Using eye gaze and speech to simulate a pointing device

Article · March 2012

DOI: 10.1145/2168556.2168634

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

Using eye gaze and speech to simulate a pointing device

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

ABSTRACT

The performance of eye gaze and speech when used as a pointing device was tested using the ISO multi-directional tapping task. Eye gaze and speech were used for target selection as is, as well as with the use of a gravitational well and in conjunction with magnification. These selection methods were then compared to the mouse. The mouse was far superior in terms of performance when selecting targets, although the use of a gravitational well did increase the performance of eye gaze and speech. However, magnification did not improve the use of gaze and speech as a pointing device.

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

Keywords: Eye-tracking, multimodal interface, speech recognition, pointing device, ISO multi-directional tapping task.

1 INTRODUCTION

This paper will report on the results found when using the ISO multi-directional tapping test to compare the mouse with various eye gaze selection techniques. Eye gaze was combined with the use of a verbal trigger to select targets as a means to combat the Midas Touch problem. The first interaction technique tested was the use of eye gaze and speech only, the second employed the use of a gravitational well, the third used magnification. A gravitational well invisibly increases the size of the target by making the target selectable once it enters the area around the target. Magnification zooms the area directly under the gaze of the user to increase the size of targets.

Various measures were captured and analyzed to determine which interaction technique was the most usable. The following sections will discuss some background literature, the methodology and then the results of the study.

2 BACKGROUND

The most commonly used metrics to evaluate pointing devices are speed and accuracy [MacKenzie et al. 2001] which gives a good indication as to whether there is a difference between the performance of pointing devices [Hwang et al. 2004]. The International Standards Organization ratified a standard, ISO 9241-9, for testing the speed and accuracy of pointing devices for comparison and testing purposes. 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 multi-direction tapping test consists of a series of boxes placed around the circumference of a circle (see Figure 1 for an example). The participant is then required to move from the center 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 until all the targets have been clicked and the user is back at the first selected target box.

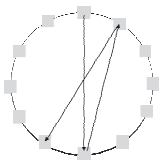

Figure 1 *Multi-directional tapping task*

One study which used the ISO 9241-9 to test eye-tracking as an input device 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 750 ms, (b) dwell time of 500 ms, (c) look-and-shoot which required participants to press the space bar to activate the target they were looking at and (d) the mouse [Zhai et al. 1999]. A head-fixed eye-tracking system with an infrared camera and a sampling rate of 30 Hz 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. The time required to press the space bar, particularly if users keep their hand on it, should be shorter than the dwell time, which was confirmed by the results of the aforementioned study [Zhai et al. 1999].

3 METHODOLOGY

3.1 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. In this study, variable size targets were used, but in order to reduce the time required to complete a test the distance between targets was not adjusted during this testing.

Standard Windows icons are 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. [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 visual angle of 0.85°. The eye-tracker used in the current study was a Tobii

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

T120 with a 17'' monitor where the resolution was set to 1280×1024. In order to replicate the visual 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. [Miniotas et al. 2006]. Following the example set by Miniotas et al. [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 used in this study had sixteen square targets situated on the edge of a circle with a diameter of 800 pixels.

Target acquisition was either via eye-tracking and speech recognition (denoted by ETS for the purpose of this paper) or the mouse (M). The mouse was used to establish a baseline for selection speed. When using a verbal command to select a target, the subjects had to say "go" out loud in order to select the target that they were looking at. This method of pointing could therefore be considered analogous to look-and-shoot.

Magnification (ETSM) and the use of a gravitational well (ETSG) can be used to combat various shortcomings of using eye gaze for target selection, namely the instability of the eye gaze and the difficulties experienced in selecting small targets. The default zoom factor for the magnification in this study enlarged the area to double its actual size within a 400×300 window while the gravitational well was activated within a distance of 50 pixels from the side of each button. This implies that when the eye gaze is detected within 50 pixels around a target, the target is acquired and is "clickable". The target button which had to be clicked was denoted by an "X". When a target received focus, visual feedback was given by framing the acquired button in a green frame. When a button was clicked, an audible click sound was played to inform the participant that the button had been clicked. A balanced Latin square [Edwards 1951] for all trial conditions was obtained and participants were randomly assigned to a Latin square condition for each session.

3.2 Participants

Participants were senior students at the university at which the study was conducted. Each participant completed three sessions and each session consisted of all trials. In total there were 15 participants who completed all three sessions. Eleven of the participants were male and 4 were female. The average age of the participants was 22.3 (standard deviation = 1.9). All participants could be ranked as having high computer expertise. Similarly, all participants ranked as having high mouse expertise and none had either eye-tracking or speech recognition experience.

4 RESULTS

4.1 Throughput

The following graph (Figure 2) depicts the resulting throughput for all interaction techniques and all sessions. The interaction techniques are the mouse (M), eye gaze and speech (ETS), ETS with magnification (ETSM) and with a gravitational well (ETSG). The mouse is the topmost line on the graph, ETSG is just below that, followed by ETS and finally ETSM.

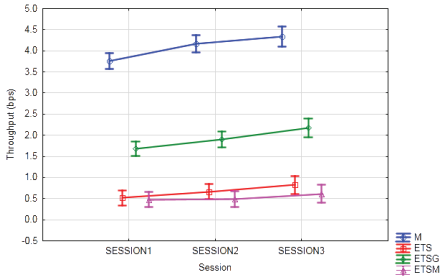

Figure 2 *Throughput for all interaction techniques*

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. This analysis showed that there was a significant difference in the throughput of the different interaction techniques for all sessions. Post-hoc tests identified that it was only ETS and ETSM that did not differ significantly from each other. Therefore, it could be concluded that the mouse had a significantly higher throughput than all the other tested interaction techniques.

For all eye gaze and speech interaction techniques there was a significant difference between the sessions in that the first session and second session differed significantly from the third session. 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 and MacKenzie, 2004]. Therefore it could be said that the observed values correspond to the expected values. 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.

4.2 Time to complete trial

The next measurement to be analyzed was the time taken to complete the trial. Although throughput includes both speed and accuracy it seemed prudent to analyze the time taken to complete the trials separately. This is especially important since some of the interaction techniques allowed for larger "clickable" areas, which were not visible to the participant. This effectively means that the target could be selected without the eye gaze actually being positioned precisely on the button (the eye gaze was not snapped to the center of 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. The graph below (Figure 3) shows the time taken to complete each trial for all sessions. In this instance, ETSM is the topmost line on

the graph, followed by ETS, ETS and finally the mouse as the bottommost line.

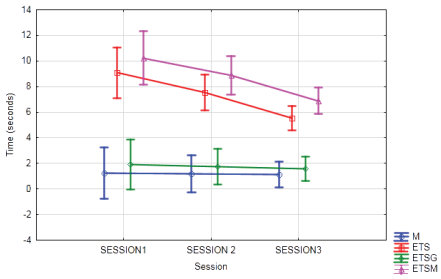

Figure 3 *Time to complete trial for all interaction techniques*

Similar to the throughput, the mouse and the use of the gravitational well resulted in the better performance for all sessions. There does, however, appear to be improvement in the performance with the other two interaction techniques from one session to the next.

At an α-level of .05, the difference between the interaction techniques was significant ($F(3, 53) = 305.767$). Subsequently it could be concluded that the interaction technique significantly affected the time required to complete the trials. Similarly, at a significance level of .05, the trial session significantly affected the time required to complete the trial ($F(2, 106) = 24.128$).

Post-hoc tests were conducted in order to determine which of the sessions and interaction techniques contributed to the significant difference. From these it could be concluded that the longer the user was exposed to the interaction technique, the more they learnt to use the device and the faster they were able to complete a point-and-click trial.

The times achieved with the mouse differed significantly from all other techniques, with the mouse being notably faster than the other interaction techniques. ETSG also differed significantly from the other interactions while ETS and ETSM did not differ significantly from each other. Accordingly, while the presence of a gravitational well significantly enhanced the performance of ETS, the magnification of targets did not.

The following graph (Figure 4) plots the time to selection for the consolidated interaction techniques. Time to selection was measured as the time between when the final target acquisition was performed and when the target was actually clicked or selected. The final target acquisition was defined as the last time the button received focus before being clicked. The mouse (bottommost line) appears to have the fastest time to selection while the remaining three interaction techniques are virtually indistinguishable from one another.

At an α-level of .05, there was a significant difference between the interaction techniques ($F(3, 54) = 196.605$) but not between the sessions ($F(2, 108) = 0.543$).

Post-hoc tests indicated that the mouse differed significantly from all other techniques. The selection time for the mouse was, on average, lower than those for the other interaction techniques; therefore the selection time for the mouse was 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 was 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.

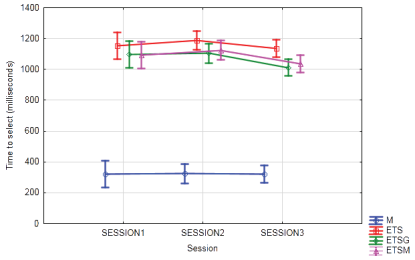

Figure 4 *Time to selection for all interaction techniques*

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 graph below (Figure 5) clearly shows that, on average, ETSG had a lower final acquisition time than the mouse.

For overall time to target selection, the mouse was significantly faster than ETSG. However, when selection time was divided into final target acquisition and time to selection, it was found that ETSG had a significantly faster final target acquisition but a significantly slower time to selection. Therefore, the time to selection was so much slower that the overall time differed significantly. It would now seem that final target acquisition would have to improve dramatically to achieve better performance with ETSG. Since acquisition times did improve over time, this remains a viable possibility for improved overall selection times.

5 Discussion

The mouse had a significantly higher throughput than the other interaction techniques. The use of a gravitational well caused a significant improvement to the throughput of eye gaze and speech as an interaction technique. However, magnification did not positively influence the throughput of eye gaze and speech as an interaction technique.

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

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 was perhaps the interaction technique which most closely resembles the expectations of the user in terms of their perceived behavior. The gravitational well also inspired more confidence in the users as they are unaware of the larger selectable area but they were achieving the desired results with minimal effort. It also allowed for a more aesthetically pleasing interface as the widgets were kept to a smaller size, although they had to be spaced further apart to make provision for the gravitational well.

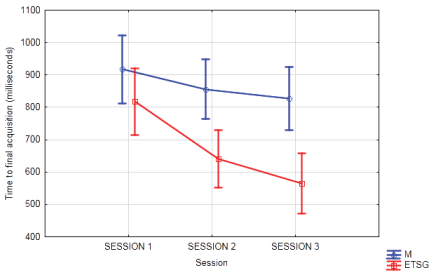

Figure 5 *Final acquisition time for mouse and ETSG*

These findings to an extent confirm 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 required that the user fine-tune the selection of the target within a magnified area. However, this still facilitated 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.

In terms of comparison with previous studies, not many previous studies comparing eye gaze selection with the ISO test were found and none on the scale of the current study. 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) [Zhang and MacKenzie 2007]. 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 and MacKenzie 2007] was the most appropriate. If one assumes that target acquisition will be similar then the speech takes twice as long as dwell time. It could therefore be concluded that using speech may be less efficient than using dwell time although studies must be conducted to verify this.

6 Future research

Further research can be conducted for interaction techniques using the ISO pointing device test. Future experimental setups can rather concentrate on changing the distance between targets and the size of targets. Additionally, other selection means may be considered as a way to counteract the significantly slower time to selection of the speech commands. Furthermore, 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.

REFERENCES

ASHMORE, M., DUCHOWSKI, A. AND SHOEMAKER, G. 2005. Efficient Eye Pointing with a Fisheye Lens. In *Proceedings of Graphics Interface 2005*, 203-210.

CARROLL, J.M. 2003. *HCI Models, theories, and frameworks: Towards a multidisciplinary science*. San Francisco: Morgan Kaufmann.

EDWARDS, A.L. 1951. Balance Latin-square designs in psychological research. *The American Journal of Psychology*, 64(4), 598-603.

HWANG, F., KEATES, S., LANGDON, P. AND 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.

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.

MACKENZIE, I.S., KAUPPINEN, T. AND SILFVERBERG, M. 2001. Accuracy measures for evaluating computer pointing devices. In *Proceedings of SIGCHI '01*, Seattle, Washington, United States of America, 9-16.

MINIOTAS, D., ŠPAKOV, O., TUGOY, I. AND 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.

SOUKOREFF, R.W. AND 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.

ZHAI, S., MORIMOTO, C. AND 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, X. AND MACKENZIE, I.S. 2007. Evaluating eye tracking with ISO 9241 – Part 9. In J. Jacko (Ed.), *Human Computer Interaction*, 779-788.