Evaluating Multimodal Interaction with Gestures and Speech for Point and Select Tasks

Alvin Jude
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
alvin_jude@baylor.edu

G. Michael Poor
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
michael_poor@baylor.edu

Darren Guinness
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
darren_guinness@baylor.edu

NordiCHI '14, Oct 26-30 2014, Helsinki, Finland
ACM 978-1-4503-2542-4/14/10.
http://dx.doi.org/10.1145/2639189.2670267

Abstract

Natural interactions such as speech and gestures have achieved mainstream success independently, with consumer products such as Leap Motion popularizing gestures, while mobile phones have embraced speech input. In this paper we designed an interaction style that combines both gestures and speech to evaluate point and select interaction. Our results indicate that while gestures are slower than the mouse, the introduction of speech allows for selection to be performed without negatively impacting navigation. We also found that users can adapt to this interaction quickly and are able to improve their performance with minimal training. This lays the foundation for future work, such as mouse replacement technologies for those with hand impairments.

Author Keywords

Multimodal Interaction; Gestural Interaction; Speech Input;

ACM Classification Keywords

H.5.2 [User Interfaces]: Evaluation/methodology, Input devices and strategies

Introduction

Paired speech and gestural interaction has potential as a next-generation user interaction technique. Both are considered natural forms of interaction, and previous user studies demonstrated the combination was preferred over each modality alone [1].

However, interest in gesture and speech interactions have declined since the 1990's due to a few drawbacks, including: (1) Gestures were deemed too hard for automatic processing systems, (2) Technical implementations that are over-simplified tend to end up rigid and brittle [3] and (3) speech-based input makes it difficult to multitask as it is difficult for people to think while speaking [5]. This causes higher error rates when using speech as input, leading to higher resistance from users. It was proposed that speech-based interaction would need more realistic goals, and better models for multitasking to allow better user-acceptance.

We designed an interaction method with the above problems in mind. Our interaction allowed the user to perform either speech-only, or gestures-only at any given time without the explicit need to perform both. We hypothesized that this would result in a better multitasking model and would allow speech and gesture multimodal interaction to be better accepted and used in practice.

Related Works

A common issue with speech based input is hyperarticulation [4]. This refers to a stylized and clarified form of pronunciation following a recognition failure. It tends to cause a cycle of recognition errors due to the difference from the original training data. This problem is usually observed in commands that have more than one word. Meanwhile, speech processors only need to understand a small number of total words, since most commands only require an utterance of 3 words [1]. We designed our commands to have fixed 2-word utterance in order to observe this phenomenon and it's effect in the event of a speech recognition failure.

A common problem with touchless, mid-air hand gestures is fatigue. In our previous work, we introduced the Personal Space approach to gestural interaction; it was demonstrated to reduce fatigue by allowing users to rest their elbow on a surface and defining their own interaction space through a calibration step, as shown in figure 1.

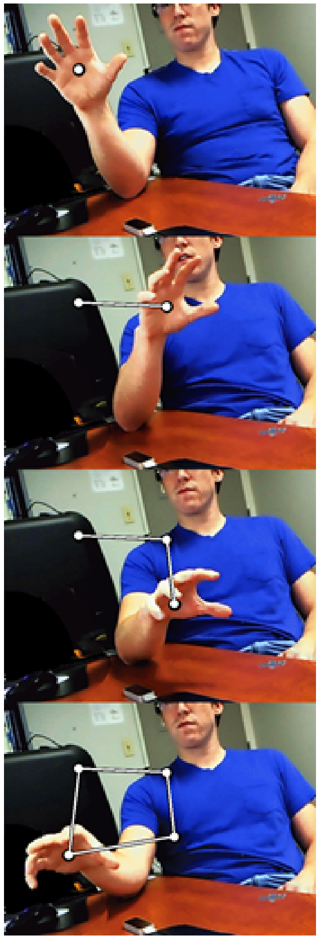

Figure 1: During calibration, the software guides users to position their hand at the 4 corners. Users can then rest their elbow on the table during the task, which reduces fatigue.

Methods

Participants

A total of 7 participants (M=3, F=4) were recruited to participate in a within-subjects experiment. Participants were between 19 to 23 years of age with a median of 20 years. None had prior experience with gestural interaction. Participants were compensated for their participation.

Task

The experiment consisted of 2 types of tasks: (1) navigation only task, and (2) navigation and speech multimodal task. In the latter, speech-based input was used to issue commands while mouse or gestures were used for navigation. Each task consisted of 70 trials, each task was performed with both the mouse and gesture-based input. This resulted in every participant performing 4 tasks per experiment. In each trial, a square target appeared at locations designed to look random. Participants were required to navigate to the target and perform a particular action that caused the target to disappear and a new target to appear. In the navigation-only task, a hover action was used, where users hover on the target for 500 milliseconds. In the multimodal task, subjects would issue a speech command of "left click" or "right click" or perform a hover

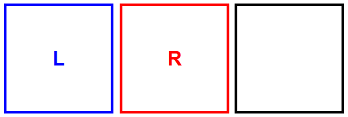

Figure 2: The targets expect 1 of 3 possible actions. From left to right: left-click, right-click, hover.

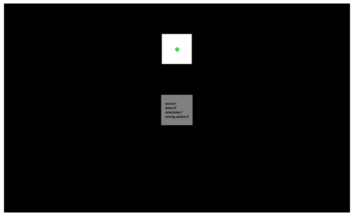

Figure 3: Target shown here expects a hover action. 70 targets are shown per task. The gray box at the center of the screen shows the users' score per task.

depending on the target. Multimodal tasks contained 40 hover targets and 30 selection targets. The targets were labeled with different colors as well as with the initial of the action: "L" for left-click, "R" for right click while the hover target had no label. Both tasks had square targets of four different sizes, each with sides of size: 220px, 160px, 120px, 90px.

Input Methods
A regular off-the-shelf mouse was used for mouse tasks. Gestural navigation used the Personal Space approach, as it has been demonstrated to reduce fatigue [2]. This approach creates a quadrilateral flat plane in 3-dimensional space which is then affine-mapped to the display screen. The commercial software "e-Speaking" was used to process speech input.

Results and Analysis

Adapting Speech to Gestures
Participants generally move the cursor to the target before beginning utterance of the command. This behavior was in fact expected and resulted in a mean speech overhead of 1814ms. When using gestures with speech however, participants eventually learned to optimize their movement by adapting for these overheads. 4 of the 7 participants used "pre-emptive utterance", where they began the utterance of the speech commands before the cursor was on the target. These users had a lower mean speech overhead per trial of 856ms, 990ms, 1000ms and 1102ms respectively. While participants without pre-emptive utterances had a mean speech overhead per trial of 1438ms, 1453ms and 1344ms respectively.

Hyperarticulation
We observed that participants would hyperarticulate in the event of an error, which is consistent with previous work [4]. A normal utterance of the "left click" or "right click" command tended to take approximately 850 milliseconds. If the command was not recognized, subsequent utterances would be hyperarticulated, and therefore take a longer utterance time – approximately 1300 milliseconds. However, we did not observe a "cycle of recognition failure" as mentioned in the previous study. This could be attributed to improvements in speech-recognition and fewer words uttered per command.

Throughput
Throughput is the standard used to evaluate performance of pointing tasks [6] and is defined as the ratio between the Index of Difficulty (*ID*) over Movement Time (*MT*) measured in bits per second (bps): Throughput $= \left(\frac{ID}{MT}\right)$. ID is defined as the ratio of the distance travelled over the width of the target: $ID = log_2 \left(\frac{D}{W} + 1\right)$.

Throughput was used to measure the performance of the navigational aspect of our implementation as it only considers the movement of the cursor, without taking the overhead of the hover or speech into consideration. This allowed us to measure if introducing speech caused a deterioration in the performance of navigation.

Throughput means per interaction are listed in table 1. A t-test showed no statistically significant difference when speech was added to mouse navigation $(t(12) = -1.1766,\ p = .26)$ nor gestural navigation $(t(12) = -0.7357, p = 0.48)$. This tells us that the introduction of speech into navigation tasks does not cause a degradation in performance measured in bps.

This demonstrates that usable gesture and speech multimodal interaction can be improved given specific design considerations, despite suggestion to the contrary as outlined in the related works section. In our experiment, these design considerations used were fewer words per command, a limited set of commands, and a constant number of utterance per command.

Completion Time

Although navigation performance itself was not affected by the introduction of speech, it is very evident that speech adds an overhead in terms of completion time. Additionally gesture and speech interaction takes almost twice as long as mouse only, as shown in Table 1.

Interface	Completion		Throughput	
	Mean	Stdev	Mean	Stdev
Mouse only	74s	5.19	4.33bps	0.35
Mouse & speech	114s	9.74	4.58bps	0.46
Gestures only	116s	14.76	2.66bps	0.40
Gestures & Speech	136s	14.82	2.83bps	0.49

Table 1: Completion time (seconds), and throughput (bps)

Discussion and Future Work

We intend to investigate this interaction as a form of assistive technology for users with carpal-tunnel, arthritis, muscle-dystrophy or any other hand impairments This interaction will be compared with other interactions used by users with the aforementioned impairments to gather both quantitative and qualitative results.

One interesting finding in this study is pre-emptive utterance of commands. We believe this behavior is likely easier to learn when the length of the commands are equal. The commands "right click" and "left click" used in our experiment are both two-syllable phrases which take equal time to vocalize. Future research will study the effect of having commands with differing lengths (eg: "right", "left", "middle click"), and if they interfere with this optimization learnt by users. Future research will also investigate why this behavior was only noticed with gestures, but not the mouse. We also intend to study the effect of target distance on this behavior.

Although the likelihood of speech-recognition failure in our experiment was low, the cost of this failure was quite high due to hyperarticulation following the recognition failure. This high cost occurs even without the cycle of recognition being present. Future work will look at ways to reduce hyperarticulation, such as user training and system feedback.

References

[1] Hauptmann, A. G., and McAvinney, P. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies 38*, 2 (1993), 231–249.

[2] Jude, A., Poor, G. M., and Guinness, D. Personal space: User defined gesture space for gui interaction. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, ACM (New York, NY, USA, 2014), 1615–1620.

[3] Kopp, S. Giving interaction a hand: Deep models of co-speech gesture in multimodal systems. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, ACM (New York, NY, USA, 2013), 245–246.

[4] Oviatt, S. User-centered modeling for spoken language and multimodal interfaces. *IEEE multimedia 3*, 4 (1996), 26–35.

[5] Shneiderman, B. The limits of speech recognition. *Commun. ACM 43*, 9 (Sept. 2000), 63–65.

[6] Soukoreff, R. W., and MacKenzie, I. S. Towards a standard for pointing device evaluation, perspectives on 27 years of fitts law research in hci. *International Journal of Human-Computer Studies 61*, 6 (2004), 751–789.