Designing Speech-Based Interfaces for Telepresence Robots for People with Disabilities

Katherine M. Tsui, Kelsey Flynn, Amelia McHugh, and Holly A. Yanco
University of Massachusetts Lowell
Lowell, MA 01854
Email: {ktsui, kflynn, amchugh, holly}@cs.uml.edu

David Kontak
Crotched Mountain Rehabilitation Center
Greenfield, NH 03047
Email: david.kontak@crotchedmountain.org

***Abstract*—People with cognitive and/or motor impairments may benefit from using telepresence robots to engage in social activities. To date, these robots, their user interfaces, and their navigation behaviors have not been designed for operation by people with disabilities. We conducted an experiment in which participants (n=12) used a telepresence robot in a scavenger hunt task to determine how they would use speech to command the robot. Based upon the results, we present design guidelines for speech-based interfaces for telepresence robots.**

I. Introduction

Designing a telepresence user interface for people with cognitive and/or motor impairments is a first step towards having our target population take the active role of the robot operator. Telepresence robots may then be used as a means for social engagement. For example, some people may wish to tour a museum or attend an art exhibit opening, concert, sporting event, or theatre performance [1]. Others may simply want to be present in a space to feel more included in an activity, like attending high school via telepresence robot (see [2]).

Telepresence robot systems have been used as healthcare support tools (see [2] for a brief overview); the design of the robots and their interfaces has been focused on the doctor, healthcare staff, or family caregiver. Many commercial telepresence robot interfaces are designed for teleoperation from a computer using a combination of key presses and/or mouse clicks, mostly with low level forward, back, left, and right (FBLR) commands (e.g., Giraff, QB, R.BOT 100, Texai, VGo). For example, a robot will move forward when the up arrow key is pressed, remain moving forward until the key is released, and then stop. Input devices for these commercial telepresence robots require a fine level of manual dexterity, which may not be suitable for people with motor impairments [3]. It is difficult to keep a telepresence robot driving straight down a hallway, rather than zig-zagging, when using teleoperation due to network lag and dynamic environments [4]. Additionally, people with cognitive impairments may have difficulty decomposing complex tasks [5].

Our research investigates how people with cognitive and motor impairments would want to a direct telepresence robot in a remote environment, and specifically focuses on a speech-based interface. A number of corpora have been constructed investigating spoken spatial commands between people (e.g., [6]–[9]), and spoken by a human to a robot (e.g., [10], [11]). However, none of the corpora and data sets involved people with disabilities as participants, who are often overlooked when developing interfaces for human-computer interaction (HCI) [12]. It is unknown how our target audience, particularly those with cognitive impairments, would want to direct robots in a remote environment. Also, there is an underlying assumption in the human-human corpora that people will talk to robots in the same manner that people talk with other people (e.g., [6]–[9]). We conducted an experiment to investigate the differences and similarities between participants from the target audience giving spatial commands to a person versus a remote robot, which was perceived to be autonomous through "Wizard of Oz" control [13].

II. Scavenger Hunt Experiment

We designed an experiment in which people gave verbal instructions to guide a remote shopping assistant within a space similar to a retail store. Our goal was to see how the language used changed if the participants thought that 1) they were commanding an autonomous robot in the environment, or 2) a person was moving a camera in the environment for them.

A. Recruitment and Participants

We recruited 12 participants for our between-subjects study. Participants were members of the Crotched Mountain Rehabilitation Center community, including inpatient clients from the Brain Injury Center and participants in the residential program. Each spoke English and had a condition that significantly limited their ability to travel and maintain contact with important individuals. Their medical conditions included amyotrophic lateral sclerosis (ALS), cerebrovascular accident (CVA, or stroke), muscular dystrophy (MD), spina bifida (SB), spinal cord injury (SCI), traumatic brain injury (TBI). We excluded people with blindness, low arousal levels, or other conditions preventing benefits from use of the robot. People with severe cognitive challenges who were unlikely to understand that the robot was a representation of themselves, as opposed to a TV show or video game, were also excluded.

Ten men and two women participated in the experiment; the average age was 40.3 years (SD=17.3). A summarized description of their abilities is given in Table I. All participants had functional vision ability; one participant had visual field loss on his right side (P9) and another had significant hemispheric neglect (P10). Eleven participants had intact literacy ability; P7 had moderate literacy. All participants had intact speech ability. P2 and P8 used tracheostomy tubes for assisted respiration. P5 drew oxygen from a tube and had limited breath support;

TABLE I
PARTICIPANT DESCRIPTIONS

P#	Age	Gender	Medical Condition	Cognitive Impairment	Literacy Ability	Speech Ability	Vision Ability
P1	36	M	TBI	moderate	intact	intact	functional when corrected with glasses
P2	63	M	ALS	none	intact	intact; uses trachostomy tube	functional when corrected with glasses
P3	67	M	TBI	moderate	intact	intact	functional
P4	45	M	TBI	moderate	intact	intact	functional when corrected with glasses
P5	32	M	MD	none	intact	intact; limited breath support	functional
P6	24	F	TBI	moderate	intact	intact	functional when corrected with glasses
P7	20	F	SB	mild	moderate	intact	functional
P8	22	M	SCI	none	intact	intact; used trachostomy tube	functional
P9	32	M	TBI	mild	intact	intact	visual field loss on right side
P10	52	M	TBI	moderate	intact	intact	functional vision when corrected with glasses; significant hemispheric neglect
P11	64	M	CVA	moderate	intact	intact; clarity slightly affected	functional when corrected with glasses
P12	27	M	TBI	moderate	intact	intact	functional

consequently, he had a quiet voice. P11 had a cerebral stroke, which slightly affected the clarity of his speech.

P2, P5, and P8 had intact cognition. P7 and P9 had mild cognitive challenges (able to function in most environments independently and execute activities of daily living with assistance from memory aids). The remaining seven had moderate cognitive challenges (may have significant memory loss but able to perform most activities of daily living with minimal support, except cooking and bathing for safety reasons).

Four participants had prior experience with voice recognition. P2 and P8 had used Dragon NaturallySpeaking, and P12 used Dragon Dictate. P5 noted that he had experience with a voice activated computer but did not list specific software.

B. Experimental Design

Participants provided their informed consent or assent, as appropriate. Then, the experimenter read a script describing the remote shopping experience. The premise was that participants were to host a themed party. They would shop for costumes, food, drinks, party games, and movies to show at their parties (Task 1, 15 min.) with the help of a remote shopping assistant located at "KAS Party Central." The remote shopping assistant would show the store to the participants using a webcam. At the end of the scenario, they would finalize their choices at a checkout station with a party planner (Task 2, 5 min.).

The live video from the webcam on our VGo robot, Margo (Fig. 1 left), was maximized on a 22 inch (55.9 cm) Dell monitor. The webcam was described as having 2-way audio, so that they could hear what was going on at the store, and the remote shopper could hear anything they said to it. Participants were told to talk to tell the shopper where to go and what they were looking for. Participants were also told that there was a few second delay between when they spoke and when the remote shopper heard them.

The experimenter provided the participants with a store directory (Fig. 1 center) and a written description of the two tasks, including the shopping list (i.e., 2 types of snacks, 1 drink, 1 movie as recommended by a party planner, 1 party game, 1 costume). After completing both tasks, participants were asked to draw their path on a map and engaged in a post-experiment interview. At the end of the session, the telepresence robot and wizard were introduced to the participant, and the researchers answered any remaining questions about the study. Estimated time for this study was 60 minutes, and participants were compensated with a $20 gift certificate.

It should be noted that a training period was not provided as the purpose of this study was to understand how people provide verbal navigation instruction. Our experiment featured a single wizard whose role was to interpret the participants' verbalized navigation instructions. The wizard then controlled the telepresence robot in accordance with the given instruction. For both the robot-agent and human-agent conditions, the instructions were interpreted in the same manner, and the wizard moved the VGo telepresence robot using a USB gamepad.

C. KAS Party Central

We divided a space (approximately 20×40 ft (6.1×12.2 m)) into four sections. There were two large sections which each had two subsections; the snacks and drinks were located in the "grocery" section, and the movie posters and party games in the "entertainment" section (Fig. 1 center). The party planning station was co-located with the checkout in an adjacent room. The costumes were grouped together in one section. Large, green signs denoted the costume, grocery, entertainment, and party planning sections, and smaller, blue signs denoted the subsections within grocery and entertainment (Fig. 1 right). A call box was set in the middle of the store, out of the line-of-site of the party planning station.

We populated the store with party refreshments and theme items. There were multiple images associated with each choice, as the participants may not have been familiar with any one in particular. It was not important for the participant to select one specific image within a choice. Fig. 1 (top right) shows examples of the two types of the four drink choices (orange soda, fruit punch, milk, and water) on one side of the grocery section. The snack choices were cupcakes, cookies, pretzels, and apples. We created four party themes (i.e., robot, Halloween, Christmas, circus), and each theme was assigned three times. There was one choice for each theme's costume and party game, and three movie choices per theme. Costumes were comprised of a t-shirt and a mask, wig, or hat (Fig. 1 top right). Movie choices were shown as large poster and grouped according to theme (Fig. 1 bottom right).

Fig. 1. (**Left**) Margo, a modified VGo telepresence robot [14]. (**Center**) "KAS Party Central" store directory. Numbered triangles indicate location and orientation for "go to named destination" instructions for the wizard. (**Right**) View of the games table and movies. Best viewed in color.

D. Wizarding

We anticipated that the participants would use a wide range of language including low level FBLR directives (e.g., turn right, go right), relative descriptions using local area information (e.g., take the next left), and global destinations (e.g., go to a named location). In addition to the robot moving as directed by the participant, our wizard provided a limited amount of scripted verbal feedback. Koulouri and Lauria [15] found that when a small set of limited feedback is provided to a robot operator, the operators reverted to giving low-level FBLR commands, ignored the feedback, and focused on the robot's movement. In another experiment, their wizard was additionally allowed to "request clarification" from and "provide information" to the robot operator in an open text format. Our verbal protocol expanded on this level of feedback, which we detail as we define how the wizard controlled the robot.

In general, the wizard rephrased the participant's command with acknowledgement. For example, our wizard responded "going to <named destination>" when initiating movement, and "Ok, I'm here" when the robot was positioned and oriented. The choices for each of the shopping list items were not known to the participants beforehand; the shopping assistant started each run as if the list were unknown to the robot or person as well. Thus, "go to <named object>" and "find the <named object>" instructions were not valid (e.g., "go to the robot movies," "find the cupcakes"), as prior knowledge about the environment was not a dependent variable in this experiment. The wizard responded "I don't understand" or "I don't know" to invalid or ambiguous instructions.

FBLR command actions were set according to what people would expect For example, "turn right" resulted in the robot turning at most 90 degrees, as opposed to turning to the right continuously until commanded to stop. If the wizard was instructed to drive for a potentially long distance (e.g., "drive straight," "drive forward," "drive down the hall"), the wizard followed the instruction until a wall, shopping display, or obstacle was encountered. The wizard did not provide verbal feedback for FBLR directives.

Our verbal protocol included a number of additional feedback messages. When a participant indicated an item selection, the wizard responded by saying "picture taken." The wizard prompted the participant at the beginning of Task 1 ("how can I help you?") if he or she did not initiate instruction, and also during the task if the participant was silent for 60 seconds after the last robot movement or command ("what would you like me to do?"). If a participant chained multiple commands, the wizard acknowledged the sequence, rephrased and acknowledged the first command (e.g., "going to ..."), acknowledged the completion of the first command, rephrased and acknowledged the second command (e.g., "now going to ..."), and so on. The wizard incorporated awkward silences between words in the robot agent condition; otherwise, the wizard used her regular speaking voice.

E. Data Collection and Analysis

We recorded video and audio for each session, and the recordings were transcribed using CastingWords [16]. We developed and refined a categorical coding scheme through open and axial coding [17] based on the participants' utterances (Table II). Utterances are separated by a verbal response from the wizard, the start of the command action by the wizard, or elapsed silence by the participant of at least 10s. Cohen's kappa for inter-rater reliability was computed as κ=0.86 (excluding chance).

Eleven of the twelve participants completed the primary shopping task within the 15 minutes allotted; we have removed P10 from the statistical analysis due to non-completion of the task. We coded 312 total utterances. As utterances could contain more than one sentence or phrase, we considered the whole utterance and noted all appropriate categories. Participants in the human agent condition (n_H=6) spoke a total of 178 utterances ($\bar{x}_H$=29.67, SD=12.88), and participants in the robot agent condition (n_R=5) spoke a total of 134 utterances ($\bar{x}_R$=26.80, SD=15.12). Unless otherwise noted, we computed two-tailed Student's unpaired t-tests with a confidence interval of 95% (α=0.05) on the categorial frequency count between the human agent and robot agent conditions.

III. Results and Discussion

Overall, speech used to direct the human remote shopper and the robot remote shopper had few statistically significant differences (Table III). We believe that this result is due to the experimental design. In both conditions, participants were given a very simple and limited description of the remote shopper's capabilities. We did not provide details in the scenario description as to how the remote shopper would

TABLE II
CATEGORY CODING DEFINITIONS; κ=0.86 (EXCLUDING CHANCE)

Category	Description	Examples
ENVIRONMENTAL KNOWLEDGE		
None	The requested command can be done regardless of the robot's location (disregarding obstacle avoidance)	Stop, forward, back, left, right, tilt, zoom
Local	The requested command requires information from local sensors (i.e., camera view)	Take a picture
Global	The requested command requires knowledge beyond local sensors	Go to <named destination> not in the camera's current field of view
SENTENCE COMPLEXITY		
Simple	One independent clause:[1] simple or compound subject and simple or compound verb	Go to the snacks.
Compound	Two independent clauses joined using: for, and, nor, but, or, yet, so	Go to the snacks and turn right.
Complex	One independent clause joined with one or more dependent clause(s) using a subordinating conjunction (after, although, as, because, before, if since, though, unless, until, when, whenever, where, whereas, wherever, while) or relative pronoun (that, what, who, which)	Go to the snacks, which are on the other side of groceries from drinks.
SENTENCE TYPE		
Declarative	Sentence that makes a statement	Ex. 1: I need a snack. Ex. 2: The first item on the list is a snack. Ex. 3: I choose pretzels.
Imperative	Expresses a command, request, or selection. Subject may be implicit ("you") or explicit (proper name of remote shopper). Verb may be implied when the predicate is only an adverb phrase or direct object.	Ex. 1: [You,] go to the snacks. Ex. 2: [You, turn] left. Ex. 3: [You, go to] snacks, please.
Interrogative	Sentence that asks a question	Could you go to the snacks?
Interjection	A single word or non-sentence phrase which is not grammatically related to the rest of the sentence	Ok, all right, yes, well, hi, thanks
FEEDBACK TO REMOTE SHOPPER		
Praise	Utterance includes positive feedback given to remote shopper when beginning, while executing, or completing an instruction.	Good, excellent
Confirmation/ acknowledgement	Utterance includes neutral feedback given to remote shopper when beginning, while executing, or completing an instruction.	Ok, yes
SOCIAL ETIQUETTE		
Greeting	Utterance includes acknowledgment of remote shopper.	Hello, hi
Expressing polite request	Utterance includes "please."	Please
Expressing polite gratitude	Utterance includes "thank you."	Thank you, thanks
Addressing by name	Utterance includes the remote shopper's name.	Margo (robot), Kelsey (human)
OTHER		
Not to remote shopper	The utterance is not directed at or spoken to the remote shopper.	
No code	There is no appropriate category for this utterance.	

[1]A clause has a subject and a predicate; a predicate minimally contains a verb. An independent clause is a sentence.

move in the environment, if the remote shopper implicitly knew to avoid obstacles, and if the remote shopper knew what items were located in the store and where they were located. Additionally, we did not ask the participants any pre-experiment questions to prevent biasing their language style. We solicited the participants' experience with voice recognition at the end.

The differences in the participants' verbal instructions can largely be attributed to personal style. Some participants in both conditions primarily gave imperative commands that did not require any environmental knowledge to fulfill the request (Fig. 2 left: left, right, turn, stop). P2 in the robot agent condition gave three times as many commands that did not require environmental knowledge (n=33) than global ones (n=11), and P5 in the human agent condition gave more than four times as many (n=40 and 9, respectively). All participants utilized commands that required global environmental knowledge (Fig. 2 right: go to). P3 (human agent condition) primarily used declarative language, while P6 and P8 (robot agent condition) used interrogative language. Two participants, P2 and P5 (one in either condition), praised the remote shopper's completed actions. The majority of the utterances were simple sentences; two participants, P5 and P6 (one in either condition), provided the majority of the compound and complex language.

A. Declaring Item Selection

Participants in the human agent condition spoke significantly more declarative utterances (n_H=67, $\bar{x}_H$=11.2, SD=8.2) than participants in the robot agent condition (n_R=13, $\bar{x}_R$=2.6,

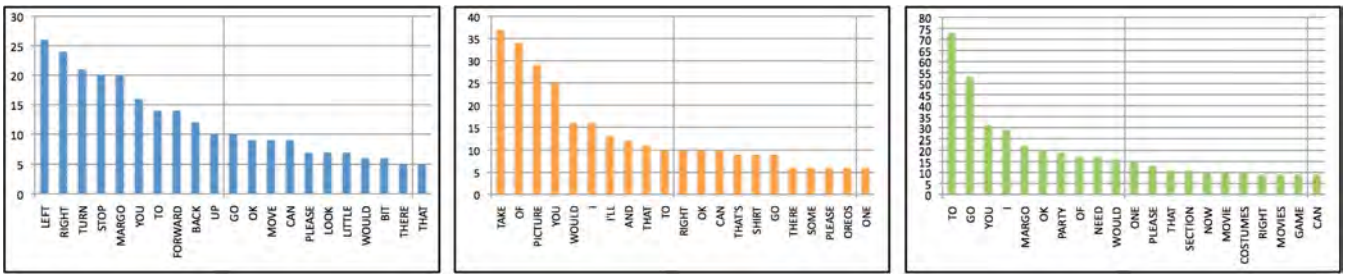

Fig. 2. Unique word histograms of top 21 utterances coded by levels of environmental knowledge. (**Left**) None: Directives require no environmental knowledge, and the request can be performed regardless of the robot's location; n=384, 67% representation. (**Center**) Local: Directives require information from the robot's sensors at an instance in time; n=573, 51%. (**Right**) Global: Directives require knowledge beyond the robot's local sensors; n=802, 51%.

TABLE III
RESULTING FREQUENCY COUNTS FROM CATEGORICAL CODING OF PARTICIPANTS' TRANSCRIPTS, EXCLUDING P10

Category	Condition	n	$\bar{x}$	SD	p
ENVIRONMENTAL KNOWLEDGE					
None	H	51	8.5	15.5	0.75
	R	57	11.4	13.4	
Local	H	46	7.7	3.2	0.55
	R	46	9.2	4.6	
Global	H	45	7.5	3.7	0.85
	R	39	7.8	0.4	
SENTENCE COMPLEXITY					
Simple	H	134	22.3	11.3	0.67
	R	132	26.4	17.8	
Compound	H	6	1.0	2.0	0.70
	R	3	0.6	1.3	
Complex	H	18	3.0	1.9	0.56
	R	9	1.8	4.0	
SENTENCE TYPE					
Declarative	H	67	11.2	8.2	**0.05***
	R	13	2.6	1.9	
Imperative	H	79	13.2	18.0	0.71
	R	89	17.8	21.8	
Interrogative	H	22	3.7	5.2	0.24
	R	49	9.8	9.3	
Interjection	H	79	13.2	6.9	**0.02***
	R	18	3.6	3.2	
FEEDBACK TO REMOTE SHOPPER					
Praise	H	5	0.8	2.0	0.83
	R	3	0.6	1.3	
Confirmation	H	58	9.7	4.8	**0.01****
	R	9	1.8	2.9	
SOCIAL ETIQUETTE					
Greeting	H	3	0.5	0.5	0.77
	R	3	0.6	0.5	
Please	H	8	1.3	2.2	0.21
	R	17	3.6	3.0	
Thank you	H	4	0.7	1.2	0.92
	R	3	0.6	0.9	
Address by name	H	6	1.0	0.9	0.30
	R	50	10.0	16.9	

SD=1.9); p=0.05 with t(9)=2.25. We found that 55 of the 67 declarative utterances in the human agent condition (82.1%) and 8 of the 13 in the robot agent condition (61.6%) were first person declarative sentences (e.g., "I want," "I need," "I'll take," "I'll choose"). Participants in the human agent condition spoke significantly more first person declarative utterances ($\bar{x}_H$=9.7, SD=7.2) than the participants in the robot agent condition ($\bar{x}_R$=1.6, SD=1.8); p<0.05 with t(9)=2.27. The functions of these first person declarative utterances related to item selection and specifying locations, which require local and global knowledge, respectively.

We further investigated the function of the commands that required local environmental knowledge, which included item selection by taking a picture. Seventy-seven of the 92 utterances that required local environment knowledge (83.7%) involved item selection. Participants in the human agent condition had a similar number requiring local knowledge (n_H=40, $\bar{x}_H$=6.7, SD=2.7) as the participants in the robot agent condition (n_R=37, $\bar{x}_R$=7.4, SD=2.7); p=0.67.

We then looked at the sentence type corresponding to item selection and found that participants spoke a total of 24 declarative utterances (n_H=13, n_R=1), 38 imperative (n_H=17, n_R=21), and 23 interrogative (n_H=6, n_R=17). Again, participants in the human agent condition spoke significantly more declarative utterances ($\bar{x}_H$=3.8, SD=3.1) than those in the robot condition ($\bar{x}_R$=0.2, SD=0.4); p=0.03 with t(9)=2.52. Participants in the robot agent condition spoke a greater number of imperative utterances ($\bar{x}_R$=4.2, SD=4.1) and interrogative utterances ($\bar{x}_R$=3.4, SD=3.2) than those in the human condition ($\bar{x}_{H_{imp}}$=2.8, SD=2.7; $\bar{x}_{H_{int}}$=1.0, SD=2.4, respectively), though not significantly so (p=0.54 and p=0.21, respectively).

Finally, we note that for all utterances involving item selections, all but four identified the specific item. Participants also provided descriptive information including the color of the item (e.g., P7: "I'll do the black shirt with the skeleton"), its location in the camera's field of view (e.g., P6: "And also the milk that's on the right side of the bottom?"), and its location with respect to other items (e.g., P9: "Up above the black shirt, there was a mask. Can you take a picture of that?").

B. Interjections Indicating Confirmation

In addition to participants in the human agent condition using more declarative statements, they also used more interjections ($\bar{x}_H$=13.2, SD=6.9) than those in the robot agent condition ($\bar{x}_R$=3.6, SD=3.2); p=0.02 with t(9)=2.90. Fifty-eight of the 79 (73.4%) interjections in the human agent condition were categorized as confirmation or acknowledgment feedback to the remote shopper, in contrast to 9 of the 18 (50%) interjections spoken to the robot shopper. This difference was also significant (p=0.01 with t(9)=3.25; $\bar{x}_H$=9.7, SD=4.8; $\bar{x}_R$=1.8, SD=2.9). We believe this difference is due to the perception of giving

instructions to a person versus robot. In both conditions, the wizard performed commands and provided feedback in the same manner. The webcam was used solely to provide the participant a view of the remote environment and not used to coordinate with the remote shopper in the human agent condition. Verbal acknowledgements are a compensatory strategy used in human-human remote collaboration (e.g., [7], [18], [19]).

C. Margo, stop!

It is imperative that the robot is able to stop on command. There were 23 utterances that contained the keyword "stop" (n_H=4, n_R=19). In 21 of these 23 utterances (91.3%), the participant was directing the remote shopper to cease the current movement. There was no statistical difference between the human agent condition ($\bar{x}_H$=0.3, SD=0.8) and the robot agent condition ($\bar{x}_R$=3.8, SD=6.9); p=0.36. Additionally, there were five utterances containing an implied stop command (n_H=4, n_R=1). Colloquialisms including "that's good" (n=1) and "[*blank*, hold it, stay] right there" (n=4) should also be given the same importance as "stop."

D. Social Etiquette

Six of the eleven participants greeted their remote shopper by name (n_H=3, n_R=3). Two participants introduced themselves by name as well: one in either condition (P9, P12). Four participants thanked their remote shoppers (n_H=4, n_R=3). P1 and P9 (human agent condition) said "thank you" once and three times, respectively; P2 and P12 thanked their robot shopper once and twice, respectively. Six participants said "please" a total of 26 times (human agent: n_{P3}=3, n_{P5}=5; robot agent: n_{P4}=2, n_{P6}=3, n_{P8}=5, n_{P12}=8). There was no significant difference between the human agent ($\bar{x}_H$=1.3, SD=2.2) and robot agent ($\bar{x}_R$=3.6, SD=3.0) conditions; p=0.20.

There were 47 instances in which a participant addressed the robot by its name, in addition to the three greetings. Addressing the robot by name was one strategy for giving a new command. Three participants said "Margo" in this manner (n_{P2}=40, n_{P6}=2, n_{P8}=2). It was also used for checking if the robot was still awaiting commands by P6 and P8 (n=3); participants gave a subsequent request following the robot's acknowledgement. In the human agent condition, the remote shopper's name "Kelsey" was spoken in 6 instances: three times as a greeting and three times at the start of a declarative sentence (P1, P3). There was no significant difference between the human agent and robot agent conditions (p=0.30). In a one-to-one scenario, addressing the remote shopper by name may have been considered superfluous; vocative expression is an explicit manner of direct address which people use in group situations [20].

IV. Guidelines for HRI Design

The use of a speech interface ultimately depends on a person's communication skills including comprehension, vocabulary, speech clarity, and rate of speech. Speech recognition software, such as Dragon Naturally Speaking, has been utilized by people with physical disabilities who cannot use traditional physical computer access methods (e.g., [21]) and people with mild, moderate, and even severe speech impairments (e.g., [22]–[24]). Finally, speech interfaces overall are increasing in popularity. Voice-activated personal assistants exist on contemporary smartphones (e.g., Apple's Siri [25], Android's Andy [26]). Based upon the results of our experiment, the following guidelines should be used for creating interfaces for remote telepresence robots:

Guideline 1: Levels of feedback expected from the remote robot and given back to the robot are not equal. Verbal responses and the robot's actions provide feedback that the given commands have been understood (or not). Participants were explicitly told of delays between giving an instruction and the remote shopper hearing it; Green et al. [27] note that the time elapsed to a system's response may lead to repeating the desired command. Sufficiently detailed feedback must be given by the robot in order to elicit spoken spatial navigation commands similar to those given by one person to another; otherwise, robot operators revert to FBLR directives [15]. In the post-experiment interview, P8 commented specifically on the robot's verbal feedback: "It [Margo] gave the feedback I asked it to do. [It] went to the place I asked. When it got there, it said it arrived to confirm."

However, the robot should not rely on feedback from a person for confirmation of its actions. Participants in the human agent condition spoke 58 utterances (of 312, 18.6%) indicating acknowledgement, as opposed to 9 utterances (2.9%) from participants in the robot agent condition. It should be noted that the remote shopper always acted in accordance with requests, thus it is unknown if or how a person might reprimand the robot (e.g., "no!" [28]) and correct the current command. Kim et al. [28] discuss how people are willing to provide feedback to a robot teaching it a task. Robot operators should not be expected to give the same level of explicit confirmation in order for the robot to act on every command (e.g., [27], [29]). Sugiura et al. [30] have begun investigating when and how to confirm a robot's pending action(s).

Guideline 2: Expect simple commands with phrase variations corresponding to the same underlying command. It is unrealistic to expect a user to memorize and recall verbatim large numbers of commands, destination labels, etc. Human mobility follows a power-law distribution (e.g., by vehicle [31], walking [32], activities of daily living at home [33]). We observed that the participants' word choices in their utterances also followed this pattern considering the participants gave navigation instructions in an unconstrained manner (Fig. 2).

We found that the majority of the utterances were categorized as simple sentences (85.3% overall). There were 312 total utterances; participants in the human agent condition spoke 134 simple sentences (42.9%), and those in the robot agent condition spoke 132 (42.3%). The majority of the types of sentences spoken to a robot in the scavenger hunt experiment were imperative (n_R=89); for example, "go to the snacks." Imperative commands can be rephrased into interrogative questions thereby increasing its politeness; for example, "could you go to the snacks?" Interrogative statements were the second most frequent utterance spoken to the robot (n_R=49). Additionally, imperative commands can be rephrased into

declarative sentences expressing desired functionality; for example, "I want you to go to the snacks" or "You can go to snacks." It was rare, however, for a participant to speak a declarative utterance to the robot (n_R=13), and, in this respect, there was a statistically significant difference between speaking to a robot versus to a human (n_H=67; p=0.05).

Guideline 3: Halt the robot whenever "stop" is parsed. The robot's safety should not be solely the responsibility of the user, and the robot should proactively keep from harming itself and the remote environment. Telepresence robot operators from our target audience may not be able to perceive environmental hazards (e.g., obstacles, cliffs) and comprehend if evasive action (e.g., stop) is needed in a timely manner depending on their cognitive ability [34]. Further, they may not be able to articulate the evasive action due to speech and/or language disorders; real-time speech generation is similar to real-time control of a physical device (e.g., power wheelchair).

However, it would be prudent to immediately halt motor commands when the keyword "stop" is parsed, even if the language processing is incomplete. Although "stop" was only spoken in less than 0.1% of the utterances (23 of 312), the resulting action of the robot ceasing its current movement would have been correct 91.3% of the time (21 of 23). We did not observe any instances in which a participant spoke a "stop" command with intense, negative affect [28] to prevent the robot from bumping into an obstacle or wall; only P8 stated that he "thought the robot might crash" in his interview.

Guideline 4: Be cognizant of the addressee. A voice command interface for giving spatial navigation commands does pose a "Midas touch" issue, as the primary purpose for telepresence robots is communication using two-way audio and video. There is potential for confusion as to whom the robot operator is speaking: the robot or a person in the remote environment. Dowding et al. [35] trained a language model on known robot-direct speech and investigated if robot-directed speech could be distinguished in predominantly human-human dialogues. Trafton et al. [19] have investigated perspective taking as a means to establish if the robot's perspective is the same as the person's. Although only a few participants addressed the remote shopper by name, it may be appropriate to leverage vocative expression to indicate direct address. There are a number of human-robot communications that utilize calling the robot by its name or simply "robot" as a conversation marker (e.g., [11], [19], [27], [36]). However, it should not be required to start every robot command with the robot's name.

V. Conclusions and Future Work

The primary focus of our current research with the telepresence robots is to assist users with spatial navigation tasks. We conducted a formative assessment of user expectation utilizing a participatory approach, performed an experiment with twelve participants from our target audience, and collected a corpus of their first-hand spatial commands. We can now begin to understand how people from our target audience would direct telepresence robots. Overall, our analysis of the corpus showed few statistically significant differences between speech used in the human and robot agent conditions. We believe that, for the task of directing a telepresence robot's movements in a remote environment, people will speak to the robot in a manner similar to speaking to another person.

One limitation of our study is the small sample size with respect to the number of participants, which prohibited analyzing the corpus of utterances using the participants' level of cognitive impairments and/or medical condition as an independent variable. Three participants had intact cognition, and nine had cognitive challenges ranging from mild to moderate, which we described in terms of a person's independent functional ability as opposed to specific aspect(s) of cognitive impairment (i.e., attention, memory, perception, organization, and/or abilities to problem-solve and conceptualize [37], [38]). Additionally, it should be noted that all of our study participants had intact speech ability. Cognitive-communication disorders are commonly associated with neurologically impaired medical conditions such as traumatic brain injury and stroke [38]. People with severe physical impairments (e.g., amyotrophic lateral sclerosis, cerebral palsy, muscular dystrophy) may also have speech disorders (e.g., dysarthria). The guidelines for developing speech-based robot interfaces described above are applicable to all telepresence robot systems, and may increase the ease of use for typically abled people as well.

Our corpus provides insights regarding the level of navigation functionality these robots are expected to have, specifically the ability to understand low level FBLR commands in addition to local interest points within the robot's field of view and global destinations. Robot teleoperation through a constrained view of the remote environment [39] is difficult and requires the operator to give low level FBLR commands while remembering the end goal and any subgoals. In our study, we found that participants in both the robot and human agent conditions used this approach. For a speech-based interface, low level directives can be cumbersome as a user may repeat the same directive several times (e.g., left, left, left), or rely on timing to give another low level command to interrupt the one executing (e.g., left, stop); P10 primarily employed this approach, but was not able to finish the task within the given time allocated.

Instead, higher level commands requiring more robot autonomy can be used as macros or shortcuts. This approach can reduce the overall task complexity and minimize the number of correctly sequenced directives given, provided that the user has the cognitive ability to both understand the robot's autonomous behavior(s) and appropriately make use of them. By investigating a speech-base interface and providing sufficient feedback in our Wizard of Oz study, we found that all participants elicited directives requiring global environmental knowledge, although the variety of utterances was limited given the scope of the experiment and simplistic store directory.

We believe that the understanding of a robots autonomous capabilities should be facilitated by the HRI interface presentation and system feedback. Our next step is to develop a supplementary augmented-reality graphical user interface that provides cognitive support for our target audience. Simple language and familiar real world analogies may allow robot

drivers recognize how to use the interface rather than having to recall how to use it from training and/or their own experience [40]. Hints about the robot's autonomous navigation capabilities and the robot's local and global environmental knowledge will be overlaid on the robot's video. For example, our telepresence robot interface will provide cognitive support by displaying the path to the robot's current global destination projected as an overlay on the robot's video in a manner similar to Google Street View [41] or a car GPS navigation system. Local points of interest within the robot's field of view could be highlighted and labeled with descriptive titles. We will then conduct a usability case study with three users from our target audience to evaluate our telepresence robot's ease of use.

ACKNOWLEDGMENTS

This research has been funded by NSF (IIS-1111125). The authors would like to acknowledge Adam Norton, Eric McCann, Munjal Desai, Dan Brooks, Mikhail Medvedev, Jordan Allspaw, and Sompop Suksawat of UMass Lowell.

REFERENCES

[1] J. Beer and L. Takayama, "Mobile remote presence systems for older adults: Acceptance, benefits, and concerns," in *Proc. of ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*, 2011.

[2] K. Tsui, A. Norton, D. Brooks, H. Yanco, and D. Kontak, "Designing telepresence robot systems for use by people with special needs," in *Proc. of Quality of Life Technologies (QoLT) Intl. Symp.: Intelligent Systems for Better Living, held in conjunction with RESNA and FICCDAT*, 2011.

[3] K. M. Tsui, D.-J. Kim, A. Behal, D. Kontak, and H. A. Yanco, ""I want that": Human-in-the-loop control of a wheelchair-mounted robotic arm," *Journal of Applied Bionics and Biomechanics*, vol. 8, no. 1, 2011.

[4] K. Tsui, M. Desai, H. Yanco, and C. Uhlik, "Exploring use cases for telepresence robots," in *Proc. of ACM/IEEE Intl. Conf. on HRI*, 2011.

[5] K. Tsui and H. Yanco, "Prompting devices: A survey of memory aids for task sequencing," in *Proc. of Quality of Life Technologies (QoLT) Intl. Symp.: Intelligent Systems for Better Living, held in conjunction with RESNA*, 2010.

[6] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, 1991.

[7] K. M. Eberhard, H. Nicholson, S. Kbler, S. Gundersen, and M. Scheutz, "The Indiana "Cooperative Remote Search Task" (CReST) corpus," in *Proc. of Intl. Conf. on Language Resources and Evaluation (LREC)*, 2010.

[8] A. Gargett, K. Garoufi, A. Koller, and K. Striegnitz, "The GIVE-2 corpus of giving instructions in virtual environments," in *Proc. of LREC*, 2010.

[9] L. Stoia, D. Shockley, D. Byron, and E. Fosler-Lussier, "SCARE: A situated corpus with annotated referring expressions," in *Proc. of LREC*, 2008.

[10] G. Bugmann, E. Klein, S. Lauria, and T. Kyriacou, "Corpus-based robotics: A route instruction example," in *Proc. of Intelligent Autonomous Systems*, 2004.

[11] M. Marge and A. Rudnicky, "The TeamTalk corpus: Route instructions in open spaces," in *Proc. of RSS Wksp. on Grounding Human-Robot Dialog for Spatial Tasks*, 2011.

[12] E. Bergman and E. Johnson, "Towards accessible human-computer interaction," 2008, http://www.sun.com/accessibility/docs/access_hci.jsp. Accessed Aug. 2009.

[13] J. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Trans. on Information Systems*, vol. 2, no. 1, 1984.

[14] K. Tsui, A. Norton, D. Brooks, E. McCann, M. Medvedev, and H. Yanco, "Design and development of two generations of semi-autonomous social telepresence robots," in *Proc. of IEEE Conf. on Technologies for Practical Robot Applications (TePRA)*, 2013.

[15] T. Koulouri and S. Lauria, "A WOz framework for exploring miscommunication in HRI," in *Proc. of AISB Symp. on New Frontiers in HRI*, 2009.

[16] CastingWords, "Audio transcription services: MP3s, video, CD/DVDs," 2012, http://castingwords.com. Accessed Sept. 2012.

[17] B. Glaser and A. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. Aldine de Gruyter, 1967.

[18] D. Gergle, R. Kraut, and S. Fussell, "Action as language in a shared visual space," in *Proc. of Computer Supported Cooperative Work*, 2004.

[19] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 35, no. 4, 2005.

[20] H. Akkeer, "On addressee prediction for remote hybrid meeting settings," Master's thesis, University of Twente, 2009.

[21] R. DeRosier and R. Farber, "Speech recognition software as an assistive device: A pilot study of user satisfaction and psychosocial impact," *Work: A J. of Prevention, Assessment and Rehabilitation*, vol. 25, no. 2, 2005.

[22] C. Havstam, M. Buchholz, and L. Hartelius, "Speech recognition and dysarthria: A single subject study of two individuals with profound impairment of speech and motor control," *Logopedics Phonatrics Vocology*, vol. 28, no. 2, 2003.

[23] K. Hird and N. Hennessey, "Facilitating use of speech recognition software for people with disabilities: A comparison of three treatments," *Clinical Linguistics and Phonetics*, vol. 21, no. 3, 2007.

[24] C. Vaquero, O. Saz, E. Lleida, and W. Rodríguez, "E-inclusion technologies for the speech handicapped," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2008.

[25] E. Sadun and S. Sande, *Talking to Siri: Learning the language of Apple's intelligent assistant*. Que Pub, 2012.

[26] Andy (Android Siri Alternative), "Android voice control software!" 2013, http://www.andyforandroid.com. Accessed Feb. 2013.

[27] A. Green and K. Eklundh, "Designing for learnability in human-robot communication," *IEEE Trans. on Industrial Electronics*, vol. 50, no. 4, 2003.

[28] E. Kim, D. Leyzberg, K. Tsui, and B. Scassellati, "How people talk when teaching a robot," in *Proc. of ACM/IEEE Intl. Conf. on HRI*, 2009.

[29] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu, "Integrated natural spoken dialogue system of Jijo-2 mobile robot for office services," in *Proc. of AAAI/IAAI*, 1999.

[30] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura, "Active learning of confidence measure function in robot language acquisition framework," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010.

[31] M. Gonzalez, C. Hidalgo, and A. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008.

[32] I. Rhee, M. Shin, S. Hong, K. Lee, S. Kim, and S. Chong, "On the Levy-walk nature of human mobility," *IEEE/ACM Trans. on Networking (TON)*, vol. 19, no. 3, 2011.

[33] R. Aipperspach, E. Cohen, and J. Canny, "Modeling human behavior from simple sensors in the home," *Pervasive Computing*, 2006.

[34] M. Endsley, "Design and evaluation for situation awareness enhancement," in *Proc. of Human Factors and Ergonomics Society (HFES) Annual Meeting*, vol. 32, no. 2, 1988.

[35] J. Dowding, R. Alena, W. Clancey, M. Sierhuis, and J. Graham, "Are you talking to me? Dialogue systems supporting mixed teams of humans and Rrobots," in *Proc. of Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems: AAAI Fall Symp.*, 2006.

[36] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Robotics and Autonomous Systems*, vol. 42, no. 3, 2003.

[37] G. Vanderheiden and K. Vanderheiden, "Guidelines for the design of consumer products to increase their accessibility to persons with disabilities or who are aging," 1992, http://trace.wisc.edu/docs/consumer_product_guidelines/toc.htm. Accessed Oct. 2009.

[38] College of Audiologists and Speech-Language Pathologists of Ontario, "Preferred practice guideline for cognitive-communication disorders," Sept. 2002.

[39] M. Voshell, D. Woods, and F. Phillips, "Overcoming the keyhole in human-robot coordination: Simulation and evaluation," in *Proc. of HFES Annual Meeting*, vol. 49, no. 3. SAGE Publications, 2005.

[40] J. Nielsen, "Enhancing the Explanatory Power of Usability Heuristics," in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, 1994.

[41] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google Street View: Capturing the world at street level," *Computer*, vol. 43, no. 6, 2010.