20

User Intent Communication in Robot-Assisted Shopping for the Blind

Vladimir A. Kulyukin[1] and Chaitanya Gharpure[2]
[1]*Utah State University*
[2]*Google, Inc.*
USA

1. Introduction

The research reported in this chapter describes our work on robot-assisted shopping for the blind. In our previous research, we developed RoboCart, a robotic shopping cart for the visually impaired (Gharpure, 2008; Kulyukin et al., 2008; Kulyukin et al., 2005). RoboCart's operation includes four steps: 1) the blind shopper (henceforth the shopper) selects a product; 2) the robot guides the shopper to the shelf with the product; 3) the shopper finds the product on the shelf, places it in the basket mounted on the robot, and either selects another product or asks the robot to take him to a cash register; 4) the robot guides the shopper to the cash register and then to the exit.

Steps 2, 3, and 4 were addressed in our previous publications (Gharpure & Kulyukin 2008; Kulyukin 2007; Kulyukin & Gharpure 2006). In this paper, we focus on Step 1 that requires the shopper to select a product from the repository of thousands of products, thereby communicating the next target destination to RobotCart. This task becomes time critical in opportunistic grocery shopping when the shopper does not have a prepared list of products. If the shopper is stranded at a location in the supermarket selecting a product, the shopper may feel uncomfortable or may negatively affect the shopper traffic.

The shopper communicates with RoboCart using the Belkin 9-key numeric keypad (See Fig. 1 right). The robot gives two types of messages to the user: synthesized speech or audio icons. Both types are relayed through a bluetooth headphone. A small bump on the keypad's middle key (key 5) allows the blind user to locate it. The other keys are located with respect to the middle key. In principle, it would be possible to mount a full keyboard on the robot. However, we chose the Belkin keypad, because its layout closely resembles the key layout of many cellular phones. Although the accessibility of cell phones for people with visual impairments remains an issue, the situation has been improving as more and more individuals with visual impairments become cell phone users. We hope that in the future visually impaired shoppers will communicate with RobotCart using their cell phones (Nicholson et al., 2009; Nicholson & Kulyukin, 2007).

The remainder of the chapter is organized as follows. In section 2, we discuss related work. In sections 3, we describe our interface design. In section 4, we present our product selection algorithm. In section 5, we describe our experiments with five blind and five sighted, blindfolded participants. In sections 6, we present and discuss the experimental results. In section 7, we present our conclusions.

Source: Advances in Human-Robot Interaction, Book edited by: Vladimir A. Kulyukin,
ISBN 978-953-307-020-9, pp. 342, December 2009, INTECH, Croatia, downloaded from SCIYO.COM

Fig. 1. RoboCart (left); RoboCart's handle with the Belkin 9-key numeric keypad (right).

2. Related work

The literature on communicating user intent to robots considers three main scenarios. Under the first scenario, the user does not communicate with the robot explicitly. The robot attempts to infer or predict user intent from its own observations (Wasson et al., 2003; Demeester et al., 2006). Under the second scenario, the user communicates intent to the robot with body gestures (Morency et al., 2007). The third scenario involves intent communication and prediction through mixed initiative systems (Fagg et al., 2004). Our approach falls under the second scenario to the extent that key presses can be considered as body gestures.

Several auditory interfaces have been proposed and evaluated for navigating menus and object hierarchies (Raman, 1997; Smith et al., 2004; Walker et al., 2006). In (Smith et al., 2004), the participants were required to find six objects from a large object hierarchy. The evaluation was done to check for successful completion of the task, and was not evaluated for time criticality. In (Brewster, 1998), the author investigated the possibility of using nonspeech audio messages, called *earcons*, to navigate a menu hierarchy. In (Walker et al, 2006), the authors proposed a new auditory representation, called *spearcons*. Spearcons are created by speeding up a phrase until it is not recognized as speech. Another approach for browsing object hierarchies used conversational gestures (Raman, 1997), such as *open-object*, *parent*, which are associated with specific navigation actions. In (Gaver, 1989), generic requirements are outlined for auditory interaction objects that support navigation of hierarchies. While these approaches are suitable for navigating menus, they may not be suitable for selecting items in large object hierarchies under time pressure.

In (Divi et al., 2004), the authors presented a spoken user interface in which the task of invoking responses from the system is treated as one of retrieval from the set of all possible responses. The SpokenQuery system (Wolf et al., 2004) was used and found effective for searching spoken queries in large databases. In (Sidner & Forlines, 2004), the authors propose the use of subset languages for interacting with collaborative agents. One advantage of using

subset language is that it can easily be characterized in a grammar for a speech recognition system. One disadvantage is that the users are required to learn the subset language that may be quite large if the number of potentially selectable items is in the thousands.
In (Brewster et al., 2003) and (Crispien et al., 1996) the authors present a 3-D auditory interface and head gesture recognition to browse through a menu and select menu items. This approach may be inefficient for navigating large hierarchies because of the excessive number of head gestures that would be required. A similar non-visual interface is also described in (Hiipakka & Lorho, 2003).
Another body of work related to our research is the Web Content Accessibility Guidelines (W3C, 2003) for making websites more accessible. However, since these guidelines are geared toward websites, they are based on several assumptions that we cannot make in our research: 1) browsing a website is not time critical; 2) the user is sitting in the comfort of her home or office; and 3) the user has a regular keyboard at her disposal.

3. Interface design

Extensive research has been done regarding advantages of browsing and searching in finding items in large repositories (Manber et al., 1996; Mackinlay & Zellweger, 1995). It is often more advantageous to combine browsing and searching. However, when the goal is known, query-based searching is found to be more efficient and faster than browsing (Manber et al., 1996; Karlson et al., 2006). Since this case fits our situation, because the user knows the products she wants to purchase, we designed a search-based interface with two modalities: typing and speech. In both modalities, the shopper can optionally switch to browsing when the found list of products is, in the shopper's judgement, short and can be browsed directly. Our interface also supports a pure browsing modality used as the baseline in our experiments. We used the following rules of thumb to iteratively refine our design over a set of user trials with a visually impaired volunteer.

- **Learning:** The amount of learning required to use the interface should be minimal. Ideally, the interface should be based on techniques already familiar to the shopper, e.g. browsing a file system or typing a text message on a mobile phone.
- **Localization:** The shopper must know the state of the current search task. While browsing, the shopper should be able to find out, at any moment, the exact place in the hierarchy. While typing, the shopper should be able to find out, at any moment, what keywords have been previously typed. Similarly, in the speech modality, the shopper should be able to access the previously spoken keywords.
- **Reduced cognitive load:** The cognitive load imposed by the interface should be minimal. For browsing, this can be done by categorizing the products in a logical hierarchy. For typing and speech, continuous feedback should be provided, indicating the effect of every shopper action, e.g. character typed or word spoken.
- **Timestamping:** Every step during the progress of the search task should be timestamped, so that the shopper can go back to any previous state if an error occurs. The shopper should be allowed to delete the typed characters or misrecognized words that returned incorrect results.

3.1 Browsing

The keypad layout for browsing is shown in Fig. 2. The *UP* and *DOWN* keys are used to browse through items in the current level in the hierarchy. The *RIGHT* key goes one level

deeper into the hierarchy, and the *LEFT* key - one level up. Visually impaired computer users use the same combination of keys for browsing file systems. Holding *UP* and *DOWN* pressed allows the shopper to jump forward or backward in the list at the current depth in the hierarchy. The length of the jump is proportional to the time for which the key is pressed. A key press also allows the shopper to localize in the hierarchy by informing the shopper the current level and category. The *PAGE-UP* and *PAGE-DOWN* keys allow the shopper to go a fixed number of items up or down at the particular level in the hierarchy. Auditory icons, short and distinct, are provided when the shopper wraps around a list, changes levels, or tries to go out of the bounds of the hierarchy.

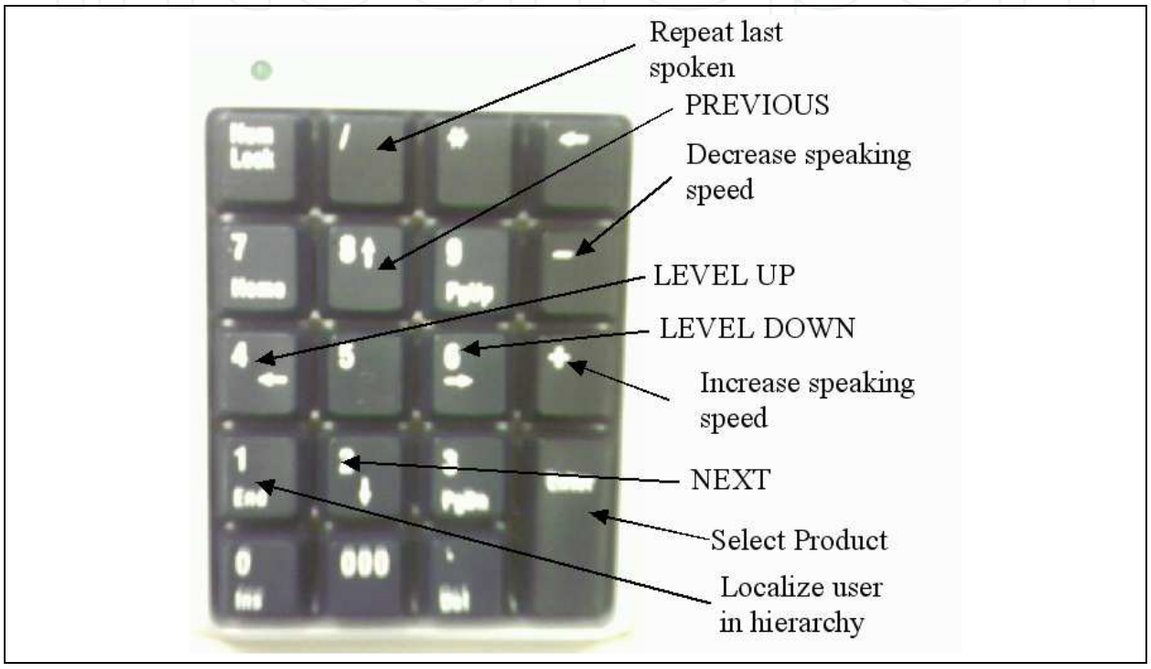


Fig. 2. Keypad layout for the browsing interface.

3.2 Typing

The keypad layout used for the typing interface is shown in Fig. 2. In the typing modality, the shopper is required to type a query string using the 9-key numeric keypad. This query string can be complete or partial. Each numeric key on the keypad is mapped to letters as if it was a phone keypad. Synthesized speech is used to communicate the typed letters to the shopper as the keys are pressed. The *SELECT* key is used to append the current letter to the query string. For example, if the shopper presses key 5 twice followed by the *SELECT* key, the letter *k* will be appended to the query string. At any time the shopper can choose to skip typing the remaining word by pressing the *space* key and continue typing the next word. Every time a new character is appended to the query string, a search is performed and the number of returned results is reported back to the shopper. The partial query string is used to form the prediction tree which provides all possible complete query strings. If the shopper feels that the number of returned results is sufficiently small, she can press *ENTER* and browse through each product using *NEXT* and *PREVIOUS* to look for the desired item.

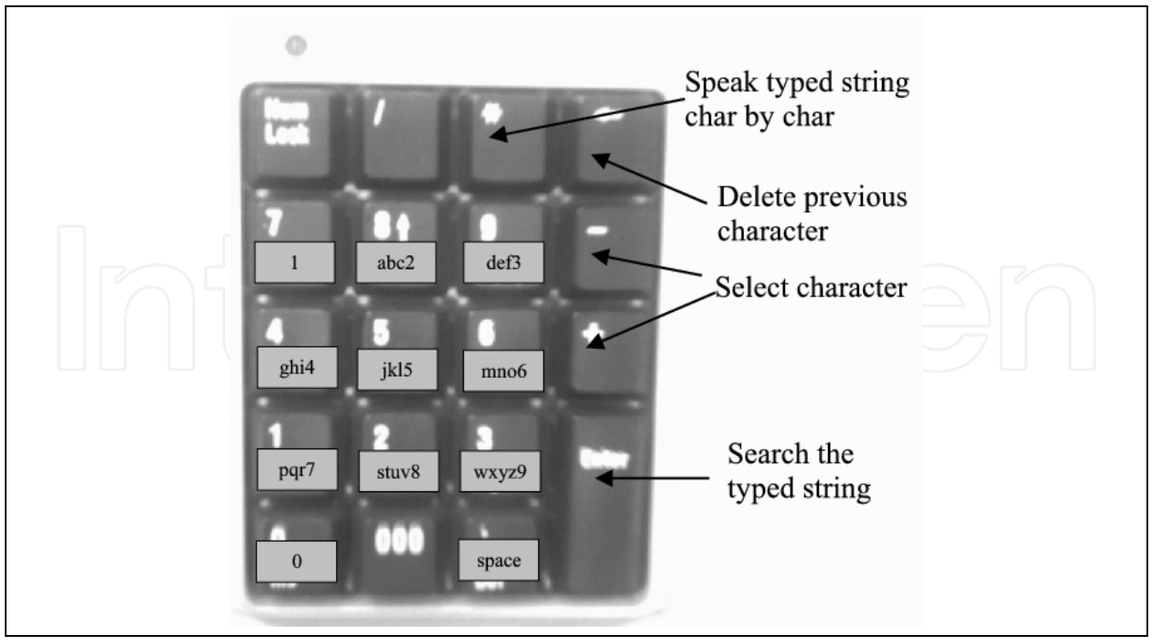


Fig. 2. Keypad layout for the typing interface.

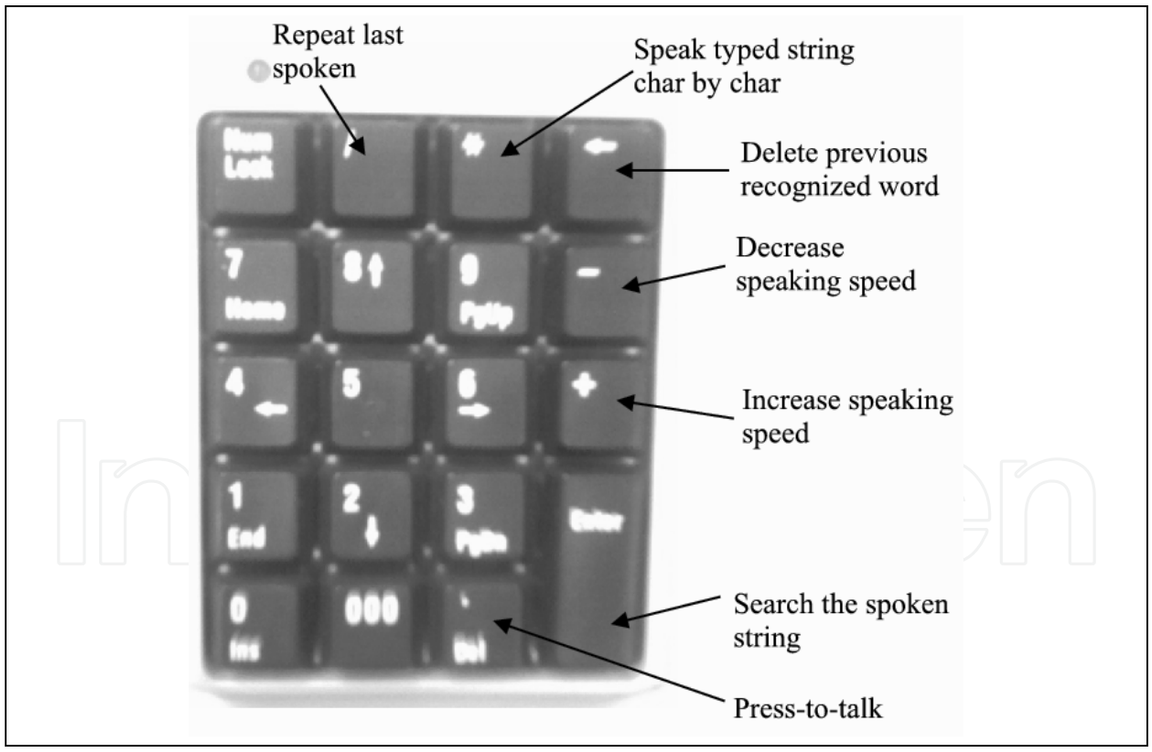


Fig. 3. A partial product hierarchy.

3.2 Speech

Our speech-based modality is a simplified version of the Speech In List Out (SILO) approach proposed in (Divi et al., 2004). The keypad layout for the speech-based modality is shown in

Fig. 3. The query string is formed by the words recognized by a speech recognition engine. The shopper is required to speak the query string into the microphone, one word at a time. A list of results is returned to the shopper, through which the shopper can browse to select the desired item. The grammar for the speech recognition engine consists of simple rules made of one word each, which reduces the number of speech recognition errors. To further reduce the number of false positives in speech recognition due to ambient noise, we provide a press-to-talk key. The shopper is required to press this key just before speaking a word. We use Microsoft's Speech API (SAPI) which provides alternates for the recognized word. The alternates are used to form the prediction tree which, in turn, is used to generate all possible query strings. The prediction tree concept is explained in the next section.

4. Product selection algorithm

Our product selection algorithm is used in the typing and speech modalities. The algorithm can be used on any database of items organized into a logical hierarchy. Each item title in the repository is extended by adding to it the titles of all its ancestors from the hierarchy. For example, in Fig. 4 the item *Kroger Diced Pineapples (0.8lb)* is extended to *Canned Products*, *Fruits*, *Pineapple*, *Kroger Diced Pineapples (0.8lb)*.

Each entry in the extended item repository is represented by an N-dimensional vector where N is the total number of unique keywords in the repository. Thus, each vector is an N-bit vector with a bit set if the corresponding keyword exists in the item string. The query vector obtained from the query string is also an N-bit vector. The result of the search is simply all entries i, such that $P_i \,\&\, S = S$, where P_i is the N-bit vector of the i-th product, S is the N-bit query vector, and & is the bit-wise and operation.

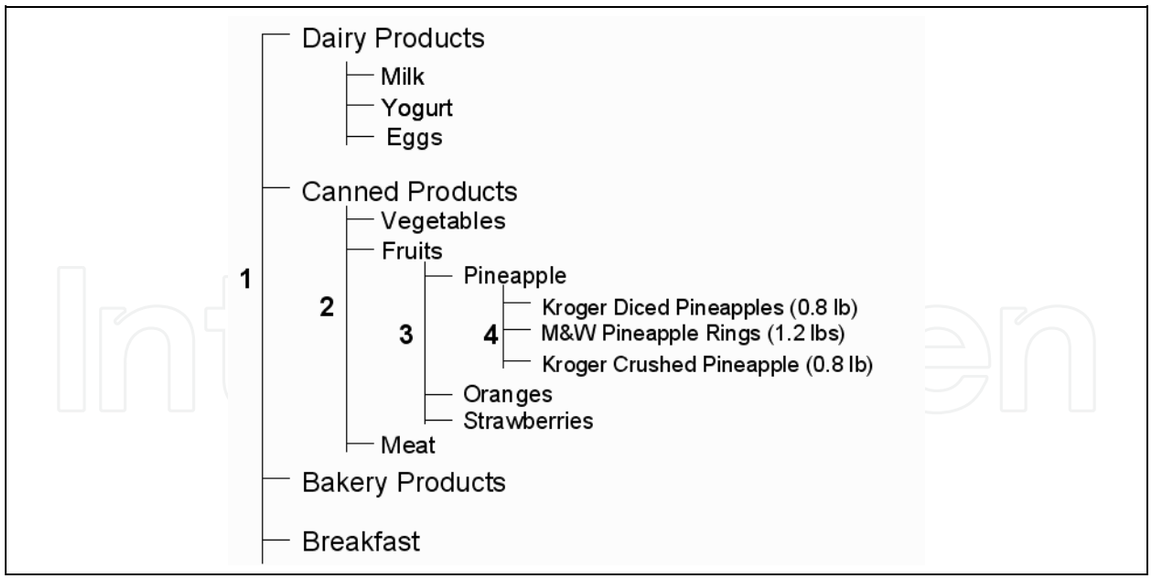


Fig. 4. A partial product hierarchy.

This approach, if left as is, has two problems: 1) the shopper must type complete words, which is tedious using just a numeric keypad or a cell phone; and 2) the search fails if a word is spelled incorrectly. To solve the first problem, we use word prediction where the whole word is predicted by looking at the partial word entered by the shopper. However,

instead of having the shopper make a choice from a list of predicted words, or waiting for the user to type the whole word, we search the repository for all predicted options. To solve the second problem, we do not use the spell checker, but instead provide the shopper with continuous audio feedback. Every time the shopper types a character, the number of retrieved results is reported to the shopper. At any point in a word, the user can choose not to type the remaining characters and proceed to the next word.

The predictions of partially typed words form a tree. Figures 5 and 6 show the prediction tree and the resultant query strings when the shopper types "deo so ola." The sharp-cornered rectangles represent the keywords in the repository, also called keyword nodes. The round-cornered rectangles are the partial search words entered by the shopper, also called the partial nodes. Keyword nodes are all possible extensions of their (parent) partial node, as found in the keyword repository.

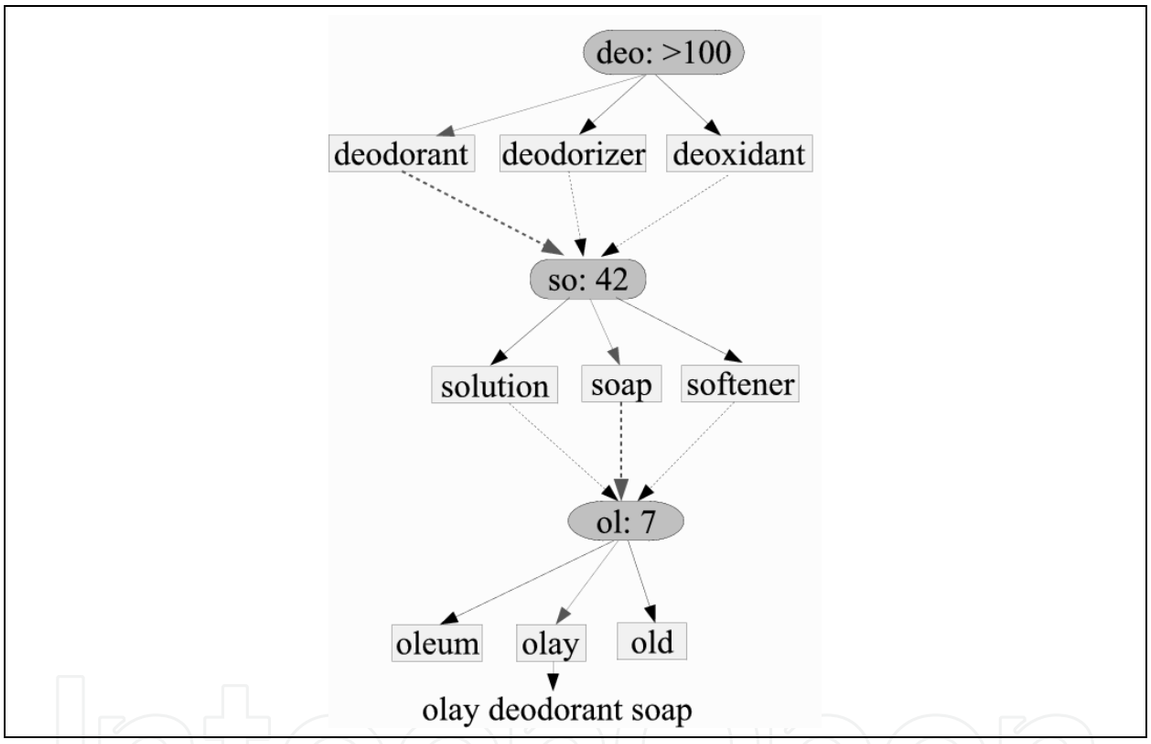


Fig. 5. A sample prediction tree.

Each keyword node is associated with multiple query strings. Every path from the root of the prediction tree to the keyword node forms a query string by combining all keywords along that path. For example, in the prediction tree shown in Fig. 5 there will be three query strings associated with the keyword node solution: deodorant solution, deodorizer solution, and deoxidant solution. The prediction subtree is terminated at the keyword node where the associated query string returns zero results. For example, in Fig. 5, the subtree rooted at solution, along the path deodorant-solution will be terminated since the search string *deodorant solution* returns zero results. Fig. 6 shows the possible query strings for the prediction tree in Fig. 5. The numbers in the parentheses indicate the number of results returned for those search strings. The number after colons in the partial nodes indicates the total results returned by all query strings corresponding to its children (keyword) nodes.

```
Possible search strings
for the Prediction tree
deodorant solution (0)
deodorant soap oleum (0)
deodorant soap olay (2)
deodorant soap old (2)
deodorant softener (0)
deodorizer solution (0)
deodorizer soap oleum (0)
deodorizer soap olay (1)
deodorizer soap old (2)
deodorizer softener (0)
deoxidant solution (0)
deoxidant soap (0)
deoxidant softener (0)
```

Fig. 6. Possible strings for the sample prediction tree.

In addition to implementing the algorithm on a Dell laptop that runs on the robot, we also ported the algorithm to a Nokia E70 cell phone that runs the Symbian 9 mobile operating system. The algorithm was modified when the interface was implemented on the cell phone. The memory and processor speed restrictions on the cell phone made us optimize the algorithm. To reduce space requirements, each word in the product repository was replaced by a number depending upon the frequency of occurrence of that word in the repository. The algorithm for assigning these codes is given in Fig. 7. The procedure SortByFrequency sorts the elements of the set of unique words (W) in the decreasing order of the frequency of occurrence.

```
1.  W = Set of Unique Words
2.  PROD = Set of Products
3.  for each w in W
4.        FREQ(w) = 0
5.  for each p in PROD
6.        for each word in W_p
7.              FREQ(word) = FREQ(word)+1
8.  SortByFrequency(W)
9.  for each word in W
10.       CODE(word) = INDEX(word)
```

Fig. 7. Frequency-Based Encoding of Words.

The product selection algorithm implemented on the Nokia E70 mobile phone is given in Fig. 8. A set P(w) is a set of indices of products containing word w. Initialized to empty set. PROD is the set of all products. S is the set of keywords in the user query. Q is the set of products containing the word S[i]. 1. R is intersected with Q for each S[i] and eventually the filtered set of products is obtained.

```
1.  W = Set of Unique Words
2.  PROD = Set of Products
3.  for n = 1 to W.length
4.     P(W[n]) = {}
5.  for i = 1 to PROD.length
6.     W_p = Set of words in PROD[i]
7.  for j = 1 to W_p.length
8.     P(W_p[j]) = Union(P(W_p[j]), {i})
9.  R = PROD
10. S = {keyword1, keyword2, ..., keywordk}
11. for i = 1 to S.length
12.    Q = {}
13.    for j = 1 to W.length
14.       if W[j] startswith S[i]
15.             Q = Union(Q, P(W[j])
16.    R = Intersection(R, Q)
17. return R
```

Fig. 8. Possible strings for the sample prediction tree.

4.1 Procedure

As mentioned above, we used the product repository of 11,147 products that we obtained from www.householdproducts.nlm.nih.gov. The following procedure was followed for each participant. After arriving at the lab, the participant was first briefly told about the background and purpose of the experiments. Each participant recieved 20 minutes of training to become familiar with the interface and the modalities. As part of the training procedure, the participant was asked to find three products with each modality.

Session 1 started after the training session. Each task was to select a product using a given modality. A set of 10 randomly selected products (set-1) was formed. Each participant was thus required to perform 30 tasks (10 products x 3 interfaces). Because of his schedule, one of the participants was unable to perform the browsing modality tasks due to a scheduling conflict. The product description was broken down into 4 parts: product name, brand, special description (scent/flavor/color), and the text that would appear in the result communicated to the participant with synthetic speech. Table 1 gives an example. In the course of a task, if the participants forgot the product description, they were allowed to revisit it by pressing a key.

PRODUCT NAME	BRAND	DESCRIPTION	RESULT TEXT
Liquid Laundry Detergent	Purex	Mountain Breeze Bleach Alternative	Purex Mountain Breeze with Bleach Alternative Liquid Laundry Detergent

Table 1. A product description.

For Session 2, another 10 products (set-2) were randomly selected. After the initial 30 tasks in Session 1, 20 more tasks were performed by each participant (10 products x 2 interfaces). We skipped the browsing modality in Session 2, because our objective in Session 2 was to

check if and how much the participants improved on each of the two modalities, relative to the other. The dependent variables are shown in Table 2. Some variables were recorded by a logging program, others by a researcher conducting the experiment. Since all the tasks were not necessarily of the same complexity, there was no way for us to check the learning effect.
All experiments were first conducted with 5 blind participants and then with 5 sighted, blindfolded participants. After both sessions, we conducted a subjective evaluation of the three modalities by administering the NASA Task Load Index (NASA-TLX) to each participant. The NASA-TLX questionnaires were administered to eight participants in the laboratory right after the experiments. Two participants were interviewed on the phone, one day after the laboratory session.

BROWSING	TYPING-BASED	SPEECH-BASED
Time to selection	Typing errors	Recognition errors
Wrong selection	Time to type	Time to speak
Failed search	Time to selection	Time to selection
	Number of returned results	Number of retruned results
	Wrong selections	Wrong selections
	Failed search	Failed search
	Number of chars typed	Number of spoken words

Table 2. Observations for product retrieval interface experiments.

4.2 Data analysis

Repeated measures analysis of variance (ANOVA) models were fitted to the data using the SAS™ statistical system. Model factors were: modality (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).
The 10 products within each set were replications. Since each participant selected each product in each set, the 10 product responses for each set were repeated measures for this study. Since the browsing modality was missing for all participants for set-2 products, models comparing selection time between sets included only typing and speech modalities. The dependent variable was, in all models, the product selection time, with the exception of analyses using the NASA-TLX workload measure. The overall models and all primary effects were tested using an α-level of 0.05, whenever these effects constituted planned comparisons (see hypotheses). However, in the absence of a significant overall F-test for any given model, post-hoc comparisons among factor levels were conducted using a Bonferroni-adjusted α-level of 0.05/K, where K is the number of post-hoc comparisons within any givenmodel, to reduce the likelihood of false significance.

5. Experiments

Experiments were conducted with 5 blind and 5 sighted, blindfolded participants. The participants' ages ranged from 17 years through 32 years. All participants were males. To avoid the discomfort of wearing a blindfold, for sighted participants the keypad was

covered with a box to prevent them from seeing it. The experiment was conducted in a laboratory setting. The primary purpose behind using sighted, blindfolded participants was to test whether they differed significantly from the blind participants, and thus decide whether they can be used in future experiments along with or instead of blind participants. We formulated the following research hypotheses. In the subsequent discussion, H1-0, H2-0, H3-0 and H4-0 denote the corresponding null hypotheses.

Hypothesis 1: (H1) *Sighted, blindfolded participants perform significantly faster than blind participants.*

Hypothesis 2: (H2) *Shopper performance with browsing is significantly slower than with typing.*

Hypothesis 3: (H3) *Shopper performance with browsing is significantly slower than with speech.*

Hypothesis 4: (H4) *Shopper performances with typing and speech are significantly different from each other.}Equations are centred and numbered consecutively, from 1 upwards.*

6. Results

For an overall repeated measures model which included the effects of modality, condition, and participant (nested within condition), and the interaction of modality with each of condition and participant, using only set-1 data, the overall model was highly significant, $F(26,243) = 7.00, P < 0.0001$. The main effects observed within this model are shown in Table 3. All the main effects were significant. Interaction of modality x condition, F(2, 243)=0.05, P = 0.9558 and modality x participant, F(14, 243)=1.17, P = 0.2976 was observed. Thus, the mean selection time differed significantly among modalities, but the lack of interactions indicated that the modality differences did not vary significantly between blind and sight, blindfolded groups, nor among individual participants. In the ANOVAs, note that the DoF for the error is 243, because one of the participants did not perform the browsing tasks.

SOURCE	MAIN EFFECTS (ANOVA)
Interface	F(2,243)=42.84, P<0.0001
Condition	F(1,243)=9.8, P=0.002
Participant	F(8,243)=9.88, P<0.0001

Table 3. Main effects.

The mean selection time for the group of blind participants was 72.6 secs versus a mean of 58.8 secs for sighted-blindfolded participants, and the difference in these means was significant (t = 3.13, P = 0.0029). As might be expected, participants differed on mean selection time. However, 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 modality with condition and participant also remained non-significant. It appears that, on average, when the outlier (participant 5) was removed, blind and sighted-blindfolded participants did not really differ. Thus, there was no sufficient evidence to reject the null hypothesis H1-0.

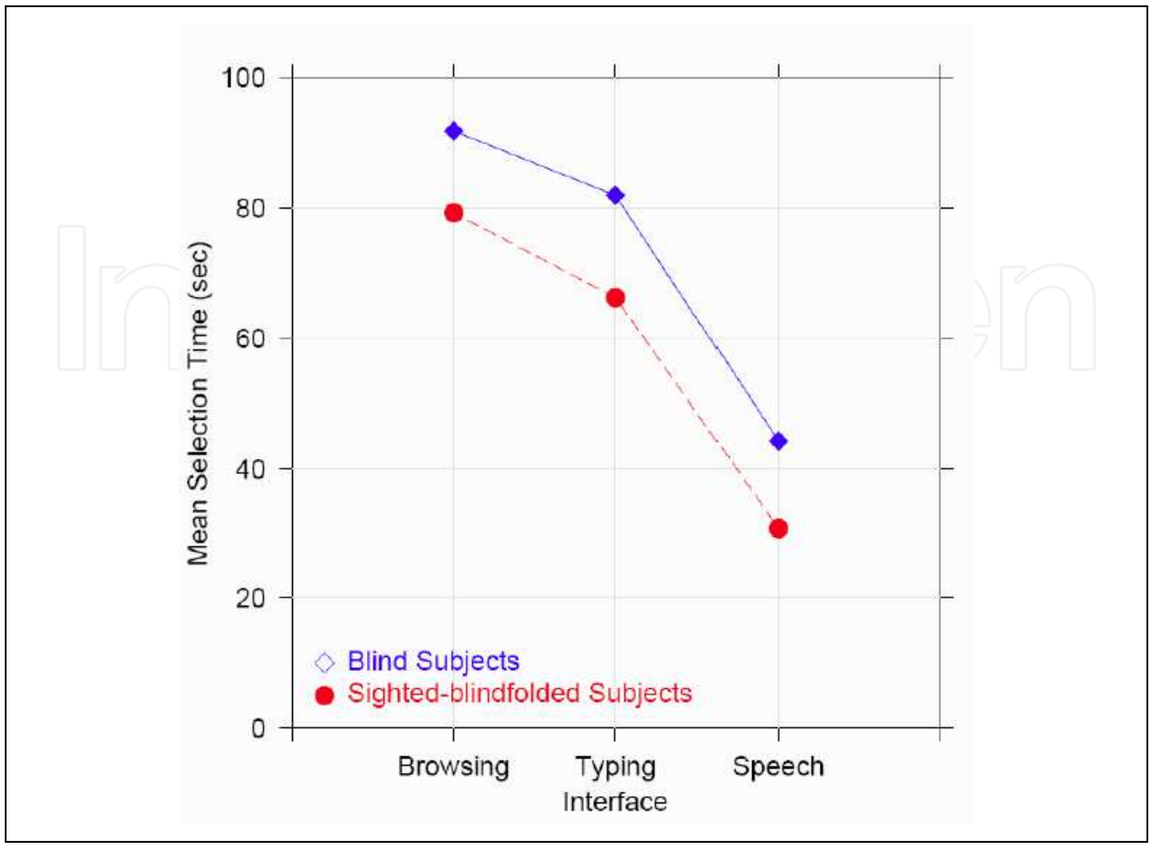


Fig. 9. Mean selection times for blind and blindfolded sighted participants against all interfaces.

A graph of the mean selection times of the blind and the sighted, blindfolded participants for each modality is shown in Fig. 9. The almost parallel lines for the blind and sighted-blindfolded participants suggest that there is no interaction between the modality and the participant type, which is also confirmed by the ANOVA result presented earlier. In other words, the result suggests that the modality which is best for sighted, blindfolded shoppers may also be best for blind shoppers.

The main effect of modality, as shown in Table 3, suggests that, on average (over all participants), two or more modalities differ significantly. Mean selection times for browsing, typing, and speech were: 85.5, 74.1, and 37.5 (seconds), respectively. Post-hoc pairwise t-tests showed that typing was faster than browsing ($t = 2.10$, $P = 0.0364$), although statistical significance is questionable if the Bonferroni-adjusted is used here. We, therefore, were unable to reach a definite conclusion about H2. Both browsing and typing were significantly slower than speech ($t = 8.84$, $P < 0.0001$, and $t = 6.74$, $P < 0.0001$, respectively). This led us to reject the null hypotheses H3-0 and H4-0 in favor of H3 and H4.

Since we were primarily interested in the difference between typing and speech, we decided to compare the modalities on the measures obtained from Session 2. Set-2 was significantly faster than set-1, averaged over the two modalities and all participants ($t = 6.14$, $P < 0.0001$). Since we did not have a metric for the task complexity, we were unable to infer if this result reflected the learning effect of the participants from Session 1 to Session 2. However, a

significant interaction of modality x set, F(1, 382)=13.8, P=0.0002 was observed. The graph of the selection times during Sessions 1 and 2, against the modality type is shown in Fig. 10. It appears from the graph that the improvement with typing was much larger than that with speech. The reduction in selection times from Session 1 to Session 2 varied significantly for typing and speech ($P < 0.0001$). This was probably because the participants were already much faster with speech than typing during Session 1 and had much less room to improve with speech during Session 2.

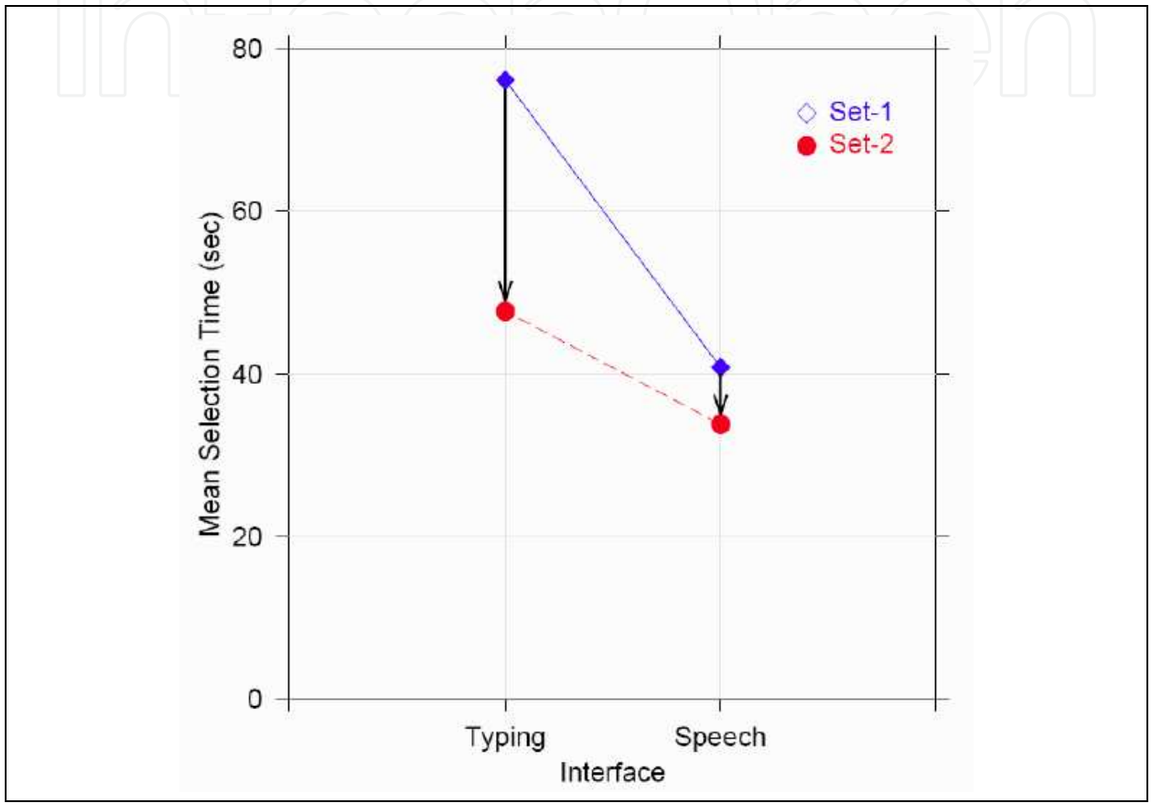


Fig. 10. Change in mean selection times for typing and speech interfaces from Session 1 and Session 2.

A strong Pearson's product moment correlation was found between selection time and query length for both typing and speech, 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 average, selection time increases with the number of characters typed or words spoken.

We used a between-subjects design to study the data obtained from the NASA TLX questionaire. The *modality type* was the independent variable and mental *demand*, *frustration*, and *overall workload* were the dependent variables. A one-way ANOVA indicated that there was a significant difference among the three modalities in terms of the mental demand, frustration, and overall workload, ($F(2, 27) = 16.63$, $P < 0.0001$), ($F(2, 27) = 16.63$, $P < 0.0001$), and ($F(2, 27) = 10.07$, $P = 0.0005$) respectively). Post-hoc pair-wise t-tests for the three dependent variables with Bonferoni adjusted α-level of 0.016 are shown in Table 4. The

mean values of mental demand, frustration and overall workload for the three modalities are shown in Table 5.

	Browsing x Typing	Browsing x Speech	Typing x Speech
Mental Demand	t=1.075, P=0.2962	* t=3.822, P=0.0012	* t=4.011, P=0.0008
Frustration	* t=6.974, P<0.0001	t=1.348, P=0.1833	* t=4.428, P=0.0004
Overall Workload	*t = 3.369, P=0.0034	* t=4.126, P=0.0006	t=0.9910, P=0.3348

Table 4. Post-hoc t-tests to study workload, mental demand, and frustration imposed by the interfaces (* indicates a significant test).

On the basis of the results reported in the literature we expected browsing to be slower than the other two modalities since the search goal was known. This expectation was confirmed in our experiments. The participants were much slower with typing than speech during Session 1. However, in Session 2, they made a significant improvement with typing.

The improvement was not so significant with speech. We conjecture that, with more trials, typing will improve until it is no longer significantly slower than speech. It is unlikely that this effect will be observed with browsing, because, unlike typing and speech, browsing does not involve any learning. The only part of browsing that may involve learning is the structure of the hierarchy. However, it is unclear how much this knowledge will help the shopper if new tasks are presented to the shopper, i.e., the tasks requiring to use previously unexplored parts of the hierarchy.

	Browsing	Typing	Speech
Mental Demand	45.6	35.9	13.4
Frustration	47.8	1.8	34
Overall Workload	12.88	8.33	7

Table 5. Mean values of mental demand, frustration, and workload.

Unlike browsing, typing and speech involve some learning due to several factors, such as using the multi-tap keypad, speaking clearly into the microphone, and many other search-specific strategies. For example, we observed that while typing and speaking, the participants understood, after a few trials, that using the product's special description for the search narrowed down the results much faster. They also gradually learned they saved time by typing partial keywords, as the trailing characters in a keyword often left the results unchanged.

Though browsing provided features like jumping forward/backward in the current level, localizing, changing speed of text-to-speech synthesis, none of the participants used those features. When the search target is known, pure browsing is cumbersome, because it involves traversing a large hierarchy and guessing the right categories for the target.

The administration of the NASA TLX to the participants revealed that in spite of the significantly slower performance with typing as compared to speech, the workload imposed by the two modalities did not differ significantly. Browsing imposed a significantly higher

workload than either typing or speech. Browsing and typing were significantly more mentally demanding than speech. It was surprising that in spite of the low mental demand, speech caused significantly more frustration than typing. User comments, informally collected after the administration of NASA-TLX, revealed speech recognition errors to be the reason behind the frustration. Though the participants expressed the desire for an hybrid interface, in absence of one, most participants (9 out 10) indicated in their comments that they would prefer just typing.

7. Conclusion

This paper discussed user intent communication in robot-assisted shopping for the blind. Three intent communication modalities (typing, speech, and browsing) are evaluated in a series of experiments with 5 blind and 5 sighted, blindfolded participants on a public online database of 11,147 household products. The mean selection time differed significantly among the three modalities, but the lack of interactions indicated that the modality differences did not vary significantly between blind and sighted, blindfolded groups, nor among individual participants. Though it was seen that speech was the fastest, in real life, the shopper may prefer to use typing as it helps to be more discrete in a public place like a supermarket. A hybrid interface might be desirable. If the exact intention is not known, i.e. when the shopper does not know what she wants to buy, an interface with a strong coupling of browsing and searching is an option. Since it is difficult to evaluate how such a hybrid interface would perform in real life, evaluating the components independently, as was done in this paper, gives us insights into how user intent should be communicated in robot-assisted shopping for the blind.

8. Acknowledgements

This research has been supported, in part, through NSF grant IIS-0346880 award. We would like to thank all participants for volunteering their time for experiments. We are grateful to Dr. Daniel Coster of the USU Department of Mathematics and Statistics for helping us with the statistical analysis of the exeperimental data.

9. References

Gharpure, C. (2008). *Design, Implementation and Evaluation of Interfaces to Haptic and Locomotor Spaces in Robot-Assisted Shopping for the Visually Impaired,* Ph.D. Thesis, Department of Computer Science, Utah State University, Logan, UT, USA.

Kulyukin, V., Gharpure, C., and Coster, D. (2008). Robot-Assisted Shopping for the Visually Impaired: Proof-of-Concept Design and Feasibility Evaluation. *Assistive Technology*, Volume 20.2/Summer 2008, pp. 86-98. RESNA Press.

Kulyukin, V.; Gharpure, C. & Nicholson, J. (2005). Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, Canada, July 2005, IEEE.

Nicholson , J. & Kulyukin, V. (2007). Shoptalk: Independent blind shopping = verbal route directions + barcode scans. *Proceedings of the 2007 Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) Conference*, avail. on CDROM, Phoenix, AZ, USA, 2007, RESNA.

Kulyukin, V. and Gharpure, C. (2006). Ergonomics-for-One in a Robotic Shopping Cart for the Blind. *Proceedings of the 2006 ACM Conference on Human-Robot Interaction (HRI 2006)*, pp. 142-149. Salt Lake City, UT, USA, March 2006, ACM.

Kulyukin, V. (2007). Robot-Assisted Shopping for the Blind: Haptic and Locomotor Spaces in Supermarkets (Extended Paper Abstract). *Proceedings of the AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics Stanford University*, pp. 36-38. Palo Alto, California, March 26-28, 2007, AAAI Press.

Gharpure, C. & Kulyukin, V. (2008). Robot-Assisted Shopping for the Blind: Issues in Spatial Cognition and Product Selection. *International Journal of Service Robotics*, Volume 1, Number 3, July 2008, DOI 10.1007/s11370-008-0020-9, Springer.

Nicholson, J., Kulyukin, V., and Coster, D. (2009). ShopTalk: Independent Blind Shopping Through Verbal Route Directions and Barcode Scans. *The Open Rehabilitation Journal*, ISSN: 1874-9437 Volume 2, 2009, DOI 10.2174/1874943700902010011.

Wasson, G., Sheth, P., Alwan, M., Granata, K., Ledoux A., Ledoux, R., & Huang, C. (2003). User Intent in a Shared Control Framework for Pedestrian Mobility Aids. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS 2003)*, pp. 2962 – 2967, Las Vegas, NV, USA, October 2003, IEEE.

Demeester, E., Huntemann, A., Vanhooydonck, D., Vanacker, G., Degeest, A., Van Brussel, H., & Nuttin, M. (2006). Bayesian Estimation of Wheelchair Driver Intents: Modeling Intents as Geometric Paths Tracked by the Driver. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS 2006)*, pp. 5775-5780, Beijing, China, October 2006, IEEE.

Morency, L. P., Sidner, C., Lee, C. & Darrell, T. (2007). Head Gestures for Perceptual Interfaces: The Role of Context in Improving Recognition. Artificial Intelligence, Volume 171, pp. 568-585, Elsevier.

Fagg, A., Rosenstein, M., Platt, R., & Grupen, R. (2004). Extracting user intent in mixed initiative teleoperator control. *Proceedings of the American Institute of Aeronautics and Astronautics Intelligent Systems Technical Conference*, Chicago, IL, USA, September 2004, AIAA.

Raman, T. V. (1997). Auditory User Interfaces, Kluwer Academic Publishers, ISBN: 0-7923-9984-6Boston, USA.

Smith, A., Cook, J., Francioni, J., Hossain, A., Anwar, M., Rahman, M. (2004). Nonvisual tool for navigating hierarchical structures. *Proceedings of the ACM SIGACCESS Accessibility and Computing Conference*, pp. 133-139, Atlanta, GA, USA, October 2004, ACM.

Walker, B., Nance, A. & Lindsay, J. (2006). Spearcons: speech-based earcons improve navigation performance in auditory menus. Proceedings of the 12th International

Conference on Auditory Display (ICAD2006), pp. 63-68, London, UK, 2006, CS Department, Queen Mary, University of London, UK.

Brewseter, S. (1998). Using nonspeech sounds to provide navigation cues, *ACM Transactions on Human-Computer Interaction*, Volume 5, Issue 3, September 1998, pp. 224-259, ISSN: 1073-0516, ACM.

Gaver, W. (1989). The SonicFinder: An interface that uses auditory icons, *Human Computer Interaction*, Volume 4, Number 1, pp. 57-94, July 1989, ACM, ISSN: 0736-6906.

Divi, V., Forlines, C., Gemert, J., Raj, B., Schmidt-Nielsen, B., Wittenburg, K., Woelfel, P., & Zhang, F. (2004). A Speech-In List-Out Approach to Spoken User Interfaces, *Proceedings of Human Language Technologies*, Boston, MA.

Wolf, P., Woelfel, J., Gemert, J., Raj, B., & Wong, D. (2004). *Spokenquery: An alternate approach to choosing items with speech*. Mitsubishi Electric Research Laboratory, TR-TR2004-121., Cambridge, MA, USA, 2004.

Sidner, C. & Forlines, C. (2002). Subset language for conversing with collaborative interface agents. Mitsubishi Electric Research Laboratory, TR-TR2002-36., Cambridge, MA, USA, 2002.

Brewster, S., Lumsden, J., Bell, M., Hall, M., Tasker, S. (2003). Multimodal 'eye-free' interaction techniques for wearable devices, *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 473-480, Ft. Lauderdale, Florida, USA, ACM, ISBN: 1-58113-630-7.

K. Crispien, K. Fellbaum, A. Savidis, & C. Stephanidis. (1996). A 3d-auditory environment for hierarchical navigation in non-visual interaction. Proceedings of International Conference on Auditory Displays (ICAD), November 1996, ACM.

Hiipakka, J. & Lorho, G. (2003). A spatial audio user interface for generating music playlists. *Proceedings of the 2003 International Conference on Auditory Display*, Boston, MA, USA, July 2003.

W3C. (2003). Web content accessibility guidelines 1.0. In Web Accessibility Initiative.

Manber, U., B. Gopal, B., & Smith, M. (1996). Combining browsing and searching. *Proceedings of the W3 Distributed Indexing/Searching Workshop*, MIT, Boston, USA.

Mackinlay, J. & Zellweger, P. (1995). Browsing vs search: Can we find a synergy? (panel session). *Proceedings of the International Conference on Computer Human Interaction (SIGCHI)*, Palo Alto, CA, USA.

Karlson, A.; Robertson, G. ; Robbins, D. ; Czerwinski, M. & Smith, G. (2006). Fathumb: A facet-based interface for mobile search. *Proceedings of the International Conference on Computer Human Interaction (CHI)*, Montreal, Quebec, Canada, 2006.

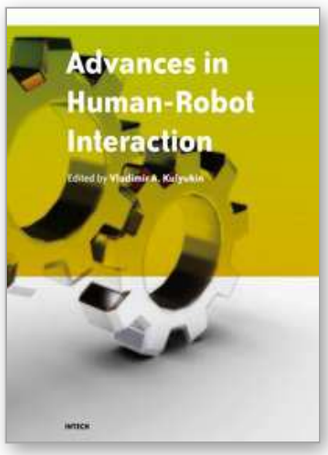
Divi, V.; Forlines, C.; van Gemert, J.V.; Raj, B.; Schmidt-Nielsen, B.; Wittenburg, K.; Woelfel, J.; Wolf, P. & Zhang, F. (2004) A Speech-In List-Out Approach to Spoken User Interfaces, *Proceedings of the Human Language Technology Conference*, May 2004 (HLT 2004), Boston, MA, May 2004, ACM.

Household Product Database. www.householdproducts.nlm.nih.gov, 2004.

Li, B.; Xu, Y. & Choi, J. (1996). Title of conference paper, *Proceedings of xxx xxx*, pp. 14-17, ISBN, conference location, month and year, Publisher, City

Siegwart, R. (2001). Name of paper. *Name of Journal in Italics,* Vol., No., (month and year of the edition) page numbers (first-last), ISSN

Arai, T. & Kragic, D. (1999). Name of paper, In: *Name of Book in Italics,* Name(s) of Editor(s), (Ed.), page numbers (first-last), Publisher, ISBN, Place of publication



Advances in Human-Robot Interaction

Edited by Vladimir A. Kulyukin

ISBN 978-953-307-020-9
Hard cover, 342 pages
Publisher InTech
Published online 01, December, 2009
Published in print edition December, 2009

Rapid advances in the field of robotics have made it possible to use robots not just in industrial automation but also in entertainment, rehabilitation, and home service. Since robots will likely affect many aspects of human existence, fundamental questions of human-robot interaction must be formulated and, if at all possible, resolved. Some of these questions are addressed in this collection of papers by leading HRI researchers.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen