Conveying Shopper Intent in Robot-Assisted Shopping for the Blind

Chaitanya P. Gharpure
Computer Science Assistive
Technology Laboratory
Utah State University, USA

Vladimir A. Kulyukin
Computer Science Assistive
Technology Laboratory
Utah State University, USA

Daniel Coster
Department of Mathematics
and Statistics
Utah State University, USA

ABSTRACT

This paper addresses the problem of quickly selecting the desired command (shopper intent) for RoboCart, a robotic shopping assistant for the visually impaired. The product titles act as the commands for RoboCart, in the sense that they tell RoboCart which location to navigate to. We present two quick item retrieval (QIRI) interfaces: typing interface and speech interface, and evaluate them against a purely browsing interface. Results of a pilot study with 5 blind and 5 sighted-blindfolded participants who used the interfaces on a publically available online database of 11,147 household products is presented. No statistically significant difference was found in the performance of the blind and sighted participants. It was found that the participants were slowest with the browsing interface. The NASA Task Load Index (NASA-TLX) was administered to the participants to assess a subjective evaluation of the task load imposed by each interface. It was found that the browsing and speech interfaces imposed higher frustration on the participants than the typing interface.

1. INTRODUCTION

Communicating user intent to a robot efficiently is a major problem in human-robot interaction. Several approaches like button presses, speech, gestures, and eye gaze, have been used to solve this problem. The problem is tractable when the user has an adequate mental model of the robot and the repertoire of commands is small. However the aforementioned approaches quickly become intractable as the repertoire grows to several thousand commands. The problem worsens when the user is visually impaired (VI), because all visual modes of interaction are ruled out. In this paper, we elaborate on this problem in the context of robot-assisted shopping for the VI.

1.1 Motivation

The motivation for this research came from our previous work on assisted shopping for the VI. [4, 5]. Toward this, we have developed a prototype of a robotic shopping cart for the VI (RoboCart) [1, 4]. RoboCart helps the VI shop, in four steps: The VI shopper (the shopper henceforth) selects the desired product; RoboCart guides the shopper to the product; The shopper finds the product from the shelf and places it in the basket; RoboCart guides the shopper to the cash register, and then to the exit. First step essentially requires the shopper to select the desired product from the repository of thousands of products. This paper addresses the following problem: How can a VI user *quickly* retrieve a desired item from a large repository of items. The issues associated with the remaining 3 steps have been addressed elsewhere [4].

The task of item selection is time critical for following reasons. First, selecting a product is merely an act of conveying shopper's intent to the device - Take me to <product> - and therefore should not be temporally expensive. Second, if the VI shopper is stranded at a position in the supermarket trying to select a product, it might negatively affect the shopper traffic and also make the VI shopper uncomfortable.

1.2 Interaction Hardware

The shopper can communicate with RoboCart using a 9-key numeric keypad or a microphone. The feedback is in form of synthesized speech relayed through a pair of bluetooth headphones. The 9-key numeric keypad is attached to the right side of the handle onto which the shopper holds while navigating. The microphone is attached with a pair of bluetooth headphones and is worn by the user. A small bump on the middle key '5' allows the VI user to locate that key and the other keys with respect to the middle key.

We ruled out the possibility of installing a full keyboard on the robot, due to following reasons. First, we hope that in the future, the VI shopper would communicate with the robot through her own mobile phone. Therefore, we use the 9-key numeric keypad which closely resembles with a cell phone's keypad layout. Second, we intend to employ the same mode of interaction for our wearable assisted navigation device, ShopTalk [5]. A full keyboard will obviously be bulky for a wearable device.

2. INTERFACE DESIGN

The item repository is organized into a hierarchy, the leaf nodes in which represent actual products while the internal nodes represent the higher level categories to which the underlying products belong. All paths in the hierarchy from the top level to the lowest level are of equal length. Thus if there are 4 levels in the hierarchy, all products are located at level 4. In this paper we are not concerned about the optimal categorization of items in the repository. Therefore, for experimental purposes, we chose to use the household product database [3], consisting of 11,147 products categorized into a 4-level hierarchy.

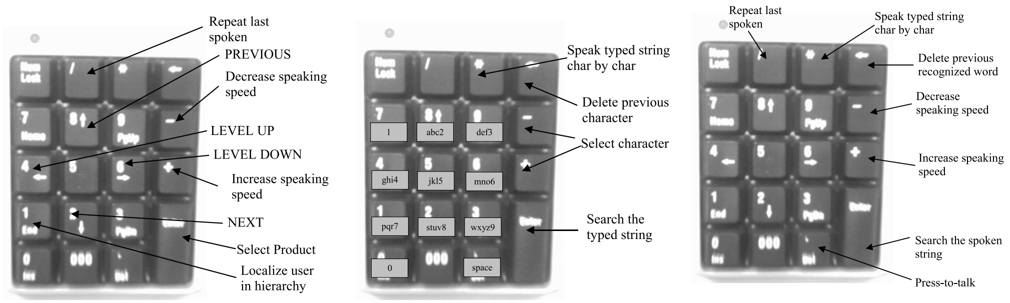

Figure 1: Keypad layout for browsing, typing and speech interfaces (L to R)

In interest of space, we will skip the detailed explanation of the three interfaces and the quick item retrieval algorithm. Please refer to figure 1 to understand the mapping of keys to their functionality. The browsing interface employs a simple directed browsing behavior. Features to move multiple items forward or backward and to localize the user in the hierarchy are provided. The typing and speech interfaces employ an incremental keyword search behavior followed by a list browsing behavior. The users can save time because they are allowed to type partial keywords and are provided with continuous feedback about the number of matching results. A quick item retrieval algorithm is used, which builds a prediction tree from the partial query. The prediction tree gives all possible complete queries which when executed, return a list of results. The user can either choose to browse the list, or continue entering more keywords to narrow the search.

3. EXPERIMENTS

Experiments were conducted with 5 blind and 5 sighted, blindfolded subjects. The subjects' age ranged from 17 years through 32 years and all subjects were males. To avoid the discomfort of wearing a blindfold, the keypad was covered with a box to prevent the sighted subjects from seeing it. The experiment was conducted in a laboratory setting.

During session-1, each subject was required to perform 10 unique tasks, using the typing, speech and browsing interfaces, in that order. For each task, the subject was provided the description of a product in form of its name, brand, speciality (color/scent/flavor), display text, and was asked to find the product in the large hierarchy using the three interfaces. During session-2, each subject performed 10 new tasks using the typing and speech interfaces. NASA Task Load Index was monnitored on the subjects after both the sessions.

4. RESULTS

4.1 User Performance

Repeated measures ANOVA was fitted to the data with selection time as the dependent variable. Model factors were: *interface* (3 levels: browsing, typing, speech), *condition* (2 levels: blind, sighted-blindfolded), *participant* (10 levels: nested within condition, 5 participants per blind / sighted-blindfolded condition), and *set* (2 levels: set-1 and set-2, each containing 10 products). Main effects for interface: F(2,243)=42.84, P<0.0001, condition: F(1,243)=9.8, P = 0.002, and participant: F(8,243)=9.88, P<0.0001 were observed. Interaction of *interface x condition*, F(2, 243)=0.05, P = 0.9558 and *interface x participant*, F(14, 243)=1.17, P = 0.2976 was observed. Thus, mean selection time differed significantly among interfaces, but the lack of interactions indicated that the interface differences did not vary significantly between blind and sight-blindfolded groups, nor among individual participants.

The majority of the differences among participants arose from blind participant 5, whose mean selection time of 120.9 (s) differed significantly from the mean selection time of all others participants (whose mean times were in the 53-63 secs range) (P < 0.0001 for all comparisons between blind participant 5 and all other participants). When blind participant 5 was dropped from the analysis, main effect of both condition and participant(condition) became non-significant (F(1, 216) = 0.16, P = 0.6928, and F(6,216) = 0.44, P = 0.8545, respectively). The interactions of interface with each of condition and participant remained non-significant also. It appears that on average, when the outlier (participant 5) was removed, blind and sighted-blindfolded participants did not really differ.

The almost parallel lines for the blind and sighted-blindfolded participants in figure 2(L), suggest that the interface which is best for sighted-blindfolded users will also be best for blind users. We therefore take the liberty not to make any explicit distinction between the blind and sighted-blindfolded participants, during the remaining analysis in this paper.

Mean selection times for the 3 interfaces were, respectively: 85.5, 74.1, and 37.5 (seconds). Post-hoc pairwise t-tests showed that the typing interface was faster than the browsing interface (t = 2.10, P = 0.0364), although statistical significance is questionable if the Bonferroni adjusted α-level is used here. Each of the browsing and typing interfaces was significantly slower than the speech interface (t = 8.84, P < 0.0001, and t = 6.74, P < 0.0001, respectively).

A significant interaction of *interface x set*, F(1, 382)=13.8, P=0.0002 was observed. It appears from the graph (figure 2) that the improvement with the typing interface was much larger than that with the speech interface. The reduction in selection times from session-1 to session-2 varied significantly for the typing and speech interface. This was probably because, since the participants were already much faster with the speech interface than the typing interface during session-1, they had much less room to improve with

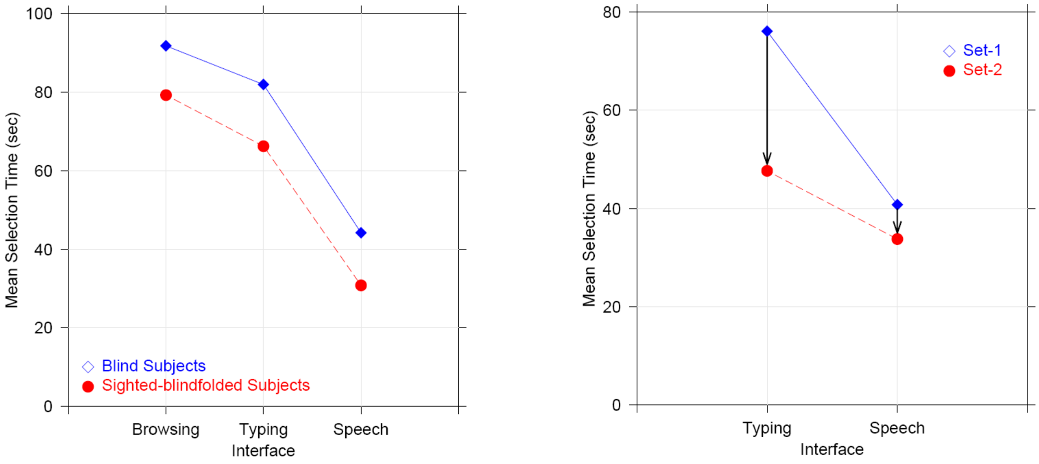

Figure 2: (L) Mean selection times for blind and sighted-blindfolded participants against all interfaces. (R) Change in mean selection times for typing and speech interfaces from Session-1 to Session-2.

the speech interface during session-2.

Finally, though we did not find other recorded observations much useful for our analysis, a strong Pearson's product moment correlation was found between selection time and query length for both typing and speech interfaces, with *r = 0.92* and *r = 0.82* respectively. To calculate the PPM correlation, we averaged the selection times over all products having the same query length. This just confirms the obvious that on an avergage, selection time increases with number of characters typed or words spoken.

4.2 Subjective Evaluation

Next the interfaces were subjectively evaluated by monitoring the NASA TLX [2] questionaire. A repeated measures ANOVA was also fitted to the data obtained with total workload as the dependent variable. Model factors were: *interface* (3 levels: browsing, typing, speech) and *condition* (2 levels: blind, sighted-blindfolded). The overall model was significant, F(5, 24) = 6.67, P = 0.0005. Both main effects of interface and condition were significant (F(2,24) = 12.25, P = 0.0002, and F(1,24) = 7.30, P = 0.0124, respectively). The interaction of interface with blind was not signficiant, so that differences between interfaces are effectively the same for blind and sighted-blindfolded groups.

The workload means for the three interfaces were 12.88, 8.33 and 7.00 for browsing, typing, and speech, respectively. From the pairwise t-tests among interface means, browsing workload was significantly greater than typing or speech, P = 0.0013 and $P < 0.0001$, respectively, but typing and speech mean workloads did not differ significantly, P = 0.2948. A little surprisingly, the mean workload for blind subjects (8.03) was significantly less than the mean workload for sighted-blindfolded subjects (10.78), t = 2.70, P = 0.0124. Component-wise analysis showed that the browsing and speech interfaces were far more frutrating than the typing interface. The participants who found speech interface frustrating gave the reason to be the high number of speech recognition errors.

5. CONCLUSIONS

With the complexity and task repertoire of intelligent personal/assistive robots increasing, there is a need for human-robot interfaces which enable the user to quickly select the desired action to be performed by the robot. This paper presented two such QIRI interfaces (typing and speech). Though it was seen that the speech interface was the fastest, in real life, the user might prefer to use the typing interface as it helps to be more discrete in a public place like a supermarket. We feel that an hybrid interface, a combination of typing and speech would be desirable. Also if the exact command is not known, and the user wishes to browse the command hierarchy, an interface with a strong coupling between browsing and searching would be desirable. Since it is difficult to evaluate how such a hybrid interface would perform in real life, we feel that evaluating the components independently, as done in this paper, gives us insights into how more complete interfaces should be designed.

6. REFERENCES

[1] C. Gharpure and V. Kulyukin. Robot-assisted shopping for the blind: Issues in spatial cognition and product selection. *Journal of Intelligent Service Robotics*, Springer Berlin / Heidelberg, Under Review, 2007.

[2] S. Hart and L. Staveland. Development of nasa-tlx: results of empirical and theoretical research. In *In Hancock, P. and Meshkati, N. (Eds.) Human mental overload*, 1988.

[3] Household-Products-Database. 2004. www.householdproducts.nlm.nih.gov.

[4] V. Kulyukin, C. Gharpure, , and C. Pentico. Robots as interfaces to haptic and locomotor spaces. In *Proceedings of the 2007 ACM Conference on Human-Robot Interaction*, Arlington, Virginia, 2007.

[5] J. Nicholson and V. Kulyukin. Shoptalk: Independent blind shopping = verbal route directions + barcode scans. In *Proceedings of the RESNA Conference*, Phoenix, AZ, USA, 2007.