On the Feasibility of Using a Standardized Test for Evaluating a Speech-Controlled Smart Wheelchair

Joelle PINEAU, Robert WEST, Amin ATRASH, Julien VILLEMURE, and François ROUTHIER

Abstract—Many people who have to rely on powered wheelchairs find it hard to fulfill daily navigation tasks with their chairs. The SmartWheeler project aims at developing an intelligent wheelchair that minimizes the physical and cognitive load required in steering it. In this paper we briefly outline the SmartWheeler project and its goals. We then argue that it is important to have a standardized test to evaluate intelligent wheelchairs in terms of performance and safety. No such test exists as yet for intelligent wheelchairs, but there has been an effort in the clinical community to design tests for conventional wheelchair usage. We discuss the existing Wheelchair Skills Test (WST). We then suggest a paradigm that allows us to use this test to benchmark the quality of intelligent wheelchairs, and in particular their interface, in a task context that is relevant to clinical practice in rehabilitation.

Index Terms—smart wheelchairs, human-robot interaction, assistive robotics, dialogue management

1. INTRODUCTION

ANY people who suffer from chronic mobility impairments, such as spinal cord injuries or multiple sclerosis, use a powered wheelchair to move around their environment. However, factors such as fatigue, degeneration of their condition and sensory impairments often limit their ability to use standard powered wheelchairs. It has been reported that as many as 40% of powered wheelchair users surveyed found daily steering and maneuvering tasks to be difficult; and according to the clinicians who treat them, nearly half of those patients unable to control a powered wheelchair by conventional methods would benefit from an automated navigation system [1].

Such numbers make it seem likely that intelligent wheelchairs catering to those patients' needs would have a deep societal impact. One might argue that the transition to wheelchairs that cooperate with the user is at least as important as that from manual to powered wheelchairs—possibly even more important since this would mark a paradigmatic rather than merely a technological shift. However, to the best of our knowledge, there is no general method of evaluating the performance of intelligent wheelchairs yet [2]. And in particular, no formal tools exist to evaluate the interaction between the intelligent wheelchair and its operator. In this paper, we try to make a first step by suggesting a methodology based on work done in the clinical rehabilitation community. In particular, we investigate the use of a specific corpus of tasks, as defined by the Wheelchair Skills Test (WST) [3], [4], [5]. The use of such a well-defined set of tasks (or skills) has many advantages for the objective evaluation of intelligent wheelchairs. It ensures the evaluation criteria is relevant to the end-user (since the task domain was originally defined for conventional wheelchair usage), it provides a repeatable evaluation protocol between test subjects, and it admits an objective performance measure.

We first describe the SmartWheeler project and the intelligent wheelchair developed by our research team. The rationale that supports the use of a standardized test and the relevant literature are exposed. Finally, we present results from the evaluation of the SmartWheeler's human-robot interaction architecture using the WST (version 4.1) protocol.

2. THE SMARTWHEELER PROJECT

The SmartWheeler project [6], [7] aims at developing in collaboration with engineers, rehabilitation clinicians and rehabilitation researchers—a prototype of a multi-functional intelligent wheelchair to assist individuals with mobility impairments in their daily locomotion, while minimizing physical and cognitive loads. The projected was initiated in 2006, and a first prototype, shown in Figure 1, was built in-house at McGill's Centre for Intelligent Machines.

This prototype is built on top of a commercially available Sunrise Quickie Freestyle powered wheelchair. The added robotic components include two SICK laser range-finders and two wheel encoders to permit autonomous navigation functionalities, as well as a multi-modal communication interface including a microphone and touch-sensitive display. The added components are not specific to the Sunrise platform, and could be easily installed on most commercially available powered wheelchairs.

A second intelligent powered wheelchair prototype was built in 2008 by research collaborators at Ecole Polytechnique de Montréal. This second robot, described in recent publications [8] and shown in Figure 3, has very similar hardware and software components. It uses the smaller Hokuyo laser range-finders, and thus the overall design is better suited to being used (and evaluated) by individuals with disabilities.

Most of the software components governing the autonomous navigation are being developed by some of our collaborators [8]. The main contribution of the authors, and the main subject of the evaluation presented below, is towards the

Manuscript received December 10, 2010; revised April 11, 2011. This work was supported by the National Science and Engineering Research Council Canada.

J. Pineau, R. West, A. Atrash and J. Villemure are with the Center for Intelligent Machines at McGill University, Montréal, Canada.

F. Routhier is with the Centre for interdisciplinary research in rehabilitation and social integration (CIRRIS) at Institut de réadaptation en déficience physique de Québec, Québec, Canada

development of the human-robot interface for the intelligent wheelchair. Figure 2 presents an overview of the software architecture controlling the human-robot interface on-board the robot. The primary mode of interaction is a two-way speech interface. Speech recognition is achieved through the commercially available Dragon NaturallySpeaking (version 9). We then employ a number of natural language technologies to achieve robust interaction, including automatic grammatical parsing [9] and high-level dialogue management using Partially Observable Markov Decision Processes [10]. A tactile/visual interface system is also installed, and used primarily for provide visual feedback to the human regarding the state of the dialogue system.



Fig. 1. The SmartWheeler robot platform.



Fig. 2. The SmartWheeler Interaction Architecture.

Speech provides a natural interface for human operators. It has been used as the primary mode of input in a few intelligent wheelchairs to date [2]. Yet speech-based interactions are subject to significant failure rates due to the noise and ambiguity inherent in speech-based communication, and it has not been clearly demonstrated whether this is an effective mode of communication for intelligent wheelchairs. In addition, it is important to realize that the choice of tasks and physical environment can significantly affect the performance of an automated dialogue system. Thus it is imperative to be able to carefully quantify the performance of our speech-based interface in the context of natural interactions and in a realistic environment. This is one of the primary contributions of the present paper.

3. IN SUPPORT OF STANDARDIZED TESTING FOR INTELLIGENT WHEELCHAIRS

All engineered research needs to be assessed in terms of the results it produces. In this section, we list the major reasons we see for adopting a standardized test for intelligent robotic wheelchairs.

In general, quantifying the performance and safety of a robotic device designed to aid people is a necessary step in the evaluation of its impact. A machine will only be accepted by people if it is of use to them. A rigorous, controlled evaluation is essential in the context of health-related projects like SmartWheeler because it provides a benchmark for proving that appropriate performance and safety requirements are met before a new technology, such as an intelligent powered wheelchair, can be deployed in real-world situations. This is especially true for intelligent powered wheelchairs, which will eventually act at least partly in an autonomous manner. As more control is taken from the user and given to the on-board intelligent system, it becomes more important to make guarantees about its performance and safety. Certainly a standard evaluation scheme is a crucial step if the use of intelligent powered wheelchairs is to be funded by public health services and insurance companies.

Also, one generally strives to supply a person with the wheelchair that fits them best. This is true for regular powered wheelchairs, and it applies equally to intelligent wheelchairs. For instance, certain features (e.g. an eye tracker) might be expensive, so one would like to dispense with them if they are not necessary. A standardized test might help figure out the best configuration for a user. Again, this will be essential for funding purposes.

Moreover, as more projects of the kind described above come into being, it will be helpful to benchmark the efficacy of the technologies employed. On the one hand, this can serve to assess how well the algorithms and hardware being developed within one research project work in a setting that is close to the real world, which can guide researchers towards the problems that have to be addressed next. On the other hand, a standardized test facilitates the comparison of similar projects by different research teams, thus highlighting the most promising approaches.

Finally, from a practical point of view, a standardized test can be helpful during the development process because it focuses the objectives of the research team. This has inherent limitations in the long-term. Nonetheless keeping the test in mind can help by providing a useful basis for thinking of possible deployment scenarios. For instance, a first English grammar for the natural language understanding component of a voice recognition system could cover the set of commands that represent the skills required in a standardized test.

4. Relevant literature

In a recent review of intelligent wheelchair projects Simpson concludes that, "[while] there has been a significant amount of effort devoted to the development of smart wheelchairs, scant attention has been paid to evaluating their performance. [...] Furthermore, no smart wheelchair has been subjected to a rigorous, controlled evaluation that involves extended use in real-world settings" [2]. Since 2005, two research teams have undertaken the task of formally evaluating their intelligent wheelchair prototype with the target population. In the first case, the evaluation protocol involved 30 disabled users which tested two tasks: forward motion through a corridor, and passing through a door [11], [12]. In the second case, the evaluation involved 17 able-bodied and 17 disabled subjects, and required each of them to complete a number of point-to-point navigation tasks (indoor and outdoor, through corridors and among obstacles), though the tests were restricted to small environments [13]. Yet both of these evaluations primarily used metrics that are typical of robotics research, and did not explicitly focus on clinically relevant protocols of evaluation.

If a test is to be used for the reasons listed above, it should be valid and reliable from a clinical point of view, i.e. it should actually measure what it is intended to measure, and do this in a reproducible way. Designing such a test can be difficult for a computer scientist or engineer lacking the necessary background in clinical rehabilitation. Yet several manual and powered wheelchair skills tests have been proposed in the literature, and Kilkens *et al.* [14] and Routhier *et al.* [15] fairly recently provided the first systematic overviews. In this section we will briefly summarize their results. In the following section we describe the test we chose for evaluating the SmartWheeler project, and how we are applying it in practice.

In their review, both Kilkens *et al.* [14] and Routhier *et al.* [15] come to the conclusion that no standard test to measure wheelchair skill or performance exists as yet, despite a considerable clinical and academic need for such a measure. From a clinical point of view, a standard test should allow for extrapolation of test results to assess subjects in their everyday activities, in order to guide training and facilitate the selection of a suited wheelchair. From an academic perspective, a standard test would alleviate the current difficulty in comparing study results due to the lack of a common benchmark. As a first step towards standardization, both articles give surveys about existing non-standardized or standardized wheelchair tests.

Kilkens *et al.* [14] conclude that, while more research is needed to identify the skills to be included in a standard test, out of the 24 tests they reviewed only the WST has been adequately tested on both validity and reliability (for the results on this see [3], [4], [16]). Note that Kilkens *et al.* center their discussion on manual wheelchairs, not including powered wheelchairs.

The article by Routhier *et al.* [15] is slightly more general in that it considers tests for manual as well as powered wheelchairs and reviews not only controlled environments (as Kilkens *et al.* [14] do) but also distinguishes between three categories of test environments:

- Real environments (observing subjects' daily wheelchair activities).
- 2) Controlled environments (e.g. obstacle courses).
- 3) Virtual environments (using a simulator).

Routhier et al. [15] recommend the controlled-environment paradigm. One important reason that makes the WST particularly appropriate is that, unlike many other tests, it has not been designed for a specific target group (e.g. stroke patients) but for wheelchair users in general (manual and powered). This is important if the intelligent wheelchair shall serve as an aid to more than just a fraction of patients. Furthermore, the WST version 4.1 is conceived for powered wheelchairs as well as manual wheelchairs (whereas previous versions were aimed only at manual wheelchairs), and thus is possibly suitable for intelligent powered wheelchairs.

5. THE WHEELCHAIR SKILLS TEST

The WST, currently in version 4.1 [5], is being developed as part of the Wheelchair Skills Program (WSP) at Dalhousie University in Halifax, Canada, as a "standardized evaluation method that permits a set of representative manual and powered wheelchair skills to be objectively, simply and inexpensively documented [5]." Extensive information about the test can be found at the WSP website (www. wheelchairskillsprogram.ca). A copy of the evaluation grid (including a list of all tasks) is included in the Appendix.

The developers envision several situations in which to apply the WST:

- 1) In the early rehabilitation process it can serve to identify the skills that should be addressed during training.
- 2) It can serve as an outcome measure to compare a subject's performance before and after rehabilitation.
- It can be used to test research hypotheses and to assist engineers in the development of new technologies.

Since the WST strives to be as general as possible, it specifies four test categories, one for each combination of wheelchair type (manual vs. powered) and test subject (wheelchair user alone vs. wheelchair user with caregiver). Some of the tasks do not apply to all of the four categories (e.g. 'Picks object from floor' is not applicable if a caregiver is present, since it is assumed that the latter rather than the wheelchair user will do this when the situation arises.)

Tasks covered: The powered wheelchair version of the WST (version 4.1, formally identified as the WST-P-WCU), covers 32 skills which are considered representative for general activities. The assumption is that a person doing well (in terms of performance and safety) on the 32 tasks included in the WST can be considered a skilled powered wheelchair user, because the situations he/she encounters on a daily basis will resemble those tested. In other words, the WST abstracts from a real-world setting to measurable powered wheelchair skills. It is based on realistic scenarios but is still standardized enough to allow for precise measurements. As one would expect, most tasks test navigation skills (e.g. 'Rolls forward 10 m in 30 sec', 'Gets over 15-cm pot-hole'), but there are some other actions as well, e.g. those concerning the wheelchair configuration, like 'Controls recline function'. Figure 3 shows an experimenter undergoing some of the skills included in the test. One pass over all tasks takes about 30 minutes [4].

Evaluation method: The test evaluates (separately) both skill performance and safety. Each skill is graded in terms of these two criteria in a binary manner: a person either



Fig. 3. Various skills of the Wheelchair Skills Test, with an intelligent wheelchair. (a) The wheelchair must travel along a 5^{o} sloped platform (the platform is angled along the wall). (b) The wheelchair must be aligned to the left wall. (c) The wheelchair must move forward through a door. (d) The wheelchair must travel through increased rolling resistance (in this case, gravel).

passes or fails a task, and he/she does so either in a safe or in an unsafe way. The Total Percentage Score consists of two numbers, which are simply: one indicates the proportion of applied skills that were successfully passed, the other one states how many of the skills were carried out safely. A task is considered unsafe if injury to the patient or to a bystander seems likely or actually occurs during task completion. The pass/fail grading method makes the evaluation simple and as objective as possible.

The WST requires the presence of a tester (giving instructions and being in charge of conducting the test) and a spotter (ensuring safe test execution); both roles can, however, be assumed by the same person.

To summarize, the WST takes little time and effort and is easy to evaluate. More important, we think that it makes most sense to adopt a test developed by the rehabilitation community and emphasize that the latter seems to converge on the WST as a standard. This is why we have decided to use this test in order to evaluate the SmartWheeler project and propose that it be used by similar projects, too.

6. PROPOSAL OF A TEST PARADIGM

As mentioned above, it is desirable to use a test developed by the rehabilitation research community to evaluate the performance and safety of intelligent wheelchairs. The WST (like the other tests reviewed in [14] and [15]) was designed principally to evaluate the joint performance of the disabled person with their wheelchair, rather than evaluating specifically the person, or wheelchair, alone. This is an important aspect, one that is worth considering also in the context of evaluating intelligent wheelchairs.

The expected outcome of applying the WST consists of two numbers indicating the percentage of skills that were accomplished successfully and safely, respectively. These numbers are absolute though, and there is no obvious way of interpreting them. For instance, what does it mean if a disabled person in an intelligent wheelchair (or, to stay within our paradigm, rather the intelligent wheelchair in cooperation with a disabled person) achieved a score of 60%? Is 60% a good or a bad score? In order to attribute more meaning to the result, one should apply the test under different conditions.

A standard way of doing this in clinical practice is to use the WST with a given individual using a variety of wheelchairs. This setup measures the change in the skills exhibited by the person on-board the various wheelchair platforms, and can allow the selection of a wheelchair matched to a person's needs. In the context of intelligent wheelchairs, the WST could be applied to compare the performance and safety achieved by an individual using both a conventional powered wheelchair and an intelligent wheelchair. The difference between the two outcomes measures how helpful the intelligent software was to the wheelchair user.

The WST was also developed for assessing the efficacy of rehabilitation, by comparing the results of taking the test before and after training or modification to the wheelchair. The WST can thus be applied to evaluate the impact of incorporating different intelligent systems on-board the smart wheelchair (e.g. speech vs. tactile interface, semi-autonomous vs. fully autonomous navigation, etc.) The WST can be further used to evaluate the efficacy of the training phase, when the human is becoming acquainted with the intelligent wheelchair.

Alternately, the WST could be used to facilitate the comparison of results produced by different research teams working on similar projects.

Finally, in the context of evaluating the interaction manager of an intelligent wheelchair, the WST can be used to verify whether the interface is suitable for completing the requisite set of tasks, as well as sufficiently usable for the target population. As well, it provides a well-defined set of tasks to evaluate the robustness of our voice-activated system in a clinically-relevant context.

Yet there are limitations to using such a constrained evaluation procedure. The set of tasks included in the WST is precisely defined and constrained, which makes it difficult to test the system for higher-level tasks such as 'Leave the house'. We will touch on this problem in the discussion. Another limitation is that the presence of a qualified tester/spotter is required. However, dispensing with such personnel is possible only if an experiment does not involve actual patients. We see our methodology in between these two extremes: We are not arguing that the WST be the only evaluation tool used to validate intelligent wheelchairs, but rather that it serves a useful purpose to benchmark systems at a crucial point in their development, namely when the state of the project already warrants experiments with real patients, without being as advanced yet as to necessitate long-term studies in real environments.

7. EVALUATION OF THE INTELLIGENT WHEELCHAIR INTERFACE

A preliminary evaluation was conducted to evaluate the initial design and implementation of the communication interface of the intelligent wheelchair. Seven healthy subjects, all of them university students without involvement in the project, were asked to go through the tasks of the WST, using appropriate vocal commands to communicate each task. The physical robot was not involved in this task; the only measures of interest were the performance of the speech recognition and the dialogue management modules through the set of WST skills. These results were reported in earlier publications [6]. The results from these preliminary experiments served to improve the interaction system, and in particular helped us collect data that was used in the probabilistic language models that underlie the interaction manager.

In this section, we describe the evaluation of the improved intelligent wheelchair interface, focusing primarily on the ability of the interaction system to correctly recognize the tasks of the WST. These experiments were performed on the second robotic platform developed at École Polytechnique de Montréal (see Sec 2 above, and [8]). Subjects were asked to execute the full set of skills in the WST-P-WCU by controlling the robot only through vocal commands. Subjects could monitor execution of the tasks by observing the robot's behavior, as well as via feedback provided on the touchscreen (usually just showing the action selected by the robot). Prior to running the test, subjects received a 30-60min. introduction to the intelligent powered wheelchair, and were allowed to try out a number of commands. Subjects also underwent a 10-min. speech collection procedure (standard procedure for the NaturallySpeaking software package) designed to briefly customize the speech recognition to the subject's vocal characteristics. Subjects were not required to memorize any special set of commands, but rather were encouraged to use whatever commands they felt were most appropriate.

Experiments were conducted with two types of subjects. The first group consisted of eight healthy subjects, all of them clinicians in local rehabilitation centers, but without involvement in the project. The second group consisted of nine individuals with mobility disorders resulting from various conditions, including stroke, spinal cord injury, multiple sclerosis, and arthritis. These subjects ranked in age from 31 to 85 (mean=58), and had been using a conventional powered wheelchair for anywhere between 2 to 17 years (mean=6.8). Each subject completed the test only once; we did not consider the effect of training on performance at this stage in the development.

The results are presented in Tables I and II. The first column shows the subject id. The second column shows the number of vocal commands issued by the user throughout the test. The third column reports the raw speech recognition error rate (deletions, additions and substitutions). The fourth column shows the number of clarification queries issued by the robot in cases where the command was misunderstood or ambiguous. The fifth column presents the number of correct actions carried out by the robot, as identified by human labeling of video sequences. Finally, the last column reports the number of times the robot selected an incorrect action; users were instructed to recover from such situations by issuing a Stop command, or starting a new command. The results presented here were acquired during experiments involving the full robot capabilities, from the robust communication to autonomous navigation. However the results presented here focus primarily on the speech interface, which is the primary contribution of the authors. Results for the performance of the full robotic system are not available yet, as their analysis is subject to clinical consideration.

All subjects, with one exception, were able to complete the test. Subject 7 in Table II was unable to complete the test due to fatigue and health issues unrelated to our experiment. The remaining subjects used between 94 and 219 commands to complete the test. The word error rate for some subjects was quite high (up to 30.9% for subject 5 in Table II). However the appropriate use of queries allowed the system to reach a performance level comparable to that of other users, as shown by the low incidence of incorrect actions. The number of queries used tended to be proportional to the word error rate; subjects with high word error rates required a larger number of queries to clarify their intent. In general, the average word error rate was higher for the disabled subjects (mean=18.5%)

than for the healthy subjects (mean=13.9%). Yet the number of incorrect actions remained low for all subjects, and furthermore we observed no significant difference in terms of the number (p = 0.35) or percentage (p = 0.27) of incorrect actions between the two types of subjects.

Subject	Number of	Word	Number of	Number of	Number of
Id	commands	error rate	queries	correct	incorrect
				actions	actions
1	136	19.2%	10	121	5 (3.7%)
2	159	13.8%	18	136	5 (3.1%)
3	165	13.5%	11	152	2 (1.2%)
4	201	23.6%	37	155	9 (4.5%)
5	114	6.2%	13	97	4 (3.5%)
6	219	2.3%	10	208	1 (0.5%)
7	210	13.1%	25	175	10 (4.8%)
8	141	19.3%	26	111	4 (2.8%)

TABLE I

PERFORMANCE OF THE INTERACTION MANAGER FOR THE TASKS OF THE WHEELCHAIR SKILLS TEST WITH HEALTHY SUBJECTS.

Subject	Number of	Word	Number of	Number of	Number of
Id	commands	error rate	queries	correct	incorrect
				actions	actions
1	149	12.1%	13	131	5 (3.4%)
2	145	14.5%	20	119	6 (4.1%)
3	94	7.2%	8	85	1 (1.1%)
4	122	22.9%	24	96	2 (1.6%)
5	149	30.9%	38	106	5 (3.4%)
6	117	5.2%	2	115	0 (0.0%)
7	16	18.9%	1	14	1 (6.3%)
8	120	23.1%	27	86	7 (5.8%)
9	149	32.7%	35	104	10 (6.7%)

TABLE II



Overall, the subjects indicated that they were largely satisfied by the functionality of the robot's vocal interface. While the word error rate was in some cases quite high, the use of probabilistic techniques allowed the system to maintain a low rate of incorrect actions for all subjects, thus providing satisfactory performance overall. Some subjects felt they needed more time to get familiar with the platform to exploit it more successfully (recall that training time for all subjects was on the order of 30 minutes). This will be addressed when designing experiments that evaluated the longterm performance and safety of our system.

8. DISCUSSION

The WST has been designed for evaluating skill performance and safety in a controlled environment. We believe that this paradigm is generally well-suited for the purpose of testing intelligent wheelchairs, especially at the advanced prototyping stage. In our experiments, we applied the WST as a benchmark for testing the interaction architecture implemented on-board the SmartWheeler. We observed that even though both ablebodied and disabled subjects had a high rate of word recognition error, the natural language processing applied in the interaction architecture was able to prevent a large number of mistakes in selecting the robot's actions. The results presented above are limited to the evaluation of the SmartWheeler's interface, and do not report on the performance of the full system (including autonomous navigation). This will be the subject of further analysis, in collaboration with clinical partners, to determine clinically relevant notions of performance and safety for our intelligent powered wheelchair.

We have argued above in favor of using the WST for evaluating intelligent powered wheelchairs. However, as we further increase the amount of autonomy onboard smart wheelchairs, it is worth asking whether the types of tasks that will become standard on such platforms will remain similar to the set of tasks that are characteristically required of conventional powered wheelchairs. Given a speech interface, is it easy for a user to request "Take me to the kitchen". This high-level type of task is obviously not currently included in the repertoire of the WST. There are important issues to explore in terms of developing standards specifically for intelligent powered wheelchairs and related devices. The ISO 13482 standard (currently under development) may provide a good model of a safety standard for personal care robots, including smart wheelchairs [17]. It may be useful to develop other standards, including possibly a version of the WST for intelligent powered wheelchairs, which provide a set of target tasks that are relevant for evaluating both safety and performance.

Another aspect which we have not yet evaluated concerns whether the intelligent components provided onboard the SmartWheeler are an effective means of reducing the physical and cognitive burden of operating a powered wheelchair. Investigating this question requires a longer-term interaction between the user and the intelligent powered wheelchair; currently, because the subjects are not accustomed to the intelligent interface, we hypothesize that it causes a significant cognitive burden (though we did not measure this formally). Future experiments will consider the long-term impact of embedding an intelligent system on-board a powered wheelchair.

Regarding the long-term evaluation of the SmartWheeler and similar systems, it is worth referring to the distinction made in [15] (and summarized in section 4) regarding different choices of procedures and environments. We now briefly comment on two test categories that will be the subject of future work:

Real environments: Observing users in their everyday setting in order to assess their performance in a standardized manner is difficult both practically and theoretically. First, from a practical point of view, it is time-consuming and thus expensive, as a clinician would have to examine the test subject's daily wheelchair performance over a sufficiently long period of time. Second, the high variance in terms of environment properties makes it conceptually hard to compare scores. Coping with this high variance is, however, one of the foremost challenges in the development of an intelligent wheelchair, so evaluating how well the device can deal with it is crucial for assessing the success of the project. Consider, for instance, a user utterance like "I'm hungry." There is no standardized way of benchmarking the wheelchair's reaction in such a situation because the best reaction depends very much on the setting: In an urban setting, the best option might be to ask the user which restaurant he/she wants to go to, whereas at home, it might be best to take him/her to the kitchen. A

modified test paradigm will be necessary to rate the quality of intelligent control software in such real environments. But to rigorously assess more basic performance and safety quality, we need the more restricted and controlled type of scenario presented above.

Virtual environments: Virtual tests involving a simulator are probably even cheaper to conduct than controlled-environment tests as proposed in this article. Routhier et al. [15], however, state that such tests have demonstrated a "limited applicability to assessment" mainly due to technical weaknesses of the simulators used. However, since big parts of the technology developed for intelligent wheelchairs (e.g. the interaction manager) are software rather than hardware, it might indeed make sense to evaluate these parts in a simulator. In fact, we did just this in the preliminary development phase, as outlined in [6]. In the long term, as the technology underlying virtual environments advances, such evaluations will clearly become more feasible and more realistic than they are today. But to assess the entire project it will be necessary to evaluate the interplay of both software and hardware. This is why evaluation of the full system in a controlled-environment test like the WST continues to be important for time being.

9. CONCLUSION

In this paper we have suggested a methodology to quantify the performance of intelligent wheelchairs, and have applied this test to the evaluation of the speech interface of an intelligent wheelchair. Rather than designing a test from scratch we are building on work done by specialists (clinicians and researchers) in the field of rehabilitation. We have picked the WST, which seems to emerge as a de facto standard in the clinical and research communities. It is based on situations occurring in the daily lives of wheelchair users but still abstract enough to allow for precise measurements, and has been checked for validity and reliability by the developing team, which is crucial both principally and practically if outcomes of that test are to be used as evidence that a wheelchair is ready to be deployed and funded by public health services and insurance companies. In this sense, a strict evaluation is a critical step towards both establishing and gauging the efficacy of intelligent wheelchairs.

ACKNOWLEDGEMENT

Active participants in the project: The project was carried out in close collaboration with researchers at the École Polytechnique de Montréal (Paul Cohen, Sousso Kelouwani, Hai Nguyen, Patrice Boucher) who developed the intelligent wheelchair platform used for the human subject experiments. Researchers at the Universite de Montreal (Robert Forget) and at the Centre de réadaptation Lucie Bruneau (Wormser Honoré) acted as the primary clinical investigators. **Collaborators:** The project also benefited from contributions by researchers at McGill University (Robert Kaplow, Sylvie Ong) and at the Université de Montréal (Louise Demers), and also clinicians and technicians at the Centre de réadaptation Lucie Bruneau (Claude Dufour) and at the Constance-Lethbridge Rehabilitation Centre (Paula Stone, Daniel Rock, Jean-Paul Dussault). **Sponsors:** Financial support was provided through the Canadian Foundation for Innovation, the Natural Sciences and Engineering Council of Canada, and the Fonds québécois de la recherche sur la nature et les technologies. Additional funding was provided by the Fondation du Centre de réadaptation Lucie Bruneau, the Fondation Constance Lethbridge, and Robovic. Materials were donated by Sunrise Medical Canada and LiPPERT.

APPENDIX: WHEELCHAIR SKILLS TEST 4.1

Reproduced from: *http://www.wheelchairskillsprogram.ca*

Wheelchair Skills Test 4.1 Power Wheelchair - Wheelchair Use

Tower wheelchan - wheelchan Oser						
Name:		Scoring Guide (see over for details) ✓ = pass safe			Type of Test	
		 ✗ = pais, saic ✗ = fail, unsafe 			Objective - Capacity Ouestionnaire - Capacity	
Date	1 ester:	NP = no part (onl	y for indicat	Questionnaire - Performance		
Time	start: Time finish:	I E = testing erro	r	[
	Individual Skills	Capacity/ Performance	Safety		Comments	
1.	Moves controller away and back			No Part?		
2.	Turns controller on and off					
3.	Selects drive modes and speeds			No Part?		
4.	Controls tilt function			No Part?		
5.	Controls recline function			No Part?		
6.	Disengages and engages motors					
7.	Operates battery charger					
8.	Rolls forward 10m					
9.	Rolls forward 10m in 30s					
10.	Rolls backward 5m					
11.	Turns 90° while moving forward L&R					
12.	Turns 90° while moving backward L&R					
13.	Turns 180°in place L&R					
14.	Maneuvers sideways L&R					
15.	Gets through hinged door in both					
16.	Reaches 1.5m high object					
17.	Picks object from floor					
18.	Relieves weight from buttocks					
19.	Transfers from WC to bench and back					
20.	Rolls 100m					
21.	Avoids moving obstacles L&R					
22.	Ascends 5° incline					
23.	Descends 5° incline					
24.	Ascends 10° incline					
25.	Descends 10° incline					
26.	Rolls 2m across 5° side-slope L&R					
27.	Rolls 2m on soft surface					
28.	Gets over 15cm pot-hole					
29.	Gets over 2cm threshold					
30.	Ascends 5cm level change					
31.	Descends 5cm level change					
32.	Gets from ground into wheelchair			1		
	Total Percentage Scores					

Additional comments:

FORM_WST_P_WCU_4.1.13 February 14, 2011

REFERENCES

- L. Fehr, E. Langbein, and S. B. Skaar, "Adequacy of power wheelchair control interfaces for persons with severe disabilities: A clinical survey," *Journal of Rehabilitation Research & Development*, vol. 37, pp. 353– 360, 2000.
- [2] R. C. Simpson, "Smart wheelchairs: A literature review," Journal of Rehabilitation Research & Development, vol. 42, pp. 423–436, 2005.
- [3] R. L. Kirby, J. Swuste, D. J. Dupuis, D. A. MacLeod, and R. Monroe, "The Wheelchair Skills Test: A pilot study of a new outcome measure," *Archives of Physical Medicine and Rehabilitation*, vol. 83, pp. 10–18, 2002.
- [4] R. L. Kirby, D. J. Dupuis, A. H. MacPhee, A. L. Coolen, C. Smith, K. L. Best, A. M. Newton, A. D. Mountain, D. A. MacLeod, and J. P. Bonaparte, "The Wheelchair Skills Test (version 2.4): Measurement properties," *Archives of Physical Medicine and Rehabilitation*, vol. 85, pp. 794–804, 2004.

Note: The WST 4.1 Manual should be consulted for scoring details (www.wheelchairskillsprogram.ca)

Scale for Scoring Skill Capacity/Performance

- Pass: (on the Data Collection Form, record "P" or ~)
 - Task independently and safely accomplished. Unless otherwise specified, the skill may be performed in any manner. The focus is on the task requirements, not the method used. Aids may be used (section 2.23).
- A pass may be awarded if the subject passed a more difficult version of the same skill (e.g. if a subject successfully ascends a 15cm curb, a pass may be awarded on the 5cm level change without the subject needing to actually perform the latter).
 Fail: (on the Data Collection Form, record "F" or X)
 Task incomplete:
 Unsafe performance (as defined in section 2.25).
 Unsafe performance (as defined on ether alter of the diministence transfer of the public of the authority in the more section 2.25).

 - Likely to be unsafe in the opinion of the clinician or tester (e.g. on the basis of the subject's description of how a task will be attempted).

 - a task win to an express. Unwilling to the university of the same skill (e.g. if the subject cannot roll forward 10m [#5.8], he/she need not be asked to roll 100m [#5.21]).
 - If a caregiver is the subject of testing, he/she may not ask the wheelchair occupant for advice or physical assistance in the performance of the skill unless specifically permitted in the caregiver section of the individual skill descriptions (section 5). Wheelchair part malfunction.

Scale for Scoring Skill Safety

Safe: (on the Data Collection Form, record "S" or ♥)

- None of the unsafe criteria were met Although a failing capacity score will be awarded in such circumstances, a safe score can be awarded to a person

- Although a failing capacity score will be awarded in such circumstances, a safe score can be awarded to a person who states that he/she cannot do and/or will not attempt a skill.
 Unsafe: (on the Data Collection Form, record "US" or X)
 Subject requires appropriate significant spotter intervention to prevent acute injury to the subject or others (section 2.26). Performing a skill quickly is not, in and of itself, unsafe. A significant intervention is one that affects performance of the skill.
 A significant acute injury occurred. This includes sprains, strains, fractures or head injury, but does not include minor blisters, abrasions or superficial lacerations. Poor technique that may or may not lead to overuse injury at a later time should be noted in the comments section, but does not varrant awarding an unsafe score.
 During screening questions (section 3), the subject describes a method of performing a skill that the tester considers dancerosing.
 - onsiders dangerous. If a caregiver creates more than minimal discomfort or potential harm (e.g. using excessive force with the knew
- If a caregiver creates more than minimal disconfort or potential harm (e.g. using excessive force with the knee against a flexible backrest of the wheelchair to help push the wheelchair through gravel).
 Specific risks and whether they warrant an unsafe score or merely a recorded comment can be found later in the section on individual skills (section 5).
 Note: If an easier version of the skill has been failed, the skill under consideration is not objectively tested, so the tester needs to determine whether the attempt would have been safe or unsafe on the basis of interview only.

%

For both Capacity/Performance and Safety No Part: (on the Data Collection Form, record "NP") As for Capacity scoring (Table 1). Testing Error: (on the Data Collection Form, record "TE") As for Capacity scoring (Table 1).

 Total Capacity/Performance Score = # passed skills ____/ (32 - # NP - #TE) X 100% =

 Total Safety Score = # safe skills ____/ (32 - # NP - #TE) X 100% = _____%

FORM_WST_P_WCU_4.1.13 February 14, 2011

- [5] R. L. Kirby, Wheelchair Skills Program (WSP), Version 4.1. Wheelchair Skills Test (WST) Manual, Website, 2007, http://www .wheelchairskillsprogram.ca/eng/4.1/WST_Manual_ Version_4.1.pdf (accessed 04/27/2008).
- [6] A. Atrash, R. Kaplow, J. Villemure, R. West, H. Yamani, and J. Pineau, "Towards the deployment of an intelligent wheelchair in a standardized rehabilitation environment," Interaction Studies, pp. 345-356, 2009.
- J. Pineau and A. Atrash, "SmartWheeler: A robotic wheelchair test-[7] bed for investigating new models of human-robot interaction," in AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics, 2007, pp. 59-64.
- [8] W. Honoré, A. Atrash, P. Boucher, R. Kaplow, S. Kalouwani, H. Nguyen, J. Villemure, R. West, F. Routhier, P. Stone, C. Dufour, J.-P. Dussault, D. Rock, P. Cohen, L. Demers, R. Forget, and J. Pineau, "Human-oriented design and initial validation of an intelligent powered
- wheelchair," in *RESNA Annual Conference*, 2010. M. White, "Open ccg: The open nlp ccg library." Software available at [9] http://openccg.sourceforge.net, 2001.
- [10] N. Roy, J. Pineau, and S. Thrun, "Spoken dialog management using probabilistic reasoning," in Annual Meeting of the Association for Computational Linguistics (ACL), 2000.
- [11] J. Urdiales, C. a nd Peula, M. Fdez-Carmona, C. Barrue, E. Perez, I. Sanchez-Tato, J. del Toro, F. Galluppi, R. Cortes, C. Annichiaricco, C. Caltagirone, and F. Sandoval, "A new multi-criteria optimization strategy for shared control in wheelchair assisted navigation," Autonomous Robots, 2010.
- [12] R. Annicchiarico, U. Cortes, A. Federici, F. Campana, C. Barrue, A. B. Martinez, and C. Caltagirone, "The impact of cognitive navigation assistance on people with special needs," Computational and Ambient Intelligence, Lecture Notes in Computer Science, vol. 4507, 2007.
- [13] J. S. Ju, Y. Shin, and E. Y. Yi Kim, "Vision based interface system for hands free control of an intelligent wheelchair," Journal of Neuroengineering and Rehabilitation, vol. 6, 2009.

- [14] O. J. E. Kilkens, M. W. M. Post, A. J. Dallmeijer, H. A. M. Seelen, and L. H. V. van der Woude, "Wheelchair skills tests: A systematic review," Clinical Rehabilitation, vol. 17, pp. 418-430, 2003.
- [15] F. Routhier, C. Vincent, J. Desrosiers, and S. Nadeau, "Mobility of wheelchair users: A proposed performance assessment framework," Disability & Rehabilitation, vol. 25, pp. 19-34, 2003.
- [16] N. J. Lindquist, P. E. Loudon, T. F. Magis, J. E. Rispin, R. L. Kirby, and P. J. Manns, "Reliability of the performance and safety scores of the wheelchair skills test version 4.1 for manual wheelchair users," Archives of Physical Medicine and Rehabilitation, 2010.
- [17] "ISO/CD 13482: Robots and robotic devices-safety requirementsnon-medical personal care robot," Website, http://www.iso.org/iso/ catalogue_detail.htm?csnumber=53820 (accessed 02/21/2011).

Joelle Pineau is Associate Professor of Computer Science at McGill University. She received her PhD from Carnegie Mellon University in 2004. Her research interests are in artificial intelligence, probabilistic machine learning, and robotics.

Robert West received a Diplom in Computer Science from the Technische Universität München, followed by an MSc in Computer Science at McGill University (2010). He is now a PhD candidate at Stanford University. His research interests are in artificial intelligence, machine learning, natural language understanding, and information network analysis.

Amin Atrash is a PhD candidate at McGill University. He received his BSc and MSc in Computer Science at the Georgia Institute of Technology. His research interests focus on planning under uncertainty and human-robot interaction.

Julien Villemure completed his M.Sc. in Computer Science at McGill University in 2010. He is now a software developer in Montreal, Canada.

François Routhier is a research scholar at the Centre for interdisciplinary research in rehabilitation and social integration (CIRRIS) at Institut de réadaptation en déficience physique de Québec. He received his BSc and MSc degree in Mechanical Engineering, and his PhD in Experimental Medicine (rehabilitation) from Université Laval, and was subsequently a postdoctocal fellow at Dalhousie University and Université de Montréal. His research interests are related to the measurement of the impacts of manual and powered wheelchair skills, and to the development of the Wheelchair Skills Program.