Measuring the performance of gaze and speech for text input

Article · March 2012

DOI: 10.1145/2168556.2168631

2 authors:

Tanya Beelders
University of the Free State
21 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Pieter Blignaut
University of the Free State
53 PUBLICATIONS **213** CITATIONS
SEE PROFILE

Measuring the performance of gaze and speech for text input

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

ABSTRACT

A popular word processor application was adapted to include the use of eye gaze and speech as a modality for text entry. An onscreen keyboard was used whereby users were expected to focus on the desired character and then issue a verbal command in order to type the character in the document. Measures of speed and accuracy were captured and analyzed. Results indicate that the keyboard is superior to the gaze and speech entry method in terms of both speed and accuracy. Keyboard button sizes and spacing between the buttons did not affect either measure in any way.

CR Categories: H.5 [Information systems]: User interfaces

Keywords: Eye-tracking, multimodal interface, text input.

1 INTRODUCTION

Communication between humans and computers is considered to be two-way communication between two powerful processors over a narrow bandwidth [Jacob and Karn 2003]. Most interfaces today utilize more bandwidth with computer-to-user communication than vice versa, leading to a decidedly one-sided use of the available bandwidth [Jacob and Karn 2003]. An additional communication mode will invariably provide for an improved interface [Jacob 1993] 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 and Karn 2003].

Eye gaze has already been successfully used for text entry in a variety of ways such as dwell time and gaze gestures (cf. [Hansen et al. 2001]; [Wobbrock et al. 2008]) as has speech recognition (cf. [Klarlund 2003]). The current study will include eye gaze as an input technique but will require the use of an additional trigger mechanism, namely speech, in order to determine whether the accuracy and speed of the text entry method can be increased in this manner.

2 BACKGROUND

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 and Tatler 2009]. Psycholinguistic studies have also shown that there is a temporal relationship between eye gaze and speech (cf. [Just and Carpenter 1976]; [Tanenhaus et al. 1995]), often referred to as the eye-voice span. 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 synchronize the two. However, it has also been found that for the majority of verbal requests, users were looking at the object of interest when the command was issued [Maglio et al. 2000].

In order to maximize the disambiguation of both modalities in this study, the user will be expected to maintain eye gaze on the desired object whilst issuing the verbal command to interact with that object.

In terms of data entry, eye gaze and speech recognition have been implemented, with great success, to complete a television license application form in the United Kingdom [Tan et al. 2003]. Another means of data entry is the RESER and SPELLER systems [Tan et al. 2003]. 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 recognize the word and offer a suggestion once it can recognize 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 aptly named Speech Dasher extends the capabilities of the original Dasher by including speech recognition [Vertanen and 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 and 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.

The current study built on the idea that eye gaze can 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. 2003] that speech could be used to activate the focused key. However, it also assumes that some users

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00

may have limited vocabularies and may not be able to vocalize all alphabetic letters. Therefore, a single command, which can be customized 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.

3 MULTIMODAL WORD PROCESSOR

Visual Studio Tools for Office (VSTO) allows developers to create extensions to the Office Suite with customized functionality [Anderson 2009]. Therefore, VSTO was used to add multimodal functionality to Microsoft Word® in the form of speech commands and direct eye gaze interaction.

Speech commands were used to facilitate navigation through, editing and formatting of document content. The results achieved with the speech commands when used for this purpose are beyond the scope of this article and will not be discussed any further.

Eye gaze was incorporated into the multimodal interface (see Figure 1) to facilitate hands-free typing in Word. An on-screen keyboard was available which was overlaid on the bottom of the current document. The size of the keys was adjustable to allow the user to select a key size which they are comfortable with. Additionally, users could choose between an alphabetic layout and a standard QWERTY layout. Users could choose an activation mechanism of dwell time, using the Enter key as a trigger, blinking or using a speech command. This article will focus on the use of speech as a trigger.

When typing using eye gaze and speech, the users were required to gaze at the desired key on the keyboard. The key which was currently under the gaze of the user was framed with a dark green border. If the user wished to type that letter in the document, then they could issue a speech command and the associated letter would be typed at the current position of the cursor. Auditory feedback, in the form of a beep, was given to alert the user that the character had been typed. This should allow them to continue typing without having to glance back at the document for confirmation.

In addition to the adjustable key sizes, there was also a magnification tool available which when activated magnified the area under the mouse or eye gaze. The magnified area was clickable and fully interactive so the user need not deactivate magnification before issuing a command.

Therefore, the multimodal interface was highly customizable and offered a variety of choices and settings to allow users to tailor the interface to their needs and preferences.

4 ANALYSIS

4.1 Participants

A total of 25 participants participated in the 10-week long study. The first week was simply an introductory session and the data collected there was not included in the analysis. Furthermore, the data of three participants had to be discarded from the sample due to various reasons. Of the remaining 22 participants, only 8 completed all sessions on the onscreen keyboard and 14 with the traditional keyboard. These were the participants who were included in the analysis.

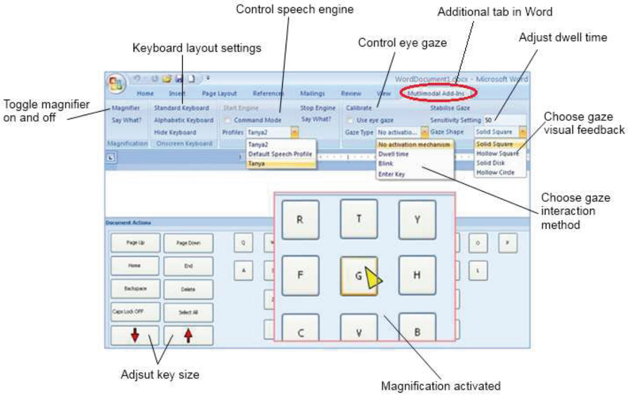

Figure 1 *Multimodal add-in for Microsoft Word*

5 Tasks

Each participant had one session per week during which time they were expected to complete a series of tasks on the adapted word processor. Three of these tasks were typing tasks with the onscreen keyboard and two with the traditional keyboard. For each session, the buttons of the keyboard were sized at 60×60 pixels (≈1.55° visual angle). Buttons were spaced 60 pixels apart with a gravitational well of 20 (≈0.52 ° visual angle) pixels on all sides of each button. The gravitational well effectively increased the selection area of each button since once the gaze was detected within the bounds of the gravitational well it was pulled onto the button. Participants were not aware of the gravitational well as it was not visible.

The typing tasks required the participant to type a phrase that was randomly selected from a set of 35 phrases. The phrase set used was a subset of the 500 as determined by [MacKenzie and Soukoreff 2002] to be everyday phrases which are commonly used.

6 Measures

Since both input methods, namely typing with the traditional and the onscreen keyboard, were character based, the measures were focused on characters. Therefore, the character error rate (CER) and characters per second (CPS) were selected as effectiveness and efficiency measures for the input techniques.

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. 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 and Soukoreff 2002]. 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 and Soukoreff 2002]. The mean length of the optimal alignments was used in this study.

The (CPS) measure 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.

7 Results of primary study

7.1 Character Error Rate

The mean CER was calculated for all sessions, for each participant. Chart 1 shows the mean CER for all sessions, where the solid blue line is for the eye gaze and speech and the red dotted line is for the traditional keyboard input.

Using the mean CER as depicted in Figure 2 it can be extrapolated that the speech interaction, on average, caused a higher error rate than the keyboard. This observation holds for all sessions, although the error rate for the speech interaction technique did improve steadily as the amount of exposure increased.

A repeated-measures ANOVA was used to determine whether there was a significant difference in the error rates of the keyboard or when using the eye gaze and speech. At an α-level of .05, it was found that there was a significant difference (F(1,19) = 14.406). It was also shown that the session had a significant effect on the error rate (F(8,152) = 5.092).

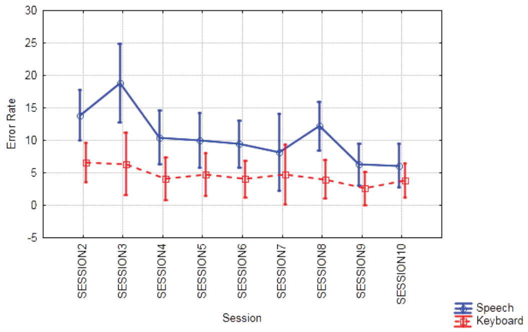

Figure 2 *Mean character error rate*

7.2 Characters per second

The line graph (Chart 2) below plots the mean characters per second for all sessions. The red line with the square indicator plots the speed of the traditional keyboard and the blue with the circular indicator plots the eye gaze and speech input.

From the graph 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 could be achieved using eye gaze and speech remained fairly constant throughout the sessions, displaying only mild improvement as the exposure increased. At an α-level of 0.05, there was a significant difference between the typing techniques (F(1,21) = 54.704, p < 0.05), with the keyboard resulting in a significantly higher typing speed. At the same α-level, there was no significant difference between the sessions (F(8, 168) = 1.385, p > 0.05).

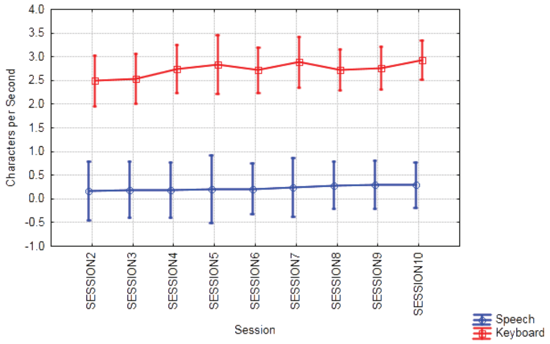

Figure 3 *Characters per second*

8 Discussion

It was found that the eye gaze and speech interaction technique had a significantly higher error rate than that of the keyboard. In terms of efficiency, the keyboard also outperformed 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. Therefore, the keyboard is far more effective and efficient than the eye gaze and speech interaction techniques when used for text input.

No similar studies were found with which these results could be compared. 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 for context switching was 12 WPM [Morimoto and

Amir 2010] and 9 WPM for symbol creator [Miniotas et al. 2003], the range in this study was much lower than these previous studies. A previous study [Majaranta 2009] showed that the use of both visual and auditory feedback increased the entry speed to 7.55 WPM which is still faster than the speeds achieved in this study. Speech Dasher achieved much higher speeds (40 WPM), quite possibly due to the inclusion of dictation capabilities. Therefore, character-based typing using eye gaze and speech is initially slower than these other techniques, but extensive practice could have an effect on its efficiency.

9 Future 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.

The use of eye gaze and speech for text entry can also be compared to other gaze modalities such as dwell time or blinking in order to determine its performance compared to these interaction techniques.

REFERENCES

ANDERSON, T. 2009. *Pro Office 2007 development with VSTO*. United States of America: APress.

HANSEN, J.P., HANSEN, D.W. AND 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.

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. AND 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.

JUST, M.A. AND CARPENTER, P.A. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8, 441-480.

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.

LAND, M.F. and TATLER, B.W. 2009. *Looking and acting: Vision and eye movements in natural behaviour*. United States of America: Oxford University Press.

LEVENSHTEIN, V.I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk*, 163, 845-848.

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. AND 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. AND 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. AND 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.

MINIOTAS, D., ŠPAKOV, O. AND 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.

MORIMOTO, C.H. AND AMIR, A. 2010. 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.

READ, J. 2005. On the application of text input metrics to handwritten text input. Text Input Workshop, Dagstuhl, Germany.

READ, J., MACFARLANE, S. AND 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.

TAN, Y.K., SHERKAT, N. AND ALLEN, T. 2003. 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.

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. AND SEDIVY, J. 1995. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268, 1632-1634.

VERTANEN, K. AND 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.

WOBBROCK, J.O., RUBINSTEIN, J., SAWYER, M.W. AND 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.