Using Vision and Voice to Create a Multimodal Interface for Microsoft Word 2007

T.R. Beelders and P.J. Blignaut
The University of the Free State, Bleomfontein, South Africa
{beelderstr; pieterb}@ufs.ac.za

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

1 Introduction

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, 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. Therefore, one aim is to determine whether it is possible to customize an interface to such an extent that all user groups are catered for with an all-inclusive interface.

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.

2 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.

3 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

ETRA 2010, Austin, TX, March 22 – 24, 2010.
© 2010 ACM 978-1-60558-994-7/10/0003 $10.00

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.

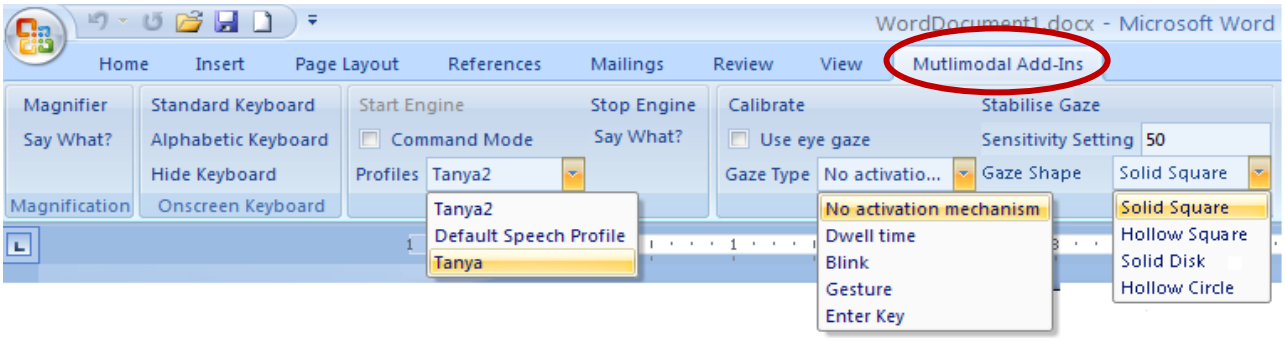

Figure 1: *Multimodal Add-ins for Word 2007*

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. Magnification can be toggled using either a verbal command or the available button on the ribbon.

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 clip-

board 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.

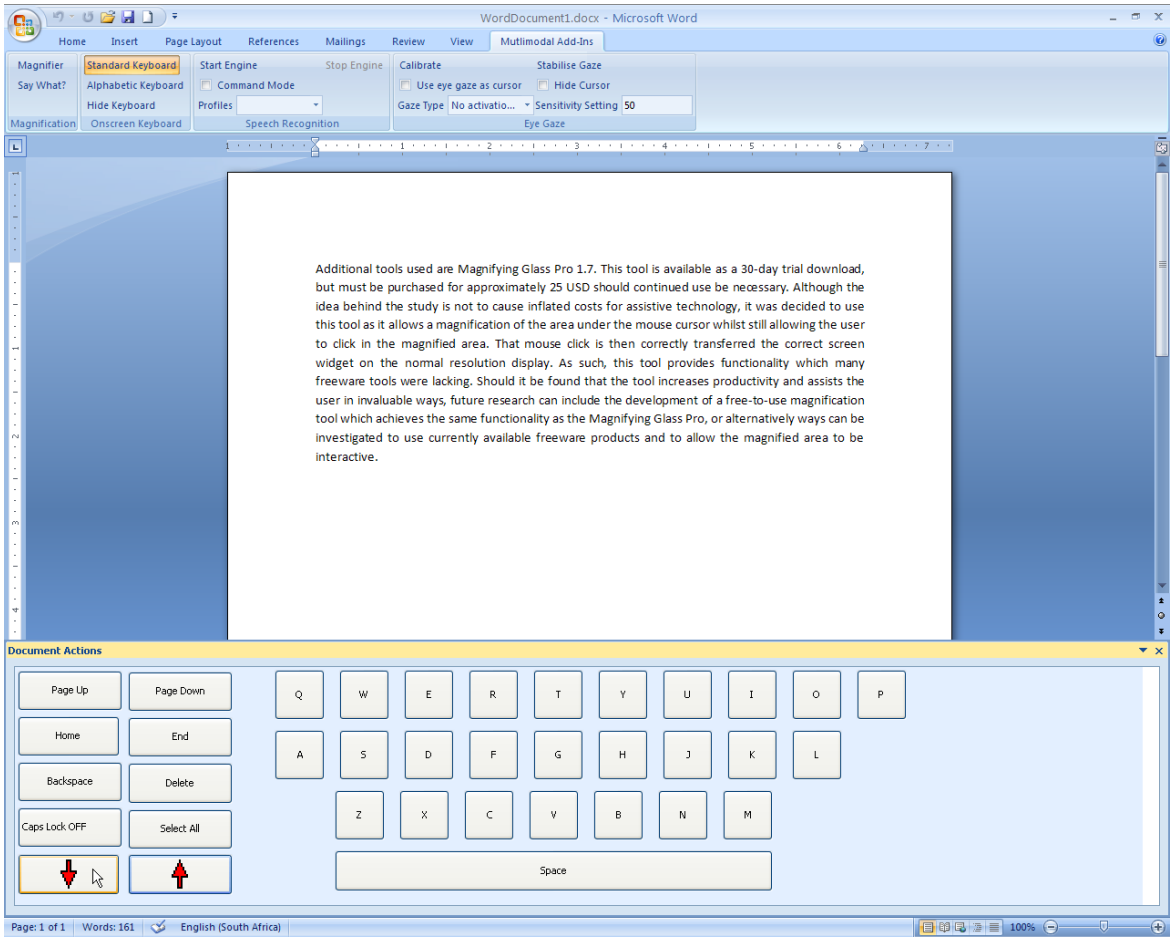

Figure 2: *Adapted interface of Word 2007 when the onscreen keyboard is activated*

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 on-screen 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.

4 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.

5 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 exteneded 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.