CONSIDERING SELF-REPORT IN THE INTERPRETATION OF OBJECTIVE PERFORMANCE DATA IN THE COMPARISON OF HMI SYSTEMS

Bruce Mehler[1,*], Bryan Reimer[1], Chaiwoo Lee[1], David Kidd[2], Ian Reagan[2]

[1]MIT AgeLab & N.E. Univ. Transportation Center, Cambridge, MA USA
[2]Insurance Institute for Highway Safety, Arlington, VA USA
[*]Corresponding author: bmehler@mit.edu

Summary: Driver interaction with two production voice-command interfaces representing differing user interface design approaches were compared under on-road highway driving conditions. A sample of 80 drivers was randomly assigned to drive each vehicle (40 per vehicle). During voice-based phone contact calling and destination address entry, participants in one vehicle showed, on average, statistically significant "better" performance in terms of task completion time, mean glance duration, total off-road glance time, and total number of glances. However, these objective measures do not fully characterize the overall experience of participants. An analysis of error rates and subjective report of attitudes, effects on driving behavior, and behavioral intentions relative to their exposure to the two systems provided important, complementary and sometimes contrasting data about the relative advantages of each implementation.

INTRODUCTION

A primary design goal in the integration of auditory-vocal (voice command) interfaces into embedded vehicle systems is to reduce distraction associated with visual-manual engagement. The potential benefits of voice-based interactions relative to visual-manual interaction have been relatively well demonstrated in simulated or prototype implementations (e.g., see reviews in Andre & Wickens, 1995; Lo & Green, 2013; Reimer, Mehler, Dobres & Coughlin, 2013). However, relatively few studies are available on the extent to which these benefits are observed with actual production systems, and even fewer have examined production systems under on-road conditions (see summaries in Mehler et al., 2016; Reimer et al., 2013). A number of researchers have raised questions regarding the extent to which variations in speech generation, speech comprehension or other aspects of mental load might impact the driver (e.g., Blanco, Biever, Gallagher & Dingus, 2006; Lee, Caben, Haake & Brown, 2001; Strayer, Cooper, Turrill, Coleman & Hopman, 2015). Additional data on how drivers engage with voice-command systems under on-road conditions could prove helpful in improving designs.

Prior to conducting such a study, a structured, hierarchical assessment of interaction in a number of representative production voice systems was undertaken under static conditions to identify differences in design approaches that might be useful to compare (Reagan & Kidd, 2013). One approach was characterized by a layered-menu structure where users stepped incrementally through a series of selections (e.g. specify a desire to select a phone contact, specify a contact name, specify a desired phone for that contact if they have multiple phones such as home, work, mobile). The other was built around a "one-shot" design style where a single compound command could be used to execute most or all of a task in a single step (e.g., "Call x at work"). Two vehicle systems were then selected for further study that most typified each approach.

The follow-on on-road study (Mehler et al., 2016; Reimer, Mehler, Reagan, Kidd & Dobres, 2016) considered task performance (task time, errors), visual engagement (glance metrics), driving performance, and indices of workload (heart rate, skin conductance, and subjective ratings) in a sample of 80 drivers randomly assigned to each vehicle (40 each). Both voice systems were found to reduce visual demand relative to the respective visual-manual interface for placing a call using a stored phone contact list. The voice-based interface for entering a destination address into each vehicle's embedded navigation system was also evaluated. Distinct advantages in terms of total task time and various measures of visual engagement appeared for the one-shot style interface (Chevrolet MyLink) compared to the layered-menu design (Volvo Sensus) (see Table 1). In contrast, the layered-menu based interface resulted in far fewer user and system related errors during the longer address entry task. While the metrics of visual engagement provide important objective measures of behavior, they do not provide a full characterization of the drivers' overall experience and its possible impact on future behavior relative to system use (Andre & Wickens, 1995). The present report expands upon that evaluation of objective data by considering additional self-report data.

Table 1: Mean and (standard error) of task time and off the forward roadway glance metrics for voice-based multi-modal and visual-manual interfaces across two vehicles (data drawn from Mehler et al., 2016).

HMI Task & Vehicle	Task Completion Time (seconds)		Mean Single Glance Duration (seconds)		% Long Duration Glances (>2seconds)		Total Eyes Off Road Time (seconds)	
	Voice	Manual	Voice	Manual	Voice	Manual	Voice	Manual
Phone Contact Calling Volvo Sensus	38.17 (6.9)	32.90 (12.8)	0.79 (0.1)	0.94 (0.2)	0.75 (1.9)	3.39 (5.6)	10.22 (4.3)	16.39 (6.6)
Phone Contact Calling Chevrolet MyLink	21.61 (8.3)	26.24 (7.6)	0.60 (0.2)	0.92 (0.2)	0.29 (1.5)	2.46 (3.7)	3.33 (2.6)	13.63 (5.3)
Address Entry Volvo Sensus	80.60 (1.71)		0.82 (0.02)		1.27 (0.36)		22.56 (1.43)	
Address Entry Chevrolet MyLink	66.68 (2.85)		0.74 (0.02)		1.02 (0.29)		14.28 (1.22)	

METHOD

This study reports on self-report data collected as part of a larger research project (Mehler et al., 2016). Participants ranged in age from 20 to 66 years and were equally distributed by age and gender across the two vehicles. Participants drove one of two standard production vehicles, a 2013 Chevrolet Equinox equipped with the MyLink infotainment system (the "one-shot" style voice system implementation) and a 2013 Volvo XC60 equipped with the Sensus system (the layered-menu style voice system implementation). In brief, details of the implementation characteristics are provided in (Mehler et al., 2016). Post-experimental subjective ratings were made using a paper questionnaire with visual-analog scales (using rankings from 1 to 10) and other items with descriptive scalers providing an ordered range of 5 choices as shown in Table 2.

RESULTS

Self-report ratings from the Volvo Sensus and Chevrolet MyLink experimental groups were statistically compared using independent t-tests to examine the extent to which participants' perceptions and attitudes differed; no adjustments for multiple tests were applied (Table 2).

Table 2. Mean and (standard deviation) of self-report ratings by independent groups of drivers following exposure to the 2013 Volvo Sensus (N=40) and Chevrolet MyLink (N=40) infotainment interfaces for phone contact calling (manual & voice) and full destination address entry (voice) (*: p<0.05).

Category	Variable	Question	Scale Range	Chevrolet	Volvo	p-value
Attitude	Overall impression	What was your overall impression of the vehicle you drove today?	1: Not at all positive 10: Very positive	**7.30** (1.95)	**8.80** (1.09)	**0.000***
	Trust	How has your experience today influenced your level of trust in new technologies that are being introduced into cars?	1: Much less trusting 5: Much more trusting	**3.40** (0.71)	**3.98** (0.89)	**0.002***
	Confidence	Based on your experience today, has your sense of your ability to learn new technologies increased, stayed about the same, or decreased?	1: Decreased 10: Increased	6.46 (1.61)	7.06 (1.66)	0.105
Effect on driving behavior (voice)	Eyes on road	To what extent do you feel that the car voice-based interface you used today allowed you to keep your eyes on the road?	1: Not at all 10: A lot	**7.95** (1.87)	**8.73** (1.50)	**0.044***
	Hands on wheel	To what extent do you feel that the car voice-based interface you used today allowed you to keep your hands on the wheel?	1: Not at all 10: A lot	**8.20** (1.90)	**9.13** (0.99)	**0.008***
Behavioral intention (voice)	Likely to use - call	How likely would you be to use the voice-based car interface to lookup a phone number and place a call?	1: Not at all likely 10: Very likely	**7.35** (2.66)	**8.95** (1.50)	**0.002***
	Likely to use - nav	How likely would you be to use the car interface to enter an address into the navigation system?	1: Not at all likely 10: Very likely	**6.38** (2.83)	**7.73** (1.87)	**0.014***
	Likely to recommend	How likely is it that you would recommend to a friend or family member that they consider buying a car with the voice control technology you used today?	1: Not at all likely 10: Very likely	**6.40** (2.56)	**8.40** (1.58)	**0.000***
	Like to have	If any limitations in the voice control interface you worked with today were solved, to what extent would you like to have a voice command interface in your next car?	1: Not at all likely 10: Very likely	**7.60** (2.53)	**9.10** (1.55)	**0.002***
Ratings not involving car voice interface	Likely to use (touch phone)	How likely would you be to use the touch based phone interface to lookup a phone number and place a call?	1: Not at all likely 10: Very likely	4.70 (2.51)	4.85 (2.74)	0.799
	Likely to use (touch car)	How likely would you be to use the knob/button/touch based car interface to lookup a phone number and place a call?	1: Not at all likely 10: Very likely	4.40 (2.76)	4.83 (2.98)	0.510
	Prefer	Do you have a preference for using touch screen vs. traditional buttons and knobs?	1: Prefer touch screen 5: Prefer buttons	2.55 (1.26)	2.71 (1.19)	0.576

The groups differed in terms of attitudes toward the vehicle driven and the technologies they interacted with during the study. Participants who interacted with the Volvo rated their overall impression of the vehicle more positively ($p < 0.001$) compared to those who interacted with the Chevrolet. Participants who worked with the Volvo Sensus infotainment interface reported being more trusting toward new car technologies ($p < 0.01$). The magnitude of perceived effect of the voice interface on driving behavior also differed. Participants in the Volvo Sensus group reported higher scores for the effect of using the voice interface on promoting better driving behavior in terms of enabling them to their keep eyes on road ($p < 0.05$) and hands on wheel ($p < 0.01$) compared with the Chevrolet MyLink group, although the rating was quite positive in both.

Participants in the Volvo Sensus group were also more likely to report more positive behavioral intentions related to use of the vehicle voice interface after the experiment. For both types of voice interface implementation, phone contact calling and address entry into the navigation system, participants in the Volvo Sensus group responded more positively as to their likelihood of using each voice interface if they were to have access to the same technologies / interfaces they had just used while driving on their own compared to those in the Chevrolet MyLink group. The Volvo Sensus group also responded, on average, as being significantly more likely to recommend the voice interface to others ($p < 0.001$) and more positively in terms of the extent to which they would like to have a voice interface in their next car if any limitations they experienced were solved ($p < 0.01$). The final 3 items presented in Table 2 represent ratings unrelated to interaction with the voice interfaces. The differences in the ratings for these items were not statistically significant between the Volvo and Chevy.

Table 3 considers the extent to which there was a relationship between experiencing errors working with the voice interface and ratings of the behavioral intention items (Kendall's rank correlation coefficients). In this analysis, an error could be by the user (using the wrong command syntax) or by the system (voice recognition error). Negative values indicate that positive endorsement decreases as the number of trials with errors increased.

Table 3. Correlation between the number of voice interaction trials in which an error occurred and behavioral intention ratings. Within each grouping, the column on the left does not include navigation cancel trials; the column of the right does include canceling navigation. (*: p<0.05; ** p<0.01).

	Variable	Across Systems		Chevrolet		Volvo	
Behavioral intention (voice)	Likely to use -call	**-.295****	**-.207***	-.225	**-.270***	-.097	.004
	Likely to use -nav	**-.230***	-.149	-.170	-.151	-.046	-.018
	Likely to recommend	**-.328****	**-.239****	**-.259***	**-.354****	-.009	.111
	Like to have	**-.318****	**-.248****	-.225	-.238	-.096	-.055

DISCUSSION

In summary, self-reported perceptions and attitudes toward the vehicles and the embedded voice systems were more positive among those who interacted with the Volvo 2013 XC60 equipped with the Sensus infotainment system than was the case for the 2013 Chevrolet Equinox with the MyLink system. The results from this study contrast with objective findings in some ways. Specifically, participants who used the "lower rated" Chevrolet MyLink voice interface for phone contact calling and destination address entry showed, on average, statistically significant "better" objective performance compared to those in the Volvo Sensus in terms of shorter task completion times as well as lower mean single glance durations, percentage of long duration glances, total off-road glance times, and number of off-road glances (Mehler et al., 2016). Thus, consistent with Andre and Wickens (1995), depending on whether self-report or these frequently used objective metrics of system performance and visual demand are considered, one might develop a very different impression of participants' experiences with the two vehicles and their associated voice command systems. Andre and Wickens argue that when a dissociation is found between a performance metric and preference reports, then a careful analysis should be undertaken to try to understand the factors responsible.

A suspect factor in the apparent discrepancy between objective glance metrics and self-report measures in this study is system errors. While voice system recognition errors were relatively low for voice phone contact calling for both systems (3.1% Chevrolet MyLink vs. 1.3% Volvo Sensus), voice recognition errors for address entry were quite high in the Chevrolet compared to the Volvo (31.7% vs. 4.2%) (Mehler et al., 2016). Moreover, combined user and system error rates were 51% and 10.1% for the Chevrolet vs. Volvo. Thus, while interaction time and visual engagement were lower with the MyLink interface, the likelihood of a problematic interaction was notably higher when attempting to enter an address into the navigation system. This likely impacted the subjective ratings, and may be most prominently reflected in the divergence in the rating of how willing a participant indicated they were to recommend purchase of the voice technology they experienced to a friend or family member (6.4 vs. 8.4 on a 10 point scale). The negative correlations between number of trials in which an error was experienced and the degree of positive endorsement of the behavioral intention measures supports this interpretation.

In our earlier reporting (Mehler et al., 2016; Reimer et al., 2016), we raised the possibility that the higher error rates in the Chevrolet Equinox might not be due solely to the capabilities of voice recognition system and technical demands of using a one-shot input approach. For reasons detailed there, it was suggested (and later documented through sound level measurements) that road noise was higher in the Chevrolet and this might have impacted the voice system. In reviewing open ended questions in the post-experimental questionnaire for the development of this paper, it was found that five of the 40 participants who drove the Chevrolet commented that background noise might have contributed to voice recognition errors; no such statements were made by Volvo participants. This is a good example of potentially useful subjective input that can be obtained from participants in addition to their objective behavior. One resulting take-away is that it may be useful to evaluate upgraded soundproofing as a means to improve performance.

As raised in Mehler et al. (2016), another factor that may play a role is the issue of ease of initial learning vs. ease of use with experience. The multi-step, layered-menu design of the Volvo Sensus system readily guided a new user through the voice task and broke the audio inputs into content segments that may have reduced the challenge for the speech parsing algorithms. This approach seems to be associated with a higher likelihood of initial success with the system. As noted, the trade-off is a longer task time and, in this design, more extended visual engagement. In contrast, the one-shot, combined command approach used by the MyLink system appeared to present a greater challenge for participants in discovering the way in which the system was optimized to parse speech input (see also McAnulty, Dobres, Mehler, & Reimer, 2017). The overall balancing may be between shorter task time and less visual engagement when the input is recognized correctly vs. the possibility of frustration when encountering errors. Such frustration may result in discontinuing use of the technology before the design concept mental model and the experience necessary to take full advantage of the system potential is developed. As expressed by one participant concerning the one-shot design, “when it actually understood me it felt like less work.” There may be value in evaluating the extent to which a design that supports both approaches offers advantages across individuals and experience level.

Although error rates for the address entry task and ease of learning and use likely explain some of the variance observed between ratings of the two voice interfaces, participants who drove the Chevrolet had significantly lower overall impressions of their assigned vehicle than those who

drove the Volvo. Differences in impressions may be owed to differences in brand luxury, and such impressions may have led to a response bias. On the other hand, the magnitude of the differences in ratings for the voice interface characteristics between the two vehicles were statistically significant while the ratings of the manual interface characteristics (bottom of Table 2) did not even approach what might be considered a meaningful trend (p values ranging from 0.501 to 0.799). This increases to some extent the confidence level that a meaningful portion of the differences in the relative positiveness of the ratings in the two groups are related to actual differences in user experiences with the voice-interfaces. Nonetheless, the current study is limited in that the relative contribution of these factors to the subjective ratings is unknown.

While risk reduction evaluations might logically emphasize relatively well-established objective risk characteristics such as total task time (Burns, Harbluk, Foley & Angell, 2010) and glance metrics (Fitch et al., 2013; Klauer, Dingus, Neale, Sudweeks & Ramsey, 2006; Victor et al., 2014), it is likely that user perception continues to be an important consideration in the comprehensive assessment of user interfaces. Error rates or other factors, such as ease of initial engagement (Harvey, Stanton, Pickering, McDonald, & Zheng, 2011), may impact actual use of one design over another and should ideally be taken into account. Self-report data cannot provide the same level of confidence as naturalistic observation, but may provide a pragmatic perspective when appropriately developed and collected. Self-report data also adds to the face validity of assumptions regarding subjective experience. Realized advantages of actual driver adoption and appropriate use of vehicle technologies are difficult to evaluate, but are ultimately critical in understanding functional risk reduction.

ACKNOWLEDGMENTS

Support for data collection was provided by the Insurance Institute for Highway Safety (IIHS). Additional support for this analysis was provided by US DOT's Region I New England University Transportation Center and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel.

REFERENCES

Andre, A. D., & Wickens, C. D. (1995). When users want what's not best for them. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *3*(4), 10-14.

Barón, A., & Green, P. (2006). *Safety and usability of speech interfaces for in-vehicle tasks while driving: a brief literature review*. Ann Arbor, MI: The University of Michigan Transportation Research Institute (UMTRI).

Blanco, M., Biever, W.J., Gallagher, J.P., & Dingus, T.A. (2006). The impact of secondary task cognitive processing demand on driving performance. *Accident Analysis and Prevention, 38,* 895-906.

Burns, P., Harbluk, J., Foley, J., & Angell, L. (2010). The importance of task duration and related measures in assessing the distraction potential of in-vehicle tasks. *Proceedings of the Second International Conference of Automotive User Interfaces and Interactive Vehicle Applications (Automotive UI 2010)*, November 11-12, Pittsburgh, PA, USA.

Fitch, G. M., Soccolich, S. A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., Hanowski, R. J., Hankey, J. M., & Dingus, T. A. (2013). *The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk. (Report No. DOT HS 811 757)*. Washington, DC: National Highway Traffic Safety Administration.

Harvey, C., Stanton, N. A., Pickering, C. A., McDonald, M., & Zheng, P. (2011). In-vehicle information systems to meet the needs of drivers. *Intl. Journal of Human–Computer Interaction, 27*(6), 505-522.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., & Ramsey, D.J. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data (DOT HS-810-594)*. Washington, DC: National Highway Traffic Safety Administration.

Lee, J. D., Caven, B., Haake, S., Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors, 43(4),* 631–640.

Lo, V.E-W., & Green, P.A. (2013). Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature. *International Journal of Vehicular Technology, 2013,* ID 924170. http://dx.doi.org/10.1155/2013/924170

McAnulty, H., Dobres, J., Mehler, B., & Reimer, B. (2017, January). Characterization of errors encountered when interacting with an auditory-vocal in-vehicle interface during highway driving. *Proceedings of the Transportation Research Board 96th Annual Meeting*, Washington D.C., January 8-12, 2017.

Mehler, B., Kidd, D., Reimer, B., Reagan, I., Dobres, J. & McCartt, A. (2016). Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems. *Ergonomics, 59*(3), 344-367. doi:10.1080/00140139.2015.1081412

Reimer, B., Mehler, B., Dobres, J. & Coughlin, J.F. (2013). *The effects of a production level "voice-command" interface on driver behavior: reported workload, physiology, visual attention, and driving performance*. MIT AgeLab Technical Report No. 2013-17A. Massachusetts Institute of Technology, Cambridge, MA.

Reimer, B., Mehler, B., Reagan, I, Kidd, D., & Dobres, J. (2016). Multi-modal demands of a smartphone used to place calls and enter addresses during highway driving relative to two embedded systems. *Ergonomics*, 59(12), 1565-1585. doi:10.1080/00140139.2016.1154189.

Reagan, I.J. & Kidd, D.G. (2013). Using hierarchical task analysis to compare four vehicle manufacturers' infotainment systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 1495-2599. Santa Monica, CA.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2015). *Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems*. Washington, DC: AAA Foundation for Traffic Safety.

Victor, T., Bärgman, J., Boda, C., Dozza, M., Engström, J., Flannagan, C., Lee, J.D. & Markkula, G. (2014). *Analysis of naturalistic driving study data: safer glances, driver inattention, and crash risk. (SHRP 2 safety project S08A prepublication draft)*, Gothenburg, Sweden: Safer Vehicle and Traffic Safety Centre at Chalmers.