EXTENDING THE LIMITS FOR GAZE POINTING THROUGH THE USE OF SPEECH

Darius Miniotas[1], Oleg Špakov[1], Ivan Tugoy[1], I. Scott MacKenzie[2]

[1]*Unit for Human-Computer Interaction, University of Tampere*
FIN-33014 Tampere, Finland

[2]*Department of Computer Science, York University*
Toronto, Ontario M3J 1P3, Canada

Abstract. Eye trackers have been used as pointing devices for a number of years. Due to the inherent limitations in the accuracy of eye gaze, however, interaction has been limited to targets that are at least one degree of visual angle in size. Consequently, targets in today's gaze-based interfaces have sizes and layouts quite distant from what is perceived as "natural settings". To cope with the accuracy constraints, we developed a multimodal pointing technique combining eye gaze and speech inputs. The technique was tested in a user study on pointing at multiple targets. Results suggest that pointing accuracy is 93% for targets subtending 0.85 degrees and 0.3-degree gaps between them. User performance is thus shown to approach the limit of practical pointing. Effectively, developing a user interface that supports the hands-free style of interaction and has a design similar to that of today's common interfaces seems a feasible task.

Keywords: multimodal input, eye movements, eye tracking, speech recognition, pointing, human performance.

1. Indroduction

Within the community of HCI researchers and system designers, developing an efficient user interface alternative to the traditional manually operated interfaces has been a major challenge for a number of years. Such an interface should not be dependent solely on inputs from the keyboard and conventional pointing devices such as a mouse. Instead, the interface should be able to employ as inputs other, more natural, communication abilities of the user. Speech, gestures and eye gaze are considered most frequently as candidates for the new type of interface. Despite the fact that each of these inputs alone is inherently ambiguous, interaction can still be made feasible by combining two or more inputs in an appropriate way.

Among the options used in novel designs, speech and eye gaze became the most popular couple. This could be attributed to the strong synergetic effect obtained through combining these two input modalities. With eye gaze employed in spatial location of objects and speech as the entry mode for commands, a fully functional input device can be built. Indeed, several workers demonstrated that integrating eye tracking and speech recognition technologies allowed achieving a reasonable amount of hands-free control over a graphical user interface [2, 5]. Practical application of such multimodal interfaces, however, still presents a challenge as described below.

In the field of eye gaze-based interfaces, there were some successful implementations manifesting the ability of eye gaze to function as a pointing device [1]. Nevertheless, the design of those user interfaces renders them quite distant from what is perceived as "natural settings" (i.e., today's standard GUIs with their widgets). One of the major differences is the size of objects interacted with.

Most of the standard GUI widgets (e.g., icons in a toolbar, checkboxes, etc.) are less than one degree of visual angle in size. For instance, a toolbar's icon in a standard MS Windows™ application (e.g., MS Word™) is 24 by 24 pixels in size, which translates to approximately 0.7 degrees for a 17-inch monitor with a resolution of 1024 x 768 and a viewing distance of 70 centimeters. Meanwhile, the size of a button in a window's title bar is even smaller (only 16 by 16 pixels, or 0.46 degrees). Moreover, icons in a toolbar are usually aligned side by side – there are no spaces between them.

From the traditional viewpoint of applied eye tracking research, however, targets below the one-degree limit are considered too small for facile eye gaze interaction [1, 3]. Consequently, gaze-operated objects are made substantially bigger to ensure facile interaction (i.e., bring gaze pointing to the level of practical accuracy). This measure is undertaken to accommodate the calibration errors of the eye tracker as well as the inherent limitations in the accuracy of eye gaze.

For the same reason, objects are also spaced on the screen at relatively large distances one from another. In turn, this poses problems in managing the real estate of the screen.

Given the constraints on the accuracy of gaze-based pointing, it is intriguing to explore to what extent user performance could be pushed towards the level of practical pointing when eye gaze is supplemented with other input modality such as speech. Despite the high interest of recent workers in multimodal applications, there are only few empirical studies aiming to evaluate user performance in multimodal pointing tasks.

Recently, Zhang et al [5] experimented with a multimodal system involving eye gaze and speech. Their setup included a 6 x 5 grid of geometric figures used as targets to be selected. The figures varied in shape (rectangle, oval, and triangle), size (two levels), and color (10 levels). The size of the smaller figures was 13 x 9 mm (1.1 x 0.74 degrees of arc at a viewing distance of 70 cm).

This presents an interesting case as the target's size approaches the critical one-degree barrier. On the other hand, the distance between the centers of adjacent targets in the grid was substantially bigger than that: 40 mm (3.27 degrees) horizontally and 27 mm (2.19 degrees) vertically. In turn, this makes the overall layout used in [5] not very suitable for modeling interactions similar to those present in conventional GUIs.

To obtain a more relevant model, we developed a gaze-based interface featuring tightly spaced targets reasonably close in size to that of the smallest GUI widgets. To meet the challenge of pointing at targets smaller than the one-degree limit, eye gaze input was augmented by speech.

This paper presents an experiment conducted to compare user performance in a point-select task using two modes of interaction: unimodal (i.e., gaze-only) and multimodal (gaze and speech).

2. Method

2.1. Participants

Twelve unpaid volunteers (6 male, 6 female) participated in the study. All were employees at a local university aged 22 to 43. All but one had prior experience with eye tracking, whereas only one of them had ever used speech as computer input before. One participant specified English as her first language, whereas all the rest were non-native speakers of English. Six participants wore glasses, whereas the other six required no correction of vision.

2.2. Apparatus

A remote eye tracking system *iView*X™ from SensoMotoric Instruments was used for collecting gaze data. Eye gaze input and associated events were recorded using experimental software developed in our laboratory. Speech input was recorded with a conventional microphone and processed using Microsoft SAP Interface 5.0.

2.3. Procedure

Participants were seated at a viewing distance of approximately 70 cm. The experiment used a point-select task. At the onset of each trial, a 30-by-30-pixel home box appeared on the screen (Figure 1). In motor space, however, the home box was expanded to 100 pixels on each side to facilitate homing.

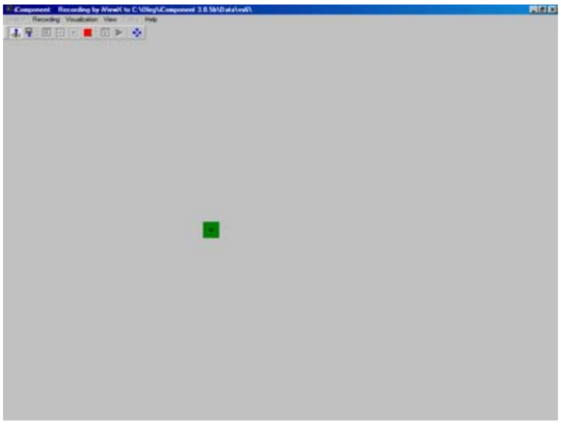


Figure 1. Home box at the onset of trial

Upon fixating on the home box for one second, a matrix of 5 x 5 squares appeared to the right of the home box (Figure 2). One of the squares was the target to be selected (marked with a cross). Participants were instructed to look at the target as quickly as possible (timing started), and fixate upon it until selection (timing ended). A window of five seconds was given to complete a trial. If no selection occurred within five seconds, an error was recorded. Then, the next trial followed.

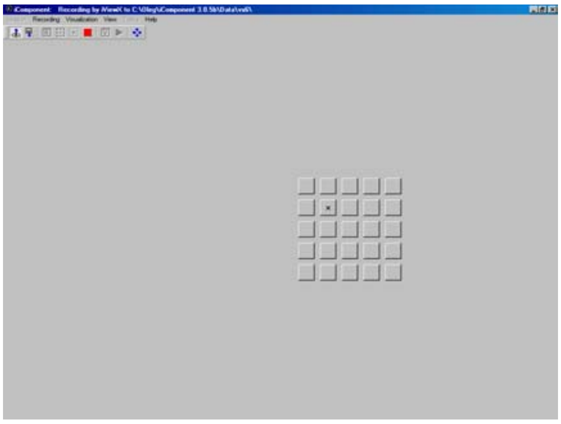


Figure 2. Matrix of 5 x 5 squares. The target for selection is the square marked with a cross

We defined the eye's region of interest (ROI) representing the focus of visual attention as a 100-by-100-pixel square with its center attached to the current gaze point location. As the gaze approached the target, the ROI began to overlap with the matrix area. The squares within the matrix that were encompassed by the overlapping area became highlighted in different colors (Figure 3).

The color-coding scheme included fifteen colors listed in the following order: red, green, blue, yellow, purple, aqua, orange, brown, pink, lime, gray, olive, magenta, sky-blue (vocally referred to as "sky"), and black. The coding was arranged so that the first color in the list (i.e., red) was assigned to the first square in the matrix to enter the ROI. Then, the second square encompassed by the ROI was highlighted in green, and so on. If more than fifteen squares got into the ROI (this quite often being the case for the smallest target size used in the experiment), only the first fifteen were highlighted in corresponding colors, whereas the remaining ones stayed unchanged.

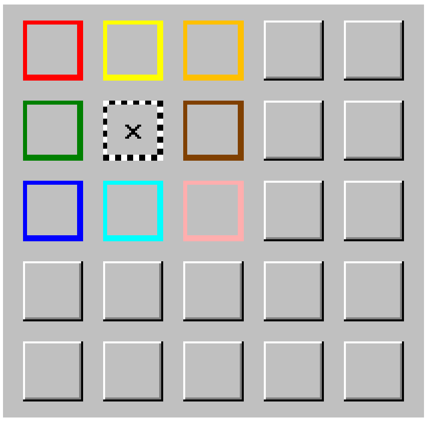


Figure 3. Highlighted squares signaling overlap of the eye's region of interest with the matrix. The black dashed outline shows the current gaze point location

Moreover, the color-coding scheme used was tolerant to instabilities in the gaze point location caused by the inherent eye jitter (see Jacob, 1991). As the ROI is centered on the current gaze point, random shifts in the spatial location of the ROI are also inevitable. In turn, this would cause the squares in the matrix looked at to flicker in different colors, were no measures taken.

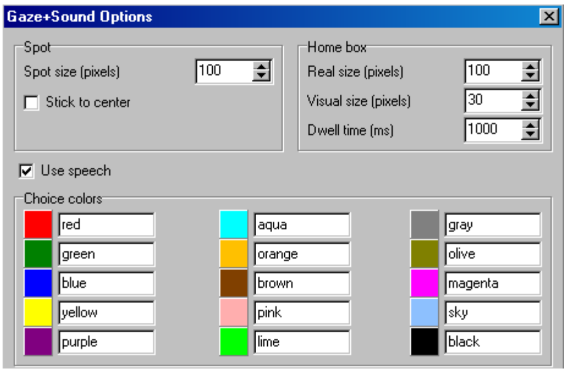
To avoid this, the same color stayed with a target for the rest of the trial once mapped initially as long as the attention was not shifted to other areas of the screen (i.e., no saccade – sudden motion of the eye – occurred in between). If at any point during the trial the attention was directed away from the current selection of the matrix squares, the squares were de-highlighted releasing the colors for subsequent selections.

No visual feedback was provided for the gaze point unless it entered the matrix. Whenever the gaze point landed on a square in the matrix, the square was highlighted with black dashed outline (Figure 3).

Prior to the first session, participants were shown a table displaying all the fifteen colors used in the color-coding of the matrix squares (Figure 4). They were asked to memorize the colors along with their names needed for the experimental condition involving speech input. After this initial introduction, participants practiced one block using speech commands. Then, data recording began.

Participants were given an opportunity to look at the table with the colors to refresh their memory when needed before a block of trials started.

The strategies to be used by participants for target selection depended on the available input modalities. In the combined gaze and speech condition, if the square with the dashed outline was other than the target, participants were to say aloud the color of the target's highlight.



Gaze+Sound Options ☒

Spot
Spot size (pixels) 100
□ Stick to center

Home box
Real size (pixels) 100
Visual size (pixels) 30
Dwell time (ms) 1000

☑ Use speech

Choice colors
red
green
blue
yellow
purple
aqua
orange
brown
pink
lime
gray
olive
magenta
sky
black

Figure 4. Table shown to participants with the fifteen colors used in the color-coding of the matrix squares

This way they were given an opportunity to compensate for the inherent limitations in the accuracy of eye gaze, as well as the drift in the eye tracker's calibration encountered most of the time. Meanwhile, in the gaze-only condition, participants could do very little to prevent an erroneous selection if the.

2.4. Design

The experiment was a 2 x 3 x 3 x 3 x 9 repeated measures factorial design. The factors and levels were as follows:

Pointing modality	gaze & speech, gaze-only
Dwell time (*DT*)	1000, 1500, 2000 ms
Target size (*S*)	20, 30, 40 square pixels
Inter-target gap (*G*)	0, 10, 20 pixels
Trial	1, 2…9

Here, *G* denotes the gap between the sides of adjacent squares in the matrix.

Participants were randomly assigned to one of three groups. Each group received the dwell time conditions in a different order using a Latin square. Order of presenting the pointing modality conditions was also counterbalanced among participants.

For each *DT* condition, participants performed 6 blocks of trials (3 blocks per modality) in one session. The three sessions were run over consecutive days with each lasting approximately 20 minutes. Each block consisted of the 9 *S-G* conditions presented in random order. For each *S-G* condition, 3 trials were performed in the same block (in total, 3 trials x 3 blocks = 9 trials). Thus, a block consisted of 27 trials. The conditions above combined with 12 participants resulted in 5832 total trials in the experiment.

The dependent measures were movement time (*MT*) and error rate (*ER*).

3. Results

3.1. Pointing Performance

The grand means on the two dependent measures were 3029 ms for *MT* and 34.3% for *ER*. The main effects and interactions on each dependent measure are presented below.

3.1.1. Speed

The mean *MT* was 3449 ms in the gaze-only condition and 2609 ms in the gaze & speech condition. Thus, with addition of speech, *MT* decreased on the average by 24%. The difference was statistically significant ($F_{1,11} = 45.3, p < .001$).

As expected, the 1000-ms *DT* condition was the fastest with a mean *MT* of 2605 ms. The 1500-ms *DT* condition was slower by 17% (3041 ms), and the 2000-ms *DT* condition by 32% (3442 ms). The main effect for *DT* was statistically significant ($F_{2,22} = 94.4$, $p < .001$), as was the input modality x *DT* interaction ($F_{2,22} = 5.9, p < .01$). The main effects and interaction are illustrated in Figure 5.

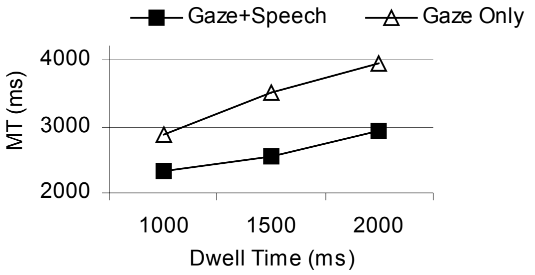


Figure 5. *MT* vs. *DT* for the two input conditions

As seen in Figure 6, the target size also significantly influenced pointing time ($F_{2,22} = 63.3$, $p < .001$). The input modality x target size interaction was significant as well ($F_{2,22} = 25.0$, $p < .001$). For the largest target (40 x 40 pixels), *MT* was on average 3132 ms in the gaze-only condition, whereas with addition of speech it dropped to 2515 ms (a reduction by 20%). As expected, the benefit of combined input was the highest for the smallest target (20 x 20 pixels): 3793 ms vs. 2741 ms (a reduction in *MT* of 28%).

3.1.2. Accuracy

The mean *ER* was 51.1% in the gaze-only condition and 17.4% in the gaze & speech condition. Thus, with addition of speech, *ER* dropped on the average by as much as 66%. The difference was statistically significant ($F_{1,11} = 48.8, p < .001$).

The lowest error rate was in the 1500-ms condition (32%). It was followed by the 2000-ms condition at 35.2% errors, and the 1000-ms condition at 35.6%. The differences were not significant ($F_{2,22} = 0.8$, ns). The input modality x *DT* interaction, however, was significant ($F_{2,22} = 5.8$, $p = 0.01$). In the gaze-only condition, more errors occurred as dwell time increased (Figure 7). With addition of speech, however, error rate decreased markedly as dwell time increased from 1000 ms to 1500 ms, and then remained at the same level with a further increase in dwell time by 500 ms.

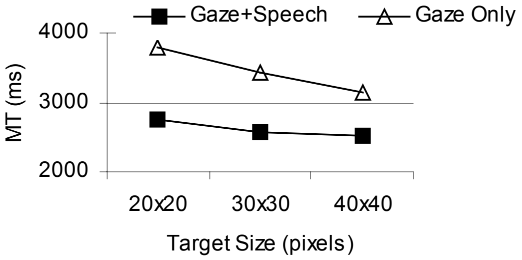


Figure 6. *MT* vs. target size for the two input conditions

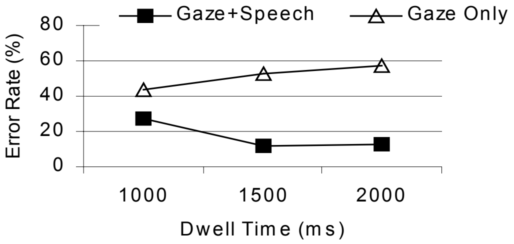


Figure 7. *ER* vs. *DT* for the two input conditions

As with pointing time, target size had a significant effect on error rate, too ($F_{2,22} = 77.6$, $p < .001$). The input modality x target size interaction was significant as well ($F_{2,22} = 27.3$, $p < .001$). For the largest target (40 x 40 pixels), error rate was on average 34.2% in the gaze-only condition, whereas with addition of speech it dropped to 12.1% (a reduction by 65%). For the two smaller sizes, a similar reduction in error rate was observed with speech employed (Figure 8).

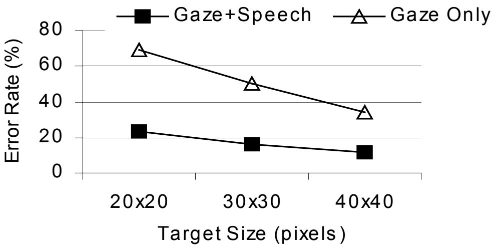


Figure 8. *ER* vs. target size for the two input conditions

In the combined input condition, inter-target gap also significantly affected error rate ($F_{2,22} = 14.3$, $p < .01$). It is not surprising that pointing accuracy was relatively poor when targets were side by side (0-pixel gap). Interestingly, however, there was no significant difference between the error rates obtained for the 10-pixel and 20-pixel gap conditions (Figure 9). In other words, for the purpose of practical pointing, a 10-pixel gap between targets is almost as good as a gap twice that size.

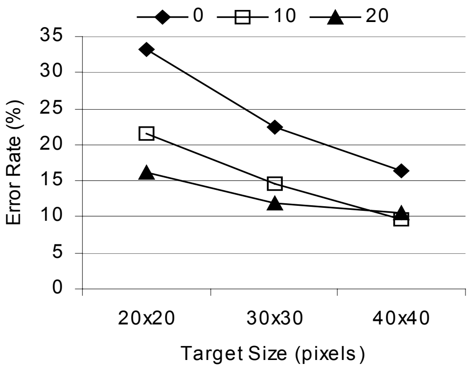


Figure 9. *ER* vs. target size and inter-target gap (in pixels)

A further insight into the extended limits of pointing accuracy with speech-augmented eye gaze input can be obtained when error rate is plotted as a function of target size and inter-target gap for the three *DT* conditions separately (Figure 10). When the shortest dwell time (1000 ms) was used for target selection, error rates for different combinations of target size and inter-target gap levels ranged from 15% to 50%. Error rates obtained for the other two *DT* conditions are much lower and do not significantly differ from one another between the conditions. They range from 2% to 26% and from 7% to 24% for the 1500-ms and 2000-ms conditions, respectively.

3.2. Performance by Colors

Some additional information on the factors contributing to the occurrence of the errors in the present study can be obtained by taking a closer look at the performance of the color-coding scheme for target identification.

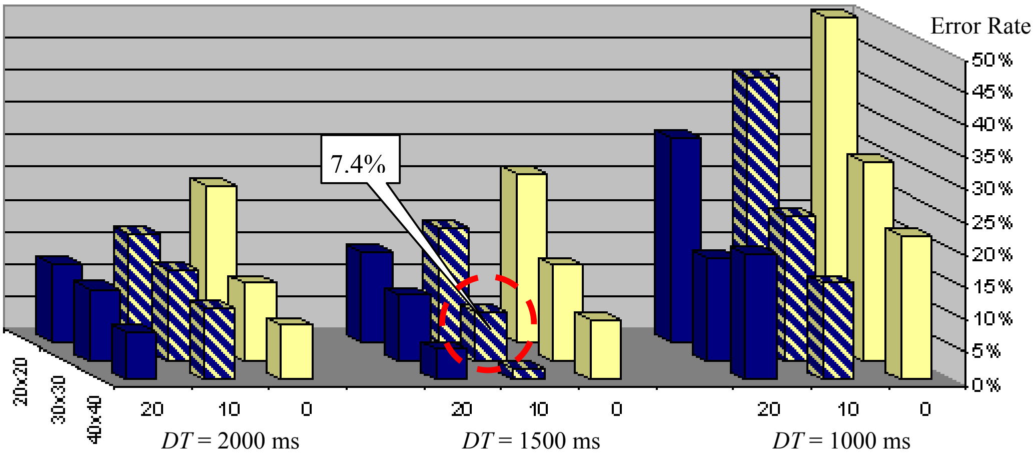


Figure 10. Error rate vs. dwell time, target size, and inter-target gap for the gaze & speech condition

The five colors most frequently mapped to the target for selection were: green, blue, yellow, purple, and aqua. They each had a share of over 8% with the total share equal to 51.7%. The remaining 48.3% were split between the other ten colors (Figure 11).

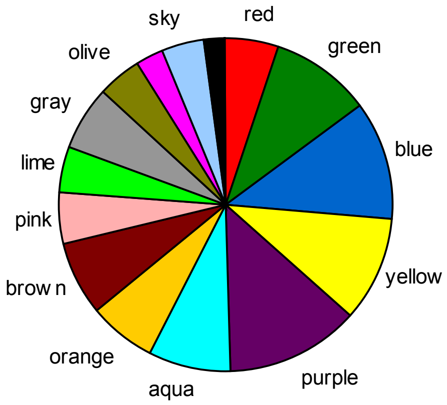


Figure 11. Distribution of target highlight colors

The percentage of the target highlight colors correctly specified by participants and recognized by the system varied among the colors. The colors with the correct selection rate above 90% were: green, blue, yellow, orange, brown, lime, sky-blue, and black (Figure 12). On the other hand, the most problematic colors for the international pool of our participants were magenta (65.1%) and purple (64.3%).

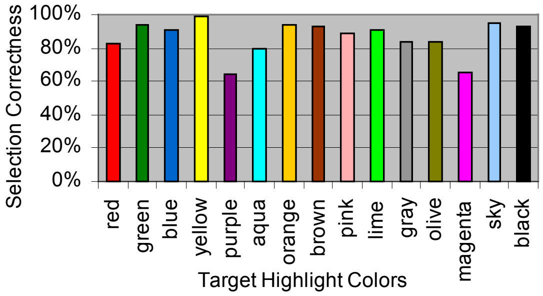


Figure 12. Correct target selection rate as a function of its highlight color

Out of all the cases involving erroneous identification of magenta, pink had the biggest share. This is an indication that for participants the less-familiar color magenta associated with the better-known pink. Meanwhile, purple was most frequently mistaken for blue or brown. Moreover, we have noticed that the false identifications of purple were quite often due to the speech recognition errors, as opposed to the actual vocal input from participants.

These observations demonstrate that there are areas of improvement for the color-coding scheme. In our future designs we will be more careful when selecting the colors to be used in the scheme, so that the challenges for the cognitive and vocal abilities of international users are minimized.

4. Conclusions

Our results suggest that the best performance (in terms of speed-accuracy tradeoff) can be expected using the following combination of the factor levels: 1500-ms dwell time, 30-by-30-pixel target size, and 10-pixel inter-target gap (shown by dashed circle in Figure 10). In our study, this combination yielded an average error rate of 7.4% (a mean of 108 trials in total: 12 participants x 9 trials). To better visualize the geometry on the real scale, Figure 13 displays a fragment of the experimental setup (the lower part) placed next to the actual GUI controls in the experimental software's window (the upper part).

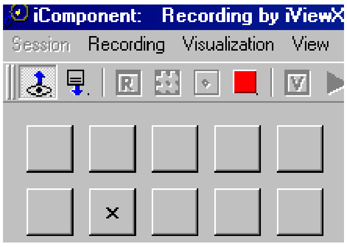


Figure 13. The size and layout suggested for buttons in a gaze-and-speech interface compared to a toolbar's buttons in a common manually operated GUI

This is a very important finding since user performance in a gaze-based selection task is shown to approach the limit of practical pointing. Moreover, the finding is consistent with the level of accuracy reported for the gaze-assisted manual pointing [4]. In effect, it means that in terms of accuracy there are no fundamental limits for combined gaze and speech input to become an alternative pointing technique just as good as manual pointing with devices such as an isometric pointing stick in notebook computers [4].

The major shortcoming of the speech-augmented gaze pointing technique presented in this study is relatively low speed. To match the cognitive demands largely associated with recalling the target's referential attribute (color in the current implementation) and then producing vocal output, dwell time for selection had to be increased substantially.

According to our data, accuracy becomes satisfactory when dwell time reaches 1500 ms. That is, of course, in sharp contrast to the common setting for the gaze-only modality, which is of the order of a few hundred milliseconds. The cost in speed, however, is offset by a dramatic reduction in error rate. As this study shows, that employment of speech allows bringing the overall error rate down by almost two thirds compared to the outcome for pointing by eye gaze alone.

By improving the scheme for target coding, we expect to be able to significantly reduce dwell time while maintaining pointing accuracy at the level currently achieved. In turn, this will allow improving the overall speed-accuracy tradeoff.

Another important issue is an adequate definition for the extent of the eye's region of interest (ROI). In the current implementation, we used a fixed value of 100 by 100 pixels for all target sizes. Intuitively, however, the extent of the ROI should depend on target size: the smaller the target, the smaller the region should be to accommodate the same number of objects within the region (in other words, to keep the probability of erroneous selection at the same level). We intend to find the best solution for defining the ROI in our future studies.

Acknowledgements

This research was funded by grant #103174 from the Academy of Finland, the UCIT Graduate School, and CIMO (Finland).

References

[1] **R.J.K. Jacob.** The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems* 9, 1991, 152-169.

[2] **D.B. Koons, C.J. Sparrell, K.R. Thorisson.** Integrating simultaneous input from speech, gaze, and hand gestures. *In M.T. Maybury (Ed.), Intelligent Multimedia Interfaces. MIT Press,* 1993.

[3] **C. Ware, H.H. Mikaelian.** An evaluation of an eye tracker as a device for computer input. *Proc. Conference on Human Factors in Computing Systems and Graphics Interfaces (CHI + GI 1987). ACM Press,* 1987, 183-188.

[4] **S. Zhai, S. Conversy, M. Beaudouin-Lafon, Y. Guiard.** Human on-line response to target expansion. *Proc. Conference on Human Factors in Computing Systems (CHI 2003). ACM Press,* 2003, 177-184.

[5] **Q. Zhang, A. Imamiya, K. Go, X. Mao.** Resolving ambiguities of a gaze and speech interface. *Proc. Eye Tracking Research and Applications Symposium (ETRA 2004), ACM Press,* 2004, 85-92.