limitation of speech recognition is the fact that it may have to be used in an open-plan environment or public forum which could infringe on the privacy of the user (Suhm, 2008). Furthermore, a user may feel that speaking to a computer is unnatural and may find it embarrassing talking to a machine, although they may quickly become accustomed to it. This sentiment was evident in one user's response to using speech in a study conducted by Nelson (1986) - "at first it was kind of strange and almost like you were sitting there talking to yourself, but once we got used to it and I started working with it full time, it was a lot faster".

A Wizard of Oz (WOZ) experiment entails a simulation of the intended environment through having a facilitator respond to user commands as though the environment was doing so. Such an experiment was conducted in order to determine whether these limitations out-weigh the advantages, whether users were prepared to use speech as an input technique and in what context they would use it (Berg, Gröber & Weicht, 2010). Results indicated that users had a tendency to make use of the more familiar GUI interaction, except when faced with a complex task. Under these circumstances, users switched to verbal communication. User reaction was positive to the use of speech recognition and many users indicated that they would like to use it in future even though it may be unnatural or embarrassing. Thus, the heavy reliance on the GUI was seen as a direct result of familiarity with such interaction.

Results on the type of commands which were issued were inconclusive, with some users resorting to issuing commands based on menu wording and others using more task-oriented commands by translating the task they wished to complete into complex instructions (Berg et al., 2010). Therefore, since it is unclear whether grammars should be structured in a menu-orientated or task-orientated manner, the current study will resort to a menu-oriented grammar in an attempt to lessen the learning curve and memorisation required.

The use of speech recognition as an input method in popular graphics editors was found to be feasible by Yankelovich (2008). However, this implementation did not incorporate actual speech recognition but used a WOZ experiment where a facilitator took over control of the mouse. This setup should minimise errors since the facilitator is aware of the intentions of the user and the actions the user is currently busy with. Furthermore, interpretation is easier since the "speech engine" has human interpretive capabilities and does not rely on a fixed grammar or require training in order to be able to understand the user. Therefore, while showing that such an interface is feasible, it does not conclusively prove that speech recognition will be successful when incorporated into such an environment as the limitations associated with such technology were not present with the human facilitator.

Speech recognition has also been successfully used in a digital music retrieval system (Forlines et al., 2005) and even in the specialised domain of computer programming, which is not suitable to speech recognition in its natural state. VoiceCode allows code to be dictated in a straightforward manner, including the dictation of variable names which would not normally be recognised by a standard ASR engine (Désiltes, Fox & Norton, 2006).

Early speech recognition studies yielded promising results and highlighted the advantages of using speech recognition. A 96.8% accuracy rate and 17.5% reduction in completion time was achieved for speech recognition commands in a simulated military command and control application (Poock, 1982). Speech recognition significantly decreased the error rate in an airline baggage handling system (Nye, 1982) and in a language-directed program editor (Leggett & Williams, 1984). Conversely, studies have shown that dictation does not achieve the same rate of words per minute as when speaking in a natural forum, mainly due to the high incidence of recognition errors, and the difficulties experienced in correcting these via voice commands, although dictation speed does increase with the experience level of the user (Feng & Sears, 2004).

The positive results achieved in some studies urge exploration into using speech recognition in everyday applications, particularly as a means to increase efficiency and learnability of these applications. If the use

thereof is successful and also serves to limit the physical consequences of extended computer use, the speech recognition could meet the expectations of future generation of user interfaces. Therefore, the current study will investigate the possibility of using such a technology within a word processor.

2.7.5 Speech-enhanced word processing

Although ASR has been used in applications such as form filling and personal digital assistant (PDA) applications, the situation is much more complex in an editing environment as there is no natural language to express editing commands (Klarlund, 2003). It has therefore been suggested that natural language is unsuitable and inefficient for use in an editing application (Klarlund, 2003). For example, in order to insert three exclamation marks one would have to issue commands such as “insert three exclamation marks” or “exclamation mark, exclamation mark, exclamation mark”; this requires far more time and effort than simply typing the three exclamation marks in successive key strokes (Klarlund, 2003). However, by making use of symbolisation, it may be possible to lighten the load on the user and make natural language commands more intuitive and more efficient than keyboard input (Klarlund, 2003). To test this premise, ShortTalk uses symbolised editing concepts that can be concatenated into phrases. For example, ShortTalk uses such commands as “Goop” for “go up”, “Loon” for “New line” and “Go aift hello” for “place the cursor to the left of the occurrence of hello”. A counting system is also provided through the commands “Ain, Twain, Traio, Fairn, ...”. The belief is that learning this symbolisation language will be easier than learning an editing language, as evidenced by the human ability to string letters into words and words into sentences (Klarlund, 2003). The afore-mentioned symbolisation system seems to require much memorisation and an unnatural way of expressing yourself and it seems doubtful that many people would go to the lengths required to learn the symbolisation system. Therefore, such a method will not be considered for the current study.

In a study conducted by Karl et al. (1993), four word processing tasks were identified that were deemed suitable for voice commands. Sixteen users, fifteen of whom were new to speech recognition, performed four simple word processing tasks using both speech recognition to issue commands and traditional direct manipulation. The first task could be completed using voice commands or the mouse – no typing was required. Users were given an unformatted document and required to reformat using predefined styles of “bullet”, “figure”, “figure-label”, “text”, and two section header styles “level one” and “level two”. The mouse was used to select the appropriate text and then the user either had to issue a verbal command or use the mouse to navigate to the correct menu item depending on which input group they were part of. The second task required the users to type a formula containing subscripted and superscripted letters, bold text and Greek letters. The ratio of keystrokes to uttered commands was measured at 1.65 to 1. The third task was to create a table of symbols using only copy, paste, up and down commands. The ratio of minimum required keystrokes to minimum required voice commands was 5.57 to 1. The fourth task required subjects to type a short paragraph which contained such word processing elements as bold and italicised text, subscripted and superscripted letters. It was compulsory to type the paragraph from left to right and activate the commands as they were needed. The keystroke to command ratio was 12.4 to 1. Performance time was reduced by 18.6% when using speech recognition and was significantly faster than direct manipulation. While error rates remained the same for both input groups, significantly more memorisation errors were made when using speech recognition. Users were enthusiastic about using speech recognition for command activation but expressed hesitation over concerns for region accuracy, background noise, inadequate feedback and slow response time. Negative comments about the speech technology included the low reliability of command recognition as well as the possibility of inadvertently inserting unwanted text into your document when engaging in conversation separate from your task at hand. These findings would seem to support the call for a function to switch the speech recognition off in a quick and easy manner.

The lack of feedback also caused some concern amongst the participants (Karl et al., 1993). The commands used in the study of Karl et al. (1993) largely corresponded with word processing terminology. Clearly, the commands are more efficient than direct manipulation in terms of the number of actions and completion time of tasks. A drawback of this study was that navigation was not accommodated through speech recognition and for those purposes the mouse still had to be used. The fact that this allowed text to be selected and a voice command to be immediately issued could be the reason for the faster completion time as use of the mouse only would require the users to move the mouse from the selection to the required icon or menu item. It would have been more beneficial to require text selection through the use of speech recognition as well and then compare the time required to complete the task. As reasoned in a prior section and in keeping with the study of Karl et al. (1993), the current study will also compile a grammar which resembles standard word processing terminology with which the participants should be familiar. Cursor control and navigation will also be provided for in the grammar so that a completely hands-free environment can be created and the usability thereof tested. The means through which cursor control can be used will be discussed in more detail in the following section. However, the current study will not allow participants to mix the modalities so that the test conditions can be controlled for and comparisons between the interaction techniques can be made. An additional drawback of this study was that participants were only tested once therefore the learning curve of the speech commands could not be established. The current study will compare measures across a period of time, thereby allowing the learning curve of the proposed system to be determined.

Children using word processors perform text entry faster when using speech recognition as opposed to keyboard or mouse entry or handwriting recognition (Read, MacFarlane & Casey, 2001). However, text entry via speech recognition also has the highest error rate of these entry methods (Read et al., 2001). The implication of this study is that, for children, the learning curve for text entry via speech recognition is not as steep as the traditional methods. Since it can be assumed that children would not have the keyboard proficiency of more mature users it can be stated that the younger users were still experiencing learning for keyboard typing. The fact that they could achieve faster speeds with speech recognition indicates that the naturalness with which this interaction technique can be adopted is swifter than learning to type fast on a keyboard. The higher error rates could decrease over time as the language skills of the children improve and the speech engine becomes more accustomed to their individual profiles. In that case it is imperative that a satisfactory method of error correction be found to negate the consequences of the errors.

In conclusion, in terms of word processing speech recognition appears to be a viable option for both younger and more mature users although some shortcomings (i.e. navigation and memorisation) do need to be overcome. Solutions to these obstacles have been proposed and the current study will attempt to test empirically whether these solutions are viable.

2.7.6 Using speech recognition to control the cursor

One of the most fundamental usability problems associated with ASR in text editing is the need to correct, manipulate and format text after it has been dictated (Vergo, 1998). This is an important consideration since (i) people seldom dictate grammatically correct, well-organised text, (ii) ASR is not 100% accurate so there are errors in the dictated text and (iii) most users prefer to edit text after they have dictated as opposed to while they are dictating it (Oviatt et al., 2000). The previous discussion has shown that using speech recognition can increase task completion times. However, if the user is tasked with positioning the cursor and correcting text errors, then task completion with speech recognition is slower and less accurate than with keyboard input (Haller, Mutschler & Voss, 1984). Furthermore, it was found that error correction using speech recognition is somewhat problematic, particularly for novice users who often get caught in a web of corrections, whereas more experienced users revert back to keyboard use to correct errors (Shneiderman, 2000).

For the purpose of correcting text, cursor control is often needed. The main types of cursor control are defined as being either target-based or direction-based (Sears, Lin & Karimullah, 2002). Target-based cursor control receives and reacts to commands which explicitly identify a target, such as a word that must be selected (Sears et al., 2002). Conversely, direction-based cursor control involves directing the movement of the cursor in a certain direction, for example, “move the cursor three words left” (Sears et al., 2002).

An associated problem with continuous cursor movement is that the cursor may overshoot the target since the application requires time to react to a command. Multiple cursors can be used to overcome this – one to indicate the current cursor position and one to indicate the position where the cursor will stop (Karimullah & Sears, 2002). Continuous cursor movement will also probably be slower than manual manipulation of the mouse as the cursor will have to move slowly enough to be tracked by the user. Therefore, it is not an ideal solution.

Alternatively, a grid-based cursor control system has been proven to work effectively (Dai, Goldman, Sears & Lozier, 2004). In this scenario, the screen is divided into a grid with each grid square being allocated a number. Should the user wish to place the cursor over a target, they simply vocalise the number of the square in which the target is situated. The grid gets progressively smaller with each vocalisation until the target can be acquired. The grid can also be moved by issuing vocal commands. The grid based solution has an increased speed of 33% over other cursor control systems and error rates were 70% lower with large targets and 85% with smaller targets. Therefore, it would seem as though the grid offers a potentially accurate method of controlling cursor movement with voice commands.

However, the presence of the grid on an already full screen will likely hamper the performance of the user as it will infringe on the relevant information which must be displayed, particularly in terms of a word processor which already has a full task bar. Furthermore, the editing area should, as far as possible, be left free of irrelevant clutter to allow users to perform at maximum potential. The use of a grid remains a possibility if the visibility thereof can be switched on and off as needed. Perhaps when the first cursor control command is issued, the grid can be overlaid on the working area and when the navigation and editing is complete the grid can be hidden from view automatically. Even so, the fact that the grid may have to shrink to the size of a character from the relatively large area of the original document implies that a substantial amount of effort may be required to position the cursor correctly. Therefore, the grid-based navigation as proposed by Dai et al. (2004) will still have to be coupled with some other mechanism.

Another possibility of combating the problem of cursor movement is through the use of a multimodal interface, such as the case with the Human-Centric Word Processing (HCWP) system. This system combines ASR and natural language understanding with pen-based pointing and selection gestures which allow the user to issue editing commands which are sensitive to location based utterances (Oviatt et al., 2000). These locations (for example, “this”, “here” and “there”) are then interpreted using the position and nature of the pen-based gesture. Eye gaze would also be an ideal solution to this problem as the user could simply look at the required location when issuing the command and the system could interpret it accordingly. The added advantage of using eye gaze and not gestures is, once again, the provision of a completely hands-free environment.

The simultaneous use of eye gaze or gestures and speech seems to be the most promising of the cursor control methods discussed here. Since some studies (cf. Bolt, 1981; Oviatt et al., 2000) have already been conducted which investigate this type of communication and shortcomings have been identified for the other discussed methods, the current study will try a new approach which was not encountered in the literature. While standard commands to move the cursor (e.g. left, right, up and down) will be provided, more complex selection commands will also be catered for. These commands will be analogous to moving the cursor efficiently using the keyboard only. Since efficient and effective cursor movement can be achieved without having to remove one’s hands from the keyboard, providing equivalent speech commands could achieve the

same results. This study will determine whether these will be sufficient to facilitate efficient cursor movements or whether they are dependent on the knowledge of the user with regard to navigation using a keyboard.

2.8 Eye-tracking

2.8.1 Hardware

The device used to measure eye movements is known as an eye-tracker (Duchowski, 2007) where eye-tracking is the measurement of "the spatial direction (gaze and eye fixation) of where the eyes are pointing" (Dvorak, 2007, p. 283). There are two ways in which eye movement can be monitored. Firstly, the position of the eye can be determined relative to the head and secondly, the point of regard, which is the orientation of the eye in space, can be determined (Duchowski, 2007).

The study of eye gaze predates computing technology by many decades, dating back as far as 1878 (Jacob & Karn, 2003). Consequently, eye-tracking research can be compartmentalised into distinct eras, the first of which is widely accepted to have begun in 1879 when Javal recorded observations concerning the function of eye movements in reading (Rayner, 1998). This era extended till 1920, and concentrated on the discovery of basic facts concerning eye movement such as saccadic suppression and latency, and perceptual span (Rayner, 1998). The second era spanned a period of applied focus, coinciding with the behaviourist movement in experimental psychology (Rayner, 1998). From the late 1950s to the mid-70s very little research was done concerning eye movements. A resurgence of interest in the mid-70s resulted in improvement in the eye movement recording systems with the result that the measurements are easier to obtain and far more accurate (Rayner, 1998). The development of eye-tracking hardware can be subdivided into four distinct groups (Duchowski, 2007), namely:

- Scleral contact lens/search coil and electro-oculography
- Those that use video- and photo-oculography
- Eye-trackers which are analogue video-based combined with pupil/corneal reflection
- Those using digital video-based combined pupil/corneal reflection which can be augmented by computer vision techniques and digital signal processors

The use of electro-oculography requires that electrodes be placed around the eye which allow the electric potential differences of the skin to be measured (Duchowski, 2007). Scleral contact lenses are a very precise eye-tracking method which necessitates that a contact lens be placed directly on the eye (Duchowski, 2007).

The second category of eye-trackers measure distinguishable features of the eye under rotation/translation while video-based eye-trackers are capable of measuring the point of regard (Duchowski, 2007). This can be done in one of two ways, namely the head must remain stationary so that the point of regard and the eye's position relative to the head are identical or head movement must be disambiguated from eye rotation by measuring a number of ocular features, for example corneal reflection and the pupil centre (Duchowski, 2007).

When using corneal reflection, light sources, for example infrared, are shone into an individual's eyes. The technology behind the infrared eye-trackers is based on the fact that an infrared LED that is shone on the human eye causes a reflection spot that remains static (Figure 2.3) regardless of the direction the eye is looking (Drewes & Schmidt, 2007). Specifically, light falling on the curved cornea is reflected back, creating the four Purkinje images, the first of which is tracked by video-based eye-trackers (Duchowski, 2007).

The fourth category of eye-trackers are based on the same principles as video-based corneal reflection eye-trackers, but they use digital optics instead of analogue video (Duchowski, 2007). The use of digital signal

processors significantly increases the accuracy and the usability of the eye-trackers whilst simultaneously causing a decrease in the cost (Duchowski, Cournia & Murphy, 2004).

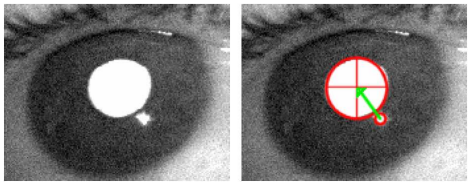


Figure 2.3: Video-based eye-tracking using the reflection of an infrared light source and the centre of the pupil to calculate the direction of the eye gaze

Source: Drewes and Schmidt (2007)

There are also gaze estimation methods which use visible light and not infrared light or which do not extract features of the eye, such as the dual- Purkinje method (for a more in-depth discussion see Hansen & Ji, 2010).

Eye-trackers can differ in terms of position of the cameras used, the illumination used as well as the type of data they produce and how the data is analysed (Holmqvist et al., in press). In broad terms they can be categorised as head mounted eye-trackers, head-trackers and static eye-trackers (Holmqvist et al., in press). The head mounted eye-tracker is worn on the head of the participant and uses a camera to record the stimulus (Holmqvist et al., in press). A head-tracker adds the functionality of tracking the position of the head to the head-mounted eye-tracker (Holmqvist et al., in press). The so-called static eye-tracker can be subdivided into a remote or tower-mounted eye-tracker (Holmqvist et al., in press). Both use illumination, such as infrared lighting, and a camera which are positioned in front of the participant. However, the tower-mounted eye-tracker requires physical contact with the participant and restrains head movement, while the remote eye-tracker requires no contact with the participant (Holmqvist et al., in press). The remote eye-tracker offers the advantage of non-invasiveness but is generally less accurate than tower-mounted and head-mounted eye-trackers (Li, Winfield & Parkhurst, 2005).

A remote video-based corneal reflection eye-tracker will be used in this study.

2.8.2 Eye-tracking applications

Applications using eye-tracking can broadly be classified as either diagnostic or interactive. Within the field of HCI, diagnostic applications are used to determine and record the eye gaze of the user for post-trial assessment and analysis (Jacob & Karn, 2003). In these instances the system is not required to react to the perceived eye gaze (Duchowski, 2002). In contrast, HCI interactive systems use eye gaze as an input modality and the system is required to respond to the eye gaze in an appropriate manner (Duchowski, 2002).

Duchowski (2002) subdivides interactive applications into either selective or gaze-contingent applications. Selective applications use eye gaze as an input device, specifically in terms of a pointing device. Gaze-contingent applications use the eye gaze information to facilitate rapid rendering of a complex display as the information in the peripheral vision and extending beyond that is degraded so as not to consume resources (Duchowski, 2002). Since information in peripheral vision is not processed it seems reasonable to suppress this information to lessen the load on the users' cognitive perception. In this way, it may be possible to provide users with a frame of reference whereby they are certain that the area they are fixating on is correctly tracked. However, the size of objects could mean that there may be multiple objects in the gaze-contingent window, which fails to give concrete affirmation to the user that the correct button or object will be manipulated, just that the general area is correct.

For the current study eye gaze will be used in an interactive, selective capacity.

2.8.3 Activation mechanisms

Eye movement-based human-computer interaction can be classified as either requiring natural or unnatural movements (Jacob & Karn, 2003). Interfaces using natural eye movements respond to eye movements which are natural to the user, such as normal scanning of an interface (Jacob & Karn, 2003). Unnatural eye movements are those that must be executed in a specific way in order to elicit reaction from the system (Jacob & Karn, 2003), such as gaze gestures. A number of different mechanisms for eye gaze activation will be discussed with some examples being given of their use in other applications.

2.8.3.1 Dwell time

> "At first, it is empowering to be able simply to look at what you want and have it happen, rather than having to look at it (as you would anyway) and then point and click it with the mouse or otherwise issue a command. Before long, though, it becomes like the Midas Touch. Everywhere you look, another command is activated; you cannot look anywhere without issuing a command. The challenge in building a useful eye tracker interface is to avoid the Midas Touch problem." (Jacob, 1991, p. 156)

The most natural method to trigger a response from a system via eye gaze is that of dwell time as it is highly intuitive and requires no training (Stampe & Reingold, 1995). Dwell time is the duration of a fixation, or the length of time that the user must continuously gaze at an object, in order to trigger a response (Jacob, 1991). A drawback of using dwell time is that it can soon escalate into the aptly named Midas touch problem (see quote above). This problem can be overcome by lengthening the dwell time required to activate commands or using a secondary action such as a mouse click or button press (Ashmore, Duchowski & Shoemaker, 2005). However, if the dwell time is too long then the speed advantages of using eye gaze are lost. Additionally, the user may become frustrated at having to maintain a stable gaze for a protracted length of time.

The ideal solution would be to react to eye gaze when appropriate but also allow the user to glance at the interface without activating commands when they so desire (Jacob, 1991). Subtle feedback can also be given, for example highlighting a button when it is about to be activated rather than simply executing the command suddenly without first giving the user some form of feedback (Jacob & Karn, 2003).

The Risø National Laboratory at Roskilde in Denmark developed a system called EyeCatcher which makes use of EyeCons. Glenstrup and Engell-Nielsen (1995) report on the use of EyeCons in their thesis as a successful solution to the Midas touch problem. These are gaze sensitive buttons that are placed next to a selectable area. Users are required to fixate on the EyeCon to activate it and by placing the EyeCon next to the selectable area the risk of the user accidentally activating the command is reduced. As an added measure, continuous feedback is given to users to enable them to judge when the activation will occur. An animation of an eye closing (Figure 2.4) is played on the EyeCon and only once the eye is closed, is the action triggered. For inexperienced users, the optimal dwell time for EyeCons was found to be 500 milliseconds while the dwell time could be set shorter for more experienced users, but was still dependent on the individual.

Figure 2.4: EyeCon animation of eye closing

Source: Glenstrup and Engell-Nielsen(1995)

No reasons were given as to why an animation of an eye blink was used for dwell time. A colour gradient or some other indicator without such strong connotations may be more usable.

Typically, dwell times range from 400-1000 milliseconds (Špakov & Miniotas, 2003) and oftentimes are accompanied by an option to change the dwell time to better suit the needs of the user. This requires that users set the dwell time according to their perceived needs. However, it is possible to continually evaluate the speed with which users' type with their eyes and adjust the dwell time accordingly (Špakov & Miniotas, 2003). This method proved popular and alleviates the need for the user to adjust the dwell time until a suitable time is found for their individual use. These finding were confirmed during experimentation with dwell time periods where it was found that for novice users the dwell time should be longer than 500 ms but that this time should be adjustable and adaptive (Hansen, Johansen, Hansen, Itoh & Mashino, 2003). Furthermore, the dwell time of a button should be dependent on the button and the amount of time required to perceive the button – for example, does it contain images, full-length text or just letters (Hansen et al., 2003). Therefore, the amount of information on the button that must be processed affects the interaction technique directly as users will require more time to process the information before a decision can be made as to whether the button must be selected. Once the users become accustomed to the interface, recognition will play a role in deciding which button is required and the dwell time can be significantly reduced. There are also various methods of implementing dwell time such as continuous or accumulated dwell time (Hansen et al., 2003).

The multimodal interface in the current study will provide dwell time capability although it will not be included as a factor in the study. Additionally, since the aim of the study is not to determine optimal dwell times for the task at hand, users will be permitted to adjust the dwell time to suit their individual needs. This will also provide better control for the users who will be able to determine for themselves how much time they need before activating the key. Since this activation method will primarily be used for typing and the keyboard buttons will contain only a single letter, the amount of time required to process the button test should be negligible. Empirical testing of the dwell time will not be done as many such studies have already been conducted.

2.8.3.2 Blinking

Buttons and other widgets can also be selected by responding to blinking. However, blinking is not necessarily the ideal solution to the Midas touch problem as blinking is not always voluntary (Ashmore et al., 2005) and the rate of blinking is also affected by the user's workload (Jacob & Karn, 2003).

As an example, Špakov (2005) showed that chess pieces can be moved by first selecting the piece and the destination square using one of three selection modes, namely dwell time, blinking and gaze gestures. Of these, dwell time seemed to be the most attractive to the users, while blinking and gaze gestures were considered to be tiring. The fact that blinking is considered to be tiring could be due to the natural reaction of the users to try and stop blinking as they are nervous that they may inadvertently activate a command. The effort required to stop blinking and the resultant eye fatigue caused by not blinking may prohibit blinking from becoming an acceptable means of communication. Although similar to any other interaction technique, practice and extended exposure could eliminate the problem. An additional consideration for using blinking is the frequency of the blinking that will be required since blinking may be useful if used in moderation.

To provide more options to the users, blinking will be incorporated as an interaction technique in the multimodal interface of the current study. However, in an effort to overcome a user's urge to suppress blinking, a pronounced blink will be required in order to activate targets. This means that both eyes must be kept closed for a protracted period of time. This should allow distinction to be made between natural blinking and a command blink. The time required for the pronounced blink should ideally be much shorter than the threshold used for activation through dwell time to facilitate faster selection speeds. Users will be allowed to change the duration for which the blink must be executed as their expertise and needs change. The testing of the blinking interaction technique is beyond the scope of the study although it may be worthwhile to test the usability and user reaction to the pronounced blink.

2.8.3.3 Look-and-shoot

When using the look-and-shoot method of eye gaze interaction (De Luca, Weiss & Drewes, 2007), the user is required to gaze at an interactive object whilst simultaneously pressing a button in order to trigger a system response (Ware & Mikaelian, 1987). Using such an activation method has the advantage that the time lost to dwell time activation is no longer applicable and the full benefit of high-speed eye movements can potentially be exploited. It can also be considered analogous to normal mouse use, where the user has to click a mouse button to elicit a response from the system. While pressing a keyboard key requires some form of automotive control, look-and-shoot can use any form of trigger, such as blowing in a pipe or issuing a speech command.

Although look-and-shoot (using the Enter key) will be available for use in the multimodal interface of this study, the use thereof will not be included in the user testing. The proposed interaction technique of eye gaze and speech is similar to look-and-shoot and will be discussed in more detail in a following section.

2.8.3.4 Gestures

Analogous to mouse gestures, the use of eye gestures have been suggested as a means of communication with the interface via eye gaze. In terms of applying this concept to eye gaze, gestures require users to perform a predefined set of eye movements that can be interpreted by the system as a command being issued. The use of gestures is a fairly novel idea in terms of eye gaze but it has already been implemented with some success both for numeric (cf. De Luca et al., 2007) and alphabetic input (cf. Huckauf & Urbina, 2008; Porta & Turina, 2008; Wobbrock, Rubinstein, Sawyer & Duchowski, 2008). Gestures are not ambiguous and there is less risk that a command be accidently executed when using a gesture, as opposed to dwell time and blinking.

Gaze gestures have been used to input numerical PINs and were found to be more robust against erroneous input although they did require more time than a standard keypad entry (De Luca et al., 2007). Such an input technique could increase the security of PINs as gaze gestures will be difficult to detect by an observer as no visual feedback will be given. Forty percent of participants stated that they would be able to use mnemonic aids, such as a shape traced on a numeric keypad, to remember their PIN gaze gestures (De Luca et al., 2007). Unfortunately, a PIN contains far fewer digits than the alphabet and it remains to be seen whether the same mnemonic benefit can be gained from alphabetic input. EyeWrite (Figure 2.5) is an example of an application which makes use of alphabetic gaze gestures. A gaze sensitive application interprets gaze gestures as alphabetical characters and then relays these to any standard Windows text editor (Wobbrock et al., 2008).

Users of EyeWrite did require some time to become accustomed to the use of the system and while they could not attain the same input speeds as when using on onscreen keyboard, there were significantly fewer uncorrected errors in the transcribed text (Wobbrock et al., 2008). Users also preferred using EyeWrite to an onscreen keyboard as they perceived it to be faster, less error-prone and less fatiguing on the eyes (Wobbrock et al., 2008).

Keyboard shortcuts can significantly reduce the time required to complete a task. Whilst observing a number of users during a document handling task, it was noted that the majority of users remove their hands from the keyboard to save the document using the toolbar shortcut (Drewes & Schmidt, 2007). This presented an ideal situation to remove the added effort of using the mouse to invoke a command by making a gaze gesture available. Subjects were immediately able to execute the gaze gesture to save the document but indicated that they would prefer to use the keyboard shortcut as an alternative to using the mouse. When closing a dialog box using gaze gestures, users were able to perform this task in the same time as it would take when using a mouse. Gaze gestures were also found to be reliable regardless of the background, including such complex backgrounds as tables, text or web pages.

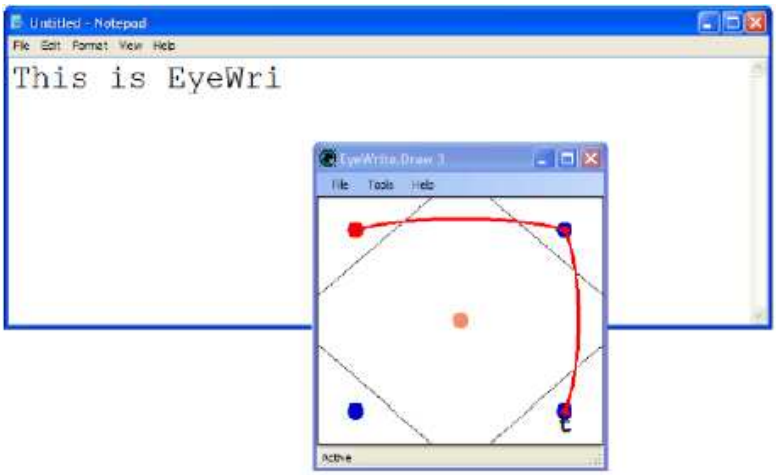


Figure 2.5: EyeWrite being used with Microsoft Notepad

Source: Wobbrock, Rubinstein, Sawyer and Duchowski (2008)

Gaze gestures have even been implemented on the limited screen space of mobile phones – a novel idea which users found attractive (Drewes, De Luca & Schmidt, 2007). A promising finding for the adoption of gaze gestures is that users have displayed an uncanny ability to switch between natural eye movements and eye movements that they surmised would elicit a response from the system (Hyrskykari, Majaranta & Räihä, 2003). When using iDict, a system developed by Hyrskykari et al. (2003), the system automatically determines if the reader is having difficulty comprehending a part of the text based on their reading pattern. Users quickly realised, without prompting, that by manipulating their eye gaze they could "force" the system to provide assistance (Hyrskykari et al., 2003).

The recent interest in gaze gestures has yielded some very encouraging results both in terms of user acceptance and objective usability measures. Since users have been able to adopt the use of gaze gestures with relative ease, gestures have become an exciting prospect albeit one which still requires in-depth investigation. Gaze gestures were considered as an activation mechanism in this study but due to time constraints they were eventually not included in the multimodal interface.

2.8.3.5 Pupil size

As most eye-tracking devices automatically measure and record pupil size during interaction, it could be suggested that this is an alternative means to influence interaction. It has been found that subjects are able to voluntarily control their pupil size (Ekman, Poikola, Mäkäräinen, Takala & Hämäläinen, 2008). However, it remains to be seen if this is a viable interaction method as the type of task currently being completed might also influence the pupil size of users. Pupil size was not included in the study as an interaction technique due to the potential inaccuracy of measuring pupil size with an eye-tracker as well as the learning which will be required for users to control their pupil size.

2.8.4 Using eye gaze in user interfaces

> "… to load the visual perception channel with a motor control task seems fundamentally at odds with users' natural mental model in which the eye searches for and takes in information and the hand produces output that manipulates external objects. Other than for disabled users, who have no other alternative, using eye gaze for practical pointing does not appear to be very promising" (Zhai, Morimoto & Ihde, 1999, p. 247)

> "Human beings look with their eyes and typically, when they want to point to something, they look before they point *(citation)*. Therefore, using eye gaze as a way of pointing on a computer seems like a natural extension of our human abilities" (Kumar, Paepcke & Winograd, 2007, p. 421).

The most trivial interactive use of an eye-tracker in HCI is to substitute the mouse with the eye-tracker and use the eye gaze of the user to determine the movement of the mouse cursor and to execute clicks (Jacob & Karn, 2003). The above quotes represent the conflicting views of researchers in the field of eye gaze technology. In the ensuing discussion it will be seen that eye gaze has been successfully implemented as an input channel in diverse environments ranging from gaming and virtual reality to so-called EyePliances and text editors.

2.8.4.1 Replacement of the cursor

Naturalness is one of the main aims of the next generation interfaces (Jacob, 1993a) and without a doubt it can be said that object manipulation using a mouse is far more natural than using eye gaze (Hyrskykari, 1997). Moreover, all mouse-controlled functions, namely clicking, double-clicking, right-clicking, dragging and releasing, must be able to be executed (Su et al., 2005) – actions for which the eyes have no natural mechanism.

A seemingly inconsequential but important design aspect is whether or not to allow the mouse cursor to be slaved to the eye gaze. If eye-trackers were 100% accurate then the mouse cursor would always be stationary on the retina, thereby becoming completely imperceptible (Jacob, 1995a; Jacob, 1995b). However, few eye-trackers, if any, are capable of such precision, rendering this particular facet of the problem unresolved. It does, however, present a secondary problem in that the cursor may be slightly offset from the centre of the gaze, thereby drawing the attention of the user and causing them to look at the cursor which would in turn cause the cursor position to change accordingly creating a scenario where the user effectively chases the cursor (Jacob, 1995a; Jacob, 1995b). Furthermore, the movement of the cursor as it follows the eye would be prolific and fairly distracting to the user (Jacob, 1995a; Jacob, 1995b).

The afore-mentioned arguments, together with other reasons (cf. Jacob & Karn, 2003; Jacob, 1995a), leave many people wondering why it is necessary to explore the idea of eye gaze controlled interfaces. Firstly, it may improve utilisation of the bandwidth from user to computer which, in WIMP interfaces, is under-utilised to a large degree (Jacob, 1993a). Secondly, in order to meet the obligation of providing an equivalent experience to all users regardless of their abilities, it is necessary to develop interfaces for users with motor disabilities but who still retain full control of their oculomotor facilities. Thirdly, it may increase the speed of interaction since eye movement is faster (Hyrskykari, 1997) and consequently target acquisition may be faster than with a mouse (Oyekoya & Stentiford, 2006). Fourthly, since it is natural to look at a target before attempting to select it, one could assume that the eye was moving to and acquiring the target. Therefore, it could just as well be used for input purposes without having to require an additional manual movement which some feel could provide more natural communication with a computer.

These contradictory views leave room for exploration of the use of eye gaze and how it can best be utilised to improve the usability of user interfaces. A number of studies, some of which will be discussed in the following section, have already been conducted to provide empirical evidence to confirm or refute the theoretical opinions.

2.8.4.2 Target selection

Fitts' Law and its applicability to eye pointing was established by Ware and Mikaelian (1987) but refuted by Zhai et al. (1999) who found only low correlation between eye input and Fitts' Law and then again confirmed by Miniotas (2000) who later found high correlation with a variation of Fitts' Law. Sibert and Jacob (2000) established that the further away the target is, the greater the advantage of using eye gaze because the cost remains constant irrespective of distance. Regardless of the Fitts' law variations, Ashmore et al. (2005) insist that the underlying theory behind Fitts' Law is as applicable to eye pointing as to manual input. Therefore, mean selection time can be reduced either through expansion of the targets or by reducing the distance to said target (Ashmore et al., 2005).

Manual And Gaze Input Cascaded (MAGIC) pointing involves warping the mouse cursor to the eye gaze (Zhai et al., 1999). Consequently, when the user focuses on a target, the mouse cursor is automatically moved to that position so that very little manual movement is required. Alternatively, the mouse cursor is positioned on the boundary of the eye gaze when the user initiates movement of the mouse. The mouse pointer must then manually be moved over the target. The former method, dubbed liberal MAGIC pointing, was faster for target selection than traditional manual mouse selection, while the latter, called conservative MAGIC pointing, was slower than manual pointing. This finding was verified in another study where direct manipulation with eye gaze proved to be faster with a mouse although selection speed slows over time as a possible consequence of eye fatigue (Sibert & Jacob, 2000). The fact that the conservative MAGIC pointing was slower than manual pointing raises the question as to whether the participant looked at the mouse pointer and then back at the destination point before adjusting the mouse cursor onto the target. Since eye gaze and mouse movement are closely related it would seem highly plausible as it would be natural to glance at the mouse cursor once it appears and before mouse interaction commences. The distance threshold used to identify a new object should prevent a readjustment to the cursor position in this instance. Another explanation for the difference between liberal and conservative MAGIC pointing is the fact that more manual movement could be required for the conservative MAGIC pointing as the mouse pointer was warped to the boundary of the gaze area and not the calculated gaze position as with liberal MAGIC. The actual gaze position should be near to or on the target, thereby significantly reducing the amount of manual movement required.

In order to negate the effect that moving the mouse has on the positioning of the mouse cursor, Drewes and Schmidt (2009) used MAGIC pointing with a touch sensitive mouse. When the user touches the mouse the pointer will move to the current gaze position. Their findings were that participants preferred using gaze to position the mouse rather than moving the mouse. This type of interaction could save a substantial number of mouse movements without placing additional strain on the eyes since a user will have to look at a target before clicking it. The combination of gaze and a touch sensitive mouse offers speeds that are superior to that of a mouse (Drewes & Schmidt, 2009). The interface did not provide feedback to the user as to where the gaze position was detected. Instead, the positioning of the mouse pointer at the gaze position was regarded as the visual feedback and the onus was on the user to verify that the correct target had been acquired before it could be clicked. While this could explain the increased accuracy which was achieved in the study since there were fewer incorrect clicks, it could also have an impact on the speeds achieved. Nevertheless, this was a successful combination of the advantages of the liberal and conservative MAGIC pointing by preventing unwanted cursor movement but still positioning the cursor at the eye gaze which reduces the amount of manual movement required.

2.8.4.2.1 Using an ISO standard to assess a pointing device

The International Standards Organisation ratified a standard, ISO 9241-9, for testing the speed and accuracy of pointing devices for comparison and testing purposes. The details of this ISO standard will be discussed in depth in Chapter 3, but some results will be discussed here. The first study to test eye-tracking as an input

device using ISO 9241-9 was conducted in 2007 by Zhang and MacKenzie (2007). This test used the multi-directional tapping test across four conditions, namely (a) a dwell time of 750ms, (b) dwell time of 500ms, (c) look-and-shoot which required participants to press the Space bar to activate the target they were looking at and (d) the mouse (Zhang & MacKenzie, 2007). A head-fixed eye-tracking system with an infrared camera and a sampling rate of 30Hz was used for the study. The look-and-shoot method was the best of the three eye-tracking techniques with a throughput of 3.78 bps compared to the mouse with 4.68 bps. Throughput (defined in Chapter 3) is a measurement which incorporates both speed and accuracy of use and can be used to measure the usability of pointing devices.

The fact that the look-and-shoot method is the most efficient activation mechanism is not surprising since the selection time of a target is not dependent on a long dwell time and theoretically target *acquisition* times for all interaction techniques should be similar. The time required to press the space bar, particularly if users can keep their hand on it, should be shorter than the dwell time, which was confirmed by the results of the aforementioned study (Zhang & MacKenzie, 2007). Recommendations stemming from the study included that a dwell time of 500 ms seemed the most appropriate so as to avoid the Midas touch problem whilst simultaneously ensuring that participants did not get impatient waiting for system reaction (Zhang & MacKenzie, 2007). Increasing the width of the target reduced the number of errors made but had no effect on the throughput. Participants indicated that the high speed positioning of the eye-tracker is desirable but that it causes eye fatigue, dry eyes and discomfort. Since the eye-tracker used was head-fixed, neck and shoulder fatigue was also a source of concern for respondents.

In a comparable study, the ISO standard was used to compare four pointing devices which could serve as a substitute mouse for disabled users (Man & Wong, 2007). The four devices tested were the (i) CameraMouse, which was activated by body movements captured via a USB web cam, (ii) a Head-Array Mouse Emulator, (iii) a CrossScanner, which has a mouse-like pointer activated by a single click and an infrared switch and (iv) a Quick Glance Eye Gaze Tracker which allows cursor placement through use of eye movement (Man & Wong, 2007). Targets had a diameter of 20 pixels and the distance between the home and the target was 40 pixels. Two disabled participants, both with dyskinetic athetosis and quadriplegia, were tested over a period of eight sessions with two sessions per week. Each participant was analysed separately and it was found that the CrossScanner was suitable for both participants although the ASL Head-Array was also suitable for use by one of the participants. This study is a prime example of the difficulties associated with testing disabled users. The disabilities are often specific to the user and wide-ranging customisation will have to be provided to ensure that one interface can cater for a diverse group of disabled users. The findings cannot be generalised to any population and also cannot serve to confirm or refute many other studies. The current study aims to provide a highly customisable multimodal interface which will allow a number of different interaction techniques to be used according to the preference and capabilities of users. However, due to the intricacies involved, disabled users will not be tested at this stage but deferred until after the usability of the multimodal interface has been established for able bodied participants.

Since the use of the ISO standard for testing eye gaze has been established it will be used to compare the various selection techniques which will be discussed in the next section.

2.8.4.2.2 Increasing accuracy

An additional problem of gaze based interfaces is the size of targets that are required (Drewes & Schmidt, 2009). For example, suppose the resolution of the screen is 1024×768 pixels and that the user sits, on average, 60 centimetres from the screen, then the minimum size of the targets will have to be approximately 31 pixels which is larger than the standard widgets in current windowed environments, which are generally 24×24 pixels in size. Therefore, most standard GUI elements are less than one degree visual angle in size. The ribbon concept which has recently been adopted by Microsoft to replace standard menus may offer some hope as the

majority of the icons are now much larger than in previous toolbars. However, they are closely spaced and may still present a challenge to select accurately with eye gaze. Additionally, the most common tasks in Microsoft Word such as justification and formatting still use the smaller icons. However, the common tasks all have keyboard shortcuts and can also easily be accommodated in a speech grammar, which was done in this study.

The most natural response to this problem would be to increase the size of onscreen targets. This, however, creates the problem that much more screen real estate is used for onscreen widgets – leaving less room for the working area of the user. These designed interfaces are often viewed as unnatural and are rarely used by users other than the disabled (Špakov & Miniotas, 2005). 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 et al., 2005). These include expansion of the target, expanding or zooming into the entire display uniformly or expanding a portion of the display through the use of a fisheye lens (Ashmore et al., 2005). The following sections will discuss these suggestions, a number of which will be tested in a comparative investigation during the current study. A comparison will also be drawn with the mouse as a pointing device. Additionally, some of these proposals will be included in the multimodal interface for the word processor.

2.8.4.2.2.1 Expansion and magnification of targets

Expansion of the targets can be either visible or invisible, 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. To illustrate a concrete example (Figure 2.6), Miniotas, Špakov and MacKenzie (2004) investigated target selection through invisible target expansion. In the experiment, the target was displayed to the user as a 20×20 pixel square icon (dark solid square) when in reality the selection area was a 120×120 pixel area (demarcated with a dotted line).

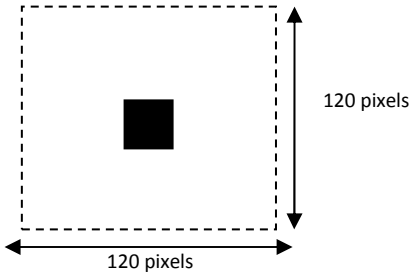


Figure 2.6: Invisible expansion of targets

Target selection was found to be both faster and more accurate with invisible target expansion. Visual feedback was provided by highlighting the target when selection was acquired. As the spatial cost of such a design is permanent (Miniotas et al., 2004), it has limited applicability in standard interface elements where the expanded target area will overlap adjacent interactive targets, such as on a toolbar. Additionally, using invisible target expansion does not reduce the amount of screen real estate that is required. It may however, have the advantage that users will have more confidence as the visually small targets will be easier to acquire and select.

A gravitational well is similar to invisible expansion of the target as the cursor is automatically pulled towards the target once the cursor is close enough to the target (Keates et al., 2002). Testing of a gravitational well indicates that users noticed the presence thereof and started to rely on this feature, thus freeing themselves

to concentrate more on moving the cursor to the desired target (Keates et al., 2002). Additionally, the gravitational well substantially improved the selecting prowess of the users (Keates et al., 2002).

Invisible expansion of the target, specifically in the form of a gravitational well, will be used in this study to facilitate a larger selection area of the target without having to visibly increase the size of the targets. This method was chosen since it may serve the dual purpose of making target selection easier and simultaneously serve to boost the confidence and satisfaction of the user who will be able to select smaller targets with relative ease. Moreover, interface elements will not be oversized, adding to the aesthetic appeal of the interface.

An alternative to static expansion is dynamic expansion of the targets whereby the target is physically expanded to a more click-friendly size when users indicate their intention to interact with the target (Miniotas & Špakov, 2004). Ashmore et al. (2005) compared an omnipresent fisheye lens, a fisheye lens which appears only after a fixation is detected (a MAGIC fisheye lens), a Grab-and-Hold (GHA) fisheye lens, which works much the same as the MAGIC lens but is fixed in place after fixation onset, and no lens. A fisheye lens shows the area in proximity to the gaze in great detail, usually through magnification, while areas in the periphery are degraded systematically (Furnas, 1986). The GHA lens counteracts the effect of jitter since it remains fixed once it is activated and allows the user to look around the magnified area without the lens moving. Using dwell time it was found that the GHA and MAGIC lenses led to significantly faster selection times than when selection is completed with no lens or with the omnipresent lens (Ashmore et al., 2005). When measuring accuracy of selection it was found that the MAGIC and omnipresent lens improved selection accuracy. Additionally, the use of a grab-and-hold algorithm when used in conjunction with invisible expansion resulted in a significant decrease in errors (Miniotas et al., 2004). Unfortunately, no comparison was made to a manual selection technique so while the MAGIC and GHA lenses were superior to the other two, their performance with manual selection techniques cannot be verified. The current study will provide a more comprehensive comparison of selection techniques by requiring selection with manual and the proposed multimodal interaction technique.

Miniotas and Špakov (2004) express scepticism as to the viability of dynamic expansion due to a number of considerations. Firstly, eye movement alternates between fixations and saccades, the latter of which are mostly ballistic in nature and during which all visual perception is suppressed (Miniotas & Špakov, 2004). Therefore, it would be wasteful to expand targets during saccadic motion. Secondly, the presence of fixational eye movements means that the eyes are never really completely at rest which Miniotas and Špakov (2004) take to mean that a visually expanding target would distract a user. Lastly, the eye gaze serves to replace the cursor which means that cursor movement cannot be tracked in order to determine the area that must be expanded (Miniotas & Špakov, 2004).

The above-mentioned reasoning of Miniotas and Špakov (2004) arguably has merit but not enough to summarily dismiss the notion of dynamic expansion; rather they should be used to adapt dynamic expansion accordingly. For instance, replacement of the cursor with the eye gaze simply means that the area situated around the current eye gaze position will be expanded. There is also no need for expansion during saccadic motion; therefore the detection of a fixation could trigger the target expansion. The drawback of such a method is that the expansion of an area may cause its position to move relative to its actual position on the screen. This could be disruptive to users as they would then have to readjust their gaze to suit the positioning of the expanded area. Therefore, careful consideration must be given to whether expansion should only be triggered after a fixation is detected. Furthermore, smoothing and stabilisation algorithms can be used to avoid jumpy visual feedback. All of these factors can be negated by expanding not only a single target at a time, but rather a realistic area, for example a 200×200 (≈6.5°) pixel area as suggested, around the current fixation point, after which the user will be required to confirm the selection by fixating on one of the expanded targets. This may require more effort as a result of the repositioning of the eye gaze to acquire the expanded area.

Taking the cautionary advice of Miniotas and Špakov (2004) as well as the above-mentioned proposed solutions into account, dynamic expansion of the interface was built into the multimodal interface of the current study. A third-party application, capable of magnifying the area under the mouse cursor was used for this purpose. The magnification tool can be switched on or off as the needs of the user change, but when switched on it will be omnipresent until it is switched off. The magnification will be slaved to the eye gaze of the user, therefore wherever the eye gaze is currently detected, the area directly below that will be magnified whether the gaze is stationary or not. This will circumvent the need for readjustments to the eye gaze on the magnified area as the magnification will always be slaved to the eye movement and the user will not be aware of the spatial shift in relation to the non-magnified area. A smoothing and stabilisation algorithm (section 3.3.5) will be used to avoid jerky movement of the magnified area. The tool used for magnification provides an adjustable zooming factor and display window which provides further customisation for the user who can then adjust the zooming factor as required. The default zoom factor is set to enlarge the area to double its actual size within a 400×300 window. In order to preserve the naturalness of the interface and the ease with which a user can read the contents of a document in a word processor, a fisheye lens which degrades the information in the periphery, was not considered for inclusion in the interface.

2.8.4.2.2.2 Zooming the entire display

Another means of increasing accuracy of eye gaze pointing is to magnify the entire display in response to the position of the eye gaze. A drawback of this approach is that contextual information beyond the zoom region is lost (Ashmore et al., 2005). Positive results have, however, been achieved for target acquisition in a word processing application as well as a web browser using eye gaze and entire display zooming (Bates & Istance, 2002). Zooming of the entire display will not be used in this study.

2.8.4.2.2.3 Applicability to the current study

The method of text entry which will be proposed for this study entails that eye gaze be used in the capacity of a pointing device. Furthermore, one of the integral interaction means within a word processor is the selection of a target, such as an icon or a menu item. Therefore, it is imperative that the proposed interaction technique of eye gaze and speech be investigated for its usability as a pointing device and as a technique to select targets.

A number of methods with which the accuracy of pointing and selecting could be increased, were discussed in the previous sections. To facilitate accurate pointing and clicking, the nature of eye gaze requires that screen widgets often have to be larger than the standard, but the resultant loss of screen real estate and oversized targets could make the interface undesirable for users. Therefore, other techniques such as zooming of targets and magnification are suggested.

For the purposes of the current study, a gravitational well will be employed in order to increase the selectable size of the targets without causing a visible increase in target size. While this should increase the accuracy of eye gaze as a pointing device, it will also boost the confidence of the users as they will feel that they are mastering eye pointing without added enhancements to the interface. Omnipresent magnification, with a smoothing and stabilisation algorithm, will be used to overcome the limitations and problems associated with zooming of the interface.

As far as could be ascertained, there is no study where a direct comparison between invisible and visible expansion of the target and magnification was included. Therefore, the current study will address this shortcoming by comparing the differences in usability with and without target expansion and magnification as well as the differences between these methods. In order to overcome identified shortcomings in previous

studies, these methods will also be compared to the mouse as a pointing device. In order to control for external variables as much as possible, this testing will not be conducted in the word processor but will rather make use of the ISO tests as previously discussed. The complete experimental design will be discussed in section 3.4.2.2.

2.8.5 Gaze-based user interfaces in practice

As previously discussed, gaze can be used in user interfaces in two ways, namely for selective or gaze-controlled interfaces and for gaze-contingent interfaces. A gaze-controlled system typically facilitates the use of eye gaze as a selection technique, while a gaze-contingent interface responds to the eye gaze by providing informative information at the point of the gaze but degrading the information at the periphery of the gaze in some way (Duchowski, 2002; Rayner, 1998). The following sections will provide an overview of some of the gaze-based applications which are available, both in general applications and more specifically for text entry.

2.8.5.1 Eye typing

Eye typing is a means whereby text is output using eye gaze as an input (Majaranta, MacKenzie, Aula & Räihä, 2006). According to Majaranta & Räihä (2007) text entry methods can be classified into distinct groups based on the input technique which is used. The first category uses direct gaze pointing, where the user has to "press" keys on an onscreen keyboard. The second category utilises eye switches, the third requires discrete, consecutive gaze gestures and the final group continuous gaze gestures.

An example of an eye typing application which requires direct gaze pointing is GazeTalk, which also employs a smaller keyboard which predicts the six most likely letters to be needed based on the previously typed letters (Hansen, Hansen & Johansen, 2001). GazeTalk uses dwell time as a selection method. A novel direct gaze method, called context switching, was developed to overcome the Midas Touch problem associated with eye typing (Morimoto & Amir, 2010). Context switching makes use of two keyboards and requires two eye movements in order to achieve key-focus and key-selection. Key-focus is achieved through the use of a short dwell time and then in order to type the focused key, that is to activate key-selection, the context must be switched. This implies that the user must perform a saccade into the other keyboard or context. Context switching resulted in faster typing speeds than simple dwell time with an average typing speed of 12 words per minute for participants after eight sessions with the context switching system (Morimoto & Amir, 2010). A major drawback of this type of typing system is of course, the space that is used in order to display two keyboards in a single work area.

Blinking, winks or coarse eye movements can be used as eye switches in eye typing systems. Examples of such systems are the "eye-switch controlled communication aids" (Ten Kate et al., 1979) and I4Control (Fejtová et al., 2004). EyeWrite (Wobbrock et al., 2008), as discussed in section 2.8.3.4 is an example of a gesture-based eye typing system which uses discrete gestures.

The gaze-enabled Quikwriting application uses continuous gaze gestures. All characters are placed within an inner resting area but grouped in such a way that the location indicates the gesture required (Bee & André, 2008). The cursor must be moved to successive sections in order to indicate which character is to be typed. When the user returns their gaze to the centre resting area, it signifies the end of the gesture and the associated character is typed.

The problems associated with eye typing are threefold. Firstly, eye typing using an eye-tracker generally makes use of a full-size onscreen keyboard which covers a large part of the screen. The impracticality of this will perhaps ensure that eye typing will never become a mainstream activity for computer users (Isokoski, 2000).

Secondly, eye typing is fundamentally incapable of achieving the same speeds as keyboard typing which can process parallel key presses at a high speed, such as in the case of a touch typist (Stampe & Reingold, 1995; Isokoski, 2000). For example, if dwell time is used, the speed of typing is restricted by the dwell time (Majaranta et al., 2006), which will probably never match the speeds that can be attained using a keyboard. Thirdly, eye typing compels the user to look at the keyboard, thus preventing them from looking at the text and typing simultaneously which can easily be achieved by most typists using a keyboard input device (Isokoski, 2000).

Solutions to the first of the afore-mentioned problems could be to provide keyboards with varying key sizes, where the commonly used keys are larger than the lesser used keys (Istance et al., 1996). Alternatively, the keys can expand dynamically as they receive focus (Istance et al., 1996). Another consideration to minimise the amount of screen real estate used by the visual keyboard is to use a cluster keyboard, similar to cell phone keypads which use predictive algorithms such as T9. These cluster keyboards partition the letters of the alphabet onto several keys, where each key contains more than a single letter. The cluster keyboard can be used by selecting the key which contains the desired letter only once, regardless of its position on the key (Klarlund & Riley, 2003). When the spacebar is pressed to indicate the end of the word has been reached, the sequence of the keys is translated into a list of words which have that sequence of digits (Klarlund & Riley, 2003). If there is more than one word with that sequence then the most probable word is selected (Klarlund & Riley, 2003). This suggested solution could serve to simultaneously reduce the amount of screen real estate required as well as speed up text entry since fewer keys have to be inspected and the desired key only has to be selected once.

Off-screen targets can also be used to save screen space by spreading the letters around the sides of the screen instead of as a bulk keyboard (Isokoski, 2000). Once the user glances off the screen, the eye-tracker is still able to track it to a certain degree albeit with a much degraded signal (Isokoski, 2000). The screen can also be divided into zones and gestures can be interpreted as letters, as in the case of QuikWriting (Isokoski, 2000). The implementation methods suggested for off-screen targets require a fair amount of learning on the part of the user before they will be able to use the system, therefore they don't appear to be the most ideal solution.

Another way to save screen space is to make use of a scrollable keyboard where only a portion of the keyboard is visible at any given moment and the user must scroll the keyboard to access other letters (Špakov & Majaranta, 2008). While screen real estate is gained, the more users have to scroll and the slower their resultant typing speed becomes. Špakov and Majaranta (2008) also found that typing speeds varied from an average of 15.06 wpm for a full keyboard, 11.12 for a 2-row scrollable keyboard to 7.29 wpm for a 1-row scrollable keyboard. These speeds were achieved after 8 sessions using the scrollable keyboards. However, participants were all experienced with eye typing before the commencement of the study. The amount of experience and exposure was, however, not specified and therefore it is difficult to gauge how comparable these results will be with the current study. Another shortcoming of the Špakov and Majaranta (2008) study is that no indication is given of the learning curve which will be experienced by the general users who have not had prior experience with eye typing.

In order to speed up the typing process, it is possible to implement word completion algorithms which will eventually reduce the total amount of typing time (Isokoski, 2000). Early forays into text entry mechanisms for the disabled arranged letters in order of their frequency of use (Istance et al., 1996), which would necessitate the need to adapt to a constantly changing keyboard layout which could confuse the users. Instead of enlarging common keys as suggested previously, the keys used most regularly by the current user could be enlarged. This would be especially helpful for different languages as the frequency of characters differs between languages. Another mechanism which could minimise eye fatigue is to make the layout dynamic and rearrange the letters according to their probability of next use. Once again, this could hamper the user more as the layout of the keyboard would constantly change and the user may have to search for the required letter. It

could be expected though that the user would very quickly become familiarised with this practice and at least know where to locate a list of probable characters.

When using dwell time, the dwell time can be automatically adjusted according to the capabilities of the user (Majaranta, Ahola & Špakov, 2009). Using an adjustable dwell time, Majaranta, Ahola and Špakov (2009) found that typing speeds improved from 6.9 wpm on average to 19.9 wpm while at the same time the dwell time decreased from 876 milliseconds to 282 milliseconds. There was also a marked improvement in the error rate. These findings bode well for the learnability of eye typing using dwell time since the usability measurements all improved over time.

Based on the motivation that the majority of quadriplegics and paraplegics sustain their injuries in their late teens or early twenties and thirties, Miniotas, Špakov and Evreinov (2003) developed a system of eye typing that is founded on the Latin cursive style of writing since by then users would have learnt to write in cursive. The system, called Symbol Creator, uses the fact that a set of basic elements can be identified into which most of the characters can be decomposed into. These basic elements were then used to create a limited set of segments which can be combined into cursive letters. An eighth symbol was added to signal the end of a character. Users must piece together the segmented symbols in order to type a cursive letter. The system provides hints to the user as to which key is required to complete a character and disables those keys that cannot be used in conjunction with the previously selected key. The system also provides for the all-important feature of feedback by highlighting keys when they are activated by the eye gaze and then highlighting them in a different manner when the dwell time expires to indicate to the user that the key has been selected.

A second text entry method was also incorporated into Symbol Creator which used the idea of the cluster keyboards found on cell phones (Miniotas et al., 2003). Results indicated that the cluster and the Symbol Creator system have mean entry times of 9 words per minute with a very low error rate. Overall the users preferred using Symbol Creator to the cluster keyboard (Miniotas et al., 2003). Additionally, very little screen real estate was used and only horizontal saccades were required as all the keys were situated in a single line. The idea behind the horizontal saccades is an interesting one as it could substantially lessen the fatigue experienced by users required to manipulate the mouse pointer with their eyes. The scrollable keyboard will, to some extent, mimic the need for horizontal saccades as in some instances there is only one row of characters on the keyboard at any time. The speeds achieved with Symbol Creator are between the speeds of the 2-row and the 1-row scrollable keyboard.

Perhaps it will also be possible to reduce eye fatigue by placing the enabled symbols which can be used together at a specific position on the keyboard, as previously suggested for onscreen keyboards. The user will soon learn where the cluster of allowable keys is situated and it is only when a new character is started that a larger search will be needed to locate the start symbol. Unfortunately, the researchers did not report on the mental effort or memorisation that was required for Symbol Creator as this could have an impact on the speeds achieved. Even though older users may have knowledge of cursive writing, the segmentation process is not a natural means of writing and may require some cognitive processing. In this regard, the disabling of invalid symbols may provide invaluable assistance to the user.

To overcome the problem that users have to look at the keyboard, confirmation feedback can be given to the user so that they are aware when the character has been typed. This will reduce their need to look at the document. Eye typing speeds have been shown to be higher when both visual and auditory feedback is given with users achieving speeds of 7.55 words per minute after four sessions (Majaranta, 2009). Therefore, both these feedback mechanisms will be used in the application for the current study.

Visual feedback of eye gaze position is also an important aspect of eye typing, both to indicate selection and focus as it impacts on accuracy, typing speed, gaze behaviour and subjective satisfaction (Majaranta et al., 2006). Feedback is essential when using any type of gaze controlled interface and not only when it is used for

typing (Drewes & Schmidt, 2009). However, as previously discussed, using a gaze pointer is challenging since any inaccuracy may cause the user to chase the pointer and the pointer could also have jerky movement.

Eye typing is of paramount importance to the study at hand since typing with the multimodal interface is one of the key aspects of the study. Some of the problems which were identified in the discussed studies will be circumvented in this study through various means. The issue of wasted screen real estate remains a problem but the onscreen keyboard provided in this study will have adjustable keys, meaning that users can increase or decrease the size of the keys as it suits their needs. The size of the keyboard will fluctuate with the size of the keys thereby providing the opportunity to reduce the amount of screen space that the keyboard occupies. Whether screen space is "wasted" on an onscreen keyboard may simply be a subjective issue and not really a usability issue. Therefore, questionnaires can elicit user reaction to the onscreen keyboard after the initial use thereof and once users have become accustomed to the new layout. Users will also be allowed to toggle between different keyboard layouts which will increase the customisability of the keyboard in an effort to increase the acceptance of the proposed application. Furthermore, since users will be required to look at the keyboard in order to type, audio feedback will be given when a letter is typed – thus eliminating the need to continually look at the document for confirmation of typed letters. So as not to distract the user, visual feedback of the current selection will only be given when gaze is within the bounds of an onscreen button and will remain stable for the entire time that the gaze is detected on the button. While dwell time, blinking and look-and-shoot will be provided for selection of onscreen keys, it is the combination of eye gaze and speech which is of interest in this study. Using verbal commands is analogous to look-and-shoot which has proven to be faster than dwell time selection. Since the target acquisition times should theoretically be the same, regardless of the activation mechanism, using speech commands could prove to be faster than dwell time. Therefore, by using speech, the increased selection time caused by the dwell time is avoided, the unnatural feeling of blinking to select a key is also avoided and at the same time, the requirement that the user is able to locate and press a keyboard key is no longer required. The use of speech will also mean that the need for double-clicking and right-clicking falls away. Commands which are reached through double- or right-clicking can be provided for in the speech grammar. Whether this proposed method of interaction will be able to achieve the typing speeds possible with a keyboard will be determined though comparative user testing.

2.8.5.2 Other applications of gaze-interaction

The Gaze Enhanced UI Design (GUIDe) project in the HCI Group at Stanford University successfully designed a number of applications which did not overload the visual channel and exploited the natural use of eye gaze to facilitate everyday computing (Kumar & Winograd, 2007). These applications were EyePoint, which facilitated pointing and selecting, EyeExposé which used eye gaze to switch applications and EyeScroll which scrolls screen content based on the reading speed of the user (Kumar & Winograd, 2007).

EyePoint (Kumar, Paepcke & Winograd, 2007) provided for a left and right mouse click, a double click, mouse-over action and click-and-drag action (requires a start and end drag action) by assigning a keyboard key to each one of these. Users were expected to gaze at the link or button and simultaneously press the desired key to execute a mouse click, right-click or double-click. This then causes the immediate area around the eye gaze to be magnified at which time the user can focus on the magnified target and release the key to execute the command. Task analysis showed that EyePoint performs similarly to a mouse and can be faster than the mouse in some instances. However, users did report having to concentrate more when using eye gaze but since the tasks were strenuous and had to be completed over a short period, it is not expected to have the same effect in normal use. Participants did express a high preference for EyePoint.

The fact that EyePoint could achieve speeds faster than the mouse is undeniably very promising for the use of eye gaze as a pointing device. The additional fact that all mouse clicks could be emulated is also very positive. However, the design of EyePoint necessitates that in order to use it to its full potential the user must be able

to press (and hold) six different keys spread out on the numeric keypad. This assumes that the user has control over at least one of their limbs to such an extent that they can accurately locate and press a key. For many disabled users this is not possible and, furthermore, EyePoint also fails to provide a completely hands-free environment for users with busy-hands tasks.

A system which does provide completely hands-free interaction and which has proven to be useful for disabled users is the EagleEyes system of Gips and Olivieri (1996) (Figure 2.7). This system relies on the measurement of the electro-oculographic potential (EOG), and requires that five electrodes be positioned on the head of the user. The user can then control the cursor through movement of the eyes while keeping the head stationary or through movement of the head while keeping the position of the eyes relatively stationary or alternatively, a combination of both. A wide range of applications can be controlled through EagleEyes such as educational and entertainment software, messages can be spelled out and users can also navigate on the Internet. The EagleEyes system runs as a background application and captures eye gaze coordinates every $1/60^{th}$ of a second. These coordinates are then treated as though they were mouse coordinates and not gaze coordinates. Optionally, dwell time can be used to simulate a mouse click.

Figure 2.7: EagleEyes application in use

Source: Gips and Olivieri (1996)

The EagleEyes system was successfully developed for users with severe disabilities and can take anywhere from 15 minutes to a few months to master (Gips & Olivieri, 1996). One of the greatest advantages of EagleEyes is that it is actually a background function which is application independent which means it has the potential to be used in conjunction with any application. The small targets which are present in modern day applications will still pose difficulties for accurate selection which may limit its use in standard windowed applications. The more invasive use of electro-oculographic potential can also be a drawback as users may not relish the idea of having electrodes positioned on their face which is unnatural and not necessarily suited to use in all environments. The advancement of technology has allowed for eye-tracking systems which are far less invasive than the EagleEyes system and therefore provides much potential as an alternative means of input. The possibility of adapting EagleEyes for use with other eye-tracking technology is yet to be explored.

GazeSpace also provides a hands-free gaze based selection system to browse content spaces, such as blogs, news pages, video and image clips (Laqua, Bandara & Sasse, 2007). GazeSpace utilises the centre of the display as a main area in which information is displayed. Smaller contextual navigational elements surround this main area and users can navigate to one of these using eye gaze. Upon selecting one of the navigational elements it will be enlarged and moved into the centre area. Three different selection means were provided, including a type of accumulative dwell time where users were allowed to look at another element and then return to the original element without any of the accumulated time being lost. User preference was highest for this accumulated threshold since it facilitated faster selection. Overall user reaction to GazeSpace was very

positive. A shortcoming which can be identified in this study is that the accumulated dwell time may eventually result in the Midas touch problem unless a mechanism is provided to cancel all accumulated dwell time should the user deem it necessary.

EyeDraw (Hornof et al., 2004) is another example of a gaze-sensitive application. EyeDraw is a drawing application where the cursor is controlled by the eye movements of the user who can toggle between looking at the drawing or drawing on the canvas by enabling and disabling a gaze sensitive button. Dwell time was used as an activation mechanism and once the drawing command had been executed, auditory feedback was given. Ten fully-able children were used to test EyeDraw and the quality of the drawings produced indicated that EyeDraw can be used to draw pictures with eye gaze.

Other applications in which eye-tracking has been used to enhance human-computer interaction are RealTourist (Qvarfordt, Beymer & Zhai, 2005), virtual reality (Jacob, 1993b), a number of gaming genres (cf. Jönsson, 2005; Špakov, 2005) and EyePliances (Shell et al., 2003b).

2.8.6 Market trends of eye-tracking

> "Eye tracking technologies could transform the lives of tens of thousands of people ... The most extreme example of how this technology is used is its ability to give voice to people who are 'locked-in', people who can only move their eyes and only communicate with their gaze" (Kari-Jouko Räihä as quoted in COGAIN, 2006, p. 1).

The application of eye-tracking in the field of human-computer interaction started in the 1980s after a surge in the popularity of eye-tracking in the 1970s (Jacob & Karn, 2003). Each resurgence in eye-tracking led to more advances in both the application thereof and the hardware and technology needed for successful use of the eye-tracker (Jacob & Karn, 2003). Consequently, eye-tracking has been viewed as a promising technology for decades without ever quite living up to the expectations, similar to speech recognition. Jacob and Karn (2003) offer the assurance that for a technology to be seen as promising for so long it must be very promising otherwise it would have long since been abandoned; they caution however that there must be something holding back eye-tracking from reaching its full potential. Possible reasons for this could be technical problems, the labour-intensive data extraction and the difficulties experienced with data interpretation (Jacob & Karn, 2003). Resolutions to these problems have slowly been forthcoming which could possibly lead to the eventual adoption of eye-tracking as a mainstream usability evaluation tool and/or input device.

Currently available eye-tracking technologies are expensive and beyond the financial reach of most users. This poses the greatest obstacle in preventing widespread adoption of eye-tracking technology (Kumar, 2006). Even amongst disabled users, perhaps the group which stands to benefit the most from such technology, the use is limited to the select few who can afford the acquisition of the expensive equipment (Kumar, 2006). The development and availability of an application, dubbed a "killer application", which uses eye-tracking technology will increase demand for eye-trackers and cause a substantial reduction in prices (Kumar, 2006). The successful incorporation of eye gaze into a mainstream application could give the acceptance of such technology the boost that it needs.

In September 2004, the Communication by Gaze Interaction (COGAIN) undertook an ambitious project to establish standard control software that need not be tied to the proprietary software of eye-tracking vendors (COGAIN, 2006). In this way, COGAIN aims to make development of eye gaze software accessible to a wider audience (COGAIN, 2006). Additionally, COGAIN aims to develop a more affordable eye-tracker solution using standard webcam and ambient light (COGAIN, 2006). Should this be achieved, the software capabilities should be ready to handle the deluge of demand for software to use with newly available and affordable hardware. This places the onus on the HCI community to ensure that these hardware advances will not be in vain but that the relevant software will be in place to exploit these technologies. Therefore, studies such as the current one

are pertinent to the advancement of the software to ensure that they progress to meet the capabilities and availability of the hardware.

Apart from such aforementioned initiatives, cost effective means of providing such technology have been explored with great success. As would be expected, it is suggested that by using cheaper produced eye gaze technology, the accuracy and performance indicators of such systems would be substandard to the more state-of-the-art systems, which employ the use of high-precision cameras, eye recognition firmware and video processing systems. So it would hardly be worth the effort to produce lower cost hardware if the resulting software developed was inferior in terms of accuracy, speed and other performance indicators. However, contrary to these assumptions, it has been found that through use of a standard PC and a fairly inexpensive widely available webcam, it is possible to achieve acceptable selection performance results (Corno, Farinetti & Signorile, 2002). A reliable and inexpensive eye-tracker which is capable of using infrared light to track eye gaze in real time has already been developed (Haro, Essa & Flickner, 2000). GazeTalk is a gaze-based typing communication tool designed specifically for people with Amyotrophic Lateral Sclerosis (ALS) who are unable to communicate in any way but with their eyes (Hansen et al., 2003). The system is capable of functioning on a standard computer using commercially available digital camera technology (Hansen et al., 2003) which makes the system very attractive in terms of affordability. Furthermore, a low-cost gaze-enabled user interface was developed which achieved rates of up to 99% accuracy in locating the eyes correctly (Su et al., 2005).

The slow acceptance rate should not be disheartening as there is often a fairly long time span between invention and widespread use of devices. For example, ten years after the mouse was invented, it could only be found in a handful of research laboratories and it was only after twenty years that it was found in arenas other than research laboratories (Jacob & Karn, 2003). The acceptance of eye-tracking recently received a boost in the form of the world's first eye-controlled laptop which was released on 1 March 2011 (Tobii, 2011). In collaboration with Lenovo, Tobii has provided control for icon selection, zooming and centring of the working area. The screen is also capable of auto-dimming and brightening depending on whether the user's eyes are recognised (Tobii, 2011). Although the technology is still expensive at this stage, the fact that a fully-functioning prototype could be developed is most encouraging for eye-tracking technology.

In conclusion, the cost associated with such technologies should not be disheartening to the research community who must continually strive to find ways in which the technology can be used. This could result in a decrease in the cost of the equipment, but even in the event that this does not occur, cheaper highly accurate eye-tracking devices may be available to fill the void. The current study therefore aims at the advancement of the software applications of the technology, specifically in a multimodal capacity.

2.9 Multimodal interfaces

Previous sections discussed the shortcomings and limitations of using eye gaze and speech as isolated interaction techniques. This study proposes to combine these two in a multimodal capacity to determine whether the shortcomings of one can be compensated for through the use of the other. Previously, fears were expressed that using two error-prone interaction techniques together in a multimodal interface would result in an interface that compounded the errors but it has since been proven that the multimodal interface is in fact more robust. Some examples will be discussed in a subsequent section. Moreover, suggestions have been made to abandon the possibility of eye gaze as the only input device, but rather to use it as one modality in a multimodal interface (Hyrskykari, 1997).

Speech and gesture seem to be a very popular choice for multimodal interfaces as is evidenced in the number of applications thereof (cf. Bolt, 1980; Hauptmann, 1989; Latoschik, Fröhlich, Jung and Wachsmuth, 1998; Oviatt et al., 2000). This is perhaps because humans tend to talk with their body and in human-human

communication it is not only the speech content which plays a role in the understanding but also body language. In order to truly emulate human-human communication it may be necessary to interpret a full set of gestures, including hands, head and eye gaze, with speech.

Eye gaze systems are an attractive alternative to direct manipulation with a mouse since users naturally look at an object of interest and they are accustomed to completing other tasks while looking (Sibert & Jacob, 2000). A multimodal interface which uses eye gaze as one of the modalities will enrich user experience as it will serve to reduce ambiguous commands (Sibert & Jacob, 2000). Although it would seem a natural assumption that a computer user would look at a target before attempting to click on it, Smith, Ho, Ark and Zhai (2000) found variability in hand eye coordination during target selection using various devices, to the extent that the same individual exhibited different tendencies even when using the same input device. Coordination techniques varied between eye gaze preceding the cursor, the cursor preceding eye gaze and the eye gaze switching between the cursor and the target until the target was reached (Smith et al., 2000). Therefore, it is imperative that researchers first determine the capacity in which eye gaze can be interpreted within a multimodal interface.

2.9.1 Classification of multimodal interfaces

Coutaz and Caelen (1991) offer a taxonomy for multimodal interfaces based on the number of modalities which can be used simultaneously. An exclusive multimodal user interface offers a choice of modalities although input is obtained from one modality only. In contrast, a synergic multimodal user interface also offers various modalities but input is built from multiple modalities being used in unison. For example, when speech and gestures are used, a verbal command such as “Put that there” can be interpreted through simultaneous interpretation of gestures.

Multimodal interfaces can also be tactile, auditory or visual. Tactile interfaces require physical contact, auditory use some form of speech or sound detection while visual interfaces detect human movement in some way, for example eye gaze.

The current study will compare a synergic visual and auditory multimodal interface with a standard tactile interface where the user types using a keyboard and uses a mouse device for pointing and selecting. Options for an exclusively visual interface will also be available.

2.9.2 Implementation of multimodal interfaces

Bernhaupt et al. (2007) conducted a study with two mice and speech recognition and found that such an interface was quickly adopted as a natural means of interaction for a satellite monitoring application. However, during the course of two-handed interaction, participants often neglected to make use of speech interaction (Bernhaupt at al., 2007). Nevertheless, task solution was more efficient when using two mice than with only one and also resulted in lower cognitive load (measured as the number of fixations during task completion) on the users. However, user perception of the cognitive load experienced was not measured. A possible reason for not using speech when engaged in multi-mouse manipulation could be linked to the strain the user felt while concentrating on moving two mice. Overall, the cognitive load could be lower when using two mice but expecting users to use two mice and speech commands could be unrealistic. The study could have been enhanced by eliciting user satisfaction with the system and asking users why they neglected to use speech commands under certain conditions. The introduction of an additional modality could also be delayed to the stage when the user has completely mastered working with two mice.

A vision and gesture-based application has been offered as a possible multimodal interface as body posture, in general, and pointing are natural modalities (Wachs et al., 2011). However, these mean nothing if the system

is not aware of what is being pointed at, hence the need for combination with vision. This universal concept should be intuitive and hand gestures have been suggested as a universal language (Wachs et al., 2011). However, cultural references and context may play a huge role in the definition of a particular hand gesture and this modality may not be as universal or intuitive as suspected.

QuickSet is an application for use by the military for force "laydown" (Cohen et al., 1998). QuickSet uses a multimodal interface consisting of a pen and voice. Objects can be placed on a map by simultaneously drawing and speaking. Users of QuickSet were able to achieve substantially higher speeds than when using a traditional GUI interface (Cohen et al., 1998). Users also indicated a preference for the multimodal interface over direct manipulation (Cohen et al., 1998).

Intellectual Computer Assistant for disabled operators (ICANDO) uses head movements and speech commands while Multimodal Oral With Gesture Large display Interface (MOWGLI) uses gesture recognition and speech recognition to create a collaborative environment in which two users can work together always being aware of the other's actions (Karpov, Carbini, Ronzhin & Viallet, 2008). Using Fitts' experiments it was found that MOWGLI performed better as a pointing device than ICANDO.

Speech and gestures have also been used for gaming through the development of a gaze-aware table (Tse, Greenberg, Shen & Forlines, 2006).

2.9.3 Eye gaze and speech multimodal interfaces

Previous sections have discussed the use of eye gaze as well as speech recognition as an input modality. Insofar as can be ascertained these particular modalities are not often used in combination for multimodal text input. When used in isolation, these and other alternative modalities such as gestures are often ambiguous but when appropriately used in combination, they could result in effective interaction methods (Oviatt, 1999).

The goal of speech and vision multimodal interfaces is to emulate the ease and robustness of human communication through the integration of automatic speech recognition (ASR) and the nonverbal communication afforded by the use of eye gaze (Pireddu, 2007). In particular, the foremost aim of multimodal interaction is to integrate interaction methods to allow the advantages of one to supersede the drawbacks of another. Amongst the variety of available alternative input modalities, the combination of speech recognition and eye gaze has not gained much popularity, yet when eye gaze is used for locating objects and speech for issuing commands a fully functional system is entirely feasible (Miniotas et al., 2006). Given the inherent problems associated with target selection via eye gaze (section 2.8.4), it seems plausible that an additional modality might make selection easier and more feasible. For example, the Midas Touch problem will be minimised as two inputs will be required before the application will respond. Furthermore, the ambiguity caused by inaccurate eye-tracking could be negated if an additional modality was available to infer user intention. To date, though, there have been very few empirical studies conducted to explore this phenomenon.

Gaze as an input medium has the advantage of being a reliable indicator of the current focus of attention and since it is a natural input medium it does not require any hand-eye coordination to be learnt (Kaur et al., 2003). The eyes are expressive during conversation, which is often punctuated with gestures as well (Kaur et al., 2003). Therefore, it would seem natural to combine eye gaze and speech in a multimodal environment. However, when doing so, it is imperative that eye gaze and speech be synchronised so that intention can be inferred correctly (Kaur et al., 2003). As recently as 2009, it was said that an interface capable of improved human-computer communication through the use of eye gaze and speech is still a long way from being possible and requires further research to investigate the possibilities (Drewes & Schmidt, 2009).

Zhang, Imamiya, Go and Mao (2004) confirmed that a multimodal interface using eye gaze and speech yielded better performance than a speech-only interface. Their application responded to speech commands and used eye gaze to resolve ambiguities or to verify what the intended target was, based on its proximity to the eye gaze of the user.

When implementing a system using a combination of eye-tracking and speech recognition, Castellina, Corno and Pellegrino (2008) advocate that there are three aspects which play a simultaneous role at any given moment in the users' interaction with the system, namely the objects, the context and the commands. The objects are the widgets that are available on the screen, for example, icons, buttons or menus. The context is the area which is identified by the eye-tracker as where the eye gaze of the user is focused. This context is also referred to as the gaze window. The commands are a list of possible objects or action names within the gaze window. The use of a gaze window, which may contain one or more objects, eliminates the error created by the detection of the direction of the user's gaze. The user will utter a command name after gazing at a certain area on the screen and the application will then match the utterance to a list of commands/objects contained within the gaze window and generated as a VoiceXML grammar. The ambiguity of the uttered command will thus be eliminated. Tests indicated that the combination of the two modalities succeeded in overcoming the inherent ambiguities present in each (Castellina et al., 2008).

The nature of the speech commands and their use in the current study precludes the ability to generate a grammar based on the gaze window. During normal text input in a word processor, the user may often want to change the formatting or perform other related editing tasks. While it may be possible to change the grammar based on whether the eye gaze is within the bounds of the onscreen keyboard or not, it was decided that in order to increase the naturalness of the application and reduce the number of eye movements required, these commands could be issued regardless of where the eye gaze was at that given moment. Therefore, while the context will be established, the grammar will not be dependent on the gaze window but typing commands will be processed dependent on the gaze position.

2.9.3.1 Acquisition and spacing of targets

The Portable Interactive Command Console (PICC) is used for crisis management, manpower and equipment deployment in the field and an experimental interface included the use of gaze and speech to move objects around the interface (Kaur et al., 2003). Results indicated that the correct fixation to use for identification of the target object was the one which was acquired, on average, 630 milliseconds before the verbal command was issued. The interface of the current study will interpret the target as the one which has focus when the command is processed. Results will determine whether this intended method is sufficient for this type of multimodal interface.

To investigate the feasibility of small, closely spaced targets using speech and eye gaze combined, Miniotas et al. (2006) required participants to select a single button in a 5×5 matrix of small closely spaced buttons. All squares encompassed within the region of interest (ROI), as detected by the currently detected eye gaze, were highlighted by outlining each in a different colour (Figure 2.8). In a mixed modality trial, a participant could verbalise the colour of the desired square aloud to select it regardless of whether that square was the selected square or not. It was determined that the ideal setup is to have icons sized 30×30 pixels (≈1°) with a 10 pixel (≈0.3°) space between them (Miniotas et al., 2006). The study determined that there is high accuracy of target selection to such an extent that user performance approaches that of manual pointing.

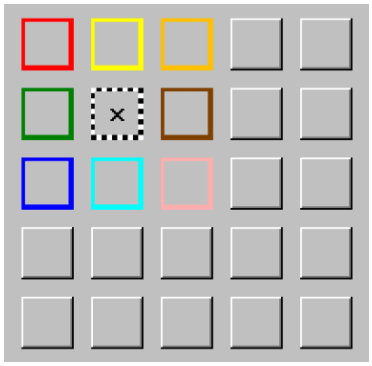


Figure 2.8: Matrix with ROI squares each outlined in a different colour

Source: Miniotas et al. (2006)

The use of speech to select a target could possibly be faster than the dwell time of 1500 ms which was used and it is suspected that accuracy will also be much higher since the correct button can be selected without the eye gaze having to be positioned on it. Therefore, in this instance it would appear that the use of the multimodal interface would be more usable than just eye gaze. The colours used must be easy to identify and vocalise and the utterances must differ enough to avoid confusion and facilitate adequate recognition. For such an application a multimodal interface seems to be a better solution than one using eye gaze in isolation. It is suspected that users would also have preferred to use speech and not only to resolve ambiguities. In terms of motivation for the current study, this study proves that despite its relative unpopularity as a multimodal interaction means, speech and gaze input can be successfully combined to create an environment which can be manipulated as accurately as with manual pointing and selection. Secondly, the optimal size identified by the study for targets will be used as a basis for the targets in the current study.

2.9.3.2 Applications

An implementation of a multimodal interface by Zhang et al. (2004) used eye gaze and speech and required users to select differently sized, shaped and coloured figures. Speech commands based on colour, colour and shape as well as colour, shape and size were available in order to select an appropriate object. The position of the gaze when commands were issued was then used to determine which object had to be selected in combination with the speech command received. The use of both eye gaze and speech was found to be more robust than using only eye gaze or speech.

The superior usability of eye gaze and speech in this instance bodes well for the acceptance of these modalities, however care must be exercised that the grammar required is not too complicated which will result in additional memorisation for the user. The grammars used in this instance may be too limited to include in a fully functioning application. Therefore, the limited number of colours which highlight objects within the current region of interest of the user seems to be a better solution when viewed in the context of a large scale application. However, both of these presume that users have a certain command of the language and are able to verbalise a wide-ranging grammar, which might not be the case. The current study will make no such assumption when it comes to using eye gaze and speech for pointing purposes.

EyeTalk is a voice and eye gaze 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 & Jenkins, 1997). 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. 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 & Jenkins, 1997).

Although EyeTalk is application independent and can potentially be used with a multitude of applications it was only tested with aviation displays and may not be suitable for use with standard windowed applications.

Kammerer et al. (2008) tested menu selection with eye gaze only and with eye gaze and speech combined. Three different menu designs were used, namely, a linear menu, a full-circle menu and a semi-circle menu. Eye gaze was used to establish the menu item that was to be selected and then either gaze or speech was used for selection purposes. Results indicated that accuracy with the linear menu was significantly lower than with the other two menus but that the input device did not affect the accuracy with which a menu item could be selected. The semi-circle menu yielded the fastest selection time of the three different kinds of menus. In terms of the interaction technique, the eye gaze and speech interaction techniques had a significantly longer selection time than eye gaze only (Kammerer et al., 2008). These findings are contrary to what would be expected for selection time but perhaps not for accuracy. Since the menu items were identical for both the interaction techniques, the accuracy with which the correct menu item could be acquired when using only eye gaze and when using eye gaze and speech should be the same. The time to acquisition can be assumed to be the same for the different interaction techniques. Since selection with the dwell time is faster, it implies that when using speech, the time required to issue the command, synthesise it and react to it, is significantly longer than 750 ms.

EyeCook is an attentive multimodal cookbook using eye gaze and speech which changes the display based on whether the user's attention is focused on the book or not (Shell, Bradbury, Knowles, Dickie & Vertegaal, 2003a). If the user is looking at the cookbook then the recipe is shown on one page, otherwise the recipe is broken into multiple cards with enlarged text. Speech can then be used to issue verbal commands which are context sensitive based on the position of the eye gaze at the moment when the command is issued (Shell et al., 2003a).

Eye gaze and speech have also been effectively combined to design a user interface with a 360° panoramic view (Stiefelhagen & Yang, 1997) for implicit and explicit command activation in aircraft (Schnell, 2000), attentive television and AuraLamp, an eye gaze and speech enhanced lava lamp (Shell et al., 2003b).

These studies highlight the potential uses of eye gaze and speech for a multimodal interface and encourage researchers to continue to find other uses for these modalities. Specifically, eye gaze and speech will be used for text entry and the next section will discuss some of the text entry applications which employ the use of these modalities.

2.9.4 Text and data entry using eye gaze and speech

In terms of data entry, eye gaze and speech recognition have been implemented, with great success, to complete a television licence application form in the United Kingdom (Tan, Sherkat & Allen, 2003a). The edit box which the user is looking at receives focus and then dictation can be used to complete the forms. This method was compared to the mouse and keyboard, handwriting and speech only. Even though eye gaze and speech was neither the fastest nor the most accurate, it was the most preferred method of data entry (Tan et al., 2003a). This could be attributed to the naturalness of completing a form in this manner. Another possibility which could be investigated is the use of eye gaze to set the focus and the keyboard to enter data.

Another means of data entry is the RESER and SPELLER systems (Tan, Sherkat & Allen, 2003b). The keyboards used in these systems are cluster keyboards and users are required to look at the relevant key on the keyboard and then speak the letter that they wish to type. The RESER system will attempt to recognise the word and offer a suggestion once it can recognise the word that is being typed. The user must then give confirmation as to whether or not that was the intended word. The SPELLER system, on the other hand, requires users to spell out the entire word. Visual feedback to indicate focus is through highlighting the button on the keyboard. For

text entry, users preferred the mouse and the keyboard while speech and eye gaze was the preferred means of data recovery.

The fact that a button must receive focus significantly reduces the size of the potential vocabulary which must be recognised. It is surmised that the accuracy rate of such a method would be much higher than if a full-length vocabulary was present at all times. In addition, the use of fewer buttons but with more letters on reduces the amount of screen real estate which is required by the onscreen keyboard thus lessening one of the associated disadvantages of eye gaze as an input device. The use of a grammar comprising alphabetic characters and the fact that a visual cue is also available reduces the amount of learning and memorisation that is required by the user. The use of multiple modalities should also increase the accuracy for text entry than simply using speech recognition. A combination of the colour scheme used by Miniotas et al. (2006) for non-alphabetic characters with this text entry method could provide an all-encompassing interface for eye gaze and speech. Of course, this assumes that the user has a wide ranging vocabulary and is capable of all the speech utterances which is not always the case.

Dasher is a text entry interface which uses continuous pointing gestures to facilitate text entry (Ward, Blackwell & MacKay, 2000). All letters start on the right hand side of the screen and a user must point at the desired character to cause the area around that character to grow larger. The character then also starts moving towards the left side of the screen and once it crosses the centre of the screen it is accepted for text entry. The size of the letters is also adjusted according to the probability of the letter being selected next in order to speed up typing. When using a mouse as the pointing device, users were able to achieve typing speeds of 34 words per minute compared to traditional keyboard input speeds of 40-60 words per minute. Dasher has since been modified to use eye gaze as a pointing device for a hands free environment (Tuisku, Majaranta, Isokoski & Räihä, 2008). During the first session with the modified version, typing speed was 2.5 WPM while after the tenth session of working with Dasher, users were able to type an average of 17.3 WPM which indicates that learning is required to achieve acceptable speeds with the application. The result is nevertheless a promising one as it offers an intuitive hands-free means of text entry for a variety of users. Comparison with a mouse showed significantly slower entry rates with eye gaze than with the mouse but no significant difference in error rate was detected between the two pointing devices (Tuisku et al., 2008). Only one session was completed with the mouse where participants achieved an entry rate of 20.69 WPM which was only slightly higher than that of eye gaze after ten sessions. However, it cannot be said that eye gaze and mouse input are comparable in this instance since the previous Dasher study showed that speeds of 34 WPM were possible after extended practice with the mouse. It can, however, be concluded that eye gaze may require more practice than the mouse but whether the speeds will be comparable once both modalities reach a plateau, remains to be seen.

Dasher has been proven to respond well to control via a brain-computer interface (Felton, Lewis, Wills, Radwin & Williams, 2007) and the aptly named Speech Dasher extends the capabilities of Dasher even further by including speech recognition as well (Vertanen & MacKay, 2010). Speech Dasher uses the same selection technique as the original Dasher but allows the user to zoom through entire words. The word set is obtained through speech recognition where the user speaks the text they would like to enter. With an error recognition rate of 22%, users were able to achieve typing speeds of 40 WPM (Vertanen & MacKay, 2010) which is similar to keyboard text entry. Speech Dasher is an example of a multimodal interface where gaze is used to enhance the capabilities of speech recognition. In the current study, eye gaze and speech will be used simultaneously in such a manner that the disadvantages of one are counteracted by the other.

The current study will build on the idea that eye gaze will be used to establish which keyboard button is required by the user. However, instead of relying on the inaccurate or time-consuming methods of eye gaze only, an additional modality is suggested. The use of look-and-shoot with a physical trigger assumes that the user may have some mobility although it may be possible to use a triggering mechanism such as blowing in a pipe. Instead, this study will remove the reliance on physical dexterity and will build on the idea proposed by

Tan et al. (2003a) that speech could be used to activate the focused key. However, it also assumes that some users may have limited vocabularies and may not be able to vocalise all alphabetic letters. Therefore, a single command, which can be customised to meet the abilities of the user, will be used to activate the key which currently has focus. Through this means it will be possible to provide text entry capabilities using eye gaze and speech. This method of text entry may eventually prove to be more accurate than dictation giving the inherent recognition error rate with dictation systems. Furthermore when accounting for time spent on error correction, it may also be faster. The scope of this study will only encapsulate the comparison of this entry method with the traditional keyboard but comparison with dictation is proposed for future research.

2.10 Summary

This chapter discussed some of the relevant literature on which the current study was based. Based on the literature review it was hypothesised that multimodal interfaces may offer a more intuitive and natural means of human-computer communication. Modalities on their own oftentimes have associated disadvantages which can prohibit widespread acceptance and use of the modalities. However, instead of aggravating the problems experienced with the individual modalities, a multimodal interface can potentially compensate for disadvantages of one by drawing on the advantages of the other. Many examples of this were discussed in this chapter.

Since eye gaze and speech were the chosen modalities of the multimodal interface in the current study, each of these was discussed in detail. Interaction methods as well as methods to increase the accuracy of eye gaze were discussed. The identified activation mechanisms for eye gaze, namely dwell time, blinking and look-and-shoot, will all be included in the multimodal interface which will be developed as well as some of the mechanisms suggested to increase the accuracy of eye gaze. However, in order to negate the disadvantages of eye gaze, it will be coupled with speech. Negativity about continued use of speech recognition may stem from the high memorisation rate and system response and capability as well as the environments which are best suited to speech recognition. Memorisation of single commands using word processor terminology is not considered a serious consideration as word processors have a language which is unique to their environment and which has not curbed the popularity of the software, as is evidenced by its widespread use. Novice users may experience some difficulty, but it should not be more than the normal learning curve experienced when using the software. An additional mnemonic strain could occur if users are expected to remember an entire sequence of commands without visual feedback. However, if the speech commands are closely coupled with the naming of menus and tabs and executed in the same sequence as mouse clicks (for example, the menu name which causes the expansion of the menu, the menu command which opens the dialogue box and then utterance of a command), with the same level of visual feedback, it should not be a problem. The use of speech recognition will also alleviate the need to provide alternatives for all types of mouse clicking with eye gaze only, as commands can be provided to circumvent this. Using speech as an activation method for eye gaze could improve the speed and accuracy of other activation methods.

Many multimodal interfaces have already been empirically investigated but the combination of eye gaze and speech is a relatively new area, particularly when used for text entry. Some results have been forthcoming in this area but it remains to be seen whether eye gaze and speech will be able to achieve the speeds of more traditional means of input. Insofar as can be ascertained, the multimodal interactions have never been fully integrated into a mainstream application or a fully functional word processor. The development of such an application will be discussed in the following chapter as well as the methodology which will be followed to test the usability of the multimodal word processor.

CHAPTER 3

EXPERIMENTAL DESIGN AND METHODOLOGY

3.1 Introduction

The previous chapter discussed some of the available literature which was used to motivate the study and upon which it was based. Different types of gaze interaction methods were identified and the advantages and disadvantages of both eye gaze and speech were discussed. This chapter will discuss the experimental design which will be used to answer the research questions which have been posed. In particular, details of the actual tests used and the procedures followed will be elaborated upon.

3.2 Experimental design

The main aim of the study was to determine the feasibility and usability of eye gaze and speech when used as an interaction technique in a word processor (Section 1.2). Therefore, the study can be divided into two main parts, namely the feasibility and the usability of such a multimodal interface. In order to evaluate the feasibility of such an application, two phases were identified which had to be completed, namely:

1. The proposed application had to be developed in order to tentatively verify the feasibility of incorporating a multimodal interface into a word processor.
2. The feasibility had to be verified through a more concrete means than simple development.

In order to meet the requirements of the first phase, an application was developed which incorporated all the features of the proposed multimodal interface. This will be discussed in section 3.3. The second phase was achieved by conducting a feasibility test using Human-Computer Interaction (HCI) researchers as a sample. The experimental design of this phase will be discussed in section 3.4.1.

Once the feasibility has been established, the usability of the application will have to be tested. For these purposes, the functions that needed to be performed were identified. The experimental design for this phase will be discussed in sections 3.4.2 and 3.4.3.

3.3 Development of the application

3.3.1 Motivation

As previously discussed, a word processor application is one of the most popular and widely used applications. Furthermore, Microsoft Word® is the leader in the word processor market with high market penetration (Bergin, 2006a; Bergin 2006b). The interface used by Microsoft in previous versions has become the *de facto* standard for interfaces of similar packages as well as other types of applications and it may be said that it paves the way for the establishment of trends. Therefore, it stands to reason that Microsoft Word would play a central role in this study. Additionally, disabled users are often relegated to using software which has been specially developed for them. This software often does not encompass the full functionality of a product such as Microsoft Word. Moreover, the support and availability of such software is frequently not of the same

standard as offered by more mainstream suppliers. Consequently, Microsoft Word provides an ideal environment for the development of a truly multimodal interface.

3.3.2 Hardware

The eye-tracker used during the study was a Tobii T120 eye-tracker (www.tobii.com). This eye-tracker was chosen due to its availability at the university at which the study was conducted. The data rate of the T120 eye-tracker is 120Hz and the accuracy is measured at 0.5 degrees. The results obtained during the study should be interpreted with reference to the eye-tracker used as there are trackers available with both higher precision and accuracy. Since the only eye-tracker that was available for the study was the Tobii T120, the tests could not be conducted on a range of eye-trackers in order to determine the effect that it may have on the results.

A Logitech webcam with a built-in microphone was used to capture verbal utterances for speech recognition. The computer used had a quad core i7 processor with 4 GB of RAM. The screen resolution was set to 1280×1024 at all times and the participants were requested to sit approximately 60 cm from the screen.

3.3.3 Development tools

Visual Studio Tools for Office (VSTO), allows .NET developers to customise not only the interface of the Office suite but also to add the functionality that is required (Anderson, 2009). 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. The programming language was C# using the .NET framework 3.5.

In order to include the interaction techniques of speech and eye-tracking, some third party tools were required. The Microsoft speech application programming interface (SAPI) is the native speech API for Windows (Microsoft, nd) and provides access to text-to-speech (TTS) engines as well as automatic speech recognition (ASR) engines (Simon, 2002). The SAPI software development kit (SDK) provides samples and tools to incorporate speech capability in developed applications (Microsoft, n.d.). The SAPI allows the use of dictation in an application or specialised grammars can be created for use within the application. The Microsoft SAPI is free to download and since its capabilities were deemed sufficient for the purposes of this study, it was used to provide speech capabilities for the multimodal interface.

In order to provide eye-tracking capabilities, the Tobii® SDK was used. The availability of the Tobii eye-tracker at the university at which the study was conducted was the overwhelming factor in selecting its associated SDK for use.

Magnification capability was also provided in the application. For these purposes, the relatively inexpensive Magnifying Glass Pro® (www.workerscollection.com/wcollect/english/html/mg_pro.html) was used. This tool allows the magnification of the area directly under the mouse cursor. Furthermore, it is one of the only tools discovered which allow for capturing of mouse clicks on the magnified area. These mouse clicks are then automatically transferred to the underlying area and the application responds appropriately without having to close the current magnified area. As such, this tool provides functionality which many freeware tools were lacking. Should it be found that the tool increases productivity and assists the user in invaluable ways, future research can include the development of a free-to-use magnification tool which achieves the same functionality as Magnifying Glass Pro, or alternative ways can be investigated to use currently available freeware products and to allow the magnified area to be interactive.

3.3.4 Interaction techniques

The interaction techniques of eye gaze and speech were proffered as solutions to create a highly customisable, hands free multimodal interface which could potentially cater for a diverse group of users with varying capabilities and levels of expertise. The aim was to make an all-encompassing application through which all of the necessary interaction could take place so as to minimise disruption for the user who is then not required to switch between multiple applications. To this end, all tools such as the calibration, setting of the gaze interaction sensitivity and others were all included in the application. For example, Figure 3.1 shows the results of a calibration contained within a separate window in Microsoft Word. The only external tool which was required was the training wizard for the speech engine. For this the user must use the wizard through normal Windows interaction and not through the developed application.

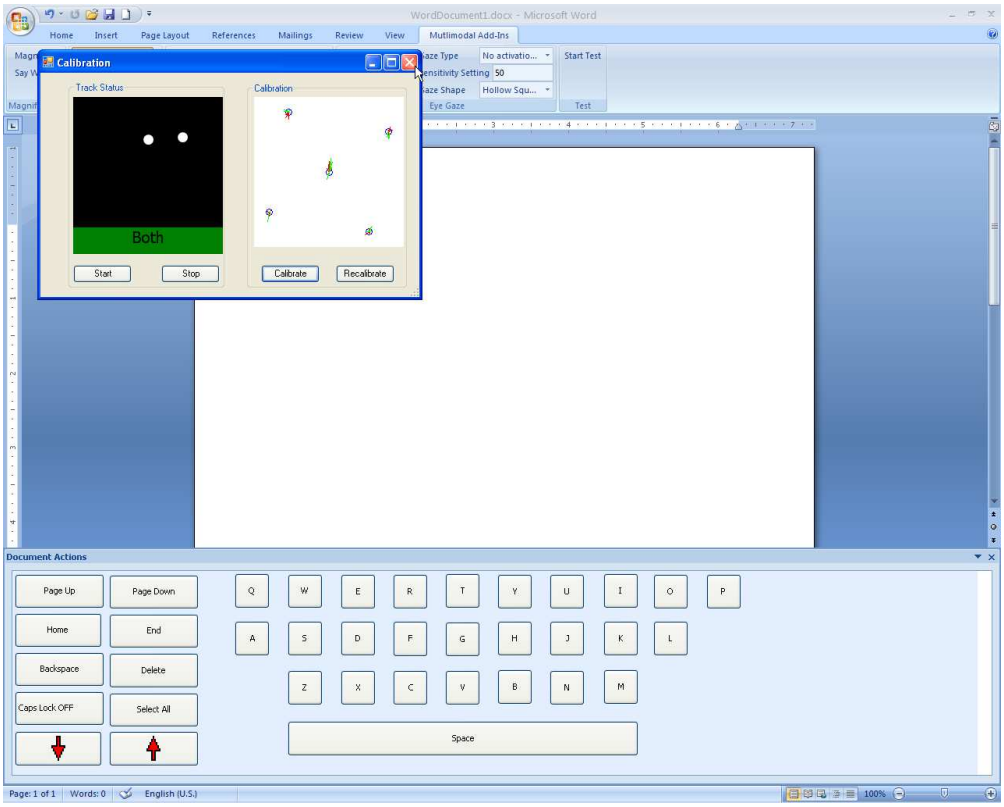


Figure 3.1: Calibration process in Microsoft Word

The discussion in Chapter 2 highlighted various ways in which eye gaze could be used as an input technique, namely dwell time, look-and-shoot, gaze gestures and blinking. Apart from gaze gestures which were not implemented due to time constraints, all the other means of communication were incorporated into the application. Look-and-shoot uses the Enter key as an activation mechanism. Furthermore, the sensitivity of the dwell time can be set to allow further customisation. Additionally, eye gaze and speech recognition can be used in combination as an interaction technique similar to look-and-shoot but where the activation mechanism is a speech command and not a physical device.

In order to be able to use these interaction techniques for text input, onscreen keyboards are available and are displayed as a panel in Microsoft Word. The figure below shows the onscreen QWERTY keyboard. The keyboard not only has alphabetic characters but also provides commonly used keys such as *Page Up*, *Page Down*, *Home*, *End* and *Delete* keys. The user is also able to activate and deactivate *Caps Lock*. A *Select All* button is provided to allow an easy method of selecting a large amount of text with a single click. When a button on the onscreen keyboard is pressed using any activation mechanism, audible feedback is given in the

form of a clicking sound. Therefore, it will not be necessary for users to look at the document in order to obtain confirmation that the button has been pressed.

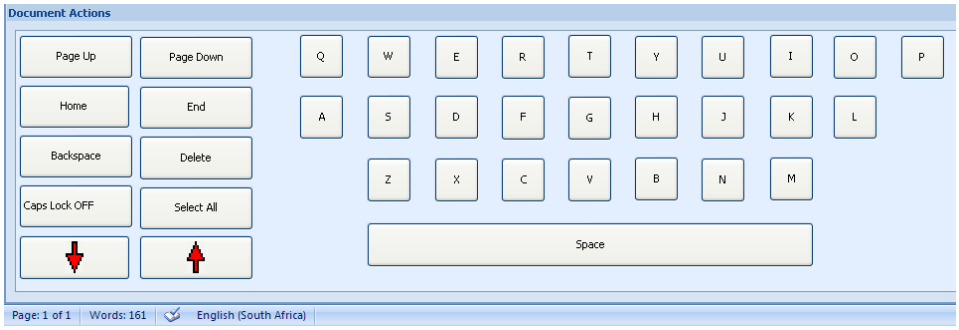


Figure 3.2: Onscreen QWERTY keyboard

One implication of using eye gaze to simulate a pointing device is that buttons and selectable targets must be larger than in standard interfaces. Additionally, spacing between targets might have to be adjusted in order to allow a margin of error around each selectable target so that it can be identified accurately for activation based on proximity of the eye gaze to each selectable target in the area immediately surrounding the eye gaze. Therefore, the buttons on the keyboard were larger than standard Windows buttons and were also more widely spaced. The arrow buttons on the bottom left of the keyboard allow the user to respectively decrease and increase the size of the keyboard keys. Resizing the keyboard keys also causes the size of the lettering on the keys to be resized proportionally. Magnification was also proposed as a possible solution to decrease the amount of screen real estate which is lost whilst still enjoying the advantages of larger selectable targets. The use of the magnification tool is shown in the figure below. The yellow arrow on the figure indicates the current position of the mouse cursor. The default zoom factor enlarged the area to double its actual size within a 400×300 window.

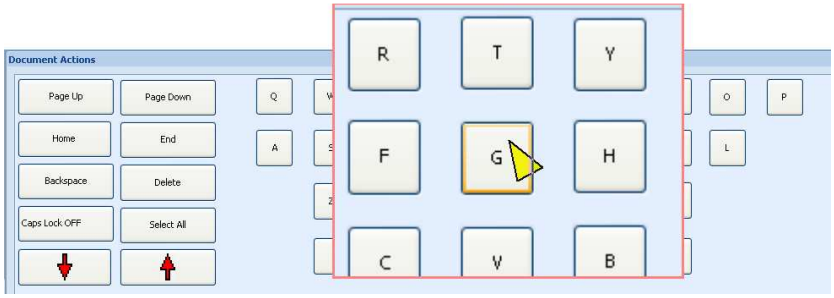


Figure 3.3: Magnification of the onscreen keyboard

It was established in a prior section that feedback is vital when using eye gaze as a pointing device so that the user is aware of the position at which the eye gaze is being detected. However, a gaze indicator which is slaved to the eye gaze may disturb the user as the gaze indicator will never be still if it accurately reflects eye movements. Taking this into account, it was decided that the gaze indicator should remain stable within the confines of the closest selectable target. Consequently, the gaze indicator does not mimic eye movement but is rather stabilised on a selectable target for as long as the eye gaze is situated closest to that target. In order to further minimise the negative impact of a gaze indicator, gaze position is indicated only when the gaze is on the onscreen keyboard and not when it is situated on other parts of the document.

Originally, gaze position was indicated by centring a 10 x 10 pixel solid square on the button directly under the gaze of the user (Figure 3.4a). During the feasibility testing, before the formal user testing commenced, it was recommended that the square not obscure the letter on the key, as this requires that the user look away and

then back to confirm that the correct key has been located. Consequently, the square was positioned slightly off-centre leaving the letter on the button completely visible (Figure 3.4b).

Figure 3.4: (a) Centred and (b) off-centre gaze position indicator

Although the off-centre button allowed the letter on the button to be visible it was feared that novice users would follow the gaze indicator and not focus on the button. However, since the square stays in a stable position on the button for as long as the eye gaze is on the button or in nearby proximity, even should the user follow the indicator, it should have no effect whatsoever. However, to allay concerns that the off-centre indicator may distract the users or prevent them from accurately seeing the letter, even though this should comfortably be perceived in their peripheral vision, other options were explored for gaze indicators. This included a hollow circle (Figure 3.5a) or square (Figure 3.5b) which surrounds the letter on the button, where the width and colour of the shape could be set by the user.

Figure 3.5: (a) Hollow circle and (b) square used as gaze indicators

However, the most aesthetically pleasing option, with the simultaneous benefit of providing unequivocal confirmation of which button was receiving focus, were the two options included in the final application. These were a frame which was drawn around the border of the button and the inverting of the button colour. The frame is green in colour which provides ample contrast to alert the user as to which button currently has focus. It also has the added advantage that the letter on the button is completely visible and contained within the frame. The second option provides visual feedback by inverting the colour of the button which has focus. This means that when a button has focus, its background colour is a darker grey and the colour of the letter is white. The figure below shows a framed button on the left and on the right, the use of the inverted colour.

Figure 3.6: Visual feedback on a selectable target through (a) framing and (b) inverting colours

Speech recognition was also incorporated as a standalone interaction technique to facilitate a means of navigation, editing, selection and manipulating text. The speech engine provides a means for both dictation and commands to be issued. A specialised grammar was developed for use within a word processor. This grammar allows for formatting controls, document handling, basic and complex cursor control and mouse manipulation. The complete set of available commands is tabulated below.

Table 3.1: Verbal commands

	Command	Application reaction	Current key press
Formatting commands	Bold	Activate/deactivate bold	[CTRL] + B
	Italic Emphasise	Activate/deactivate italic	[CTRL] + I
	Underline	Activate/deactivate underline	[CTRL] + U
Document handling commands	Cut	Cut the current selection	[CTRL] + X
	Copy	Copy the current selection	[CTRL] + C
	Paste	Paste the current clipboard item at the cursor position	[CTRL] + V
	Undo	Undo the previous action	[CTRL] + Z
	Delete Remove	Delete text to the right of the cursor or a current selection if present	[DELETE]
Basic cursor control	Down	Move the cursor one position down	[DOWN] arrow
	Left	Move the cursor one position to the left	[LEFT arrow
	Right	Move the cursor one position right	[RIGHT] arrow
	Up	Move the cursor one position up	[UP] arrow
	Home	Move the cursor to the start of the current line	[HOME]
	End	Move the cursor to the end of the current line	[END]
Complex cursor control and selection techniques	Select line	Select the entire line that the cursor is currently on	[HOME] and then [SHIFT] + [END] OR Requires left mouse click in the left margin
	Select word	Select the word subsequent to the cursor or the current selection	[SHIFT] + [CTRL] + [RIGHT] arrow
	Select word back	Select the word prior to the cursor or the current selection	[SHIFT] + [CTRL] + [LEFT] arrow
	Shift down	Move the cursor down as though the Shift key is in	[SHIFT] + [DOWN] arrow
	Shift left	Move the cursor left as though the Shift key is in	[SHIFT] + [LEFT] arrow
	Shift right	Move the cursor right as though the Shift key is in	[SHIFT] + [RIGHT] arrow
	Shift up	Move the cursor one position up as though the Shift key is in	[SHIFT] + [UP] arrow
	Select All	Selects all the text in the document	[CTRL]+[A]
Mouse manipulation	Click Activate Select Go	Left mouse click	

An extra tab was added to the ribbon in Microsoft Word (Figure 3.7) to accommodate all the additional features which were added. This tab is called *Multimodal Add-Ins* and allows the user of the application to set the interaction techniques as desired. Table 3.2 summarises the settings and explains the options provided on the multimodal tab.

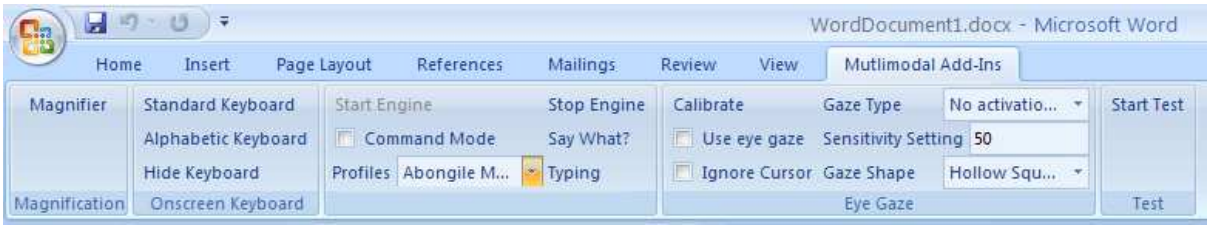


Figure 3.7: Multimodal Add-Ins tab in Microsoft Word

As can clearly be seen from the summary in Table 3.2, the multimodal interface is highly customisable to suit the expertise and the current needs and environment of the users. The interaction techniques were added in order to complement the existing input methods and were not intended to replace them. Therefore, the multimodal interface meets the requirement of having to provide alternative means of input to prevent overuse of a single one (Oviatt & Cohen, 2000). The next section will discuss some technical specifications of the developed application.

3.3.5 Technical specifications

As mentioned in a previous section, VSTO was used to modify the interface of the Word environment. Third-party tools and SDKs were then used to add the required functionality. These tools included the Tobii SDK, Microsoft speech API and Magnifying Glass Pro. Therefore, apart from using VSTO, these additional tools had to be managed and the functionalities they provided had to be programmed into the VSTO solution. Figure 3.8 illustrates the classes used in the application to get the complete set of interaction techniques. The class diagram does not contain insignificant class attributes, for example, those used to monitor the status of the interaction techniques. Similarly, no properties are indicated on the class diagrams. The class diagrams are used simply to show the essential functionality of the classes.

The `cCommandList` class controls the panel which is displayed to show the command list for the speech recognition grammar.

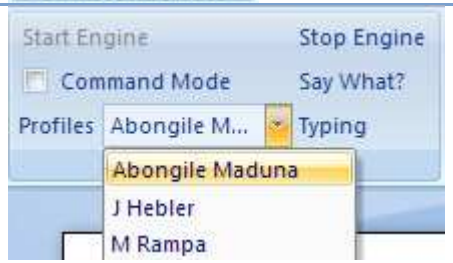
The `cKeyboard` class maintains the layout of the standard and alphabetical keyboards. It is also responsible for the resizing and consequent spacing of the keys on the keyboard. The `KeyPrint` method is used by all key presses to type the letter associated with the pressed button in the document at the current cursor position.

The `cRibbon` class was used for the design and basic on`Click` events for the Multimodal Add-Ins tab.

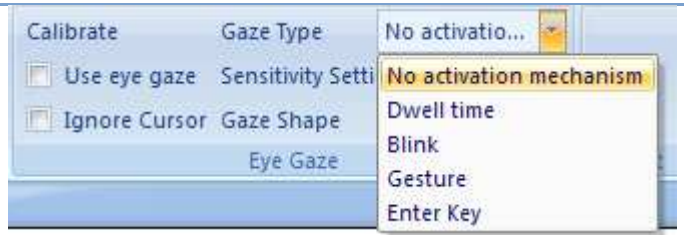
The `DocumentContentManager` class executes all commands that are issued within the current Word document. For example, the `SelectLine` method will select the line of the document on which the cursor is currently situated. The `DocumentBoldCommand` will send a command to the document to toggle bold formatting on or off depending on its current status.

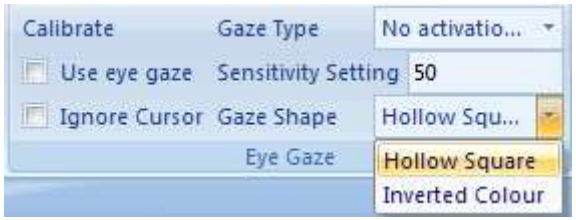
The `cSpeechController` class uses the Microsoft speech API to handle both the dictation and command mode. The class manages the toggling between dictation and command mode, starts and stops the speech engine and loads and manages the use of the Windows speech profiles. The class is responsible for building the grammar required, capturing verbal utterances and responding to them via a `DocumentContentManager` object. The commands captured via a `cSpeechController` object will invoke the correct method in the `DocumentContentManager` object which will in turn send the correct command through to the Word document.

Table 3.2: Multimodal Add-Ins tab functions

Group	Screenshot	Explanation
Magnification		This *Magnification* button allows the user to toggle the magnification on and off. A standard Microsoft Office toggle button is used to ensure that the user is always aware of the status.
Keyboard		This group allows the user to control the display of the onscreen keyboards. When the user presses the *Standard Keyboard* button, a standard QWERTY layout keyboard is displayed along the bottom of the Word document (Figure 3.2). When the user presses the *Alphabetic Keyboard* button, an onscreen keyboard is displayed along the bottom of the Word document configured in alphabetical order.
Speech		This group allows the user to control the use of speech within Microsoft Word. The *Start Engine* and *Stop Engine* are two mutually exclusive buttons used to control whether the speech engine is active or not. When the user presses the *Start Engine* button, the speech engine is activated and the interface will react to any verbal utterance that is captured. The user must then press the *Stop Engine* button in order to deactivate the speech engine. The *Profiles* drop-down box loads all the trained profiles from the list that Windows maintains. The user can then select the profile that they would like to use for the speech engine component of the multimodal interface. When the *Command Mode* check box is selected, the speech engine only responds to words contained within the word processor grammar (Table 3.1). Otherwise, when the speech engine is on, the speech engine is in dictation mode and any verbal utterances captured are written to the current document through the Speech-To-Text engine. The *Say what?* button shows a list of acceptable verbal commands which can be issued in order to perform common word processing tasks. The *Typing* button allows the grammar to be minimised to commands for selection of onscreen targets only. Therefore, all the formatting, text selection and navigation commands are disabled when the feature is activated.

Gaze





The *Calibrate* button allows the Tobii calibration process to start. The calibration is required for a new user to ensure that the tracking of the eye gaze is accurate. Calibration occurs exclusively through the Word interface (Figure 3.1).

The *Use eye gaze* checkbox provides a quick mechanism for the user to toggle the reaction to eye gaze on and off.

The *Sensitivity Setting* allows the user to determine the length (in milliseconds) of the dwell time and the sensitivity of system response to user blinking.

The *Gaze Type* drop-down box allows the user to choose how the system must react to the eye gaze.

- No activation mechanism – This allows the user to use eye gaze and speech together. The eye gaze of the user is tracked and when a verbal command is issued, the command is executed at the current position of the eye gaze.
- Dwell time – the system responds to dwell time as set by the sensitivity setting. If the user gazes at a particular area for the length of the sensitivity setting, then a left mouse click is executed at that location. The length of the dwell time can be set to increase the customisation of the application.
- Blink – the system responds to blinks of the user. The blink must be more pronounced than an involuntary blink so the natural blinking process should not interfere with the interaction technique.
- Gesture – the system will respond to an eye gaze gesture by executing a left mouse click at the current location of the eye gaze. Gestures were not implemented due to time constraints.
- Enter Key – this implements the look-and-shoot method of interaction. When the user presses the Enter key, a left mouse click is executed at the current location of the eye gaze. The gaze location is only interpreted should it be located on the onscreen keyboard. Therefore, look-and-shoot does not interfere with normal typing on the document area should the user also wish to use the keyboard for some typing.

The *Gaze Shape* drop-down box allows the user to specify the form of the visual feedback on the onscreen keyboard. By default, the button is framed but if users so prefer, they can also choose to invert the colour of the button which is currently being gazed at.

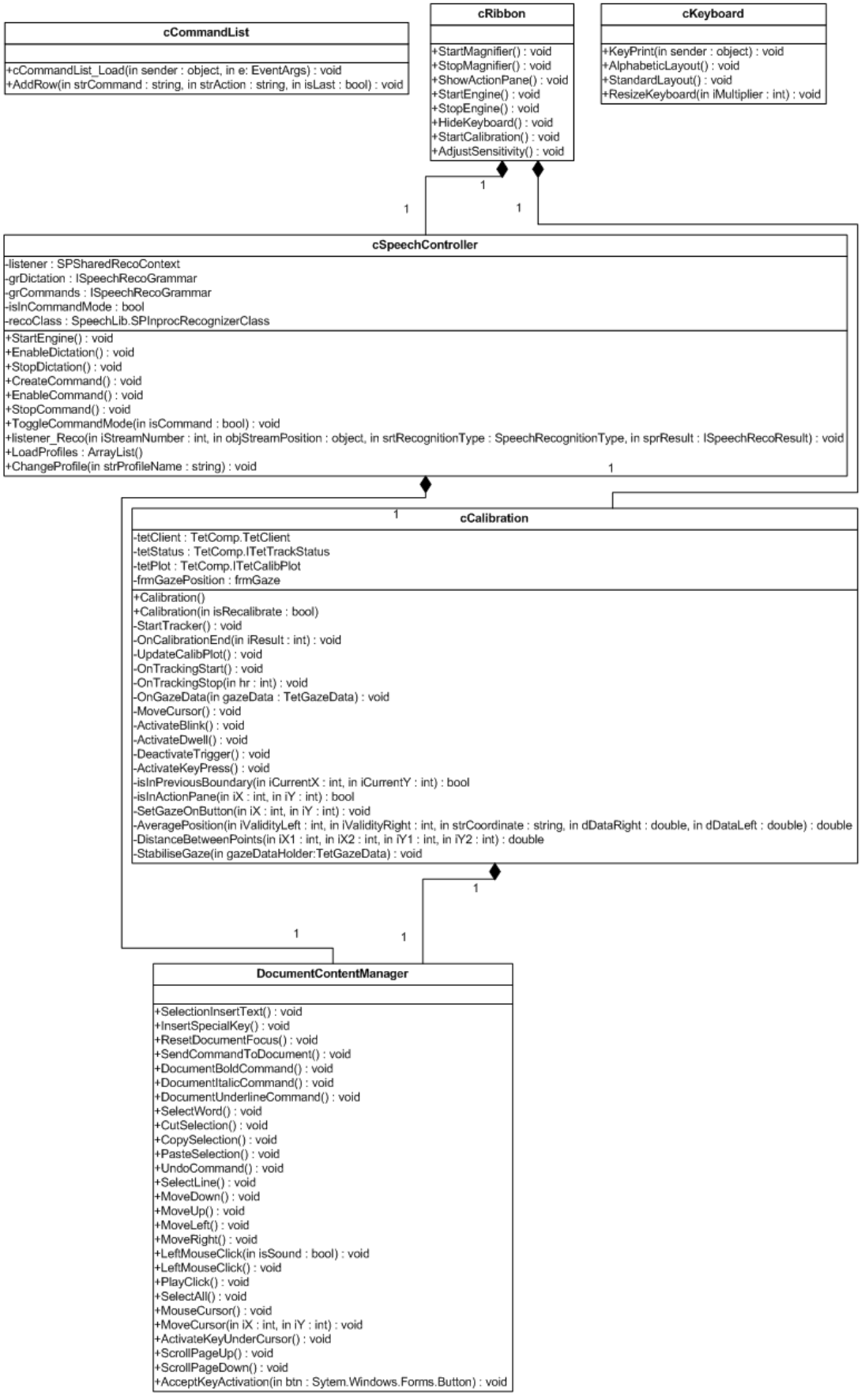


Figure 3.8: Class diagram of developed application

The `cCalibration` class is responsible for the management of all eye-tracker generated data. At the most basic level, it handles the calibration process for the Tobii eye-tracker and stores the results of the calibration process. It is also responsible for turning the eye-tracking function on and off. Thereafter, if it is required it will monitor the eye gaze of the user and respond to blinks and dwell time appropriately. When look-and-shoot or speech is used, this class is used to determine the position of the eye gaze so as to interpret where the left mouse click must be executed. Furthermore, the class also controls the visual feedback of the gaze indicator by determining whether the eye gaze of the user is currently positioned over the onscreen keyboard. In order to do this the class employs the use of a gaze stabilising algorithm. Since the eye is subject to noisy fixational eye movements, slaving the gaze indicator to the eye gaze in a dedicated fashion will result in a fairly jumpy gaze indicator which might distract the users more than it will assist them. Therefore, extracts of the smoothing algorithm of Kumar (2007) were used to stabilise the gaze on the button nearest the current eye position. This algorithm smooths the data in real time by determining whether the most recent point is the start of a saccade, whether it belongs to the current fixation or whether it is an outlier. For these purposes, if the distance between two points is more than a previously defined threshold then a saccade is detected. The algorithm is robust to noise since it measures "the displacement of each eye relative to the current estimate of the fixation location rather the to the previous measurement" and movements one movement ahead which are over the threshold are rejected (Kumar, 2007).

Together with this smoothing algorithm, the fixation points were calculated as the weighted mean of all the points in the fixation window using the following formula (Kumar, Klinger, Puranik, Winograd & Paepcke, 2008) where *p* is a data point within a fixation window with *n* points:

$$p_{mean} = \frac{1p_0 + 2p_1 + \cdots + np_{n-1}}{(1 + 2 + \cdots + n)}$$

The result of the implementation of this algorithm is a much smoother movement of the gaze indicator and more accurate determination of the eye gaze position.

3.3.6 Resulting multimodal interface

The previous sections detailed the development of the application, both in technical terms as well as giving a visual and all-encompassing discussion on the interaction techniques provided and how they could be used within the environment of a word processor. The resulting application was one which was highly customisable and provided a multitude of interaction means as replacement options for the traditional keyboard and mouse. All of these additional functions could be provided to the end-user through the well-known Microsoft Word.

This development of the complete solution positively answered the first research question which was posed in section 1.5, namely whether it was possible to provide a highly customisable multimodal interface using eye gaze and speech within a mainstream word processor. The next sections will discuss how the remaining four research questions will be answered.

3.4 Resolving the empirical research questions

Research questions 2 and 3 (with the three secondary research questions) must still be answered. However, they necessitate that a robust experimental design be established before they can be resolved. Research questions 2 and 3 as defined in section 1.5 are as follows:

2. How feasible is such an interface and in which context is it feasible?
3. How usable is the multimodal interface compared to the traditional interaction techniques?
    a. How usable is the combination of eye gaze and speech when used to simulate a pointing device?
    b. How usable are speech commands for performing common word processing tasks?
    c. How usable is the combination of eye gaze and speech when used for text entry?

The experimental design required to answer research question 2 will be discussed in the following section. Thereafter, section 3.4.2 will discuss the methodology which will be used to answer question 3a. The approach employed to answer questions 3b and 3c will be discussed in section 3.4.3

3.4.1 Feasibility study

In order to explore objective 2 as detailed in Chapter 1 and answer research question 1 above, a feasibility study will be conducted using the application developed. The scope of a feasibility study is not to identify usability problems, but rather to determine whether the envisaged system has long-term potential as a viable multimodal interface within the realm of modern-day word processing.

Due to the nature of the study, five participants are sufficient for such an undertaking (Nielsen, 2000). Therefore, five senior members of the lecturing staff of the university where the study was conducted, who are proficient in the field of HCI will be approached to participate in the study. A pre-test questionnaire will be used to measure their level of expertise and exposure to the technologies used in the multimodal questionnaire and is contained in Appendix A.

The participants will be given a thorough demonstration of the functionalities of the application and then allowed some time to become familiar with the system. They will then be requested to complete some simple open-ended tasks (Appendix B). The tasks will be left open-ended as the results of the tasks will not be evaluated and the premise of this initial study is simply to observe user interaction and allow participants the freedom to use the system in order for them to form an objective opinion of the system. At the end of each participant's session, the participant will be required to complete the post-test questionnaire as contained in Appendix C. Results of the feasibility study will be discussed in detail in Chapter 4.

3.4.2 Pointing and clicking

The next research question which must be answered is in regard to the usability of eye gaze and speech when used to simulate a pointing device. A suitable means of user testing must first be determined which will facilitate the collection of data. The data must fulfil the requirement that at least one measurement per usability component must be analysed.

3.4.2.1 Assessment of a pointing device

The most commonly used metrics to evaluate pointing devices are speed and accuracy (MacKenzie, Kauppinen & Silfverberg, 2001) which give a good indication as to whether there is a difference between the performance

of pointing devices (Hwang, Keates, Langdon & Clarkson, 2004). In 1954, Paul Fitts proposed a relationship between target size and distance to the target which could effectively predict movement time from the current position to the targeted position (Fitts, 1954). This relationship was henceforth known as Fitts' Law. Since major pointing devices are used to position a cursor over a target using hand movement (Shneiderman, 1998), Fitts' Law has often been applied to HCI. The application by HCI pundits was generally in one of two ways, namely to predict the time required to position a cursor over some target based on the distance to travel and the size of the target or as a means to derive the throughput (discussed below) by measuring movement times and then determining how the different conditions affect the coefficients in Fitts' Law (Soukoreff & MacKenzie, 2004). In this way, it became possible to establish effective efficiency between various pointing devices.

However, since pointing devices are no longer only used to point but also to draw, write and navigate through nested menus, Fitts' law presents somewhat of a limited methodology in terms of pointing devices (Accot & Zhai, 1999). These additional uses for pointing devices are all trajectory-based, a feature which Fitts' Law is ill-equipped to evaluate. Therefore Fitts' Law alone, when used to test pointing devices, neglects to test the quality of trajectories produced by these pointing devices (Accot & Zhai, 1999) and a more comprehensive set of tests is required to strengthen the comparison.

The inclusion of Fitts' Law in an International Standards Organisation (ISO) standard ISO 9241-9 (ISO, 2000) confirmed its pre-eminence as the leader in the evaluation of pointing devices whilst also providing for trajectory-based testing to be performed on pointing devices through extension of Fitts' Law. The ISO standard uses a throughput metric which encapsulates both speed and accuracy (ISO, 2000) in order to compare pointing devices and is measured using any one of six tasks including three point-and-click tasks which conform to Fitts' Law (Carroll, 2003). The six tasks included in ISO 9241-9:

1. Tapping tests (one-directional and multi-directional)
2. Dragging tests
3. Path-following tests
4. Tracing test
5. Free-hand input test
6. Grasp and park test

The one-directional tapping test requires the participant to move from a home area to a target and back. In contrast, the multi-direction tapping test consists of 24 boxes placed around the circumference of a circle. The participant is then required to move from the centre of the circle to a target box. From there the participant must move to and click in the box directly opposite that box and then proceed in a clockwise direction around the circle (Figure 3.9) until all the targets have been clicked in and the user is back at the first selected target box. The target which should be selected next should always be graphically highlighted for the user (Soukoreff & MacKenzie, 2004).

The dragging test is a variation of the one- and multi-directional tapping test where the user is required to drag an object and drop it in the destination target box. The path-following test requires the participant to trace or steer along a pre-defined path of a certain width. The fourth test is the tracing test which requires that the participant follow a circular path whilst attempting to stay within the bounding circles. The fifth task as set out by ISO 9241-9 is designed to test the effectiveness of the pointing device for entering free-hand text or pictures (Douglas, Kirkpatrick & MacKenzie, 1999).

The final test is the grasp and park test during which "the subject performs a simple pointing task and operates a key on the keyboard between each pointing with the same hand" (ISO, 2000). This task is also referred to as a device switching task (Douglas et al., 1999), as it requires the subject to point at a target and then press a key on the keyboard using the same hand as was used to point at the target object. The tasks as set out in ISO

9241-9 can be used to evaluate and compare pointing devices and can be selected according to their applicability to the pointing devices in question.

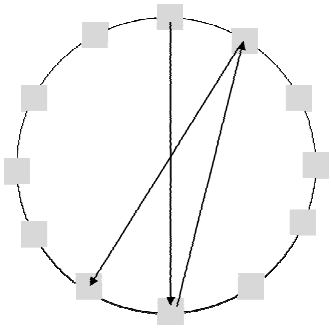


Figure 3.9: Multi-directional tapping test using ISO9241-9

Throughput is measured using the tests as set out in ISO 9241-9 and is reported as bits per second (bps). The equation for calculating throughput is Fitts' Index of Performance, with the exception that an effective index of difficulty is used (Zhang & MacKenzie, 2007). The equation for throughput is (Natapov, Castellucci & MacKenzie, 2009):

$$Throughput = ID_e / MT \quad (1)$$

where MT is the mean movement, in seconds, for all trials within the same condition and

$$ID_e = log_2(D_e / W_e + 1) \quad (2)$$

where De is the effective distance to the target and We is the effective width of the target and is calculated as

$$W_e = 4.133 * SD_x \quad (3)$$

The effective distance to the target (De) is the distance the subject traversed along the task axis (Natapov et al., 2009). In turn, the task axis is measured as the straight line from the centre of the source to the centre of the target (Natapov et al., 2009). The term SDx is the standard deviation of the selection coordinates (Douglas et al., 1999).

Apart from the standardised tests to measure throughput, ISO 9241-9 also provides a questionnaire designed to assess aspects of the operation, fatigue, comfort and overall usability (ISO, 2000) of the pointing devices.

Since its inception in draft form in 1998, ISO 9241-9 has been used to compare a multitude of pointing devices ranging from joysticks and touchpads (Douglas et al., 1999), to mouse emulators (Man and Wong, 2007), hand gestures and even video game controllers (Natapov et al., 2009). It was originally designed to apply to a mouse, trackballs, light-open and styli, joysticks, touch-sensitive screens, tablet-overlays, thumbwheels, hand-held scanners, pucks, hand-held bar code readers and remote-control mice (Douglas et al., 1999). It was not designed to cover input devices such as speech activators, head-mounted controllers, data gloves, devices for disabled users or foot-controlled devices (Douglas et al., 1999). However, its compatibility with eye-trackers and testing with disabled users has since been established through a number of studies (Keates et al., 2002; Zhang & MacKenzie, 2007; Man & Wong, 2007; Gajos et al., 2008).

The ISO tasks that have been selected to be included in this study are as follows:

1. Multi-directional tapping task

The task was selected based on its applicability to the study and based on the functions the interaction techniques will eventually fulfil in the word processor application. For example, there is no use including the

grasp and park since the test requires that a keyboard key must be pressed between selecting targets using the same hand as for manipulation of the pointing device. Even though the user will be able to use both eye-tracking and speech recognition in combination with the keyboard, the eye-tracking and speech recognition provide a completely hands-free environment and users will never have to switch devices using their hands.

While ISO9241-9, similar to Fitts' Law, is undoubtedly a step in the right direction, allowing researchers to establish whether there are differences in speed and accuracy between various pointing devices, it does however fail to determine why these differences exist (Keates & Trewin, 2005). MacKenzie et al. (2001) propose seven additional measures which will provide more information as to why differences are detected between performance measures of pointing devices. These measures are designed to complement the measures of speed, accuracy and throughput and to provide more insight into why differences exist between pointing devices. The seven measures as proposed by MacKenzie et al. (2001) are as follows:

1. Target re-entry
    a. If the pointer enters the area of the target, leaves it and then re-enters it, a target re-entry has occurred.
2. Task axis crossing
    a. A task axis crossing is recorded if the pointer crosses the task axis on the way to the target. The task axis is normally measured as a straight line from the centre of the home square to the centre of the target (Zhang & MacKenzie, 2007).
3. Movement direction change
    a. Each change of direction relative to the task axis is counted as a movement direction change.
4. Orthogonal direction change
    a. Each change of direction along the axis orthogonal to the task axis is counted as an orthogonal direction change.
5. Movement variability
    a. This "represents the extent to which the sample points lie in a straight line along an axis parallel to the task axis".
6. Movement error
    a. This is measured as the average deviation of the sample points from the task axis, regardless of whether these sample points are above or below the task axis.
7. Movement offset
    a. This is calculated as the mean deviation of sample points from the task axis.

The ISO9241-9 multi-directional tapping task was used to verify these metrics with 16 circular targets, each 30 pixels in diameter and placed around a 400 pixel diameter outer circle (MacKenzie et al., 2001). These seven metrics, as well as throughput, movement time and missed clicks were used in a study to determine the difference in cursor movement for motor-impaired users (Keates et al., 2002).

A further six metrics which could assist in determining why a difference exists, were specifically designed for use with disabled users and were proposed by Keates et al. (2002). These measures will not be used during this study as they are not considered relevant. An additional metric measuring the number of clicks outside the target is also suggested in order to measure the performance of pointing devices (Keates et al., 2002).

Additional measurements will be analysed in an effort to explain the difference in performance if such a difference exists between the interaction techniques. These additional measurements will either be some of the afore-mentioned measurements or they will be derived from these measurements. Therefore, the total task completion time will be measured as well as the task completion time from when the target is highlighted to when it is clicked, the number of target re-entries, the number of incorrect targets which are acquired during task completion and the number of incorrect clicks. This will allow efficiency and effectiveness to be

tested. Furthermore, the ISO device assessment questionnaire, which is reproduced in its entirety in Appendix E (Questions 1-9), will be used to test satisfaction to a degree.

3.4.2.2 Experimental design

The ISO test requires that the size of the targets and the distance between targets be varied in order to measure the throughput. Therefore, variable size targets will be used, but in order to reduce the time required to complete a test the distance between targets will not be adjusted during this testing.

The smaller icon on the Word ribbon is 24x24 (visual angle ≈ 0.62°) pixels in size. This was therefore used as the base from which to start testing target selection with speech recognition and eye gaze. Miniotas et al. (2006) determined that the optimal size for targets when using speech recognition and eye gaze as a pointing device was 30 pixels. This was determined using a 17'' monitor with a resolution of 1024x768. Participants were seated at a viewing distance of 70 cm. This translated into a viewing angle of 0.85°. The eye-tracker used in this study was a Tobii T120 with a 17'' monitor where the resolution was set to 1280x1024. In order to replicate the viewing angle of 0.85° obtained by Miniotas et al. (2006), a 30 pixel target could be used but at a viewing distance of 60 cm from the screen. Therefore, the next size target to be tested in the trials was determined to be a 30x30 pixel button. It was decided to also test a larger target than that established by Miniotas et al. (2006). Following the example set by Miniotas et al. (2006) of testing target sizes in increments of 10 pixels, the final target size to be used was 40 pixels (visual angle ≈ 1.03°).

The multi-directional tapping task will have sixteen targets situated on a circle with a diameter of 800 pixels. The targets will be positioned on the edges of the circle – thereby creating an inner circle with diameter of 800 pixels. Square targets will be used and not circular targets as the buttons in the final application will be rectangular in shape. Therefore, it was decided that square targets will be more meaningful since they are also allowable under the ISO standard.

Target acquisition will either be via eye-tracking and speech recognition or the mouse. The mouse will be used to establish a baseline for selection speed. When using a verbal command to select a target, the subjects will have to say "go" out loud in order to select the target that they are looking at. This method of pointing can therefore be considered analogous to look-and-shoot. The word "go" was chosen as it was established during development that this was the word which was most accurately captured by the speech engine with minimal training. The words "select" and "click" will also be available as verbal commands.

The literature review uncovered various shortcomings of using eye gaze for target selection, namely the instability of the eye gaze and the difficulties experienced in selecting small targets. In order to combat these shortcomings, a number of solutions have been proposed. These include magnification and the use of a gravitational well. Consequently, both of these techniques will be tested during this phase of the research study. The magnification settings were the same as for the Word application while the gravitational well was activated within a 50 pixel radius around each button. Another prerequisite of using eye gaze as a pointing device is that visual feedback is given at all times. Since the final application provides a choice between inverting the colour or framing the button which has focus, both these visual feedback mechanisms will be tested in order to establish whether they affect the performance of the pointing device.

Therefore, there were essentially three varying conditions which could be combined according to the matrix as depicted below (Table 3.3). Since the mouse is the benchmark against which the alternative means of pointing and selecting must be evaluated, it was not deemed necessary to have a gravitational well with the mouse at any point. Secondly, the fact that a faster means of mouse selection was not under inspection meant that only the traditional means of mouse selecting had to be measured.

Table 3.3: Matrix of test conditions for ISO testing

	Framed visual feedback	Inverted visual feedback
Gravitational well	Eye-tracking and speech recognition	Eye-tracking and speech recognition
No gravitational well	Mouse / Eye-tracking and speech recognition	Mouse / Eye-tracking and speech recognition

Additional trials will also be included using magnification on the 24 pixel targets to determine whether magnification alone can allow users to achieve comparable speeds with the standard size icons. For this reason, magnification was not combined with the gravitational well and was also not used with both visual feedback techniques. This resulted in a total of fourteen trials per session (Table 3.4), the number of which served as motivation for not adding more trials for the mouse as this would simply prolong the session time and might cause participants to become irritable and fatigued during the session. Since this could influence the results it was decided to forgo additional mouse trials since all participants had to be proficient with the mouse and two trials with the mouse was considered sufficient to get an accurate throughput for the mouse.

Table 3.4: Multi-directional tapping trials

Group	Trial settings
M(F)	Mouse,30,Framed,No target magnification, No gravitational well
M(I)	Mouse,24,Inverted,No target magnification, No gravitational well
MM	Mouse,24,Inverted,Target magnified, No gravitational well
ETS(F)	Eye gaze and speech,30,Framed,No target magnification, No gravitational well
	Eye gaze and speech,40,Framed,No target magnification, No gravitational well
ETS(I)	Eye gaze and speech,30,Inverted,No target magnification, No gravitational well
	Eye gaze and speech,40,Inverted,No target magnification, No gravitational well
ETSG(I)	Eye gaze and speech,30,Inverted,No target magnification, Gravitational well
	Eye gaze and speech,40,Inverted,No target magnification, Gravitational well
ETSG(F)	Eye gaze and speech,30,Framed,No target magnification, Gravitational well
	Eye gaze and speech,40,Framed,No target magnification, Gravitational well
ETSM	Eye gaze and speech,24,Inverted,Target magnified, No gravitational well
	Eye gaze and speech,30,Inverted,Target magnified, No gravitational well
	Eye gaze and speech,40,Inverted,Target magnified, No gravitational well

Three sessions will be conducted in which all 14 trials will have to be completed by all participants. The first session of the study will be preceded by a pre-test questionnaire (Appendix D) which will capture user demographics and other information pertinent to the study. The full-length questionnaire is contained in Appendix E. The ISO device assessment questionnaire will be administered at the end of testing to measure subjective opinion of the pointing device.

When using a repeated measures design the dangers of asymmetric skill transfer are heightened. Asymmetric skill transfer or learning effects are often encountered due to the order in which the tasks or treatments are presented to the subject (Poulton & Freeman, 1966). The best way to counterbalance these learning effects is through the use of a balanced Latin square (Bradley, 1958; Reese, 1997). By varying the interaction techniques using a Latin square, a measure of control will also be imposed upon the results, thereby lending further credibility to the results. Therefore, a balanced Latin square for all trial conditions was obtained by following the instructions provided by Edwards (1951). Participants will be randomly assigned to a Latin square condition for each session.

The target button which must be clicked will be denoted by an "X". An example of one of the trials using inverted colour feedback, since the eye gaze is currently focused on the target button, is depicted below:

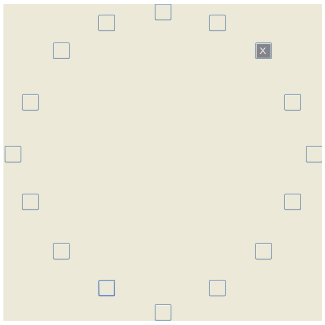


Figure 3.10: Multi-directional tapping task using eye gaze and speech with target button currently having focus

3.4.3 Word processor functions and text entry

The word processor functions of navigation, editing and formatting of text as well as text entry will be tested together during user testing.

3.4.3.1 Assessment of word processor functions

Standard usability measures should suffice to test the usability of the word processing functions of navigating, formatting and manipulating text in a word processor. A number of usability measures are advocated as a means to measure usability of a software application (cf. Bohmann, 2000; Faulkner, 2000; Nielsen, 2001a; Nielsen, 2001b; Preece et al., 1994). Usability models which consolidate numerous measurements of usability have been proposed and tested by various authors (cf. Abran, Suryn, Khelifi, Rilling & Seffah, 2003; Dix, Finlay, Abowd & Beale, 1993). These were also considered for inclusion in the study. It was however found that the vast majority of the proposed measurements were either not applicable or were very similar to the five measurable objectives proposed by Shneiderman (1998). Therefore, these usability measures were considered an acceptable foundation from which usability measures could be extracted in order to measure the aforementioned identified components of usability, namely effectiveness, efficiency and satisfaction as well as the additional component of learnability. These measureable objectives are as follows (Shneiderman, 1998):

- Time to learn – How long it takes users to learn the commands or actions necessary to complete the task.
- Retention over time – The extent to which users retain the knowledge they have gained.
- Speed of performance – Time taken to complete the task.
- Rate of errors by users – The number of errors made by the user in an attempt to carry out the task.
- Subjective satisfaction – The level of satisfaction, or how much they enjoyed working with the system or parts thereof.

Time to learn and retention over time are learnability factors. This portion of the research study will make use of longitudinal testing, which requires that the same tasks are completed during a number of sessions. This will allow the learnability of the interaction technique to be tested. The speed of performance or time to complete the task will be used as a measure of efficiency, as will the additional measurement of the number of actions required to complete the tasks. Effectiveness will be analysed in terms of the correctness with which the task could be completed as opposed to the number of errors made. Furthermore, subjective satisfaction will be tested through the use of a questionnaire.

3.4.3.2 Assessment of text entry

Measuring the efficiency and effectiveness of text entry requires other measurements to be analysed. The underlying concept of testing text entry is to present some text which must be entered using the interaction technique which is to be tested. The resultant text which is then entered, often referred to as the transcribed text, is then compared to the presented text in order to determine how different they are. The minimum string distance (Levenshtein, 1965) or so-called Levenshtein distance can be used for this purpose. This distance is calculated as the minimum number of corrections which must be made in order to transform one string into another (Wobbrock, 2007). The operations which can be used to transform the strings are the insertion of a character (i), the deletion of a character (d) and substituting one character for another (s).

This Levenshtein distance can then be used to determine the effectiveness measurements of character error rate (CER) and percentage correctness measures (Read et al., 2001). The Levenshtein distance is divided by the number of characters to obtain the CER (Equation 1).

$$CER = \frac{s+d+i}{n} * 100 \quad (1)$$

The character error rate is a negative effectiveness measurement which can then be transformed into the positive measurement percentage correctness measurement, denoted by PCM (Read et al., 2001):

$$PCM = 100 - CER \quad (2)$$

The number of characters typed per second (Equation 3) have been successfully used as measures of efficiency for word processing text entry methods (Read et al., 2001).

$$CPS = \frac{Characters\ input}{Time\ taken} \quad (3)$$

For an accurate measure of characters per second, the number of characters should be measured as the number of characters in text input – 1 (MacKenzie, 2002). This is to compensate for the fact that the preparation time for the first character cannot be accurately measured. Therefore, by discarding a single character the time is measured as starting from the preparation time for the second character until the last character is input.

Other text entry measurements include keystrokes per character (Soukoreff & MacKenzie, 2001), gestures per second (Wobbrock, 2007) and corrected, uncorrected and total errors as well as efficiency of error correction (Soukoreff & MacKenzie, 2003). Apart from these there are a number of other measurements which are available for use but were not deemed applicable to this study.

In order to test the usability of the proposed text entry method only one measurement per usability component was considered namely, characters per second as an efficiency measurement and character error rate as an effectiveness measurement. User demographics (Appendix F) and the satisfaction experienced by users will be measured through the use of a questionnaire (Appendices G and H) and learnability will be monitored as the progress that is made over a number of sessions.

3.4.3.3 Experimental design

Each participant will be required to complete a task list comprising representative tasks in a word processor environment (Table 3.5).

Table 3.5: Word processor functions and text entry testing task list

Task no	Task text	Task type	Skill being tested
1	Underline the first three lines of text using speech recognition.	Line selection and formatting	Selection Formatting
2	Italicise the last three lines of text using the keyboard.	Line selection and formatting	Navigation Selection and formatting
3	Use speech recognition to select all the text in the document and delete it.	Select all text and remove	Selection Editing/Manipulation
4	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing	Typing
5	Use speech recognition to select the first two words of the sentence and make them bold.	Select words and format	Navigation Selection Formatting
6	Use the keyboard to select the whole paragraph and then to cut it.	Select all text and remove	Selection Editing/Manipulation
7	Type the following phrase using the keyboard: *<randomly selected phrase>*	Typing	Typing
8	Use the keyboard to select the first two words in the document and then to make them bold.	Select words and format	Navigation Selection Formatting
9	Use speech recognition to italicise all the text.	Select all and format	Selection Formatting
10	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing	Typing
11	Paste the previously cut text using the keyboard.	Paste	Editing/Manipulation
12	Undo your previous action using speech recognition.	Undo	Editing/Manipulation
13	Paste the previously cut text using speech recognition.	Paste	Editing/Manipulation
14	Undo your previous action using the keyboard.	Undo	Editing/Manipulation
15	Type the following phrase using the keyboard: *<randomly selected phrase>*	Typing	Typing
16	Use speech recognition to select the last word and to copy it.	Select word and copy	Navigation Selection Editing/Manipulation
17	Use the keyboard to insert the copied word after the second word.	Position and paste	Navigation Editing/Manipulation
18	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing	Typing
19	Use the keyboard to select the last word and to copy it.	Select word and copy	Navigation Editing/Manipulation
20	Use speech recognition to insert the copied word after the second word.	Position and paste	Navigation Editing/Manipulation

The task list was compiled in such a way as to include elements of all of the required functions. Each task will also specify the interaction technique which must be used to complete the task. In order to perform meaningful comparative analysis between the traditional methods of input and the proposed interaction technique, similar tasks will have to be performed using both these interaction techniques. For example, the participant will be required to position the cursor correctly, select some text and then copy it using either the keyboard or the mouse. A similar task will then have to be performed using speech commands. A window containing the task instruction will be overlaid on the top-right hand corner of the Word window.

There are also a number of text entry tasks, some of which must be completed using the keyboard and others using speech recognition and eye gaze. For each of these tasks, the sentence which must be input will be randomly chosen from a set of 35 pre-selected phrases. These phrase sets were chosen from the 500 as determined by MacKenzie and Soukoreff (2003) to be everyday phrases which are commonly used. The subset was selected based on its applicability to the setting of the study as well as for their length, character set and level of difficulty. Phrases with unusual words or hard to spell words were omitted from the list as this did not conform to the aim of the study. Some phrases were, however, included based on whether they contained double letters. In order to emulate the study conducted by Karl et al. (1993) which also tested text entry within a word processor and verbal commands to complete formatting, the phrases were also chosen for their memorability and familiarity so that participants could easily remember the phrase to be entered and would not have to continually refer back to a hard copy or, in this case, the instructional window.

The phrase set which was chosen was:

- Time to go shopping
- Elephants are afraid of mice
- You must be getting old
- I agree with you
- Take a coffee break
- Fish are jumping
- I am wearing a tie and a jacket
- All together in one big pile
- Goldilocks and the three bears
- My favourite web browser
- Have a good weekend
- This is a very good idea
- User friendly interface
- It is very windy today
- Zero in on the facts
- Universities are too expensive
- A picture is worth many words
- The dog buried the bone
- The daring young man
- Prepare for the exam in advance
- A dog is the best friend of a man
- That is a very odd question
- Rapidly running short on words
- Dolphins leap high out of the water
- Nothing finer than discovering a treasure
- The location of the crime
- Luckily my wallet was found
- They watched the entire movie
- Sit at the front of the bus
- The elevator door appears to be stuck
- With each step forward
- Wishful thinking is fine
- What goes up must come down
- Insurance is important for bad drivers
- Tell a lie and your nose will grow

Analysis of the phrase set reveals the descriptive statistics tabulated in Table 3.6. The five most frequently occurring letters are summarised in Table 3.7.

The letter “E” and “T” are the first and second most frequently used letters in English text, with “I”, “A” and “O” are in the group of third most frequently used letters in English text (Oxford Dictionary, 2011). When analysis letter frequency in main dictionary entries and not printed text, the letter “E” is the most frequently used letter, “A” the second, “I” the fourth, “O” the fifth and “T” the sixth (Oxford Dictionary, 2011). Therefore, the phrase set selected for use closely resembles the English language in terms of letter frequency.

Table 3.6: Descriptive statistics for phrase set

Descriptive statistic	Measurement
Number of phrases	35
Minimum characters in phrase (excluding spaces)	13
Maximum characters in phrase (excluding spaces)	36
Minimum characters in phrase (including spaces)	20
Maximum characters in phrase (including spaces)	41
Average ± Standard Deviation of characters in phrase (excluding spaces)	22.29±5.27
Average ± Standard Deviation characters in phrase (including spaces)	26.6±6.16
Number of words	186
Number of unique words	100
Minimum word length	1
Maximum word length	12

Table 3.7: Frequencies with which letters occur in selected phrase set

Letter	Frequency
E	72
T	58
I	52
A	48
O	47

The most frequently occurring words are:

Table 3.8: Most frequently occurring words in selected phrase set

Word	Frequency
the	14
a	11
is	7
of	5
in	3
are	3
and	3
very	3

The words “the”, “a”, “is”, “of”, “in” and “and” are in the top 7 most commonly used words in English, while the word “are” is the fifteenth most commonly used word and “very” the 127th most commonly used English word (Fry, Kress & Fountoukidis, 1993; word-english, 2003). Therefore, the most frequently occurring words in the phrase set used are also some of the most commonly used words in the English language.

As suggested by MacKenzie and Soukoreff (2003), no capitalisation or punctuation will be expected when entering the phrase sets.

In order to gauge learnability of the application using the new interaction techniques, the participants will be required to complete 10 sessions so that their progress can be measured. Each participant will have to attend one session per week. During the first session, participants will complete a pre-test questionnaire, designed to elicit the participant’s expertise with a word processor. Since expertise is a measurement of both the frequency and length of use (Rosson, 1984), the questions will be phrased so as to gauge both of these aspects of the participant’s expertise. A number of other demographics will also be captured through this

questionnaire. The complete questionnaire can be seen in Appendix F. During subsequent sessions, the participants will have to complete the task list as set out in Table 3.5 to the best of their ability. A post-test questionnaire (Appendix H) will be administered during the final session to gauge user satisfaction after prolonged use of the application.

3.5 Statistical analysis

Inferential statistics will be used to analyse the data and investigate the stated hypotheses. The notation H_0 will be used to denote the null hypothesis or the hypothesis of no difference. Where there are multiple null hypotheses under investigation, the notation $H_{0,i}$ will be used to denote the i^{th} null hypothesis. For example, $H_{0,1}$ is the first null hypothesis and $H_{0,2}$ is the second null hypothesis.

Since most of the data captured will follow a within-subjects experimental design, the data will be in the form of repeated measures, where a number of measures are taken for each participant over a number of sessions, for the same condition. In order to determine whether there is a significant difference between measures taken over a number of sessions with the same participants, a repeated measures analysis must be used. Normality tests will be performed on the data in order to verify whether the data is normally distributed or not. If the data is normal a suitable parametric test will be used, otherwise an equivalent non-parametric test will be used.

Since the same participants will be tested multiple times, paired tests can be used for analysis. Where there are only two dependent variables, a paired t-test can be used. The non-parametric equivalent of a paired t-test is the sign test (Whitley and Ball, 2002) or the Wilcoxon test (Motulsky, 1995). If there are multiple independent variables or more than two dependent variables then a within-subjects repeated measures ANOVA can be used.

Repeated measures ANOVA assumes normality and sphericity of data (Minke, 1997). A non-parametric alternative to the repeated measures ANOVA is the Friedman test (Mutolsky, 1995). However, since the ANOVA is robust to violations of normality, it will be used in all instances regardless of the distribution of the data. Mauchley's sphericity test will be used to verify whether assumption of sphericity is met before analysis commences. Sphericity can be compared to the homogeneity of variance in the between-groups ANOVA (Field, 1998). When the assumption of sphericity is not met, there are a number of corrections which can be applied to the degrees of freedom, such as the Geisser-Greenhouse correction, the Huynh-Feldt correction and the Lower Bound correction (StatSoft, 2010). The closer the Greenhouse-Geisser estimate is to 1 the more homogeneous the variances of differences are and the more spherical the data is (Field, 1998). For a Greenhouse-Geisser estimate larger than 0.75, the more conservative Huyn-Feldt adjusted correction should be applied (Girden, 1992; Nimon & Williams, 2009). If the assumption of sphericity is not met, the F ratio is positively biased which increases the chances of rejecting falsely (Maxwell & Delaney, 2004). Some texts advocate reporting the results of both the univariate and multivariate approaches (Minke, 1997), while others advise using the multivariate approach wherever possible (StatSoft, 2010) since these tests are not dependent on the assumption of sphericity (Field, 1998). Therefore, when adjusted corrections are required, the results of the multivariate test will also be reported for the sake of completeness and to ensure that the results are not compromised by the lack of sphericity.

Line graphs will used to graphically illustrate the measures over time and for comparison purposes. Where line graphs are used, the vertical bars will denote a 95% confidence interval. The confidence interval is calculated as follows (StatSoft, 2010):

$$Confidence\ interval = \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

When significant differences are detected, post-hoc tests will be conducted to determine which factors were the cause of the significant difference. Tukey's HSD test will be used for post-hoc tests.

Where the data is summarised in a frequency table, the Chi-Square test will be used to test whether two variables are independent of one another (StatSoft, 2010).

Before the commencement of any statistical analysis, outliers will be removed from the data. A data point will be determined to be an outlier if one of the following conditions held (StatSoft, 2010):

1. data point value > UBV + o.c.*(UBV - LBV)
2. data point value < LBV - o.c.*(UBV - LBV)

where UBV was the upper value or 75th percentile, LBV was the lower bound or the 25th percentile and o.c. was the outlier coefficient which was set to 1.5.

3.6 Summary

The development of a customisable, highly inclusive multimodal interface for a mainstream word processor was discussed in this chapter. The development of such an application, together with all the functionality which was provided, was discussed in detail. The resulting application offers a multimodal interface which can be adjusted to meet the needs and circumstances of a wide group of users. However, the usability of the proposed interaction techniques must still be established. To this end, the experimental methodology which will be followed in order to answer the remaining research questions was also elaborated upon. The settings for the user testing as well as the tasks which must be completed were discussed together with the procedure which will be followed during user testing. The following chapter will discuss the results of the first of these experiments, namely the feasibility testing which was conducted.

CHAPTER 4

FEASIBILITY TESTING OF THE MULTIMODAL INTERFACE

4.1 Introduction

In the previous chapter the development of the multimodal interface for Word was discussed. Using a variety of tools, eye gaze and speech were incorporated into the popular word processor application. It was found that in this way it was possible to create a highly customisable, hands-free multimodal interface for a mainstream application.

Since the multimodal interface has successfully been included in Microsoft Word, the next step was to determine if the interface was a feasible solution. A feasibility test is aimed at determining whether the proposed interface is viable and whether it could offer a potentially usable interface to any users. Therefore, contrary to a more formal usability study, it does not require that objective measurements be captured and analysed statistically. A number of Computer Science lecturers, who were familiar and comfortable within the field of Human-Computer Interaction (HCI), were approached to complete a questionnaire designed to elicit their reaction to, and assessment of the proposed system. This chapter reports on the results of this feasibility testing.

4.2 Participants

It was established in Chapter 3 that five participants were sufficient for such a study. The five participants included in this study all had extensive experience in the field of HCI. These particular participants were targeted for inclusion in this sample based on the fact that they had the experience and foresight which is required to objectively judge the long-term viability of an application. Since the aim of this phase of the study was not to test the actual usability of the application but rather to determine whether such an application has potential, sampling was performed under this premise. Two of the participants specialised in HCI research, one in e-learning, one in web programming and security and the other in general computer programming. Moreover, all had experience in the field of HCI and were comfortable with the terminology and principles of this field. Two of the participants were female and the average age of the participants was 33.8 (standard deviation = 6.6).

4.3 Tasks

Participants were required to complete the pre-test questionnaire as contained in Appendix A. This made it possible to determine if they did in fact fall into the required target group. Once this was verified, participants were given a short demonstration on the use of the application and the various customisable features which were available. The command list in Table 3.1 was provided to them and Appendix B offered some suggestions for them to familiarise themselves with the application. They were then allowed to interact freely with the application, encouraged to explore the various options and to make full use of the functionalities offered by the application. Thereafter they all completed the post-test questionnaire in Appendix C.

4.4 Limitations

One limitation of this study is the small sample size, which has the consequence that statistical analysis could not be done on the results of the questionnaire. Therefore, the results will be reported on in an anecdotal manner which precludes the possibility of generalising to the population. Additionally, the fact that all participants were involved in HCI research could mean that responses to the questionnaire are biased. Should a different sample be approached then the results may differ substantially. Furthermore, the questions posed are very subjective in nature and also mean that they not be able to be generalised and may be very biased.

However, since the purpose of this part of the study was to set the stage for the larger study and to provide substantiation for the study these limitations were not of high consequence.

4.5 Results

Results of the questionnaires will not be statistically analysed since the sample size is too small for meaningful analysis. Instead the responses will be inspected and reported on in an anecdotal manner.

Four of the respondents were initially excited by the system, predominantly due to the possibilities it offered to disabled users in particular. Interaction with the application did not change the viewpoint of any of the participants – including the single respondent who was sceptical about the use of such a system. The main concern of this participant was the lack of control one has over one's eye gaze. Unintentional and natural movements of eye gaze are hard to suppress and do cause a dilemma for researchers. However, there are possibilities of overcoming such shortcomings, such as smoothing and stabilisation algorithms which can be applied to the eye gaze response. The Midas touch problem can also be neutralised to a degree through the correct adjustment of dwell time settings or use of mechanical activation. Therefore, the main concern of the sceptic can be countered.

All participants agreed that the time has come for a paradigm shift in user interface design and that a multimodal interface may be the way of the future. To this end, the combination of eye gaze and speech was met with enthusiasm and optimism although concern was voiced that some practice would be needed to become accustomed to the interface. Since most new systems require some level of training and practice in order to master, this should not be considered a problem unique to the proposed system. Additionally, the naturalness and intuitive means of interaction could prove more of an advantage to the application although there are of course the inherent problems associated with the interaction techniques, such as the Midas touch, which have to be compensated for.

Chart 4.1 shows the spread of the responses to a number of questions designed to gauge the subjective feelings of the participants towards the system.

In most instances the response to the combination of eye gaze and speech and to the provision of a multimodal interface in a mainstream application are positive. However, the majority of the respondents felt that they could not navigate through the document very easily but that with extended use and practice they could improve their speeds to a satisfactory level.

Since there are various options available to facilitate interaction with eye gaze, namely dwell time, blinking, look-and-shoot as well as combining it with speech, participants were asked to rank these interaction techniques according to their own preference and then also according to their usability within a word processor. One participant did not answer these questions. The highest (n = 3) preference was for look-and-shoot, followed by the combination of eye gaze and speech, followed by blinking and finally dwell time. This is quite understandable as look-and-shoot and the combination of eye gaze and speech probably offer the highest feeling of control over the system. For instance, blinking is a natural occurrence and is difficult to

control. While the system does require a more pronounced blink to be executed before it responds, this could still lead to a feeling that blinking is not an allowable action. This perception would presumably change as more practice with the system is gained. Similarly, dwell time might appear to place more strain on the eyes as it requires a stable gaze to be maintained on an object of interest for a specified time.

In terms of usability, look-and-shoot was seen as the most usable of the interaction techniques, followed by the combination of eye gaze and speech. Dwell time and blinking tied for the third most usable interaction technique.

In conclusion, participants see the value of such an application, most notably as a means of absorbing disabled users into the mainstream user group. All respondents were in agreement that a multimodal interface for Word is a desirable development and that eye gaze and speech offered a viable multimodal solution. Overall, look-and-shoot was viewed as the preferred method of interaction as well as the most usable interaction technique of the four. While this might be true, some users may not have the mobility to perform such an action and should they have a limited vocabulary the combination of eye gaze and speech could offer a more usable and quicker means of interaction than dwell time and blinking.

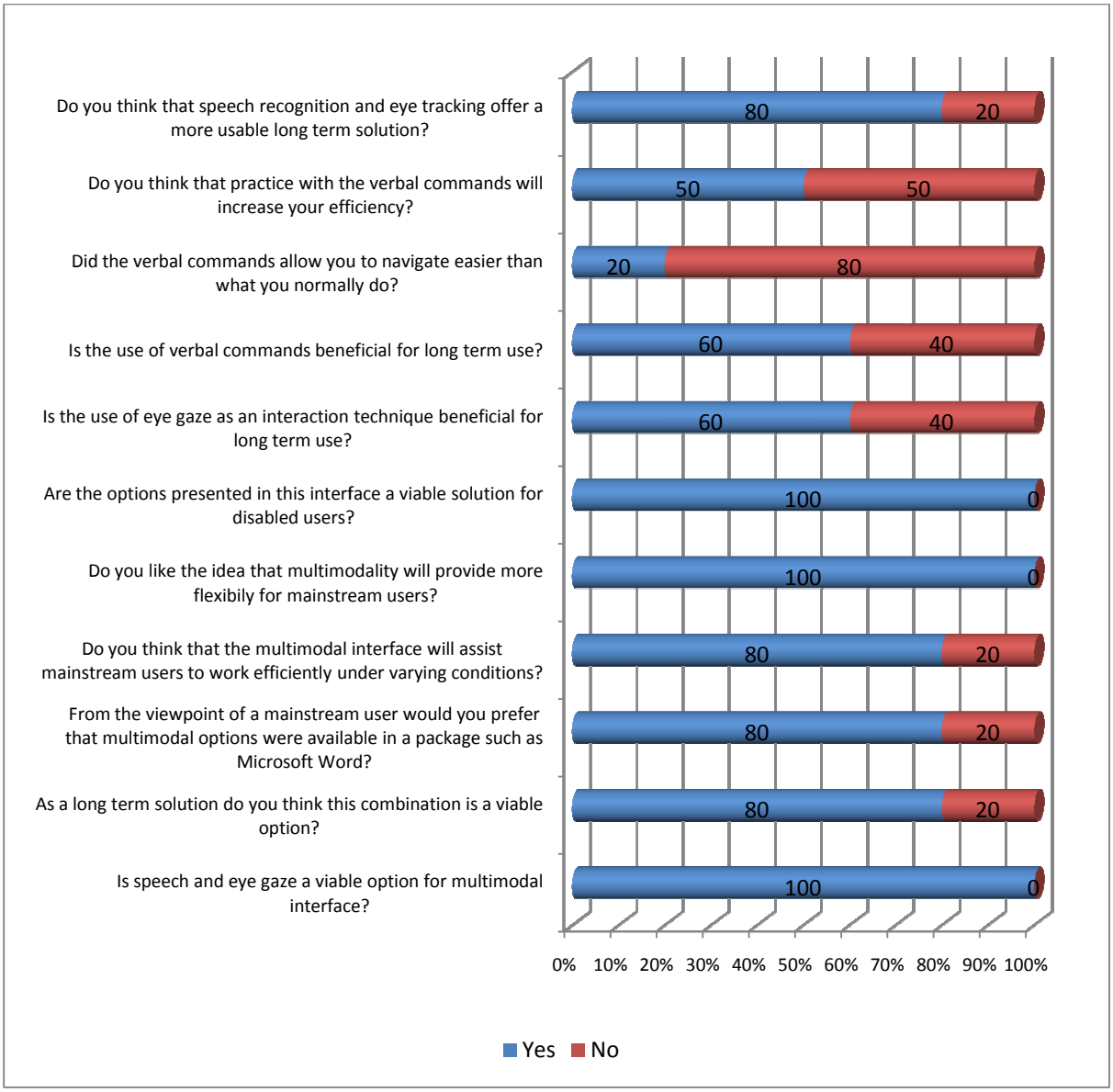


Chart 4.1: Responses to questionnaire

4.6 Conclusion

This chapter reported on the results of the responses of a number of lecturers, all of whom were comfortable with HCI research, to the proffered multimodal interface for Word. The overall reaction was a positive one, particularly in light of the possibilities it will offer to disabled users who will be able to interact with a mainstream application. Additionally, in the opinion of the participants, a multimodal solution of eye gaze and speech is possibly an acceptable solution for a diverse group of users and not only for disabled users.

Since these results indicate optimistic subjective feelings towards the system, it now remains to be seen whether the objective usability goals can be met. The next chapter will report on the first of these, namely the testing of eye gaze and speech when used for pointing-and-clicking purposes.

CHAPTER 5

ANALYSIS OF EYE GAZE AND SPEECH TO SIMULATE A POINTING DEVICE

5.1 Introduction

The previous chapter reported on the feasibility review which was conducted. Overall, the reaction of the reviewers to the vision and voice word processor was positive. It must now be investigated as to whether this word processor will provide a suitable experience to end-users. The first step in this investigation is to determine whether the combination of eye gaze and speech can effectively be used to simulate a pointing device. Section 1.5 identified that one of the common means of interaction with a word processor was through the selection of icons and menus via the use of a pointing device. Furthermore, when using the onscreen keyboard, eye gaze and speech will be used to select keys, therefore serving the purpose of pointing and clicking. The usability of a pointing device can be established by using the International Standards Organisation (ISO) pointing device test, ISO 9241-9, to compare interaction techniques.

This chapter will report on the results of such an ISO test which was conducted using an eye-tracker and speech recognition as an alternative to a mouse. A number of trial conditions were arranged and a group of participants each completed a few sessions with each trial. The results will give an indication of the viability of eye gaze and speech recognition to simulate a pointing device. The standard ISO measurement of throughput was analysed. Thereafter, the time was analysed as a separate variable since the nature of some of the interaction techniques could negatively influence the throughput. Some other measurements, which were identified in previous chapters, were also analysed, namely the number of target re-entries, the number of incorrect target acquisitions, the number of incorrect clicks and the time to selection of the designated target. All these measurements will be defined and then analysed in terms of their differences between interaction techniques. The chapter provides an in-depth analysis of all these measurements in an attempt to scrutinize the viability of the proposed interaction techniques.

5.2 Participants

A convenience sample was used for this part of the study as participants were sourced from the student population of the University of the Free State. Participants were expected to be competent with the computer and mouse and therefore they were chosen based on their exposure to a computer. Consequently, the sample consisted of senior students (no first year students).

Each participant completed three sessions and each session consisted of all fourteen trials, which will be discussed in the following section. For the first session, there were 20 participants. However, five of the participants did not return for the second or third sessions for various reasons. Therefore, in total there were 15 participants who completed all three sessions and only the data of these fifteen participants was included in the final analysis.

Eleven of the participants were male and 4 were female. The average age of the participants was 22.3 (standard deviation = 1.9). Analysis of the indicated computer expertise on the pre-test questionnaire (Appendix D) indicated that all participants could be ranked as having high computer expertise. Similarly, all participants ranked as having high mouse expertise. All participants indicated that they had neither eye-tracking nor speech recognition experience. No previous eye-tracking or speech recognition was required as it

was considered more desirable that participants did not have prior exposure to these technologies. Therefore, the sample was well set up with regard to computer and mouse expertise and lack of eye-tracking and speech recognition experience.

5.3 Trials

As discussed in Chapter 3 there were 14 trial conditions (Table 3.4). To recap the trial conditions were as follows:

1. Eye gaze and speech with no added features.
2. Since the accuracy of eye gaze and speech could influence the ability of users to select small targets, magnification of the target can be achieved in two ways, namely:
    a. through "invisible" expansion of the target by using a gravitational well. This means that the selectable area of the button is larger than the size of the target as it is portrayed in the interface. When a stable eye gaze is detected within this larger area, the eye gaze is pulled onto the target, thereby creating a gravitational well; and
    b. through magnification of the area directly under the eye gaze.
3. Visual feedback is essential in the use of eye-tracking as a pointing device; therefore different means of visual feedback were investigated.
4. As the primary pointing device used in Word, the mouse was included for comparative purposes.

While it is acknowledged that the interaction technique is essentially the mouse or eye gaze and speech, a distinction will also be made based on the visual feedback that was provided as it cannot be assumed that this did not affect user performance. Proper statistical analysis will be performed on the data before consideration will be given to disregarding the visual feedback used.

Since the mouse is the interaction technique which is regarded as the benchmark, the condition of MM is not of importance to the scope of the study and will not be included in the analysis.

5.4 Sessions

Each session required that the participant complete all fourteen trials. Participants were randomly assigned a Latin square condition for each session. No participant was assigned to the same Latin Square condition more than once.

The first session commenced with each participant giving informed consent to participate in the research study. Participants then completed pre-test questionnaire (Appendix D). The purpose of the study was then explained and a quick overview of the procedure and requirements was given. The first and second sessions were, unfortunately, spaced 10 weeks apart as the laboratory was occupied between the sessions. The second and third sessions were two days apart. Admittedly, the uneven spacing between the sessions is not advisable. However, since the factor of interest was the interaction technique and not the learning effect over time it was decided that the period between session 1 and session 2 should not have an effect on the analysis or results of the interaction technique. At the end of the third session, each participant completed a post-test questionnaire (Appendix E) to gauge subjective reaction to the proposed interaction technique.

Since the participants were sourced from the university, they were all fluent in either English or Afrikaans as these are the tuition languages of the university. Each session was conducted in the language that the participant was most comfortable with. The participants received an incentive, in the form of a gift voucher, for each session they completed.

Some problems were experienced with the equipment during the second session of two participants. This resulted in no data being captured for some of the mouse tasks for these two participants. The M(F) task of two participants had to be discarded and the M(I) task for one of these participants also had to be excluded from the analysis. Therefore, their mouse throughput will be calculated from only two sessions. The remainder of the trials as well as the rest of the participants will all be calculated using all three session's data.

5.5 Device movement

Complete statistical analysis of the data must be performed, but first the data was inspected visually. For this purpose, the path of each trial was traced and drawn as an overlay over the trial setup. The images were only extracted for illustration purposes to serve as a visual representation of the trial completion. They also serve to give an idea of how much movement was required to complete the task. The set of images contained in Figure 5.1 below show some paths that were traced as the participant completed a trial. The first image (a) is for the mouse, with framed feedback. The second (b) is for eye gaze and speech with no gravitational well, the third (c) is for a trial with a gravitational well. The fourth image (d) is for a magnification trial. The blue lines signify mouse movement and the red lines eye movement.

The black circles indicate captured data points. It is important to note that not all data points are represented as it was not essential to capture the actual movement; therefore not all data points were saved, although they were reacted to in real-time during the course of the test. The numbers indicate the sequence of the clicks and are placed at the exact position where the click occurred. Note that when the gravitational well is activated, it is effectively possible that the click occurs outside of the button since the button is essentially larger than what is actually represented on the screen. Also take note that the buttons are redrawn as graphical squares for the purposes of this representation and that during the trial the buttons were displayed as standard Windows buttons. These squares are redrawn to be the exact size of the buttons during the trial and the decision to draw them as squares was purely to facilitate a simplified drawing process.

Figure 5.1 are visual representations generated for a participant who did not struggle to complete the trials. They represent the first session for this participant.

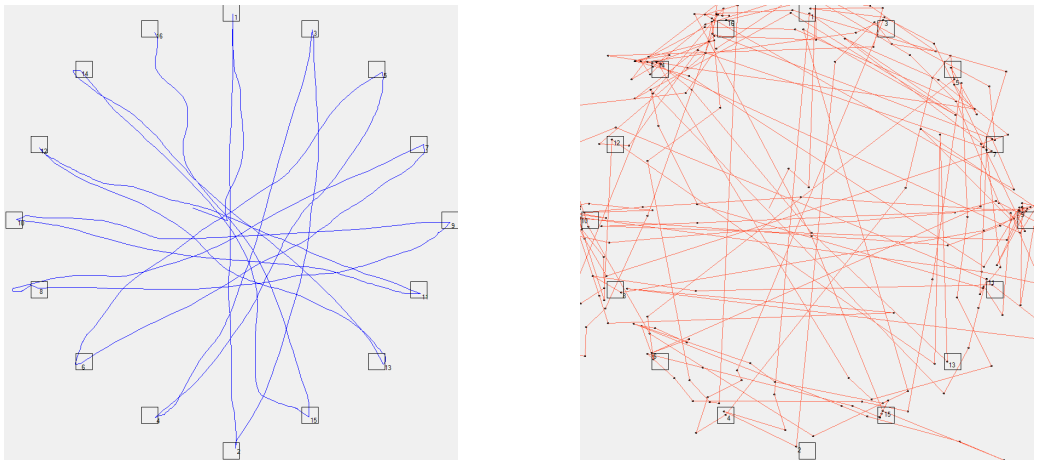


Figure 5.1(a): Mouse path and (b) Eye-tracking (without gravitational well) path of a single participant

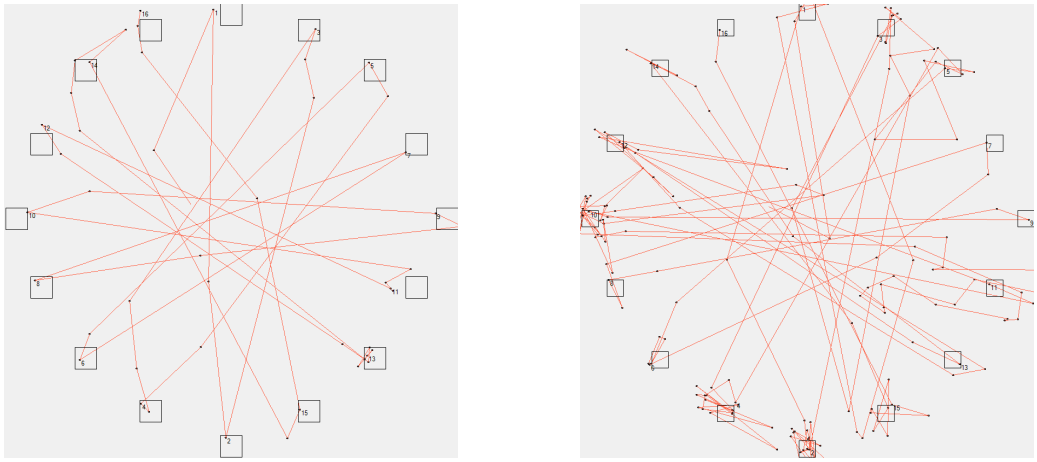


Figure 5.1(c): Eye-tracking (with gravitational well) path and (d) Eye-tracking, with magnification, path of a single participant

The following image set is for the same tasks as above but for another participant. This time it is for a participant who struggled more with the trials. These were also for the first session.

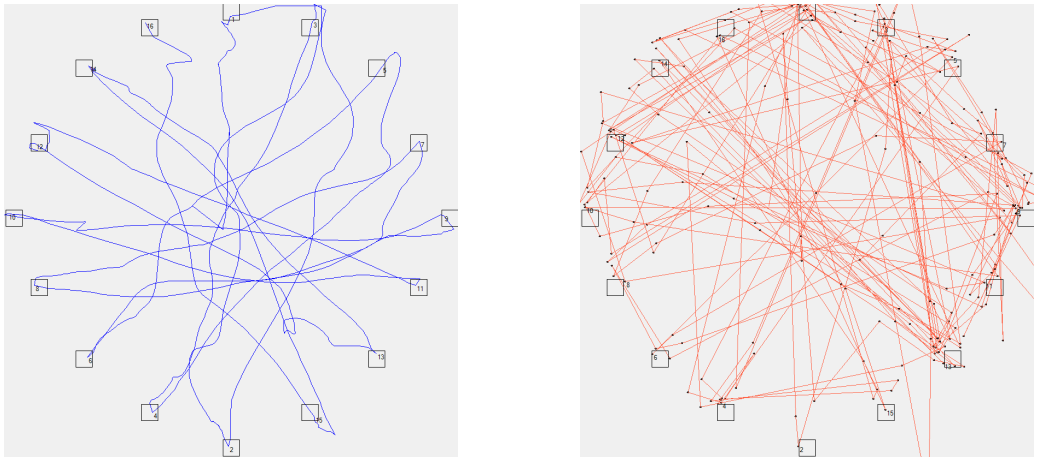


Figure 5.2(a): Mouse path and (b) Eye-tracking (without gravitational well) path of a single participant

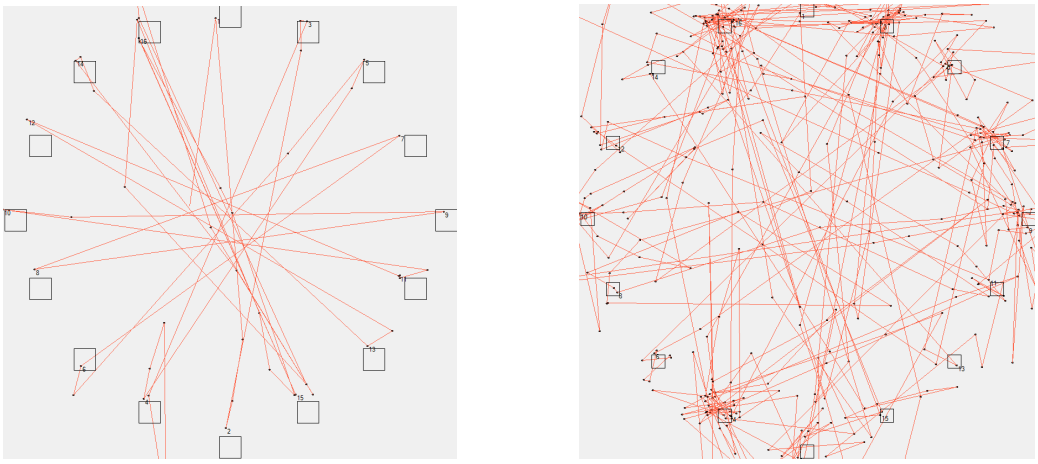


Figure 5.2(c): Eye-tracking (with gravitational well) path and (d) Eye-tracking, with magnification, path of a single participant

These images are representative of the other participants and there was no noticeable difference between the sessions. As can clearly be seen from the figures, the amount of movement needed when using eye-tracking and speech recognition, appears to be substantially more than when using a mouse. The magnification does not seem to lessen the effort required to click on all the buttons and in the case of the second participant, actually seemed to hamper the selection more. The cleaner lines when using the gravitational well appears to almost rival those of the mouse, which would seem to indicate that the gravitational well may enhance the use of the eye-tracking as an interaction technique. These qualitative observations must still be statistically verified.

5.6 Analysis of the throughput

The output of the ISO tests is throughput, as calculated by the formula explained in Section 3.4.2.1. The throughput was calculated for each interaction technique per participant and per session. The first step in the analysis was to determine if M(F) and M(I) could be combined. Following this it will be determined if ETS(F) and ETS(I) can be combined into a single interaction technique. The same analysis will be conducted with the gravitational well interaction technique. Once the allowable combinations have been conducted, the final analysis will be performed on all the remaining interaction techniques. The subsequent section will discuss this analysis in depth.

5.6.1 Combining the interaction techniques

Since the mouse was tested with both an inverted and a framed visual feedback cue, it was decided to first determine if this influenced the participant's mouse throughput. Should this not be the case, then the trials for the mouse could be grouped together for further analysis. The reasoning behind this is that the mouse should be the device which remained fairly consistent throughout the trials. The initial setup of the trials did not allow the throughput to be calculated per mouse interaction technique using the ISO standard for each session. In retrospect, it would be advisable not to have changed the visual feedback cue since it could be argued that this could affect the throughput. Since the mouse was the benchmark interaction technique it was also not entirely necessary to provide different visual feedback cues but rather just to change the size of the targets and subsequently calculate throughput for the mouse using different target sizes. Since all participants were competent with the mouse, it was unlikely that they would "significantly learn" to use the mouse better over the three sessions, therefore it was decided that the throughput could be calculated using the data for the three sessions and then only distinguish between the mouse interaction techniques.

A paired t-test was used to determine if there was a significant difference between throughput for the inverted colour feedback and the framed button feedback. The throughput for these two conditions was calculated over all sessions per participant. The following hypothesis was formulated:

1. H_0: There is no difference between the throughput achieved with the mouse when using inverted colour feedback or framed button feedback.

The normality of the data was first verified using the Shapiro-Wilks normality tests. Since the p-values for framed feedback ($W = 0.97$, $p > 0.05$) and inverted feedback ($W = 0.95$, $p > 0.05$) were larger than the α-value, it could be accepted that the data was normally distributed and a paired t-test could be used for analysis.

The average throughput per participant was calculated over the three sessions for both M(F) and M(I). The mean for M(F) was 4.164 and that of M(I) was 4.170 and the standard deviation was 0.535 and 0.539 respectively. Since $p > 0.05$, the null hypothesis cannot be rejected ($t = 2.14$, $df = 14$, $p > 0.05$). Therefore, the conditions M(F) and M(I) can be combined into a single condition M.

The new groupings are summarised in the table below. If the mouse column is not ticked then the interaction technique is eye gaze and speech and if the framed column is not ticked, then the visual feedback is inverted.

Table 5.1: Grouped interaction techniques

Group	Mouse	Pixel Size	Framed	Magnification	Gravitational well
M	✓	30	✓		
	✓	24			
ETS(F)		30	✓		
		40	✓		
ETS(I)		30			
		40			
ETSG(I)		30			✓
		40			✓
ETSG(F)		30	✓		✓
		40	✓		✓
ETSM		24		✓	
		30		✓	
		40		✓	

Since the throughput of the mouse was not affected by the type of visual feedback given, it was considered worthwhile to determine if the other interaction techniques were affected by the visual feedback. The average throughput for these interaction techniques is tabulated below:

Table 5.2: Average throughput for all interaction techniques prior to consolidation

	Session 1	Session 2	Session 3
ETS(I)	0.633	0.752	0.851
ETS(F)	0.643	0.839	1.027
ETSG(I)	1.813	2.197	2.389
ETSG(F)	1.883	1.994	2.378

Chart 5.1 gives a visual representation of the data.

Since the interaction techniques of ETS(I) and ETS(F) follow the same trend, it was considered worthwhile to determine whether these two interaction techniques differed significantly in terms of throughput. Since the only difference between them was the type of visual feedback, such an analysis could determine whether the type of visual feedback has an impact on the throughput levels achieved. If not, then the two interaction techniques could be amalgamated into a single interaction technique. The same logic applies to the interaction techniques of ETSG(I) and ETSG(F).

In order to determine whether the interaction techniques of ETS(F) and ETS(I) could be combined into a single interaction technique, the following hypothesis was formulated:

1. H_0: There is no difference between the throughput of the different interaction techniques.

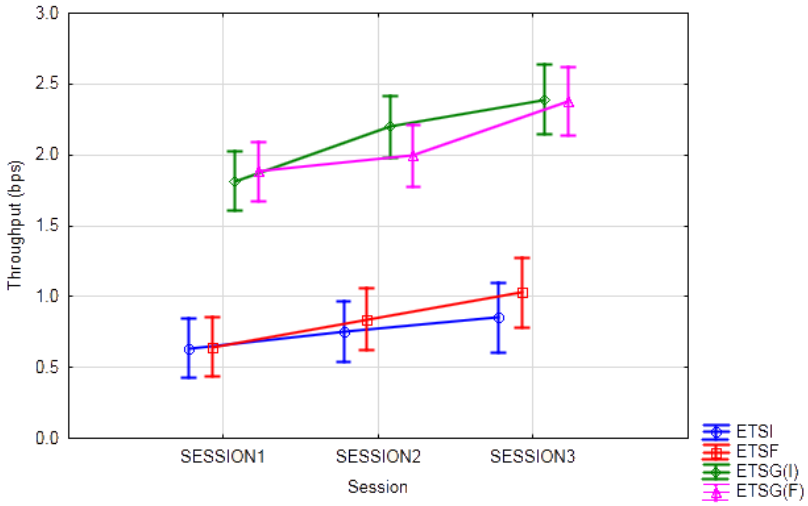


Chart 5.1: Average throughput for all interaction techniques prior to consolidation

Since each participant completed all three sessions, a repeated-measures ANOVA (section 2.5) was required for analysis. Therefore, the normality and sphericity of the data had to be verified before analysis could commence. Shapiro-Wilks was used to test normality and it was found that the data was not normally distributed. The results of the Shapiro-Wilks test are summarised in Table 5.3. According to the Kolmogorov-Smirnov normality test all three sessions were normally distributed and since the ANOVA is robust to violation of normality (Section 3.5), it was decided to continue with the ANOVA analysis. Mauchley's sphericity test confirmed that the assumption of sphericity was met ($\chi^2(2) = 1.846$, $p > 0.05$).

Table 5.3: Results of normality tests for ETS(F) and ETS(I) throughput

Session 1	Session 2	Session 3
$W = 0.880$, $p < 0.05$	$W = 0.908$, $p < 0.05$	$W = 0.879$, $p < 0.05$

A within-subjects repeated measures ANOVA showed that the null hypothesis could not be rejected at an α-level of 0.05 ($F(1, 28) = 0.456$, $p > 0.05$). Therefore, the type of visual feedback did not affect the throughput that could be achieved with the eye gaze and speech interaction technique. Therefore, the two interaction techniques could be consolidated into a single interaction technique which will be called ETS.

The next step was to perform the same analysis for ETSG(F) and ETSG(I). In order to determine whether the feedback affected the throughput achieved when the gravitational well was present, the following hypothesis was formulated:

1. H_0: There is no difference between the throughput of the different interaction techniques.

The normality of the data was confirmed using the Shapiro-Wilks test, the results of which are tabulated below. The data met the assumption of sphericity ($\chi^2(2) = 2.375$, $p > 0.05$) for the repeated-measures ANOVA.

Table 5.4: Results of normality tests for ETSG(F) and ETSG(I)

Session 1	Session 2	Session 3
$W = 0.969$, $p > 0.05$	$W = 0.985$, $p > 0.05$	$W = 0.976$, $p > 0.05$

It was found that the null hypothesis cannot be rejected at an α-level of 0.05 ($F(1, 28) = 0.185$, $p > 0.05$). Therefore, the type of visual feedback does not affect the throughput of the interaction technique.

As a result of not being able to reject H_0, the interaction techniques of ETSG(I) and ETSG(F) can be considered to be one interaction technique. Subsequent analyses need not distinguish between them and all throughput measurements for these techniques will be combined and referred to as ETSG. The subsequent section will provide an in-depth discussion of this analysis.

5.6.2 Analysing throughput

In light of the findings in the previous section, the throughput was recalculated for each participant and for each session, taking into account that M(F) and M(I) as well as ETS(I) and ETS(F) were respectively combined as M and ETS and that ETSG(I) and ETSG(F) were now only ETSG. The underlying averages now apply to the interaction techniques:

Table 5.5: Average throughput for the consolidated interaction techniques for all sessions

	Session 1	Session 2	Session 3
M	3.77	4.17	4.36
ETS	0.52	0.67	0.82
ETSG	1.68	1.90	2.17
ETSM	0.48	0.48	0.61

Chart 5.2 gives a graphical representation of the throughput for the interaction techniques over all three sessions. A 95% confidence interval is superimposed on each point.

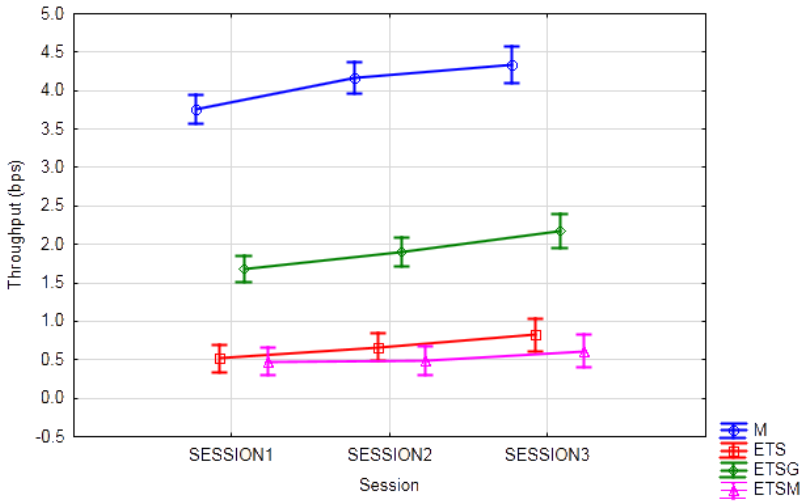


Chart 5.2: Average throughput for consolidated interaction techniques over all sessions

Table 5.6 summarises the results of the normality tests for the data. The sphericity was confirmed ($\chi^2(2) = 0.387$, $p > 0.05$) before analysis commenced in order to inspect the following hypotheses:

1. $H_{0,1}$: The interaction technique has no effect on the throughput achieved.
2. $H_{0,2}$: The session has no effect on the throughput achieved.

The results of the Shapiro-Wilks normality test are tabulated below, with the Kolmogorov-Smirnov also shown where the Shapiro-Wilks failed to verify normality of the data.

Table 5.6: Results of the normality tests conducted on the throughput of all interaction techniques

	Session 1	Session 2	Session 3
Shapiro-Wilks	W = 0.847, p < 0.05	W = 0.841, p < 0.05	W = 0.865, p < 0.05
Kolmogorov-Smirnov	d = 0.180, p < 0.05	d = 0.175, p > 0.05	d = 0.193, p < 0.05

A within-subjects repeated measures ANOVA showed there was significant interaction between the session and interaction technique ($F(6, 108) = 2.598$, $p < 0.05$) therefore separate analyses had to be conducted by isolating each factor in turn.

$H_{0,1}$ could be rejected at an α-level of 0.05 for all three sessions (Table 5.7). During all three sessions, it was only ETS and ETSM that did not differ significantly from each other. Since the mouse has, on average, the highest throughput it can be said that the mouse yields the best throughput of the tested interaction techniques. ETSG differed significantly from both ETS and ETSM and since, on average, ETSG has a higher throughput it implies that ETSG allows for faster, more accurate pointing than the other two interaction techniques.

Table 5.7: Results of separate ANOVA on throughput for consolidated interaction techniques

	Session 1	Session 2	Session 3
ANOVA	$F(3, 56) = 325.024$, $p < 0.05$	$F(3, 54) = 309.927$, $p < 0.05$	$F(3, 56) = 255.637$, $p < 0.05$

$H_{0,2}$ could also be rejected for all interaction techniques (Table 5.8), with the first session and second session differing significantly from the third session for ETS, ETSG and ETSM. When evaluating the mouse, the first session differed significantly from the second and third sessions. The expected average throughput rate of a mouse is between 3.5 and 4.5 bps (Soukoreff & MacKenzie, 2004). Therefore it could be said that the observed values correspond to the expected values, although they are slightly higher for the final two sessions. The fact that even the throughput of the mouse increased would suggest that some improvement could be attributed to a learning effect for the *test* and not the *pointing device*. The use of the Latin Square allows the probability of the learning effect to be negated in terms of preventing one interaction technique outperforming the others by virtue of its position in the test as opposed to its actual usability. Therefore, if the learning effect is to be solely attributed to the users becoming accustomed to the test and not the interaction technique, then the level of improvement should be somewhat consistent for all interaction techniques.

Table 5.8: Results of separate ANOVA on throughput for sessions

	M	ETS	ETSG	ETSM
ANOVA	$F(2, 24) = 10.872$, $p < 0.05$	$F(2, 28) = 4.269$, $p < 0.05$	$F(2, 28) = 14.253$, $p < 0.05$	$F(2, 28) = 5.064$, $p < 0.05$

Inspection of the significant difference indicates that there is not uniform improvement. Consequently, the level of improvement is not similar for the interaction techniques across the sessions and therefore improvement cannot be said to be caused only due to familiarisation with the test but also with the interaction techniques.

It could also be said that the unintentional long gap between session 1 and session 2 did not negatively impact performance. ETSG and ETS also appear to increase at a more rapid rate than ETSM, which signifies that they were easier to learn than ETSM. It would be interesting to increase the number of sessions in order to see

whether a more prolonged exposure could eventually lead to a throughput which is comparable to that of the mouse for one of the interaction techniques or whether they will eventually plateau at a steady throughput after a few sessions.

5.7 Analysis of the time

The next measurement to be analysed was the time taken to complete the trial. Although throughput includes both speed and accuracy it seems prudent to analyse the time taken to complete the trials separately. This is especially important since some of the interaction techniques allow for larger "clickable" areas, which are not visible to the participant. This effectively means that the target can be selected without the eye gaze actually being positioned precisely on the button. This could negatively influence the throughput because of the measurement of the distribution of the click position. Consequently, the time taken for each interaction technique was calculated per session for each participant. This analysis should provide more insight into the usability of the different interaction techniques and serve as confirmation of the throughput results. An analysis was first performed to determine whether there was a possibility of combining interaction techniques, similar to that for the throughput.

5.7.1 Combining the interaction techniques

Similar interaction techniques were first isolated and analysed to determine whether the time for these could be combined. The following procedure was followed to determine if it was allowable to combine the interaction techniques:

- Averages for each session and each interaction technique were calculated.
- Averages were visually inspected for conformity to a general trend.
- If such a trend was identifiable, the normality of the data was investigated. Shapiro-Wilks was used as the preferred normality test, with an additional test being conducted with Kolmogorov-Smirnov if necessary. Since time is rarely normal and the problem can easily be solved by converting the measurements to 1/time, this standard practice was employed in an attempt to normalise the data if the original time measurements were not normal. In the instances where time was converted to 1/time, normality of 1/time was tested again.
- Sphericity of the data was confirmed using Mauchley's sphericity test.
- A within-subject repeated-measures ANOVA was conducted on the data to determine if the interaction technique significantly influenced the throughput yielded during the trials.

This procedure was followed for M(F) and M(I), ETS(F) and ETS(I) as well as for ETSG(F) and ETSG(I). For the sake of brevity, only the final results of the ANOVA will be reported here. The following hypotheses were formulated:

1. $H_{0,1}$: There is no difference between the time required to complete the trials when using M(F) or M(I).
2. $H_{0,2}$: There is no difference between the time required to complete the trials when using ETS(F) or ETS(I).
3. $H_{0,3}$: There is no difference between the time required to complete the trials when using ETSG(F) or ETSG(I).

The results of the ANOVA show that $H_{0,1}$ could not be rejected therefore there was no difference between the time required for the trials when using M(F) and M(I) ($F(1, 24) = 2.530$, $p > 0.05$). Since the null hypothesis $H_{0,1}$ could not be rejected, the visual feedback did not affect user performance in terms of time. The interaction

techniques of M(F) and M(I) can therefore be combined into a single interaction technique of M for the analysis of the time taken to complete the trials.

$H_{0,2}$ could also not be rejected at an α-level of 0.05 ($F(1, 28) = 0.002$, $p > 0.05$) which means the time required for trial completion with ETS(F) did not differ significantly from that for ETS(I).

Similarly, $H_{0,3}$ could not be rejected at an α-level of 0.05 ($F(1, 28) = 0.141$, $p > 0.05$), which implies that ETSG(F) and ETSG(I) did not require significantly different times to complete the trials.

As a consequence of not rejecting $H_{0,2}$, the interaction techniques of ETS(F) and ETS(I) can be combined into a single interaction technique for subsequent time analysis. This interaction technique will be referred to as ETS. ETSG(F) and ETSG(I) can likewise be combined into a single interaction technique, called ETSG for further time analysis.

The conclusion that can be drawn from this analysis is that the visual feedback does not affect the time in which users can complete point-and-click tasks when using eye gaze and speech recognition as an interaction technique. This holds both for when a gravitational well is present or not.

5.7.2 Analysing Time

The average times for the trial completion for all three sessions are tabulated below, with the graphical representation in Chart 5.3 below.

Table 5.9: Average times for consolidated interaction techniques

	Session 1	Session 2	Session 3
M	1.254	1.187	1.155
ETS	9.071	7.556	5.527
ETSG	1.919	1.745	1.576
ETSM	10.226	8.867	6.900

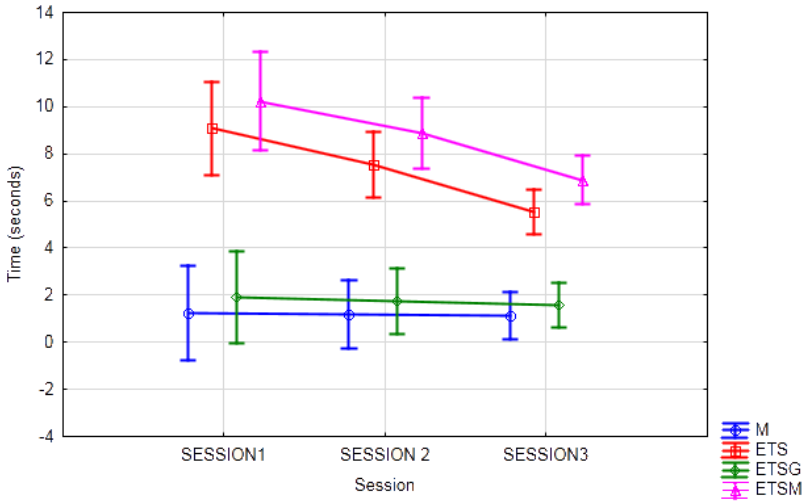


Chart 5.3: Average times for consolidated interaction techniques

From Table 5.9 and Chart 5.3 it can clearly be seen that the mouse and ETSG have the most rapid completion times for the trials. Furthermore, ETSG performs on a level which appears to be comparable to that of the

mouse. Although ETS and ETSM have a much longer completion time, there is noticeable improvement between the times achieved over the sessions.

The session data was tested for normality and it was found that none of the sessions were normally distributed. The times were therefore converted to 1/time and normality was tested again. The results of the Shapiro-Wilks tests for both time and 1/time are tabulated below:

Table 5.10: Results of normality tests on time for consolidated interaction techniques

	Session 1	Session 2	Session 3
Time	$W = 0.786, p < 0.05$	$W = 0.822, p < 0.05$	$W = 0.834, p < 0.05$
1 / Time	$W = 0.884, p < 0.05$	$W = 0.894, p < 0.05$	$W = 0.890, p < 0.05$

Neither time nor 1/time was normally distributed for any of the sessions. Since the ANOVA is robust to violations of normality (Section 3.5) and 1/time measurements are generally "more normalised" than time, the analysis was conducted on the 1/time measurements. Mauchley's sphericity test indicated that the assumption of sphericity was met ($\chi^2(2) = 0.717$, $p > 0.05$). The following hypotheses were formulated for the analysis of the time:

1. $H_{0,1}$: The interaction technique has no effect on the trial times.
2. $H_{0,2}$: The trial times did not differ between the sessions.

$H_{0,1}$ could be rejected since the resultant p-value was less than the significance level ($F(3, 53) = 305.767$, $p < 0.05$). The subsequent conclusion is that the interaction technique significantly affects the time required to complete the trials. Similarly, $H_{0,2}$ could be rejected at a significance level of 0.05 ($F(2, 106) = 24.128$, $p < 0.05$), with the conclusion that the trial session significantly affected the time required to complete the trial.

Post-hoc tests were conducted in order to determine which of the sessions and interaction techniques contributed to the significant difference. Tukey's HSD test indicates that the first session differs significantly from both session 2 and session 3. The second session also significantly differs from the third session. Since the session averages decrease as time went by, it can be concluded that there is an element of learning over time. Therefore, the longer the user is exposed to the interaction technique, the more they learn to use the device and the faster they are able to complete a point-and-click trial.

The times achieved with the mouse differed significantly from all other techniques. The averages indicate that the mouse has lower times, which means the mouse is notably faster than the other interaction techniques. ETSG also differs significantly from the other interaction techniques and closer inspection of the times achieved indicates that the gravitational well significantly decreases the time required to complete the trials when using eye gaze and speech. ETS and ETSM do not differ significantly from each other. Accordingly, while the presence of a gravitational well significantly enhances the performance of ETS, the magnification of targets does not.

According to Chart 5.3, the average times of the mouse remain fairly consistent, while those of ETSG systematically improve over time. ETSG also comes very close to the average times of the mouse. Although analysis shows that the difference between these two interaction techniques is still significant it might be worthwhile to extend the number of trials in order to investigate whether ETSG can ever achieve times which are comparable to the mouse. It would be expected that the times of the interaction technique, similar to the mouse, would eventually reach a fairly constant performance time. Whether this performance time will be higher, the same or less than the mouse will have to be determined. Furthermore, ETS experienced quite a rapid drop in time from session 2 to session 3 and it would be interesting to determine whether this rate of performance improvement will continue over an extended period.

5.8 Analysis of other measurements

Apart from the throughput and time to complete the trials, a number of other measurements can also be used to compare the effectiveness of eye gaze and speech as a pointing device to that of the mouse. These measurements were identified and discussed in Section 3.4.2.1. The measurements deemed appropriate to this study were target re-entries, incorrect target acquisitions, incorrect clicks and time to selection.

5.8.1 Target re-entries

Target re-entries are defined as the number of times the designated target was acquired and then lost before the user was able to click on it.

5.8.1.1 Combining the interaction techniques

Following the same procedure as in the previous analyses (without conversion to 1/measurement), similar interaction techniques were isolated and analysed in order to determine if they could be combined. The following null hypotheses were formulated:

1. $H_{0,1}$: There is no difference between the number of target re-entries for M(F) and M(I).
2. $H_{0,2}$: There is no difference between the number of target re-entries for ETS(F) and ETS(I).
3. $H_{0,3}$: There is no difference between the number of target re-entries for ETSG(F) and ETSG(I).

It was found that $H_{0,1}$ could not be rejected at an α-level of 0.05 ($F(1, 25) = 0.038$, $p > 0$ .05). Therefore, the number of target re-entries did not differ significantly between M(F) and M(I) and these two interaction techniques can be regarded as a single technique in subsequent analyses. $H_{0,2}$ could also not be rejected at a significance level of 0.05 ($F(1, 28) = 0.001$, $p > 0.05$), therefore there is no notable difference between the number of target re-entries for ETS(I) and ETS(F) and they can be combined into ETS for further analysis. Similarly, $H_{0,3}$ could not be rejected as the p-value was larger than the α-level of 0.05 ($F(1, 28) = 0.222$, $p > 0.05$). This result allowed ETSG(F) and ETSG(I) to be consolidated into a single interaction technique of ETSG.

The total number of target re-entries per session and per interaction technique (after consolidation), is shown Table 5.11, together with a number of other descriptive statistics.

Table 5.11 clearly shows that ETSM had a much higher average of target re-entries than any of the other interaction techniques. ETS has a lower average, followed by ETSG and finally the mouse had the least number of target re-entries. Whether these differences are significant will be determined in the following section.

5.8.1.2 Analysis of target re-entries

The final analysis of the target re-entries therefore had the four interaction techniques of M, ETS, ETSG and ETSM. The average number of target re-entries per session and per interaction technique is summarised in Table 5.12.

Table 5.11: Descriptive statistics for the number of target re-entries

		Session 1	Session 2	Session 3
M	Total	111	77	87
	Mean	7.4	5.1	5.8
	Min	3	0	0
	Max	18	10	12
	Std Dev	3.9	3.6	2.9
ETS	Total	1577	1063	646
	Mean	105.1	70.9	43.1
	Min	22	10	3
	Max	364	347	99
	Std Dev	116.9	87.2	28.0
ETSG	Total	294.0	184.0	133.0
	Mean	19.6	12.3	8.9
	Min	6	0	0
	Max	81	37	23
	Std Dev	19.0	9.3	5.5
ETSM	Total	2873	2635	2101
	Mean	191.5	175.7	140.1
	Min	59	96	60
	Max	702	306	347
	Std Dev	166.9	66.0	70.9

Table 5.12: Average target re-entries for consolidated interaction techniques

	Session 1	Session 2	Session 3
M	7.4	5.1	5.8
ETS	110.4	75.2	45.3
ETSG	19.6	12.3	8.9
ETSM	191.5	175.7	140.1

Chart 5.4 shows the plot of the interaction techniques against the session. It can clearly be seen that ETSM has a much higher number of target re-entries than the remainder of the interaction techniques.

For the sake of conciseness, the results of the normality tests will henceforth not be reported even though the tests were conducted in all instances.

A repeated-measures ANOVA was used to test the following hypotheses:

1. $H_{0,1}$: The interaction technique has no effect on the number of target re-entries.
2. $H_{0,2}$: The session has no effect on the number of target re-entries.

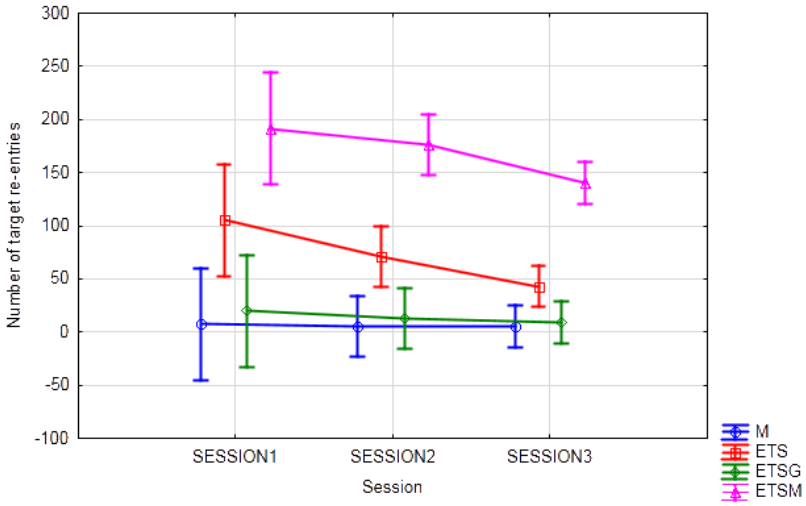


Chart 5.4: Average target re-entries for consolidated interaction techniques

The data failed to meet the assumption of sphericity ($\chi^2(2) = 16.136$, $p < 0.05$), therefore, the adjusted Geisser-Greenhouse and Huynh-Feldt corrections will also be reported for the within-effects variable. These results, as well as the multivariate results, are tabulated below:

Table 5.13: Complete repeated-measures analysis results for consolidated interaction techniques

	ANOVA	Geisser-Greenhouse	Huynh-Feldt	Multivariate
Interaction technique	$F(3, 56) = 32.071$, $p < 0.05$			
Session	$F(2, 112) = 4.249$, $p < 0.05$	$F(1.6, 89.3) = 4.249$, $p < 0.05$	$F(1.7, 96.4) = 4.249$, $p < 0.05$	$F(2, 55) = 3.783$, $p < 0.05$
Interaction technique × Session	$F(6, 112) = 1.003$, $p > 0.05$	$F(4.7, 89.3) = 1.003$, $p > 0.05$	$F(5.2, 96.4) = 1.003$, $p > 0.05$	$F(6, 110) = 0.966$, $p > 0.05$

From Table 5.13 it can be concluded that $H_{0,1}$ could be rejected at an α-level of 0.05. Therefore, the interaction technique plays a significant role in the number of target re-entries. $H_{0,2}$ could also be rejected at a significance level of 0.05, therefore the session did significantly affect the number of target re-entries.

Tukey's HSD was used for post-hoc analysis on the interaction technique. Results indicated that ETSM differed significantly from all other interaction techniques. Since ETSM has the highest number of target re-entries, on average, the use of ETSM can be said to result in significantly more target re-entries. This would imply that it is much harder to achieve a prolonged stable gaze on a button, such that the required verbal command can be issued, when the magnification tool is activated, than for any other interaction technique.

ETS also differs significantly from the mouse and ETSG. ETSG does not differ significantly from the mouse, which means that ETSG is able to perform comparably with the mouse in terms of target re-entries. This would imply that the positioning of an ETSG interaction technique is just as stable as for a mouse. The higher occurrence of re-entries for ETS indicates that the focus slips off the target fairly easily, although there is much improvement over the sessions.

5.8.2 Incorrect target acquisitions

Incorrect target acquisitions are defined as the number of times a target, which is not the designated target, is acquired. This means that in the event of the eye-tracker and speech being used, each time a button receives enough focus to give visual feedback, the incorrect target acquisitions are incremented, provided that the focused button is not the designated target. The number of incorrect target acquisitions are counted as those targets which are acquired *after* the designated target has been acquired. Therefore, the incorrect targets that are acquired cannot be attributed to normal searching for the designated target. For the purposes of this measurement, only the eye gaze and speech interaction techniques will be included in the analysis as the number of incorrect target acquisitions for the mouse interaction techniques were always zero.

Once again, before the all-inclusive analysis was analysed, similar interaction techniques were analysed in isolation to determine the viability of combining them into a single interaction technique.

5.8.2.1 Combining the interaction techniques

The same procedure as with the previous measurements was followed to determine whether the similar interaction techniques could be combined.

The following hypotheses were evaluated:

1. $H_{0,1}$: There is no difference between the number of incorrect target acquisitions when using ETS(F) or ETS(I).
2. $H_{0,2}$: There is no difference between the number of incorrect target acquisitions when using ETSG(F) or ETSG(I).

The null hypothesis, $H_{0,1}$ could not be rejected at an α-level of 0.05 ($F(1, 28) = 0.040$, $p > 0.05$), which means that ETS(F) and ETS(I) can be combined into a single interaction technique ETS. Furthermore, $H_{0,2}$ could also not be rejected ($F(1, 28) =$, $p > 0.05$). Therefore, ETSG(F) and ETSG(I) can be combined into ETSG.

The final conclusion of this analysis is that the visual feedback does not significantly impact the number of incorrect target acquisitions for any of the investigated interaction techniques. Therefore, a complete analysis will use a combined ETS(F) and ETS(I) as well as a combined ETSG(F) and ETSG(I).

Descriptive statistics of the resulting interaction techniques can be seen in Table 5.14.

5.8.2.2 Analysis of incorrect target acquisitions

The next step was to conduct an ANOVA on all the interaction techniques together, now that some of them could be combined in order to simplify the analysis. The averages of the three interaction techniques are shown in Table 5.15 with the graphical representation below that (Chart 5.5).

From these it can clearly be seen that the number of incorrect target acquisitions becomes steadily less as the amount of exposure increases. ETS has a higher number of incorrect target acquisitions than the other two interaction techniques, but it does appear to improve at a faster rate than the other two. These differences must be statistically analysed to determine whether they are significant.

The assumption of sphericity ($\chi^2(2) = 0.302$, $p > 0.05$) was, however, met. Therefore no corrections are required on the repeated-measures ANOVA for the following hypothesis:

H_0: The interaction technique has no significant impact on the number of incorrect target acquisitions.

The results of the ANOVA are summarised in Table 5.16.

Table 5.14: Descriptive statistics for the number of incorrect target acquisitions

		Session 1	Session 2	Session 3
ETS	Total	734	577	345
	Mean	48.9	38.5	23
	Min	21	7	2
	Max	95	80	61
	Std Dev	24.0	23.9	16.9
ETSG	Total	216	178	72
	Mean	14.4	11.9	4.8
	Min	4	0	0
	Max	54	36	12
	Std Dev	12.9	12.0	3.8
ETSM	Total	315	173	187
	Mean	21.0	11.5	12.5
	Min	1	1	0
	Max	84	25	35
	Std Dev	25.2	8.6	11.7

Table 5.15: Average incorrect target acquisitions for consolidated interaction techniques

	Session 1	Session 2	Session 3
ETS	48.9	38.5	23.0
ETSG	14.4	11.9	4.8
ETSM	21.0	11.5	12.5

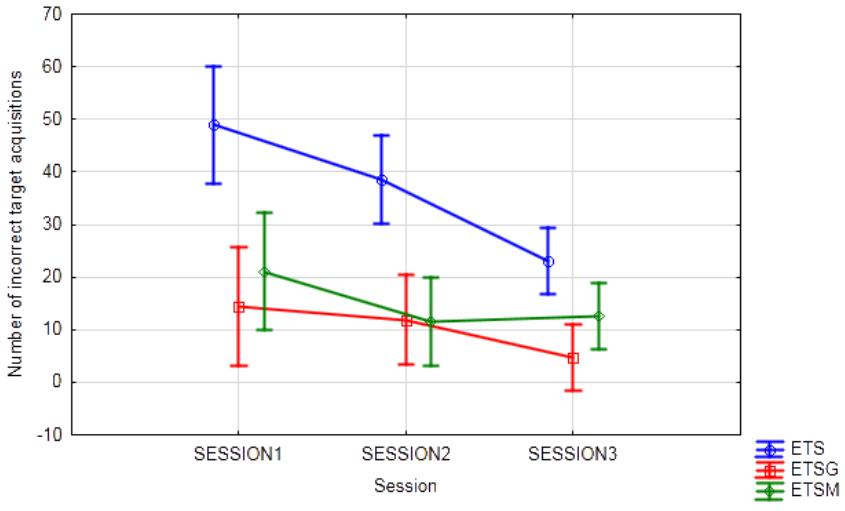


Chart 5.5: Average incorrect target acquisitions for consolidated interaction techniques

Table 5.16: Results of ANOVA on incorrect target acquisitions for consolidated interaction techniques

Factor	ANOVA
Interaction technique	$F(2, 42) = 19.327$, $p < 0.05$
Session	$F(2, 84) = 12.046$, $p < 0.05$
Interaction technique × Session	$F(4, 84) = 2.246$, $p > 0.05$

The results of the ANOVA confirm that there are significant differences between the interaction techniques and the sessions. Tukey's HSD post-hoc test was used to determine which interaction techniques contributed to the significant result in each case.

All the sessions differed significantly from one another. Since only ETSM actually increased slightly in session 3, it can be surmised that the incorrect target acquisitions are lessened at a significant rate over time. ETS in particular has a sharp decrease and it may be beneficial to increase the number of sessions so that it can be properly analysed whether it can ever reach the low values of ETSG or ETSM. In terms of the interaction techniques, ETS differs significantly from both ETSG and ETSM. ETSG and ETSM do not differ significantly from each other.

Observations made of the participants while they were completing the tasks could provide an explanation for this. Many participants soon realised that when struggling to focus on a button it was sometimes easier to focus on another button at a suitable distance from the designated one. It was not necessary to focus on this other button for a protracted time. Participants would then look back at the designated button and the extended movement seemed to provide more accuracy in focusing on the desired target rather than trying to "fine-tune" the selection within a small area around the designated button. The smoothing algorithm could have contributed to this as small movements within a certain radius are interpreted as a single fixation. Since the gravitational well effectively pulls the selection onto the nearest target once the "pointer" is within a certain distance, it becomes easier to focus on a target and no fine-tuning is required. This could explain the reason why ETSG has such a low number of acquisitions compared to ETS.

ETSM also has a lower rate and this could possibly be attributed to participants rather trying to fine-tune the selection when using the magnification. Since the buttons appear larger, participants may have perceived the fine-tuning process to be easier since a larger target could create the impression that it can be easily acquired. The high incidence of target re-entries coupled with the low number of incorrect target acquisitions may serve to substantiate the suspicion that fine-tuning was the preferred method for ETSM.

The similar pattern for ETS, regarding target re-entries and target acquisitions also corroborates the claim that the participants preferred to employ the use of a shifting of their eye gaze to focus on another button and then returning to the designated button. Closer inspection of the averages for ETS shows that incorrect target acquisitions constituted approximately half the number of re-entries for each session. This could indicate that participants would attempt to re-acquire the designated target and when they were unable to achieve a stable selection, resorted to focusing on another target before attempting to select the designated target – in contrast to the strategy employed with ETSM.

The reason for this could be that the magnification disturbs the users while they adjust their gaze and they are unwilling to move their gaze substantially because they perceive this to require more effort when magnification is activated. Another reason for the different strategies could be attributed to the fact that the magnification tool that was used has in-built visual feedback which allows the user to get an approximation of their eye gaze position, which is centred in the magnified area. Since this feedback is present, the user may

feel that fine-tuning is a better option since they can determine how close they are to the target, which is not the case with ETS. With ETS they will know they have lost the target but not how close they are to re-acquiring it, hence they feel more secure glancing at another target, establishing position and then looking at the required target again until they can maintain a stable eye gaze. Therefore, to slave a cursor to the eye gaze may be disruptive but in this instance it could tentatively be said that it may have provided useful information to the users. However, the evidence suggests that it in no way increased the efficiency or effectiveness of target selection and therefore it is not recommended for use.

The average number of target re-entries for ETSG was roughly the same as the average incorrect target acquisitions for ETSG. This could provide evidence that when using ETSG, the target was easier to acquire and keep the focus long enough to issue the required command. Since the buttons were effectively larger it would make sense that they were easier to focus on for a prolonged period of time.

5.8.3 Incorrect clicks

Incorrect clicks are determined as the number of times a target that was not the designated target was clicked during a trial.

5.8.3.1 Combining the interaction techniques

Following the same procedure as the preceding sections, similar interaction techniques were first inspected on their own to determine whether they could be combined for further analysis.

1. $H_{0,1}$: The number of incorrect clicks is not significantly different between M(F) and M(I).
2. $H_{0,2}$: The number of incorrect clicks is not significantly different between ETS(F) and ETS(I).
3. $H_{0,3}$: The number of incorrect clicks is not significantly different between ETSG(F) and ETSG(I).

$H_{0,1}$ could not be rejected ($F(1, 25) = 0.706$, $p > 0.05$). Therefore, the number of incorrect clicks is not significantly affected by the type of feedback given with a mouse. Subsequently, M(F) and M(I) will no longer be distinguished between and a single interaction technique of M will be used.

Similarly, the null hypothesis for ETS(F) and ETS(I) could not be rejected at a significance level of 0.05 ($F(1, 28) = 0.998$, $p > 0.05$). Therefore, these two interaction techniques could be combined into ETS for the number of incorrect clicks.

At a significance level of 0.05, $H_{0,3}$ could not be rejected ($F(1, 28) = 0.183$, $p > 0.05$). Therefore, the difference between the number of incorrect clicks for ETSG(F) and ETSG(I) do not differ significantly and they can be combined into a single technique, called ETSG.

As a result of these findings, there were now only four interaction techniques. Descriptive statistics are given for these interaction techniques in Table 5.17.

5.8.3.2 Analysis of incorrect clicks

The final analysis of the number of incorrect clicks included the four interaction techniques M, ETS, ETSG and ETSM, the averages of which are tabulated in Table 5.18.

Table 5.17: Descriptive statistics for the number of incorrect clicks

		Session 1	Session 2	Session 3
M	Total	2	3	2
	Mean	0.1	0.2	0.1
	Min	0	0	0
	Max	1	1	1
	Std Dev	0.4	0.4	0.4
ETS	Total	21	24	15
	Mean	1.4	1.6	1
	Min	0	0	0
	Max	4	7	2
	Std Dev	1.2	1.8	0.8
ETSG	Total	63	40	23
	Mean	4.2	2.7	1.5
	Min	0	0	0
	Max	9	6	5
	Std Dev	2.7	2.0	1.4
ETSM	Total	10	17	13
	Mean	0.7	1.1	0.9
	Min	0	0	0
	Max	2	4	4
	Std Dev	0.8	1.3	1.2

Table 5.18: Average number of incorrect clicks for consolidated interaction techniques

	Session 1	Session 2	Session 3
M	0.1	0.2	0.1
ETS	1.4	1.6	1.0
ETSG	4.2	2.7	1.5
ETSM	0.7	1.1	0.9

Chart 5.6 plots the averages of the interaction techniques against the sessions.

Almost surprisingly, it is ETSG that has the highest average number of incorrect clicks of all the interaction techniques. Due to the fact that ETSG had the lowest number of incorrect target acquisitions this observation might be considered unexpected.

The following hypothesis was formulated to determine whether the interaction technique affected the number of incorrect clicks:

H_0: The number of incorrect clicks is not significantly influenced by the interaction technique.

The assumption of sphericity ($\chi^2(2) = 4.157$, $p > 0.05$) was met. The within-subjects, repeated-measures ANOVA indicated that there was significant interaction between the two factors ($F(6, 112) = 4.689$, $p < 0.05$). Therefore, the factors had to be examined in isolation so as to control for the other factor. Since the factor of

interest is the interaction technique, the session was controlled for and three separate ANOVAs were run in order to test for significant differences. The results of these ANOVAs are summarised in Table 5.19.

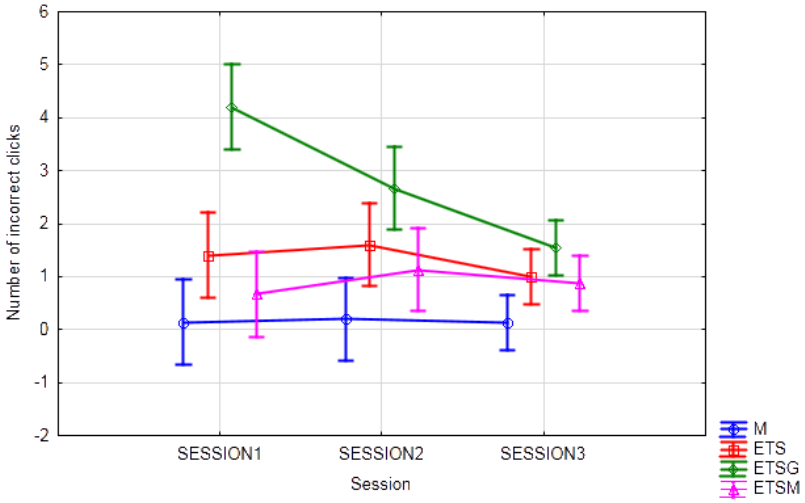


Chart 5.6: Average number of incorrect clicks for consolidated interaction techniques

Table 5.19: Results of separate ANOVA on incorrect clicks for consolidated interaction techniques

	Session 1	Session 2	Session 3
ANOVA	F(1, 3) = 20.362, p < 0.05	F(1, 3) = 6.953, p < 0.05	F(1, 3) = 4.890, p < 0.05

As can be seen from the results in the table, the interaction techniques resulted in significantly different number of incorrect clicks for all three sessions. Post-hoc analysis was conducted next so that the significant differences could be accounted for.

During session 1, ETSG differed significantly from all other techniques. Since the number of incorrect clicks was highest for ETSG, it can be concluded that ETSG caused significantly more incorrect clicks during the first session. During the second session, ETSG differed significantly from the mouse and ETSM and in the third session only from the mouse. These results clearly show that ETSG results in the highest number of incorrect clicks. Although continued practice allowed ETSG to have a comparable number of incorrect clicks to ETS and ETSM, its performance could not match that of the mouse over the three sessions. This indicates that some learning did take place over the three sessions.

Natural eye movement may provide an explanation for the observed difference. Participants could acquire the target and then issue a verbal command while already starting to look at the next target (for all eye gaze and speech interaction techniques). Since the use of the gravitational well increases the speed with which a target can be acquired, this often meant that by the time the speech engine recognised the command, the next target had already been acquired. This could account for the high number of incorrect targets for ETSG. These findings also confirm previous findings that the fixation immediately prior to the action or command being issued is usually occurs on the object of interest (Land & Tatler, 2009; Maglio et al., 2000)

Since ETSG had significantly lower incorrect target acquisitions coupled with this finding of more incorrect clicks creates the following dilemma. The use of the gravitational well increases the possibility of correctly

acquiring a target and maintaining a stable gaze on the target. This is evidenced by the fact that other eye gaze and speech interaction techniques caused participants to first glance away, acquire another target and then glance back. However, the fact that a gravitational well is present together with human tendency to start glancing at the next object of interest whilst still issuing a command to the current target, means that the next target is acquired far quicker than when no gravitational well is present. This causes the next target to be incorrectly clicked on with higher frequency for ETSG. Since participants started moving their eye gaze away from the buttons before the speech command had been executed for all eye gaze interaction techniques, it would be assumed that for ETS and ETSM, which pose greater difficulty in target acquisition, the participant would inadvertently have caused a click somewhere on the application form which was not a clickable area. This would correspond to the measurement of number of missed clicks as discussed in Section 3.4.2.1. Unfortunately, this measurement was not captured during these tests. Further research must be done in order to determine if this proposition is true.

5.8.4 Time to selection

Time to selection is measured as the time between when the final target acquisition is performed and when the target is actually clicked or selected. The final target acquisition is defined as the last time the button receives focus before being clicked. The same procedure as with the other measurements was followed. All data was measured in milliseconds and then converted to 1/time to optimise the possibility of normalisation.

5.8.4.1 Consolidating the interaction techniques

The following hypotheses were formulated:

1. $H_{0,1}$: There is no difference between the time to selection when using M(F) and M(I).
2. $H_{0,2}$: There is no difference between the time to selection when using ETS(F) and ETS(I).
3. $H_{0,3}$: There is no difference between the time to selection when using ETSG(F) and ETSG(I).

$H_{0,1}$ could be rejected at an α-level of 0.05 ($F(1, 25) = 0.473$, $p > 0.05$), therefore M(F) and M(I) can be considered as a single interaction technique for the purposes of the time to selection analysis. Similarly, both $H_{0,2}$ ($F(1, 28) = 0.647$, $p > 0.05$) and $H_{0,3}$ ($F(1, 28) = 0.406$, $p > 0.05$) could be rejected with the result that ETS(F) and ETS(I) can be combined into ETS and ETSG(F) and ETSG(I) into ETSG.

Table 5.20 tabulates the descriptive statistics for the four resulting interaction techniques.

Table 5.20: Descriptive statistics for time to selection

		Session 1	Session 2	Session 3
M	Mean	321.8	324.7	321.3
	Std Dev	48.3	93.1	77.6
ETS	Mean	1154.5	1187.5	1136.6
	Std Dev	216.1	151.4	107.9
ETSG	Mean	1097.7	1104.5	1011.4
	Std Dev	134.5	99.3	110.8
ETSM	Mean	1093.8	1123.4	1036.2
	Std Dev	213.4	129.0	125.2

5.8.4.2 Analysis of time to selection

Table 5.21 provides a summary of just the averages of the time to selection, with Chart 5.7 below that giving a graphical depiction of the averages.

Table 5.21: Average time to selection

		Session 1	Session 2	Session 3
M	Mean	321.8	324.7	321.3
ETS	Mean	1154.5	1187.5	1136.6
ETSG	Mean	1097.7	1104.5	1011.4
ETSM	Mean	1093.8	1123.4	1036.2

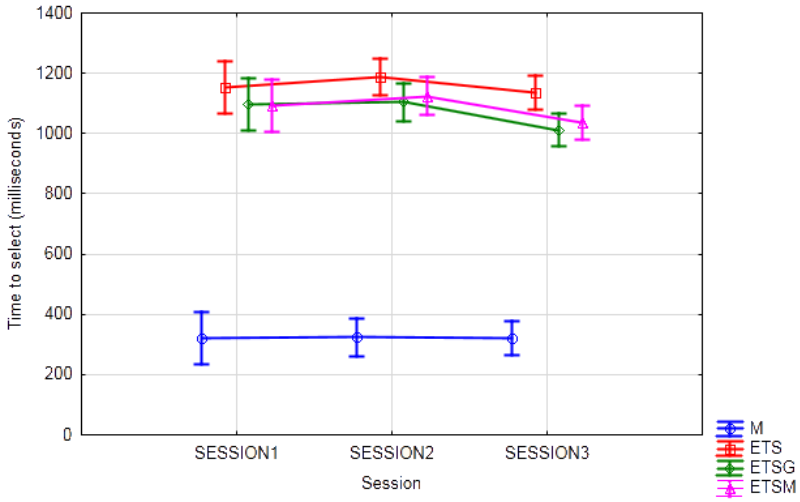


Chart 5.7: Average time to selection

As can be expected, the interaction techniques of ETS, ETSG and ETSM all have similar times to selection. Since they all have the same selection device, i.e. speech it seems appropriate that they maintain similar averages. On average, it took participants approximately 300 milliseconds to select a target with the mouse once it had been acquired. The eye gaze and speech techniques averaged a selection time of over 1 second.

The following hypothesis was inspected:

H_0: The interaction technique has no effect on the time to select an acquired target.

The assumption of sphericity ($\chi^2(2) = 8.665$, $p < 0.05$) was also violated, therefore the adjusted corrections will be reported where applicable.

Table 5.22: ANOVA results of time to selection

	ANOVA	Geisser-Greenhouse	Huynh-Feldt	Multivariate
Interaction technique	$F(3, 54) = 196.605$, $p < 0.05$			
Session	$F(2, 108) = 0.543$, $p > 0.05$	$F(1.7, 93.9) = 0.543$, $p > 0.05$	$F(1.9, 102.1) = 0.543$, $p > 0.05$	$F(2, 53) = 0.869$, $p > 0.05$
Interaction technique × Session	$F(6, 108) = 0.909$, $p > 0.05$	$F(5.2, 93.8) = 0.909$, $p > 0.05$	$F(5.7, 96.4) = 102.1$, $p > 0.05$	$F(6, 106) = 0.586$, $p > 0.05$

From the results in Table 5.22, $H_{0,1}$ could be rejected meaning that the interaction technique does have an effect on the selection time. Post-hoc tests indicate that the mouse differs significantly from all other techniques. The selection time for the mouse is, on average, lower than those for the other interaction techniques; therefore the selection time for the mouse is significantly faster than selection times for the other interaction techniques. This result has serious implications for the acceptance of eye gaze and speech as an interaction technique since it shows that even if the final acquisition can be performed in a comparable time to the mouse, thereafter the time to select will still take significantly longer. There is no noticeable improvement in the time to select over the session which is not surprising since this factor hinges on the issuing on a verbal command. The chance that a participant can improve the speed at which they utter a command, in reaction to a selection, is highly improbable.

5.8.4.3 Further analysis of selection times

This discovery led to the question being posed as to whether the final acquisition of the target differed significantly between the interaction techniques. Inspection of the overall trial times showed that only ETSG averaged in the region of the mouse. Therefore, this analysis was confined to the interaction techniques of the mouse and eye gaze and speech with a gravitational well. The time in milliseconds to achieve a final acquisition of the designated target was calculated for M(F), M(I), ETSG(F) and ETSG(I). The final acquisition was determined as the acquisition immediately prior to selection of the designated target. Analysis showed that M(F) and M(I) could be combined into M ($F(1, 25) = 3.123$, $p > 0.05$). Similarly, ETSG(F) and ETSG(I) could be combined into ETSG ($F(1, 28) = 0.174$, $p > 0.05$).

Descriptive statistics for M and ETSG are tabulated below:

Table 5.23: Descriptive statistics for final acquisition times

		Session 1	Session 2	Session 3
M	Mean	917.4	855.8	826.6
	Std Dev	104.6	113.5	130.5
ETSG	Mean	817.4	640.0	564.5
	Std Dev	246.6	204.9	209.6

Chart 5.8 clearly shows that, on average, ETSG has a *lower* final acquisition time than the mouse.

The following hypothesis can be formulated:

H_0: The interaction technique has no effect on the time to final target acquisition.

The data conformed to the assumption of sphericity ($\chi^2(2) = 0.415$, $p > 0.05$). A within-subjects repeated-measures ANOVA showed that there was significant interaction between the two factors ($F(2, 54) = 4.155$, $p < 0.05$), therefore separate ANOVAs were conducted where the session was controlled for. These results are summarised in Table 5.24.

According to Table 5.24, there was a significant difference between the interaction techniques for all three sessions. ETSG had a lower average final target acquisition in all three sessions; therefore ETSG is significantly faster in terms of final target acquisition than the mouse.

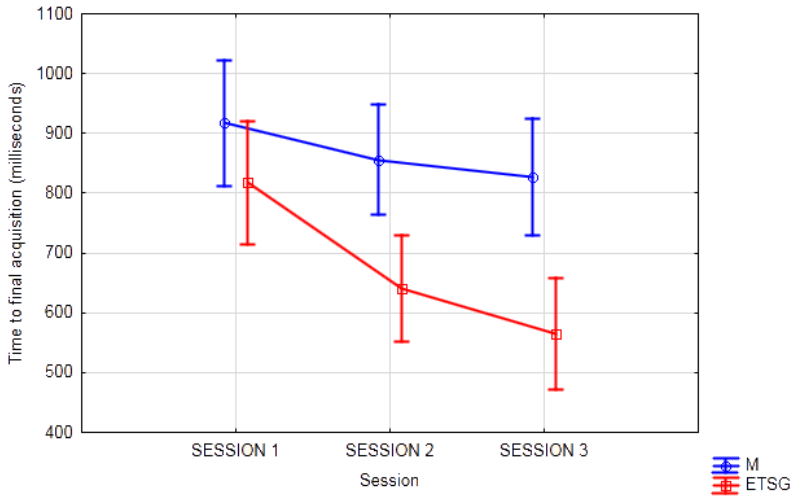


Chart 5.8: Average time to final selection for M and ETSG

Table 5.24: Separate ANOVA results for final target acquisition

Session 1	Session 2	Session 3
F(1, 28) = 4.627, p < 0.05	F(1, 27) = 10.875, p < 0.05	F(1, 28) = 19.511, p < 0.05

For overall time to target selection, the mouse is significantly faster than ETSG. However, when selection time is divided into final target acquisition and time to selection, it was found that ETSG has a significantly faster final target selection but a significantly slower time to selection. Therefore, the time to selection is so much slower that the overall time differs significantly. Section 5.7.2 posed the question as to whether ETSG could, over time, achieve the same speeds as the mouse. It would now seem that final target acquisition would have to improve dramatically to achieve this. Since acquisition times did improve over time, this remains a viable possibility for improved overall selection times. An additional option could be to explore another selection type, such as using look-and-shoot with the press of a keyboard key. This method could possibly provide a faster time to selection than the uttering of a verbal command.

5.9 Subjective device assessment

The final measurement which will be evaluated is that of subjective satisfaction. Since the subjective feeling of users does not necessarily mirror their actual performance, it is imperative that both be tested. Participants each completed the post-test questionnaire (Appendix E) which focused on the use of the eye-tracker and speech recognition to select targets. The questionnaire results are tabulated below. Each response was rated on a 5-point scale. The responses in Table 5.25 are grouped according to the number of responses for the lower range of the scale (1 and 2), the neutral or midpoint of the scale (3) and the higher range of the scale (4 and 5).

Nine participants felt that the force required to move the device was neither too low nor too high, but the average score indicates that the force required might be slightly high. The majority of the participants felt that the movement of the device was a little rough and that the mental effort required was too high. While physical effort may be low, accurate pointing is difficult and the operation speed is too fast. Neck fatigue was experienced by only a small number of the participants and the device appears to be fairly comfortable to use. In summary, it seems as though the use of the eye-tracker and speech recognition is relatively easy.

Furthermore, ten of the participants felt that with enough practice they could match the mouse speeds when using eye gaze and speech. Thirteen indicated that they enjoyed using eye gaze and speech recognition as a pointing device. Only five felt that magnification enhanced the use of the mouse, while ten participants felt that the magnification made the use of eye gaze more difficult.

In terms of the preferred appearance of the buttons, 11 participants preferred the large buttons and nine preferred the framed visual feedback over the inverted colour visual feedback.

Table 5.25: Results of the device assessment questionnaire

Question	Answer group	Number of answers	Mean	Standard deviation	Mode
Actuation force	Low	1	3.3	0.6	3.0
	Neutral	9			
	High	6			
Smoothness	Rough	6	2.9	1.1	2.0
	Neutral	4			
	Smooth	5			
Mental effort	Low	2	3.5	0.8	4.0
	Neutral	5			
	High	8			
Physical effort	Low	5	2.9	0.8	3.0
	Neutral	6			
	High	4			
Accurate pointing	Easy	3	3.6	1.1	5.0
	Neutral	4			
	Difficult	8			
Operation speed was	Fast	7	2.9	1.0	2.0
	Neutral	2			
	Slow	6			
Neck fatigue	None	11	1.8	1.0	1.0
	Neutral	3			
	High	1			
General comfort:	Uncomfortable	5	3.1	1.2	2.0
	Neutral	4			
	Comfortable	6			
Overall, the input device was	Difficult	4	3.3	1.1	4.0
	Neutral	2			
	Easy	9			

5.10 Summary of findings

This chapter had a large amount of analysis and this section will attempt to summarise all the findings. It was found that the type of visual feedback did not affect the performance of the interaction techniques with

regard to any of the measurements that were analysed. Therefore, even though users may prefer a certain type of visual feedback, their performance will not be impacted by the choice of visual feedback.

The mouse has a significantly higher throughput than the other interaction techniques. The use of a gravitational well causes a significant improvement to the throughput of eye gaze and speech as an interaction technique. Magnification does not positively influence the throughput of eye gaze and speech as an interaction technique.

The mouse is also significantly faster than the other interaction techniques and the use of a gravitational well causes a significant decrease in point-and-click time for eye gaze and speech. Furthermore, magnification of the targets does not increase the time performance of eye gaze and speech.

In terms of the other measurements that were analysed, the following discoveries were found:

1. The use of magnification causes a significant increase in the number of target re-entries made while selecting a target. This implies that when magnification is activated, it becomes harder to maintain a stable gaze on a target. This is despite the fact that the target is essentially much larger than with any other interaction technique. The high number of target re-entries, together with the relatively low number of incorrect target acquisitions suggest a desire by the participants to fine-tune the selection, possibly due to the impression that the larger the target is, the easier it is to acquire.
2. The use of a gravitational well increased the number of incorrect clicks, possibly due to the fact that the eye already starts moving to the next target while the speech command is issued. While the number of incorrect clicks does not improve, it was able to be comparable to the mouse in only three sessions. Target re-entries and incorrect target acquisitions were kept very low with this interaction technique. Moreover, the averages between target re-entries and incorrect target acquisitions remain approximately the same, providing further evidence of the assertion (when compared to other techniques) that maintaining a stable gaze is easier. The very presence of a gravitational well could well be a “double-edged sword” in the sense that it becomes easier to maintain a stable gaze but it also causes more incorrect clicks since subsequent targets are also easier to acquire – so much so that they are acquired before the verbal command is completely issued or processed.
3. ETS has significantly more incorrect target acquisitions that ETSM or ETSG.
4. ETS had roughly double the number of target re-entries compared to incorrect target acquisitions. In comparison, ETSM had a high number of target re-entries but much lower incorrect target acquisitions. This suggests that participants employed different strategies when attempting to select with ETS and ETSM. When using ETS, it would appear that participants prefer to look a distance away from the designated target, quite often inadvertently or purposefully acquiring another target. They will then look back at the designated target in order to attempt a selection. Conversely, with ETSM the high number of target re-entries indicates a high incidence of target slippage but the relatively low number of incorrect target acquisitions points to a method of fine-tuning for selection purposes.
5. ETSG has a significantly faster time to final acquisition but a significantly slower time to selection. Overall, the negative impact of the time to selection causes ETSG to be significantly slower than the mouse. However, this remains a promising discovery and possible continued practice may increase the final acquisition time, as already evidenced in the three sessions conducted. Another alternative is to provide a different means of selection, such as look-and-shoot coupled with the press of a keyboard key.

Although designed to alleviate the strain of finely focusing on small targets, the magnification tool required perhaps the most concentration and was unnatural for the majority of the participants. This could perhaps be the reason behind its poor performance against the other interaction techniques. The swift reaction of the eye gaze when employing the gravitational well could be expected by the participants as people are accustomed to rapid focusing. Additionally, the presence of peripheral vision together with the use of a physical interaction

technique negates the need for prolonged and finely tuned focusing under normal circumstances. The higher performance is undoubtedly directly related to the fact that the selectable area is much larger than with the other interaction techniques and it facilitates a smoother selection regardless of the stability of the eye gaze. Since users are not aware of their fine eye movements, the gravitational well is perhaps the interaction technique which most closely resembles the expectations of the user in terms of their perceived behaviour. The gravitational well also inspires more confidence in the users as they are unaware of the larger selectable area but they are achieving the desired results with minimal effort. It also allows for a more aesthetically pleasing interface as the widgets are kept to a smaller size, although they must be spaced further apart to make provision for the gravitational well.

These findings confirm to an extent previous findings (Ashmore et al., 2005) in the sense that omnipresent magnification does not perform as well as other pointing techniques. The GHA fisheye lens used in the study of Ashmore et al. (2005) also requires that the user fine-tune the selection of the target within a magnified area. However, this still facilitates better pointing than an omnipresent fisheye lens. The reason for this and for the performance of ETSM could be the disruption of the visual search caused by the omnipresent magnification. The current study's results also confirm those of Ashmore et al. (2005) that omnipresent magnification and no magnification have equivalent selection times for eye gaze.

Incorrect clicks were experienced with all eye gaze interaction techniques although more so with ETSG. Nevertheless, this finding corresponds with the finding of Kaur et al. (2003) that the target which was acquired a certain amount of time prior to command execution, is the target that must be selected. Although the interval was found to be 630 milliseconds (Kaur et al. 2003) this interval will have to be confirmed for use with eye gaze and speech. While natural eye gaze movement appears to dictate that the target prior to command utterance must be selected, it must still be determined whether this will appear natural to the user or whether they would prefer to adapt to the use of ETSG as it was tested in this study. Clearly, practice allows them to adjust their natural behaviour to a degree to compensate for the interaction technique as is evidenced by the improvement over the sessions. However, requiring users to change their natural behaviour is not the aim of a multimodal interface. Therefore, it becomes necessary to establish the interval required for target selection and test the usability of that compared to the standard gravitational well employed in this study.

Previous studies such as the touch sensitive mouse and MAGIC pointing warped the mouse pointer to the position of the eye gaze and then users were required to use the mouse pointer to click on the desired target. Although this exploits the high speed of eye gaze and also reduces incidences of incorrect clicks since users are not likely to click on the incorrect target when having to manually manipulate a mouse pointer, some physical dexterity is required. The solution may lie in a combination of this technique and speech. Eye gaze could be used to establish intent, a single voice command could be issued to warp the pointer to the selectable target closest to the current eye gaze and once the user has verified that the correct target is acquired, a second command can be issued to click on the target. For fine-tuning purposes of the mouse cursor, direction- or target-based navigation can also be provided.

In terms of comparison with previous studies, not many previous studies compared eye gaze selection with the ISO test and certainly none on this scale. The closest comparison would be with the look-and-shoot tests since eye gaze and speech could be considered look-and-shoot. Throughput for ETSG was much lower (2.31 bps) than the look-and-shoot using the space bar (3.78 bps). The accuracy of the speech engine could have played a significant role in this instance and it may be worthwhile investigating this supposition using a Wizard of Oz study to determine whether it can compete with dwell time and using look-and-shoot with a relatively error free activation mechanism such as a key press. In terms of selection time, ETSG averaged approximately 1000 ms while acquisition time was approximately 500 ms in the third session of the ISO test. Using the ISO test it was suggested that a dwell time of 500 ms (Zhang & MacKenzie, 2007) was the most appropriate. If one assumes that target acquisition will be similar then the speech takes double the time of using the dwell time. It

can therefore, be concluded that using speech may be less efficient than using dwell time although studies must be conducted to verify this.

5.11 Further research

Further research can be conducted for interaction techniques using the ISO pointing device test. Future experimental setups will exclude different feedback techniques and rather concentrate on changing the distance between targets and the size of targets. In this way, more trials with a single interaction technique can be added to each session without extending the time required for the session. This may also yield interesting results since more measurements per interaction technique will be available. The magnification can also be excluded since it clearly does not yield better throughput, increased speed or fewer target re-entries or other factors. More sessions may provide deeper insight into the effects of learning, particularly on ETSG. Additionally, other selection means may be considered as a way to counteract the significantly slower time to selection of the speech commands.

A more thorough examination of subjective satisfaction can also be made by requiring the device assessment to be completed for multiple interaction techniques. Furthermore, subjective satisfaction could be measured after the first exposure and then again after the final exposure in order to determine whether there is a shift in satisfaction after prolonged usage of the pointing devices.

Additionally, the results obtained could be specific to the eye-tracker used and results of ETS could be significantly different if an eye-tracker with higher accuracy and precision was used. Similarly, the gravitational well could be rendered superfluous under these conditions. Further research can be conducted whereby different eye-trackers are compared with one another in this regard.

5.12 Summary

This chapter reported on the analysis of the ISO testing which was conducted with the mouse, eye gaze and speech, eye gaze and speech with a gravitational well and also with magnification. Overall it was found that the visual feedback does not impact on performance measures in any way. Furthermore, the mouse remained the most effective means of point-and-clicking. However, the use of a gravitational well significantly increased the performance measures which can be achieved with eye gaze and speech, particularly in terms of time to acquire the designated target. Overall, the use of a gravitational well appears to provide a promising means of increasing the performance of eye gaze and speech as a pointing device. Conversely, the use of a magnification tool appears to hamper the performance of eye gaze and speech significantly since it does not provide enough stability. It remains to be seen, however, whether a more extended use of the interaction techniques can further increase the use of these interaction techniques or whether they will eventually reach a plateau of performance levels. The following chapter will report on the results of testing eye gaze and speech in Microsoft Word®.

CHAPTER 6

ANALYSIS OF SPEECH COMMANDS IN WORD

6.1 Introduction

The previous chapter explored the possibility of eye gaze and speech as a replacement for the mouse as a pointing device. It was found that when using a gravitational well, it was possible to achieve improved performance with eye gaze and speech. Additionally, the use of magnification did not enhance the use of eye gaze and speech as a pointing device. This chapter will focus on the use of a speech interaction technique within a word processing application. In particular, a number of tasks will be compared when using the traditional means of input in a word processor and when using speech as an alternative. Since the interaction techniques will be used in a well-known application, the environment will be familiar to the participants and the feasibility of using speech to complete tasks can be investigated thoroughly. A number of usability measurements will be identified and analysed and the results of the analysis will be discussed in detail.

6.2 Procedure

The longitudinal testing was conducted over a ten week period. For the purposes of this thesis, longitudinal testing refers to the fact that a series of tests were repeated and conducted over a period of time. Each participant attended one session per week at the same time and on the same day of the week. There were isolated incidents where the participant could not attend his/her scheduled session for various reasons and he/she was then accommodated in another session, which was not too close to his/her next session and not too far from his/her previous session. Of course, as students are prone to do, some did not make alternative arrangements and simply did not attend some of their sessions. Therefore, there were some weeks where not all 25 students participated. The students were paid a cash incentive for each session that they attended.

During the first session, participants completed the pre-test questionnaire (Appendix F). Following this, they each trained their speech profile using the Microsoft speech training wizard. The training wizard requires that the user read a large amount of pre-defined text. The wizard then attempts to recognise the spoken words and in so doing, build a profile for the reader based on their pronunciation and enunciation of the words. The participants were then introduced to the multimodal Word that they would be using for the next few weeks. They were also given a brief tutorial of the speech grammar which was available for use in Word (Table 3.1). 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. Once they were comfortable with the application, they were given an explanation regarding what the next few sessions would consist of. To conclude the first session, each participant completed the post-test questionnaire as shown in Appendix G.

Every subsequent session had the same procedure which was followed, which was to complete the tasks as set out in Section 3.4.3.3 (Table 3.5). Each individual task was displayed to the participant using a small window overlaid on the word processor. The window did not obstruct any part of the word processor. The participant had to complete as many tasks as they could manage in their half hour slot. After completion of their tenth session, all participants completed the more comprehensive post-test questionnaire (Appendix H). Therefore, according to this setup each participant could complete a maximum of nine tests with the application.

6.3 Participants

In total there were 25 students who participated in the longitudinal study. They were all undergraduate students who were completing their studies at the University of the Free State. Therefore, the sampling technique used for this study was convenience sampling. A pre-requisite for participation in the study was sufficient computer literacy as well as word processor expertise. Forty percent of the sample was drawn from second year Computer Science students who were registered for a community service module. The other 60% of the sample was drawn from the student assistants for the computer literacy course of the university, with the proviso that they were not studied for a computer science or related degree. These students all had to complete the literacy course prior to becoming an assistant and they had to achieve at least 70% for a competency test of Microsoft Office applications.

In order to determine a measurement of expertise with Microsoft Word®, the question pertaining to the duration that the participants had used Word and the frequency with which they used Word were each measured on a scale of 0-4. Then the responses to these two questions were multiplied to get a measurement on a scale of 0-16. This scale was then viewed as a measurement of expertise. The expertise rating of each participant was calculated and it was found that there were 3 participants with low Word expertise (scale rating from 0 until 6), 3 with average Word expertise (scale rating from 7 until 10) and 17 with high Word expertise (scale rating above 10). Since there were other measures in place to confirm their expertise with a word processor, namely their qualification to serve as assistants, all participants were accepted into the study as competent Word users.

There were 17 male participants and 8 female participants and the average age of the participants was 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 were comfortable in either English or Afrikaans. Each session was conducted in the tuition language in which the participant was most comfortable.

Only four participants used keyboard shortcuts while working in Word and 17 preferred using the mouse rather than the keyboard to complete tasks. Only one participant, who was a Computer Science student, had had exposure to the eye-tracker but this was in the capacity of using it for research purposes. Therefore, while he was at ease with the technology he had not used it as an input technique. Five participants had used speech recognition before but only as a dictation tool.

6.4 Tasks

The task list (Table 3.5) had a total of 20 tasks, five of which were typing tasks (phrases to be typed were randomly chosen from the phrase set specified in section 3.4.3.3). Three of these typing tasks had to be completed using the onscreen keyboard with eye gaze and speech as an interaction technique. The majority of the other types of task, 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 successfully. This task list remained the same for the first four sessions. Thereafter, an additional 5 typing tasks (Table 6.1) were added to the end of the task list. These typing tasks all had to be completed using the onscreen keyboard. However, the size and spacing of the keys were adjusted in order to test the effect of different spacing and button sizes on typing. By the fourth session most participants were able to complete the original 20 tasks in less than their scheduled half an hour. Since participants were not pressured to complete all tasks but rather to complete as many as possible and as accurately as possible within their allotted time, the additional tasks placed no further pressure on them to complete more tasks. If their 30 minutes expired and

they had not yet completed all tasks, they were simply allowed to finish the task they were busy with and then the test ended.

The first additional typing task used the same settings as the original tasks. The next two used buttons that were 5 pixels smaller in both height and width, but were 10 pixels further apart. The final two tasks used buttons that were reduced a further 5 pixels in width and height but which were spaced the same as for the original typing tasks. These tasks were added in response to requests from participants that they be permitted to try smaller buttons for the typing tasks. The additional typing tasks were preceded by a new task that required the participant to remove all the text from the document. Consequently, the final five sessions had 26 tasks. The setup of the typing tasks will be discussed in greater detail in the next chapter.

The tasks could be grouped as follows:

Table 6.1: Task description and grouping

Task	Task text	Task type
22	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing with original settings
23	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing with slightly smaller buttons but spaced further apart
24	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing with slightly smaller buttons but spaced further apart
25	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing with slightly smaller buttons but spaced the same as original buttons
26	Enter the following phrase using eye gaze and speech recognition: *<randomly selected phrase>*	Typing with slightly smaller buttons but spaced the same as original buttons

A more succinct summary of the tasks is tabulated below. Since the next chapter will focus on the typing tasks and this chapter only on the other tasks, the typing tasks will be omitted for the time being:

Table 6.2: Grouped tasks as divided between interaction techniques

Task type	Keyboard	Speech
Line selection and formatting	1	1
Select all text and remove	1	2
Select words and format	1	1
Paste	1	1
Undo	1	1
Select word and copy	1	1
Position and paste	1	1
Select all and format		1

This chapter will concentrate on analysing each of these task types individually for usability and learnability. The typing tasks will be analysed in the following chapter.

6.5 Measurements

As discussed and defined in section 2.3, the four pillars of usability which are of interest in this study are effectiveness, efficiency, learnability and satisfaction. Measurements of efficiency, effectiveness and

learnability of the speech commands will be discussed in this chapter. The subjective satisfaction will be analysed and reported on in Chapter 8.

The efficiency measurements that will be analysed are the time taken to complete the task and the number of actions that were required to complete the task. Additionally, the effectiveness measurement of the percentage of the task completed correctly by the participants will also be evaluated. 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 was not a member of that set could 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 number of actions to complete and the percentage of the task completed correctly were better indicators of the effectiveness of the interaction techniques.

6.6 Limitations of this study

A limitation of the study was the sensitivity of the speech recognition which was prone to pick up ambient noise and react to it. This meant that although the participants completed the task correctly, the result was not always as expected. For example, noises picked up by the microphone might cause the cursor to move to an unexpected position, resulting in erroneous input. However, participants soon learnt that this was the case and learnt to compensate for these shortcomings somewhat. This limitation also has the associated advantage of emulating a real world environment, therefore, it was not considered to be detriment to the study.

The fact that the testing was conducted in a controlled environment could also be considered a limitation as testing was not conducted in a variety of possible environments in which it could be used.

Furthermore, there was only a subset of commands catered for in the grammar. Since the aim of the study was to test speech recognition to provide a more hands-free environment, only the basic tasks were provided for in the grammar. It should not be a problem to extend the grammar to include more commands. The only shortcoming this created was that the grammar may have been easier to memorise than a longer one. Since there were only ten weeks in which the participants could learn the grammar, this was not felt to be a major drawback as most participants had been using Word in the traditional sense for a number of years already.

6.7 Task analysis

6.7.1 Line selection and formatting

These tasks required that the participant select three consecutive lines of text and then apply formatting. The minimum number of actions required for both of these tasks was 4.

6.7.1.1 Time to complete 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.

Time to complete the task was measured in seconds and Table 6.3 summarises the number of participants (first line of each row), the mean seconds required to complete the task (second line of each row) and standard deviation (third line of each row) of the time required to complete the selection and formatting tasks.

Table 6.3: Descriptive statistics for time to complete line selection and formatting

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	25	25	12	13
	$\bar{x}$	76.2	37.3	84.2	40.4
	s	44.7	17.7	51.8	20.7
Session 3	n	23	23	12	13
	$\bar{x}$	36.9	25.5	32.0	26.2
	s	20.7	8.0	19.2	7.4
Session 4	n	24	24	12	13
	$\bar{x}$	34.9	30.5	31.7	32.4
	s	18.4	26.2	17.0	31.6
Session 5	n	23	23	12	13
	$\bar{x}$	29.1	27.0	25.1	28.6
	s	21.0	23.6	21.3	28.6
Session 6	n	23	23	12	13
	$\bar{x}$	26.8	24.5	18.0	22.3
	s	22.6	14.1	3.2	9.2
Session 7	n	22	22	12	13
	$\bar{x}$	23.6	21.1	19.1	22.5
	s	17.0	9.7	9.1	9.9
Session 8	n	20	20	12	13
	$\bar{x}$	16.5	19.3	16.3	20.9
	s	4.4	8.4	4.6	9.5
Session 9	n	22	22	12	13
	$\bar{x}$	27.6	18.1	22.2	16.3
	s	19.9	7.0	10.8	6.4
Session 10	n	24	24	12	13
	$\bar{x}$	21.0	17.7	16.6	17.2
	s	13.5	8.2	4.9	7.9

From Table 6.3 it can be seen that the times required to complete the two tasks are more or less the same over the majority of the sessions. It was only during the second session that the time for the speech task was much higher than the keyboard task. This could easily be attributed to the fact that the participants had not yet mastered the speech recognition and still had to consult the handout to determine which commands were needed. In the weeks thereafter, the recall of the commands could have been easier. In the eighth week, the speech task was, on average, completed even faster than with the keyboard. With the exception of the ninth and tenth week (and the fourth week for the keyboard), the times for both tasks steadily lessened. The fact that the time lessened with each session indicates that the commands are learnable and memorable as more exposure to the commands facilitated quicker completion of tasks.

Chart 6.1 plots the means for both interaction techniques over all sessions. The vertical bars denote a 95% confidence interval.

The time measurements were in seconds and there were a vast number of instances in which the normality tests failed for the data, often for more than one test on that range of data. In order to combat this, the time measurement was converted to 1/time and normality tests were again conducted on this data. Although

1/time will be used for the analysis of the time, the descriptive statistics and charts will be based on the original time data for the sake of clarity. This will apply to all the time analyses in this chapter. The results for the normality tests are summarised in Table 6.4.

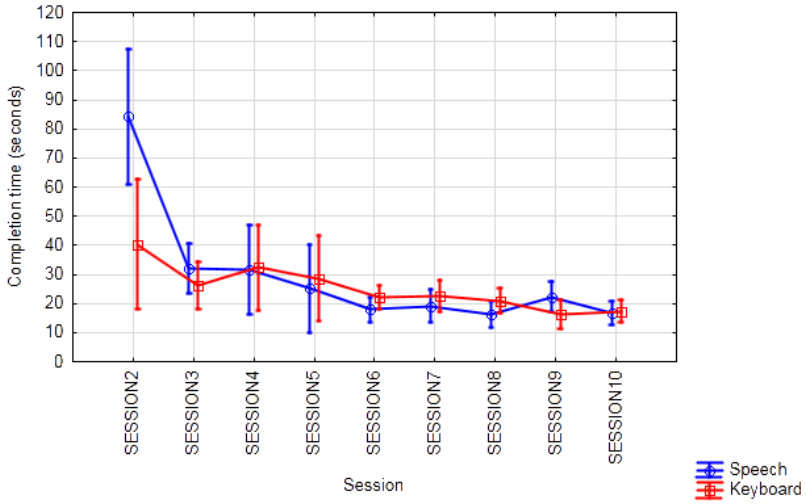


Chart 6.1: Means for completion time of line selection and formatting

Table 6.4: Normality test results from completion time of line selection and formatting

	Shapiro-Wilks	Kolmogorov-Smirnov
Session 2	W = 0.887, p < 0.05	d = 0.148, p > 0.05
Session 3	W = 0.921, p < 0.05	d = 0.111, p > 0.05
Session 4	W = 0.977, p > 0.05	d = 0.089, p > 0.05
Session 5	W = 0.968, p > 0.05	d = 0.094, p > 0.05
Session 6	W = 0.940, p > 0.05	d = 0.124, p > 0.05
Session 7	W = 0.940, p > 0.05	d = 0.124, p > 0.05
Session 8	W = 0.973, p > 0.05	d = 0.111, p > 0.05
Session 9	W = 0.896, p > 0.05	d = 0.113, p > 0.05
Session 10	W = 0.895, p > 0.05	d = 0.110, p > 0.05

According to the Shapiro-Wilks test, the time measurements of sessions 1 and 2 are not normally distributed. Owing to the robustness of the ANOVA when the data is not normally distributed, all subsequent analyses will always include the tests for normality. For the sake of conciseness, however, the individual results will not be reported.

The assumption of sphericity was also violated ($\chi^2(35) = 68.969$, $p < 0.05$), therefore the adjusted corrections (StatSoft, 2010) will also be reported. The following hypotheses were tested for this analysis:

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. $H_{0,2}$: The practice obtained over the sessions has no effect on the time taken to complete the tasks.

The repeated measures ANOVA yielded the result of not rejecting the first null hypothesis at an α-level of 0.05 ($F(1, 23) = 0.286$, $p > 0.05$). Therefore, it can be concluded that there is no difference between the time required to complete line selection and formatting when using the different interaction techniques. Therefore,

using speech commands is just as fast as using the mouse and keyboard. Moreover, this indicates that the use of speech commands is a viable alternative to the mouse and keyboard, in terms of time required.

$H_{0,2}$ could be rejected ($F(8, 184) = 14.040$, $p < 0.05$) which indicates that practice significantly affects the time required to complete the task. The interaction between the two factors of session and interaction technique was not significant ($F(8, 184) = 1.722$, $p > 0.05$). The results of the adjusted univariate results as well as the multivariate tests are shown in the table below:

Table 6.5: ANOVA results for the completion time of line selection and formatting

	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Session	$F(4.0, 92.2) = 14.040$, $p < 0.05$	$F(5.1, 118.9) = 14.040$, $p < 0.05$	$F(8, 16) = 13.292$, $p < 0.05$
Session × Interaction technique	$F(1, 92.2) = 1.722$, $p > 0.05$	$F(5.1, 118.9) = 1.722$, $p > 0.05$	$F(8, 16) = 1.255$, $p > 0.05$

Tukey's HSD was used for post-hoc analysis to establish which sessions were responsible for the significant difference. Session 2 differed significantly from all other sessions. Session 3 differed significantly from session 6 as well as from sessions 8, 9 and 10. Session 4 differed significantly from session's 8, 9 and 10. Since session 2 was actually the first session where the tasks had to be completed, the reason for the longer time could be that the participants were not familiar with the tasks and the process of the task completion. As the participants became accustomed to the session requirements, the time lessened. Therefore, the improvement from the second to the third session and the subsequent performance measured was significant. The improvement between sessions 3 and 4 and the last sessions was also significant. This indicates a significant level of learning to use the system such that performance is comparable to traditional interaction techniques.

6.7.1.2 Number of actions

The next measurement to be analysed was the number of actions that were performed during task completion. Actions are defined as any mouse click, button press or speech command that was issued during completion of the task. The number of actions was measured per interaction technique and per session for each participant and then, as for all other analyses, outliers were removed from the data set prior to analysis (Section 3.5).

The table below shows the number of participants whose data was included in the analysis, with the mean of the data in the second line of each row and the standard deviation in the third line.

The descriptive statistics show that the number of actions for the speech interaction technique was very high in the first session and then declined sharply during the second session. Thereafter, it decreased steadily until session 8 after which it stabilised somewhat around an average of 10. The actions for the keyboard remained between a minimum of 10 and maximum of 15 for the majority of the sessions. Apart from the first session and session 8, the actions appear to be comparable for the two interaction techniques. Chart 6.2 is a plot of the mean number of actions over all the sessions.

Table 6.6: Descriptive statistics for the number of actions used for line selection and formatting

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	25	23	13	10
	$\bar{x}$	32.4	11.7	34.9	16.5
	s	23.0	11.4	25.5	15.9
Session 3	n	23	22	13	10
	$\bar{x}$	12.3	8.5	8.5	8.1
	s	8.5	9.9	5.7	8.4
Session 4	n	24	20	13	10
	$\bar{x}$	14.4	14.1	12.5	20.0
	s	9.6	14.7	7.9	16.0
Session 5	n	23	22	13	10
	$\bar{x}$	12.5	13.9	11.2	14.0
	s	9.6	15.7	9.9	14.7
Session 6	n	24	23	13	10
	$\bar{x}$	10.2	10.9	7.8	12.7
	s	6.2	9.4	2.0	12.4
Session 7	n	23	22	13	10
	$\bar{x}$	9.7	11.9	8.1	12.4
	s	7.1	12.4	3.3	11.9
Session 8	n	20	19	13	10
	$\bar{x}$	7.7	15.1	8.1	15.2
	s	2.6	14.6	2.8	12.9
Session 9	n	21	21	13	10
	$\bar{x}$	10.8	9.1	10.5	11.3
	s	4.3	7.2	4.0	8.7
Session 10	n	23	23	13	10
	$\bar{x}$	9.2	12.5	7.8	10.9
	s	3.8	10.5	2.1	8.4

The descriptive statistics show that the number of actions for the speech interaction technique was very high in the first session and then declined sharply during the second session. Thereafter, it decreased steadily until session 8 after which it stabilised somewhat around an average of 10. The actions for the keyboard remained between a minimum of 10 and maximum of 15 for the majority of the sessions. Apart from the first session and session 8, the actions appear to be comparable for the two interaction techniques. Chart 6.2 is a plot of the mean number of actions over all the sessions.

The assumption of sphericity was also not met ($\chi^2(35) = 137.094$, $p < 0.05$), which will require that adjusted corrections be made to the degrees of freedom using the Geisser-Greenhouse and Huyn-Feldt tests (StatSoft, 2010). The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the number of actions required to complete the task.
2. $H_{0,2}$: There is no difference between the number of actions required to complete the tasks over the various sessions.

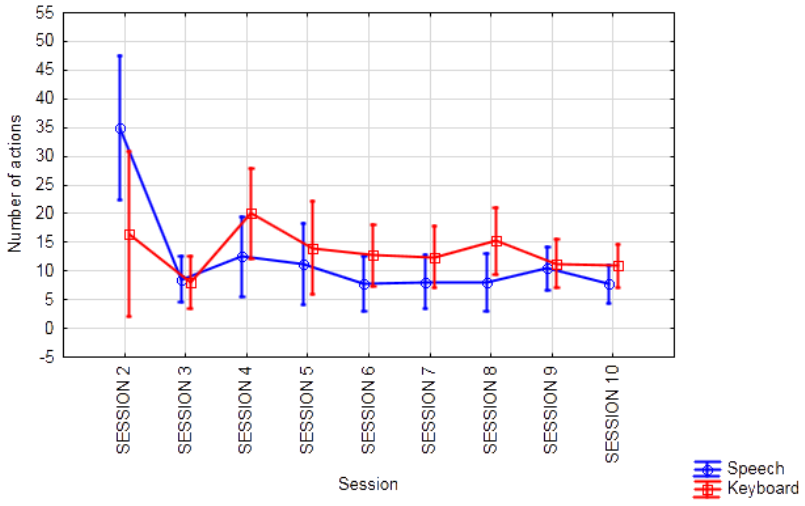


Chart 6.2: Mean number of actions required to perform line selection and formatting

At an α-level of 0.05, there is significant interaction between the two factors ($F(8, 164) = 4.105$, $p < 0.05$) which means that the results of the overall ANOVA cannot be interpreted easily, but that separate ANOVAs must be conducted so as to control for each factor individually. Consequently nine separate ANOVAs were performed to determine if the interaction technique had a significant effect on the number of actions required to complete the task. These results are tabulated below:

Table 6.7: Results of ANOVA on the number of actions required to perform line selection and formatting

	ANOVA
Session 2	$F(1, 46) = 15.260$, $p < 0.05$
Session 3	$F(1, 43) = 1.918$, $p > 0.05$
Session 4	$F(1, 42) = 0.008$, $p > 0.05$
Session 5	$F(1, 43) = 0.138$, $p > 0.05$
Session 6	$F(1, 45) = 0.105$, $p > 0.05$
Session 7	$F(1, 43) = 0.540$, $p > 0.05$
Session 8	$F(1, 37) = 4.992$, $p < 0.05$
Session 9	$F(1, 40) = 0.792$, $p > 0.05$
Session 10	$F(1, 44) = 2.006$, $p > 0.05$

Consequently, $H_{0,1}$ could only be rejected for sessions 2 and 8. During session 2, the keyboard required significantly fewer actions to complete the task and in session 8 it required significantly more actions than the speech interaction technique. Session 2 was the first session and could be viewed as a learning experience for the participants to become suitably accustomed to the speech commands available for use. After this session, the actions were reduced to such a degree that the same number of actions could be used as when using the keyboard. Combining this with the findings of the time required to complete the task it could be concluded that speech and the keyboard are equivalent in terms of efficiency and effectiveness when selecting multiple lines and applying formatting to those lines and even have the ability to surpass the keyboard with some usability measurements.

A repeated-measures ANOVA was then conducted to analyse $H_{0,2}$ for each interaction technique. The hypothesis could not be rejected for the keyboard but could be rejected for the speech commands. Tukey's post-hoc analysis was used to determine which sessions differed significantly. It was found that session 2 differed significantly from all other sessions for speech. During session 2, participants required more actions, on average, than in the other sessions. This could indicate that participants were simply unfamiliar with the test structure, particularly since these were the first two tasks in the test. At this stage, participants were also

unsure of the verbal commands and may have required some time to familiarise themselves with the available commands. The fact that the number of actions decreased over the sessions showed that the participants were able to learn, and use, the commands to select lines and format more effectively as time went by.

Since the key presses were, on average, marginally more than the speech commands, the actual keys (which were captured and stored real time during completion of the test) that were pressed were examined more closely. It was found that for each session the [**Right**] and [**Down**] keys were used a large number of times. The [**End**] key was only used during three sessions and only by a single participant during two of these sessions and two participants in the other session. This indicates that the majority of the participants either were not aware of the shortcut [**Control + End**] to move to the end of a document or preferred to navigate there using multiple key presses or the mouse. The [**Right**] key was by far pressed the greatest number of times which implies that some of the participants selected the lines character by character by holding the [**Right**] key in. Depending on the amount of text to be selected this is by far the most inefficient method of selecting text, particularly when whole words or lines must be selected. 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.

6.7.1.3 Correctness of task completion

Each of these tasks featured three distinct components which had to be performed in order for the task to be completed correctly, namely:

1. The participant had to select a portion of text.
2. The correct text had to be selected. For example, some tasks required the first two words to be selected and formatted. If the participant only applied formatting to the first word, then they would receive credit for the fact that a selection occurred (number 1) but not for this step as the formatting was not applied correctly according to the task specifications.
3. The correct formatting had to be applied.

Participants received credit for each component of the task that they completed correctly. Therefore, if the participant attempted the task a minimum of zero for each task and a maximum of 3 could be scored. The stacked bar graph below shows the number of participants who scored 0, 1, 2 and 3 for each of the tasks and for each session.

The high incidence of a zero count in a large proportion of the categories prevents a meaningful analysis from being performed; therefore no statistical inferences will be made from the data. The keyboard had a lower task completion rate than the speech. The fact that the keyboard task required participants to select the last three lines in the document and the speech task the first three lines was the cause of this difference. All the participants who scored 2 for this task lost a mark for selecting the incorrect lines. However, observation during task completion showed that the participants did in fact select three lines during the completion of the task, they just selected the first three lines of text in the document. Therefore, the lower scores for the keyboard are caused, not by the interaction technique *per se*, but rather by the participants not reading the task instructions correctly. If points were awarded less strictly and based on whether any three lines of text were selected, all participants with a current score of 2 would score 3 for the task. Consequently, the correctness of the two tasks would be identical. This leads to the conclusion that, in terms of the interaction technique, there is no difference between the keyboard and speech command with respect to the correctness with which the task can be completed.

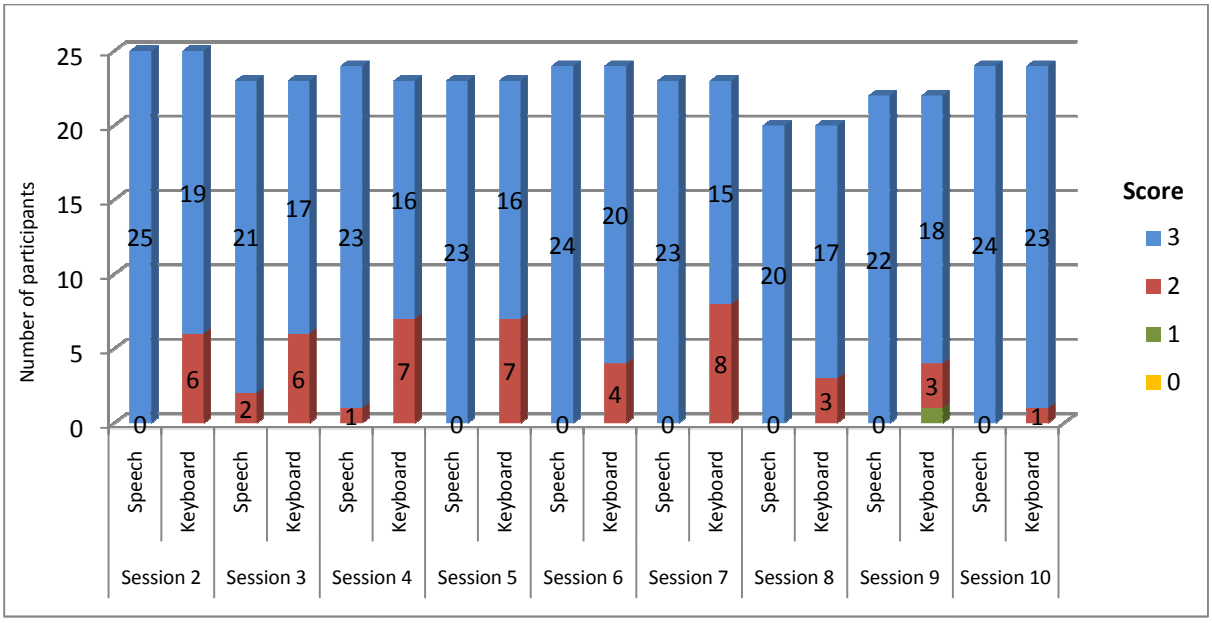


Chart 6.3: Correctness of task - Select lines and format

Also note that the document that was provided to the participants to work in had only eight lines of text and fitted comfortably within the viewing area even when the onscreen keyboard was activated. Therefore, the fact that the tasks required lines of text to be selected at the start of the document and then at the end of the document was considered trivial in terms of the possible impact it would have on task completion.

6.7.2 Select all text and remove

There was a single task which required the participants to select all text and remove it from the document when using the keyboard. There were two such tasks to be completed using speech commands. However, the second task using speech recognition was simply in place so that the participants could complete the typing tasks in a clean document. Therefore, the instruction to remove the text for this task using speech was not strictly enforced and for this reason the second of these tasks using speech recognition will not be included in the current analysis.

The keyboard task required the participants to select all the text in the document and to cut it. Alternatively, the speech task required that all text must be selected and deleted from the document, which would require the commands “Select all” followed by “Remove”. This means that the keyboard task required one more key press to complete the task successfully, namely [**Ctrl + A**] and then [**Ctrl + X**]. Nevertheless, the end result of the tasks is the same and the single extra action should not complicate the use of the keyboard to such an extent that the difference will be significant as a result thereof. Since the end result is the same the two tasks both fall under the same category of document text selection and removal.

6.7.2.1 Time to complete task

The time to complete each task was measured in seconds for each session and each participant. The assumption of sphericity ($\chi^2(35) = 47.024$, $p > 0.05$) was met, therefore the data was suitable to be analysed using a repeated-measures within-subjects ANOVA. Descriptive statistics of the data are summarised in Table 6.8, with the first line giving the number of participants who completed the task (outliers have been removed), the second the mean and the third the standard deviation.

Inspection of the means shows that both interaction techniques started with a fairly low completion time and then continued to decrease over the following sessions. Eventually, the speech interaction technique could be used to complete the task faster than with the keyboard and mouse. Chart 6.4 below gives a visual representation of the means for both interaction techniques over all sessions.

The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the time required to complete the task.
2. $H_{0,2}$: The session in which the task was completed has no effect on the time required to complete the task.

Table 6.8: Descriptive statistics for completion time of removing all selected text

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	25	25	12	13
	$\bar{x}$	28.2	25.7	24.0	20.3
	s	23.0	14.3	20.1	6.9
Session 3	n	23	23	12	13
	$\bar{x}$	17.4	17.0	16.1	14.9
	s	10.4	7.1	9.9	5.5
Session 4	n	24	24	12	13
	$\bar{x}$	13.2	16.8	12.4	13.2
	s	6.1	14.1	5.5	4.7
Session 5	n	23	23	12	13
	$\bar{x}$	13.4	14.7	10.8	12.6
	s	11.1	7.6	9.4	5.1
Session 6	n	23	24	12	13
	$\bar{x}$	13.5	12.4	10.6	11.7
	s	11.9	5.1	9.7	4.5
Session 7	n	21	22	12	13
	$\bar{x}$	7.5	12.8	8.0	11.9
	s	2.7	7.3	3.5	8.1
Session 8	n	20	20	12	13
	$\bar{x}$	10.2	13.1	8.0	11.7
	s	9.7	8.5	2.8	9.4
Session 9	n	22	22	12	13
	$\bar{x}$	8.6	12.0	7.7	12.2
	s	4.3	7.3	1.9	8.7
Session 10	n	24	24	12	13
	$\bar{x}$	7.3	12.6	6.5	10.9
	s	2.4	7.6	1.8	5.6

$H_{0,1}$ could be rejected at a significance level of 0.05 ($F(1, 23) = 4.328$, $p < 0.05$) and therefore it can be concluded that the interaction technique does have a significant effect on the time required to complete the task. When using the keyboard or mouse for this task, participants took significantly longer to complete the task in the majority of the sessions than when they made use of speech commands. This is evidence of the fact that speech can be used to make selection and removal of text more efficiently than the keyboard or mouse.

Similarly, $H_{0,2}$ could be rejected at an α-level of 0.05 ($F(8, 184) = 15.197$, $p < 0.05$), which indicates the session has a significant effect on the time taken to complete the task. Tukey's HSD test was used to determine which sessions differed significantly. It was found that session 2 differed significantly from sessions 5 to 10. Session 3 differed significantly from sessions 6 to 10 and session 4 from sessions 7 to 10. Sessions 4 and 5 differed

significantly from session 10. Similar patterns held for both interaction techniques. Since the later sessions all had an average completion time less than the earlier sessions, it could be said that the first sessions took significantly longer than the later sessions. This would indicate some measure of learning in using the interaction techniques and perhaps the application as a whole.

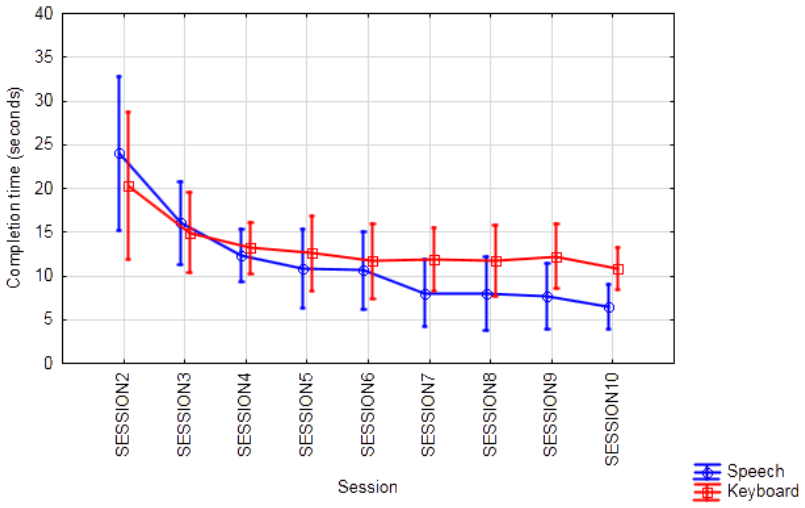


Chart 6.4: Mean plot for completion time of removing all selected text

6.7.2.2 Number of actions

The number of actions per task was measured for each participant and each session. Table 6.9 shows the descriptive statistics for the data. The first row shows the number of participants who were included in the analysis, the second row is the mean and the third the standard deviation. Chart 6.5 is a graphical representation of the mean number of actions.

From Table 6.9 it is clear that the keyboard and mouse task required more actions, on average, than the speech commands. This could indicate a higher level of efficiency for the speech interaction technique, particularly in view of the significant difference in task completion times. To determine whether these differences are significant, the following hypotheses were proposed:

1. $H_{0,1}$: The interaction technique does not significantly affect the number of actions required to complete the task.
2. $H_{0,2}$: The number of actions required to complete the tasks does not differ significantly between sessions.

$H_{0,1}$ could be rejected at an α-level of 0.05 ($F(1, 18) = 8.574$, $p < 0.05$), leading to the conclusion that the interaction technique had a significant effect on the number of actions required to complete the task. More specifically, the speech interaction technique requires significantly fewer actions to complete the task than when a keyboard and mouse are used. $H_{0,2}$ could also be rejected ($F(8, 144) = 2.562$, $p < 0.05$) indicating that there was a noticeable change in the number of actions required as exposure to the application was increased. Since the data did not meet the assumption of sphericity ($\chi^2(35) = 106.449$, $p < 0.05$), to conclude the analysis the adjusted corrections and multivariate results are summarised in Table 6.10.

Table 6.9: Descriptive statistics for the number of actions required to remove all selected text

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	24	25	12	8
	$\bar{x}$	9.2	11.8	7.4	6.4
	s	9.6	10.6	7.9	3.3
Session 3	n	23	22	12	8
	$\bar{x}$	5.2	6.7	4.2	5.8
	s	4.5	3.2	3.1	3.1
Session 4	n	24	22	12	8
	$\bar{x}$	4.8	8.2	4.7	10.8
	s	3.5	6.7	3.3	7.1
Session 5	n	23	22	12	8
	$\bar{x}$	5.3	8.9	4.5	8.8
	s	4.9	7.4	4.7	7.3
Session 6	n	22	23	12	8
	$\bar{x}$	3.6	6.3	2.9	7.1
	s	2.2	3.8	2.3	4.3
Session 7	n	21	20	12	8
	$\bar{x}$	3.3	8.4	3.3	4.9
	s	1.2	7.6	1.4	2.4
Session 8	n	20	20	12	8
	$\bar{x}$	4.2	8.8	3.8	7.9
	s	2.3	8.3	1.7	8.8
Session 9	n	22	19	12	8
	$\bar{x}$	3.2	5.0	2.9	4.0
	s	1.7	3.3	0.9	1.6
Session 10	n	24	22	12	8
	$\bar{x}$	3.3	5.9	3.1	5.0
	s	1.3	3.3	1.2	2.1

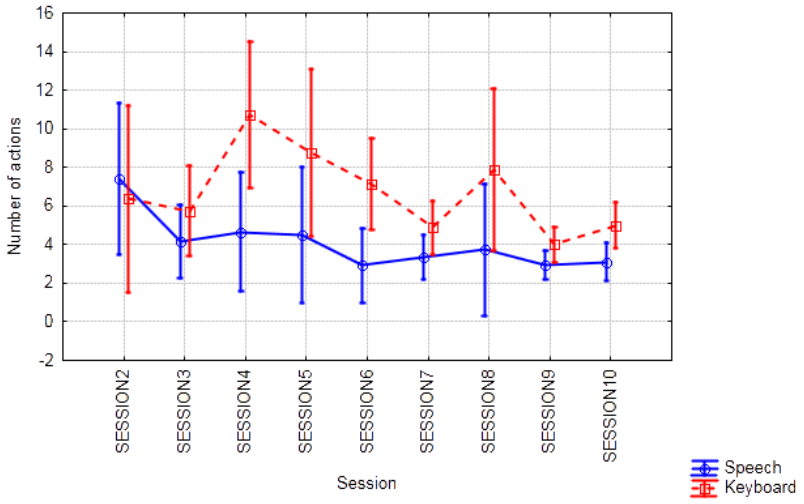


Chart 6.5: Mean plot for the number of actions required to remove all selected text

Table 6.10: Analysis results for the number of actions required to remove all selected text

	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Session	$F(3.5, 62.1) = 2.562$, $p < 0.05$	$F(4.6, 82.9) = 2.562$, $p < 0.05$	$F(8, 11) = 3.379$, $p < 0.05$
Interaction technique × Session	$F(3.5, 62.1) = 1.441$, $p > 0.05$	$F(4.6, 82.9) = 1.441$, $p > 0.05$	$F(8, 11) = 1.579$, $p > 0.05$

Once again, closer analysis of the actions for the keyboard task showed a high incidence of key presses for the [**Right**] arrow key during all sessions. Very few key presses were registered on the [**A**] key even though the combination of [**Control + A**] will select all the text in the document. This again shows that the participants do not use the keyboard shortcuts which are in place to simplify and speed up the use of the application or that they are not aware of the shortcuts which can be used. The fact that the participants appeared to select the text one character at a time could explain the high number of actions for this task. It would be interesting to conduct a study in which participants are coached in the proper use of the shortcuts and then times and actions can be measured to complete tasks using the keyboard and speech. This will allow more conclusive analysis to be conducted on the difference between the interaction techniques if the shortest method is used for both tasks.

Tukey's HSD did not highlight any significant differences between sessions but the less conservative Fisher's LSD did. Session 2 differed significantly from sessions 7, 9 and 10 and session 4 differed significantly from sessions 6, 7, 9 and 10. Sessions 2 and 4 had the highest average number of actions required to complete the tasks which could signify that for some reason the participants struggled more that week than they did in the other weeks. Session 2 could be attributed to the first-time use of the application. Apart from sessions 4, 5, and 8 all other sessions showed an improvement in or similar performance to the previous session.

The fact that this task allowed for a more efficient completion time and actions required is promising as it implies that there are circumstances under which the use of speech could be more efficient that the traditional means of interaction.

6.7.2.3 Correctness of task completion

The three components of this task that had to be completed correctly were:

1. The participant had to select a portion of text.
2. All the text in the document had to be selected.
3. The selected text had to be removed.

The chart below gives a stacked bar chart for the number of participants who scored 0, 1, 2 or 3 for either task and for all the sessions.

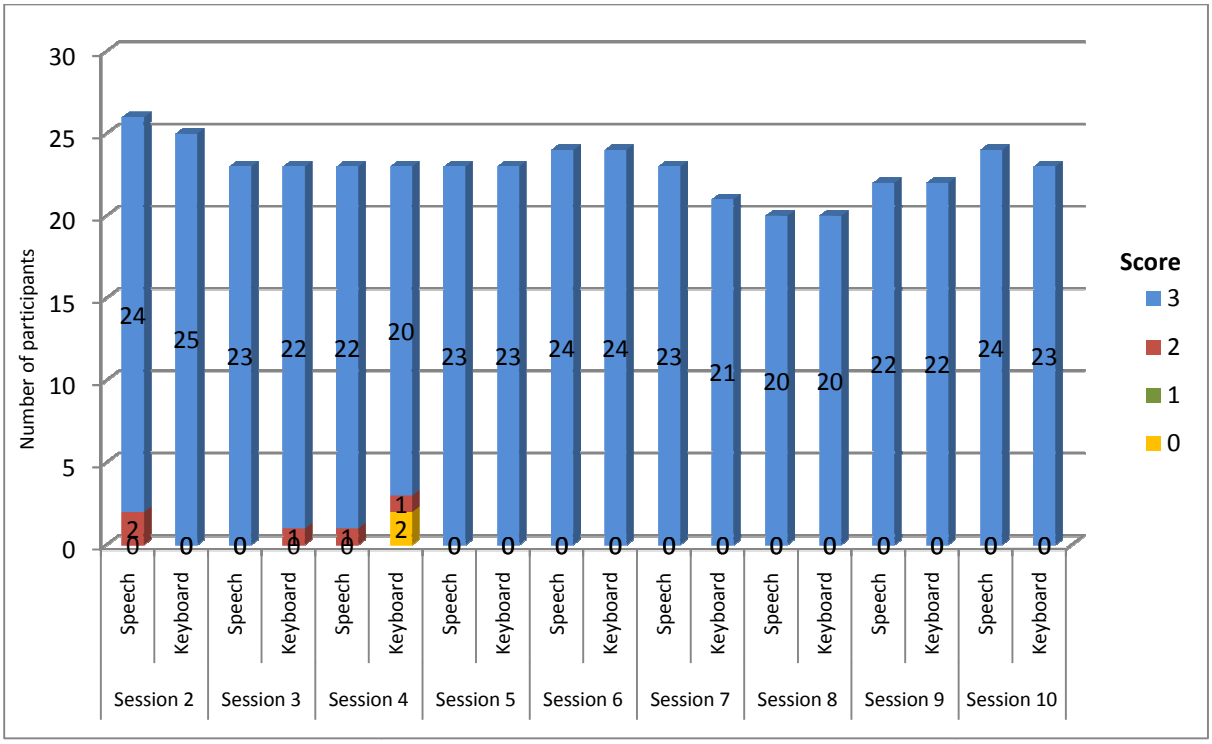


Chart 6.6: Correctness of task - Select all text and remove

Once again, the high incidence of no observations in many categories prevented a meaningful statistical analysis from being performed. However, it can clearly be seen from Chart 6.6 that the vast majority of the participants were able to complete the task 100% correctly from the very first session. There are only isolated incidents where this was not the case and it is doubtful that these will cause a significant difference between either the interaction techniques or the sessions. Therefore, it can be concluded that the correctness with which the task is completed is not affected by either the interaction technique or the session in which the task is completed.

6.7.3 Select words and format

This task required the participants to select the first two words of the current line that they were on and make them bold. The task had to be completed once with the speech interaction technique and once with the keyboard.

6.7.3.1 Time to complete the task

The number of observations, the mean and standard deviation of the completion times for each session are shown in the first, second and third line of each row respectively in Table 6.11.

This is the first task where the time for the speech task has a higher average completion time than the corresponding keyboard task. However, it could still be possible that this increased completion time is not significantly different to that of the keyboard. Chart 6.7 provides a plot of the mean for the interaction techniques across all sessions.

Table 6.11: Descriptive statistics for the completion time of formatting selected words

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	25	25	14	14
	x̄	49.4	29.3	50.1	29.7
	s	25.8	23.5	30.2	27.3
Session 3	n	23	22	14	14
	x̄	41.8	21.5	42.5	19.3
	s	28.8	13.4	30.2	10.3
Session 4	n	24	23	14	14
	x̄	31.5	17.6	27.5	16.2
	s	16.2	6.9	14.2	7.0
Session 5	n	23	22	14	14
	x̄	23.9	19.3	21.1	17.1
	s	10.4	7.7	11.7	7.1
Session 6	n	23	24	14	14
	x̄	24.9	15.1	25.1	14.4
	s	12.8	3.8	13.6	3.8
Session 7	n	21	21	14	14
	x̄	25.4	17.4	21.2	16.0
	s	15.5	8.5	11.2	7.7
Session 8	n	20	20	14	14
	x̄	21.2	14.0	18.3	12.5
	s	12.5	5.8	10.7	3.8
Session 9	n	22	22	14	14
	x̄	25.0	14.5	20.4	13.3
	s	16.1	6.6	13.5	5.9
Session 10	n	24	24	14	14
	x̄	28.1	15.6	22.9	13.5
	s	15.8	6.4	15.0	4.9

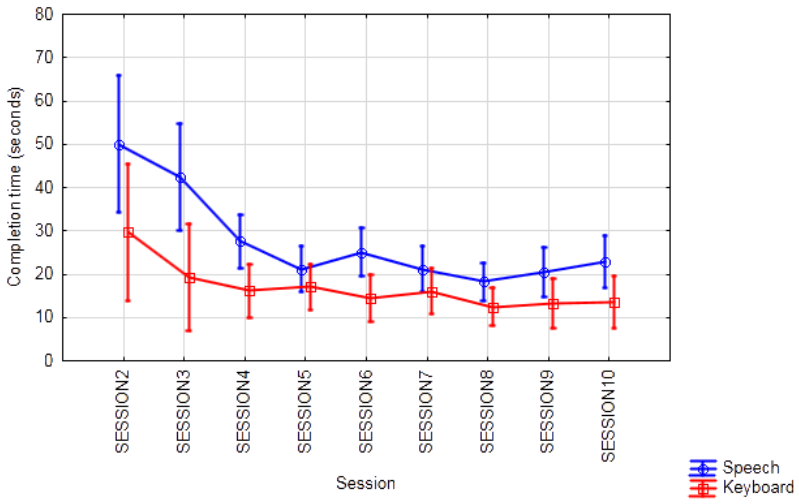


Chart 6.7: Mean plot for completion times of formatting selected words

The data did not meet the assumption of sphericity ($\chi^2(35) = 53.048$, $p < 0.05$), so an adjusted univariate analysis also had to be performed in order to analyse the following hypotheses:

1. $H_{0,1}$: The interaction technique has no effect on the time taken to complete the task.
2. $H_{0,2}$: There is no difference between the time taken to complete the task between the different sessions.

These results, together with the general ANOVA and the multivariate results, are summarised in Table 6.12.

Table 6.12: Analysis results for the completion times of formatting selected text

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 26) = 10.447$, $p < 0.05$			
Session	$F(8, 208) = 9.487$, $p < 0.05$	$F(4.9, 126.4) = 9.487$, $p < 0.05$	$F(6.3, 165.1) = 9.487$, $p < 0.05$	$F(8, 19) = 5.707$, $p < 0.05$
Interaction technique × Session	$F(8, 208) = 0.669$, $p > 0.05$	$F(4.9, 126.4) = 0.669$, $p > 0.05$	$F(6.3, 165.0) = 0.669$, $p > 0.05$	$F(8, 19) = 0.952$, $p > 0.05$

The null hypothesis, $H_{0,1}$, was rejected at an α-level of 0.05. This means that the interaction technique has a significant effect on the time taken to complete the task. Since the time to complete the task using speech was, on average, higher for all sessions it can be concluded that using speech commands to select words and apply formatting takes significantly longer than using the mouse or keyboard.

The second null hypothesis of no difference could also be rejected at a significance level of 0.05. Post-hoc tests were conducted to determine which sessions differed significantly from one another. Session 2 differed significantly from sessions 4 to 10, session 3 differed significantly from sessions 7 to 10 and session 4 differed significantly from session 8. The extended completion times could be attributed to the learning curve experienced with the verbal commands. Similar to previous tasks, session 2 has the highest time which has previously been attributed to inexperience with the application. This may be true for this task as well. There was a large improvement in task completion during session 3 but this was not a significant improvement. The keyboard also showed marked improvement after the first two sessions.

This task can be considered more complex than the previous since it requires slightly more difficult verbal commands to be issued, in sequence, in order to achieve the desired goal. Selection of a word is accomplished through the verbal command of “select word” or “select word back” or alternatively, individual characters can be selected by issuing several “shift right” or “shift left” commands. These commands may be less intuitive than the previous commands and require more time for becoming accustomed to and remembering. Even so, after a number of sessions, participants were able to achieve times that were comparable with the mouse and/or keyboard. The best performance with the speech commands was achieved during the eighth session after which the time increased again. It can also be surmised that it is possible that participants struggled to remember that there were commands available to select a single word at a time and resorted to selecting the letters one at a time. Closer inspection of the number of actions and commands issued will provide more insight into whether this was the possible reason for the extended time.

6.7.3.2 Number of actions

Table 6.13 shows the number of participants included in each session for the analysis, followed by the mean and finally the standard deviation for each session.

Table 6.13: Descriptive statistics for the number of actions required to format selected words

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	25	22	14	11
	$\bar{x}$	15.1	14.9	16.1	17.1
	s	10.7	14.0	12.4	16.0
Session 3	n	23	23	14	11
	$\bar{x}$	13.6	15.3	14.1	11.7
	s	10.0	16.8	10.6	10.3
Session 4	n	24	24	14	11
	$\bar{x}$	10.4	14.2	8.6	13.5
	s	6.2	11.8	6.2	10.1
Session 5	n	23	22	14	11
	$\bar{x}$	9.0	16.8	7.3	17.3
	s	6.1	14.6	4.2	18.7
Session 6	n	24	24	14	11
	$\bar{x}$	9.3	13.1	9.6	15.6
	s	5.8	14.2	5.8	15.1
Session 7	n	23	21	14	11
	$\bar{x}$	8.7	19.3	7.7	19.1
	s	6.1	20.1	5.2	20.1
Session 8	n	20	19	14	11
	$\bar{x}$	8.1	9.5	7.4	8.4
	s	5.2	7.8	5.5	5.1
Session 9	n	22	22	14	11
	$\bar{x}$	9.7	14.0	7.0	11.4
	s	6.8	12.0	3.6	9.3
Session 10	n	24	24	14	11
	$\bar{x}$	10.5	14.2	8.8	12.6
	s	6.1	11.8	5.5	11.7

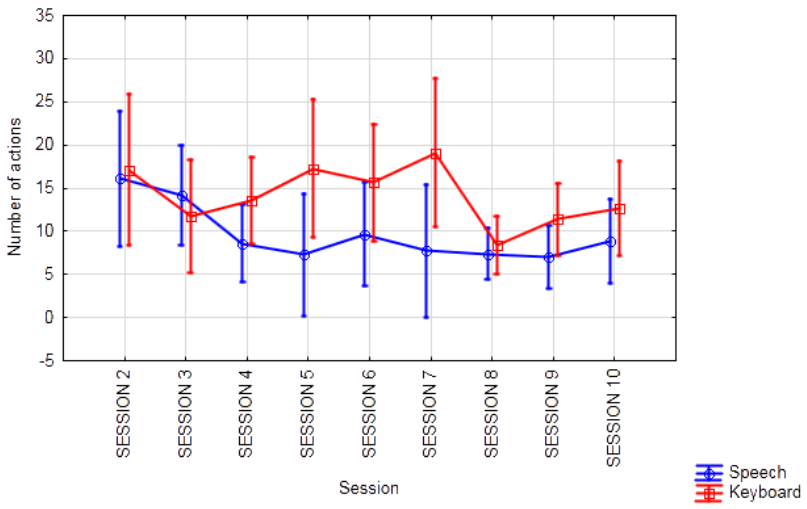


Chart 6.8: Mean plot for the number of actions required to format selected words

According to the table above, the number of actions required to complete the task using the keyboard or mouse was, on average, much higher than the number required when using the speech commands after the third session. This is contrary to one of the explanations offered for the difference in the times required to

complete the tasks. Chart 6.8 gives a visual representation of the mean number of actions for the interaction techniques over all the sessions.

The assumption of sphericity was not met ($\chi^2(35) = 93.477$, $p < 0.05$); therefore the adjusted corrections will also be reported. The following hypotheses were investigated:

1. $H_{0,1}$: The interaction technique does not have a significant impact on the number of actions required to complete the task.
2. $H_{0,2}$: There is no significant difference between the number of actions per session.

The following table summarises all the results for the repeated-measures within-subjects ANOVA.

Table 6.14: Analysis results for the number of actions required to format selected words

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 23) = 2.598$, $p > 0.05$			
Session	$F(8, 184) = 2.234$, $p < 0.05$	$F(4.1, 94.3) = 2.234$, $p < 0.05$	$F(5.3, 122.2) = 2.234$, $p < 0.05$	$F(8, 16) = 3.300$, $p < 0.05$
Interaction technique × Session	$F(8, 184) = 1.646$, $p > 0.05$	$F(4.1, 94.3) = 1.646$, $p > 0.05$	$F(5.3, 122.2) = 1.646$, $p > 0.05$	$F(8, 16) = 1.546$, $p > 0.05$

From the table it can be concluded that $H_{0,1}$ could not be rejected but $H_{0,2}$ could be rejected. This means that while the interaction technique does not have an impact on the number of actions required to complete the task, the session does. Post-hoc tests indicate that session 2 differed significantly from sessions 8 and 9. Even though there was some learning experienced, as indicated by the decrease in the number of actions for all sessions, the improvement was not significant from one session to the next. However, the level of improvement resulted in the performance in sessions 8 and 9 being significantly better than for the second session.

Since the speech interaction technique took significantly longer to complete the tasks, it was assumed that this could mean that more actions were required to complete the task. However, this analysis shows that there is no significant difference between the actions performed during the task even though the number of actions for the speech was, on average, less than that for the keyboard. It seems counterintuitive that two measures of efficiency could yield such contradictory results. The question now arises – how can the speech be significantly slower but require fewer actions, albeit not significantly fewer? A reasonable explanation for this could be that there are longer pauses between the actions performed or that the actions performed required more time to complete. The following section will investigate this supposition.

6.7.3.3 Average time between actions

The fact that speech commands result in a significantly longer time to complete the task but required fewer actions may seem contradictory. It was however, inferred that this meant that although there were fewer actions required, each action required more time than did those for the keyboard. Therefore, in an effort to explain the apparent discrepancy, the average time between actions was measured for both of these tasks. This time does not include the time from when the task started to when the first action is performed but rather measures only the difference between performed actions. The mean difference between actions and the standard deviation per session can be seen in Table 6.15.

Table 6.15: Descriptive statistics for the time difference between actions

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	$\bar{x}$	3.364	1.466	3.119	1.405
	s	1.437	1.350	0.811	1.320
Session 3	$\bar{x}$	3.215	1.549	2.972	1.418
	s	1.411	1.271	1.199	1.226
Session 4	$\bar{x}$	2.640	1.195	2.586	1.132
	s	0.747	0.986	0.657	1.196
Session 5	$\bar{x}$	2.672	1.608	2.457	1.741
	s	1.091	0.815	0.951	2.244
Session 6	$\bar{x}$	2.480	1.323	2.136	1.189
	s	0.788	1.072	0.369	1.164
Session 7	$\bar{x}$	2.233	1.452	2.176	1.060
	s	0.674	1.434	0.793	1.461
Session 8	$\bar{x}$	2.371	1.085	2.398	0.833
	s	0.633	0.774	0.745	0.519
Session 9	$\bar{x}$	2.372	1.214	2.350	0.918
	s	0.624	1.165	0.567	0.896
Session 10	$\bar{x}$	2.418	1.356	2.206	1.430
	s	0.749	1.522	0.535	1.935

Table 6.15 and Chart 6.9 clearly show that the average time between actions is a great deal less for the keyboard and mouse than for the speech commands. It is encouraging to notice that the difference between commands for the speech improved with each session, although it appears to stabilise for the speech from the seventh session. The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no noticeable impact on the average time between actions.
2. $H_{0,2}$: There is no noticeable difference between the average time between commands between sessions.

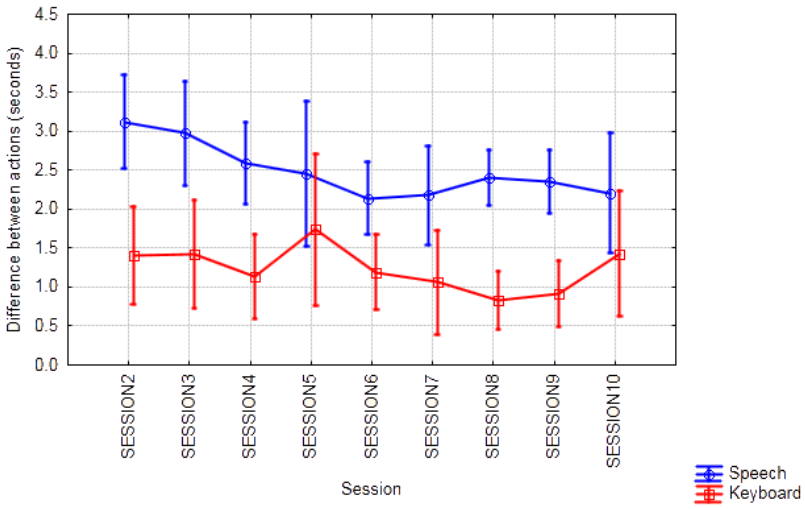


Chart 6.9: Mean plot for the time difference between actions

Using a confidence interval of 95%, both $H_{0,1}$ ($F(1, 25) = 22.307$, $p < 0.05$) and $H_{0,2}$ ($F(8, 200) = 2.037$, $p < 0.05$) could be rejected (Table 6.14 contains the adjusted corrections and results of the multivariate analysis since the assumption of sphericity did not hold – $\chi^2(35) = 76.009$, $p < 0.05$). This means that the interaction technique plays a significant effect on the time elapsed between actions. Actions can be performed in a much more rapid sequence when using the keyboard and mouse than when using speech commands.

Table 6.16: Analysis results for the time difference between actions

	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Session	$F(4.6, 116.1) = 2.037$, $p < 0.05$	$F(6.1, 151.6) = 2.037$, $p < 0.05$	$F(8, 18) = 1.972$, $p > 0.05$
Interaction technique × Session	$F(4.6, 116.1) = 1.035$, $p > 0.05$	$F(6.1, 151.6) = 1.035$, $p > 0.05$	$F(8, 18) = 1.972$, $p > 0.05$

This analysis shows that even though the number of actions required to complete the task is comparable for the two interaction techniques, the time difference between issuing commands is not negligible and has a noticeable impact on the time taken to complete the task. The improvement between the second and third sessions and the last four sessions was significant.

The time difference between the issuing of commands could be attributed to two reasons, namely either the physical utterance of the verbal command consumes more time than key presses or quite possibly the participants required more time to determine the next command to be issued for the speech than for the keyboard. The fact that the time between commands decreased as time went by, points to the second reason as it seems implausible that the participants could learn to utter a command faster as a person's speaking rate is an innate human quality. Therefore, it could be assumed that the commands used were less intuitive than the prior commands required and could have placed more strain on the memory of the participant. Since the time between commands appears to stabilise from session 7 onwards, this could be the first session where the time difference is purely because of the time required to issue the command and not recall the command. A more prolonged testing period may serve to either substantiate or contradict this statement and further research is required to test this statistically.

6.7.3.4 Correctness of task completion

This task too had a maximum score of three which could be obtained by completing the following steps correctly:

1. Any portion of text is selected by the participant.
2. The correct portion of text is selected by the participant.
3. Bold formatting is applied to the selection.

The chart below is a stacked bar graph representing the number of participants who scored zero, one, two or three for the task. Participants with zero were unable to complete any steps of the tasks correctly. The graph shows the results for both interaction techniques as well as all sessions.

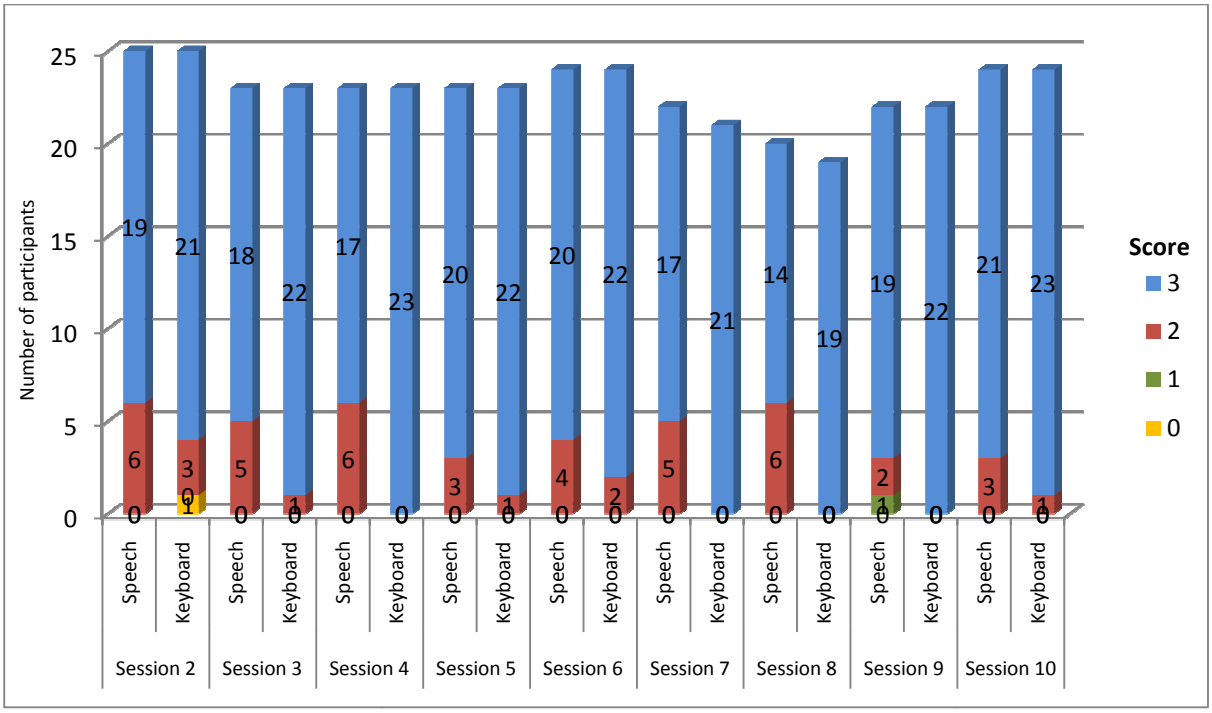


Chart 6.10: Correctness of task - Select words and apply formatting

Similar to the previous tasks, the greatest majority of participants could complete the task correctly from the very first session and irrespective of the interaction technique which was used. There is a slightly higher occurrence of participants who did not complete the task completely correctly but once again this is more due to the fact the task was not read correctly. Participants did not always select the first two words on the current line, thereby causing a decrease in the correctness with which they completed the task. Hence, if the selection of the first two words were to be disregarded then virtually all tasks would be completed with 100% correctness.

6.7.4 Paste

These tasks required the participants to paste a previously copied or cut word after the second word on the current line that they were on. Therefore, there was some navigation required, followed by a paste. Both of the tasks could be achieved in the same minimum number of actions.

6.7.4.1 Time to complete the task

The number of participants whose data was included in the analysis, the mean in seconds as well as the standard deviation of each session and for each interaction technique are summarised in Table 6.17 and the chart directly below that plots the means for the data.

Table 6.17: Descriptive statistics for paste time completion

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	23	23	13	13
	$\bar{x}$	8.6	13.6	8.2	11.9
	s	2.9	4.4	2.2	3.1
Session 3	n	23	23	13	13
	$\bar{x}$	9.1	10.6	7.2	9.3
	s	7.0	4.2	2.4	1.8
Session 4	n	24	24	13	13
	$\bar{x}$	6.2	10.2	5.6	8.4
	s	2.0	4.7	1.6	3.6
Session 5	n	23	23	13	13
	$\bar{x}$	5.6	8.0	5.6	6.7
	s	1.4	3.0	1.9	1.5
Session 6	n	23	23	13	13
	$\bar{x}$	5.0	7.8	4.7	7.1
	s	1.0	2.1	0.9	1.9
Session 7	n	22	22	13	13
	$\bar{x}$	4.8	8.3	4.3	7.0
	s	1.5	3.8	1.1	2.2
Session 8	n	20	20	13	13
	$\bar{x}$	4.6	6.9	4.3	6.0
	s	1.1	2.8	1.2	1.5
Session 9	n	22	22	13	13
	$\bar{x}$	4.7	8.7	4.5	8.6
	s	1.7	5.1	0.9	6.5
Session 10	n	24	24	13	13
	$\bar{x}$	4.4	9.3	4.2	6.7
	s	1.3	7.7	1.1	2.0

Using Table 6.15 and Chart 6.11 as a reference, it can be seen that, on average, the time to complete the task using speech commands was faster than when using the keyboard or mouse. The assumption of sphericity ($\chi^2(35) = 37.242$, $p > 0.05$) was met at a confidence interval of 95%, therefore no adjusted corrections had to be applied. The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the time required to complete the task.
2. $H_{0,2}$: The session in which the task was completed has no effect on the time required to complete the task.

When a repeated-measure within-subjects ANOVA was performed, 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 be analysed in isolation to preclude the interaction with the other factor having an effect on the analysis.

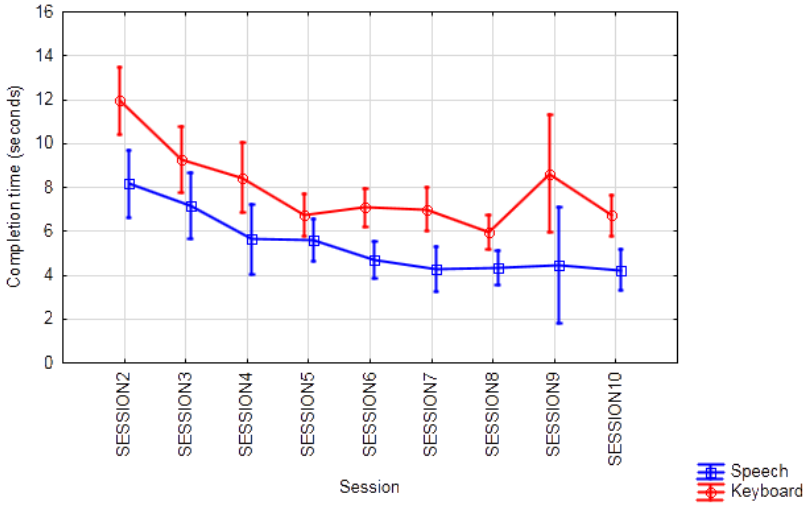


Chart 6.11: Mean plot for the paste time completion

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 ($F(8, 96) = 6.883$, $p < 0.05$). For the speech interaction technique, post-hoc tests indicated that there was a significant difference between the times of session 2 and sessions 4 to 10, as well as between session 3 and sessions 6 to 10 and between session 4 and session 7, 8 and 10. Similarly, there was a significant difference between session 5 and sessions 7, 8, and 10. These results indicated that session 2 could be viewed as a simple practice run to allow participants to become accustomed to the application and the appropriate use of the speech commands. From then onwards there was improvement in the times achieved to complete the task to such an extent that from session 4 onwards there was a significantly better completion rate. Over the subsequent three sessions there was constant improvement to the extent that they were significantly slower than the final sessions. From session 6 onwards, there was still minor improvement but not to an extent that times differed significantly.

Post-hoc tests for the keyboard and mouse showed that session 2 differed from session 4 to 10 and session 3 differed significantly from sessions 8 and 10.

6.7.4.2 Number of actions

Descriptive statistics for the number of actions are given in Table 6.18. The number of participants whose data was included in the analysis, after the removal of the outliers, is shown in the first line of each row in Table 6.18. Following this is the mean for that session and then the standard deviation in the second and third row respectively.

Table 6.18: Descriptive statistics for the number of actions to complete a paste

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	22	18	12	5
	x̄	1.9	2.4	1.8	2.0
	s	0.8	0.8	0.8	0
Session 3	n	22	20	12	5
	x̄	2.9	2.3	2.8	2.2
	s	1.5	0.6	1.5	0.4
Session 4	n	24	23	12	5
	x̄	1.8	2.4	1.5	2.0
	s	1.0	0.7	0.7	0
Session 5	n	23	22	12	5
	x̄	1.6	2.4	1.8	2.2
	s	1.0	0.9	1.4	0.4
Session 6	n	22	21	12	5
	x̄	1.5	2.1	1.4	2.0
	s	0.9	0.3	0.7	0
Session 7	n	22	19	12	5
	x̄	1.7	2.6	1.9	2.2
	s	0.9	0.8	1.2	0.4
Session 8	n	20	19	12	5
	x̄	1.6	2.5	1.4	2.2
	s	0.8	0.9	0.8	0.4
Session 9	n	22	21	12	5
	x̄	1.7	2.3	1.4	2.2
	s	0.9	0.6	0.7	0.4
Session 10	n	24	20	12	5
	x̄	1.6	2.3	1.5	2.8
	s	1.0	0.6	0.7	0.8

Chart 6.12 gives a visual presentation of the mean number of actions per interaction technique for each session.

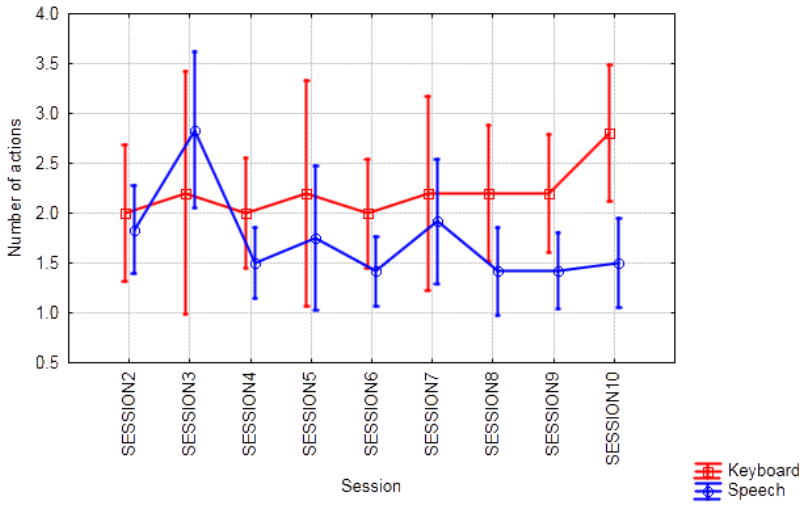


Chart 6.12: Mean plot for the number of actions to complete the paste

The minimum number of actions to complete the tasks was 1 for speech ("Paste") and 2 when using the keyboard ([**Control + V**]) and mouse; therefore if the participants used the most efficient method for each task the number of actions should be approximately the same. From Chart 6.9 it can be seen that the number of actions when using the keyboard and mouse is consistently higher than when using the speech commands for all sessions except for session 3. The mean for the speech indicates that the majority of the participants could use the most effective means when using speech commands. However, the same cannot be said for the keyboard and mouse.

The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique does not have a significant impact on the number of actions required to complete the task.
2. $H_{0,2}$: There is no significant difference between the number of actions per session.

The assumption of sphericity ($\chi^2(35) = 66.827$, $p < 0.05$) was not met, therefore an adjusted correction analysis was required (Table 6.19). The repeated-measures within-subjects ANOVA allowed for $H_{0,1}$ to be rejected ($F(1, 15) = 6.287$, $p < 0.05$) at an α-level of 0.05. Since the speech commands required less actions in all sessions (apart from session 3), it can be concluded that the keyboard and mouse required significantly more commands to complete the task than the speech. One observation that was made during data capturing was that many participants used a right click to show the context menu and then clicked on paste. The paste command is normally the third item on the menu. However, if there is a spelling or grammatical error, the paste command moves to the last item on the menu. Very often it was the case that where the paste was to occur there was an error in the document. This led to the participants' not seeing the paste option at the very end of the menu as they were not accustomed to this. The participants would then repeatedly right click the menu in an attempt to get a menu that they recognise, not realising that the paste command was in fact available. This could have significantly increased the number of actions performed by the participants when using the mouse. The fact that this behaviour was observed throughout the nine sessions also indicates that the participants did not learn that the paste option shifts to the end of the menu to accommodate corrective suggestions to the text.

The second null hypothesis could not be rejected ($F(8, 120) = 1.297$, $p > 0.05$) at an α-level of 0.05. Given that the number of actions for the speech was low throughout, this indicates that no learning was required for the paste command as it was intuitive enough to accommodate expedited completion times from the very first use of the command. Furthermore, it can also be said that no learning occurred when using the keyboard and mouse to paste a piece of text which is slightly more worrisome as it would be expected at the competency level of the participants that such a minor change in the menu arrangement would be easily noticeable. However, since it is doubtful that a paste will occur before the spelling is corrected under normal use; this observed phenomenon is perhaps inconsequential within the scope of standard word processing use.

Table 6.19: Analysis results for the number of actions to complete the paste task

	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Session	$F(4.0, 59.3) = 1.297$, $p > 0.05$	$F(5.9, 88.5) = 1.297$, $p > 0.05$	$F(8, 8) = 2.011$, $p > 0.05$
Interaction technique × Session	$F(4.0, 59.3) = 1.424$, $p > 0.05$	$F(5.9, 88.5) = 1.424$, $p > 0.05$	$F(8, 8) = 0.948$, $p > 0.05$

6.7.4.3 Correctness of task completion

For this task, participants could only score zero or one as the paste had to occur at the precise location of the cursor when the task started. Therefore, no positioning was required to complete the task. All participants

could complete the tasks correctly from the very first session and with either interaction technique. It was only in session 2 and 3 for the keyboard task where a single participant did not perform the paste correctly.

6.7.5 Undo

The undo tasks required only a single action when using both the speech and the keyboard or mouse and were designed to undo the paste of the previous task. The same analysis procedure as for the prior tasks was followed for the undo tasks.

6.7.5.1 Time to complete

The underlying table gives descriptive statistics for the completion time of the undo task. The number of observations, mean and standard deviation are listed on the first, second and third row for each session. The chart directly following that plots the mean completion times.

Table 6.20: Descriptive statistics for task completion time for the undo task

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	23	22	13	13
	x̄	11.7	10.6	8.7	9.4
	s	7.3	4.5	3.8	3.8
Session 3	n	23	23	13	13
	x̄	9.4	8.7	7.5	7.4
	s	8.4	5.1	4.7	3.9
Session 4	n	24	24	13	13
	x̄	7.7	7.4	5.8	6.3
	s	4.8	4.1	1.6	3.1
Session 5	n	23	23	13	13
	x̄	5.8	6.6	5.6	6.0
	s	2.6	3.1	3.1	2.6
Session 6	n	23	23	13	13
	x̄	5.3	5.9	4.7	5.1
	s	1.7	2.1	1.1	2.0
Session 7	n	22	22	13	13
	x̄	4.6	5.8	4.2	5.3
	s	1.0	2.6	0.7	2.4
Session 8	n	20	20	13	13
	x̄	4.3	5.3	4.1	4.5
	s	1.1	2.6	0.7	2.1
Session 9	n	22	22	13	13
	x̄	4.5	5.6	4.2	5.1
	s	1.1	2.6	1.0	2.8
Session 10	n	24	24	13	13
	x̄	4.6	4.9	4.4	4.3
	s	1.1	1.9	1.0	1.9

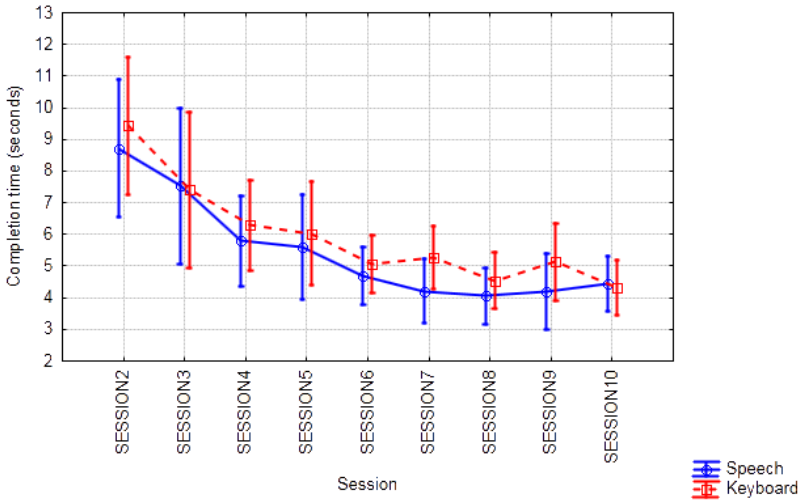


Chart 6.13: Mean plot for the completion time of the undo task

This is the first task where the speech and keyboard interaction techniques have almost identical completion times. Both interaction techniques exhibited a decrease in completion time as the sessions went by. It is really only session 7 and 9 which show a slight increase for the keyboard and 9 and 10 for the speech. However, these increases are expected to be non-significant. Furthermore, since the completion times over all sessions are approximately the same it is not expected that there will be a significant difference between them. However, it is still essential that statistical analysis be performed to verify this as well as to establish whether the decrease in completion times, as exhibited by both interaction techniques, is significant.

The assumption of sphericity for the time data was violated ($\chi^2(35) = 50.614$, $p < 0.05$), therefore the adjusted corrections will also be reported. The following hypotheses were formulated for evaluation of the undo tasks:

1. $H_{0,1}$: The interaction technique does not significantly affect the time taken to complete the task.
2. $H_{0,2}$: The time taken to complete the task does not differ significantly over the sessions.

As expected, $H_{0,1}$ could not be rejected at an α-level of 0.05, which proves that the task can be completed in the comparable times regardless of the interaction technique. The second null hypothesis could, however, be rejected at an α-level of 0.05 (Table 6.21). Therefore, the sessions differed significantly; in particular, each of sessions 2 to 5 differed significantly from all sessions that followed them. This means that, even though the improvement between the first sessions was not significant, it eventually allowed the last sessions to be significantly better than the first.

Table 6.21: Analysis results for the completion time of the undo task

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 24) = 0.001$, $p > 0.05$			
Session	$F(8, 192) = 22.148$, $p < 0.05$	$F(5.5, 131.5) = 22.148$, $p < 0.05$	$F(7.6, 182.0) = 22.148$, $p < 0.05$	$F(8, 17) = 20.036$, $p < 0.05$
Interaction technique × Session	$F(8, 192) = 0.643$, $p > 0.05$	$F(5.5, 131.5) = 0.643$, $p > 0.05$	$F(7.6, 182.0) = 0.643$, $p > 0.05$	$F(8, 17) = 0.784$, $p > 0.05$

6.7.5.2 Number of actions

As previously mentioned, the tasks respectively required 1 or 2 actions to be completed with the speech and keyboard interaction techniques. Descriptive statistics for the number of actions for this task are contained in Table 6.22.

Table 6.22: Descriptive statistics for the number of actions to complete the undo task

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	23	22	13	13
	$\bar{x}$	2.3	2.6	1.8	2.5
	s	1.3	1.9	0.9	1.5
Session 3	n	23	23	13	13
	$\bar{x}$	2.5	3.5	2.8	2.6
	s	2.4	3.5	2.3	2.3
Session 4	n	24	24	13	13
	$\bar{x}$	1.8	2.6	1.7	1.8
	s	0.6	2.3	0.9	0.4
Session 5	n	23	22	13	13
	$\bar{x}$	1.8	1.9	1.9	1.9
	s	0.5	1.1	1.3	0.5
Session 6	n	23	23	13	13
	$\bar{x}$	1.9	1.7	1.6	2.0
	s	0.7	0.7	0.5	0.6
Session 7	n	22	22	13	13
	$\bar{x}$	1.8	1.3	1.2	1.9
	s	0.6	0.5	0.4	0.5
Session 8	n	19	20	13	13
	$\bar{x}$	1.9	1.5	1.5	1.9
	s	0.5	0.6	0.7	0.5
Session 9	n	22	22	13	13
	$\bar{x}$	1.8	1.8	1.9	1.8
	s	0.4	0.7	0.8	0.4
Session 10	n	24	24	13	13
	$\bar{x}$	1.9	1.8	1.6	2.1
	s	0.7	0.9	0.8	0.6

From the table above, it can be extrapolated that the number of actions for the two interaction techniques are approximately the same for all sessions. There is no real discernible pattern that can be determined from the number of actions. The number of actions for the speech increase for session 3 and decrease sharply for session 4 after which it stabilises. Conversely, the number of actions for the keyboard and mouse start decreasing from session 4 and continues decreasing until session 7. Chart 6.14 provides a plot of the mean number of actions to assist the analysis for significant differences.

The following hypotheses were formulated for evaluation of the number of actions:

1. $H_{0,1}$: The interaction technique has no effect on the number of actions required to complete the task.
2. $H_{0,2}$: There is no difference between the number of actions required to complete the actions between the sessions.

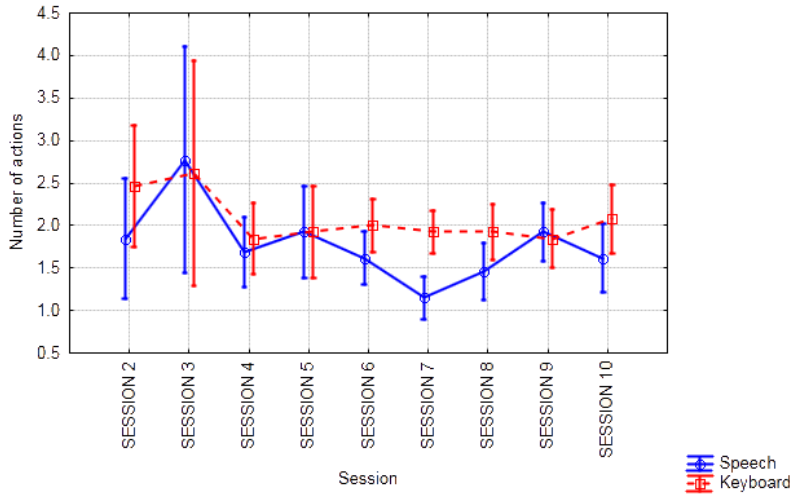


Chart 6.14: Mean number of actions to complete the undo task

The assumption of sphericity required for the repeated-measures within-subjects ANOVA was violated ($\chi^2(35) = 137.438$, $p < 0.05$), therefore the tabulated results below show the results of the ANOVA, the multivariate tests and the adjusted corrections results.

Table 6.23: Analysis results for the number of actions to complete the undo task

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(1, 24) = 2.294, p > 0.05			
Session	F(8, 192) = 2.934, p < 0.05	F(2.5, 62.4) = 2.934, p < 0.05	F(3.1, 73.5) = 2.934, p < 0.05	F(8, 17) = 3.742, p < 0.05
Interaction technique × Session	F(8, 192) = 0.690, p > 0.05	F(2.5, 62.4) = 0.690, p > 0.05	F(3.1, 73.5) = 0.690, p > 0.05	F(8, 17) = 2.904, p < 0.05

From Table 6.23 it can be concluded that $H_{0,1}$ could not be rejected meaning that there was no difference between the number of actions required to complete the task when using speech and when using the keyboard and mouse. The second hypothesis of no difference could, however, be rejected using a confidence interval of 95%. Therefore, there is a significant difference between the number of actions over the sessions. Post-hoc tests provided more insight into which sessions differed significantly. It was only session 3 which differed significantly from other sessions, namely sessions 4, 6, 7 and 8. Session 3 had a very high average number of actions hence it could be concluded that during session 3 significantly more actions were performed to complete the task than in sessions 4, 6, 7 and 8. Since this was an isolated incident the overall conclusion that could be made from these findings is that to reverse the previous action is as simple when using speech as when using the keyboard and mouse, even to such an extent that from the very first session the number of actions between the two interaction techniques is on a comparable level.

6.7.5.3 Correctness of task completion

The simplistic nature of the task required only a single action and, apart from in the final session of the keyboard task, all participants completed the task correctly. In the final session of the keyboard task, one participant did not complete the task correctly. This was due to the fact that instead of using the keyboard the

participant issued the relevant speech command to complete the task. Technically this means the task was completed correctly but the participant was penalised since the incorrect interaction technique was used.

6.7.6 Select word and copy

At the start of this task, for both the interaction techniques, the cursor should have been at the end of a line in the document, based on the task directly prior to this one. This meant that the task required the participant to select the word directly to the left of the cursor. Therefore, this task tested a different type of selection to the previous tasks which had a selection component.

6.7.6.1 Time to complete task

The underlying table contains the number of observations included in the analysis, the mean of the observations and finally the standard deviation. These are arranged in the first, second and third lines respectively of each row.

Table 6.24: Descriptive statistics for the completion time for selecting and copying a word

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	21	21	13	11
	$\bar{x}$	37.6	17.6	29.3	15.6
	s	18.1	6.0	11.2	5.0
Session 3	n	22	21	13	11
	$\bar{x}$	31.5	19.0	31.2	17.9
	s	19.4	8.7	20.4	9.1
Session 4	n	24	24	13	11
	$\bar{x}$	29.6	23.6	24.8	16.1
	s	22.5	12.9	23.4	7.2
Session 5	n	22	22	13	11
	$\bar{x}$	23.1	15.3	25.1	13.6
	s	14.1	7.4	17.3	9.0
Session 6	n	22	22	13	11
	$\bar{x}$	15.3	14.5	12.7	11.9
	s	6.7	5.5	5.2	3.4
Session 7	n	22	22	13	11
	$\bar{x}$	23.9	17.9	23.2	14.6
	s	17.2	8.2	15.5	6.3
Session 8	n	20	20	13	11
	$\bar{x}$	19.9	17.9	19.3	15.3
	s	14.9	7.0	17.6	6.8
Session 9	n	22	22	13	11
	$\bar{x}$	20.2	15.9	19.8	13.1
	s	14.8	6.1	16.0	5.9
Session 10	n	24	24	13	11
	$\bar{x}$	20.7	17.0	16.5	13.9
	s	10.9	8.8	5.1	6.1

Through inspection of Table 6.24, it can be inferred that the speech interaction technique required more time to complete the task for all of the sessions. There are isolated sessions, for example session 6, where the

completion times appear to be on a more comparable level between the two interaction techniques. Chart 6.15 is a plot of the mean number of actions for both interaction techniques over all sessions.

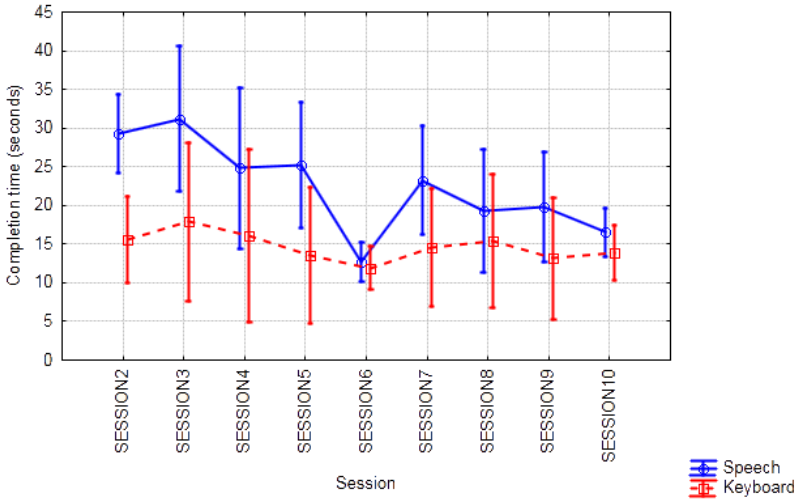


Chart 6.15: Mean plot for the completion time for selecting and copying a word

A repeated-measures within-subject ANOVA will be used to determine whether these differences are significant. The following hypotheses were formulated:

1. $H_{0,1}$: There is no difference in the time to complete the task when using the different interaction techniques.
2. $H_{0,2}$: There is no difference between the time to complete the task over the different sessions.

The assumption of sphericity was met ($\chi^2(35) = 39.456$, $p < 0.05$) at an α-level of 0.05. Using a confidence level of 95%, $H_{0,1}$ could not be rejected ($F(1, 22) = 3.655$, $p > 0.05$) but $H_{0,2}$ could be rejected ($F(8, 176) = 3.470$, $p < 0.05$). The multivariate tests confirmed that $H_{0,2}$ could be rejected ($F(8, 15) = 3.103$, $p < 0.05$).

These results show that the interaction technique does not impact the time required to complete the task of selecting a word and copying it. Since there were multiple sessions, a post-hoc test was required to determine which sessions differed significantly. Tukey's HSD test showed that session 2 differed significantly from sessions 6 and 8. Session 3 also differed significantly from session 6. Session 2 had a significantly longer completion time than the other two and session 3 had a significantly longer time than session 6.

To conclude, it can be said that although the speech commands necessitate a longer time to select a word and copy it, this longer time is not significantly different to that of the keyboard and mouse. Therefore, with regard to selecting text to the left of the cursor, the same efficiency can be achieved with the two interaction techniques.

During the previous selection task the keyboard was significantly faster when used to select a word and apply formatting. This task also required a word to be selected, and in this instance to the left of the cursor, it was surmised that it might be slightly more complicated. The additional components to the task, namely bold and copy were considered to be inconsequential as both are very common tasks and required the same number of actions. The complexity of the task was viewed to be the selection of the required text. The fact that there was no significant difference could possibly be attributed to two reasons. Firstly, it could be that the selection of a word to the left of the cursor provided more of a challenge with the keyboard and mouse than selecting to the right. Secondly, it could possibly be that the previous task jogged the memory of the participants enough that

they could effectively recall the required command – so much so that they could select words on a comparable rate to that of the keyboard and mouse. Closer inspection of the mean times for both of these tasks indicated that the completion rate was fairly similar for the keyboard but was lower for the speech for the second task. This would seem to imply that the second proposition is more plausible than the first. Therefore repetition of similar tasks allows the commands for the latter tasks to be recalled and executed easier than the first time the task is encountered in each session of use. However, bearing in mind that the previous task required selection of two words and this task the selection of only a single word, credence could be lent to the first supposition as well. It would seem that both suppositions have some merit. Of course, this has not been analysed statistically and is based purely on observation of the data spread and speculation about the findings of the analysis.

7.7.6.2 Number of actions

The minimum number of actions required for the completion of this task was, once again, similar for the two interaction techniques. Table 6.25 summarises the descriptive statistics for the number of actions for this task.

Table 6.25: Descriptive statistics for the number of actions to select and copy text

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	22	20	13	8
	$\bar{x}$	14.0	17.0	8.4	13.0
	s	12.2	22.6	5.1	11.3
Session 3	n	23	21	13	8
	$\bar{x}$	10.9	24.9	10.8	11.5
	s	7.5	32.2	8.0	11.5
Session 4	n	24	24	13	8
	$\bar{x}$	10.0	38.6	8.7	22.0
	s	9.3	42.4	8.9	30.7
Session 5	n	23	20	13	8
	$\bar{x}$	9.5	12.7	9.3	12.5
	s	9.5	13.3	7.6	14.7
Session 6	n	23	20	13	8
	$\bar{x}$	6.1	7.5	4.8	6.6
	s	6.1	5.8	2.7	4.2
Session 7	n	22	21	13	8
	$\bar{x}$	9.0	16.4	8.5	14.9
	s	9.0	17.9	7.2	18.1
Session 8	n	20	19	13	8
	$\bar{x}$	7.3	20.9	7.5	16.8
	s	7.3	23.8	8.2	18.9
Session 9	n	22	20	13	8
	$\bar{x}$	7.5	13.3	7.9	10.5
	s	7.5	11.9	7.7	10.1
Session 10	n	24	24	13	8
	$\bar{x}$	7.9	19.4	6.6	19.6
	s	7.9	22.2	3.3	23.2

The number of actions for the speech interaction technique varies in the same range from session 4 onwards. Sessions 2 and 3 have a slightly higher number of actions and session 6 has the lowest mean number of actions for the speech interaction technique. The number of actions for the keyboard, on the other hand, fails to

stabilise and continues rising and falling sharply throughout the sessions (see Chart 6.16 for the mean graph). Nevertheless, the keyboard task has, on average, more actions than that of the speech interaction techniques for all sessions.

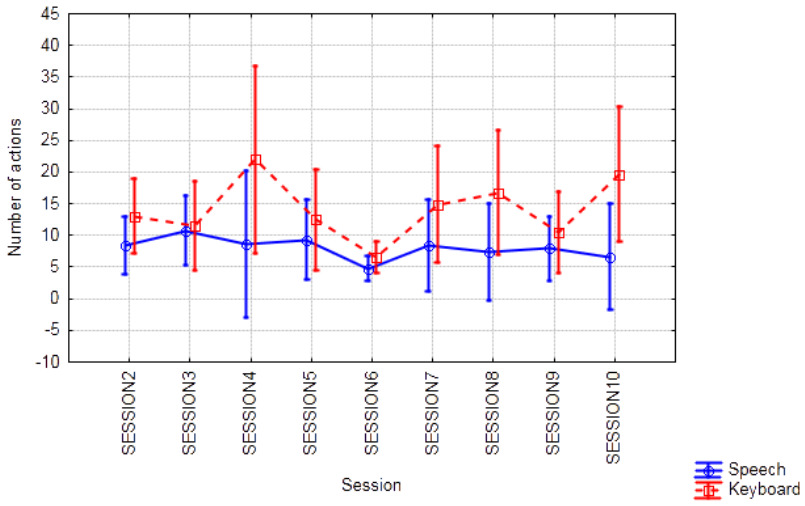


Chart 6.16: Mean for the number of actions to select and copy text

The hypotheses below were formulated to determine if these differences were non-significant:

1. $H_{0,1}$: The interaction technique has no effect on the number of actions required to complete the task.
2. $H_{0,2}$: The number of actions does not differ between the sessions.

The assumption of sphericity was not met by the spread of data for the number of actions ($\chi^2(35) = 126.721$, $p < 0.05$). Table 6.26 below contains the results of the required analyses to evaluate the afore-mentioned hypotheses.

Table 6.26: Analysis results for the number of actions required to select and copy text

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(1, 19) = 3.498, p > 0.05			
Session	F(8, 152) = 1.378, p > 0.05	F(3.2, 60.4) = 1.378, p > 0.05	F(4.1, 77.8) = 1.378, p > 0.05	F(8, 12) = 1.801, p > 0.05
Interaction technique × Session	F(8, 152) = 1.099, p > 0.05	F(3.2, 60.4) = 1.099, p > 0.05	F(4.1, 77.8) = 1.099, p > 0.05	F(8, 12) = 0.074, p > 0.05

Neither $H_{0,1}$ nor $H_{0,2}$ could be rejected at an α-level of 0.05. Therefore, it could be concluded that neither the interaction technique nor the session affects the number of actions required to complete the task. Similar to a previous task, this task displays the same phenomenon that the time for the speech interaction technique is more than for the keyboard but that the actions are fewer. Nonetheless, the differences are not significant, so when selecting a word and copying an equivalent efficiency is achieved. The fact that the number of actions for the keyboard is higher than that for the speech could again be attributed to the selection technique used. Similar to previous selection tasks, the selection with the keyboard is achieved by selecting the characters individually instead of using a more efficient means such as the combination of the [**Control**] and [**Shift**] keys or

a mouse selection. However, since the difference is not significant, the selection method used is of little consequence to the efficiency of task completion in terms of the number of actions.

An observation that was made during data capturing was that participants expected some sort of feedback when issuing the verbal copy command even though there is no such feedback with the counterpart action for the keyboard. This may well be due to the fact that the user will at least be sure that they had clicked the correct menu option or used the correct keyboard shortcut but they could not be sure that the speech command had been correctly interpreted by the speech engine. Therefore, it becomes imperative that for commands with no visible result there must be feedback of some sort so that the user can be reassured that the command has been executed.

6.7.6.3 Correctness of task completion

The steps required to complete this task were as follows:

1. A portion of text must be selected.
2. Specifically the last word on the current line must be selected.
3. The selection must be copied to the clipboard.

Chart 6.17 below is a stacked bar graph which shows the number of participants in each score category for both interaction techniques and all sessions.

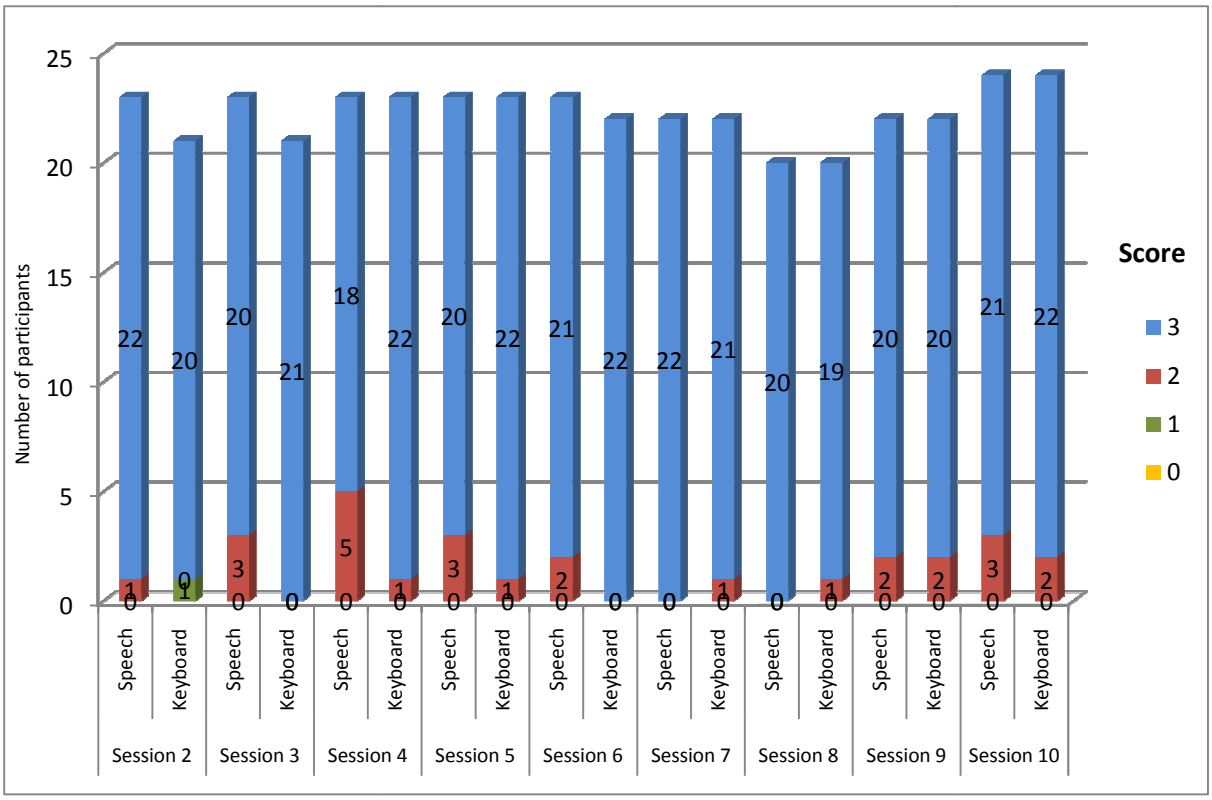


Chart 6.17: Correctness of task completion - Select word and copy

Similar to previous tasks, the majority of the participants completed the task 100% correctly with either interaction technique and from the very first session. As with the previous selection tasks, the reason for the lower scores was usually because the participant selected text other than that was specified in the task instruction. This happened for both interaction techniques and should not significantly affect the use of the interaction techniques.

6.7.8 Position and Paste

This task required that the previously copied word be pasted after the second word of the current line. Therefore, both tasks required that the cursor be correctly positioned and then the contents of the clipboard had to be inserted at that position. Again, the minimum number of actions required to complete the tasks were similar for the different interaction techniques.

6.7.8.1 Time to complete the task

The table below summarises the descriptive statistics for the completion rate of the task.

Table 6.27: Descriptive statistics for completion time to position cursor and paste text

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	21	22	11	13
	$\bar{x}$	53.0	23.1	59.1	19.8
	s	40.7	10.1	47.6	8.7
Session 3	n	21	22	11	13
	$\bar{x}$	28.1	17.8	22.9	14.7
	s	15.4	7.7	11.5	4.7
Session 4	n	23	24	11	13
	$\bar{x}$	38.8	14.4	31.5	11.9
	s	31.6	8.8	20.0	4.0
Session 5	n	22	23	11	13
	$\bar{x}$	25.8	14.7	26.0	13.4
	s	16.1	8.0	13.6	7.4
Session 6	n	22	23	11	13
	$\bar{x}$	31.9	11.9	21.8	10.8
	s	18.1	3.6	11.7	3.1
Session 7	n	22	22	11	13
	$\bar{x}$	29.4	14.3	21.1	14.0
	s	18.0	5.9	11.3	6.4
Session 8	n	20	20	11	13
	$\bar{x}$	27.4	13.6	27.1	12.7
	s	15.0	5.8	17.3	6.5
Session 9	n	22	22	11	13
	$\bar{x}$	25.9	12.2	21.2	12.2
	s	15.2	4.0	8.6	4.7
Session 10	n	24	24	11	13
	$\bar{x}$	24.8	12.1	23.7	11.3
	s	17.5	4.3	16.2	4.1

Sessions 2 and 4 are the only sessions where the speech interaction technique has a completion time that does not appear to be comparable to that of the keyboard. Nevertheless, throughout all the sessions, the speech

interaction technique had a higher average completion time than the keyboard and mouse. Similar to previous tasks, there is continual improvement in the completion rate for the speech interaction technique as exposure to the application is prolonged. Chart 6.18 provides a plot of the means for the two interaction techniques over all the sessions.

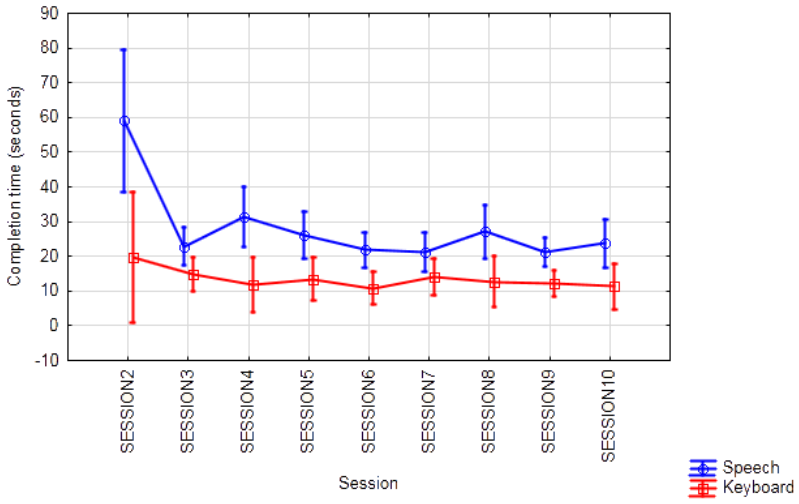


Chart 6.18: Mean plot for completion time to position cursor and paste text

The following hypotheses will be used to determine whether the difference is significant:

1. $H_{0,1}$: The interaction technique has no effect on the time taken to complete the task.
2. $H_{0,2}$: The time taken to complete the task does not differ significantly between the sessions.

The assumption of sphericity was not met ($\chi^2(35) = 71.833$, $p < 0.05$), therefore Table 6.28 contains the results of the ANOVA and the multivariate tests as well as the adjusted corrections.

Table 6.28: Analysis results for completion time to position cursor and paste text

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 22) = 15.448$, $p < 0.05$			
Session	$F(8, 176) = 5.123$, $p < 0.05$	$F(4.0, 89.0) = 5.123$, $p < 0.05$	$F(5.3, 116.4) = 5.123$, $p < 0.05$	$F(8, 15) = 5.705$, $p > 0.05$
Interaction technique × Session	$F(8, 176) = 0.936$, $p > 0.05$	$F(4.0, 89.0) = 0.936$, $p > 0.05$	$F(5.3, 116.4) = 0.936$, $p > 0.05$	$F(8, 15) = 1.986$, $p > 0.05$

From the table above it can be extrapolated that both $H_{0,1}$ and $H_{0,2}$ could be rejected. This leads to the conclusion that the interaction technique does significantly affect the time taken to complete the task; specifically when using the keyboard and mouse, the task can be completed in a noticeably faster time than when using the speech interaction technique. Tukey's HSD post-hoc test showed that session 2 differed significantly from all other sessions where session 2 required a longer time to complete the task than any other session. Once again, the fact that the first session where the test was completed required more time to complete, can be attributed to the participant's lack of experience with the application.

6.7.8.2 Number of actions

The task could be completed using the same minimum number of actions for both interaction techniques. However, closer inspection of Table 6.29 shows that the speech interaction technique resulted in more actions being performed in order to complete the task. This observation holds for all sessions, although it is encouraging to see that the number of actions decreased with each session and stabilises within the same range from session 5 onwards.

Table 6.29: Descriptive statistics for the number of actions to position the cursor and paste text

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	21	22	9	9
	$\bar{x}$	25.1	6.6	21.3	6.3
	s	23.3	4.7	10.8	3.6
Session 3	n	21	22	9	9
	$\bar{x}$	12.2	5.4	9.7	4.6
	s	6.8	3.1	6.3	1.7
Session 4	n	24	21	9	9
	$\bar{x}$	21.2	4.1	15.2	4.1
	s	20.6	2.6	10.8	1.5
Session 5	n	22	21	9	9
	$\bar{x}$	10.9	4.6	10.4	3.4
	s	7.5	3.9	6.1	1.9
Session 6	n	22	20	9	9
	$\bar{x}$	11.5	3.3	9.0	2.8
	s	6.3	1.6	6.1	0.8
Session 7	n	22	22	9	9
	$\bar{x}$	11.1	5.2	7.8	5.9
	s	6.3	4.2	4.5	4.3
Session 8	n	20	20	9	9
	$\bar{x}$	9.4	4.0	7.6	3.7
	s	7.7	2.9	4.2	1.9
Session 9	n	22	21	9	9
	$\bar{x}$	10.2	4.6	7.1	5.3
	s	6.6	2.3	4.2	3.0
Session 10	n	24	23	9	9
	$\bar{x}$	9.2	4.0	7.9	3.7
	s	6.3	1.8	4.6	1.7

Chart 6.19 plots the means for the number of actions over all sessions.

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

1. $H_{0,1}$: The interaction technique does not significantly affect the number of actions required to complete the task.
2. $H_{0,2}$: The session has no effect on the number of actions performed to complete the task.

The repeated-measures within-subjects ANOVA showed that there was significant interaction ($F(8, 128) = 4.256$, $p < 0.05$) between the two factors at an α-level of 0.05. Therefore, in order to compensate for the effect of the interaction, separate ANOVAs had to be performed for one factor whilst the other factor is controlled for.

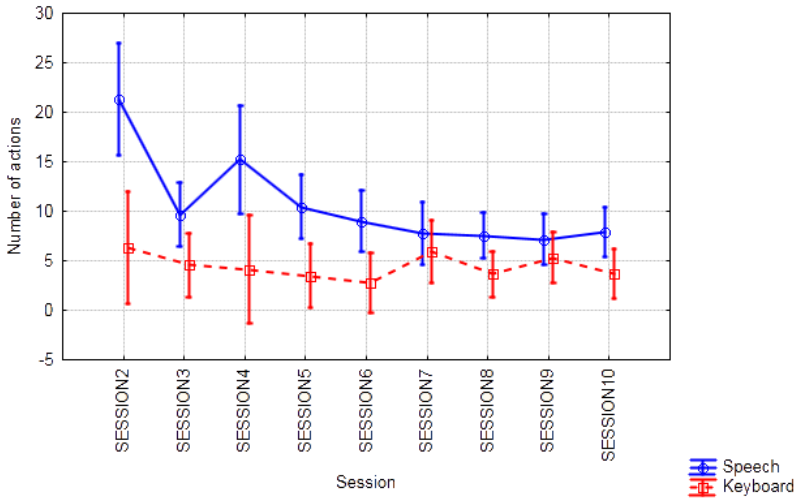


Chart 6.19: Mean number of actions to position the cursor and paste text

When investigating the first null hypothesis, it was found that there was a significant difference between the interaction techniques over all sessions. In particular, this means that the speech commands required significantly more actions than the keyboard. Even though the number of actions decreased over the sessions, which indicates learning, the learning does 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” to move the cursor to the right one word at time. This would offer a far quicker way to move the cursor and would contribute only a single action to the task completion. Since the keyboard and mouse offer 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 the most efficient method of task completion. Moreover, it could mean that they do not explore the use of commands that may yield the same final result but not the same intermediate results. In other words, “Select word” followed by the command “Right” will allow a user to move to the right, one word at a time yielding the desired final result. However, since it selects the word it does not have the same intermediate result of moving the cursor that the “Right” command does and it may have escaped the notice of participants as a possible command to complete the task. Familiarity with general cursor movement seems the obvious choice for the participants within the scope of the task and they failed to explore more efficient commands. This observation may also hold for the previous tasks which could also be completed with fewer actions but a more obscure string of actions, but where participants still navigated using a simpler but longer string of commands. The shorter route for completing the task was communicated to the participants if they were unable to discern this themselves. This could account for the lower number of mean actions as the exposure increased.

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. Similarly, $H_{0,2}$ could be rejected for the keyboard and mouse ($F(8, 64) = 2.287$, $p < 0.05$). Tukey’s HSD post-hoc test did not highlight any significant difference between any sessions; therefore the less conservative Fisher’s LSD post-hoc test was used. This showed that session 2 differed significantly from sessions 5, 6, 8 and 10. Session 9 also differed significantly from session 6. Therefore, there was a noticeable effect of learning between session 2 and the remainder of the sessions and since session 2 was the first time the tasks were completed, the higher number of actions

could easily be attributed to the first-time use of the application. The shorter means mentioned previously of positioning the cursor was communicated to participants and while there were fewer incidents of moving the cursor a character at a time, this still occurred during all sessions. Nevertheless, the number of actions decreased with significant effect between the first and latter sessions and this does indicate an increased familiarity with the application such that a significant learning effect is observed.

6.7.8.3 Correctness of task completion

The task correctness was evaluated according to the following criteria:

1. An insertion took place, regardless of its position.
2. The insertion was after the second word.

The chart below is a stacked graph showing the results of both interaction techniques for all sessions.

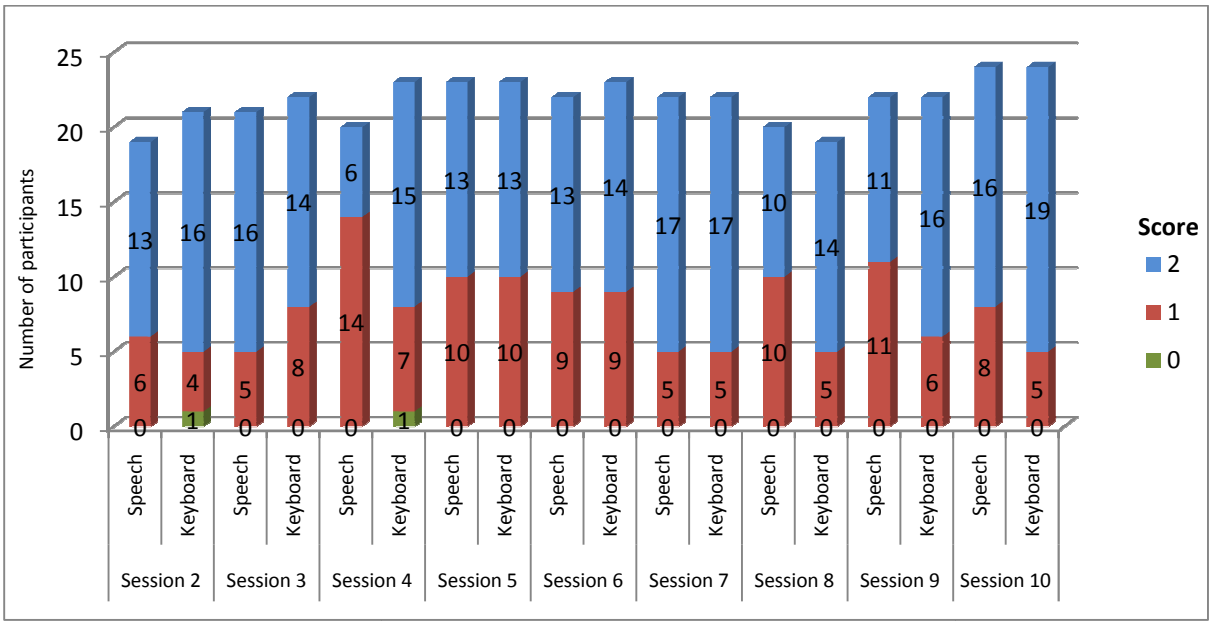


Chart 6.20: Correctness of task completion - Position and paste

This task has the poorest results in terms of correctness of task completion. There is an approximately even split between participants scoring 1 and those scoring 2. The main reason for this was the fact that the task instruction was to paste the word *after* the second word in the current line. Most participants pasted the text as the second word. Therefore, the wording of the task was perhaps not so clear which resulted in a lower correctness for the task. Consequently, it is not the interaction technique which caused the lower correctness and it can be concluded that the interaction technique does not affect the correctness of task completion when positioning and pasting.

6.7.9 Select all and format

This task required that all the text in the document be selected and italicised through the use of speech commands. There was no counterpart using the keyboard and mouse. Therefore, the analysis had to determine whether there was an improvement or decline in the time taken to complete the task over the sessions.

6.7.9.1 Time to complete task

The number of participants who completed the task, the mean time required to complete the task and the standard deviation for the data are summarised in Table 6.30.

Table 6.30: Descriptive statistics for the completion time to select and format all text

		All participants	Participants completing all sessions
		Speech	Speech
Session 2	n	25	13
	x̄	30.7	27.1
	s	20.9	17.6
Session 3	n	23	13
	x̄	18.0	16.0
	s	9.6	9.3
Session 4	n	24	13
	x̄	15.1	12.0
	s	7.3	5.2
Session 5	n	23	13
	x̄	12.9	11.5
	s	5.1	3.6
Session 6	n	24	13
	x̄	13.2	13.2
	s	8.3	9.6
Session 7	n	22	13
	x̄	9.9	9.9
	s	3.2	3.9
Session 8	n	20	13
	x̄	12.9	11.8
	s	5.2	5.4
Session 9	n	22	13
	x̄	10.5	10.4
	s	5.4	5.7
Session 10	n	24	13
	x̄	9.8	10.4
	s	4.8	6.4

Chart 6.21 provides a visual representation of the spread of data as it is a plot of the means for all sessions.

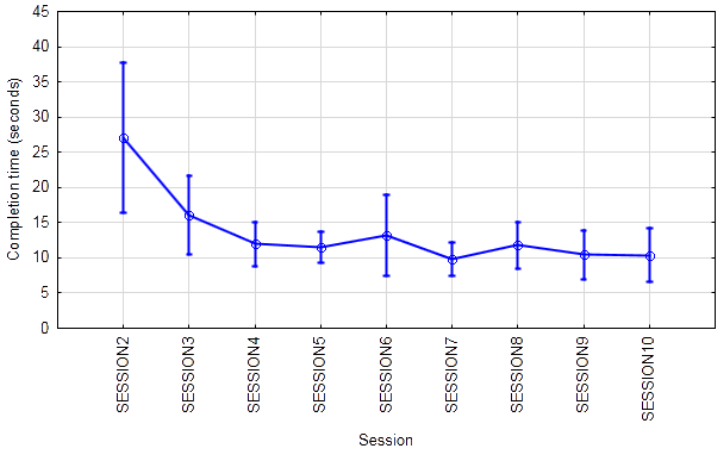


Chart 6.21: Means for the completion time to select and format all text

The table and chart show that there is a steady decrease in the completion time of the task for the first few sessions, thereafter stabilising to a relatively steady pace. The following hypothesis was evaluated:

H_0: The session has no effect on the task completion time.

The assumption of sphericity ($\chi^2(35) = 24.843$, $p > 0.05$) was met and H_0 ($F(8, 96) = 5.351$, $p < 0.05$) could be rejected at an α-level of 0.05. Therefore, there was a significant decrease in completion time. In particular, the completion time of session 2 was significantly higher than the time for session 4 to 10. This provides evidence that there was significant improvement in task completion time as exposure to the application increased.

6.7.9.2 Number of actions

The difference in the number of actions was also analysed in order to determine whether there were any differences between the sessions. The table below summarises the descriptive statistics for the number of actions performed. The underlying chart is a plot of the mean number of actions for all sessions.

Table 6.31: Descriptive statistics for the number of actions to select and format all text

		All participants	Participants completing all sessions
		Speech	Speech
Session 2	n	25	14
	$\bar{x}$	10.7	8.9
	s	8.7	6.3
Session 3	n	23	14
	$\bar{x}$	6.3	6.0
	s	3.9	4.6
Session 4	n	24	14
	$\bar{x}$	5.3	4.6
	s	3.2	3.1
Session 5	n	23	14
	$\bar{x}$	4.7	4.2
	s	2.8	2.7
Session 6	n	24	14
	$\bar{x}$	4.8	4.7
	s	3.3	3.5
Session 7	n	22	14
	$\bar{x}$	3.8	3.9
	s	1.7	1.8
Session 8	n	20	14
	$\bar{x}$	5.6	5.1
	s	2.7	2.8
Session 9	n	22	14
	$\bar{x}$	4.0	4.3
	s	2.5	3.0
Session 10	n	24	14
	$\bar{x}$	4.5	4.9
	s	2.8	3.4

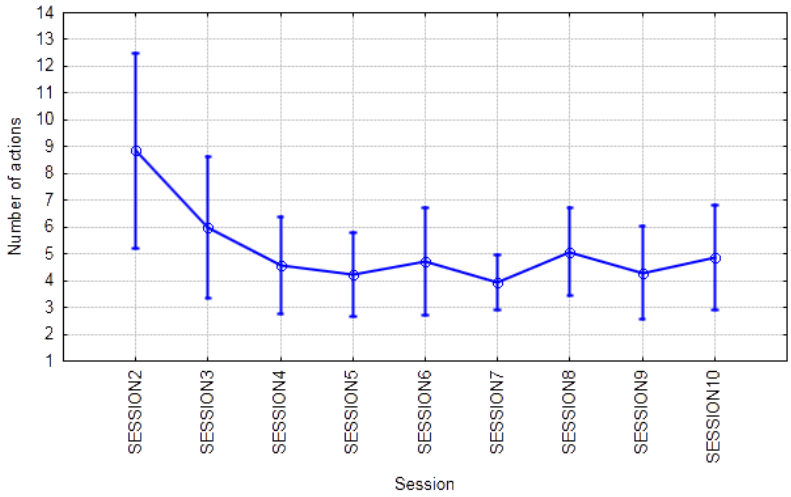


Chart 6.22: Mean number of actions to select and format all text

As with the time, there is a sharp decline in the number of actions performed in the first few sessions, after which the mean stabilises within a certain range. The following hypothesis was formulated:

H_0: The session has no effect on the number of actions required to complete the task.

The assumption of sphericity required for a repeated-measures ANOVA was violated ($\chi^2(35) = 54.057$, $p < 0.05$). H_0 could be rejected at an α-level of 0.05 ($F(8, 104) = 2.718$, $p < 0.05$) meaning that the session does have an impact on the number of actions required to complete the task. The adjusted corrections were as follows: $F(4.1, 53.1) = 2.718$, $p < 0.05$ and $F(6.2, 80.5) = 2.718$, $p < 0.05$.

Tukey's HSD post-hoc test indicated that session 2 differed significantly from sessions 4 to 7 and 9. Session 2 required significantly more actions to complete the task than was the case in the other sessions. Overall, this shows that only the first session where the tasks were completed had a significantly higher completion time and number of actions. This is sufficient evidence to show that there is ample improvement in the efficiency of task completion.

6.7.9.3 Correctness of task completion

This task had the following criteria against which it was evaluated:

1. A portion of text had to be selected.
2. The selection should include all text in the document.
3. Italic formatting must be applied.

All participants, apart from one individual in session 3 who scored 2, managed to complete this task correctly for all sessions. Therefore, the interaction technique allows for the task to be completed correctly from the first session.

6.8 Summary of results

This chapter contains a substantial number of analyses and results, and this section will be used to summarise these findings in a comprehensive but much more succinct way. Table 6.32 contains a summary of the results

for each of the tasks. Where there was a significant difference detected between the interaction techniques an S will be used if the speech interaction technique had a lower mean completion measurement and a K will be used if the keyboard and mouse had a lower mean completion measurement. A blank cell indicates that there was no difference between the interaction techniques. The same technique will be used for the session, where a tick indicates there was improvement over the sessions for both interaction techniques. Note that while there are significant differences indicated for the completion times between sessions, this does not imply that all sessions differed significantly but rather that there were some sessions which facilitated faster completion times as exposure increased.

Table 6.32: Summary of significant results

	$H_{0,1}$: Interaction technique		$H_{0,2}$: Session	
	Completion time	Number of actions	Completion time	Number of actions
Line selection and formatting			✓	S
Select all and remove	S	S	✓	✓
Select words and format	K		✓	✓
Paste	S	S	✓	
Undo			✓	✓
Select word and copy			✓	
Position and paste	K	K	✓	✓

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 fewer 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 measurements.

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 that there was 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. At the very least the final sessions usually showed significantly better performance than the first sessions. The fact that improvement is shown over the sessions is testimony to the fact that the speech commands are easy to learn and remember. Few participants had to refer back to the command list provided after the first session and could manage to complete the tasks with speeds comparable to those achieved with the keyboard and mouse. Therefore, the use of a menu-orientated grammar allowed the speedy adoption thereof and did not appear to place additional strain on the users. Although the menu-orientated grammar was not compared with a task-orientated grammar, the fact that the grammar was so quickly learnt is motivation enough to recommend the use of a menu-orientated grammar for a word processor. Even though users tend to resort to task-orientated commands when faced with a complex task (Berg et al., 2010), the assumption that was made that the terminology of the word processor is unique and part of the initial learning process was proven in the current study. It is therefore, surmised that complex tasks will also be facilitated in this way since even under those circumstances there is a unique grammar already in place which can quickly be learnt by the users.

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 − the most effective method is not always chosen. Rather the user chooses the method which results in an intermediate action which is closer to the final result even though in reality there is a shorter method that can be used. However, the immediate responses may be contradictory to the final solution or

simply not appear to move the user closer to the desired goal. For example, in order to position the cursor, the user may prefer to move the cursor one character in the desired direction consecutive times since the intermediate results move the task closer to completion. In contrast, selection of a word does not immediately appear to move the cursor when in effect it does and at a much faster rate than moving one character at a time.

Successive tasks that make use of similar speech commands also appear to show that the initial task jogs the memory of the participant for the current session to such an extent that the subsequent tasks can be completed more efficiently. This observation was however not analysed statistically and is only suspicion based on the available data and observations of the participants during test completion.

The fact that the speech commands resulted in fewer actions may be attributed to the fact that the grammar that was used was fairly simple and provided commands to complete basic operations only. The complexity of the options provided by Word is much higher than that accommodated in the grammar. This may have led to expedited speeds and actions to complete tasks as there was, in many instances, only a single command available to complete a task. In contrast, when using Word in the normal capacity there is, more often than not, at least three 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. However, judging by the results tabulated above, this did not have an impact on the results as, in most instances, the interaction techniques could perform at a level that was comparable. The tasks where the speech outperformed the keyboard perhaps had the most intuitive speech commands: this could explain why the speech task could be completed quicker and with fewer actions than its keyboard counterpart. The remainder of the grammar was less intuitive and may have required some time to memorise and learn which could account for the fact that the speech was not, in these instances, faster than the keyboard. However, in these instances it was at least comparable to the keyboard.

In terms of the correctness of the tasks, there was very little difference between the interaction techniques. Furthermore, the majority of the tasks could be completed correctly from the very first session and using either interaction technique. The instances where there were tasks which were not completed correctly were largely due to the participants not reading the instructions properly or the fact that the wording of the instructions may not have been very clear. Owing to the large number of zeros in the categories, meaningful statistical analysis could not be conducted. Even so, it could be inferred that the interaction technique did not affect the level of correctness with which the task could be completed. Furthermore, since high correctness measurements were achieved from the very first session there was also no learning required in order to complete the task correctly.

Overall, this study disproved the notion that speech commands cannot be used effectively and efficiently in an editing environment (Klarlund, 2003). The efficiency was for the most part similar or superior to the keyboard and mouse. Effectiveness for the two interaction techniques was on an equivalent level. Therefore, there is confirmation of the findings by Karl et al. (1993) that speech is able to provide a more efficient and effective means of completing word processing tasks. The study of Karl et al. (1993) allowed the modalities of speech and mouse/keyboard to be mixed and also did not provide for text selection using speech commands. The current study therefore also improved on these prior findings.

6.9 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 less 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 grammar which includes only a smaller grammar and then, when required by the user, the extended grammar can be activated.

The results of the study indicate that using 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. The participant would then be able to issue verbal commands during standard interaction with the application to apply formatting, corrections and move around the document. Usability measures for such a task can then be compared to measurements recorded when speech commands are not available but the participant has to complete the same task. This would give a clearer indication as to whether the incorporation of speech commands in a word processing application is a viable alternative to the mouse and keyboard.

Whether or not an extended grammar is considered, further research will have to be done where the exposure to the application is extended in order to determine whether the learning effect can continue to an even greater degree. This could mean that the speech could perform on a similar level to the mouse and keyboard on a number of other tasks – or eventually even better. Such a study could use a smaller sample as it has already been established that it is possible to use this interaction technique effectively.

6.10 Summary

This chapter reported on the results of similar 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. The correctness with which the tasks could be completed was also measured and analysed. For the majority of the tasks it was found that the interaction techniques could compete on a comparable level, particularly as the time the participant used the application increased. Therefore, there was a definite improvement in user performance as the use of the application was extended. This indicates that the application was indeed learnable. Since the speech interaction techniques could also be used with the same efficiency as the keyboard and mouse, the proposed use of speech commands within a word processor application is viable. The correctness with which the task could be completed was neither affected by the interaction technique nor the amount of exposure to the system.

In conclusion, although the interaction technique affects the time with which some tasks can be completed and sometimes the number of actions required it does not affect the correctness with which the task is completed. Additionally, the time to complete a task generally improved as exposure to the application increased as did the number of actions required. However, the correctness of the task completion was high from the very start so although it took longer and required more effort in the first sessions, the tasks were still completed correctly.

The following chapter will report on the analysis of using the onscreen keyboard to type through using eye gaze and speech recognition as an interaction technique.

CHAPTER 7

ANALYSIS OF TYPING TASKS

7.1 Introduction

The previous chapter concentrated on the analysis of the use of speech commands for formatting, navigating a document and other common word processor tasks. During the longitudinal testing, participants were also required to enter text using both the keyboard and eye gaze and speech (analogous to look-and-shoot). The buttons on the onscreen keyboard used varying sizes and spacing for the typing tasks. This chapter will analyse and discuss the effectiveness and efficiency of eye gaze and speech when used for text input, as compared to a traditional keyboard.

7.2 Participants

Since these tasks were part of the task list set out for the longitudinal testing of the multimodal Word interface, the participants for this analysis were the same as in the previous chapter. Therefore, there were 25 participants who completed the typing tasks. There were, however, three participants who were unable to type using eye gaze and speech for various reasons. The first participant was unable to maintain a stable eye gaze on any of the buttons on the onscreen keyboard. This behaviour was observed for all ten sessions and since there was no improvement and the participant was unable to type even a single character, this participant will not be included in this chapter's analysis. The second participant experienced the same problem as the first. This participant wore glasses with thick lenses and a very wide frame. Although an acceptable calibration was achieved by the participant, he was unable to select any of the onscreen buttons. It is quite possible that his glasses interfered with his ability to type. Therefore, his data was also excluded from the analysis in this chapter.

The third participant could manage to maintain a stable eye gaze on the onscreen buttons but the speech engine was unable to recognise the commands he issued to select the button. The participant had an unusual pronunciation of some words and also did not enunciate very clearly. Repeated measures were taken to attempt to correct this. Firstly, this participant completed additional training sessions to improve the accuracy of his speech profile. When this was unsuccessful, special commands were added specifically for this participant but while this was initially successful, the participant quickly slipped back into his normal speaking tone and his enunciation of the special commands was degraded to such a degree that they no longer worked. Therefore, it was considered prudent rather to discard the data of this participant.

Consequently, the sample size for the typing tasks was twenty-two, comprising of 14 males and 8 females. There were 6 English-speaking participants, 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) and there were 9 Computer Science students and 13 non-Computer Science students.

7.3 Tasks

In total there were two typing tasks using the keyboard and three using the eye gaze and speech. 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 gravitational well of 20 (≈0.52 ° visual angle) pixels on all sides of each button. Since the

results of Chapter 6 showed that the gravitational well was the most effective means of increasing the usability of eye gaze and speech as a pointing device, a gravitational well was included in the onscreen keyboard. The larger the gravitational well is, the more widely spaced the buttons must be. Consequently, no screen real estate is gained through the use of a gravitational well and in order to optimise the aesthetic appeal of the onscreen keyboard it was decided rather to decrease the gravitational well so that the buttons could be closer together and then to enlarge the buttons to make selection easier.

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. Therefore, the analysis included two typing tasks to be completed with a keyboard and two with eye gaze and speech.

Additional typing tasks were added from the fifth session onwards in order to test varying sizes and spacing between buttons. These additional tasks were added to the end of the existing task list. By then the majority of the participants were completing the current task list in less than 30 minutes. No pressure was placed on the participants to complete all tasks within their scheduled time so it was felt that adding additional tasks to the end of the test would not unduly cause any more anxiety or place more strain on the participants. 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) pixels in size and spaced 70 (≈1.80° visual angle) pixels apart. Following this there were another two tasks which had to be completed using buttons that were 50×50 pixels in size but were spaced 60 pixels apart. For all typing tasks a gravitational well of 20 pixels on all sides of the buttons were employed.

The use of the keyboard will be denoted by K and the larger originally sized buttons by Speech-L. Speech-L was used for the first two typing tasks using eye gaze and speech for text input. From session 5 onwards there were two typing tasks using smaller buttons which were widely spaced, which will be denoted by Speech-SW, and then a final two using smaller buttons which were spaced closer together, namely Speech-SC.

All text that had to be typed was selected randomly for each task from the set of 35 pre-selected phrases (Section 3.4.3.3). Similar to the previous tasks, all the typing tasks were displayed to the participant using a window overlaid over the word processor application.

7.4 Measurements

The measurements that were selected for analysis were the character error rate and the characters typed per second. Since both input methods, namely typing with the traditional and the onscreen keyboard, were character based, the character error rate and characters per second are a more applicable means of measuring the effectiveness and efficiency of the interaction technique (Read, 2005). The character error rate (CER) measures how many insertions, deletions and substitutions have taken place between the presented text and the transcribed text (Read, 2005). This measurement, which is effectively the minimum number of insertions, substitution and deletions, is synonymous with the Levenshtein distance between two strings. As discussed in section 3.4.3.2, the Levenshtein distance (Levenshtein, 1965) measures the difference between two strings in terms of the minimum number of insertions, substitutions and deletions required to transform one string (in this case the presented text) into another (in this case the transcribed text). This sum is then divided by the number of characters to give a character error rate (Read, 2005). Since there are multiple ways in which the presented text can be transformed into the transcribed text, using the same minimum number of edits, a more accurate means of calculating this character error rate is to determine the number of ways in which the transformation can occur (MacKenzie & Soukoreff, 2003). These possible transformations are called the

optimal alignments. Once these optimal alignments have been identified, their mean length is calculated and then the Levenshtein distance is divided by, this mean length to give an error rate (MacKenzie & Soukoreff, 2002). For example (the example is taken from MacKenzie & Soukoreff, 2002), suppose the presented text is the word “quickly” and the test participant types “qucehkly”. The Levenshtein difference between these two strings is 3, but there are four different ways in which “quickly” can be transformed into “qucehkly” by making only 3 errors. These four different ways are referred to as the optimal alignments. The mean length of these optimal alignments is then used to divide the Levenshtein distance by to give a more accurate error rate. This error rate measurement will be analysed in this chapter as a measure of effectiveness.

In order to measure the efficiency of the interaction techniques, the characters per second (CPS) measurement will be used. This measurement literally measures the number of characters that were typed and then divides it by the time taken to type the characters, measured 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, which is indistinguishable from the time taken to locate the first character that must be typed – which is then also excluded. As a consequence of measuring the time in this manner, the number of characters becomes n-1.

Corrections to transcribed text were not captured as it was felt that the correction would either be another key press or speech command that would be issued. Since these were analysed separately in the previous chapter it was decided not to include these measurements in the typing tasks as well.

7.5 Analysis

Analysis will first only be conducted between the tasks which used K and Speech-L. All the chosen measurements will be analysed for these tasks. Following this, all measurements will be analysed for the two typing tasks (K), the original two eye gaze and speech (Speech-L), the two typing tasks for the smaller buttons widely spaced (Speech-SW) and the two using the smaller buttons closer together (Speech-SC). Since speech-SW and speech-SC were only included from the fifth session onwards and it was not judicious simply to discard the first few sessions, it was decided to rather conduct two separate analyses.

7.5.1 Analysis of keyboard and large buttons

The error rate measurement was calculated individually for all typing tasks and then averaged over the typing tasks for each interaction technique and for each participant. As previously mentioned, the first typing task using speech and eye gaze was not considered for analysis as this was viewed as a practice task entry for each session. The participants were not informed of this and were instructed to complete all tasks to the best of their ability.

7.5.1.1 Error rate

The average error rates, as discussed in a prior section, for each participant and each interaction technique were calculated for all sessions. The descriptive statistics for the two keyboard typing tasks and the two speech-L typing tasks are tabulated below. The first line of each row is the number of participants who were included in the analysis, the second row the mean error rate and the third the standard deviation.

Table 7.1: Descriptive statistics for keyboard and speech-L error rate

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	17	21	8	13
	x̄	15.3	7.9	13.9	6.6
	s	7.7	6.7	4.6	5.6
Session 3	n	20	20	8	13
	x̄	17.7	5.3	18.8	6.4
	s	8.2	5.8	10.0	6.9
Session 4	n	20	21	8	13
	x̄	15.0	4.0	10.4	4.1
	s	11.2	4.3	6.9	4.7
Session 5	n	21	21	8	13
	x̄	14.2	6.5	10.1	4.7
	s	9.0	6.7	4.2	6.4
Session 6	n	20	21	8	13
	x̄	12.8	5.0	9.5	4.0
	s	7.2	5.0	6.4	3.6
Session 7	n	18	20	8	13
	x̄	12.2	4.9	8.2	4.8
	s	9.2	7.1	7.3	8.3
Session 8	n	17	18	8	13
	x̄	12.0	5.8	12.2	4.0
	s	6.5	5.2	5.6	4.8
Session 9	n	18	19	8	13
	x̄	9.3	3.4	6.3	2.6
	s	6.5	4.6	5.7	3.5
Session 10	n	17	21	8	13
	x̄	9.3	3.9	6.1	3.8
	s	5.9	3.8	4.9	4.2

Chart 7.1 is a plot of the mean error rate for the interaction technique across all sessions.

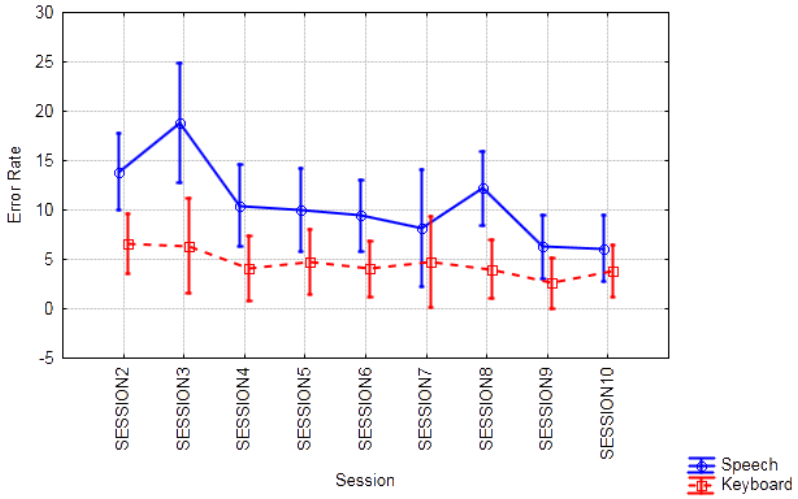


Chart 7.1: Mean error rate of keyboard and speech-L

From Table 7.1 and Chart 7.1 it can be extrapolated that the speech interaction, on average, caused a higher error rate than did the keyboard. This observation holds for all sessions, although the error rate for the speech interaction technique does improve steadily as the amount of exposure increases. The following hypotheses were formulated for this analysis:

1. $H_{0,1}$: The error rate is not affected by the interaction technique.
2. $H_{0,2}$: The error rate is not affected by the session during which the task was completed.

The same procedure as in the previous chapter was followed for the analysis of the data (see Section 3.5 for an explanation of data analysis)

In this instance, the assumption of sphericity ($\chi^2(35) = 54.795$, $p < 0.05$) was not met at an α-level of 0.05. Therefore, Table 7.2 reports the results of the repeated-measures ANOVA, the multivariate tests as well as the adjusted corrections.

Table 7.2: Results of error rate analysis for keyboard and speech-L

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 19) = 14.406$, $p < 0.05$			
Session	$F(8, 152) = 5.092$, $p < 0.05$	$F(4.8, 90.7) = 5.092$, $p < 0.05$	$F(6.9, 131.1) = 5.092$, $p < 0.05$	$F(8, 12) = 4.818$, $p < 0.05$
Interaction technique × Session	$F(8, 152) = 1.860$, $p > 0.05$	$F(4.8, 90.7) = 1.860$, $p > 0.05$	$F(6.9, 131.1) = 1.860$, $p > 0.05$	$F(8, 12) = 1.633$, $p > 0.05$

The results contained in the table above show that both null hypotheses could be rejected. Therefore, the interaction technique had a significant effect on the error rate of the typed sentence. Since the keyboard had a consistently lower average than the eye gaze and speech this means that the use of eye gaze and speech interaction for text input results in a higher error rate.

Post-hoc tests were required to determine which sessions differed significantly. Tukey's honestly significant difference (HSD) test was used to establish the cause of the differences. It was found that session 2 differed significantly from session 9 and session 3 differed significantly from sessions 6, 7, 9 and 10. Since the average error rate for sessions 2 and 3 was higher than for the remainder of the sessions, this indicates some degree of learning over time. Although sequential sessions improved, the rate of improvement was not significant. However, the overall improvement from the first sessions to the last was significant. In particular, the first few sessions with speech-L differed significantly from all later sessions. The rate at which the eye gaze and speech interaction technique improves over time is an encouraging observation and hints that the error rate could possibly reach a comparable level with that of the keyboard. More typing sessions would have to be tested and analysed in order to verify this supposition.

The average error rate for each session was then inspected more closely to determine how many participants were able to type a completely error-free sentence. The results, for all participants, are shown in the bar graph below.

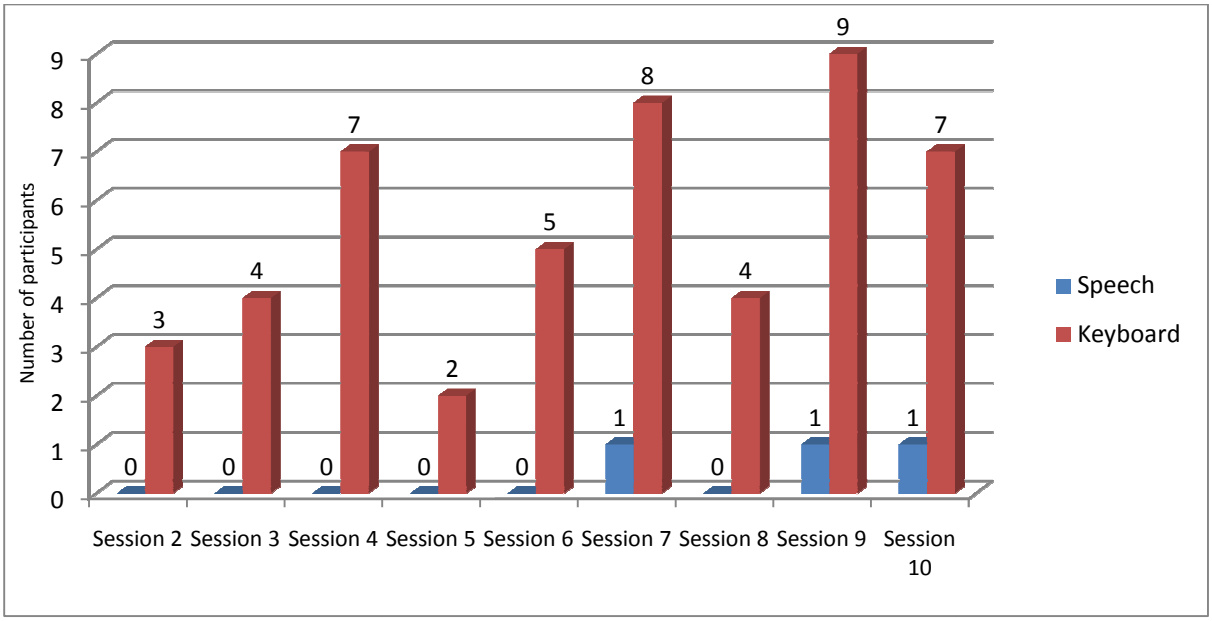


Chart 7.2: Error-free transcribed text for keyboard and speech-L

The graph shows that each session for the keyboard had at least two error-free transcribed text strings. There were only three sessions in which an error-free transcribed text string was achieved through the use of the eye gaze and speech and in each only by a single individual (although not the same individual more than once). This may not bode well for the effectiveness of speech-L for text input, although it is still entirely possible that prolonged use may result in decreased error rates as users will become accustomed to the use of the onscreen keyboard. Overall, it is the fact that the effectiveness increased as time went by that is of more importance than the fact that completely error-free transcription could not be achieved. If the effectiveness continues to improve, then eventually an error-free transcribed text should be achieved. Additionally, there are always mechanisms available to correct errors and should this be used then the end result of transcribed text may be error-free.

7.5.1.2 Breakdown of error rates

During ISO testing it was observed that the number of incorrect target clicks was significantly higher when using eye gaze and speech than when using the mouse. This was attributed to a tendency by the participants to acquire the intended target, start issuing the command and then move the eye gaze to the next intended target before the command had been processed. This resulted in the next target being selected instead of the designated target. If this behaviour was emulated during the typing tasks, it would manifest in a higher error rate, which was discovered to be the case. However, clicking the wrong target would specifically result in either an insertion error – if the participant realised the error and then inserted the correct character after the incorrect character, or a substitution error – if the participant did not realise the error and return to insert the correct character. Therefore, in order to determine whether this was true the error rate measured for the interaction techniques was broken down into the number of insertions, deletions and substitutions which could have occurred in order to transform the presented text into the transcribed text. Each of these was expressed as a percentage of the total error rate percentage as illustrated by MacKenzie and Soukoreff (2002).

For illustration purposes, the first session's data, as broken down into the categories of insertions, deletions and substitutions errors, is shown as a stacked bar graph in Chart 7.3. As can clearly be seen, the highest number of edits was insertions, followed by substitutions and finally deletions. This was true for both the

speech-L interaction technique as well as the keyboard interaction technique. More specifically, the eye gaze and speech had a total character error rate of 15.4% which consisted of 6.6% insertions, 5.3% substitutions and 3.4% deletions.

For further illustrative purposes, the same information is provided for the last session (Chart 7.3). While the average error rate for the speech decreased, the majority of the errors were still caused by insertions. The second most errors were substitutions followed closely by deletions. The same pattern was observed for the keyboard interaction technique.

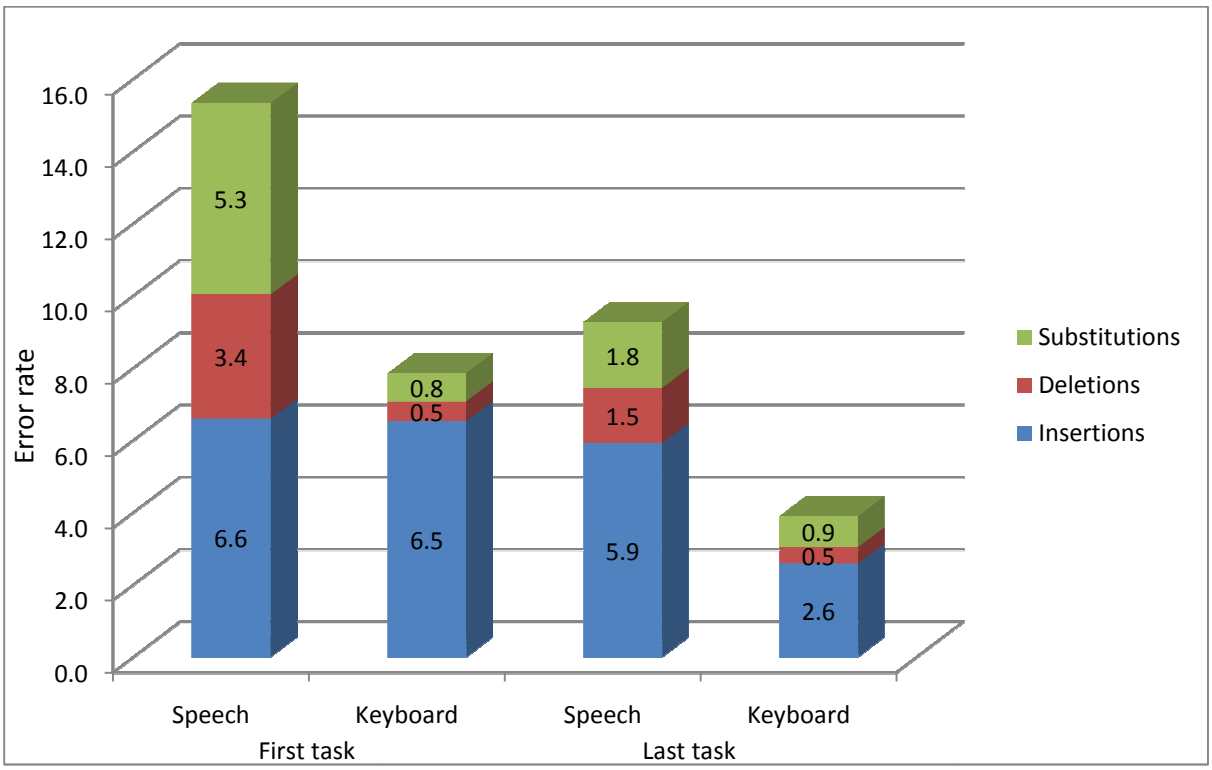


Chart 7.3: Breakdown of first and last task's error rates for keyboard and speech-L

Each of the types of edits was analysed separately, the results of which are discussed in the following sections.

7.5.1.2.1 Insertion error percentage

The ratio of insertions was calculated for each participant on each interaction technique and for all sessions (Table 7.3). These could then be analysed statistically to determine whether the afore-mentioned finding of incorrect clicks when using eye gaze and speech has an impact when typing.

Chart 7.4 shows the plot for the mean average insertion percentage for both interaction techniques over all sessions.

Table 7.3: Descriptive statistics for insertion errors of keyboard and speech-L

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	17	21	10	13
	x̄	6.6	6.5	6.5	6.3
	s	5.0	6.4	5.2	5.6
Session 3	n	20	20	10	13
	x̄	9.8	2.4	9.5	3.3
	s	6.9	2.8	7.5	3.1
Session 4	n	21	21	10	13
	x̄	5.8	2.9	6.0	2.8
	s	4.9	3.8	5.9	4.2
Session 5	n	21	21	10	13
	x̄	5.9	4.5	4.5	3.0
	s	4.6	4.3	4.8	3.4
Session 6	n	21	21	10	13
	x̄	5.7	3.7	5.3	3.2
	s	4.3	5.0	5.3	3.3
Session 7	n	19	20	10	13
	x̄	5.7	2.7	4.6	1.9
	s	4.7	3.5	4.6	3.2
Session 8	n	18	18	10	13
	x̄	7.9	3.7	8.0	2.7
	s	6.6	5.2	7.0	4.9
Session 9	n	19	19	10	13
	x̄	6.2	2.4	4.9	1.7
	s	4.3	3.4	4.6	2.6
Session 10	n	18	18	10	13
	x̄	6.2	2.6	4.4	2.2
	s	5.2	3.1	4.8	2.7

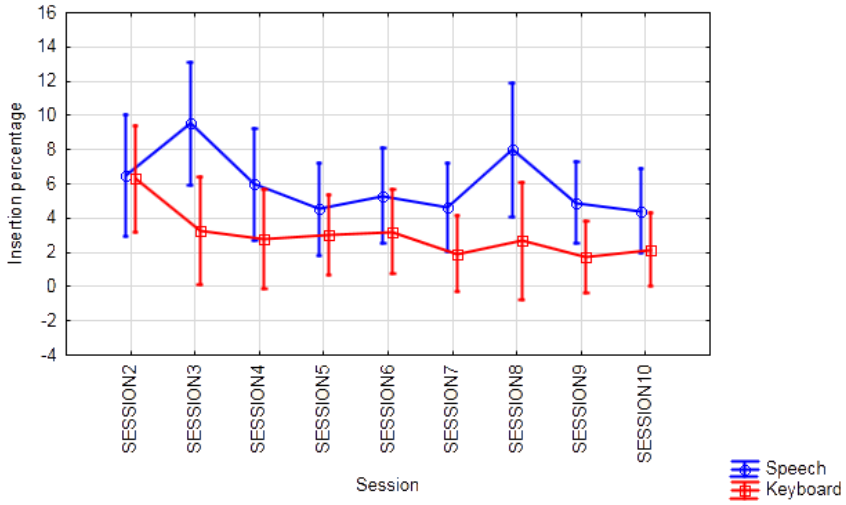


Chart 7.4: Mean insertion error percentage of keyboard and speech-L

From both the chart and the table it is clear that the speech-L interaction technique had a higher insertion error percentage than did the keyboard. There was, however, some improvement over the sessions, apart from session 8 where there was a sharp increase in the error rate. The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the insertion errors percentage.
2. $H_{0,2}$: The session has no effect on the insertion errors percentage.

The assumption of sphericity was violated ($\chi^2(35) = 61.167$, $p < 0.05$), therefore adjusted corrections were applied to the degrees of freedom. Table 7.4 contains the results of all the required analysis.

Table 7.4: Analysis results for insertion error percentage of keyboard and speech-L

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(1, 21) = 6.516$, $p < 0.05$			
Session	$F(8, 168) = 2.278$, $p < 0.05$	$F(5.1, 107.4) = 2.278$, $p < 0.05$	$F(7.3, 152.9) = 2.278$, $p < 0.05$	$F(8, 14) = 1.687$, $p > 0.05$
Interaction technique × Session	$F(8, 168) = 1.236$, $p > 0.05$	$F(5.1, 107.4) = 1.236$, $p > 0.05$	$F(7.3, 152.9) = 1.236$, $p > 0.05$	$F(8, 14) = 1.646$, $p > 0.05$

Using a 95% confidence interval, both $H_{0,1}$ and $H_{0,2}$ could be rejected. Therefore, there is a significant difference between the percentage of insertion errors made when using the different interaction techniques. Since the interaction technique of eye gaze and speech has, on average, more insertion errors, it could be concluded that the prior supposition was indeed correct.

Tukey's post-hoc test did not indicate significant differences between any sessions, but the less conservative Fisher's LSD test did. There was a significant difference between session 2 and sessions 5, 7, 9 and 10. Session 3 also differed significantly from sessions 7, 9 and 10. In particular, it was session 3 of the speech-L which was significantly higher than the majority of the other sessions with either interaction technique. Session 3 with the speech-L had a very high percentage of insertion errors, therefore during that session, the participants made significantly more insertion errors than in any other session with either interaction technique.

7.5.1.2.2 Substitution error percentage

The same procedure as in the previous section was used to determine the ratio of substitution errors. Descriptive statistics are tabulated below.

Using Chart 7.5 and Table 7.5 as a reference it is clear that for the first seven sessions, the eye gaze and speech averages a much higher substitution percentage than the keyboard. It was only during the final two sessions that the number of substitutions for the interaction techniques reached levels that are possibly comparable to one another.

The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the percentage of substitution errors made.
2. $H_{0,2}$: There is no difference between the percentages of substitution errors made between the sessions.

Table 7.5: Descriptive statistics for substitution error percentage of keyboard and speech-L

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	17	21	8	13
	$\bar{x}$	5.3	0.8	5.2	0.3
	s	5.4	1.3	4.6	0.7
Session 3	n	19	20	8	13
	$\bar{x}$	4.3	0.9	3.4	0.9
	s	4.6	2.3	4.6	2.4
Session 4	n	19	21	8	13
	$\bar{x}$	2.9	0.5	3.1	0.4
	s	3.4	0.9	2.9	0.9
Session 5	n	21	21	8	13
	$\bar{x}$	6.5	0.9	2.4	0.6
	s	7.5	1.8	3.0	1.5
Session 6	n	20	21	8	13
	$\bar{x}$	3.3	0.5	1.2	0.5
	s	4.9	1.0	2.4	0.9
Session 7	n	20	20	8	13
	$\bar{x}$	4.9	0.8	3.8	1.0
	s	5.2	1.8	4.6	2.1
Session 8	n	18	18	8	13
	$\bar{x}$	2.2	0.3	1.1	0
	s	2.5	0.8	1.7	0
Session 9	n	18	19	8	13
	$\bar{x}$	1.72	0.51	1.7	0.5
	s	2.51	1.28	1.9	1.2
Session 10	n	18	21	8	13
	$\bar{x}$	2.41	0.86	0.9	1.1
	s	3.42	1.73	1.3	2.0

When analysing the null hypothesis, it was found that there was significant interaction ($F(8, 152) = 2.205$, $p < 0.05$) between the two factors of interaction technique and session. Therefore, each session was analysed individually to determine whether there was a significant difference between the percentage of substitutions for each interaction technique during that session. It was found that $H_{0,1}$ could be rejected for all sessions other than the last two. Therefore, for sessions 2 to 8 the use of the speech-L interaction technique resulted in participants making significantly more substitution errors.

Since it was only the interaction technique of eye gaze and speech that was of interest, the second null hypothesis of no difference was only applied to the speech-L interaction technique. The percentage of substitution errors violated the assumption of sphericity ($\chi^2(35) = 54.808$, $p < 0.05$), thus Table 7.6 reports all required analysis that was performed on the data.

The second null hypothesis could not be rejected, therefore the percentage of substitution errors does not improve as the use of the eye gaze and speech for text input increases.

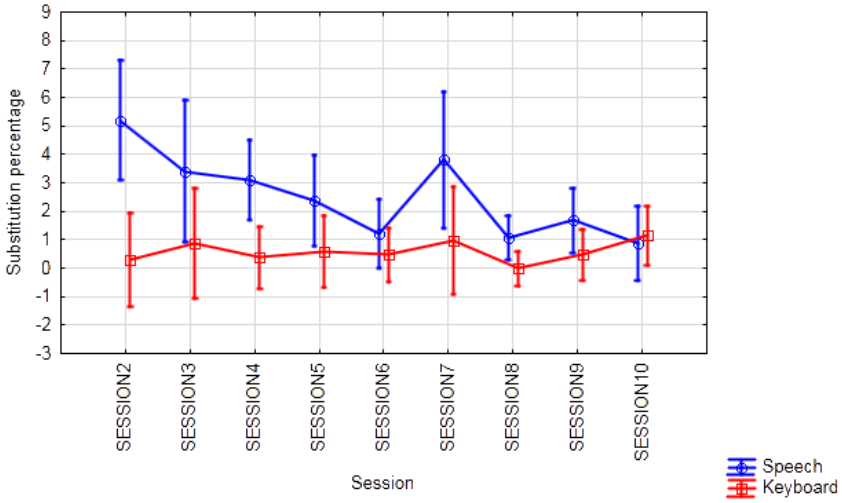


Chart 7.5: Mean substitution error percentage of keyboard and speech-L

Table 7.6: Results for the analysis of session for speech-L substitution errors percentage

	ANOVA	Geisser-Greenhouse	Huyn-Feldt
Session	F(8, 56) = 1.623, p > 0.05	F(2.8, 19.3) = 1.623, p > 0.05	F(4.7, 33.0) = 1.623, p > 0.05

7.5.1.2.3 Deletion error percentage

The final category of error percentages was the deletion percentage. The percentage of possible deletion errors was calculated for all participants for each session's typing tasks using both interaction techniques. Descriptive statistics for the data are summarised in Table 7.7.

Chart 7.6 and Table 7.7 indicate that the deletion percentages for the keyboard do not follow a generalised trend, but fluctuate erratically between sessions. The eye gaze and speech interaction technique is somewhat more stable across the sessions. From Chart 7.6 it appears as though the percentage of deletion errors is comparable for the keyboard and speech-L.

The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the percentage of deletion errors made.
2. $H_{0,2}$: The session has no effect on the percentage of deletion errors made.

Table 7.8 contains the results of the repeated-measures within-subjects ANOVA, the multivariate tests and the adjusted corrections since the assumption of sphericity was not met ($\chi^2(35) = 75.912$, $p < 0.05$).

Table 7.7: Descriptive statistics for the deletion error percentage of keyboard and speech-L

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	17	21	8	13
	$\bar{x}$	3.5	0.5	3.2	0
	s	4.1	1.5	3.6	0
Session 3	n	20	20	8	13
	$\bar{x}$	2.5	1.9	1.5	2.2
	s	3.7	3.9	2.5	4.7
Session 4	n	19	21	8	13
	$\bar{x}$	2.8	0.7	1.5	0.9
	s	4.0	1.6	2.8	2.0
Session 5	n	21	21	8	13
	$\bar{x}$	1.8	1.2	1.9	1.1
	s	2.7	2.9	3.0	3.4
Session 6	n	20	21	8	13
	$\bar{x}$	2.4	0.8	2.0	0.4
	s	4.3	1.7	4.1	0.7
Session 7	n	18	20	8	13
	$\bar{x}$	1.3	1.5	1.8	1.9
	s	2.9	3.2	4.1	3.9
Session 8	n	18	18	8	13
	$\bar{x}$	2.8	1.8	2.8	1.3
	s	4.7	2.5	4.5	2.2
Session 9	n	18	18	8	13
	$\bar{x}$	1.0	0.5	0.7	0.4
	s	1.5	1.1	1.0	1.0
Session 10	n	17	21	8	13
	$\bar{x}$	1.0	0.5	0.9	0.5
	s	2.0	0.8	2.1	0.8

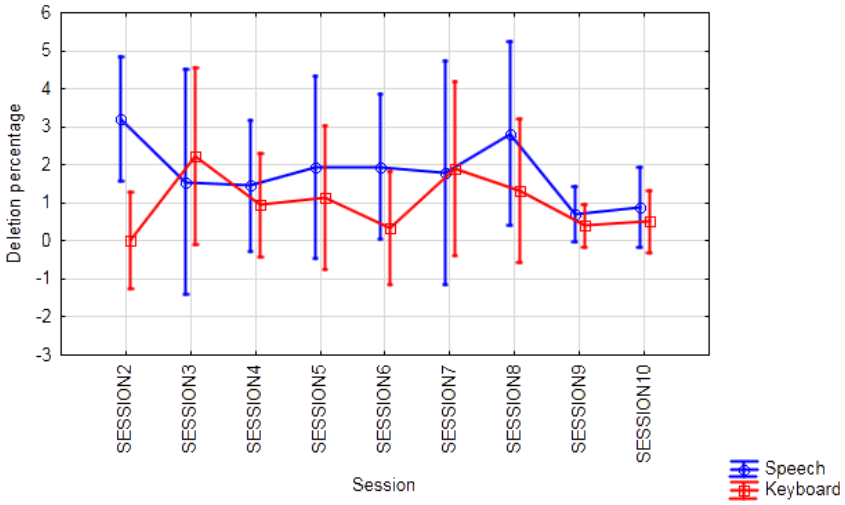


Chart 7.6: Mean deletion errors percentage of keyboard and speech-L

Table 7.8: Analysis results for deletion error percentage of keyboard and speech-L

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(1, 19) = 1.760, p > 0.05			
Session	F(8, 152) = 0.809, p > 0.05	F(4.2, 79.1) = 0.809, p > 0.05	F(5.8, 109.3) = 0.809, p > 0.05	F(8, 12) = 1.437, p > 0.05
Interaction technique × Session	F(8, 152) = 0.937, p > 0.05	F(4.2, 79.1) = 0.937, p > 0.05	F(5.8, 109.3) = 0.937, p > 0.05	F(8, 12) = 1.951, p > 0.05

Neither of the null hypotheses could be rejected at an α-level of 0.05, which indicates that the interaction technique has no noticeable impact on the number of deletions which occur when transcribing text. Therefore, the deletions percentages are not affected by the interaction technique.

7.5.1.3 Characters per second

Characters per second (CPS) for all sessions, each participant and each interaction technique were calculated. The number of observations, the mean and standard deviation of these observations are tabulated below.

Table 7.9: Descriptive statistics for characters per second of keyboard and speech-L

		All participants		Participants completing all sessions	
		Speech	Keyboard	Speech	Keyboard
Session 2	n	18	21	10	13
	$\bar{x}$	0.2	2.2	0.2	2.5
	s	0.1	1.1	0.1	1.2
Session 3	n	20	20	10	13
	$\bar{x}$	0.2	2.4	0.1	2.5
	s	0.1	1.0	0.1	1.2
Session 4	n	21	21	10	13
	$\bar{x}$	0.2	2.5	0.2	2.7
	s	0.1	1.0	0.1	1.2
Session 5	n	21	21	10	13
	$\bar{x}$	0.2	2.6	0.2	2.8
	s	0.1	1.2	0.1	1.4
Session 6	n	20	21	10	13
	$\bar{x}$	0.2	2.4	0.2	2.7
	s	0.1	1.0	0.1	1.1
Session 7	n	20	20	10	13
	$\bar{x}$	0.2	2.6	0.2	2.9
	s	0.1	1.1	0.1	1.2
Session 8	n	18	18	10	13
	$\bar{x}$	0.3	2.6	0.3	2.7
	s	0.1	1.0	0.1	1.0
Session 9	n	19	19	10	13
	$\bar{x}$	0.3	2.6	0.3	2.8
	s	0.1	0.9	0.1	1.0
Session 10	n	21	21	10	13
	$\bar{x}$	0.2	2.7	2.9	2.9
	s	0.1	0.9	0.1	1.0

The chart below is a plot of the mean characters per session for each interaction technique over all sessions.

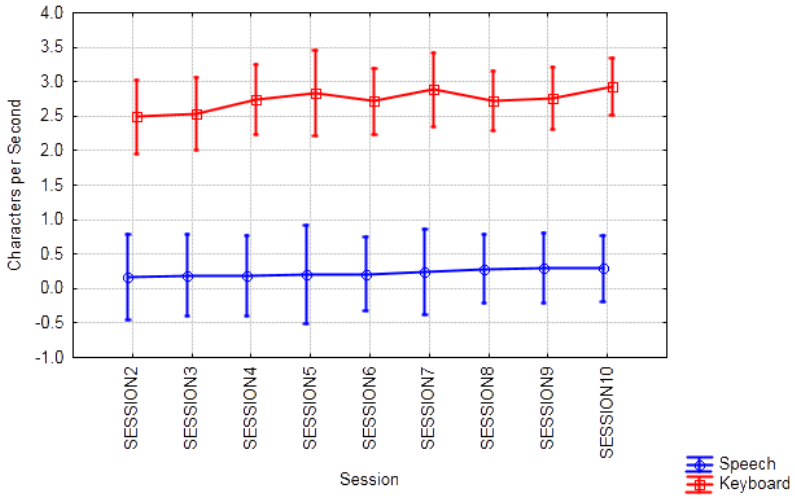


Chart 7.7: Mean characters per second of keyboard and speech-L

From the graph and the table it can be seen that when typing with the keyboard, participants were able to type at a faster rate than when using eye gaze and speech. The speed with which typing can be achieved using eye gaze and speech remains fairly constant throughout the sessions, displaying only mild improvement as the exposure increases. The following hypotheses were formulated:

1. $H_{0,1}$: The number of characters per second that can be typed is not influenced by the interaction technique.
2. $H_{0,2}$: The number of characters per second that can be typed is not influenced by the session in which the task was completed.

The assumption of sphericity was not met ($\chi^2(35) = 136.334$, $p < 0.05$), therefore Table 7.10 contains both the results of the repeated-measures ANOVA as well as the adjusted corrections that were required. To complete the analysis, the results of the multivariate tests are also reported.

Table 7.10: Analysis results for characters per second of keyboard and speech-L

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(1, 21) = 54.704, p < 0.05			
Session	F(8, 168) = 1.385, p > 0.05	F(3.6, 75.5) = 1.385, p > 0.05	F(4.6, 97.3) = 1.385, p > 0.05	F(8, 14) = 5.866, p > 0.05
Interaction technique × Session	F(8, 168) = 0.660, p > 0.05	F(3.6, 75.5) = 0.660, p > 0.05	F(4.6, 97.3) = 0.660, p > 0.05	F(8, 14) = 3.105, p > 0.05

The results in the table above show that $H_{0,1}$ could be rejected at an α-level of 0.05. Therefore, typing speed, as measured by characters typed per second, is significantly slower than when using a keyboard. The fact that $H_{0,2}$ could not be rejected shows that, although there is slight improvement between sessions, the improvement is not significant.

7.5.2 Analysis of all typing tasks

The previous section analysed the typing tasks which were completed using the keyboard and the original sized onscreen keyboard. As previously mentioned, there were four additional typing tasks added from the fifth session onwards. Two of these were performed with buttons which were smaller and closer spaced and the other two also with smaller buttons but spaced further apart. These two interaction technique will be referred to as Speech-SC and Speech-SW respectively. The original sized buttons will be referred to as the Speech-L interaction technique for the duration of the analysis. The same measurements as with the previous analysis will be analysed, namely error rate and characters per second. This will allow both the effectiveness and efficiency to be analysed.

7.5.2.1 Error Rate

The average error rate of each participant was calculated for each interaction technique and each session. The number of observations for each session, the mean and the standard deviation are tabulated below.

Table 7.11: Descriptive statistics for error rates of all interaction techniques

		All participants				Participants completing all sessions			
		Speech – L	Keyboard	Speech – SW	Speech – SC	Speech – L	Keyboard	Speech – SW	Speech – SC
Session 5	n	21	21	19	17	11	15	12	9
	$\bar{x}$	14.2	6.5	15.1	15.3	13.2	5.7	15.5	14.7
	s	9.0	6.7	6.3	5.7	9.5	6.4	6.9	7.0
Session 6	n	20	21	18	16	11	15	12	9
	$\bar{x}$	12.8	5.0	16.3	17.1	11.5	5.3	17.7	16.7
	s	7.2	5.0	9.6	10.0	7.1	5.6	10.9	11.3
Session 7	n	18	20	19	17	11	15	12	9
	$\bar{x}$	12.2	4.9	15.6	13.5	10.3	5.7	13.3	12.3
	s	9.2	7.1	8.4	6.7	8.3	8.1	8.4	8.9
Session 8	n	17	18	18	17	11	15	12	9
	$\bar{x}$	11.9	5.8	13.1	15.1	13.3	4.8	11.2	14.5
	s	6.5	5.2	8.6	9.9	5.8	4.9	8.6	11.9
Session 9	n	18	19	18	17	11	15	12	9
	$\bar{x}$	9.3	3.4	12.5	15.5	7.7	3.0	12.9	17.8
	s	6.5	4.7	8.4	10.6	6.7	4.1	8.2	11.9
Session 10	n	17	21	21	20	11	15	12	9
	$\bar{x}$	9.3	3.9	14.2	13.4	8.9	4.4	10.3	12.7
	s	5.9	3.8	11.9	8.4	6.5	4.3	8.5	9.3

Chart 7.8 is a plot of the means for the interaction techniques over all sessions.

From Table 7.11 it can be determined that the keyboard had the lowest error rate of all interaction techniques for all sessions. Thereafter, speech-L had the next lowest error rate while the smaller buttons, both speech-SW and speech-SC, caused the highest error rates for all sessions. The latter two seem to cause approximately the same error rates while typing, however, the widely spaced buttons have an improved error rate during the later sessions while the error rates for the closely spaced buttons increased over the same period.

The following hypotheses were formulated:

1. $H_{0,1}$: The error rate while typing is not affected by the interaction technique.
2. $H_{0,2}$: There is no difference between the sessions in the error rate while typing.

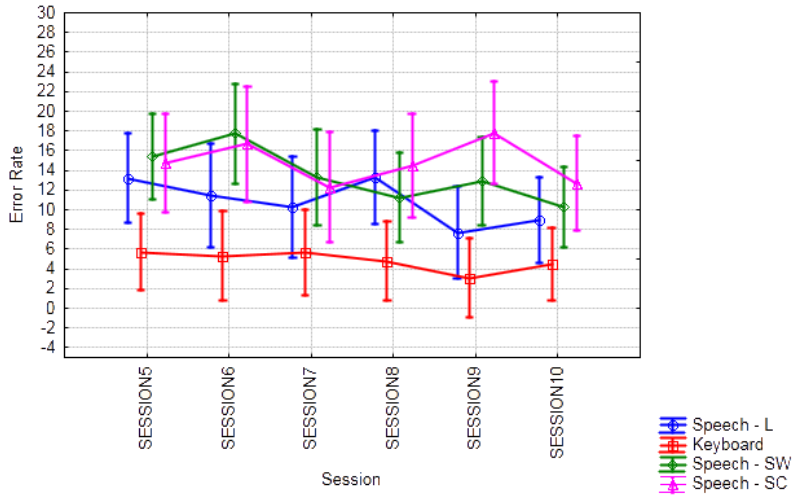


Chart 7.8: Mean error rate for all interaction techniques

The data met the condition of sphericity ($\chi^2(14) = 17.521$, $p > 0.05$) which meant that no adjusted corrections had to be applied to the degrees of freedom. The required results for the analysis are shown in Table 7.12.

Table 7.12: Analysis results of error rates for all interaction techniques

	ANOVA	Multivariate
Interaction technique	$F(3, 43) = 7.303$, $p < 0.05$	
Session	$F(5, 215) = 2.530$, $p < 0.05$	$F(5, 39) = 2.599$, $p < 0.05$
Interaction technique × Session	$F(15, 215) = 1.212$, $p > 0.05$	$F(15, 108) = 1.544$, $p > 0.05$

Consequently, both null hypotheses could be rejected using an α-level of 0.05. Therefore, the interaction technique has a noticeable impact on the error rate while typing and the error rate differs significantly between sessions.

Tukey's post-hoc test was used to determine which interaction techniques were responsible for the significant difference. It was found that the keyboard differed significantly from both speech-SW and speech-SC. Therefore, when using either speech-SW or speech-SC participants had a significantly higher typing error rate than when using a keyboard. In this instance, it is encouraging to determine that speech-L does not differ significantly from the keyboard in these later sessions. This would seem to indicate that after some practice with the larger buttons, the number of errors made decreases. The same cannot be said of the smaller buttons.

Session 6 differed significantly from session 10; in particular the error rates for speech-SW during session 6 were significantly higher than the error rates for the keyboard for all sessions. Furthermore, session 9 of Speech-SC differed significantly from all sessions with the keyboard.

The number of participants who were able to transcribe the text error-free was determined next. The results are shown in Chart 7.9.

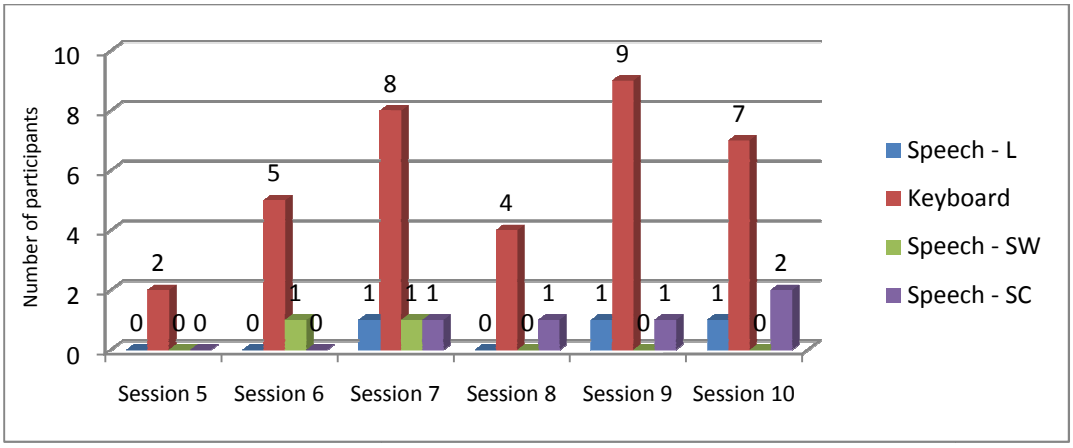


Chart 7.9: Error-free transcribed text for all interaction techniques

The keyboard clearly outperformed the other interaction techniques in this regard. For each session there were at least two participants who transcribed the text error-free when using the keyboard. In contrast, the speech interaction techniques had either zero or only one error-free transcribed text string in each session. It was only in the final session, that speech-SC had a higher number of error-free transcribed text strings, but even then it was only two participants who could manage that feat.

7.5.2.2 Breakdown of error rate

Each of the error rates could be further subdivided into the percentage of insertion errors, the percentage of substitution errors and the percentage of deletion errors. The graph below is a stacked bar graph for the first task (first four stacks) and the last task (last four stacks).

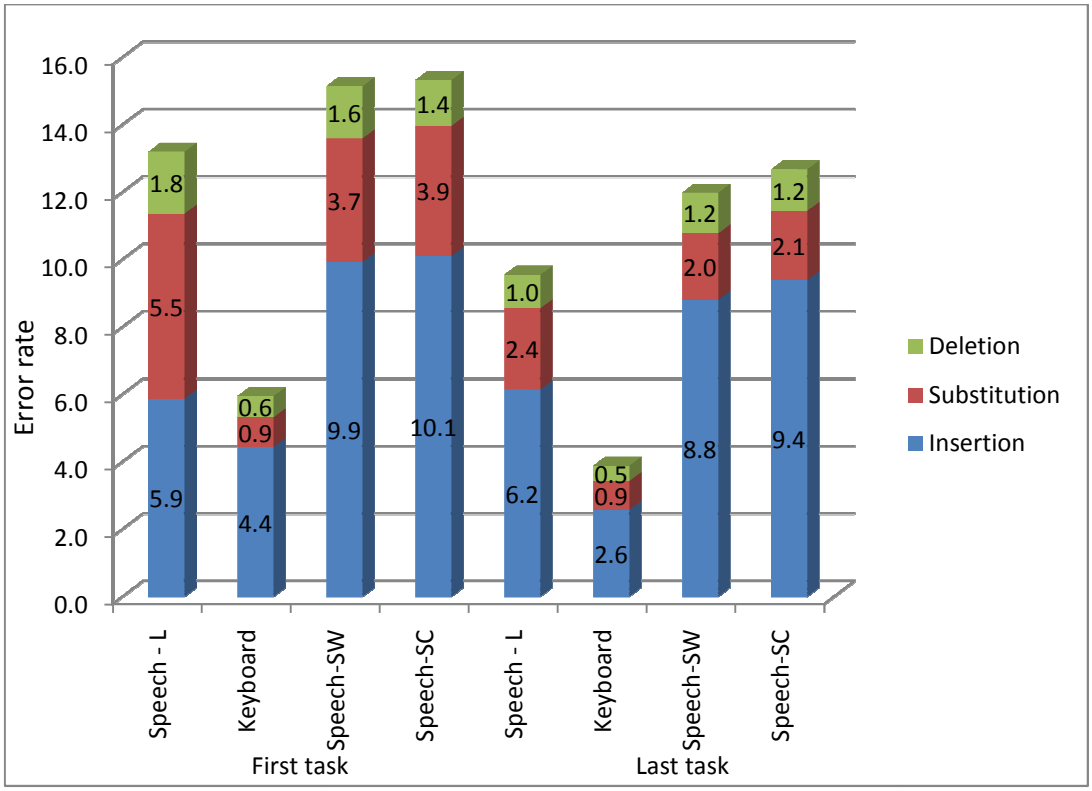


Chart 7.10: Breakdown of first task and last task's error rate for all interaction techniques

The percentage of insertion errors was the highest for all interaction techniques for both of these sessions. The interaction techniques of speech-SW and speech-SC have very similar distributions over the number of insertions, substitutions and deletions. In order to determine the significance of these distributions, each was analysed individually. The first of these was the percentage of insertion errors which will be discussed in the following section.

7.5.2.2.1 Percentage of insertion errors

Descriptive statistics for the insertion errors percentage of all the interaction techniques and for all sessions are summarised in Table 7.13.

Table 7.13: Descriptive statistics for insertion errors percentage of all interaction techniques

		All participants				Participants completing all sessions			
		Speech-L	Keyboard	Speech-SW	Speech-SC	Speech-L	Keyboard	Speech-SW	Speech-SC
Session 5	n	21	21	19	17	14	15	11	8
	$\bar{x}$	5.9	4.4	9.9	10.1	6.1	3.6	8.7	7.7
	s	4.6	4.3	5.9	5.6	5.1	3.9	6.1	4.8
Session 6	n	21	21	18	16	14	15	11	8
	$\bar{x}$	5.7	3.7	10.6	13.2	5.0	4.2	11.4	9.1
	s	4.3	5.0	9.2	10.0	4.9	5.7	11.4	9.5
Session 7	n	19	20	19	17	14	15	11	8
	$\bar{x}$	5.7	2.7	11.4	9.3	6.2	2.7	7.9	5.5
	s	4.7	3.5	7.7	6.9	5.2	4.0	5.7	5.8
Session 8	n	18	18	18	17	14	15	11	8
	$\bar{x}$	7.9	3.7	9.3	10.8	7.6	2.6	5.5	9.7
	s	6.6	5.2	6.6	8.1	6.4	4.6	3.4	11.0
Session 9	n	19	19	18	17	14	15	11	8
	$\bar{x}$	6.2	2.4	10.2	9.7	5.7	2.3	8.1	7.2
	s	4.3	3.4	8.0	9.1	4.4	3.6	4.5	6.2
Session 10	n	18	21	21	20	14	15	11	8
	$\bar{x}$	6.2	2.6	8.9	9.4	6.4	3.0	6.9	6.3
	s	5.2	3.1	6.3	7.9	5.9	3.5	7.0	5.0

Chart 7.11 is a plot of the mean percentage of insertion errors for all typing tasks.

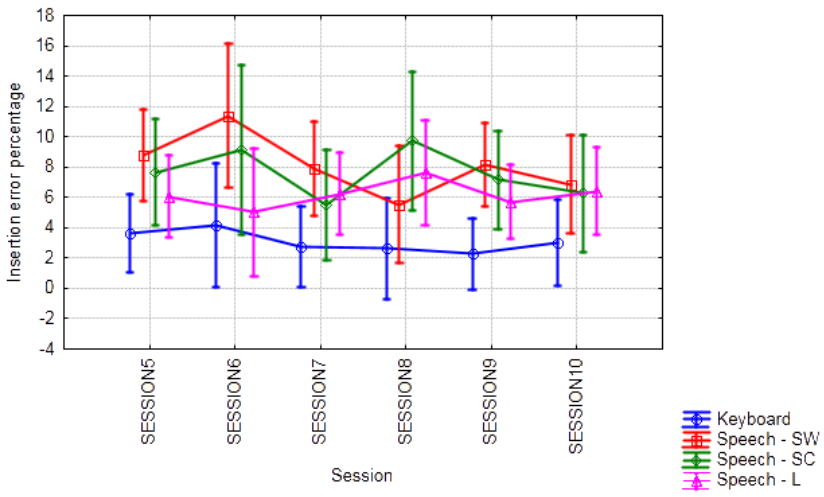


Chart 7.11: Mean insertion errors percentage for all interaction techniques

The keyboard had the lowest percentage of insertion errors, followed by speech-L. Once again, the interaction techniques of speech-SW and speech-LC were, for the most part, barely distinguishable from each other. The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the percentage of insertion errors made.
2. $H_{0,2}$: There is no difference between the percentage of insertion errors made between the sessions.

The assumption of sphericity was violated ($\chi^2(14) = 35.538$, $p < 0.05$), therefore Table 7.14 contains the results of the adjusted corrections as well as the other required analyses.

Table 7.14: Analysis results for insertion errors percentage of all interaction techniques

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(3, 44) = 4.100, p < 0.05			
Session	F(5, 220) = 1.056, p > 0.05	F(3.9, 169.9) = 1.056, p > 0.05	F(4.6, 200.9) = 1.056, p > 0.05	F(5, 40) = 5.866, p > 0.05
Interaction technique × Session	F(15, 220) = 1.002, p > 0.05	F(3.9, 169.9) = 1.002, p > 0.05	F(4.6, 200.9) = 1.002, p > 0.05	F(15, 111) = 0.811, p > 0.05

The first null hypothesis could be rejected as the p-value was less than the α-level used. Tukey's HSD post-hoc test indicated that the speech-SW interaction technique resulted in a significantly higher percentage of insertion errors than did the keyboard. There was no other significant difference between any other interaction techniques. The second null hypothesis could not be rejected; therefore there was no significant change in the percentage of insertion errors as the amount of exposure to the application increased.

7.5.2.2.2 Percentage of substitution errors

The number of included observations, the mean and the standard deviation of the percentage of substitution errors are tabulated below.

Chart 7.12 is a plot of the mean percentage of substitution errors for all interaction techniques and all sessions.

The percentage of substitution errors was lowest for the keyboard over all sessions. Of the interaction techniques which incorporated speech and eye gaze, speech-SW had the lowest mean percentage of substitution errors for the majority of the sessions. Apart from session 9, the percentage of substitution errors for speech-SC steadily improved as time went by. The following hypotheses were formulated to determine whether the differences noticed on Chart 7.14 and Table 7.15 were significant.

1. $H_{0,1}$: The interaction technique has no effect on the percentage of substitution errors made.
2. $H_{0,2}$: There is no difference between the percentages of substitution errors made between the sessions.

When performing a repeated-measures within-subjects ANOVA it was found that there was significant interaction ($F(15, 210) = 2.228$, $p < 0.05$) between the factors of session and interaction technique. Subsequently, separate ANOVAs had to be performed in order that the influencing factors could be controlled for.

Table 7.15: Descriptive statistics for substitution errors percentage of all interaction techniques

		All participants				Participants completing all sessions			
		Speech-L	Keyboard	Speech-SW	Speech-SC	Speech-L	Keyboard	Speech-SW	Speech-SC
Session 5	n	20	21	19	17	12	15	11	8
	$\bar{x}$	5.5	0.9	3.7	3.9	3.3	0.7	3.4	3.4
	s	6.1	1.8	4.9	4.3	3.4	1.6	4.8	4.3
Session 6	n	20	21	18	16	12	15	11	8
	$\bar{x}$	3.3	0.5	3.9	3.3	1.0	0.6	3.4	3.0
	s	4.9	1.0	3.5	5.7	2.0	1.0	2.6	5.3
Session 7	n	20	20	19	17	12	15	11	8
	$\bar{x}$	4.9	0.8	1.9	2.6	4.6	1.0	1.2	2.3
	s	5.2	1.8	2.8	3.3	5.0	2.0	1.5	2.3
Session 8	n	18	18	18	16	12	15	11	8
	$\bar{x}$	2.2	0.3	1.3	1.4	1.6	0.3	0.8	0.5
	s	2.5	0.8	2.0	2.3	1.6	0.8	1.1	0.9
Session 9	n	18	19	18	17	12	15	11	8
	$\bar{x}$	1.7	0.5	1.5	4.1	1.1	0.4	1.1	4.3
	s	2.5	1.3	2.9	5.7	1.7	1.2	1.7	5.7
Session 10	n	18	21	20	19	12	15	11	8
	$\bar{x}$	2.4	0.9	2.0	2.1	1.6	1.0	0.8	1.2
	s	3.4	1.7	3.1	3.0	3.5	1.9	1.3	1.9

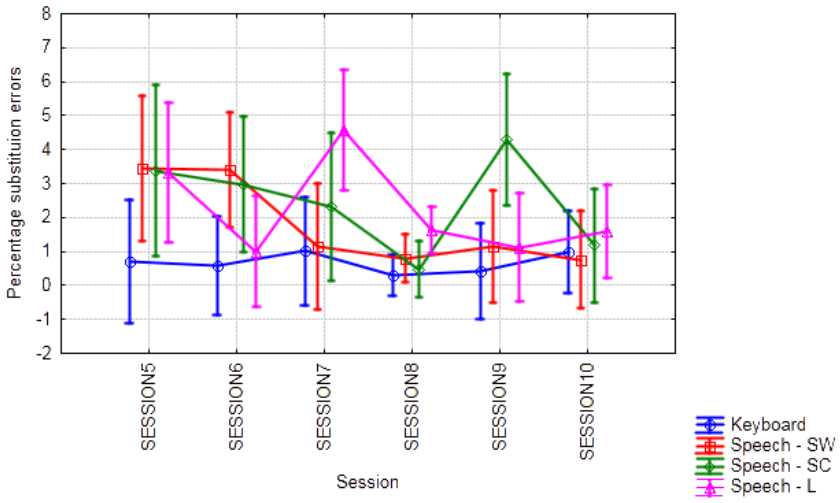


Chart 7.12: Mean substitution errors percentage of all interaction techniques

$H_{0,1}$ could be rejected for session 5 ($F(3, 73) = 3.690$, $p < 0.05$), where the keyboard had a significantly lower percentage of substitution errors than speech-L. Similarly, $H_{0,1}$ could be rejected for session 6 ($F(3, 71) = 2.862$, $p < 0.05$) as the keyboard had a significantly lower percentage of substitution errors than all other interaction techniques. During session 7 ($F(3, 72) = 5.040$, $p < 0.05$) when participants used the speech-L interaction technique they had a significantly higher percentage of substitution errors than when using the speech-SW interaction technique and the keyboard. The keyboard also resulted in lower percentages of substitution errors than the speech-SC interaction technique during session 9 ($F(3, 68) = 3.442$, $p < 0.05$). The null

hypothesis could not be rejected for either session 8 ($F(3, 66) = 2.671$, $p > 0.05$) or 10 ($F(3, 74) = 1.113$, $p > 0.05$) at an α-level of 0.05.

The second null hypothesis could be rejected at α-level of 0.05 for speech-L and speech-SW. The results of all the analyses to investigate $H_{0,2}$ are tabulated below.

Table 7.16: Analysis results of substitution errors percentage for all interaction techniques

	Mauchley's	ANOVA	Geisser-Greenhouse	Huyn-Feldt
Speech-L	$X^2(14) = 18.507$, $p > 0.05$	$F(5, 55) = 2.824$, $p < 0.05$		
Keyboard	$X^2(14) = 33.646$, $p < 0.05$	$F(5, 70) = 0.738$, $p > 0.05$	$F(2.7, 37.7) = 0.738$, $p > 0.05$	$F(3.4, 47.5) = 0.738$, $p > 0.05$
Speech-SC	$X^2(14) = 31.840$, $p < 0.05$	$F(5, 35) = 1.400$, $p > 0.05$	$F(2.7, 18.6) = 1.400$, $p > 0.05$	$F(4.4, 31.1) = 1.400$, $p > 0.05$
Speech-SW	$X^2(14) = 22.982$, $p > 0.05$	$F(5, 50) = 3.285$, $p < 0.05$		

From the table it can be concluded that $H_{0,2}$ could be rejected for speech-L and speech-SW. For the large buttons, the percentage of substitution errors was, on average, lower for session 6 than for session 7. When using eye gaze and speech with the smaller buttons which were widely spaced, session 5 and session 6 differed significantly from sessions 7 to 10. The latter sessions had a lower average than the first two sessions which indicated that when using this interaction technique there was some measure of learning as the exposure to the application increased.

7.5.2.2.3 Deletion errors percentage

Table 7.17 contains a summary of the number of observations, mean and the standard deviation of the deletion errors percentage.

Table 7.17: Descriptive statistics of deletion errors percentage for all interaction techniques

		All participants				Participants completing all sessions			
		Speech-L	Keyboard	Speech-SW	Speech-SC	Speech-L	Keyboard	Speech-SW	Speech-SC
Session 5	n	21	20	19	17	9	14	11	9
	$\bar{x}$	1.8	0.6	1.6	1.4	2.4	0.6	2.1	1.9
	s	2.7	1.6	2.4	2.1	3.2	1.5	2.5	2.1
Session 6	n	17	21	17	16	9	14	11	9
	$\bar{x}$	0.7	0.8	1.3	0.6	0.2	0.5	1.0	0.5
	s	1.3	1.7	2.1	2.0	0.6	1.1	1.9	0.9
Session 7	n	17	19	18	16	9	14	11	9
	$\bar{x}$	0.7	0.8	1.5	1.1	0.5	1.1	2.2	0.9
	s	1.4	1.7	2.3	2.2	1.0	1.9	2.5	1.4
Session 8	n	18	18	18	17	9	14	11	9
	$\bar{x}$	2.8	1.8	2.5	2.2	2.3	1.8	3.5	1.4
	s	4.7	2.5	4.6	2.9	4.4	2.6	5.6	2.8
Session 9	n	18	19	18	17	9	14	11	9
	$\bar{x}$	1.0	0.5	0.9	1.7	0.2	0.4	0.8	1.9
	s	1.5	1.1	1.6	2.6	0.5	1.0	1.5	3.1
Session 10	n	17	21	20	20	9	14	11	9
	$\bar{x}$	1.0	0.5	1.2	1.2	0.8	0.5	0.8	1.2
	s	2.9	0.8	2.4	2.2	2.0	0.7	2.5	2.0

The chart below is a plot of the mean for all interaction techniques across all sessions.

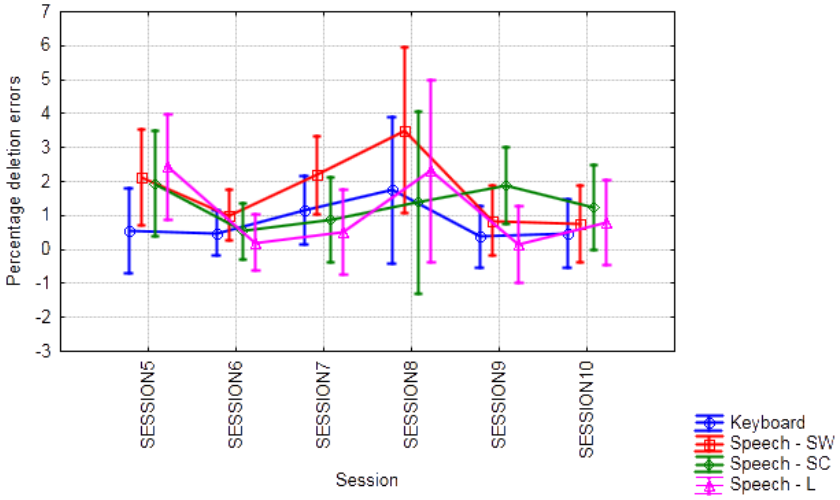


Chart 7.13: Mean deletion errors percentage for all interaction techniques

Inspection of Table 7.17 and Chart 7.13 shows that the number of deletions is approximately the same for all interaction techniques and across all sessions. Apart from session 8, which shows a sharp spike in the number of deletions, the deletion errors percentage remains fairly stable throughout. The following hypotheses were formulated:

1. $H_{0,1}$: The interaction technique has no effect on the percentage of deletion errors made.
2. $H_{0,2}$: There is no difference between the percentages of deletion errors made between the sessions.

The assumption of sphericity which is required for a repeated-measures, within-subjects ANOVA was not met ($\chi^2(14) = 67.342$, $p < 0.05$). This required additional analyses to be performed on the data. The results of all the tests are tabulated below.

Table 7.18: Analysis results of deletion errors percentage for all sessions

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	F(3, 39) = 1.638, p > 0.05			
Session	F(5, 195) = 3.450, p < 0.05	F(2.6, 103.2) = 3.450, p < 0.05	F(3.1, 120.0) = 3.450, p < 0.05	F(5, 15) = 4.221, p < 0.05
Interaction technique × Session	F(15, 195) = 0.766, p > 0.05	F(7.9, 103.2) = 0.766, p > 0.05	F(9.2, 120.0) = 0.766, p > 0.05	F(15, 97) = 0.491, p > 0.05

With a confidence interval of 95%, $H_{0,1}$ could not be rejected but $H_{0,2}$ could be rejected. As could be expected, Tukey's HSD post-hoc test indicated that session 8 differed significantly from sessions 6, 9 and 10. Session 8 had a much higher average deletion percentage than the other sessions.

7.5.2.3 Characters per second

The typing speed, measured as characters per second, was calculated for each participant and each typing task. The average per session was then calculated for each interaction technique. The descriptive statistics for this measurement are summarised in the table below.

Table 7.19: Descriptive statistics of characters per second for all interaction techniques

		All participants				Participants completing all sessions			
		Speech-L	Keyboard	Speech-SW	Speech-SC	Speech-L	Keyboard	Speech-SW	Speech-SC
Session 5	n	21	19	19	17	14	13	12	9
	x̄	0.2	2.2	0.2	0.2	0.2	2.4	0.2	0.2
	s	0.1	0.7	0.1	0.1	0	0.8	0	0.1
Session 6	n	20	21	18	16	14	13	12	9
	x̄	0.2	2.4	0.2	0.3	0.2	2.3	0.3	0.3
	s	0.1	1.0	0.1	0.1	0.1	0.7	0.1	0.1
Session 7	n	20	19	19	17	14	13	12	9
	x̄	0.2	2.5	0.2	0.3	0.2	2.5	0.3	0.3
	s	0.1	0.7	0.1	0.1	0.1	0.6	0.1	0
Session 8	n	18	18	18	17	14	13	12	9
	x̄	0.3	2.6	0.3	0.3	0.3	2.6	0.3	0.3
	s	0.1	1.0	0.1	0.1	0.1	0.8	0.1	0.1
Session 9	n	19	19	18	17	14	13	12	9
	x̄	0.3	2.6	0.3	0.3	0.3	2.5	0.3	0.3
	s	0.1	0.9	0.1	0.0	0.1	0.7	0.1	0
Session 10	n	21	21	21	20	14	13	12	9
	x̄	0.2	2.7	0.3	0.3	0.3	2.7	0.3	0.3
	s	0.1	0.9	0.1	0.1	0.1	0.7	0.1	0.1

Chart 7.14 shows a plot of the means for the characters per second measurement of all interaction techniques across all sessions.

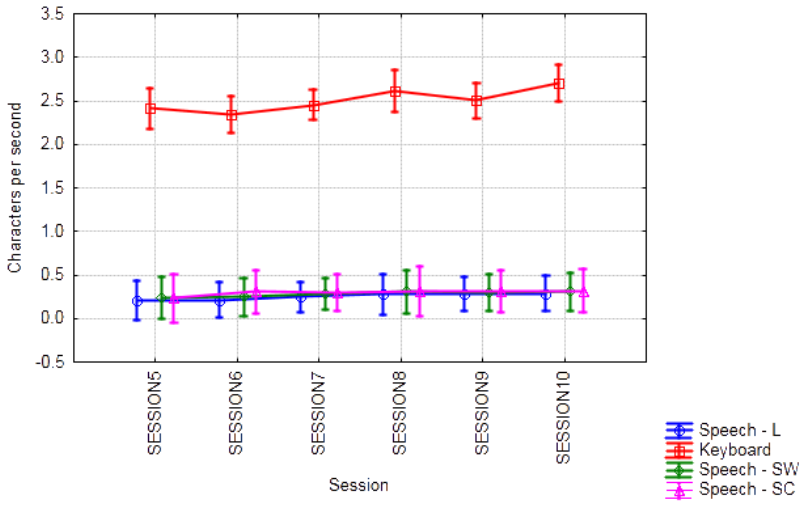


Chart 7.14: Mean characters per second for all interaction techniques

When using the keyboard, participants were clearly able to type at a much faster rate than when using eye gaze and speech with the onscreen keyboard. With reference to Table 7.19 and Chart 7.15, it appears that the

size and spacing of the buttons did not affect the speed at which typing could occur. The underlying hypotheses were formulated to analyse this statistically:

1. $H_{0,1}$: There is no difference between the number of characters per second that can be typed using the different interaction techniques.
2. $H_{0,2}$: The session has no effect on the number of characters per second that can be typed.

The assumption of sphericity was violated ($\chi^2(14) = 76.146$, $p < 0.05$), therefore Table 7.20 below shows the results of the ANOVA, the multivariate tests and the corrected adjustments required.

Table 7.20: Analysis results of characters per second for all interaction techniques

	ANOVA	Geisser-Greenhouse	Huyn-Feldt	Multivariate
Interaction technique	$F(3, 44) = 148.369$, $p < 0.05$			
Session	$F(5, 15) = 3.002$, $p < 0.05$	$F(3.3, 147.4) = 3.002$, $p < 0.05$	$F(3.9, 171.8) = 3.002$, $p < 0.05$	$F(5, 40) = 2.563$, $p < 0.05$
Interaction technique × Session	$F(15, 220) = 0.845$, $p > 0.05$	$F(10.0, 147.4) = 0.845$, $p > 0.05$	$F(11.7, 171.8) = 0.845$, $p > 0.05$	$F(5, 110) = 1.264$, $p > 0.05$

Both null hypotheses could be rejected at an α-level of 0.05. Therefore, the number of characters typed per second differs between the interaction techniques and sessions. Tukey's HSD shows that the keyboard yields a significantly faster typing rate than all other interaction techniques. Session 10 also yielded significantly faster typing speeds than sessions 5 and 6.

7.5.3 Summary of results

It was found that the eye gaze and speech interaction technique had a significantly higher error rate than that of the keyboard, undoubtedly as a result of a higher number of insertions and substitutions. This may serve as confirmation that even when using eye gaze and speech as a text input mechanism, the user is inclined to glance away before completing the issuing of the verbal command. This finding corresponds to that of the increased number of missed clicks during ISO testing of the interaction techniques. The average insertions are generally higher than the substitutions which would seem to indicate that users are aware that they have activated the incorrect character and attempt to correct it by inserting the correct character. This is encouraging as it indicates that users become familiarised with the system such that they can interpret the selection (indicated by audio feedback) and are able to make corrections to text entry. Further research could confirm these suppositions by capturing the correction of the text input as well so that it can be analysed to determine whether incorrect inputs are reversed/erased before text input is continued. Whether the buttons are large, small and widely spaced or small and closely spaced seems to be of little consequence. There was no difference between the error rates of these three interaction techniques 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 the interaction technique. If the learning effect can be maintained then more practice with the eye gaze and speech could eventually lead to an effectiveness measurement which is comparable to that of the keyboard.

In terms of efficiency, the keyboard also outperformed all the eye gaze and speech interaction techniques with significantly higher numbers of characters per second which could be typed. The typing speed of the eye gaze and speech also did not improve as exposure increased. This could indicate that either more practice is needed

to achieve increased speeds 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.

Therefore, in terms of effectiveness and efficiency, the three eye gaze and speech interaction techniques seem fairly interchangeable as they perform on comparable levels to each other. The keyboard is far more effective and efficient than any of the eye gaze and speech interaction techniques when used for text input.

There are no similar studies with which to compare the results found during this study. However, the fact that speech outperforms keyboard input for young children (Read et al., 2001) indicates that the learning curve for keyboard entry is fairly steep. This could be the same for text entry with eye gaze and speech. Although there was no significant improvement in the speed of the text entry, participants clearly became more comfortable with the use of the interaction technique. Therefore, extended practice may be required to improve speeds.

The mean entry rate of eye gaze and speech fell within the range between 0.2 and 0.3 characters per second. Considering that the entry rate was relatively low for context switching at 12 WPM (Morimoto & Amir, 2010) and 9 WPM for symbol creator (Miniotas et al., 2003), the range in this study was comparable to these previous studies. A previous study showed that the use of both visual and auditory feedback increased the entry speed to 7.55 WPM which is lower than the speeds achieved in this study. Speech Dasher achieved much higher speeds (40 WPM), while using Dasher with eye gaze also resulted in higher speeds (17 WPM). Therefore, when comparing the text entry method to studies using only eye gaze without text predictors, speech and eye gaze performs slightly better. However, the speeds are still lower than using text prediction methods and when using speech as an activator. While these comparisons are promising since they indicate that speech and eye gaze could facilitate faster entry speeds than using eye gaze only, they are discussed with caution since the text entered in the current study required only a few short phrases to be entered and more prolonged use could have an impact on the entry speed.

7.6 Further research

Further research can be conducted in terms of which 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 gravitational well will impact performance. Although the decrease of physical size and increase of gravitational 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.

7.7 Summary

This chapter reported on the results of the use of eye gaze and speech for text input when compared to a traditional keyboard. Measurements of efficiency, namely characters typed per second, and effectiveness, namely the character error rate, 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 of these usability measures far outperforms 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.

The next chapter will discuss the results of the questionnaires designed to elicit the subjective usability measurement of satisfaction. Furthermore it will also discuss the subjective feelings of participants towards using the combination of eye gaze and speech to simulate a pointing device.

CHAPTER 8

PARTICIPANT SUBJECTIVE SATISFACTION

8.1 Introduction

The previous two chapters analysed a number of objective usability measurements. These included efficiency, effectiveness and learnability measurements of using speech commands as well as typing using the interaction technique of eye gaze and speech. Speech commands were found, in some instances, to be comparable or even more efficient or effective than the traditional means of using the keyboard and mouse to complete the same task. Where typing was concerned it was found that the keyboard was far more efficient and effective than eye gaze and speech and that even with continued use, eye gaze and speech could not compete with the keyboard.

Even with this objective evidence it is important also to measure the subjective response of participants with regard to the use of the application. Subjective feelings were captured through the use of an extensive questionnaire (Appendix H) which was administered during the first and last sessions. Informal interviews were also conducted with the participants after each session. The facilitator also unobtrusively observed the participants during each session and made notes on any behaviour that was deemed noteworthy. This chapter reports on the results of these questionnaires, observations and interviews.

8.2 Procedure

Questionnaires were administered to participants of the study in order to gauge their subjective satisfaction with the application. Therefore, the sample was the same as for the previous two chapters. There was, however, a single participant who did not attend her last session and therefore did not complete the questionnaire in the final session. Therefore, the analysis in this chapter is based on the responses of the remaining 24 participants.

At the end of the very first session which contained the introduction to the application and the informal interaction with the application, the participants were required to complete the questionnaire as outlined in Appendix G. After their exposure to the application over the course of the ten week longitudinal testing, the extended questionnaire as contained in Appendix H was completed by the participants. These questionnaires were designed to elicit overall satisfaction with the application, the subjective feelings towards using eye gaze and speech to effectively point at objects and in the case of the second questionnaire, satisfaction with the typing aspect of the application as well as the speech commands which could be used. The questionnaires were compiled from the pointing device assessment questionnaire as advocated by ISO 9241-9 as well the Questionnaire for User Interaction Satisfaction (Chin, Diehl & Norman, 1988).

The ISO questionnaire consisted of 9 questions, each of which could be ranked on a 5-point scale. For analysis purposes, the 5-point scale was divided into three categories. The first category consisted of the two lowest rankings, the second category consisted of the neutral category only and the third category consisted of the two highest rankings. Although there may be a danger that respondents simply choose the neutral category as the least controversial choice, it was decided to still view this as a separate category and not to group it with the negative responses as it could not be guaranteed that a negative response was indeed the intention of the respondent.

In order to measure user satisfaction, an extract of the Questionnaire for User Interaction Satisfaction (QUIS) was used. These questions were then posed based on the system as a whole, the response to the speech commands and the response to the typing. A 5-point qualified Likert scale with explicit adjectives on either side of the scale was used as a response scale for all questions. The five-point response scale was numbered 1 – 5 for analysis purposes and responses were grouped into a negative category which consisted of the two lowest points on the scale, neutral which was the midpoint and a positive response which was any response to the two highest points of the scale. As advocated by Harper and Norman (1993), each subsection could be given an overall score for each participant by calculating the average score for that subsection.

8.3 Reaction to the application

8.3.1 Satisfaction

Both the first and second administered questionnaires contained sections to gauge the overall reaction of the participants to the application (see Appendix G and H, Part 3). The responses to each of the questions were categorised into low usability, neutral usability and high usability, as specified in the previous section. The number of responses in each category was determined after which a contingency table was created for each question. For example, the contingency table for the first question looked as follows (Yates correction was applied during analysis where necessary):

Table 8.1: Example contingency table for overall satisfaction

	Terrible Responses = 1 or 2	Neutral Responses = 3	Wonderful Responses = 4 or 5
First session	1	8	16
Last session	0	7	17

Thereafter a Chi-square test was used to evaluate the underlying hypothesis for each question.

> H_0: There is no difference between satisfaction after the first exposure to the application and after extended use of the application.

Table 8.2 contains descriptive statistics for each question as well as the result of the Chi-square test for that question. Scores are on a scale of 1 to 5, with a midpoint of 3. The first column contains the scale adjectives that were used for that particular question.

The null hypothesis could not be rejected for any of the questions at an α-level of 0.05 which means that user reaction to the application was not significantly altered over the course of the ten weeks. The null hypothesis for the second question (ranging from frustrating to satisfying) could however be rejected at an α-level of 0.10. Since the overall mean decreased from the first to the last session it would indicate that the level of frustration was slightly higher at the end of the ten week period. In fact, the mean for all the questions was lower after the last session than after the first session but not significantly so. For example, after ten weeks the participants found the application less stimulating than after the first exposure. This could be due to the fact that they had learnt to use the system and found it less stimulating once they had mastered the use of the multimodal interface. However, the adequacy and rigidity measurements were also slightly lower which could indicate that once they had explored the available options they felt they needed more freedom than was offered by the available grammar and onscreen typing. The fact that the application is highly customisable and

extendable could offer a solution to this and perhaps should have been pointed out to the participants even if they did not get the opportunity to make use of these features.

Table 8.2: Descriptive statistics for each satisfaction question for the application

	First session			Last session			Chi-square test
	Mode	Mean	Standard Deviation	Mode	Mean	Standard Deviation	
Terrible – Wonderful	4	3.8	0.9	4	3.8	0.6	$\chi^2(2) = 0.02$ *, $p > 0.05$
Frustrating – Satisfying	4	3.6	1.0	3	3.0	1.0	$\chi^2(2) = 4.9$, $0.05 < p < 0.10$
Dull – Stimulating	4	4.1	0.8	4	3.9	1.0	$\chi^2(2) = 2.4$, $p > 0.05$
Difficult – Easy	3	3.4	1.2	3	3.2	0.9	$\chi^2(2) = 0.4$, $p > 0.05$
Inadequate – Adequate	3	3.6	0.9	3	3.4	0.8	$\chi^2(2) = 0.6$, $p > 0.05$
Rigid – Flexible	4	3.6	0.9	4	3.4	0.8	$\chi^2(2) = 2.3$, $p > 0.05$

* Yates corrected Chi-square applied

The fact that the mean for each of the questions rated on the high side for both sessions is encouraging in terms of the subjective satisfaction experienced by the participants.

Each subsection of the QUIS can be given a score based on the responses of the participants (Shneiderman, 1998). Following the example of other studies (cf., Tullis & Stetson, 2004), a score for this subsection was given to each participant. This was calculated as the mean of the responses for the six questions for each participant and for each administration of the questionnaire. Table 8.3 below summarises the descriptive statistics for both sessions, which are scored on a scale of 1 – 5:

Table 8.3: Descriptive statistics for overall satisfaction with application

	First session	Last session
Mean	4.7	4.5
Standard Deviation	0.7	0.7

The average satisfaction rating was lower after the last session than after the first session, although there is a satisfactory high usability measurement achieved for both sessions. The following hypothesis was formulated to determine if the detected difference was significantly different:

> H_0: There is no difference between the overall satisfaction of the participants after the first exposure to the application and after extended use of the application.

A paired t-test was used to determine if the opinions of the participants had changed after they could interact with the application over an extended period. The null hypothesis of no difference could not be rejected ($t = 1.74$, $df = 23$), therefore there was no difference in the overall satisfaction of the participants between the first and the last session.

Although there was no difference between the sessions, it is encouraging to note that the overall reaction to the system was either neutral or positive. No question was rated on the negative side which indicates that, in general, the participants viewed their experience with the application as pleasing and satisfying.

8.3.2 Learnability

The learnability subsection of the questionnaire had four questions (Appendix H, Part 6). Responses after the first session were based on the perceived learnability after a short interaction with the system, while the responses after the last session were based on experience with the system and how the participants experienced their own learning curve. The same categorisation as in the prior section was used to categorise the responses for learnability into low, neutral and high learnability (see Table 8.4 for an example).

Table 8.4: Example contingency table for overall learnability

	Difficult	Neutral	Easy
First session	8	4	13
Last session	6	5	13

The following hypothesis was formulated for this purpose:

> H_0: There is no difference between subjective feelings of learnability after the first exposure to the application and after extended use of the application.

The afore-mentioned hypothesis was analysed for each question in the learnability subsection and then also for the overall learnability of the application. Descriptive statistics and results of the Chi-square test for each of the questions are tabulated below. The first column contains the scale that was used as well in the particular facet of learnability that was targeted by the question.

Table 8.5: Descriptive statistics for learnability questions for the application

	First session			Last session			Chi-square test
	Mode	Mean	Standard deviation	Mode	Mean	Standard deviation	
Difficult – Easy (Overall learning)	4	3.3	1.2	4	3.5	1.3	$\chi^2(2) = 0.4$, $p > 0.05$
Difficult – Easy (Getting started)	3	3.3	1.2	4	3.1	1.4	$\chi^2(2) = 8.7$, $p < 0.05$
Difficult – Easy (Learning advanced features)	2	3.0	1.1	4	3.3	1.1	$\chi^2(2) = 1.9$, $p > 0.05$
Slow – Fast (Time to learn)	3	3.3	1.2	4	3.5	1.4	$\chi^2(2) = 3.0$, $p > 0.05$

The null hypothesis could only be rejected for the question aimed at eliciting an opinion about the ease with which a user can make use of the application from the very start. After the first session the majority of the participants felt neutral about this aspect of the application but after extended use the majority felt that it was relatively easy. However, closer inspection of the spread of the responses shows that while some respondents moved from neutral to easy, there were also 5 more that felt the initial learning curve was steeper in retrospect than after their first exposure.

Each participant was given an overall score, calculated as an average of their four responses, as a measure of the learnability of the application after both the first and the last sessions. The table below is a summary of the descriptive statistics for the learnability of the application. Scores are ranked on a scale of 1 – 5.

Table 8.6: Descriptive statistics for overall learnability of the application

	First session	Last session
Mean	4.1	4.3
Standard Deviation	1.0	1.1

A paired t-test was conducted to determine if there were any differences between the two sessions. At an α-level of 0.05, the afore-mentioned null hypothesis could not be rejected (t = 1.235, df = 23). Learnability of the application, both in the short- and long-term, was rated positively by the majority of the participants, particularly after the last session.

8.4 Typing

8.4.1 Satisfaction

The questionnaire administered after the final session contained questions aimed specifically at testing user satisfaction with regard to typing using the onscreen keyboard and eye gaze and speech. The same six Likert (Olivier, 2004) scales with explicit adjectives were provided as with the questions for the whole application. An extra question was added to measure the naturalness of typing using eye gaze and speech with an onscreen keyboard.

In order to determine whether the observed number of participants for each one of the categories is significantly different from an even distribution, a contingency table as in Table 8.7 was set up for each one of the six questions.

Table 8.7: Example contingency table for Chi-square test

	Terrible	Neutral	Wonderful
Observed responses	4	7	13
Expected even distribution	8	8	8

The following null hypothesis was formulated:

H_0: The responses in the possible categories are evenly distributed.

Descriptive statistics for the responses, together with the results of the respective Chi-square tests, are summarised in Table 8.8 (scores are reported on a scale of 1 – 5).

Chart 8.1 below shows a stacked bar chart for the number of responses in each category. The numbers in the stacks indicate the number of responses in that category.

The null hypothesis could not be rejected for any of the questions at an α-level of 0.05, but that of the dull – stimulating measurement could be rejected at an α-level of 0.1. This means that significantly more respondents experienced the application as being stimulating than those who experienced it as dull. The mean score for the majority of these questions is above the midpoint of the scale which suggests that the experience of typing was a fairly pleasant one for most of the participants. The subjective feelings of difficult and frustrating had a number of responses on the low side of the scale. This is understandable as it takes some practice for the participants to become used to typing using eye gaze and speech and even after some practice it could still be considered challenging to maintain a stable gaze on a button long enough to issue the

command required to type the letter to the document. It could also be frustrating for the users in the sense that sometimes they would still glance away whilst issuing the command, thereby preventing the desired letter to be typed to the document. This could also account for the fair number of responses in the neutral category of the rigid – flexible scale. Unfortunately, the feeling of naturalness experienced during interaction is below the midpoint which indicates that the majority of participants found this method of interaction unnatural to some degree. The novelty of the interaction technique could play a significant role in this mindset and it may be necessary to increase the level of exposure and practice with the application before it could possibly match the naturalness of traditional keyboard typing.

Table 8.8: Descriptive statistics for satisfaction questions for the typing feature

	Mode	Mean	Standard deviation	Chi-square test
Terrible – Wonderful	4	3.5	1.0	$\chi^2(2) = 2.6$, $p > 0.05$
Frustrating – Satisfying	3	3.1	1.1	$\chi^2(2) = 0.1$, $p > 0.05$
Dull – Stimulating	3	3.7	1.1	$\chi^2(2) = 4.8$, $0.05 < p < 0.10$
Difficult – Easy	3	3.2	1.2	$\chi^2(2) = 0.2$, $p > 0.05$
Inadequate – Adequate	3	3.6	0.9	$\chi^2(2) = 3.1$, $p > 0.05$
Rigid – Flexible	3	3.3	1.2	$\chi^2(2) = 0.8$, $p > 0.05$
Unnatural – Natural	2	2.8	1.3	$\chi^2(2) = 0.8$, $p > 0.05$

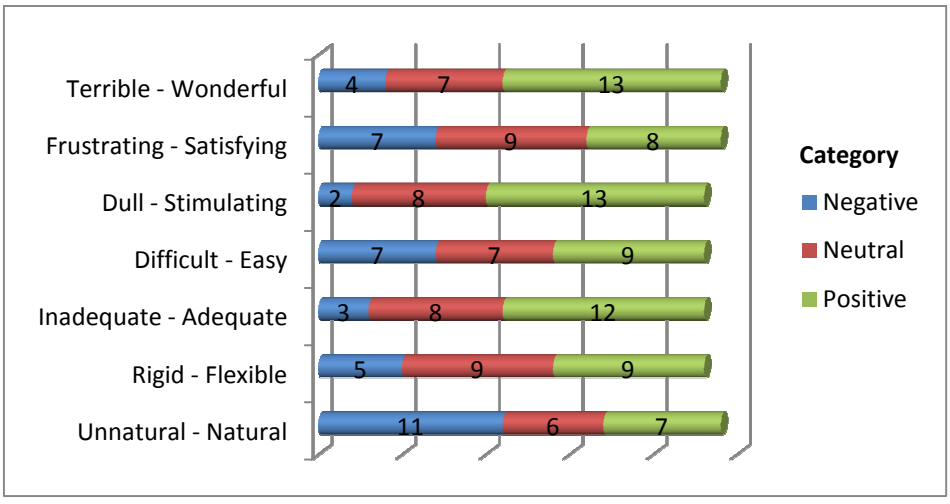


Chart 8.1: Number of responses in each category of the typing feature satisfaction questions

8.4.2 Learnability

The learnability of typing using eye gaze and speech was measured using the same four questions as for the overall usability of the application. Once again, the same grouping per category was used for the responses. Since there was only one measurement available, the average number of participants was used as a secondary measurement (see Table 8.7 in previous section for an example) in order to analyse the underlying hypothesis.

H_0: The responses in the possible categories are evenly distributed.

Descriptive statistics and the results of the Chi-square test for the learnability of typing are summarised in Table 8.9. Scores are measured on a scale of 1 – 5 and the first column contains the adjectives used on the scale for the particular question.

Table 8.9: Descriptive statistics for learnability questions for the typing feature

	Mode	Mean	Standard deviation	Chi-square test
Difficult – Easy (Overall learning)	3	3.2	1.4	$\chi^2(2) = 0.8$, $p > 0.05$
Difficult – Easy (Getting started)	3	3.0	1.3	$\chi^2(2) = 0.4$, $p > 0.05$
Difficult – Easy (Learning advanced features)	3	3.3	1.2	$\chi^2(2) = 1.0$, $p > 0.05$
Slow – Fast (Time to learn)	5	3.8	1.3	$\chi^2(2) = 5.7$, $0.05 < p < 0.10$

The null hypothesis could not be rejected for any of the questions which indicates that the opinion of the participants is not significantly skewed to any of the categories. However, interpretation of the results using an α-level of 0.10 allows the null hypothesis to be rejected for the time to learn to use the application when measured on a scale of slow to fast.

The chart below is a stacked bar graph for the number of responses in each category. The numbers in each stack indicate the number of responses in that particular category.

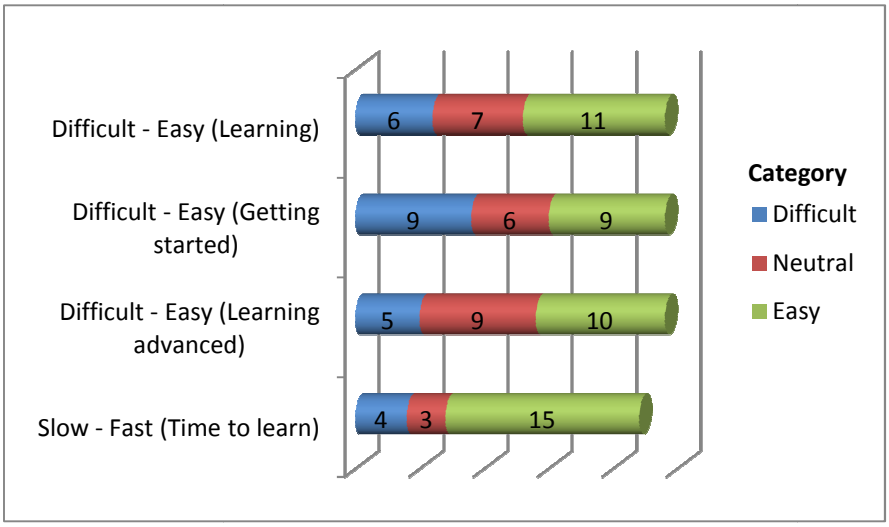


Chart 8.2: Number of responses in each category of the typing feature learnability questions

From the graph it can be seen that the vast majority of the participants felt that learning to type with the onscreen keyboard was fast. Getting started with the typing was equally considered to be easy and difficult since both categories had 9 responses. This could possibly be attributed to the fact that some users have a predisposition to embrace new advancements and find it easy and quick to get into the habit of using the new features while others may balk at the idea of having to change the way they interact with an application. In general, it seems as though participants found that learning to type with eye gaze and speech was relatively easy.

8.4.3 Preference and ease of use for typing settings

During longitudinal testing, participants used three different configurations for the onscreen keyboard, namely (1) large buttons, (2) smaller buttons which were widely spaced and (3) smaller buttons which were spaced closer together. Participants were asked to rank both their preference (Appendix H, question 21) of these three configurations as well as the order in which they found them easiest to use (Appendix H, question 22). Since subjective preference does not always mirror objective performance, it was decided to pose both these questions to determine if the preference of the participants differed from the ease with which they could use the onscreen keyboards.

Charts 8.3 and 8.4 below respectively show stacked bar graphs for the preference ranking and the ease of use ranking. The numbers indicate the number of responses in each category. The blue bar indicates the number of respondents who ranked the specific keyboard setup as their first preference, the red is their second preference and green the third most preferred setup.

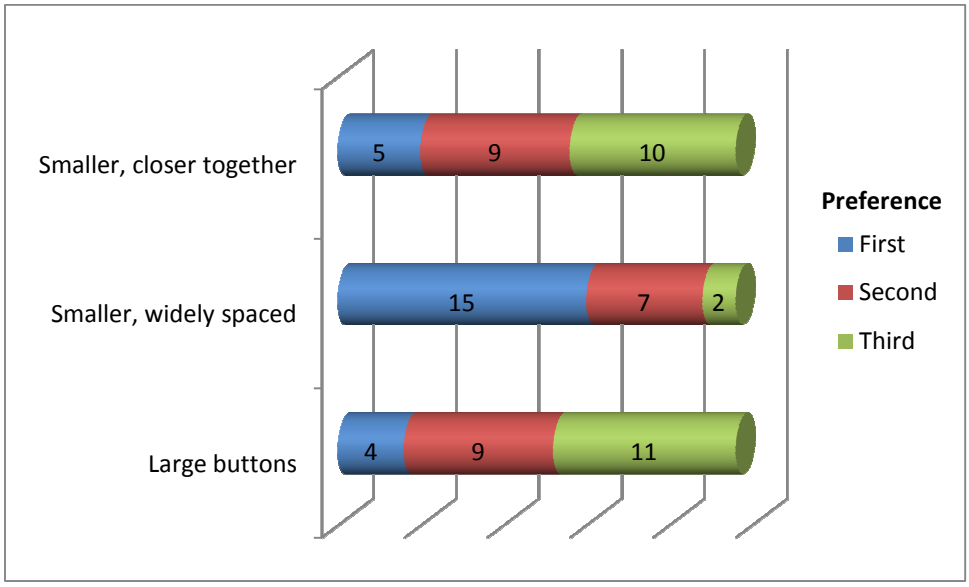


Chart 8.3: Preference ranking of the onscreen keyboard setups

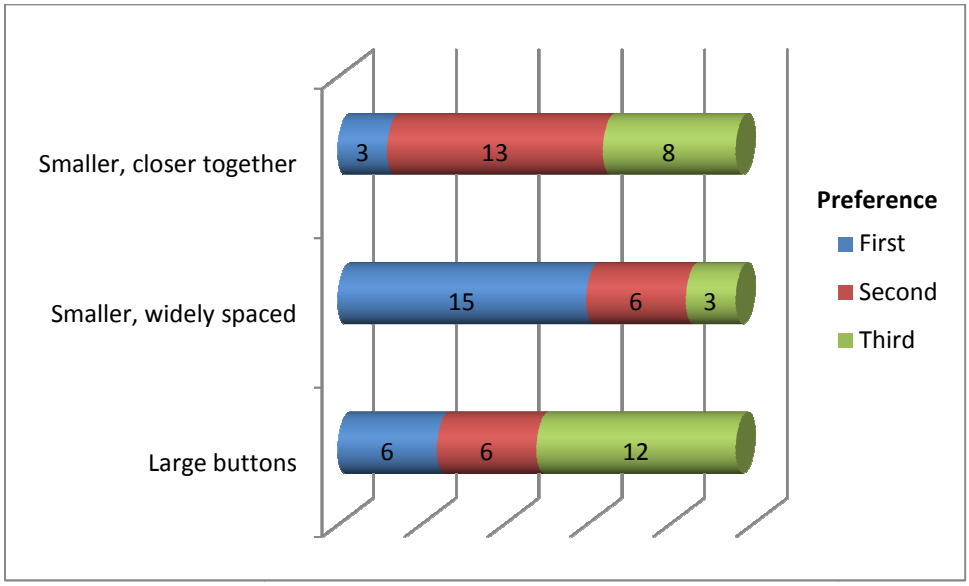


Chart 8.4: Ease of use ranking for the onscreen keyboard settings

The charts clearly indicate that preference is highest for the smaller widely spaced buttons and that the majority of the participants also found these the easiest to use. The larger buttons were the least preferred and were also judged to be the least usable by the majority of the participants. The following is the contingency table used for the participant preference:

Table 8.10: Contingency table for keyboard setup preference

	First preference	Second preference	Third preference
Small, closer together	5	9	10
Small, widely spaced	15	7	2
Large buttons	4	9	11

The following hypotheses were formulated for this analysis:

1. $H_{0,1}$: Participants' preference is independent from the keyboard setup.
2. $H_{0,2}$: The perceived ease of use is independent from the keyboard setup.

$H_{0,1}$ could be rejected at an α-level of 0.05 ($\chi^2(4) = 15.9$, $p < 0.05$) which suggests that there is a significant preference for a certain keyboard setup. The smaller widely spaced buttons were chosen by the majority of the participants as their most preferred setup. Therefore, it could be concluded that there was a significant preference for this setup although the results of Chapter 7 indicated that there was no significant difference between the effectiveness and efficiency of these setups.

The second null hypothesis could also be rejected at an α-level of 0.05 ($\chi^2(4) = 19.0$, $p < 0.05$), which indicates that a certain setup is significantly easier to use than others. Once again, the smaller widely spaced buttons were chosen as the easiest to use therefore it could be said that they were perceived as being significantly easier to use than the other setups. Similar to the preference of the setups, although there was a significant difference in the subjective usability of these setups, results from the previous chapter indicate that there was no objective difference between these three setups.

8.5 Commands

8.5.1 Satisfaction

Using the same procedure as with the previous sections, the subjective satisfaction of the participants towards using the speech commands was gauged and divided into categories. The same questions as in the previous sections were used and then categorised into the negative, neutral and positive categories. The following hypothesis was formulated:

H_0: The responses in the possible categories are evenly distributed.

A contingency table for the first question is given as an example below.

Table 8.11: Example of contingency table for satisfaction with speech commands

	Terrible Responses = 1 or 2	Average Responses = 3	High Responses = 4 or 5
Observed responses	2	8	14
Expected responses	8	8	8

Descriptive statistics for these questions are summarised in Table 8.12. The first column contains the adjectives used on the scale for the questions and the rightmost column summarises the results of the Chi-square test performed on each question.

Table 8.12: Descriptive statistics for satisfaction questions for the command feature

	Mode	Mean	Standard deviation	Chi-square test
Terrible – Wonderful	4	3.7	0.9	$\chi^2(2) = 5.2$, $0.05 < p < 0.10$
Frustrating – Satisfying	3	3.4	1.1	$\chi^2(2) = 1.0$, $p > 0.05$
Dull – Stimulating	4	4.0	0.8	$\chi^2(2) = 10.0$, $p < 0.05$
Difficult – Easy	3	3.7	1.0	$\chi^2(2) = 3.1$, $p > 0.05$
Inadequate – Adequate	4	3.5	1.0	$\chi^2(2) = 2.6$, $p > 0.05$
Rigid – Flexible	4	3.6	1.1	$\chi^2(2) = 2.2$, $p > 0.05$
Unnatural – Natural	3	3.6	0.9	$\chi^2(2) = 4.6$, $0.05 < p < 0.10$

The null hypothesis could not be rejected for any of the questions at an α-level of 0.05 which indicates that there was no significant difference in the level of satisfaction for any of the questions. The scales of terrible – wonderful and unnatural – natural could however be rejected at an α-level of 0.1. The majority of the participants felt that the use of commands was wonderful and that it felt natural. The responses of the questions were mapped to a scale of 1 to 5 and then a mean score was calculated for each question using these mappings. The mean scores for the questions, as measured on a scale of 1 to 5, were all positive for the

speech commands. This tendency was confirmed through inspection of Chart 8.5 which shows the number of responses in each category.

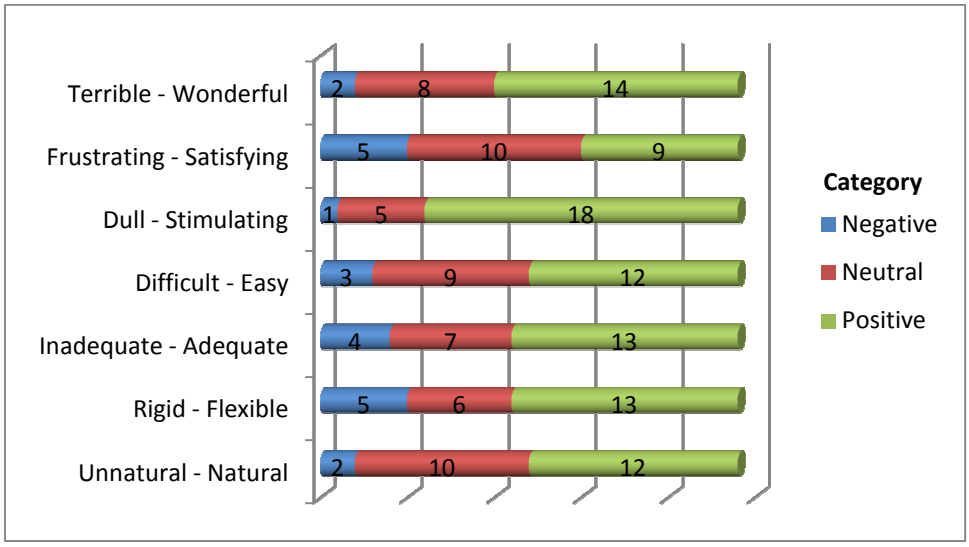


Chart 8.5: Number of responses in each category for satisfaction questions for command feature

From the graph it can clearly be seen that using speech commands was judged satisfactory by a larger number of participants for all the rating scales than the typing. Similar to the typing feature, the scale of frustrating – satisfying had the lowest number of responses for high satisfaction. This could easily be attributed to the fact that ambient noise often interfered with the speech recognition and sometimes led to unexpected responses from the application. This could at times be frustrating for the participants. Even so, the level of satisfaction experienced was more than acceptable.

In comparison with the satisfaction experienced with typing, the commands appear to be more gratifying for the participants. They were ranked as more stimulating, adequate, flexible and natural by more participants than the typing was. The objective results reported in Chapter 6 also indicated that in some instances the use of speech commands was comparable to user performance when using the keyboard, while in others the speech commands performed even better than the keyboard. Therefore, in this case user satisfaction closely mirrors the objective usability measurements.

8.5.2 Learnability

The same four questions used to gauge learnability in the previous two sections were posed to the participants in respect of the learnability of the commands. Similar to previous sections, the responses were grouped into the three categories of difficult, neutral and easy (see Table 8.7 for an example). The following hypothesis was formulated:

H_0: The responses in the possible categories are evenly distributed.

Descriptive statistics for the questions relating to learnability are summarised below. Scores are measured on a scale of 1 to 5. The first column indicates the question together with the scale used, while the final column reports the results of the Chi-square test.

The null hypothesis could not be rejected for any of the questions which indicates no significant number of responses in a particular category. However, the mean response for all of the questions ranks on the positive side of the scale which offers some encouragement as to the learnability of the commands.

Table 8.13: Descriptive statistics for learnability questions for the command feature

	Mode	Mean	Standard deviation	Chi-square test
Difficult – Easy (Overall learning)	4	3.6	1.1	$\chi^2(2) = 3.3$, $p > 0.05$
Difficult – Easy (Getting started)	3	3.5	1.2	$\chi^2(2) = 1.6$, $p > 0.05$
Difficult – Easy (Learning advanced features)	4	3.6	1.1	$\chi^2(2) = 2.6$, $p > 0.05$
Slow – Fast (Time to learn)	3	3.6	1.0	$\chi^2(2) = 4.5$, $p > 0.05$

8.5.3 Types of commands

The questionnaire also required the respondents to distinguish between the different types of actions and functions that could be achieved through the use of speech commands and to rank them on a 5-point scale ranging from difficult to easy. The four types of functions which could be completed using speech commands were navigation (moving the cursor), formatting text (for example, bold and italic), selecting text and actions such copying, cutting and pasting. The following table is the contingency table which was used to analyse the satisfaction with moving the cursor:

Table 8.14: Contingency table for satisfaction with moving the cursor

	Low	Neutral	High
Observed responses	9	2	12
Expected mean	8	8	8

The table below contains descriptive statistics for these questions as well as the Chi-square test results for each question. The following hypothesis was tested for each question:

H_0: The responses in the possible categories are evenly distributed.

Table 8.15: Descriptive statistics for satisfaction of command types

	Mode	Mean	Standard deviation	Chi-square test
Moving the cursor	2	3.3	1.4	$\chi^2(2) = 4.4$, $p > 0.05$
Formatting text	5	4.3	0.7	$\chi^2(2) = 11.2$ *, $p < 0.05$
Selecting text	5	4.3	1.0	$\chi^2(2) = 12.3$, $p < 0.05$
Cutting/copying and pasting	5	4.6	0.7	$\chi^2(2) = 16.4$ *, $p < 0.05$

* Yates corrected Chi-square applied

From the table it can be extrapolated that the null hypothesis could be rejected for all functions except for navigation. Nine participants felt that navigation was difficult while twelve felt that it was easy. With the two neutral responses this gives a fairly even spread of opinions of the difficulty level of navigating using speech commands. In contrast, twenty participants felt that formatting text was easy, 20 indicated that selecting text was easy and 23 found cutting/copying and pasting easy. Therefore, it could be concluded that functions achievable through the issuing of speech commands were considered easy by the majority of the participants except for having to navigate through the document. Since it was discovered during the analysis of the number of actions used that most participants moved the cursor one character at a time and failed to use shorter methods of moving the cursor, it is entirely plausible that participants found navigation a laborious process when having to use speech commands. Even so, when inspecting Chart 8.6 below, it is clear that when grouping the responses in the difficult, average and easy categories, most of the respondents found this function easy to master.

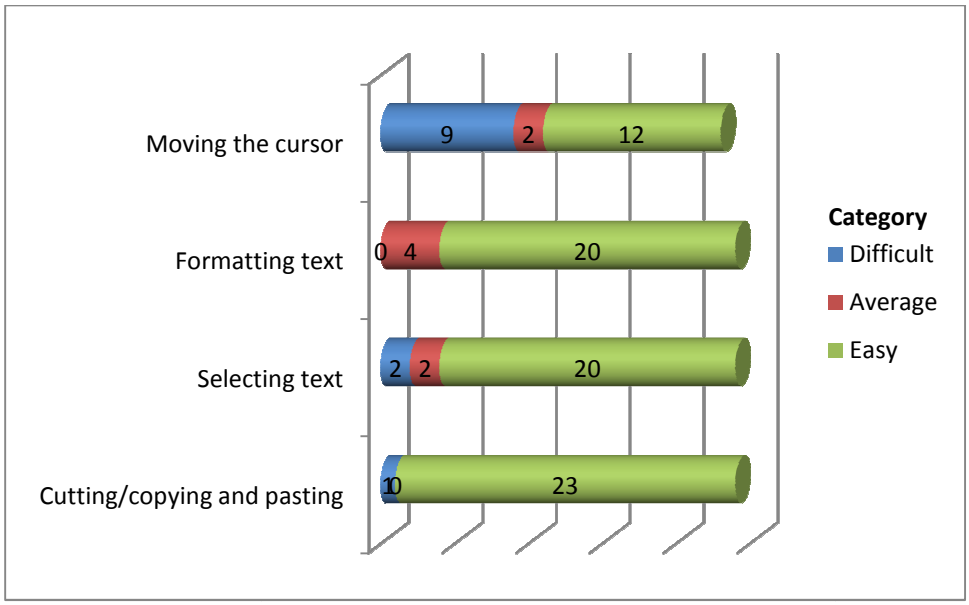


Chart 8.6: Number of responses in each satisfaction category for command types

8.6 Additional considerations

The questionnaire also contained a number of yes/no type questions to gauge user response and willingness to continue using the features of the multimodal interface (Appendix H, questions 15 – 20). Chart 8.7 shows a stacked graph of the response rate for each of these questions, which for these purposes have been rewritten in a shorter format.

The graph clearly shows that the majority (14) of the respondents would prefer to have both visual and audio feedback when a keyboard button is pressed. While the audio feedback serves the purpose of alerting the user to the fact that a character has successfully been inserted into the document (21 participants responded positively), participants seem to prefer that a visual cue be given in order to confirm which letter was inserted into the document. During development of the application it was considered sufficient that the button which had focus be framed and that this would serve as an indicator as to which character had been inserted. Clearly, while this assists the participants in the knowledge of which button has focus, they desire a different mechanism as confirmation of activation. Therefore, in this instance, user preference closely correlates to the option which was found to have the highest text entry speeds (Majaranta, 2009).

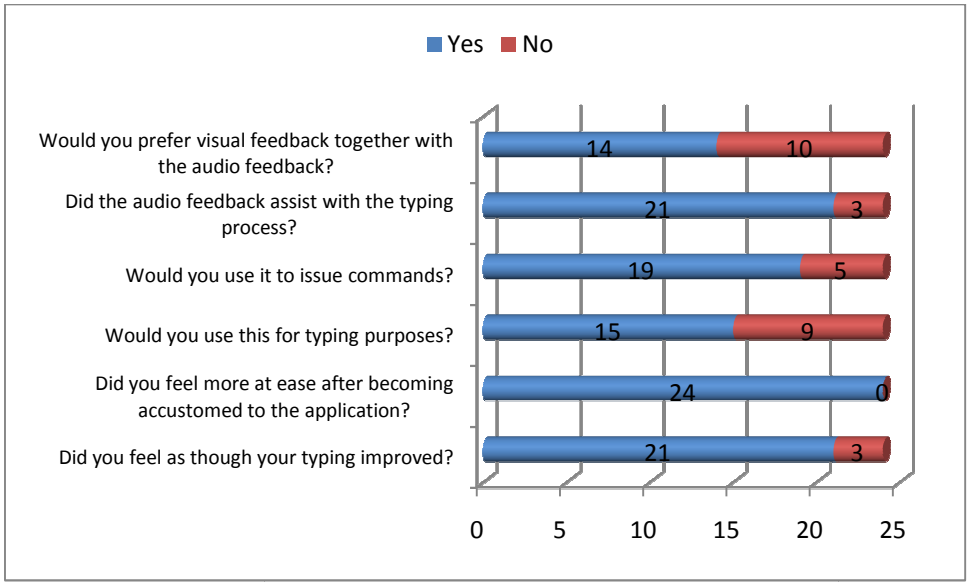


Chart 8.7: Number of responses in each category for additional considerations of using eye gaze and speech

The responses to indicate continued use of the features are very encouraging because in all instances the participants showed an overwhelming eagerness to continue using the features made available for them. Most participants would consider using the speech commands as well as the onscreen typing. Furthermore they felt more at ease with the system the more they used it which is an important factor to bear in mind in terms of product acceptance. Results in Chapter 7 indicate that there was no significant improvement in the typing speeds as the weeks progressed, but the fact that participants felt they had improved could improve the acceptance of and satisfaction with the multimodal interface.

The table below contains the results of the Chi-square test used to analyse the following hypothesis:

H_0: The responses are evenly distributed over the possible categories.

Table 8.16: Analysis results for satisfaction of additional considerations

	Chi-square test
Did you feel as though your typing improved?	$\chi^2(2) = 7.9$, $p < 0.05$
Did you feel more at ease after becoming accustomed to the application?	$\chi^2(2) = 13.4$ *, $p < 0.05$
Would you use this for typing purposes?	$\chi^2(2) = 0.8$, $p > 0.05$
Would you use it to issue commands?	$\chi^2(2) = 4.5$, $p < 0.05$
Did the audio feedback assist with the typing process?	$\chi^2(2) = 7.9$, $p < 0.05$
Would you prefer visual feedback together with the audio feedback?	$\chi^2(2) = 0.3$, $p > 0.05$

* Yates corrected Chi-square applied

The null hypothesis could be rejected for all questions except for the question pertaining to the inclusion of visual feedback. These results confirm that a significantly larger proportion of the respondents would consider continued use of the application for both commands and typing. Additionally, they felt as though their typing

skills with the onscreen keyboard had improved and that they had become more at ease with the application as they got more exposure. The findings also confirm the importance of feedback as a significant number of participants felt that audio feedback enhanced the typing process.

8.7 Pointing device

The device assessment questionnaire advocated for use by the ISO 9241-9 for the usability of pointing devices was given to the participants after the first session and the last session (Appendix H). They were instructed to complete that part of the questionnaire with reference to using eye gaze and speech for text input. The 5-point scale used was divided into three categories for each question using the two lowest scale points, the midpoint and the two highest. The table below serves as an example and is a contingency table for the first question pertaining to actuation.

Table 8.17: Example of a contingency table for device assessment questions

	Low	Neutral	High
First session	6	11	8
Last session	3	16	5

A Chi-square test was used to determine whether the opinion of the participants changed between the first and the last session. Descriptive statistics and results of the Chi-square test are summarised in Table 8.18. The first column indicates the characteristic of the device which was being evaluated as well as the scale on which it was evaluated. The following null hypothesis was formulated:

> H_0: There is no difference in the assessment of the device after the first session and after the last session.

The null hypothesis could not be rejected for the majority of the assessment questions which suggests that the opinion of the participants to the use of eye gaze and speech was not swayed during extended use thereof. The only question for which there was a significant difference in the responses was concerned with the smoothness of the operation. In this instance, participants were more positive about the smoothness of the operation after prolonged use of the application. Considering that the overall response to the eye gaze and speech was positive in terms of the effort required, the speeds achieved and the fatigue or discomfort it may cause, this is a very encouraging result.

8.8 Anecdotal observations

During each session, participants were closely observed by the facilitator as they completed the tasks. A number of general observations were made during the ten week period. The fact that the application was susceptible to ambient noise appeared to cause the greatest frustration for the participants. In the first few weeks they did resort to “hyperarticulation” (Oviatt et al., 1998) when the application did not respond as they wanted it to. Even so, after some weeks the participants learnt to compensate for these problems and did not react negatively as they had in the first few weeks. This observation confirms the suspicion that users will be flexible and resilient enough to overcome some of the limitations of the technologies (Nusbaum et al., 1995). Furthermore, the fact that some ambient noise was present provided a more natural and “real-world” environment which allowed the application to be tested in an environment resembling one in which actual use may occur, such as an office setting.

Table 8.18: Descriptive statistics for device assessment questionnaire responses

	First session			Last session			
	Mode	Mean	Std dev	Mode	Mean	Std dev	Chi-square test
Actuation (too low – too high)	3	3.2	1.1	3	3.0	0.7	$\chi^2(2) = 2.6$, $p > 0.05$
Smoothness (very rough – very smooth)	3	3.0	1.1	3	3.3	0.7	$\chi^2(2) = 7.1$, $p < 0.05$
Mental effort (too low – too high)	2	3.2	1.0	3	3.0	1.0	$\chi^2(2) = 0.6$, $p > 0.05$
Physical effort (too low – too high)	2	2.6	1.2	2	2.5	1.1	$\chi^2(2) = 2.5$, $p > 0.05$
Accurate pointing (easy – difficult)	3	3.6	1.1	4	3.6	0.9	$\chi^2(2) = 4.3$, $p > 0.05$
Operation speed (too fast – too slow)	3	2.6	0.9	3	2.7	0.8	$\chi^2(2) = 0.2$, $p > 0.05$
Neck fatigue (none – very high)	1	2.0	1.1	3	2.9	1.3	$\chi^2(2) = 4.7$, $0.05 < p < 0.10$
General comfort (very uncomfortable – very comfortable)	4	3.8	1.0	3	3.3	1.0	$\chi^2(2) = 2.5$, $p > 0.05$
Overall use (very difficult – very easy)	4	3.4	1.3	4	3.5	0.8	$\chi^2(2) = 1.370$, $p > 0.05$

Due to the orientation of the onscreen keyboard it was found that accuracy was heightened if a participant looked down at the keyboard with the screen tilted slightly downwards.

Informal conversations with the participants were also held after each session. During these conversations it became clear that participants enjoyed using the application the more they were exposed to it. They also became more comfortable issuing verbal commands and using the onscreen keyboard to type. The opinions captured during the interviews were the same for many of the participants and unfortunately did not serve to complement the study to the degree which was initially hoped when the experimental methodology was planned.

One participant die share a fairly humorous experience he had during the testing period. He had become so accustomed to using the onscreen keyboard and enjoyed using it so much that when faced with typing an assignment he found himself issuing verbal commands to complete both the typing and formatting. Naturally, the application did not respond, which led to great disappointment on his part. Quite a large majority of the participants indicated that they would use the application if they had access to it.

Observations of some of the participants over the weeks led to the supposition that they valued speed above accuracy when typing using the onscreen keyboard. When typing using the onscreen keyboard it became obvious that they were attempting to achieve high typing speeds, often to the detriment of accuracy as they scarcely gave the indicator time to stabilise before issuing commands and looking quickly at the next letter. During an interview after the ninth session, one participant commented that he found it easier to move along the keys on the keyboard one at a time than trying to focus directly on the next desired letter from the previously typed one. He perceived this as being much easier to type and maintain a stable gaze. Although this would increase his typing time, he felt the increase in the accuracy was a worthwhile trade-off. This behaviour was also observed in some of the other participants which was in direct contrast to some of the observed behaviour mentioned before. This indicates that the desired balance of accuracy and speed is a highly subjective one.

8.10 Summary

This chapter discussed the results of the questionnaires which were completed by study participants after the first and last session of the longitudinal user testing. The questionnaires were designed to elicit the subjective opinions of the participants towards the use of speech for issuing commands in a word processor, using eye gaze and speech to type and the overall satisfaction experienced with the application.

Overall, the satisfaction of the participants was positive with very little shift in opinion from the first session to the last. However, the use of eye gaze and speech to type was ranked as very unnatural – a mindset which could perhaps be changed after more exposure and practice with the system. The positive response to the satisfaction and learnability of the application is heartening as this may increase the acceptance rate of such an application by mainstream users of a word processor.

The next chapter will provide a discussion on the results of the complete study and make recommendations for further study and use.

CHAPTER 9

CONCLUSION

9.1 Introduction

The previous chapters reported on the results of the statistical analysis of the user testing. This chapter will start by providing a summary of the results and how they serve to answer the research questions which were originally posed. This will be followed by the limitations of the study and the recommendations based on the findings. Finally, the implications the study has for the future will be discussed.

9.2 Motivation

The motivation of the study was essentially threefold. Firstly, as the word processor is a popular, everyday tool of a majority of computer users, it offers an environment well-suited to improvement and exploitation of emerging technologies. Secondly, the future of user interfaces indicates that there will be a movement away from traditional GUI interfaces. This presented an opportunity to provide and test a multimodal interface for the word processor which uses non-traditional interaction techniques. Thirdly, this could offer a customisable interface which could potentially cater for a very diverse group of users. In particular, it offered the potential of including the oft-marginalised disabled users into the mainstream group of users by providing a means of interaction which is not dependent on the keyboard and mouse.

9.3 Aim

The study had three main aims which were investigated. The first aim was simply to determine whether a customisable multimodal interface could be developed for a mainstream word processor. This interface should have the potential to offer a variety of interaction means which could be set according to the needs and environment of the user. The second aim was to determine whether the interface that was developed was feasible within the confines of the mainstream word processor and whether it exhibited long term potential as a viable future interface. The final aim was to determine how usable eye gaze and speech are as an interaction technique within a word processor. This aim could be subdivided into three secondary aims based on the types of interaction required within a word processor. To begin with, it was necessary to determine whether eye gaze and speech could be used to replace a pointing device. Thereafter, it had to be determined whether common word processing tasks could be accomplished using the proposed interaction techniques. Finally, text entry is the most integral part of a word processor and it had to be established whether eye gaze and speech could be used for text entry in a usable manner and whether it was comparable to the traditional means of text entry.

9.4 Results

Each of the aims, which led to individual research questions, was explored using specific tests and experimental designs. The results of the tests will be briefly summarised in this section in the order that they were conducted to answer the research questions.

9.4.1 Multimodal word processor

Can a customisable multimodal interface be developed and successfully incorporated into a mainstream word processor?

A multimodal interface using eye gaze and speech was developed and incorporated into Microsoft Word 2007 (Chapter 3). The multimodal interface provided a number of different interaction techniques which facilitated a hands-free environment. The main aim was to provide an interface which was highly customisable since a word processor is used by a very diverse group of users and in varying environments.

As a result, the multimodal interface provided a speech grammar for common word processing tasks such as formatting, navigation, text manipulation and text selection. Eye gaze was then incorporated into the interface using identified activation mechanisms. Dwell time, look-and-shoot (with the Enter key) and blinking were all available to use as interaction techniques. The configuration of the onscreen keyboard was QWERTY by default but could also be changed to an alphabetical layout. The setting used for dwell time as well as the sensitivity of the pronounced blink could be set by the users to meet their needs.

The interaction of interest in this case was the combination of eye gaze and speech. Using eye gaze to control the focus and indicate which letter was required, the user could issue a speech command to type the letter in the current document at the current cursor position. The combination of eye gaze and speech could provide improved speed as it is not dependent on a dwell time. It could also exhibit increased accuracy as there is almost no potential for inadvertently activating the incorrect button due to prolonged eye gaze or a blink.

Potentially, in order to increase the accuracy even more, magnification could also be used to increase the area directly under the gaze of the user. All activation techniques were still available to be used while the magnification was turned on.

The resulting interface was one which incorporated speech in terms of dictation and speech commands and eye gaze which could be used in a number of ways in isolation or with speech as an activation mechanism.

The successful development and incorporation within Word 2007 proved that such a highly customisable, multimodal interface was indeed possible for both a multimodal interface and within a mainstream application.

9.4.2 Feasibility study

How feasible is such an interface and in which context is it feasible?

The development of such an interface does not implicitly mean that such an interface is feasible or viable by any means. Therefore, before user testing commenced, a feasibility study was undertaken to determine whether there was any potential for adoption of such an interface.

In order to determine this, a number of experienced HCI researchers participated in a short demonstration of the application, interacted with the system and then completed a questionnaire designed to determine whether they felt that the multimodal interface had long-term use potential. Results indicated that the interface was a move in the right direction, particularly as a means of exposing disabled users to a mainstream application. Furthermore, they were positive about the potential that the interface offered and felt that in the long-term the use thereof may be beneficial to a wide group of users.

9.4.3 User testing

How usable is the multimodal interface compared to the traditional interaction techniques?

The afore-mentioned results could not be statistically verified and served only as a means to determine whether the proposed interface was possible and viable. The next step of the study was to determine how usable the application was. However, a comparison of all the interaction techniques would require a large scale project to be undertaken which was beyond the scope of this thesis. Therefore, the proposed interaction technique of eye gaze and speech was tested as a pointing device, to format a document and for text entry. Where applicable, this was then compared to other means of input. This allowed three secondary questions to be posed in order to answer the above-mentioned research question.

9.4.3.1 Usability of eye gaze and speech as a pointing technique

How usable is the combination of eye gaze and speech when used to replace a pointing device?

The multi-directional tapping ISO assessment (ISO, 2000) was used to determine the usability of eye gaze and speech to replace a pointing device. Throughput, selection time, target re-entries, incorrect target acquisitions and incorrect clicks were compared for eye gaze and speech when using a gravitational well (ETSG), magnification (ETSM) and the absence of both (ETS). These were then compared to the mouse as a pointing device.

The feedback provided to users did not influence the usability of the interaction technique in any way although participants preferred using the framed feedback. The inverted colour results in the button being a very dark gray which may be too harsh for the majority of the users. A background colour which is not quite so dark may be more pleasant for the users.

The mouse had a significantly higher throughput than all of the eye gaze interaction techniques, but inclusion of the gravitational well allowed the eye gaze to perform significantly better than the other eye gaze interaction techniques. The mouse throughput was within the expected range as reported by Soukoreff and MacKenzie (2004) but all interaction techniques did exhibit improvement over the three sessions.

The behaviour used to stabilise and select targets in terms of incorrect target acquisitions and target re-entries suggests that ETS and ETSM require the most effort to maintain a stable eye gaze. This is contrary to expectations since the magnified button should become easier to acquire and maintain due to its larger surface area. The high incidence of target re-entries indicates that the eye gaze tends to slip off the target frequently even though theoretically the larger surface area should increase the ease with which the target can remain selected. Participants also tended to attempt to fine-tune the position when using ETSM possibly due to the disturbance experienced by the magnification, the feedback given by the mouse pointer which indicates the gaze position may be close to the button or the fact that the larger button is perceived to be easier to acquire. In contrast, when using ETS participants tended to focus on another button and then glance back to the required target. Therefore, in this instance it is possible that the feedback provided by the magnification tool may have altered the behaviour of the user. However, the evidence suggests that it in no way increased the efficiency or effectiveness of target selection and therefore it is not recommended for use. In response to the questions posed in the questionnaire it would appear that the use of the magnification is not a pleasant experience for the participants but no comments were made specifically about the visual feedback which was given.

Incorrect clicks were experienced with all eye gaze interaction techniques although more so with ETSG. Participants did, however, improve dramatically with ETSG to a level that was comparable with the other eye

gaze techniques. Incorrect clicks were most probably caused by the fact that participants have a tendency to acquire the next target whilst still issuing the verbal command. Once the verbal command has been processed, they have already achieved a stable gaze on the next target, thus causing this target to be selected. This closely resembles the behaviour seen when performing an action (Land & Tatler, 2009) or issuing verbal commands in an application (Maglio et al., 2000). If this theory holds, the number of missed clicks should be higher for ETSM and ETS as the click would occur while users were still attempting to acquire the next target. Unfortunately these were not measured and this will be held over for possible further research to confirm the proposed theory.

It was found that the use of a gravitational well significantly improves the usability of eye gaze and speech as a pointing device, with this interaction technique being the closest rival to the mouse in terms of usability. While magnification is proposed as a means to increase the accuracy of eye gaze selection, it was found in this case, that magnification did not increase the usability of eye gaze and speech as a pointing device at all.

9.4.3.2 Usability of speech commands

How usable are speech commands for performing common word processing tasks?

The next phase of the study was to test the multimodal interface within Microsoft Word. For this purpose, user testing was identified as the most suitable means of exploring the research question. Representative word processing tasks which encompassed navigation, formatting and editing were selected for use. Objective usability measurements were captured while users completed the tasks using the speech commands and when completing them with the traditional interaction techniques.

The tasks could be divided into generalised types of tasks. The table below contains a summary of the results. Where there was a significant difference between the interaction techniques, an “S” indicates that speech was significantly better while “K” indicates that the keyboard was significantly better. Where there is a tick mark it indicates that both interaction techniques improved significantly over time.

Table 9.1: Summary of results for speech commands

	$H_{0,1}$: Interaction technique		$H_{0,2}$: Session	
	Completion time	Number of actions	Completion time	Number of actions
Line selection and formatting			✓	S
Select all and remove	S	S	✓	✓
Select words and format	K		✓	✓
Paste	S	S	✓	
Undo			✓	✓
Select word and copy			✓	
Position and paste	K	K	✓	✓

In most instances, speech performed comparably with the keyboard and sometimes even outperformed it for both measurements. Selection of words and lines were completed with comparable efficiency and effectiveness with the keyboard and with speech. Since selection of a line at a time and the whole document were more efficient with the speech or similar to using the keyboard, other selection techniques such as selection of a paragraph should be easily accommodated into the grammar and adopted by the users. It was only where isolated words had to be navigated to that the speech could not compete with the keyboard.

This could be due to two reasons. Firstly, users may not have realised that using a command which had an immediate effect that was not required could eventually lead to the correct result. Secondly, it was perhaps

due to the fact that users are not able to navigate efficiently with the keyboard only and therefore could not translate the verbal commands into efficient navigation techniques. It may be advisable to first train the users in using the keyboard for navigation and then retest them. However, since the word processor enjoys such widespread use, the grammar should ideally be as intuitive as possible and not require any training. Therefore, for other means of navigation, more intuitive commands may have to be provided. To reduce the chances that unwanted connotations may be invoked or that the user won't understand the command, it is advised that a study be conducted in which participants are shown how to use the keyboard for navigation and then asked to suggest verbal commands for these navigational techniques.

The results of this study allow the conclusion to be drawn that speech commands can be used effectively and efficiently in an editing environment and that the use of a menu-orientated grammar may induce rapid learning and use of the grammar.

9.4.3.3 Usability for text entry

How usable is the combination of eye gaze and speech when used for text entry?

Text entry was also tested using longitudinal user testing by requiring users to enter phrases from a pre-selected phrase set using both the keyboard and eye gaze and speech with an onscreen keyboard. Three onscreen keyboard configurations were tested, namely large buttons, smaller widely spaced buttons and smaller closely spaced buttons. In all instances, the QWERTY keyboard layout was used.

When comparing the keyboard with the large buttons, the keyboard had significantly fewer errors than the eye gaze and speech. However, there was significant improvement between the second and last sessions and the third and last sessions. Therefore, the number of errors decreased dramatically as the participants became accustomed to the interaction technique. When inspecting the insertions, deletions and substitutions it was found that there were significant differences between the insertions and substitutions. The higher incidence of insertions and substitutions corresponds closely with the finding of the high number of incorrect clicks with this interaction technique. Users were instructed not to delete erroneous characters as this would allow an accurate measure of the types of errors that were made. Therefore, the fact that the average number of insertions was higher than the average substitutions indicates that the users at least noticed their errors and then inserted the correct character. Future studies could allow users to correct their errors and then measure both the correctness of the transcribed text as well as the number of corrections required.

Text entry using the keyboard was also significantly faster than using eye gaze and speech. There was no significant improvement over the sessions for the eye gaze and speech input.

When comparing the keyboard, large buttons, smaller widely spaced buttons and smaller, closely spaced buttons, the keyboard differed significantly from both of the smaller button configurations in terms of the number of errors made. In terms of text entry speed, the keyboard was significantly faster than all other input configurations.

9.4.3.4 Satisfaction

The overall reaction to the system was fairly positive and on a comparable level after the first exposure and extended exposure. The majority of the participants preferred using the smaller, widely spaced buttons even though they did not facilitate a faster typing speed. This preference could be due to the reduced space that is occupied by the keyboard or the fact that the smaller buttons resemble standard sized buttons more than the others. Therefore, there could be consequences for the space which is lost to the onscreen widgets required for eye gaze interaction although direct questions should be posed to elicit this.

In terms of the naturalness and satisfaction experienced when using speech commands, most participants felt that they were most natural and enjoyed using them. The types of commands which were used were also pleasant to the participants although satisfaction was lowest for the navigation using speech commands. This corresponds closely with the findings of the objective usability measures.

Many participants would have preferred having visual as well as audio feedback while typing. The use of visual feedback could increase the typing speed as it cannot be definitively stated that participants did not look at the document in order to confirm their typing progress. This could have a significant impact on the typing speed of users.

9.5 Recommendations

Eye gaze and speech appears to be a suitable combination for pointing, in particular when a gravitational well is activated around the targets.

Originally, it was assumed that the presence of the visual feedback would be enough incentive for the users to keep a steady gaze on the button until the button had been activated. However, it seemed to be a natural occurrence for the user to glance at the next target whilst still issuing the command. This resulted in significantly more incorrect clicks with the ISO testing and quite possibly a higher error rate with the text entry. A recommendation to solve this problem would be to activate the button which has focus at the start of the utterance and not at the end of the utterance. This could increase the accuracy as well as the speed with which text can be input. Higher speeds can be attained as the user will not have to wait for confirmation before proceeding. One drawback of this method is the confirmatory audio beep which was given to alert users that the button had successfully been activated. If the user is allowed to look at the next target before the click has been executed it means the auditory feedback will have to be given *before* the successful execution of the command. If the auditory feedback is given at the end, then the user may already be focusing on the next button and the feedback could cause some confusion. Since users indicated a high preference for audio feedback, the usability of no audio feedback, a more premature beep and a delayed beep will have to be investigated.

The speech grammar should be minimised to include only the commands for typing when gaze is detected on the keyboard. This should be automatic and could potentially reduce the number of errors incurred since, for example, the cursor will not move around because of triggering through ambient noise. In order to allow formatting to occur while typing, there should be a mechanism for the user to extend the grammar to its full length as well as automatically extending it when the eye gaze leaves the keyboard. Furthermore, speech is recommended for use as an input technique to accomplish common word processing tasks.

Recommendations for the use of eye gaze and speech for text entry cannot be made at this stage as results were inconclusive when comparing objective and subjective measurements. Objective measurements indicate less productivity but subjective measurements indicate that users enjoyed using it and would like to use it in future. Typing via the means of eye gaze and speech is no faster than other means using eye gaze and a more efficient means, such as word completion algorithms are needed. However, the method should not be summarily disregarded as users may simply require more practice in order to type faster. For example, the speeds which can be achieved for typing using a cell phone keyboard are often outstanding and are achieved through extended practice. Proper motivation to use the text entry method could increase the speed with which it is adopted and used.

In conclusion, the incorporation of a multimodal interface using eye gaze and speech in a mainstream word processor is recommended, as it increases the potential penetration of the application. Moreover, users are

accustomed to using technology to meet their specific needs and the use of such an interface could increase the satisfaction that users have with the application.

9.6 Implications for the future

As evidenced by the incorporation of the technologies used in the multimodal interface of this study, the time has perhaps dawned when they should be exploited as replacement interaction techniques. Speech recognition has become a standard feature in personal computers and is often available for dictation purposes. Similarly, there are packages available for purchase which can react to spoken commands (cf. Dragon, nd). Furthermore, the first fully-integrated eye-controlled laptop has recently been showcased at exhibitions (Tobii, 2011) and bodes well for the adoption of eye-tracking as a standard feature in personal computers. Cheaper, accurate eye-trackers (cf. Haro et al., 2000) are also available which could function just as well as a standard interaction technique.

Therefore, the fact that a popular mainstream application can be adapted to include a highly customisable, multimodal interface could be a step in the right direction for the next generation of interfaces. The multimodal user interface displays great potential and test results indicate that the interaction techniques can be used for pointing and selecting tasks and common word processing tasks. Moreover, it has been proven that speech recognition can indeed be used for editing commands in a word processor which was contrary to theoretical beliefs (Klarlund, 2003). This could mean that in the future a more diverse group of users can be accommodated and disabled users may no longer have to be relegated to using specialised applications.

The findings therefore suggest that the word processor is well placed to include such an interface in future developments as the technology is rapidly becoming available. As it is foreseen that access to the technologies by mainstream users is imminent, future word processors could be developed with multimodal interfaces incorporated.

9.7 Further research

The results of the study unlocked a myriad of possibilities for further research in this area. Firstly, there are a number of possibilities for testing the use of eye gaze and speech as a pointing device. For example, eye gaze and speech can be compared to dwell time activation to determine if it is comparable or even superior for pointing purposes. In order to negate the effects of a possibly slow recognition engine, a Wizard of Oz experiment can also be tested. Speech commands can also be captured and executed based on the gaze position prior to when the command was recognised and processed. Furthermore, the use of a double command system similar to the touch sensitive mouse can be investigated. Speech commands can also be tested in different environments since this study was conducted in a controlled environment which may not be indicative of the actual use of such a system. Secondly, since ribbon icons are larger and it has been shown that common tasks (which still have smaller icons) can be accommodated in speech grammar more complex tasks can be tested by expecting users to interact with the ribbon using eye gaze and speech. This will provide evidence as to whether a full-length grammar is required or whether the ribbon used in Word 2007 and onwards is conducive to eye gaze as a pointing device.

Thirdly, in terms of speeding up typing using eye gaze and speech there are a number of areas for further research. Word completion algorithms can be used to reduce the number of keys which have to be selected. Visual feedback can be coupled with the audio feedback which was used in order to confirm that a keyboard button has been pressed. Participants can be expected to practise more text entry tasks to determine whether the speed of text entry can be increased through protracted practice. Typing free text instead of presented

text can also be tested to determine whether the use of eye gaze negatively affects the compositional speed of the users.

Fourthly, a more diverse user group must be tested on the interface including disabled and aged users.

The study could also be replicated on a variety of eye-trackers which could have an impact on the results achieved. For example, a more accurate eye-tracker with higher precision could be tested. This could have a significant effect on the use of eye gaze as a pointing device both with and without the gravitational well. Furthermore, some of the cheaper, web-cam based eye-trackers could also be tested for their usability with the developed application.

9.8 Summary

Recent trends have indicated that the time has dawned to move away from the traditional direct manipulation interfaces. Non-command, attentive, perceptual and multimodal interfaces present a possible solution for the dilemma of providing a more natural and intuitive human-computer interaction.

Gestures, speech and eye gaze are some of the natural mechanisms which are used by humans during communication. These offer a means of improving human-computer interaction. Speech and eye gaze were concentrated on in this study to create a multimodal interface for a popular word processor.

The combination of eye gaze and speech could successfully be used to fulfil the needs of a pointing device, particularly when employed with a gravitational well. Furthermore, speech commands could be used to facilitate formatting of word processing documents. While text entry was slower than using a keyboard, indications are that there was an overall positive response to the interface and that it may well herald a suitable multimodal interface. The ease with which participants became accustomed to the interface is further proof of the naturalness and intuitiveness provided by speech and eye gaze. With constant progress being made in the development of the hardware required by such an interface, the proposed multimodal interface may well lay the foundation for a word processor to continue its exploitation of emerging technologies and remain a forerunner in the establishment of trends. While there is undoubtedly room for improvement and expansion, the use of eye gaze and speech has proven to be very promising.

REFERENCES

Abran, A., Suryn, W., Khelifi, A., Rilling, J. & Seffah, A. (2003). Consolidating the ISO usability models. *Proceedings of 11th International Software Quality Management Conference*, Glasgow, Scotland.

Accot, J. & Zhai, S. (1999). Performance evaluation of input devices in trajectory-based tasks: An application of the steering law. In *Proceedings of CHI 99*, Pittsburgh, Pennsylvania, United States of America, 466-472.

Al-Qaimari, G. & McRostie, D. (2001). KALDI: A CAUsE tool for supporting testing and analysis of user interaction. In A. Blandford, J. Vanderdonckt and P. Gray (Eds), *People and Computers XV – Interaction Without Frontiers: Joint proceedings of HCI 2001 and IHM 2001* (pp. 153-169). United Kingdom: Springer.

Anderson, T. (2009). *Pro Office 2007 development with VSTO*. United States of America: APress.

Ashdown, M. & Sato, Y. (2005). *Attentive interfaces for multiple monitors*. CHI 2005 Workshop on Distributed Display Environments, Portland, Oregon, United States of America.

Ashmore, M., Duchowski, A. & Shoemaker, G. (2005). Efficient Eye Pointing with a Fisheye Lens. In *Proceedings of Graphics Interface 2005*, 203-210.

Atchinson, D.A. & Smith. G. (2000). *Optics of the human eye*. Oxford: Butterworth-Heinemann.

Bahill, A.T. & Clark, M.R. (1975). Glissades – Eye Movements Generated by Mismatched Components of the Saccadic Motoneuronal Control Signal. *Mathematical Biosciences*, 26, 303-318.

Basson, S., Fairweather, P.G. & Hanson, V.L. (2007). Speech recognition and alternative interfaces for older users. *Interactions*, July/August 2007, 26-29.

Bates, R. (2002). Computer Input Device Selection Methodology for Users with High-Level Spinal Cord Injuries. In *Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT)*, 25-27 March 2002. Trinity Hall, University of Cambridge.

Bates, R. & Istance, H. (2002). Zooming interfaces! Enhancing the performance of eye controlled pointing devices. In *Proceedings of Assets 2002*, 119-126.

Bee, N. & André, E. (2008). Writing with your eye: A dwell time free writing system adapted to the nature of human eye gaze. *Perception in Multimodal Dialogue Systems*, 111-122. Berlin: Springer.

Beelders, T.R. (2009). *Graphics, text and language in a word processor interface*. Germany: VDM Verlag.

Berg, M., Gröber, P. & Weicht, M. (2010). User study: Talking to computers. In *Proceedings of the 3rd Workshop on Inclusive eLearning*, London, United Kingdom, 19-32.

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

Bergin, T.J. (2006b). The Proliferation and Consolidation of Word Processing Software: 1985-1995. *IEEE Annals of the History of Computing*, 28(4), 48-63.

Bernhaupt, R., Palanque, P., Winkler, M. & Navarre, D. (2007). Usability study of multi-modal interfaces using eye-tracking. In *Proceedings of INTERACT 2007*, 412-424.

Bevan, N. & Macleod, M. (1994). Usability measurement in context. *Behaviour and Information Technology*, 13, 132-145.

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? (Training of people from disadvantaged communities in computer usage: Is the backlog too large to overcome?) *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie,* 26(3), 216-235.

Bobick, A., Intille, S., Davis, J., Baird, F., Pinhanez, C., Campbell, L., Ivanov, Y., Schutte, A. & Wilson, A. (1999). The Kidsroom: A Perceptually-Based Interactive and Immersive Story Environment. *Presence: Teleoperators and Virtual Environments*, 8(4), 367-391.

Bohmann, K. (2000). User performance metrics. Retrieved 19 October 2005 from: http://www.bohmann.dk/articles/ user_performance_metrics.html.

Bolt, R. (1980). "Put-that-there": Voice and gesture at the graphics interface. *Computer Graphics*, 14(3), 262-270.

Bolt, R. (1981). Gaze-orchestrated dynamic windows. *Computer Graphics*, 15(3), 109-119.

Bradley, J.V. (1958). Complete counterbalancing of immediate sequential effects in a Latin square design. *Journal of the American Statistical Association*, 53(282), 525-528.

Carroll, J.M. (2003). *HCI Models, theories, and frameworks: Towards a multidisciplinary science*. San Francisco: Morgan Kaufmann.

Castellina, E., Corno, F. & Pellegrino, P. (2008). Integrated Speech and Gaze Control for Realistic Desktop Environments. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications (ETRA)*, 79-82.

Cato, J. (2001). *User-centered web design*. Great Britain: Addison-Wesley.

Chin, J.P., Diehl, V.A. & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI '88 Conference Proceedings: Human Factors in Computing Systems*, 213-218, New York: ACM/SIGCHI.

COGAIN. (2006). An affordable future for eye tracking in sight. Retrieved 29 February 2008 from http://www.cogain.org/media/ files/COGAIN-IST-Results.pdf.

Cohen, P.R., Johnston, M., McGee, D., Oviatt, S.L., Clow, J. & Smith, I. (1998). The efficiency of multimodal interaction: A case study. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 249-252.

Cohen, P.R., McGee, D.R. & Clow, J. (2000). The efficiency of multimodal interaction for a map-based task. In *Proceedings of the Applied Natural Language Processing Conference*, 331-338.

Corno, F., Farinetti, L. & Signorile, I. (2002). A cost-effective solution for eye-gaze assistive technology. In *Proceedings of the 2002 IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, 433-436.

Coutaz, J. & Caelen, J. (1991). A taxonomy for multimedia and multimodal user interfaces. In *Proceedings of the Second East-West HCI conference*, St Petersburg, Russia, 229-240.

Dai, L., Goldman R., Sears, A. & Lozier, J. (2004). Speech-based cursor control: A study of grid-based solutions. In *Proceedings of Assets '04*, Atlanta, Georgia, United States of Ameria, 94-101.

Daintith, J. & Wright, E. (Eds). (2008). *A Dictionary of Computing* (6th Ed.). New York: Oxford University Press.

De Luca, A., Weiss, R. & Drewes, H. (2007). Evaluation of eye-gaze interaction methods for security enhanced PIN-entry. In *Proceedings of OzCHI 2007*, Adelaide, Australia, 199-202.

Désiltes, A., Fox, D.C. & Norton, S. (2006). VoiceCode: An innovative speech interface for programming-by-voice. In *Proceedings of CHI 2006*, Montréal, Canada, 239-242.

Dickinson, A., Gregor, P. & Dickinson, L. (2003). SeeWord: Rethinking interfaces. Insights from word-processing software for dyslexic readers. In *Proceedings of INTERACT 2003*, Zurich, Switzerland, 615-622.

Dillon, A. (2001). The evaluation of software usability. In W. Karwowski (Ed.), *Encyclopedia of Human Factors and Ergonomics*, (pp. 1-6). London: Taylor and Francis.

Dirica, A.C. & Göktürk, M. (2009). Attentive Interfaces. In I. Maurtua (Ed.), *Human-Computer Interaction*, InTech.

Ditchburn, R.W. & Ginsborg, B.I. (1953). Involuntary eye movements during fixation. *Journal of Physiology*, 119, 1-17.

Dix, A., Finlay, J., Abowd, G. & Beale, R. (1993). *Human-computer interaction*. New Jersey: Prentice-Hall.

Douglas, S.A., Kirkpatrick, A.E. & MacKenzie, I.S. (1999). Testing pointing device performance and user assessment with the IS0 9241, Part 9 Standard. In *Proceedings of CHI '99*, Pennsylvania, United States of America, 215-222.

Dragon Naturally Speaking. (nd). History of speech and voice recognition and transcription software. Retrieved 13 February 2009 from http://www.nuance.com.

Drewes, H., De Luca, A. & Schmidt, A. (2007). Eye-gaze interaction for mobile phones. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*, 364-371.

Drewes, H. & Schmidt, A. (2007). Interacting with the computer using gaze gestures. In *Proceedings of Interact '07,* Rio De Janeiro, Brazil, 475-488.

Drewes, H & Schmidt, A. (2009). The MAGIC touch: Combining MAGIC-pointing with a touch-sensitive mouse. In *Human-Computer Interaction - INTERACT 2009. 12th IFIP TC 13 International Conference, Part II*, Uppsala, Sweden, 415-428.

Duchowski, A.T. (2002). A breadth-first survey of eye tracking applications. *Behavior Research Methods, Instruments, and Computers*, 34(4), 455-70.

Duchowski, A.T. (2007). *Eye tracking methodology: Theory and practice 2nd Edition*. London: Springer-Verlag.

Duchowski, A.T., Cournia, N. & Murphy, H. (2004). Gaze-contingent displays: A review. *CyberPsychology and Behavior*, 7(6), 621-634.

Dvorak, J.L. (2007). *Moving wearables into the mainstream: Taming the Borg*. United States: Springer.

Edwards, A.L. (1951). Balance Latin-square designs in psychological research. *The American Journal of Psychology*, 64(4), 598-603.

Eisenberg, D. (1992). Word Processing (History of). In *Encyclopedia of Library and Information Science*, 49, 268-278. New York: Dekker.

Ekman, I., Poikola, A., Mäkäräinen, M., Takala, T. & Hämäläinen, P. (2008). Voluntary pupil size change as control in eyes only interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 115-118.

Faulkner, C. (1998). *The essence of human-computer interaction*. Great Britain: Prentice-Hall.

Fejtová, M., Fejt, J., & Lhotská, L. (2004). Controlling a PC by eye movements: The MEMREC project. In *Proceedings of the 9th International Conference on Computers Helping People with Special Needs (ICCHP '04)*, 770-773. Berlin: Springer.

Felton, E.A., Lewis, N.L., Wills, S.A., Radwin, R.G. & Williams, J.C. (2007). Neural signal based control of the Dasher writing system. In *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering*, Hawaii, United States of America, 366-370.

Feng, J. & Sears, A. (2004). Are we speaking slower than we type? Exploring the gap between natural speech, typing and speech-based dictation. *Accessibility and Computing*, 79, 6-9.

Field, A. (1998). A bluffer's guide to ... sphericity. *The British Psychological Society: Mathematical, Statistical and Computing Section Newsletter*, 6, 13-22.

Fitch, W.T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Science*, 4(7), 258-267.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381-391.

Foley, J.D., Van Dam, A., Feiner, S.K. & Hughes, J.F. (1990). *Computer graphics: Principles and practice*. Reading, Massachusetts: Addison-Wesley.

Forlines, C., Schmidt-Nielsen, B., Raj, B., Wittenburg, K. & Wolf, P. (2005). A comparison between spoken queries and menu-based interfaces for in-car digital music selection. In *Proceedings of International Conference on Human-Computer Interaction (INTERACT 2005)*, 12-16.

Forsberg, M. (2003). Why is speech recognition difficult? Technical Report, Chalmers University of Technology.

Freedman, A. (1998). The Computer Glossary (8th Edition). United States of America: AMACOM.

Freudenthal, A., Keyson, D.V., DeKoven, E. & De Hoogh, M.P.A.J. (2001). Communicating extensive smart home functionality to users of all ages: the design of a mixed-initiative multimodal thermostat interface. In *OIKOS 2001 Workshop: Methodological Issues in the Design of Household Technologies*, Molslaboratoriet, Denmark, 34-39.

Fry, E.B., Kress, J.E. & Fountoukidis, D.L. (2003). *The reading teacher's book of lists*. United States of America: Center for Applied Research in Education.

Furnas, G.W. (1986). Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Boston, United States of America, 16-23.

Gajos, K.Z., Wobbrock, J.O. & Weld, D.S. (2008). Improving the Performance of Motor-Impaired Users with Automatically-Generated, Ability-Based Interfaces. In *Proceedings of CHI 2008*, Florence, Italy, 1257-1266.

Gips, J. & Olivieri, P. (1996). EagleEyes: An eye control system for persons with disabilities. In *Proceedings of the 11th International Conference on Technology and Persons with Disabilities*, Los Angeles, United States of America.

Girden, E. R. (1992). *ANOVA: Repeated measures*. Newbury Park, California: Sage.

Glenstrup, A.J. & Engell-Nielsen, T. (1995). Eye controlled media: Present and future state. Bachelor's Degree Thesis, University of Copenhagen.

Gorniak, P. & Roy, D. (2003). Augmenting user interfaces with adaptive speech commands. In *Proceedings of ICMI '03*, Vancouver, Canada, 176-179.

Gregory, R.L. (1966). *The eye and the brain: The psychology of seeing*. London: World University Library.

Griffin, Z. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82, B1–B14.

Griffin, Z. M. & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11, 274–279.

Haigh, T. (2006). Remembering the office of the future: The origins of word processing and office automation. *IEEE Annals of the History of Computing*, 28(4), 6-31.

Haller, R., Mutschler, H. & Voss M. (1984). Comparison of input devices for correction of typing errors in office systems. *Proceedings of INTERACT '84, First IFIP Conference on Human-Computer Interaction*, London, United Kingdom, 177-182.

Hansen, J.P., Hansen, D.W., & Johansen, A.S. (2001). Bringing gaze-based interaction back to basics. In C. Stephanidis (Ed.) *Universal Access in HCI (UAHCI): Towards an Information Society for All - Proceedings of the 9th International Conference on Human-Computer Interaction (HCII'01)*, 325-328. Mahwah, NJ: Lawrence Erlbaum Associates.

Hansen, D.W. & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 23(3), 478-500.

Hansen, J.P., Johansen, A.S., Hansen, D.W., Itoh, K. & Mashino, S. (2003). Command without a click: Dwell time typing by mouse and gaze selections. In *Proceedings of Human-Computer Interaction (INTERACT '03)*, Zurich, Switzerland, 121-128.

Haro, A., Essa, I. & Flickner, M. (2000). A non-invasive computer vision system for reliable eye tracking. In *Proceedings of CHI '00*, The Hague, Netherlands, 167-168.

Harper, B.D. & Norman, K.L. (1993). Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5. In *Proceedings of the 1st annual Mid-Atlantic Human Factors Conference*, Virginia Beach, Virginia, United States of America, 224-228.

Hassenzahl, M. & Tractinsky, N. (2006). User experience – a research agenda. *Behaviour and Information Technology*, 25(2), 91-97.

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*, 1-7.

Hauptmann, A.G. (1989). Speech and Gestures for Graphic Image Manipulation. In *Proceedings of the International Conference on Human-Computer Interaction*, 241-245.

He, T. & Kaufman, A.E. (1997). Virtual input devices for 3D systems. In *Proceedings of IEEE Visualization '93*, San Jose, California, 142-148.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. & Van de Weijer, J. (In press). *Eye tracking: A comprehensive guide to methods and measures*. London: Oxford University Press.

Hornof, A., Cavender, A. & Hoselton, R. (2004). EyeDraw: A system for drawing pictures with eye movements. In *Proceedings of ASSETS '04*, Atlanta, Georgia, United States of America, 86-93.

Huckauf, A. & Urbina, M.H. (2008). Gazing with pEyes: Towards a universal Input. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 51-54.

Hwang, F., Keates, S., Langdon, P. & Clarkson, J. (2004). Mouse movements of motion-impaired users: A submovement analysis. In *Proceedings of ASSETS '04*, Atlanta, Georgia, United States of America, 102-109.

Hyrskykari, A. (1997). Gaze Control as an Input Device. In *Proceedings of ACHCI '97*, University of Tampere, 22-27.

Hyrskykari, A., Majaranta, P. & Räihä, K-J. (2003). Proactive response to eye movements. In *Proceedings of INTERACT '03*, Zurich, Switzerland, 129-136.

ISO9241. (1998). ISO 9241-11: *Guidance on usability*. International Organization for Standardization.

ISO. (2000). *ISO 9241-9: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 9: Requirements for non-keyboard input devices*. International Organization for Standardization.

ISO9241-210:2010. (2010). Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. International Organization for Standardization.

Isokoski, P. (2000). Text input methods for eye trackers using off-screen targets. In *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications (ETRA)*, Palm Beach Gardens, Florida, United States of America, 15-21.

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, United Kingdom, 109-116.

Jacob, R.J.K. (1991). The use of eye movements in human-computer interaction techniques: What you look at is what you get, *ACM Transactions on Information Systems*, 9(2), 152-169.

Jacob, R.J.K. (1993a). Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. In H.R. Hartson and D. Hix (Eds), *Advances in Human-Computer Interaction*, 4, 151-190. Norwood, New Jersey: Ablex Publishing.

Jacob, R.J.K. (1993b). What you look at is what you get: Using eye movements as computer input. In *Proceedings of Virtual Reality Systems '93 conference and exhibition*, New York, New York, United States of America, 164-166.

Jacob, R.J.K. (1995a). Eye tracking in advanced interface design. In W. Barfield & T.A. Furness (Eds.), *Virtual Environments and Advanced Interface Design* (pp. 258-288). New York: Oxford University Press.

Jacob, R.J.K. (1995b) Natural Dialogue in Modes other than Natural Language. In R.J. Beun, M. Baker & M. Reiner (Eds), *Dialogue and Instruction* (pp. 289-301). Berlin: Springer-Verlag.

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 & H. Deubel (Eds) *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573-605). Amsterdam: Elsevier Science.

Jaimes, A. & Sebe, N. (2005). Multimodal human computer interaction: A survey. *IEEE workshop on human computer interaction*, Las Vegas, Nevada, United States of America, 15-21.

Jönsson, E. (2005). If looks could kill – An evaluation of eye tracking in computer games. Master's Thesis, Royal Institute of Technology, Stockholm, Sweden.

Jurafsky, J.H.M.D. (2000). *Speech and language processing: An introduction to Natural Language Processing, Computational Linguistics and Speech recognition.* New Jersey: Prentice Hall.

Just, M.A. & Carpenter, P.A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8, 441-480.

Kammerer, Y., Scheiter, K. & Beinhauer, W. (2008). Looking my way through the menu: The impact of menu design and multimodal input on gaze-based menu selection. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 213-220.

Karimullah, A.S. & Sears, A. (2002). Speech-based cursor control. In *Proceedings of ASSETS '02*, Edinburgh, Scotland, 178-185.

Karl, L., Pettey, M. & Shneiderman, B. (1993). Speech-activated versus mouse-activated commands for word processing applications: An empirical evaluation. *International Journal of Man-Machine Studies*, 39, 667-687.

Karpov, A., Carbini, S., Ronzhin, A. & Viallet, J.E. (2008). Comparison of two different similar speech and gestures multimodal interfaces. In *Proceedings 16th European Signal Processing Conference,* Lausanne, Switzerland.

Kaukènas, J., Navickas, G. &Telksnys, L. (2006). Human-computer audiovisual interface. *Information technology and control*, 35(2), 87-93.

Kaur, M., Tremaine, M., Huang, N., Wilder, J., Gacovski, Z., Flippo, F. & Mantravadi, S. (2003). Where is "it"? Event synchronization in gaze-speech input systems. In *Proceedings of ICIM '03*, Vancouver, Canada, 151 - 158.

Keates, S., Hwang, F., Langdon, P., Clarkson, P.J. & Robinson, P. (2002). Cursor movements for motion-impaired computer users. In *Proceedings of ASSETS '02*, Edinburgh, Scotland, 135-142.

Keates, S. & Trewin, S. (2005). Effect of age and Parkinson's Disease on cursor positioning using a mouse. In *Proceedings of ASSETS '05*, Baltimore, Maryland, United States of America, 68-75.

Klarlund, N. (2003). Editing by Voice and the Role of Sequential Symbol Systems for Improved Human-to-Computer Information Rates. In *Proceedings of ICASSP*, Hong Kong, 553-556.

Klarlund, N. & Riley, M. (2003). Word n-grams for cluster keyboards. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 51-58.

Kumar, M. (2006). Reducing the cost of eye tracking systems. Technical Report CSTR 2006-08, Stanford HCI Group.

Kumar, M. (2007), GUIDe Saccade Detection and Smoothing Algorithm. Technical Report CSTR 2007-03, Stanford HCI Group.

Kumar, M., Klinger, J., Puranik, R., Winograd, T. & Paepcke, A. (2008). Improving the accuracy of gaze input for interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 65-68.

Kumar, M., Paecke, A. & Winograd, T. (2007). EyePoint: Practical pointing and selection using gaze and keyboard. In *Proceedings of CHI 2007*, San Jose, California, United States of America, 421-430.

Kumar, M. & Winograd, T. (2007). GUIDe: Gaze-enhanced UI Design. In *Proceedings of CHI 2007,* San Jose, California, United States of America, 1977-1982.

Land, M.F. & Tatler, B.W. (2009). *Looking and acting: Vision and eye movements in natural behaviour*. United States of America: Oxford University Press.

Laqua, A., Bandara, S.U. & Sasse, M.A. (2007). GazeSpace: Eye gaze controlled content spaces. In *Proceedings of HCI 2007*, Beijing, China, 55-58.

Latoschik, M.E., Fröhlich, M., Jung, B. & Wachsmuth. I. (1998) Utilize speech and gestures to realize natural interaction in a virtual environment. In *IECON'98 – Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society*, 2028–2033.

Leggett, J. & Williams, G. (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies*, 21(6), 493-520.

Levenshtein, V.I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk*, 163, 845-848.

Li, D., Winfield, D. & Parkhurst, D.J. (2005). Starbursts: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Proceedings of the IEEE Vision for Human-Computer Interaction Workshop at CVPR*, Beijing, China, 1-8.

Liu, Y., Chai, J, Y. & Jin, R. (2007). Automated vocabulary acquisition and interpretation in multimodal conversational systems. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Liu, X., Crump, M.J.C. & Logan, G.D. (2010). Do you know where your fingers have been? Explicit knowledge of the spatial layout of the keyboard in skilled typists. *Memory and Cognition*28(4), 474-484.

Logan, G.D. & Crump, M.J.C. (2009). The left hand doesn't know what the right hand is doing: The disruptive effects of attention to the hands in skilled typewriting. *Psychological Science*, 20(10), 1296-1300.

MacKenzie, I.S. (2002). A note on calculating text entry speed. Retrieved 14 June 2010 from http://www.yorku.ca/mack/RN-TextEntrySpeed.html.

MacKenzie, I.S., Kauppinen, T. & Silfverberg, M. (2001). Accuracy measures for evaluating computer pointing devices. In *Proceedings of SIGCHI '01*, Seattle, Washington, United States of America, 9-16.

MacKenzie, I.S. & Soukoreff, R.W. (2002). A character-level error analysis technique for evaluating text entry methods. In *Proceedings of NordiCHI 2002*, Aarhus, Denmark, 243-246.

MacKenzie, I.S. & Soukoreff, R.W. (2003). Phrase sets for evaluating text entry techniques. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems – CHI 2003*, Fort Lauderdale, Florida, United States of America, 754-755.

Maglio, P.P., Matlock, T., Campbell, C.S., Zhai, S. & Smith, B.A. (2000). Gaze and speech in attentive user interfaces. In *Proceedings of the Third International Conference on Advances in Multimodal Interfaces*, Vancouver, Canada, 1-7.

Majaranta, P. (2009). Text entry by eye gaze. Dissertations in Interactive Technology, number 11, University of Tampere.

Majaranta, P., Ahola, U-K. & Špakov, O. (2009). Fast gaze typing with an adjustable dwell time. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, Boston, Massachusetts, United States of America, 357-360.

Majaranta, P., MacKenzie, I.S., Aula, A. & Räihä, K-J. (2006). Effects of dwell time on eye typing and accuracy. *Universal Access in the Informational Society*, 5(2), 199-208.

Majaranta, P. & Räihä, K.-J. (2007). Text entry by gaze: Utilizing eyetracking. In I. S. MacKenzie and K. Tanaka-Ishii (Eds.) *Text Entry Systems: Mobility, Accessibility, Universality*, 175-187. San Francisco: Morgan Kaufmann.

Man, D.W.K. & Wong, M-S, L. (2007). Evaluation of computer-access solutions for students with quadriplegic athetoid cerebral palsy. *American Journal of Occupational Therapy*, 61, 355-364.

Martinez-Conde, S. & Macknick, S.L. (2008). Fixational eye movements across vertebrates: Comparative dynamics, physiology, and perception. *Journal of Vision*, 8(14), 1-16.

Martinez-Conde, S., Macknik, S.L. & Hubel, D.H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3), 229-240.

Maxwell, S. E., & Delany, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associated, Publishers.

Microsoft (nd). Microsoft Speech API. Retrieved 4 May 2010 from http://msdn.microsoft.com/en-us/library/ms723627(VS.85).aspx.

Minke, A. (1997). Conducting repeated measures analyses: Experimental design considerations. In *Proceedings of the Annual Meeting of the Southwest Educational Research Association*, Austin, Texas, 23-25.

Miniotas, D. (2000). Application of Fitts' Law to eye gaze interaction. In *Proceedings of CHI 2000*, The Hague, Netherlands, 339-340.

Miniotas, D. & Špakov, O. (2004). Target expansion as a means to facilitate eye-based selection. *Elektronika Ir Elektrotechnika*, 3(25), 13-17.

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

Miniotas, D., Špakov, O. & MacKenzie, I.S. (2004). Eye gaze interaction with expanding targets. In *Extended abstracts of the ACM Conference of Human Factors in Computing Systems – CHI 2004*, Vienna, Austria, 1255-1258.

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)*, 67-72.

Morimoto, C.H. & Amir, A. Context switching for fast key selection in text entry applications. In *Proceedings of the 2010 Symposium on Eye Tracking Research and Applications (ETRA)*, 271-274.

Morrison, D.L., Green, T.R.G., Shaw, A.C. & Payne, S.J. (1984). Speech-controlled text-editing: effects of input modality and of command structure. *International Journal of Man-Machine Studies,* 21, 49-63.

Motulsky, H. (1995). *Intuitive biostatics: Choosing a statistical test*. United States of America: Oxford University Press.

Murata, A. (2006). Eye-Gaze Input Versus Mouse: Cursor Control as a Function of Age. *International Journal of Human-Computer Interaction*, 21(1), 1-14.

Natapov, D., Castellucci, S.J. & MacKenzie, I.S. (2009). ISO 9241-9 evaluation of video game controllers. In *Proceedings of Graphics Interface Conference*, Kelowna, British Columbia, Canada, 223-230.

Nelson, D. L. (1986) User acceptance of voice recognition in a product inspection environment. *The Official Proceedings of Speech Tech '86: Voice Input / Output Applications Show and Conference*, p. 62.

Nielsen, J. (1993). Noncommand user interfaces. Retrieved 11 March 2011 from http://www.useit.com/papers/ noncommand.html.

Nielsen, J. (2000). Why you only need to test with 5 users. Alertbox, 19 March, 2000. Retrieved 7 June 2010 from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2001a). Usability metrics. Alertbox, January, 2001. Retrieved 7 June 2010 from http://www.useit.com/alertbox/ 20010121.html.

Nielsen, J. (2001b). Success rate: The simplest usability metric. Alertbox, February, 2001. Retrieved 7 June 2010 from http://www.useit.com/alertbox/20010218.html.

Nielsen, J. (2006). Quantitative Studies: How Many Users to Test? Alertbox, 26 June, 2006. Retrieved 7 June 2010 from http://www.useit.com/alertbox/quantitative_testing.html.

Nijholt, A. & Tan, D. (2008). Brain-computing interfacing for intelligent systems. *IEEE Intelligent Systems*, 23(3), 72-79.

Nimon, K. & Williams, C. (2009). Evaluating performance improvement through repeated measures: A primer for educators considering univariate and multivariate designs. *Research in Higher Education Journal*, 2, 28-48.

Nusbaum, H.C., DeGroot, J & Lee, L. (1995). Using speech recognition systems: Issues in cognitive engineering. In A.Sydral, R. Bennett and S.Greenspan (Eds), *Applied Speech Technology* (pp. 127-194). Boca Raton, Florida: CRC Press.

Nye, J. M. (1982). Human factors analysis of speech recognition systems. *Speech Technology*, 1(2), 50-57.

Olivier, M. (2004). *Information technology research: A practical guide for Computer Science and Informatics* (2nd Edition). Pretoria: Van Schaik.

O'Shaughnessy, D. (1995). Speech Technology. In A.Sydral, R. Bennett and S.Greenspan (Eds), *Applied Speech Technology* (pp. 47-98). Boca Raton, Florida: CRC Press.

Oviatt, S. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the ACM SIGCHI 99*, Pittsburgh, Pennsylvania, United States of America, 576-583.

Oviatt, S. & Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(2), 45-53.

Oviatt, S., Cohen, P., Wu, L.Z., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. & Ferro, D. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human-Computer Interaction*, 15 (4), 263-322.

Oviatt, S., MacEachern, M. & Levow, G. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Community*, 24(2), 87-110.

Oxford Dictionaries. (2011). Oxford Dictionaries. England: Oxford University Press. Last accessed 14 July 2011 at http://oxforddictionaries.com.

Oyekoya, O.K. & Stentiford, F.W.M. (2006). Eye tracking – A new interface for visual exploration. *BT Technology Journal*, 24(3), 57-66.

Paluch, K. (2009). What is user experience design. Last accessed 23 July 2011 at http://www.montparnas.com/articles/what-is-user-experience-deisgn/print.

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

Poock, G. K. (1982). Voice recognition boosts command terminal throughput. *Speech Technology*, 1(2), 36-39.

Porta, M. & Turina, M. (2008). Eye-S: a full-screen input modality for pure eye-based communication. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 27-34.

Poulton, E.C. & Freeman, P.R. (1966). Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin*, 66, 1-8.

Prasov, Z., Chai, J.Y. & Jeong, H. (2007). Eye gaze for attention prediction in multimodal human-machine conversation. In *Proceedings of AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants*.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. & Carey, T. (1994). *Human-computer interaction*. England: Addison-Wesley.

Quadriplegic Association of South Africa. (nd). Available from http://quad.stormnet.co.za/index.htm.

Qvarfordt, P., Beymer, D. & Zhai, S. (2005). RealTourist – A Study of Augmenting Human-Human and Human-Computer Dialogue with Eye-Gaze Overlay. In *Proceedings of INTERACT 2005*, Rome, Italy, 767-780.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

Read, J. (2005). On the application of text input metrics to handwritten text input. Text Input Workshop, Dagstuhl., Germany.

Read, J., MacFarlane, S. & Casey, C. (2001). Measuring the usability of text input methods for children. In *Proceedings of Human-Computer Interaction (HCI) 2001*, New Orleans, United States of America, 559-572.

Reese, H.W. (1997). Counterbalancing and other uses of repeated-measures Latin-square designs: Analyses and interpretations. *Journal of Experimental Child Psychology*, 64, 137-158.

Rosson, M.B. (1984). Effects of experience on learning, using, and evaluating a text editor. *Human Factors*, 26(4), 463-475.

Rubinoff, R. (nd). How to quantify the user experience. Last accessed 24 July 2011 at http://www.sitepoint.com/quantify-user-experience.

Russel, S.J. & Norvig, P. (2009). *Artificial Intelligence: A modern approach* (3rd Edition). Prentice Hall.

Schmandt, C., Ackerman, M.S. & Hindus, D. (1990). Augmenting a window system with speech input. *Computer*, 23(8), 50-56.

Schnell, T. (2000). Applying eye tracking as an alternative approach for activation of controls and functions in aircraft. In *Proceedings of Digital Avionics Systems Conferences*, Washington, DC, United States of America, 19(2), 5A5/1-5A5/9.

Scholtz, J. (2004). Usability evaluation. Publication #545, National Institute of Standards and Technology.

Sears, A., Karat, C-M., Oseitutu, K., Karimullah, A. & Feng, J. (2001). Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society*, 1(4), 4-15.

Sears, A., Lin, M. & Karimullah, A.S. (2002). Speech-based cursor control: Understanding the effects of target size, cursor speed, and command selection. *Universal Access in the Information Society*, 2(1), 30-43.

Shackel, B. (1991). Usability – Context, framework, design and evaluation. In B. Shackel & S. Richardson (Eds), *Human factors for informatics usability* (pp.21-38). Cambridge: Cambridge University Press.

Shell, J.S., Bradbury, J.S., Knowles, C.B., Dickie, C. & Vertegaal, R. (2003a). eyeCook: A gaze and speech enabled attentive cookbook. In *Video Proceedings of Ubiquitous Computing* (Ubicomp), Seattle, Washington, United States of America.

Shell, J.S., Vertegaal, R., Mamuji, A., Pham, T., Sohn, C. & Skaburskis, A. (2003b). EyePliances and EyeReason: Using Attention to Drive Interactions with Ubiquitous Appliances. In *Extended Abstracts of UIST*, Vancouver, Canada.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd Edition). Massachusetts: Addison-Wesley.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, 43(9), 63-65.

Sibert, L.E. & Jacob, R.J.K. (2000). Evaluation of Eye Gaze Interaction. In *Proceedings of ACM CHI 2000: Human Factors in Computing Systems Conference*, The Hague, Netherlands, 281-288.

Simon, C. (2002). Com objects, C# and the Microsoft speech API. Dr Dobb's Journal, September 2002.

Smith, B.A., Ho, J., Ark, W. & Zhai, S. (2000). Hand eye coordination patterns in target selection. In *Proceedings of the Eye Tracking Research and Application Symposium (ETRA)*, Palm Beach Gardens, Florida, United States of America, 117-122.

Soukoreff, R. W. & MacKenzie, I. S. (2001). Measuring errors in text entry tasks: An application of the Levenshtein string distance statistic. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI '01)*, Seattle, Washington, United States of America, 319-320.

Soukoreff, R.W. & MacKenzie, I.S. (2004). Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' Law research in HCI. *International Journal of Human-Computer Studies*, 61, 751-789.

Špakov, O. (2005) EyeChess: The tutoring game with visual attentive interface. In *Proceedings of Alternative Access: Feelings & Games*, University of Tampere, Finland, 81-86.

Špakov, O. & Majaranta, P. (2008). Scrollable keyboards for eye typing. In *Proceedings of the 4th Conference on Communication by Gaze Interaction (COGAIN)*, Prague, Czech Republic, 63-66.

Špakov, O. & Miniotas, D. (2003). An algorithm for adjustable dwell time in eye typing systems. *Information Technology and Control*, 2(31), 49-52.

Špakov, O. & Miniotas, D. (2005). Gaze-based selection of standard-size menu items. In *Proceedings of 7th International Conference on Multimodal Interfaces (ICMI)*, Trento, Italy, 124-128.

Stampe, D.M. & Reingold, E.M. (1995). Selection by looking: A novel computer interface and its application to psychological research. In J.M. Findlay, R. Walker & R.W. Kentridge (Eds), *Eye movement research: Mechanisms, processes and applications* (pp. 467-478). Amsterdam: Elsevier Science Publishers.

StatSoft, Inc. (2010). Electronic Statistics Textbook. Last accessed 21 May 2011 at http://www.statsoft.com/ textbook/.

Stiefelhagen, R. & Yang, J. (1997). Gaze Tracking for Multimodal Human-Computer Interaction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, 140-147.

Su, M-C., Su, S-Y. & Chen, G-D. (2005). A low-cost vision-based human-computer interface for people with severe disabilities. *Biomedical Engineering: Applications, Basis and Communications*, 17(6), 284-292.

Suhm, B. (2008). IVR usability engineering using guidelines and analyses of end-to-end calls. In D. Gardner-Bonneau and H.E. Blanchard (Eds), *Human factors and voice interactive systems* (pp. 1-42). New York, NY: Springer Science+Business Media.

Tan, Y.K., Sherkat, N. & Allen, T. (2003a). Eye gaze and speech for data entry: A comparison of different data entry methods. In *Proceedings of the International Conference on Multimedia and Expo*, Baltimore, Maryland, United States of America, 41-44.

Tan, Y.K., Sherkat, N. & Allen, T. (2003b). Error recovery in a blended style eye gaze and speech interface. In *Proceedings of ICMI '03*, Vancouver, Canada, 196-202.

Tanaka, K. (1999). A robust selection system using realtime multi-modal user-agent interactions. In *Proceedings of IUI'99*, 105-108.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, 268, 1632-1634.

Ten Kate, J.H., Frietman, E.E.E., Willems, W., Ter Haar Romeny, B.M., & Tenkink, E. (1979). Eye-switch controlled communication aids. In *Proceedings of the 12th International Conference on Medical and Biological Engineering*, Jerusalem, Israel, 19-20.

Thomas, J.C., Basson, S. & Gardner-Bonneau, D. (2008). Accessibility and speech technology: Advancing toward universal access. In D. Gardner-Bonneau and H.E. Blanchard (Eds), *Human factors and voice interactive systems* (pp. 417-442). New York, NY: Springer Science+Business Media.

Tobii. (2011). Tobii unveils the world's first eye-controlled laptop. Retrieved 14 March 2011 from http://www.tobii.com/en/ eye-tracking-integration/global/news-and-events/press-releases/tobii-unveils-the-worlds-first-eye-controlled-laptop/.

Tse, E., Greenberg, S., Shen, C. & Forlines, C. (2006). Multimodal multiplayer tabletop gaming. In *Proceedings of PerGames*, Dublin, Ireland, 139-148.

Tuisku, O., Majaranta, P., Isokoski, P. & Räihä, K-J. (2008). In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 19-26.

Tullis, T. & Albert, B. (2008). *Measuring the user experience*. United States of America: Morgan Kaufmann Publishers.

Tullis, T.S. & Stetson, J.N. (2004). A comparison of questionnaire for assessing website usability. In *Connecting Communities: UPA, Network In Our Community*, Minneapolis, Minnesota, United States of America.

Turk, M. (2001). Perceptual user interfaces. In R. Earnshaw, R. Guedj, A. van Dam & J. Vince (Eds), *Frontiers of human-centred computing, online communities and virtual environments* (pp. 39-51). London: Springer-Verlag.

Turk, M. & Kölsch, M. (2004). Perceptual Interfaces. In G. Medioni and S.B. Kang (Eds), *Emerging Topics in Computer Vision* (pp. 358-403). Prentice Hall.

Unger, R. & Chandler, C. (2009). A project guide to UX design: For user experience designers in the field or in the making. United States of America: New Riders Press.

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: Springer-Verlag.

Velichkovsky, B. M., Sprenger, A., & Pomplun, M. (1997). Auf dem Weg zur Blickmaus: Die Beeinflussung der Fixationsdauer durch kognitive und kommunikative Aufgaben. In R. Liskowsky, B. M. Velichkowsky, & W. Wünschmann (Eds), *Software-Ergonomie* (pp. 317-327).

Vergo, J. (1998). A statistical approach to multimodal natural language interaction. In *Proceedings of the AAAI'98 Workshop on Representations for Multimodal Human-Computer Interaction*, Madison, Wisconsin, United States of America, 81-85.

Vertanen, K. & MacKay, D.J.C. (2010). Speech Dasher: Fast writing using speech and gaze. In *Proceedings of CHI 2010*, Atlanta, Georgia, United States of America, 595-598.

Wachs, J.P, Kölsch, M., Stern, H. & Edan, Y. (2011). Vision-based hand-gesture applications. *Communications of the ACM*, 54(2), 60-71.

Ward, D.J., Blackwell, A.F. & MacKay, D.J.C. (2000). Dasher – a data entry interface using continuous gestures and language models. In *Proceedings of UIST 2000: The 13th Annual ACM Symposium on User Interface Software and Technology*, San Diego, California, United States of America, 129-137.

Ware, C. & Mikaelian, H.H. (1987). An evaluation of an eye tracker as a device for computer input. In *Proceedings of CHI*, 183-188.

Whitley, E. & Ball, J. (2002). Statistics review 6: Nonparametric methods. *Critical Care*, 6, 509-513.

Wixon, D. & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M.G. Herlander (Ed.), *Handbook of human-computer interaction* (pp. 653-688). Holland: Elsevier.

Wobbrock, J.O. (2007). Measures of text entry performance. In I.S. MacKenzie & K. Tanaka-Ishii (Eds), *Text entry systems: Mobility, Accessibility, Universability* (pp. 47-74). San Francisco: Morgan Kaufmann.

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 Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 11-18.

Word-english. (2003). Retrieved 14 July 2011 from http://www.world-english.org/english500.htm.

wordiQ. (2010). Retrieved 16 August 2010 from http://www.wordiq.com.

www.msu.edu. Retrieved 23 February 2011 from www.msu.edu.

www.sci-info-pages.com. (nd). Retrieved 12 November 2010 from http://www.sci-info-pages.com.

www.tobii.com. Tobii Eye-tracking Technology. Retrieved 2 February 2011 from www.tobii.com.

Yale Medical Group. (nd). Accessed 23 January 2011 at www.yalemedicalgroup.org.

Yankelovich, N. (2008). Using natural dialogs as the basis for speech interface design. In D. Gardner-Bonneau & H.E. Blanchard (Eds), *Human factors and voice interactive systems* (pp. 417-442). New York, NY: Springer Science+Business Media.

Zhai, S., Morimoto, C. & Ihde, S. (1999). Manual And Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of CHI '99: ACM Conference on Human Factors in Computing Systems*, Pittsburgh, Pennsylvania, United States of America, 246-253.

Zhang, Q., Imamiya, A., Go, K. & Mao, X. (2004). Resolving ambiguities of a gaze and speech interface. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, San Antonio, Texas, United States of America, 85-92.

Zhang, X. & MacKenzie, I.S. (2007). Evaluating eye tracking with ISO 9241 – Part 9. In J. Jacko (Ed.), *Human Computer Interaction*, 779-788.

BIBLIOGRAPHY

Andrews, S. (1991). Improving touchscreen keyboards: Design issues and a comparison with other devices. *Interacting with computers*, 3(3), 253-269.

Barea, R., Boquete, L., Bergasa, L.M., López, E. & Mazo, M. (2003). Electro-oculography guidance of a wheelchair using eye movements codification. *The International Journal of Robotics Research*, 22, 641-652.

Berglund, A. & Qvarfordt, P. (2003). Error resolution strategies for interactive television speech interfaces. In *Proceedings of Interact '03*, Zurich, Switzerland, 105-112.

Cohen, M.H. Giangola, J.P. & Balogh, J. (2004). *Voice User Interface Design*. Boston: Addison-Wesley.

Coleman, J. (2005). *Introducing speech and language processing*. Cambridge: University Press.

Czerwinski, M., Smith, G., Regan, T., Meyers, B., Robertson, G. & Starkweather, G. (2003). Toward characterizing the productivity benefits of very large displays. In *Proceedings of Interact '03*, Zurich, Switzerland, 9-16.

Czerwinski, M., Robertson, G., Meyers, B., Smith, G., Robbins, D. & Tan, D. (2007). Large display research overview. In *Proceedings of CHI '06*, Quebec, Canada, 69-74.

Deng, L. & Huang, X. (2004). Challenges in adopting speech recognition. *Communications of the ACM*, 47(1), 69-75.

Drewes, H. (2006). Gaze tracking in HCI. In *Proceedings of the First International Colloquium on Pervasive Computing*.

Farid, M., Murtagh, F. & Starck, J.L. (2002). Computer Display Control and Interaction Using Eye-Gaze. *Journal of the Society for Information Display*, 10(3), 289-293.

Hatfield, F., Jenkins, E.A. & Jennings, M.W. (1996). Principles and Guidelines for the Design of Eye/Voice Interaction Dialogs. In *Proceedings of the Third Annual Symposium on Human Interaction with Complex Systems*, Dayton, Ohio, United States of America, 10-19.

Holman, D. (2007). GazeTop: Interaction techniques for gaze-aware tabletops. In *Proceedings of CHI 2007*, San Jose, California, United States of America, 1657-1660.

Hyrskykari, A., Majaranta, P. & Räihä, K-J. (2005). From gaze control to attentive interfaces. In *Proceedings of Human-Computer Interaction International (HCII)*, Las Vegas, Nevada, United States of America.

MacKenzie, I.S. (n.d.). ISO Testing of Computer Pointing Devices. Retrieved 1 February 2010 from http://www.yorku.ca/mack/.

MacKenzie, I.S. (2003). Motor behaviour models for human-computer interaction. In JM Carroll (Ed.) *Toward a multidisciplinary science of human-computer interaction* (pp. 27-54). San Francisco: Morgan Kaufmann.

Milekic, S. (2003). The more you look the more you get: Intention-based interface using gaze-tracking. Museums and the Web '03.

Miniotas, D., Špakov, O., Tugoy, I. & MacKenzie, I.S. (2005). Extending the limits for gaze pointing through the use of speech. *Information and Control*, 34, 225-230.

Modlitba, P. (2004). Audiovisual attentive user interfaces – Attending to the needs and actions of the user. T-121.900, Seminar on user interfaces and usability.

Nielsen, J. (1996). International usability testing. Retrieved 7 June 2010 from http://www.useit.com/papers/international_usetest.html.

Optimoz Project. (nd). Mouse gestures. Retrieved from http://optimoz.mozdev.org/gestures/.

Oulasvirta, A. & Salovaara, A. (2004). A cognitive meta-analysis of design approaches to interruptions in intelligent environments. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, Vienna, Austria, 1155-1158.

Pernice, K. and Nielsen, J. (2009). Eyetracking Methodology: How to Conduct and Evaluate Usability Studies Using Eyetracking. Alertbox, August, 2009. Retrieved 7 June 2010 from http://www.useit.com/eyetracking/methodology/eyetracking-methodology.pdf.

Rauterberg, M. (nd). The complete history of HCI. Retrieved 9 February 2009 from http://www.idemployee.id.tue.nl/g.w.m.rauterberg/ presentations/HCI-history_files/ frame.htm.

Roberts, T.L. & Moran, T.P. (1983). The evaluation of text editors: Methodology and empirical results. *Communications of the ACM*, 26(4), 265-283.

Rosson, M.B. (1984). Characterizing freeform editing behavior. IBM Research Report RC 10550, IBM T. J. Watson Research Center, Yorktown Heights, New York.

Sasangohar, F., MacKenzie, I. S., & Scott, S. D. (2009). Evaluation of mouse and touch input for a tabletop display using Fitts' reciprocal tapping task. In *Proceedings of the 53rd Annual Meeting of the Human Factors and Ergonomics Society – HFES 2009*, San Antonio, Texas, United States of America, 839-843.

Sears, A., Feng, J. & Oseitutu, K. (2003). Hands-free, speech-based navigation during dictation: Difficulties, consequences and solutions. *Human-Computer Interaction*, 18, 229-257.

Sears, A., Karat, C-M., Oseitutu, K., Karimullah, A. & Feng, J. (2001). Productivity, satisfaction, and interaction strategies of individual with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society*, 1(4), 4-15.

Selker, T. (2004). Visual Attentive Interfaces. *BT Technology Journal*, 22(4), 146-150.

Soukoreff, R. W., & MacKenzie, I. S. (2003). Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '03)*, Fort Lauderdale, Florida, United States of America, 113-120.

Sullivan, P. (1989). Human-computer interaction perspectives on word-processing issues. *Computers and Composition*, 6(3), 11-33.

Tuisku, O., Majaranta, P., Isokoski, P. & Räihä, K-J. (2008). In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Savannah, Georgia, United States of America, 19-26.

Vertegaal, R. (2002). Designing Attentive Interfaces. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, New Orleans, Louisiana, United States of America, 23-30.

Vertegaal, R. (2003). Attentive user interfaces: Introduction. *Communications of the ACM*, 46(3), 31-22.

Zajicek, A. & Morrissey, W. (2001). In A. Blandford, J. Vanderdonckt and P. Gray (Eds.), *People and Computers XV – Interaction without frontiers*, (pp. 503-558). London, Great Britain: Springer.

APPENDIX A

FEASIBILITY STUDY PRE–TEST QUESTIONNAIRE

University of the Free State

Department of Computer Science and Informatics

Multimodal word processors

Pre-Test Questionnaire

1. Name and surname: _____
2. Age: _____
3. Highest qualification: _____
4. Which fields do you specialise in?

_____

_____

5. Do you understand what is meant by *usability*?
   Yes No
6. Do you understand what is meant by *multimodal interfaces*?
   Yes No
7. Have you ever used speech recognition as a dictation tool?
   Yes No
8. Have you ever used speech recognition to issue commands to a computer programme?
   Yes No
9. Have you ever used eye tracking as a means to interact with an application?
   Yes No
10. When working in a word processor do you make use of shortcut keys?
    Yes No
11. Are you a touch typist?
    Yes No

APPENDIX B

EXPERT REVIEW TASK LIST

Tasks

1. Familiarise yourself with the new Word environment which makes use of the eye tracking, speech recognition and onscreen keyboard.
    a. Make sure that you are comfortable using your eye gaze and the verbal command “Go” to write a character to the document.
    b. Insert a few words of your choice into the document and try and increase your speed to a speed that is comparable with your normal typing speed.
    c. Change some of the options for interaction techniques, such as the length of the dwell time, using blinking and the enter key as well as the voice commands.
    d. Change other options such as the shape of your gaze indicator.
    e. Use the verbal commands in the table below to navigate through your document and make changes. Use them in combination to determine how they work together.

APPENDIX C

FEASIBILITY STUDY POST–TEST QUESTIONNAIRE

Post-Test questionnaire

1. When you first encountered the system, were you sceptical as to its practicality or were you excited?
   Sceptical Excited

2. Please explain the reason for your answer in question 1.

3. Did your interaction with the system change your mind?
   Yes No

4. Please explain the reason for your answer in question 3.

5. As a person involved in the IT field, do you think there is a need for multimodal interfaces?
   Yes No

6. Do you think that the combination presented to you today is a viable option for a multimodal interface?
   Yes No

7. Please explain the reason for your answer in question 6.

8. As an initial impression, did you feel excited about the opportunities presented by the system?
   Yes No

9. In terms of a more long term application, do you think the combination presented to you would be a solution to multimodal interfacing?
   Yes No

10. From the point of view of a mainstream user, would you prefer that the multimodal options were available in a widely used package such as Microsoft Word?
    Yes No

11. From the point of view of a mainstream user, do you think that the multimodal interface will assist in working more efficiently under varying working conditions?
    Yes No

12. From the point of view of a mainstream user, do you like the idea that the multimodal interface will provide more flexibility and choice of input techniques?
    Yes No

13. Disabled users who cannot make use of a keyboard and mouse are generally forced to make use of a specially designed application. As a consequence, they are not normally assimilated into the user group of the mainstream applications. Do you feel that the distinction is justified?
    Yes No

14. Are you of the opinion that the solution offered in Microsoft Word is a possible solution for disabled users?

Yes No

15. As a first impression, did you find the use of eye-tracking as an interaction technique an exciting development?

Yes No

16. As a more long term application, do you think the use of eye tracking as an interaction technique will be beneficial?

Yes No

17. Rank the eye gaze interaction techniques in the order in which you enjoyed using them where 1 is the most enjoyable and 4 the least enjoyable.

	Dwell time		Blinking		Enter key		Combined with voice commands

18. Indicate which eye gaze interaction techniques you think are the most viable/usable for a more long term use. Rank them from 1 as the most viable to 4 as the least viable.

	Dwell time		Blinking		Enter key		Combined with voice commands

19. As a first impression, did you find the use of speech recognition for verbal commands an exciting development?

Yes No

20. As a more long term application, do you think the use of speech recognition for verbal commands will be beneficial?

Yes No

21. Did you enjoy using the verbal commands?

Yes No

22. Did the verbal commands allow you to navigate easier than what you normally do?

Yes No

23. If you answered no to question 20, do you think that with practice you will be more efficient with verbal commands than the way you normally do?

Yes No

24. As a first impression, did you find the use of eye-tracking and speech recognition together an exciting development?

Yes No

25. In terms of a long-term solution, do you think eye-tracking and speech recognition together will offer a usable working environment?

Yes No

26. Please provide suggestions for improvements or changes to the application.

APPENDIX D

POINTING DEVICE STUDY PRE–TEST QUESTIONNAIRE

University of the Free State

Department of Computer Science and Informatics

Pre-test Questionnaire

ON BEHALF OF THE UNIVERSITY OF THE FREE STATE AND THE DEPARTMENT OF COMPUTER SCIENCE WE WOULD LIKE TO THANK YOU FOR PARTICIPATING IN THIS RESEARCH PROJECT.

WE OFFER OUR ASSURANCE THAT ALL INFORMATION CAPTURED AND/OR RECORDED HERE WILL ONLY BE USED FOR RESEARCH PURPOSES AND YOUR PARTICIPATION IS VOLUNTARY.

PLEASE ANSWER THE FOLLOWING QUESTIONS.

1. Subject unique identifier (***will be provided by the facilitator***):_____

2. Age: _____

3. Home Language: _____

4. For how many years have you been using a computer?

☐ Never used a computer ☐ Less than 1 year

☐ 1 – 3 years ☐ 3-5 Years

☐ More than 5 years

5. How often do you use a computer?

☐ Daily ☐ Weekly

☐ Once every two weeks ☐ Once a month

☐ Less than once a month

6. For how many years have you been using a computer mouse?

☐ Never used a mouse ☐ Less than 1 year

☐ 1 – 3 years ☐ 3-5 Years

☐ More than 5 years

7. How often do you use a computer mouse?

☐ Daily ☐ Weekly

☐ Once every two weeks ☐ Once a month

☐ Less than once a month

8. Have you ever used an eye tracker to work on a computer?

☐ Yes ☐ No

If Yes, proceed to Question 9, else proceed to Question 11.

9. Have you ever used an eye tracker as a pointing device (substitute for a mouse)?

☐ Yes ☐ No

If Yes, for how long and how often do you use it?

_____

10. Specify in what capacity you have used an eye tracker.

_____

11. Have you ever used speech recognition to work on a computer?

☐ Yes ☐ No

If Yes, proceed to Question 12, else the questionnaire is complete.

12. Have you ever used speech recognition for cursor control?

☐ Yes ☐ No

If Yes, for how long and how often do you use it?

_____

13. Specify in what capacity you have used speech recognition.

_____

APPENDIX E

POINTING DEVICE ASSESSMENT QUESTIONNAIRE

Device assessment
Please circle the x that is most appropriate as an answer to the given comment.

1.	The force required for actuation (propelling or moving the device) was				
	x	x	x	x	x
	too low				too high
2.	Smoothness during operation was				
	x	x	x	x	x
	very rough				very smooth
3.	The mental effort required for operation was				
	x	x	x	x	x
	too low				too high
4.	The physical effort required for operation was				
	x	x	x	x	x
	too low				too high
5.	Accurate pointing was				
	x	x	x	x	x
	easy				difficult
6.	Operation speed was				
	x	x	x	x	x
	too fast				too slow
7.	Neck fatigue				
	x	x	x	x	x
	none				very high
8.	General comfort:				
	x	x	x	x	x
	very uncomfortable				very comfortable
9.	Overall, the input device was				
	x	x	x	x	x
	very difficult to use				very easy to use

10. Please indicate which of the following you preferred by circling your preferred method.

Large Buttons Small Buttons

11. Please indicate which of the following you preferred by circling your preferred method.

    Framed button Inverted colour button

12. Do you think that you will eventually be able to achieve the same speeds with eye gaze and speech recognition as with a mouse? Y / N

13. Did you enjoy working with the speech recognition and eye gaze as a pointing device? Y / N

14. When using the mouse, did the magnification tool assist you to work more accurately? Y / N

Comments: _____

15. When using eye gaze and speech recognition, did the magnification tool assist you to work more accurately? Y / N

Comments: _____

Any other comments and suggestions:

_____

_____

_____

APPENDIX F

USER TESTING PRE–TEST QUESTIONNAIRE

University of the Free State

Department of Computer Science and Informatics

Pre-test Questionnaire

ON BEHALF OF THE UNIVERSITY OF THE FREE STATE AND THE DEPARTMENT OF COMPUTER SCIENCE WE WOULD LIKE TO THANK YOU FOR PARTICIPATING IN THIS RESEARCH PROJECT.

WE OFFER OUR ASSURANCE THAT ALL INFORMATION CAPTURED AND/OR RECORDED HERE WILL ONLY BE USED FOR RESEARCH PURPOSES AND YOUR PARTICIPATION IS VOLUNTARY.

PLEASE ANSWER THE FOLLOWING QUESTIONS.

1. Student Number:

_____

2. Age: _____

3. Gender: Male / Female

4. Current field of study: _____

5. For how many years have you been using a word processor?

☐ Never used a word processor ☐ Less than 1 year

☐ 1 – 3 years ☐ 3-5 Years

☐ More than 5 years

6. How often do you use a word processor?

☐ Daily ☐ Weekly

☐ Once every two weeks ☐ Once a month

☐ Less than once a month

7.
8. Do you ever use keyboard shortcuts in a word processor?

☐ Yes ☐ No

9. Do you prefer using a mouse or the keyboard to complete tasks in a word processor?

☐ Mouse ☐ Keyboard

10. Have you ever used an eye tracker to work on a computer?

☐ Yes ☐ No

If Yes, proceed to Question 10, else proceed to Question 12.

11. Have you ever used an eye tracker as a pointing device (substitute for a mouse)?

☐ Yes ☐ No

If Yes, for how long and how often do you use it?

_____

12. Specify in what capacity you have used an eye tracker.

_____

13. Have you ever used speech recognition to work on a computer?

☐ Yes ☐ No

If Yes, proceed to Question 13, else the questionnaire is complete.

14. Have you ever used speech recognition for cursor control?

☐ Yes ☐ No

If Yes, for how long and how often do you use it?

_____

15. Specify in what capacity you have used speech recognition.

_____

APPENDIX G

POST– TEST QUESTIONNAIRE – FIRST SESSION

Adapted from Shneiderman (1998). *Designing the User Interface*. p 136 – 143.

PART 3: Overall User Reactions

3.1 Overall reaction to the system:	**Terrible**				**Wonderful**
	1	2	3	4	5
3.2	**Frustrating**				**Satisfying**
	1	2	3	4	5
3.3	**Dull**				**Stimulating**
	1	2	3	4	5
3.4	**Difficult**				**Easy**
	1	2	3	4	5
3.5	**Inadequate**				**Adequate**
	1	2	3	4	5
3.6	**Rigid**				**Flexible**
	1	2	3	4	5

PART 6: Learning

6.1 Learning to operate the system	**Difficult**				**Easy**
	1	2	3	4	5
6.1.1 Getting started	**Difficult**				**Easy**
	1	2	3	4	5
6.1.2 Learning advanced features	**Difficult**				**Easy**
	1	2	3	4	5
6.1.3 Time to learn to use the system	**Slow**				**Fast**
	1	2	3	4	5

PART 7: System capabilities

7.4 Correcting your mistakes	**Difficult**				**Easy**
7.5 Ease of operation depends on your level of experience	**Never**				**Always**
	1	2	3	4	5
7.5.1 You can accomplish tasks knowing only a few commands	**With difficulty**				**Easily**
	1	2	3	4	5
7.5.2 You can use features/shortcuts	**With difficulty**				**Easily**
	1	2	3	4	5

Device assessment
Please circle the x that is most appropriate as an answer to the given comment.

1. The force required for actuation (propelling or moving the device) was

x	x	x	x	x
too low				too high

2. Smoothness during operation was

x	x	x	x	x
very rough				very smooth

3. The mental effort required for operation was

x	x	x	x	x
too low				too high

4. The physical effort required for operation was

x	x	x	x	x
too low				too high

5. Accurate pointing was

x	x	x	x	x
easy				difficult

6. Operation speed was

x	x	x	x	x
too fast				too slow

7. Neck fatigue

x	x	x	x	x
none				very high

8. General comfort:

x	x	x	x	x
very uncomfortable				very comfortable

9. Overall, the input device was

x	x	x	x	x
very difficult to use				very easy to use

10. Do you think that you will eventually be able to achieve the same speeds with eye gaze and speech recognition as with a mouse and keyboard? Y / N

11. Did you enjoy working with the speech recognition and eye gaze as a pointing device? Y / N

12. Do you think the added features in Word are useful? Y / N

13. Do you think the added features in Word make the Word application better?
Y / N

14. Do you think that the added features in Word will ever gain mainstream use?
Y / N

Give reasons for your answer:

15. Do you think that there can be a market for the added features in Word? Y / N

Give reasons for your answer:

16. Would you consider adopting the added features as a standard means of using Word for yourself?
Y / N

Any other comments and suggestions:

APPENDIX H

POST–TEST QUESTIONNAIRE – LAST SESSION

University of the Free State

Department of Computer Science and Informatics

Post-test Questionnaire

PLEASE ANSWER THE FOLLOWING QUESTIONS.

Student Number:

_____

Do you have corrected vision (glasses / contact lenses)? If so, please indicate which:

Y / N

Glasses / Contact lenses

1. Do you think that you will eventually be able to achieve the same speeds with eye gaze and speech recognition as with a mouse and keyboard? Y / N

2. Did you enjoy working with the speech recognition and eye gaze as a pointing device? Y / N

3. Do you think the added features in Word are useful? Y / N

4. Do you think the added features in Word make the Word application better? Y / N

5. Do you think that the added features in Word will ever gain mainstream use? Y / N
   Give reasons for your answer:
   _____
   _____

6. Do you think that there can be a market for the added features in Word? Y / N
   Give reasons for your answer:
   _____
   _____

7. Would you consider adopting the added features as a standard means of using Word for yourself?
   Y / N

Give reasons for your answer:

_____

_____

PART 3: Overall User Reactions – Complete system

Answer the following questions based on your experience with the system as a whole.

3.2 Overall reaction to the system:	**Terrible**				**Wonderful**
	1	2	3	4	5
3.2	**Frustrating**				**Satisfying**
	1	2	3	4	5
3.3	**Dull**				**Stimulating**
	1	2	3	4	5
3.4	**Difficult**				**Easy**
	1	2	3	4	5
3.5	**Inadequate**				**Adequate**
	1	2	3	4	5
3.6	**Rigid**				**Flexible**
	1	2	3	4	5

PART 6: Learning

6.1 Learning to operate the system	**Difficult**				**Easy**
	1	2	3	4	5
6.1.1 Getting started	**Difficult**				**Easy**
	1	2	3	4	5
6.1.2 Learning advanced features	**Difficult**				**Easy**
	1	2	3	4	5
6.1.3 Time to learn to use the system	**Slow**				**Fast**
	1	2	3	4	5

Any other comments and suggestions:

_____

_____

_____

_____

PART 3: Overall User Reactions - Typing

Answer the following questions based on your experience of typing with the eye tracking and speech recognition.

3.3 Overall reaction:	**Terrible**				**Wonderful**
	1	2	3	4	5
3.2	**Frustrating**				**Satisfying**
	1	2	3	4	5
3.3	**Dull**				**Stimulating**
	1	2	3	4	5
3.4	**Difficult**				**Easy**
	1	2	3	4	5
3.5	**Inadequate**				**Adequate**
	1	2	3	4	5
3.6	**Rigid**				**Flexible**
	1	2	3	4	5

PART 6: Learning

6.1 Learning to type	**Difficult**				**Easy**
	1	2	3	4	5
6.1.1 Getting started	**Difficult**				**Easy**
	1	2	3	4	5
6.1.2 Learning advanced features	**Difficult**				**Easy**
	1	2	3	4	5
6.1.3 Time to learn to use the system	**Slow**				**Fast**
	1	2	3	4	5
7. How natural did it feel to type using eye gaze and speech	**Unnatural**				**Natural**
	1	2	3	4	5

Any other comments and suggestions:

PART 3: Overall User Reactions - Commands

Answer the following questions based on your experience of issuing commands to the system for formatting and cursor movement.

3.4 Overall reaction:	**Terrible**				**Wonderful**
	1	2	3	4	5
3.2	**Frustrating**				**Satisfying**
	1	2	3	4	5
3.3	**Dull**				**Stimulating**
	1	2	3	4	5
3.4	**Difficult**				**Easy**
	1	2	3	4	5
3.5	**Inadequate**				**Adequate**
	1	2	3	4	5
3.6	**Rigid**				**Flexible**
	1	2	3	4	5

PART 6: Learning

6.1 Learning to issue commands	**Difficult**				**Easy**
	1	2	3	4	5
6.1.1 Getting started	**Difficult**				**Easy**
	1	2	3	4	5
6.1.2 Learning advanced features	**Difficult**				**Easy**
	1	2	3	4	5
6.1.3 Time to learn to use the system	**Slow**				**Fast**
	1	2	3	4	5
6.1.4 Time to learn to string commands together	**Slow**				**Fast**
	1	2	3	4	5
7. How natural did it feel to issue commands	**Unnatural**				**Natural**
	1	2	3	4	5

8. For each of the following command types, indicate how easy it was to use them:	**Difficult**				**Easy**
Moving the cursor	1	2	3	4	5
Formatting text (e.g. bold, italic)	1	2	3	4	5
Selecting text (e.g. line or word)	1	2	3	4	5
Cutting/copying and pasting	1	2	3	4	5

Any other comments and suggestions:

15. Did you feel your typing improved over the time period in which you used the system?

☐ Yes ☐ No

Comments:

16. Did you feel more at ease with issuing commands as you became accustomed to the system?

☐ Yes ☐ No

Comments:

17. Would you consider using a system like this for typing purposes?

☐ Yes ☐ No

Give a reason for your answer:

18. Would you consider using a system like this for issuing commands?

☐ Yes ☐ No

Give a reason for your answer:

19. Did the audio feedback when typing assist in the typing process?

☐ Yes ☐ No

Give a reason for your answer:

20. Would you have preferred having visual feedback during the typing task (e.g. button changes colour)?

☐ Yes ☐ No

Give a reason for your answer and any other suggestions for feedback you may have:

For the typing tasks, there were five sentences you had to type right at the end of the test. The first sentence used large buttons, the second and third sentence smaller buttons which were spaced further apart and the fourth and fifth sentences used smaller buttons which were spaced closer together.

21. Rank the buttons in order of your preference where 1 is the most liked and 3 the least liked:

☐ Large buttons	☐ Smaller buttons, far apart	☐ Smaller buttons, closer together

Comments:

---

22. Rank the buttons in the order in which they were easiest to use where 1 is the easiest and 3 the most difficult:

☐ Large buttons	☐ Smaller buttons, far apart	☐ Smaller buttons, closer together

Comments:

Device assessment
Please circle the x that is most appropriate as an answer to the given comment.

Answer the following questions regarding using eye gaze and speech recognition for typing:

1. The force required for actuation (propelling or moving the device) was

x	x	x	x	x
too low				too high

2. Smoothness during operation was

x	x	x	x	x
very rough				very smooth

3. The mental effort required for operation was

x	x	x	x	x
too low				too high

4. The physical effort required for operation was

x	x	x	x	x
too low				too high

5. Accurate pointing was

x	x	x	x	x
easy				difficult

6. Operation speed was

x	x	x	x	x
too fast				too slow

10. Neck fatigue

x	x	x	x	x
none				very high

11.	General comfort:				
	x very uncomfortable	x	x	x	x very comfortable
12.	Overall, the input device was				
	x very difficult to use	x	x	x	x very easy to use

Any other comments and suggestions:

APPENDIX I

PUBLICATIONS

To date, there have been four publications stemming from the research study discussed in the thesis. These publications are as follows:

Appendix I-1: Abstract is reproduced here in Afrikaans as it was originally published
Beelders, T.R. and Blignaut, P.J. (2009). A multi-modal interface for a popular word processor. *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns Studentesimposium 2009*, Bloemfontein, South Africa.

Appendix I-2
Beelders, T.R. and Blignaut, P.J. (2010). Using vision and voice to create a multimodal interface for Microsoft Word 2007. *Proceedings of the Symposium on Eye-Tracking Research and Applications (ETRA)*, Austin, Texas, United States of America, 173-176.

Appendix I-3: Abstract is reproduced here in Afrikaans as it was originally published
Beelders, T.R., Blignaut, P.J. and Greeff, F. (2010). Eye-tracking and speech recognition instead of a computer mouse. *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns Studentesimposium 2010, Pretoria*, South Africa.

Appendix I-4:
Beelders, T.R. and Blignaut, P.J. (2011). The Usability of Speech and Eye Gaze as a Multimodal Interface for a Word Processor. In I. Ipšić (Ed), *Speech Technologies* (pp. 385-404). ISBN: 978-953-307-996-7.

’n Multimodale koppelvlak vir ’n gewilde woordverwerkingspakket

’n Woordverwerker is ’n populêre rekenaarprogram wat deur ’n diverse groep gebruikers op ’n gereelde basis gebruik word. ’n Enkele program moet dus vir ’n groot verskeidenheid gebruikers, elkeen met sy eie behoeftes en voorkeure vir interaksie, voorsiening maak. Voorts is gestremde gebruikers gewoonlik beperk in hulle keuses omdat net sekere programme hulle beperkinge in ag neem. In die algemeen word programme wat deur gestremde gebruikers gebruik word, nie deur die hoofstroom gebruikers gebruik nie, maar word spesiaal vir gestremde gebruikers geskryf. Verder is die neiging om weg te beweeg van die standaard koppelvlakke met menus en ikone wat met die muis gemanipuleer word. Die fokus van hierdie studie is om nie-tradisionele interaksie tegnieke in ’n woordverwerker in te bou en dan vas te stel of dit ’n volwaardige oplossing bied om toeganklikheid vir alle gebruikers te verseker.

Daar is heelwat woordverwerkingsprogramme op die mark beskikbaar, waarvan Microsoft Word die gewildste is. Hierdie studie gebruik dus Microsoft Word 2007 as ’n basis waarin ekstra interaksie-tegnieke ingebed word om ’n multimodale koppelvlak te bied. Een van die nuwe interaksietegnieke laat toe dat ’n oog-volgapparaat (Engels “eye-tracker”) gebruik word om ’n dokument te redigeer. Vir dié doeleindes kan ’n muis kliek op verskeie maniere deur die gebruiker se oë gesimuleer word. Die tweede nuwe interaksietegniek wat in die koppelvlak ingebed word, maak voorsiening vir die gebruik van spraakherkenning om teks te dikteer, sowel as om redigeringsopdragte hardop uit te spreek.

Genoemde twee interaksietegnieke kan ook gekombineer word sodat die konteks van ’n mondelinge instruksie bepaal word deur die item waarna gekyk word. So byvoorbeeld kan die gebruiker na die “Bold” ikoon in die taakbalk kyk en dan hardop sê “click”. Verder word ’n afbeelding van ’n toetsbord onder-aan die skerm vertoon en die gebruiker kan ’n dokument in Microsoft Word tik deur slegs na die onderskeie toetse op die toetsbord te kyk. Die muiswyser volg die gebruiker se blik en die onmiddellike area onder die muiswyser kan ook vergroot word om dit vir gebruikers met swak sig makliker te maak om met die koppelvlak te werk.

Die studie beoog om die verskillende interaksietegnieke met mekaar te vergelyk om te bepaal watter kombinasie van tegnieke die bruikbaarste is. ’n Ekspertanalise is reeds gedoen om die langtermyn lewensvatbaarheid van sodanige koppelvlak te evalueer en om die eerste indrukke van die interaksietegnieke soos wat hulle in Word 2007 gebruik kan word, te kry.

Die volgende stap is om te bepaal of die nuwe interaksietegnieke produktiwiteit verhoog en of gebruikers kan leer om die tegnieke te gebruik om aan hulle bepaalde omstandighede en behoeftes te voldoen. Om dit te doen sal toetsgebruikers gevra word om verteenwoordigende take uit te voer deur van al die moontlike interaksietegnieke gebruik te maak. Die tyd wat gebruikers neem en die korrektheid waarmee take uitgevoer word, sal vergelyk word om te bepaal of die veranderde koppelvlak gebruikers toelaat om ten minste dieselfde vlak van produktiwiteit te behaal as wat met ’n standaard koppelvlak bereik kan word.

Using Vision and Voice to Create a Multimodal Interface for Microsoft Word 2007

Abstract

There has recently been a call to move away from the standard WIMP type of interfaces and give users access to more intuitive interaction techniques. Therefore, it in order to test the usability of a multimodal interface in Word 2007, the most popular word processor, the additional modalities of eye gaze and speech recognition were added within Word 2007 as interaction techniques. This paper discusses the developed application and the way in which the interaction techniques are included within the well-established environment of Word 2007. The additional interaction techniques are fully customizable and can be used in isolation or in combination. Eye gaze can be used with dwell time, look and shoot or blinking and speech recognition can be used for dictation and verbal commands for both formatting purposes and navigation through a document. Additionally, the look and shoot method can also be combined with a verbal command to facilitate a completely hands-free interaction. Magnification of the interface is also provided to improve accuracy and multiple onscreen keyboards are provided to provide hands free typing capabilities.

Keywords: Eye-tracking, speech recognition, usability, word processing, multimodal

Introduction

The word processor has become a very popular tool in the everyday use of a computer [Roberts and Moran, 1983] and by 1984, 80-100% of users' time on a computer was spent using a word processor or other editor-based application [Rosson, 1984]. The word processor application has evolved substantially since its initial inception and since then has undergone a virtual metamorphosis to achieve the capabilities that are available in these applications today. As an integral part of everyday life for many people it caters for a very diverse group of users, therefore, it is highly unlikely that only one such complex application would be able to offer the best possible experience to all users [Sullivan, 1989]. Furthermore, users with disabilities or needs other than those of mainstream users are not always taken into consideration during system development and often have to compensate by using specially designed applications which do not necessarily compare with the more popular applications. This study therefore aims to investigate various means to increase the usability of a word processor for as wide a user group as possible.

For this reason, the interface of the most popular word processor application will be extended into a multimodal interface. This interface should facilitate use of the mainstream product by marginalized users, whilst at the same time enhancing the user experience for novice, intermediate and expert users. Ideally the interface should be customizable and allow users to select any combination of interaction techniques which suit their needs. The premise of the research study is not to develop a new word processor but rather to incorporate additional interaction techniques, besides the keyboard and mouse, into an application which has already been accepted by the user community. This will allow for the improvement of an already popular product and stimulate inclusiveness of non-mainstream users into the mainstream market.

The research study is still in the beginning phase where development of the tool is underway. Therefore, for the purposes of this paper, the application as it has been developed will be the main focus. The paper will, however, conclude with a short discussion of the next phases of the research study.

Interaction Techniques

Using a physical input device in order to communicate or perform a task in human-computer dialogue is called an interaction technique [Foley, et al., 1990 as cited in Jacob, 1995]. The interaction techniques of speech recognition and eye tracking will be included in a popular word processor interface to create a multimodal interface as a means to determine whether the usability of this product can be enhanced in this way.

Although this approach has received limited attention thus far, the multimodal approach has always focused on the development of a third-party application, for example EyeTalk [Hatfield and Jenkins, 1997]. Contrary to this, this study will use an already existing application, namely Microsoft Word ©, which currently enjoys a high prevalence in the commercial market.

Development environment

The development environment used was Visual Studio 2008, making use of the .NET Framework 3.5. Visual Studio Tools for Microsoft Office System 2008 (VSTO) in C# was used for development. VSTO allows programmers to use managed code to build Office-based solutions in C# and VB.NET [Anderson, 2009]. In order to incorporate the speech recognition the Microsoft Speech Application Programming Interface (SAPI) with version 5.1 of the SDK was used. The SDK provides the capability of compiling customized grammars and accessing the functionalities of the speech recognizer. In order to provide gaze interaction Tobii SDK 1.5.4 was used. For magnification purposes, which will be discussed in an upcoming section, the commercial product Magnifying Glass Pro 1.7 was chosen as a relatively inexpensive solution but primarily based on the fact that it was one of the few applications which incorporated clickable areas within the magnified area which are then correctly transferred to the underlying area. This is essential in the developed product as the magnification will increase the accuracy of cursor positioning via eye gaze and correct interpretation of user intention and requiring the user to disable magnification before clicking on the interface would negate all the advantages gained from magnification

The aim of the development process was to incorporate speech recognition and eye tracking as additional interaction techniques in the Microsoft Word environment. The user should also be given the freedom to determine in which combination the interaction techniques must be used, while still having the option of continued use of the traditional interaction techniques. As illustrated in Figure 1, an extra tab was added to the established Microsoft Word ribbon. This tab (circled in red) was named Multimodal Add-Ins.

The new tab provides numerous options to the user to select which additional interaction techniques they would like to use (Figure 1). As is evident from Figure 1, complete customization of the techniques is allowed via selection of any combination of techniques as well as in what capacity the techniques must be implemented.

Figure 1: *Multimodal Add-ins for Word 2007*

Additional tools which are available to enhance the user experience are a magnification tool and an onscreen keyboard which can be displayed at the bottom of the Word document. The magnification tool magnifies the immediate area under the mouse cursor, thereby providing increased accuracy for users with weak eyesight and those making use of the gaze sensitive interface. Magnification is available when using the mouse or when using eye gaze as an interaction technique. The use of the magnification tool is entirely at the discretion of the user who is capable of turning magnification on and off at will or as needed.

Onscreen keyboards are available as an alternative to using a traditional keyboard. The onscreen keyboard can be used either through use of the traditional mouse or to achieve hands-free typing using eye gaze or a combination of eye gaze and speech recognition. The final adapted interface, as envisioned in use when the on screen keyboard is in use, is shown in Figure 2.

The layout of the onscreen keyboard can be changed to either a traditional QWERTY keyboard layout or to an alphabetic layout. Each keyboard contains all 26 alphabetic letters, a Space bar, Backspace and Delete keys as well as special keys which simplify movement through the document. Special keys which are provided are Page up, Page down, Home and End. The user can also toggle between upper case and lower case by activating and deactivating the CAPS lock key. A Select All key is provided as a means for the user to select all the text in the document. The two red arrows in the lower left corner of the keyboard (Figure 2) change the size of all keyboard keys in decrements and increments of 10 pixels respectively, thereby providing even more customization of the keyboard for the user. Auditory feedback in the form of a soft beep is given when a keyboard key is clicked on.

Speech recognition

The user has the option of enabling the speech engine so that Microsoft Word can respond to verbal utterances. In terms of the customizable options, the user can toggle between dictation mode and command mode. In dictation mode, the speech recognition is implemented in the well-known method of capturing vocalizations, translating those vocalizations into text and writing the result to the currently activated document in Microsoft Word. In order for the dictation mode to be effective the user must select a previously trained profile. A unique profile can be trained through the Windows Speech wizard. All the available speech profiles are provided in a drop-down box on the multimodal add-in tab for the convenience of the user.

In command mode, a grammar is activated which accepts only isolated commands and responds to these in a pre-determined manner. Command mode provides the functions of cursor control, formatting capabilities and certain document handling capabilities. Several different commands are provided which have the same application reaction, thereby contributing to further customization for the user as they can determine which the most desirable command is for them to use. Moreover, simple cursor control is provided by providing directional commands but more complex cursor control is also provided by allowing line selection and movement of the cursor as though control keys (such as Shift) are being pressed in combination with the verbal command. These types of commands will simplify selection of text and provide verbal commands for complex key combinations which are not always known to novice and intermediate users. For example, the word “Bold” causes the activation or deactivation of the bold formatting style. Similarly the words “Italic” and “Underline” activate or deactivate their formatting style. Words such as “Cut”, “Copy” and “Paste” allow for text manipulation and are their subsequent actions are of course the cutting or copying of the currently selected text and the pasting of the clipboard contents at the position of the cursor. More complex commands for text selection are available such as “Select line”, which selects the whole line on which the cursor is situated, “Select word”, which selects the word nearest to the right of the current cursor position. Cursor control is achieved through the commands “Left”, “Right”, “Up” and “Down”. Verbal commands can be issued in sequence to perform relatively complex document manipulation.

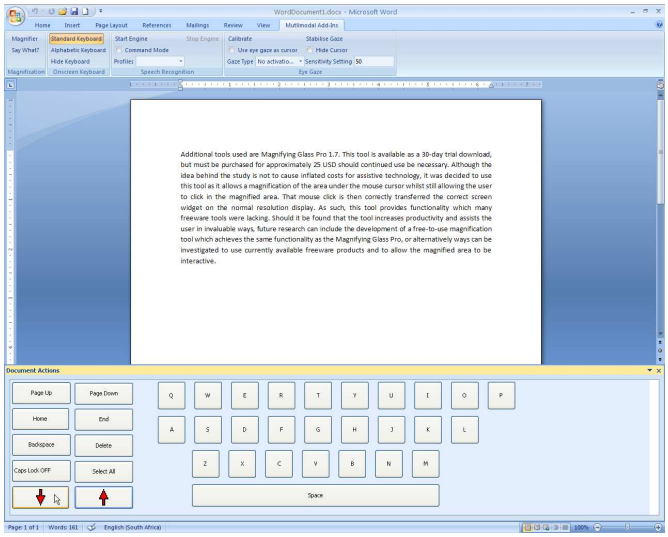


Figure 2: *Adapted interface of Word 2007 when the onscreen keyboard is activated*

Eye-tracking

The eye tracker can be calibrated for use directly through the Microsoft Word interface. This increases the usability of the application as the user is not required to move between applications to achieve their goal of using gaze as an interaction technique. Since the word processor is the focus of this study, this meets the requirement of the research question scope. The user has the option to activate eye gaze which can then be used to position the cursor in the document or over an object to be manipulated. Customization is provided by allowing the user to choose the activation method. For use purely as a single interaction method, the choices of dwell time, look and shoot and blinking are provided. When dwell time is selected, the user is able to set the interval of the dwell time (see the Sensitivity Setting text box o Figure 1). This provides additional customization as the user can determine the speed with which they are most comfortable and leaves the option for adjusting this interval as the user gains more confidence and experience with gaze based interaction. The interval can be changed at any time during the use of the application. Dwell time requires the user to fixate on a position for the set time of the dwell time interval before a left mouse click is simulated. When selecting the look and shoot method, the user can position the cursor using eye gaze and then press the Enter key to simulate a left mouse click. This would have the effect of either placing the cursor at the position of the eye gaze or clicking on the icon directly under the eye gaze of the user. The third option available to the user is that of blinking. In this scenario, the user fixates on the desired object or position and then blinks their eyes to simulate a left mouse click.

Multiple interaction techniques

When the user selects No activation (Figure 1) via eye gaze that implies that they will instead be using voice commands to respond to the current eye gaze position of the user. In this instance, the speech recognition must also be enabled and the user can then issue verbal commands to move the cursor to the current gaze position which is analogous to executing a left mouse click at that position. In this way, it is possible for the user to place the cursor at any position in the document, or to click one of the Microsoft Word icons on the ribbon. The verbal commands of “Go”, “Click” or “Select” all simulate a left mouse click at the button closest to the current gaze position. In this way, the user is free to choose the command which they find most suitable for them.

In most instances it is envisioned that the onscreen keyboard will also be activated under these circumstances. When the onscreen keyboard is activated in conjunction with the eye gaze, visual feedback is given to the user to indicate which button will be clicked when the verbal command is issued. With each fixation that is detected within the boundaries of the keyboard, the button which is closest to that fixation is determined to be the target and a shape is displayed in the centre of the button. The user can also select which shape they would like to use for visual feedback. The available shapes are a solid square, a hollow square, a solid disk and a hollow circle. The hollow shapes do not obscure the letter of the key and in so doing provide the necessary visual feedback whilst still allowing the user to see the letter which will be written to the document. Feedback is only given on the keyboard to minimize interference during normal document browsing. In order to achieve increased stabilization of the feedback within a targeted object, the algorithm as suggested by Kumar (2007) was used.

If the user is satisfied that the correct button has been determined they can then issue any of the verbal commands to simulate a left mouse click. The letter shown on the keyboard is then written to the document at the current cursor position.

Where to next?

As previously mentioned, the research study is still in the preliminary stages of an empirical study. An application has been developed to investigate the effect of multimodal interaction techniques on the usability of a mainstream word processor application. Further enhancements to the application will include the expansion of the keyboards to include numerical keys and the magnification will be refined to respond to eye gaze and voice commands. More voice commands will be provided for, particularly for commands that currently have shortcut keys assigned to them, such as Save and displaying certain dialog boxes.

Additionally, a back-end will be written for the application which will capture certain measurements which can be used for usability analysis. Measurements such as the number of errors made during a task, the number of actions required and the percentage of the task completed correctly will automatically be saved to a database for further analysis.

Once the application has been completed, user testing will commence. Both disabled and non-disabled users of a local university will be approached to participate in the study. A longitudinal study will be conducted whereby the participants will be required to spend periods interacting with the system. After each exposure to the system, users will be required to complete a number of tasks for which measurements will be captured. In this way, the learnability of the study can be measured over a period of time by comparing the results of these sessions to determine if user performance increases in correlation to user exposure to the application. Since it is expected that there will be a learning curve associated with the application, it is deemed more applicable to capture usability measurements over a period of time rather than only after a single session with the application. In order to determine whether the application succeeds in providing for disabled users whilst simultaneously providing for a better user experience for mainstream users, it is imperative that users from both these demographics be included in the sample. Furthermore, to further investigate the usability of the newly developed application, user efficiency effectiveness can be measured in a within-subjects experiment by requiring users to complete identical tasks in both the commercial Microsoft Word and the new multimodal Microsoft Word.

Moreover, the usability of the various interaction techniques will also be analyzed to determine which combination of the interaction techniques provides the most usable interface – if any. User satisfaction will be measured through means of a questionnaire in order to gauge user reaction, both in a short-term and long-term exposure period.

Summary

A multimodal interface was developed for Microsoft Word in order to eventually determine whether the usability of this application can be enhanced for mainstream users whilst simultaneously providing an adaptable and usable interface for disabled users. For these purposes, eye tracking and speech recognition capabilities were built into the Word interface. These interaction techniques can be used in isolation or in combination and the way in which they are used can be customized in a number of ways. Once the development has been completed and measurements can be captured automatically in the background during user interaction, a longitudinal usability study will be undertaken Both disabled and able-bodied users will be included in the sample and will be required to complete a number of practice sessions with the application over a prolonged period of time. After each session, participants will be required to complete a number of tasks, during which measurements will be captured for further analysis In this way, it will be possible to determine whether users are able to improve their performance on the system over an extended period – in other words, whether the system is usable. Additionally, user performance between the new application and the commercially available application will be compared to determine whether they can achieve comparable performance on both the systems. In this way, it will be possible to determine whether a popular commercial application can be fully extended into a worthwhile multimodal application which caters for a diverse group of users comprised of both disabled and able-bodied users.

References

ANDERSON, T. (2009). *Pro Office 2007 development with VSTO*. APress: United States of America.

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

JACOB, R.J. (1995). Eye tracking in advanced interface design. In *Virtual Environments and Advanced interface Design*, W. Barfield and T. A. Furness, Eds. Oxford University Press, New York, NY, 258-288.

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

ROBERTS, T.L. AND MORAN, T.P. (1983). The evaluation of text editors: Methodology and empirical results. *Communications of the ACM*, 26(4): 265-283.

ROSSON, M.B. (1984a). Characterizing freeform editing behavior. IBM Research Report RC 10550, IBM T. J. Watson Research Center, Yorktown Heights, New York.

SULLIVAN, P. (1989). Human-computer interaction perspectives on word-processing issues. *Computers and Composition*, 6(3): 11-33.

Oogvolging en spraakherkenning in plaas van ’n rekenaarmuis

Die doel van die studie was om te bepaal hoe effektief ’n oog-volgapparaat (“eye-tracker”) en spraakherkenning in plaas van ’n muis gebruik kan word om teikens op ‘n rekenaarskerm te selekteer. Die International Standards Organisation (ISO) standaard 9241-9 bestaan uit ses seleksietake. Een van hierdie take vereis dat die gebruiker die muis of alternatiewe aanwyser moet gebruik om 16 teikens in ’n sekere volgorde te selekteer. Die effektiwiteit word gemeet in terme van die spoed en akkuraatheid waarmee die seleksies gedoen word. Op hierdie wyse kan bepaal word of teikens met dieselfde effektiwiteit geselekteer kan word met oogvolging en spraakherkenning as wat die geval is met ’n muis.

Vir elk van die seleksietegnieke is die effektiwiteit verder ondersoek met betrekking tot die grootte van die teiken, die gebruik van ’n gravitasieput, ’n elektroniese vergootglas en visuele terugvoer. ’n Gravitasieput laat ’n gebruiker toe om effens buite die teiken te kliek en dan word die wyser as’t ware in die teiken ingetrek. ’n Elektroniese vergrootglas vergroot die area direk onder die wyser en visuele terugvoer behels dat ‘n raampie om die geselekteerde teiken getrek word. Elke toetspersoon het die seleksietaak met 14 verskillende kombinasies van faktore uitgevoer. ’n Gebalanseerde Latynse vierkant is gebruik om die volgorde van toetse vir elke persoon te bepaal sodat die effek van leer deur ervaring geminimaliseer word.

Twintig studente het aan die studie deelgeneem en daar is van deelnemers verwag om ten minste muisvaardig te wees. Benewens die seleksietaak wat elke deelnemer op 14 verskillende maniere moes doen, moes elke deelnemer ook ’n vraelys voltooi om subjektiewe terugvoer omtrent elkeen van die verskillende toetsvariasies te verkry.

Analise van die data sal bepaal of die kombinasie van oogvolging en spraakherkenning effektief genoeg is om as alternatiewe interaksietegniek vir rekenaargebruik te dien.

Eye gaze and speech recognition instead of a computer mouse

The combination of eye gaze and speech recognition as a selection technique was investigated using the ISO9241-9 multi-directional tapping task. Twenty participants were tested on 14 conditions with varying target size, magnification capabilities and presence of a gravity well. Analysis of the data will determine whether this is a viable alternative to the mouse.

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 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.

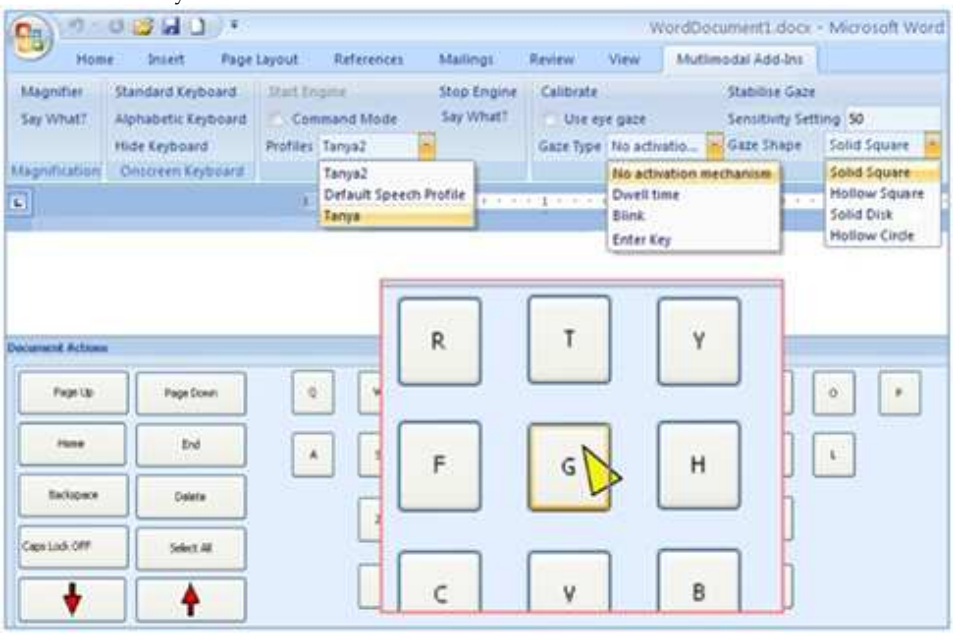


Figure 1: Multimodal Add-Ins tab in Microsoft Word

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.

Figure 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.

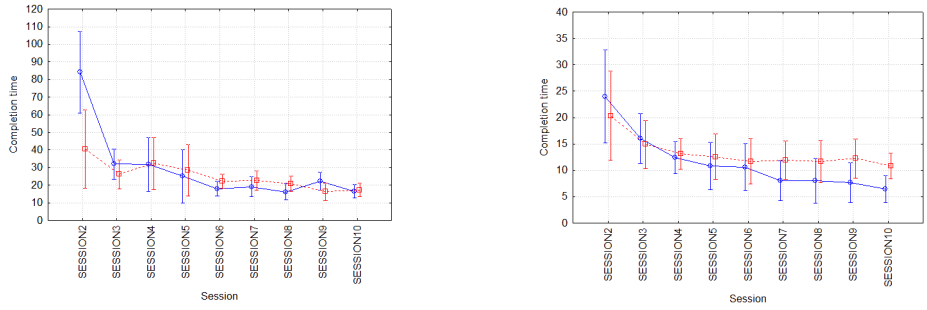


Figure 3: Average completion times for (a) line selection and formatting and (b) select all and remove

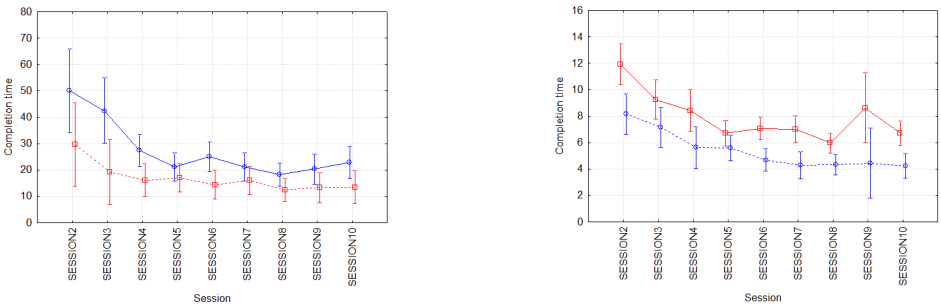


Figure 4: Average completion times for (a) select words and format and (b) paste

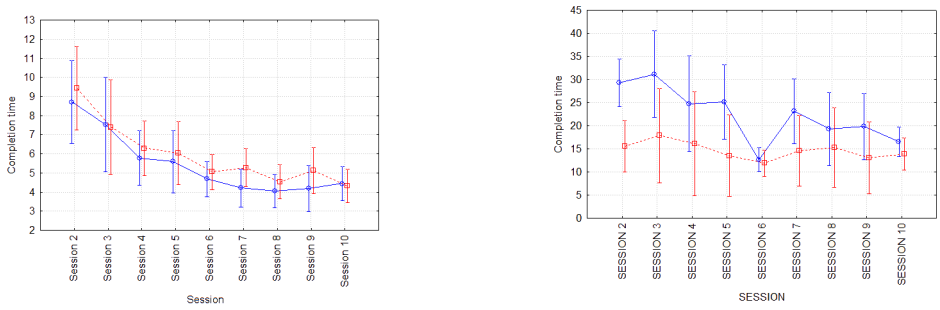


Figure 5: Average completion times for (a) undo and (b) select word and copy

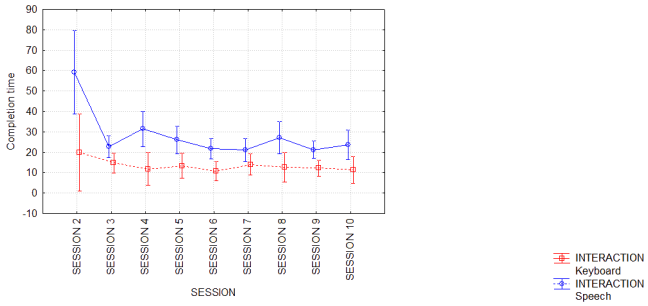


Figure 6: Average completion times for position and paste

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.

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:

3. $H_{0,1}$: There is no difference between the time required to complete the tasks when using the mouse and keyboard or speech commands.
4. $H_{0,2}$: 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.

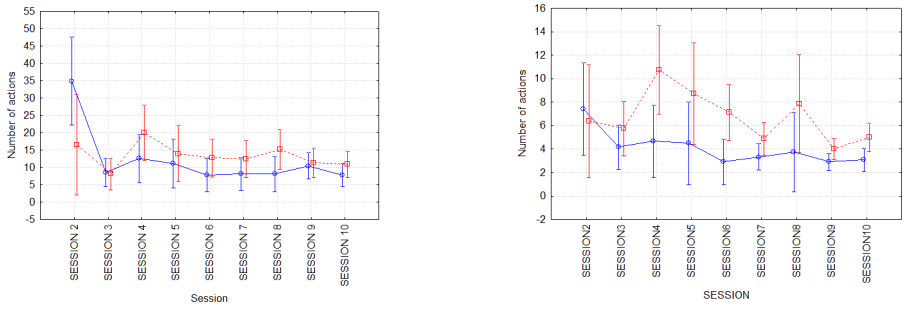


Figure 7: Average number of actions for (a) line selection and formatting and (b) select all and remove

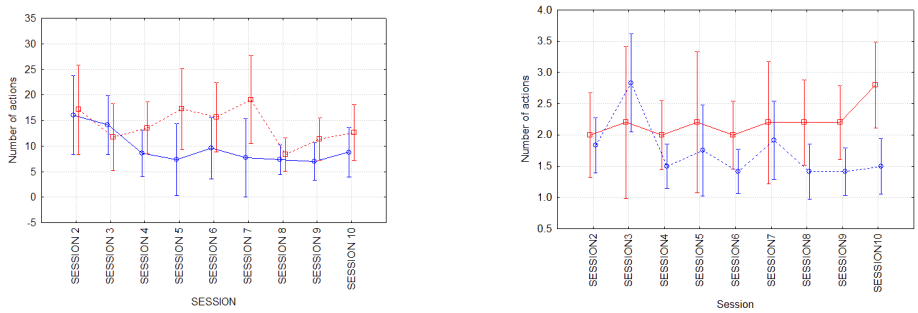


Figure 8: Average completion times for (a) select words and format and (b) paste

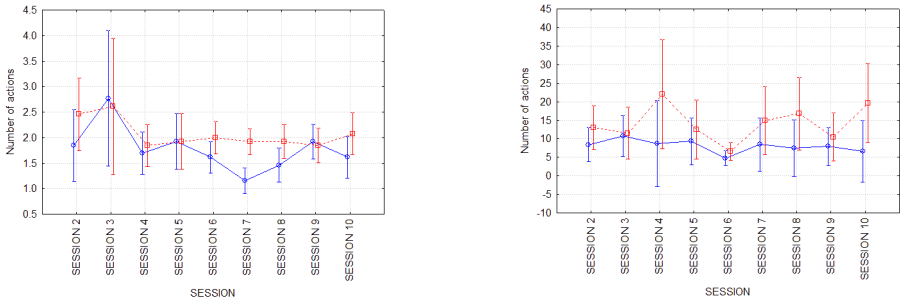


Figure 9: Average completion times for (a) undo and (b) select word and copy

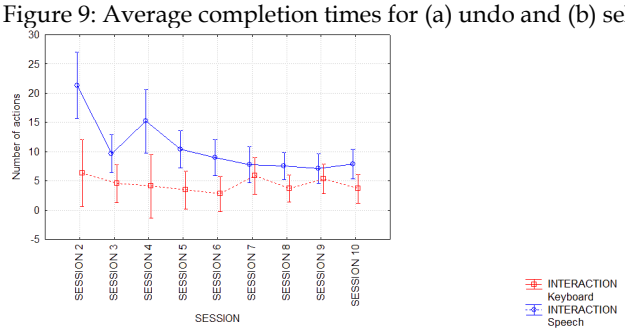


Figure 10: Average completion times for position and paste

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.

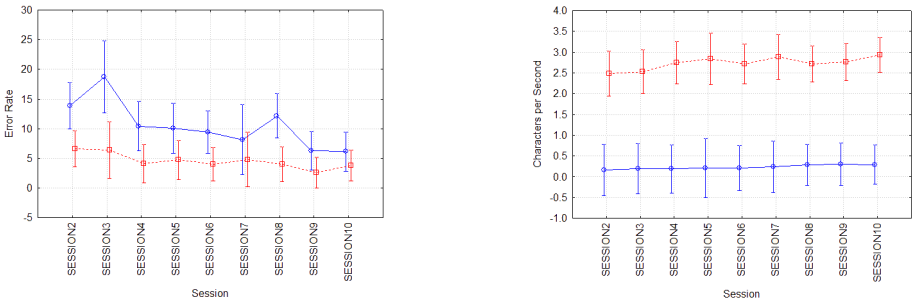


Figure 11: Least squares mean plot of character error rate and characters per second

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.

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.

The graphs below plot the error rate and characters per second for each of the text entry methods for the sessions during which they were tested.

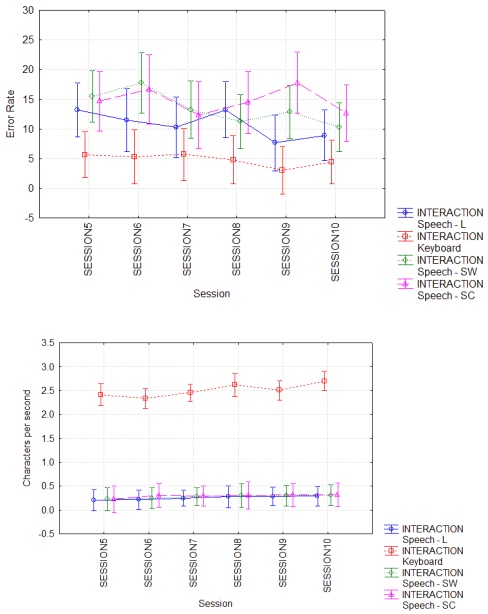


Figure 12: Least squares mean plot of character error rate and characters per second for all typing tasks

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

SUMMARY

Multimodal interfaces may herald a significant improvement on current GUIs which have been commonplace until now. It is also possible that a multimodal interface could provide a more intuitive and natural means of interaction which, simultaneously, negates the reliance on traditional, manual means of interaction. Eye gaze and speech are common components of natural human-human communication and were proposed for use in a multimodal interface for a popular word processor for the purposes of this study.

In order for a combination of eye gaze and speech to be a viable interface for a word processor, it must provide a means of text entry and facilitate editing and formatting of the document contents. For the purposes of this study a simple speech grammar was used to activate common word processing tasks, as well as for selection of text and navigation through a document. For text entry, an onscreen keyboard was provided, the keys of which could be pressed by looking at the desired key and then uttering an acceptable verbal command. These functionalities were provided in an adapted Microsoft Word 2007® to increase the customisability and possibly the usability of the word processor interface and to provide alternative means of interaction. The proposed interaction techniques also had to be able to execute typical mouse actions, such as point-and-click. The usability of eye gaze and speech was determined using longitudinal user testing and a set of tasks specific to the functionality.

Results indicated that the use of a gravitational well increased the usability of the speech and eye gaze combination when used for pointing-and-clicking. The use of a magnification tool did not increase the usability of the interaction technique. The gravitational well did, however, result in more incorrect clicks due to natural human behaviour and the ease of target acquisition afforded by the gravitational well. However, participants learnt how to use the interaction technique over the course of time, although the mouse remained the superior pointing device.

Speech commands were found to be as usable, or even more usable, than the keyboard and mouse for editing and selection purposes, although navigation was hindered to some extent. For text entry purposes, the keyboard far surpasses eye gaze and speech in terms of performance as an input method as it is both faster and results in fewer errors than eye gaze and speech. However, even though the participants were required to complete a number of sessions and a number of text entry tasks per session, more practice may be required for using eye gaze and speech for text entry. Subjectively, participants felt comfortable with the multimodal interface and also indicated that they felt improvement as they progressed through their sessions. Observations of the participants also indicated that as time passed, the participants became more adept at using the multimodal interface for all necessary interactions.

In conclusion, eye gaze and speech can be used instead of a pointing device and speech commands are recommended for use within a word processor in order to accomplish common tasks. For the purposes of text entry, more practice is advocated before a recommendation can be made. Together with progress in hardware development and availability, this multimodal interface may allow the word processor to further exploit emerging technologies and be a forerunner in the use of multimodal interfaces in other applications.

Keywords: Multimodal interfaces, gaze-controlled interfaces, speech-controlled interfaces, eye-tracking, speech recognition, word processing, usability

OPSOMMING

Multi-modale koppelvlakke kan ’n betekenisvolle bydrae lewer tot grafiese gebruikerskoppelvlakke soos wat dit die afgelope tyd bekend was. Dit is ook moontlik dat multi-modale koppelvlakke ’n meer intuïtiewe en natuurlike interaksie-medium kan bied om die afhanklikheid van tradisionele handbeheerde interaksie tegnieke te verminder. Visie en spraak is alledaagse komponente van natuurlike mens-tot-mens kommunikasie en word in hierdie studie ook voorgestel vir gebruik in ’n multi-modale koppelvlak vir ’n gewilde woordverwerkingspakket.

Om lewensvatbaar te wees in die koppelvlak van ’n woordverwerkingspakket, moet ’n kombinasie van visie en spraak die invoer van teks, redigering asook formatering van ’n dokument, fasiliteer. Vir die doeleindes van hierdie studie is ’n beperkte stel mondelinge opdragte gebruik om alledaagse woordverwerkingsopdragte, sowel as die seleksie van teks en navigering in ’n dokument, te aktiveer. Met die oog op teksinvoer is ’n visuele sleutelbord op die skerm vertoon. ’n Sleutel kon geaktiveer word deur daarna te kyk en dan ’n gepaste opdrag te uiter. Hierdie funksionaliteite is in ’n aangepaste Microsoft Word 2007® woordverwerkingspakket geïmplementeer om die aanpasbaarheid en moontlik ook die bruikbaarheid van die woordverwerkings-koppelvlak te verhoog en om alternatiewe interaksietegnieke te voorsien. Die voorgestelde interaksietegnieke moes ook geskik wees om tipiese muis-aksies, byvoorbeeld wys-en-kliek, uit te voer. Die bruikbaarheid van visie en spraak is bepaal deur longitudinale gebruikerstoetsing en ’n stel take wat op spesifieke funksionaliteite betrekking het.

Die resultate het aangedui dat die gebruik van ’n gravitasieput die bruikbaarheid van die kombinasie van spraak en visie tydens wys-en-kliek aksies verhoog het. Die gebruik van ’n vergrotingspakket het nie die bruikbaarheid van die interaksietegniek verhoog nie. Natuurlike menslike gedrag en die gemak waarmee teikens gekliek kon word deur gebruik van ’n gravitasieput, het egter veroorsaak dat die gravitasieput meer foutiewe klieks tot gevolg gehad het. Deelnemers het egter mettertyd geleer om die tegniek te gebruik, alhoewel die muis steeds die beste wysertoestel gebly het.

Dit is verder bevind dat mondelinge opdragte net so goed of selfs beter is vir redigering en seleksie as die sleutelbord en muis, alhoewel navigering in ’n mate gekortwiek is. Die sleutelbord is verreweg die beste tegniek om teks in te voer aangesien dit vinniger was en deelnemers ook minder foute gemaak het as met spraak en visie. Alhoewel deelnemers ’n aantal take uitgevoer het tydens ’n hele paar sessies, mag meer oefening nodig wees om spraak en visie vir teksinvoer te gebruik. Subjektiewe terugvoer van deelnemers het aangedui dat hulle gemaklik was met die multi-modale koppelvlak en dat hulle ervaar het dat hulle van een sessie tot die volgende verbeter het. Dit is ook waargeneem dat deelnemers meer bedrewe geraak het met oefening en die multi-modale koppelvlak mettertyd vir al die nodige interaksies kon gebruik.

Ter opsomming is dit duidelik dat spraak en visie gebruik kan word in die plek van ’n wysertoestel en dit word aanbeveel dat mondelinge opdragte gebruik word om alledaagse woordverwerkingstake uit te voer. Dit is nodig dat deelnemers meer oefening in teksinvoer moet kry voordat ’n aanbeveling gemaak kan word. Hierdie multi-modale koppelvlak kan, in samehang met die ontwikkeling en beskikbaarheid van apparatuur, die woordverwerker toelaat om nuwe tegnologieë te ontgin en die weg te baan vir gebruik van multi-modale koppelvlakke in ander toepassings.