City Browser: Developing a Conversational Automotive HMI

Alexander Gruenstein[1]
MIT - CSAIL
alexgru@mit.edu

Bryan Reimer[1]
MIT - AgeLab
reimer@mit.edu

Jarrod Orszulak[1]
MIT - AgeLab
jorszulak@mit.edu

Bruce Mehler[1]
MIT - AgeLab
bmehler@mit.edu

Sean Liu[1]
MIT – CSAIL
seanyliu@mit.edu

Stephanie Seneff[1]
MIT - CSAIL
seneff@csail.mit.edu

Shannon Roberts[1]
MIT – AgeLab
scr09@mit.edu

James Glass[1]
MIT - CSAIL
glass@mit.edu

Jeff Zabel[2]
BMW Technology Office
jeff.zabel@bmw.de

Joseph Coughlin[1]
MIT - AgeLab
coughlin@mit.edu

1. 77 Massachusetts Ave, Cambridge, MA 02139 USA
2. 555 Hamilton Ave, Palo Alto, CA 94301 USA

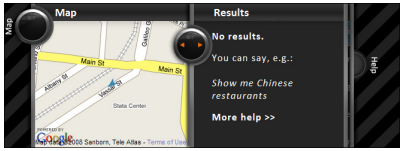
CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/09/04.

Abstract
This paper introduces City Browser, a prototype multimodal, conversational, spoken language interface for automotive navigational aid and information access. A study designed to evaluate the impact of age and gender on device interaction errors, perceptions and experiences with the system along with physiological indices of workload is outlined. Preliminary results, plans for further analysis and a larger scale user evaluation are presented.

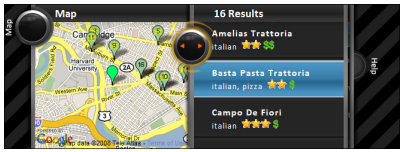
Keywords
Multimodal Interfaces, Speech I/O, User Interface Design, Usability Testing and Evaluation

ACM Classification Keywords
H5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces—graphical user interfaces, natural language, voice I/O; I2.7 Artificial Intelligence: Natural Language Processing—language parsing and understanding, speech recognition and synthesis; J4 Social and Behavioral Sciences: Psychology

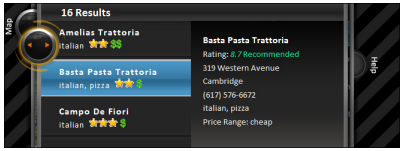
Introduction
The use of navigational devices in automobiles is increasing as drivers are exposed to the benefits of real-time access to directions and points of interest like restaurants and hotels. At the same time, interacting with navigational devices is frequently cumbersome and



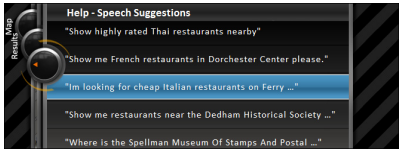
(a)



(b)



(c)



(d)

figure 1. Screenshots of the City Browser automotive interface: (a) The position of the car and brief help, (b) Results screen for *Show me Italian restaurants*, (c) Detailed information pane, and (d) Context-sensitive speech suggestions.

can be quite distracting. Interface designers have turned to spoken language input and output to alleviate some of the manual manipulations and reduce the overall difficulty of device interactions.

However, speech recognition accuracy is far from perfect, meaning that existing automotive interfaces rely on a limited set of formulaic commands. Moreover, information is often entered by voice bit by bit, *e.g.* entering an address may involve separate utterances for the street number, street name, city and state. Older adults, who are most likely to purchase more advanced systems that incorporate speech features, are likely to have the most difficulty adapting to fundamental changes in interface technology [1].

Over the last several years, several of the authors have developed a web-based, conversational natural language interface called City Browser. City Browser, via spoken language and a graphical user interface, allows users to find addresses on a map, search for points of interest like restaurants and hotels, and obtain driving directions [3,4]. In this paper, a new version of City Browser specifically targeting automobile applications is presented. Highlights from a preliminary study designed to evaluate this new version across different ages and technical backgrounds are provided. Subjects' experiences with the system will be used to further improve the HMI (human machine interface) and minimize total workload and potential for distraction. Minimizing the system's potential for distracting the operator is a critical component in the acceptability of conversational vehicle interfaces. Unlike other applications of City Browser [3,4], in an automotive environment safety is a key consideration and device operation must not negatively impact the operation of the vehicle.

As an exploratory measure, physiological indices of workload were recorded to assess two hypotheses: (1) do speech recognition errors increase frustration and consequently arousal?, and (2) do conversational tasks of different difficulty produce measurable changes in workload similar to those observed in earlier studies involving non-contextual tasks [4,5]?

Interface Overview

Using a conversational interface, City Browser allows users to search a database of 6,146 restaurants, 564 hotels, 42 museums, and 134 subway stations in the Boston metropolitan area. Users can locate addresses on a map, obtain driving directions and acquire more detailed information on the establishments. The speech recognition and natural language understanding capabilities of the system are described in more detail in [2]. Screenshots of the graphical user interface (GUI) developed specifically for the automobile appear in figure 1.

The automotive version of City Browser is deployed in a BMW 530xi sedan. It uses the car's built-in display, sound system and iDrive controller as integrated components of the HMI (see figure 2). Speech is captured through an array microphone positioned on the driver's sun visor.

In addition to via spoken language, the City Browser GUI can be navigated using the iDrive controller, a rotary knob which can spin, move laterally left, right, forward, and back, and be pushed down to select items—see figure 2(a). Users can rotate the knob to

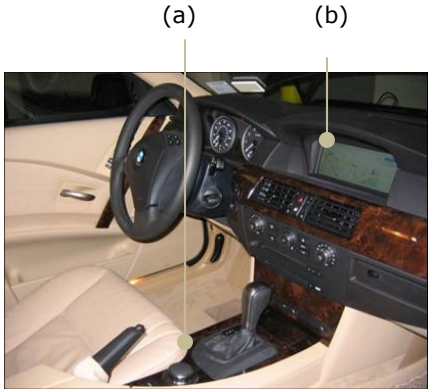


figure 2. The interior of the car, showing (a) the iDrive knob controller (b) the display

scroll through lists, translate the knob left and right to move through the screens shown in figure 1, and push down to make selections. Finally, to speak with City Browser, users press and hold the dedicated speech button. Natural language understanding is performed in context. For example: after highlighting a restaurant in the list of results, a user may say simply "give me directions" to obtain directions to the location. Speech can be conversational, as the example interactions in figure 3 demonstrate. For additional examples, see [2].

When confronted with a new speech interface, users often have difficulty learning what they can say. To alleviate this difficulty, the GUI includes a context-sensitive speech suggestions generator [3], which produces suggestions like those shown in figure 1(d). Suggestions automatically appear after two input rejections due to a low system confidence score. They can also be accessed by navigating the GUI with the iDrive knob. Content in the suggestions changes depending on the context (*e.g.* if a list of restaurants is shown, it will give an example of asking for directions to one of them). The suggestions are generated dynamically using the content of the database.

The interface stands out from both commercially available interfaces and other research prototypes because it combines a capable graphical user interface with a conversational speech interface in an actual automobile. While commercially available interfaces often support sophisticated GUIs, they typically provide no or limited speech capabilities. A similar research prototype is CHAT [6], which provides a similar set of navigational capabilities – as well as a conversational music interface. However, to our knowledge, CHAT has not been tested on a large scale in an actual automobile, and includes only a passive GUI.

Task	You're meeting up with some friends in Cambridge and want to take them to a Chinese restaurant. Find one and get directions to it.	You're picking up a friend from his apartment in Quincy at 180 Hancock Street. Find the address on the map and find a cheap restaurant near his apartment to have dinner.
Example	*U.* Show me Chinese restaurants in Cambridge *S.* There are 12 Chinese restaurants in Cambridge. [Shows a results screen similar to figure 1b] *U.* [Uses iDrive controller to browse the list, leaving All Asia Café highlighted] *U.* Give me directions. *S.* Here are directions from this location to All Asia Café. [Shows directions on the map]	*U.* Where is one eighty Hancock Street in Quincy. *S.* Here is one eighty Hancock Street in Quincy [Shows the address marked on the map] *U.* Are there any cheap restaurants near there? *S.* There are fifteen inexpensive restaurants near one eighty Hancock Street in Quincy. [Shows a results screen similar to figure 1b] *U*. [Uses iDrive controller to browse]

figure 3. Two tasks which subjects were asked to complete, and possible solutions. *S* indicates System, *U* indicates User.

Overview of an initial study designed to assess the usability of the system

A pilot study was conducted to assess the usability of the prototype, which was refined over the course of the study, and to refine the experimental protocol. A total of 33 participants were recruited from the local community. Participants were required to have a valid driver's license and be fluent in English. All participants were first provided with an overview of the project's objectives and the experimental protocol. They were then required to read and sign an informed consent approved by the local institutional review board. Physiological sensors were then placed on the participant. These included a photoelectric plethysmograph (PPG) on the left index finger to monitor the peripheral blood volume pulse wave, two gold plated contacts on the underside of the outer flanges of the middle and ring fingers to measure skin conductance level (SCL), and a respiration belt around the waist just below the rib cage. PPG was selected to measure heart rate over a more robust EKG measurement given its convenience and non-invasiveness. The physiological signals were processed using a MEDAC System/3 physiological recording unit (NeuroDyne Medical, Corp, Cambridge MA).

Following the placement of physiological sensors, participants were asked to complete a questionnaire that assessed driving habits, technology exposure and demographics. Once the questionnaire had been completed, the subject entered the instrumented car. For the purpose of this evaluation, the car was located in a parking lot and the evaluation was carried out under parked, non-driving conditions. Prerecorded instructions were provided to introduce the iDrive interface and the mode of speech interaction. When errors in operation of the iDrive controller or speech interaction occurred during training, the research assistant clarified the task in a manner similar to an interactive tutorial.

Twelve experimental tasks were then presented. The tasks were designed to encompass typical navigational and information goals. Subjects were asked to find points of interest (POIs) like restaurants, hotels, and museums, to get driving directions to particular addresses or POIs, and to obtain information such as phone numbers. Two sample tasks, and possible solutions, are shown in figure 3. The tasks were designed to be of easy, medium, or hard difficulty. The first four tasks were supposed to be easy to familiarize the subject with the interface and allow him or her to gain confidence. The subsequent eight trials comprised a mixture of easy, medium and hard tasks. In trials where a participant made no progress towards the completion of the task, the experiment progressed to the next task.

The tasks were presented in the same sequence for all participants. To minimize biasing participant's pronunciation, the tasks were printed on index cards and handed to participants. There was a 25 second pause between tasks. During the experiment, three video cameras recorded the subject's face, their hand movement and the car's display; a complete audio recording of the subject's interactions was also obtained. Following the interactive portion of the experiment, a second questionnaire was presented to gauge each subject's impression of the system, likes and dislikes, measures of effort and frustration and desire for purchasing the system.

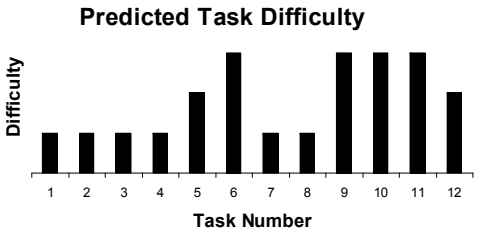


figure 4. Predicted task difficulty ranging from easy to medium to hard. Task difficulty was expected to correspond with physiological readings, except for habituation over the first

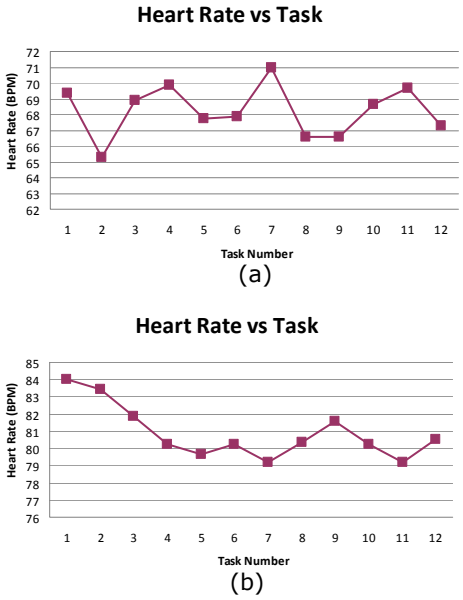


figure 5. Average heart rate per task for two subjects: (a) younger male (b) late middle aged female

Overview of findings

Age distribution consisted of 13 college age (19-21), 10 post-college (22-30), and 10 middle age (40-56) individuals. Of the 33, 24 completed all 12 tasks successfully (11 college age, 5 post-college, 8 middle age); 5 participants completed 11; one participant completed 10 tasks and another 9. Two younger female participants failed to complete five or more tasks. Overall, participants enjoyed their interaction with the system and indicated that they would be interested in purchasing this type of system. Interestingly, the most common complaint about the system was the voice of the speech synthesizer used to present output.

Across 396 tasks, 1,651 utterances were collected and transcribed. The overall speech recognition word error rate (WER) was 31.9%. If the 216 utterances from the two subjects who had great difficulty with the system are excluded, the WER drops to 26.4%. The impression of the transcriber (first author) was that generally subjects completed the majority of the tasks with few or no system errors, but often had difficulty with a few tasks. Errors often occurred on utterances involving addresses, where the street number, street name, and city must all be recognized correctly for the system to respond appropriately. This indicates that in specific instances, the ability to issue bit-by-bit information might be helpful. A number of participants appeared to use the help screen, typically resulting in utterances after which typically seemed to result in more accurate system behavior.

Figure 4 shows the predicted difficulty of each task, while figure 6 presents actual system response accuracy. Comparing figures 4 and 6 reveals some variation between expected and actual levels of difficulty. These findings were used to reorder the tasks for the follow-on study. Figure 5 shows average heart rate per task for two of the subjects. It had been hypothesized that heart rate would correspond to task difficulty; however, individual variations are evident and a more detailed analysis is planned that takes individual performance variables into account.

One observation suggests an interesting area of future study; it appeared that participants often took deep breaths before speaking to the system. It may be possible to use breathing as a cue to automatically trigger the microphone, eliminating the need for users to hold a button while speaking. If systems could be triggered to listen for input just after a user takes a deep breath, a balance between computational limitations of monitoring audio and usability may be found.

Conclusions and Future Work

This paper presents work in progress towards the development and evaluation of City Browser, a prototype multimodal conversational interface deployed in an automobile. A group of 33 pilot subjects used the interface to complete a series of 12 tasks, while physiological data was collected. Based on the experience gained in this proof-of-concept pilot study, a larger investigation with an intended sample size of 72 subjects, a more even distribution of ages, specifically targeting the following age groups: 25-34, 45-54, and 65+, and an HMI constant across all users is now in progress. By analyzing similar parameters with a larger sample size, it should be possible to determine how subjects from different demographics, specifically age and technological background, interact with the system. With the larger number of subjects, self report

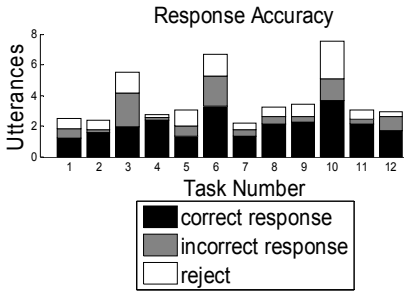


figure 6. Average number of utterances per task for the 31 subjects who successfully completed 9 or more tasks. Utterances are labeled by the accuracy of the system's response: correct, incorrect, or reject (in which the system responded with variations of "pardon me"). Although it was expected that some tasks would require multiple utterances, generally, more total utterances and more incorrect responses per task are a good indicator of task difficulty. Pre-experiment predictions regarding the difficulty of specific tasks did not match perfectly with the results obtained. For example, tasks 3 and 7 both required the subject to get directions to a specific museum, but the plot above indicates that task 3 was likely much more frustrating than task 7, even considering the user's increased experience using the system.

measures such as questionnaires will have more weight with regards to how much the user did or did not enjoy interacting with the system.

Implications for this ongoing research are potentially far reaching. The City Browser interface is a preliminary exploration of the utility of conversational interfaces in an automotive environment, so evaluating its efficacy and determining areas for improvement is crucial as part of an iterative design program. Moreover, the use of physiological measures as a design tool for assessing workload has not previously been explored in this domain—results pertaining to the impact of system accuracy should be generally useful across many application domains. Finally, as the car continues to evolve, and technologies like navigation systems becomes more common, studies which measure workload in such environments will become more important.

An important limitation of the current study is that using the City Browser interface is the primary task, given that the car remains parked through the duration of the experiment. However, in a real world implementation of such a system, driving would become the primary task and interacting with City Browser, the intended secondary task. A logical follow-on study would be to ask participants to use the system while driving to investigate the additional workload the system imposes on the driver to evaluate if the convenience of the system is outweighed by the potential distraction. In addition to quantifying work load through the use of physiology, eye tracking systems such as the faceLab system used in previous studies [6] would be especially valuable in determining the risks and benefits of a system like City Browser.

Acknowledgments

The authors gratefully acknowledge the support of the United States Department of Transportation's Region I New England University Transportation Center at the Massachusetts Institute of Technology, BMW, and T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan. We also wish to acknowledge the contributions of Alea Mehler, Eugenia Gisin, Michael Thompson, and Lisa D'Ambrosio.

Citations

[1] Coughlin, J. Not your father's auto industry? Aging, the automobile, and the drive for product innovation. *Generations: Journal of the American Society on Aging*, (2004-2005), 38-44

[2] Gruenstein, A. and Seneff, S. Context-Sensitive Language Modeling for Large Sets of Proper Nouns in Multimodal Dialogue Systems. In *Proc. SLT* (2006), 130-133.

[3] Gruenstein, A. and Seneff, S. Releasing a Multimodal Dialogue System into the Wild: User Support Mechanisms. In *Proc. SIGdial* (2007), 111-119.

[4] Mehler, B., B. Reimer, J.F. Coughlin, and J.A. Dusek. The impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Proc. Transportation Research Board of The National Academies, (2009).

[5] Reimer, B. Cognitive Task Complexity and the Impact on Drivers' Visual Tunneling. Proc. Transportation Research Board of The National Academies, (2009).

[6] Weng, F., Yan, B., Feng, Z., Ratiu, F., Raya, M., Lathrop, B., Lien, A., Mishra, R., Varges, S., Lin, F., Purver, M. Meng, Y., Bratt, H., Scheideck, T., Zhang, Z., Raghunathan, B., and Peters, S. CHAT To Your Destination. In *Proc. SIGdial* (2007), 79-86.