VOICE-CONTROLLED IN-VEHICLE SYSTEMS: EFFECTS OF VOICE-RECOGNITION ACCURACY IN THE PRESENCE OF BACKGROUND NOISE

Neil Sokol, Huei-Yen Winnie Chen, Birsen Donmez
Mechanical and Industrial Engineering, University of Toronto
Toronto, Ontario, Canada
Email: nsokol@mie.utoronto.ca, win.chen@mail.utoronto.ca, donmez@mie.utoronto.ca

Summary: This paper presents initial findings from a driving simulator study comparing user responses to a noise-robust voice-controlled system while driving to a noise-sensitive one in the presence of background noise. Twenty participants interacted with both noise-sensitive and noise-robust simulated voice-controlled infotainment systems while driving under three background noise conditions (no noise, music, and children). While both systems were viewed as useful and satisfying, user acceptance was affected by background noise with the noise-sensitive system, but not the noise-robust one. There was also no evidence that user acceptance was calibrated by having background noise as a context for varying levels of accuracy. No significant differences were observed between the two systems in driving performance metrics analyzed (average speed, speed variability, and standard deviation of lane position), but the use of either system affected driving performance compared to baseline driving. A larger sample size at the end of this study along with the analysis of a larger set of performance metrics will provide further insights.

INTRODUCTION

Voice-controlled, or voice-command, infotainment systems have become increasingly common in new automobiles. By 2019, more than half of new vehicles are expected to come with a voice-controlled system installed (IHS Technology, 2013). Such systems can perform a wide variety of tasks, such as navigation, music selection, and retrieval of contacts.

On-road studies by Ranney et al. (2005) and Reimer et al. (2013) have shown that production level in-car voice command systems performed better compared to manual interactions with infotainment systems on a number of driving performance metrics. However, voice-controlled systems are not without their limitations, as their accuracy can degrade in a noisy environment, which is often the case for driving, given the presence of audio entertainment, passenger conversation, and outside noise. Previous simulator studies by McCallum et al. (2004) and Kun et al. (2007) showed degradation in driving performance when voice recognition accuracy was reduced. However, in both studies, recognition accuracy of the system was manipulated under identical environmental conditions. In other words, there was no additional context provided to help justify the reduction in voice recognition accuracy these participants experienced. It has been suggested by Lee & See (2004) that understanding the possible influence of a system's environment on its performance may allow users to appropriately adjust their level of trust in the system.

More recently, Li et al. (2014) reviewed new techniques that may resolve the issue of sensitivity to noise in voice-controlled systems, and that may soon be implemented in commercially available vehicle systems. With sight of such development, the present study aims to compare a noise-sensitive system that degrades in accuracy due to the presence of background noise to a noise-robust one in terms of user experience and driving performance. Furthermore, by examining the voice-controlled systems in the presence of background noise, the study also aims to investigate the effects of accuracy degradation when context (i.e., background noise) is introduced. This paper presents preliminary findings of the study, focusing on user acceptance and driving performance of the first 20 participants who completed the study. In particular, user acceptance and some driving performance metrics were contrasted between the use of the noise-robust and the noise-sensitive systems under three background noise conditions (none, music, children). A baseline condition with no system assessed driving performance under the three background noise conditions.

METHODS

This study followed a 3x3 within-subjects design, manipulating the voice-controlled system used: baseline (no voice-controlled tasks), noise-sensitive, and noise-robust voice-controlled systems, as well as background noise type: none, music at 60 dB, and backseat children noise at 70 dB. The sensitive system, designed to represent accuracy degradation in the presence of disruptive background noise, had high accuracy (90%) with no background noise, medium accuracy (70%) with music, and low accuracy (30%) with the louder background noise of children talking. The robust system, representing a system unaffected by background noise, performed at 90% accuracy under all three background noise conditions. Following a 'Wizard of Oz' approach, the voice-controlled systems were simulated using pre-programmed graphics, and the accuracy level was manipulated by the experimenter without the knowledge of the participant.

Participants

Participants were recruited through social media and online classified advertisement services from the Greater Toronto Area, Canada. They were required to have normal hearing, have had a full driver license for at least three years, and have driven at least 1600 km in the last year. Participants were also required to have normal vision or be able to wear contact lenses, and they were screened for proneness to simulator sickness. The compensation was C$15/hr, with up to C$5 task bonus for secondary task performance (all participants received C$5 bonus at the end of the study as system accuracy and hence secondary task performance were controlled in the experiment). The study was approved by the University of Toronto Research Ethics Board.

Twenty-two participants have completed this on-going study, but two were excluded from analysis due to sensor issues. The analyzed sample consists of 20 participants (11 male and 9 female), aged 25-40 (M=31, SD=4.7).

Apparatus

A NADS quarter-cab MiniSim™ Driving Simulator was used, which consists of three 42" widescreen displays, with a 130° horizontal and 24° vertical field of view at a 48" viewing distance. Driving performance data was collected at 60 Hz. While not reported in this paper, the study also collected gaze data via an eye tracking system, and heart rate and Galvanic skin response using physiological sensors.

A Microsoft Surface Pro 2 was used to present the voice-controlled systems to participants. This tablet was positioned to the right of the simulator's dashboard. Background noises were played through the simulator's speakers, initiated by triggers in the driving scenarios after participants drove 1812 ft (552 m) in a scenario. Sound samples used for creating the background noise of children arguing were obtained from an open source directory (ETSI, 2006), and a song for the music background (Michael Jackson's "Billy Jean") was purchased through Apple's iTunes.

A program developed in Python ran on a separate PC allowed the experimenter to control images displayed on the Surface in real time to imitate a voice-controlled infotainment system.

Driving Scenario

All experimental drives used the same driving environment, which consisted of an approximately 22912 ft (6984 m) section (approximately 7 minutes to complete) of a 45ft (13.7m) wide, undivided four lane road in an urban environment with light oncoming traffic and a posted speed limit of 40 mph. Participants were instructed to follow a lead vehicle, which was programmed to maintain a headway time of 1.5s to the participant's vehicle except when braking. Four lead vehicle braking events occurred in each drive, at 16%, 30%, 40%, and 89% of the drive in distance. Participants undertook nine experimental drives in blocks of three, each block testing a different voice-controlled system across all three noise conditions. The voice-controlled systems and the noise conditions experienced were administered in a counterbalanced order across participants.

Secondary Tasks

Tasks were chosen to resemble typical requests of an in-car voice command system, and fell into three categories: music tasks (i.e., find this song), navigation tasks (i.e., find directions to this location or search for locations nearby), and contact tasks (i.e., find this contact's phone number or address). During each experimental drive, participants were asked to interact with the supplied "voice-controlled system," while maintaining their primary task of "driving as safely as possible." Participants were prompted verbally by the experimenter to attempt each task (e.g., "Please use the voice-controlled system to find pizza restaurants around the University of Toronto"). Participants would initiate the voice-controlled system by saying "Hey VC", and once the system responded with an alert chime and displayed "Listening", participants would continue to issue a voice command that they judged would best complete the task (e.g., "Show me pizza restaurants around the University of Toronto" or "What are the pizza restaurants around the University of Toronto?"). The experimenter controlled the secondary display to provide a canned visual response imitating the voice-controlled system.

In each drive, there were 10 interactions with the voice-controlled system (baseline drives excluded). Participants were told to repeat a task continuously until it was successfully completed. System accuracy was pre-set in terms of interactions. For example, a 70% accuracy condition would have seven successful interactions and three failures. A failed interaction could be either planned (canned erroneous display for the task, e.g., showing pizza restaurants in an incorrect but similar sounding location) or occurred when participants provided incorrect verbal input, in which case the system would display a generic "Cannot Understand, Please Repeat" error. The latter case has been exceedingly rare thus far in our data collection.

Procedure

Informed consent was followed by a short questionnaire on demographics, driving history, and technology use. The eye tracker was calibrated and physiological sensors were attached to the participants. Through a practice drive, participants were familiarized with the simulator, introduced to the voice-controlled tasks, and were monitored for signs of simulator sickness. Participants were informed that researchers were evaluating two new voice-controlled systems and were instructed on how to initiate and interact with the voice-controlled systems.

Following each voice-controlled system drive (six total), participants were asked to complete a short user acceptance questionnaire (Van Der Laan, Heino, & De Waard, 1997), a nine item scale assessing how useful and satisfying users find a system. Participants were also asked to rate the workload they experienced using the voice-controlled system; the Rating Scale Mental Effort (De Waard, 1996) was used for this purpose. After all experimental drives had been completed, participants filled out a final general survey on personality and driving behaviours.

RESULTS

While the current study collected various measures, many were outside the scope of this paper (gaze data, physiological data, and subjective responses on technology experience, driving behaviours/history, and personality). The present analysis focuses on user acceptance and driving performance during car following (excluding lead vehicle braking events).

To assess user acceptance, responses were averaged across the 5-point Likert scale items of the usefulness and satisfying subscales (Van Der Laan et al., 1997), respectively. The usefulness and satisfying scores were computed for each of the six drives that contained voice-command tasks. Driving performance measures were computed for all drives to obtain the average speed, standard deviation (SD) of speed, and SD of lane position for each experimental drive, excluding periods of lead vehicle braking (from the onset of lead vehicle brake lights to the lead vehicle regaining a speed of 40 mph).

Mixed linear models were built using R's "nlme" package for each of the dependent measures described above. Voice-controlled system type (none/baseline, noise-sensitive, and noise-robust), noise type (none, music, children), and their interaction were included as fixed effects, and participant was included as a random effect. SD speed was log transformed to correct for heteroscedasticity.

Participants viewed both systems positively in terms of usefulness and how satisfying they found the experience (Figure 1). For both usefulness and satisfying measures, there were significant interaction effects between voice-controlled system type and noise type, and significant main effects of noise type (Table 1). The sensitive system, compared to the robust system, scored lower on both usefulness and satisfying scales, under both music (usefulness: $\Delta = -0.48$, $t(95) = -4.03$, $p = .002$; satisfying: $\Delta = -0.46$, $t(95) = -3.42$, $p = .01$) and child noise (usefulness: $\Delta = -0.81$, $t(95) = -6.80$, $p = .002$; satisfying $\Delta = -1.09$, $t(95) = -8.04$, $p < .001$) conditions. Further, follow-up comparisons revealed that these differences observed in the acceptance responses were greater for the child noise condition than they were in the music condition for both usefulness ($\Delta = 0.48$, $t(95) = 4.03$, $p = .002$) and satisfaction ($\Delta = 0.54$, $t(95) = 3.97$, $p = .002$).

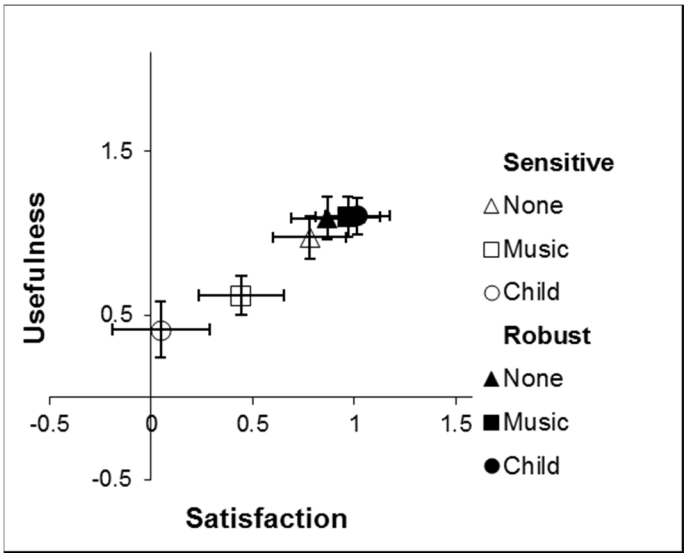

Figure 1. User acceptance results with standard error

The type of voice-controlled system participants experienced first did not affect their ratings. Welch two sample tests did not find differences in the usefulness ($t(17.4) = -0.28$, $p = .78$) and satisfying ($t(14.9) = -0.27$, $p = .79$) scores of the sensitive system between those participants who experienced the sensitive system before the robust system (n=11), and those who received robust system first (n=9). Thus, the experience with one system did not appear to create certain expectations for the performance of the next one.

Table 1. F-Statistics from mixed linear models on user acceptance and driving performance

	System Type		**Noise**		**System Type*Noise**	
Response Variable	F-value	*p*	F-value	*p*	F-value	*p*
Usefulness	F(1,95) = 1.59	.21	F(2,95) = 18.32	**<.0001**	F(2,95) = 7.68	**.001**
Satisfying	F(1,95) = 1.03	.31	F(2,95) = 23.06	**<.0001**	F(2,95) = 12.73	**<.0001**
Average Speed	F(2,156) = 13.36	**<.0001**	F(2,156) = 0.40	.67	-	
Log (SD Speed)	F(2,156) = 22.08	**<.0001**	F(2,156) = 0.26	.77	-	
SD Lane Position	F(2,156) = 4.50	**.01**	F(2,156) = 0.38	.69	-	

For driving performance measures, the interaction between system type and background noise was not significant, and was excluded from the final models. Voice-controlled system type had a significant effect on average speed, SD speed, and SD lane position (Table 1). Differences existed only between the baseline and voice-controlled system conditions. Compared to baseline, both noise-sensitive and noise-robust systems had lower average speed (sensitive: Δ= -1.73, $t(156) = 4.76$, $p < .0001$; robust: Δ= -1.50, $t(156) = 4.12$, $p < .001$) and higher SD of speed (sensitive: $\Delta = 0.15$, $t(156) = -6.17$, $p < .0001$; robust: Δ= 0.13, $t(156) = -5.22$ $p < .0001$). Finally, SD of lane position was smaller for sensitive (Δ= -0.040, $t(156) = 2.72$, $p = .02$) and robust (Δ = -0.036, $t(156) = 2.46$, $p = .04$) systems, relative to baseline driving.

DISCUSSION

Preliminary findings show that participants found both voice-controlled systems to be useful and satisfying in general, meriting the use of voice commands for in-vehicle infotainment tasks. Higher satisfaction and perceived usefulness was observed with the noise-robust system compared to the noise-sensitive one when background noise (music or children) was present; the effect was larger in the presence of child noise. Thus, it appears that system reliability affects user acceptance even if an explanation for degradation in reliability was available (i.e., the background noise). Such observations have been noted by Lee & See (2004), where automation failures, even due to environmental factors external to the automation, can cause mistrust and disuse of a system. Furthermore, as the order of experiencing noise-sensitive or noise-robust system did not affect the user acceptance scores, there was no evidence of calibration in user acceptance in the presence of a more reliable voice-controlled system. However, the user acceptance results did not consider participants' previous experience with in-car voice command systems or general technology use. It is possible that users with a more solid understanding of in-car voice command systems may be better at calibrating their trust in the system, as they may better understand its potential limitations in noisy environments. Future work should investigate if such relationships exists.

For data collected from 20 participants so far, driving performance (speed, SD of speed, and SD of lane position) was affected by interacting with a voice-controlled system relative to baseline driving, but not affected by the level of accuracy or the type of voice-controlled system used. Previous research, limited in numbers, had not been conclusive of the effects of voice recognition accuracy level on driving performance either. Kun et al. (2007), also with a sample size of 20, found larger variance in steering angles with lower voice recognition accuracy, but no effect on speed or lane position variability. McCallum et al. (2004) also did not find significant effects of voice recognition accuracy level on average speed and lane position variability, but noted a non-significant trend towards improved responses to emergency braking events with enhancing accuracy (with n=24). Current analyses did not examine lead vehicle braking events, but future analysis will investigate whether response to lead vehicle braking is affected by recognition accuracy, mediated through user acceptance. The larger sample size (n = 36) that will be obtained at the completion of this study will also provide more power for detecting any other potential effects that may exist. Analysis of other data collected in this study, such as physiological data and survey responses, will also enhance our understanding of how previous

experience, personality, and stress may have affected participants' perception of the voice-controlled systems and their driving performance while using these systems.

The current study was limited in the design of noise conditions, where two active noise conditions were of both a different nature and different volume. Music is likely to be pleasant and entertaining, while child noise may be irritating or stressful. Future research should vary the background noise in a more systematic manner in the study of voice-controlled systems. Introduction of another condition, where the accuracy levels of the voice-controlled system degrades without any background noise, can help further assess the effects of context on user acceptance.

ACKNOWLEDGEMENTS

The funding of this work was provided by MITACS in partnership with Qualcomm Canada Inc. Many thanks to Qualcomm Canada for providing advice on the design of our voice-controlled system simulation. We would also like to acknowledge Nick Mirjalali for his assistance with the experiment.

REFERENCES

De Waard, D. (1996). *The Measurement of Driver's Mental Workload*. University of Groningen, Haren, NL.

ETSI - European Telecommunications Standards Institute. (2006). Retrieved May 30, 2016, from http://www.etsi.org/

Kun, A., Paek, T., & Medenica, Z. (2007). The effect of speech interface accuracy on driving performance. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (pp. 1326–1329). Antwerp, Belgium.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Proceedings*, *22*(4), 745–777.

McCallum, M. C., Campbell, J. L., & Richman, J. B. (2004). Speech recognition and in-vehicle telematics devices: Potential reductions in driver distraction. *International Journal of Speech Technology*, *7*(1), 25–33.

Ranney, T. A., Harbluk, J. L., & Noy, Y. I. (2005). Effects of voice technology on test track driving performance: Implications for driver distraction. *Human Factors*, *47*(2), 439–454.

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). Cambridge, MA: Massachusetts Institute of Technology.

Van Der Laan, J. D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*(1), 1–10.

IHS Technology. (2013). Voice Recognition Installed in More than Half of New Cars by 2019. Retrieved October 17, 2016, from https://technology.ihs.com/427146/voice-recognition-installed-in-more-than-half-of-new-cars-by-2019