PERFORMING E-MAIL TASKS WHILE DRIVING: THE IMPACT OF SPEECH-BASED TASKS ON VISUAL DETECTION

Joanne L. Harbluk, Simone Lalande
Transport Canada
Ottawa, Ontario, Canada
E-mail: harbluj@tc.gc.ca

Summary: Drivers listened and responded to e-mail messages presented in a human voice and two types of synthetic speech (concatenative and formant) while driving a simulator. Their performance for visual event detection, vehicle control, and message responses was assessed. Results indicated that the type of speech output system affected drivers' detection of visual changes in the driving environment; they were poorer at detecting these events when either of the synthetic speech systems was used. Drivers detected fewer visual changes during the difficult messages than during the baseline driving. No effects of the speech system type or e-mail message difficulty were observed on the vehicle control measures. Drivers were also less accurate when responding to message content for messages presented in synthetic speech (concatenative) compared with recorded human voice. Subjective ratings indicated that listening to the synthetic speech required more mental effort than listening to the recorded human voice. Preference ratings for the interfaces decreased as mental effort increased. The results indicated that although drivers were not required to direct their attention away from the road, using the speech-based interfaces reduced drivers' visual event detection and their response accuracy to messages themselves.

INTRODUCTION

Speech-based interfaces are often proposed as safer alternatives to visual/manual interfaces for in-vehicle use. While these systems may provide safety improvements relative to systems requiring visual/manual interaction, research suggests that there may still be considerable cognitive costs associated with their use while driving. Lee, Caven, Haake and Brown (2001) reported a considerable increase in reaction time for braking responses when compared with driving with no task. Jamson, Westerman, Hockey and Carsten (2004) have also reported negative impacts on braking anticipation and shorter time to collision during speech interface use. Ranney, Harbluk and Noy (2005) found that, although visual target detection was better when using a speech-based interface compared with a visual/manual interface, visual target detection performance still suffered relative to a no-task baseline.

Continued interest in text-to-speech (TTS) and its various applications remains strong in the automotive sector. Producers of automatic speech recognition (ASR) and TTS systems have targeted telematics applications as one of their two primary global markets (Schalk, 2002). Specific applications include navigation systems, e-mail readers and systems that provide traffic information. The information presented by these systems is often generated by synthetic speech due to the large and/or frequently changing information databases, which make the use of recorded human speech impractical (Lai, Wood & Considine, 2000). Consequently, the comprehensibility of TTS output and its potential to impact task performance has been an area of research interest. Lai et al. (2000) reported no significant differences in the comprehension of synthetic speech for a number of text-to-speech engines. When Tsimhoni, Green and Lai (2001) extended the research to the driving domain, however, they found that drivers' comprehension of text-to-speech passages was significantly worse than their comprehension of natural speech, but they found no effects for speech type (natural or synthesized) or message type (navigation, e-mail, news) on the driving performance measures they examined (standard deviation of lane position and steering wheel angle).

There is reason to believe that listening to synthetic speech may be qualitatively different than listening to human speech. Luce, Feustel and Pisoni (1983), for example, found that listening to and processing synthetic speech can increase workload relative to human speech due to encoding difficulties and increased processing demands for what is encoded.

In the present study we wanted to address two questions. First, does using a speech-based system while driving result in differences in driver performance? Second, does the type of speech system matter? Participants drove while listening to e-mail messages read by three types of speech output systems: recorded human voice and two types of text-to-speech (TTS) systems. These were concatenative (where segments of recorded human are strung together to form words) and formant-based (where synthesis models the formant frequencies of the vocal tract). Both easy and difficult e-mail messages were presented and drivers were required to respond "yes" or "no" to the questions posed in the e-mails. Drivers' performance on event detection (changing images in side mirrors), message responses, vehicle control as well as ratings of mental effort and interface preference were measured. Comparisons were made with data obtained during baseline driving where no additional tasks were present.

METHOD

Participants. Twelve drivers (7 female and 5 male) aged 20 to 34 years old (M=24.83, SD=4.49) were recruited from a local university and paid for their participation. All were experienced drivers (minimum 3 years' driving experience; 10,000 km or more annually). All had normal or corrected to normal vision and their hearing was pretested to ensure it was in the normal range.

Equipment and Materials. A Systems Technology Incorporated fixed-base driving simulator (STISIM) was used with graphics projected to provide a 135-degree forward field of view. Auditory cues included throttle-linked engine noise and wind noise when the subject vehicle passed oncoming vehicles. Fifty-two e-mail messages, designed to represent typical business and personal messages were constructed. Baddeley (1968; Hitch & Baddeley, 1976) had demonstrated that manipulations of active/passive and affirmative/negative sentence construction impact speed and accuracy of responses. This manipulation was used to create Easy messages (characterized by an active/affirmative sentence construction) and Difficult messages (characterized by a passive/negative sentence construction; see Tables 1 and 2). Drivers were required to respond aloud "yes" or "no" confirming the last statement presented in each e-mail. Both synthesized voices were generated using Fonix iSpeak v.3 and the human voice was recorded from a male in his 20's. Mean message length was 28.34 s. An earcon signaled the beginning of each message.

Design and Procedure. The design was 4X2 repeated measures factorial with Interface Type (None, Human, Concatenative and Formant Speech Types) and Message Complexity (Easy, Difficult) as within-subject variables. Presentation order of the interface conditions and message difficulty was determined by Latin Square. Accuracy and response time were collected for both the event detection task and the e-mail responses. Driving measures were longitudinal speed (m/s, vehicle speed), lateral lane position (m, deviation from centre dividing lane) and steering wheel position (degrees, angle of turn). Mental effort ratings and interface preference ratings were obtained using a scale of 1 (lo) –10 (hi).

Table 1. Example of Easy Message
(Active/Affirmative Construction Using "Follows")

Subject: Management Course Offerings
Hello:
We are pleased to announce a new series of management courses for this session.
The course for New Managers follows the course on Managing Projects.
Please let us know your course requirements and preferred dates.
Please confirm: The order for the courses is Managing Projects then New Managers.
---
The correct answer to the question is YES.

Table 2. Example of Difficult Message
(Passive/Negative Construction Using "Is Not Preceded By")

Subject: Book Club Meetings
Hello:
Your book club is scheduling the next two monthly meetings to discuss the latest novels.
The meeting at Peter's house is not preceded by the meeting at Jeff's house.
You will have to decide which month you want the meeting to be held at your house.
Please confirm: The location of the book club meetings is Jeff's house then Peter's house.
---
The correct answer to the question is NO.

After familiarization with the procedures, the drivers drove the 10km route keeping to 80km/h. The route was a two-lane country highway with straight sections and an equal number of right- and left-hand curves. Drivers were to respond aloud to the e-mail messages when they occurred. Periodically, one of the green diamond symbols presented in one of the side mirrors would change to a triangle of the same color and remain visible for 2 seconds before reverting to its original shape. Drivers were to indicate this change by activating the appropriate turn signal (left or right) as quickly and accurately as possible. Subjective ratings of mental effort and interface preference were collected after each drive.

RESULTS

Comparisons of performance were made with the comparable route segments from the no-task baseline drive. Post hoc tests are reported at .05.

Event Detection

Event Detection Accuracy. The performance for event detection was affected by the type of interface the drivers were using ($F_{(3,33)}$ = 4.35, p<.01). As shown in Figure 1, detection performance was poorest when drivers listened to synthesized speech (concatenative 65.28, formant 71.53). This difference was significant in the comparison between none and concatenative and marginally significant in the comparison between none and formant (p=.08). Drivers detected a similar proportion of the events when listening to the messages presented in the recorded human voice (81.95) as they did when driving without any messages (81.11).

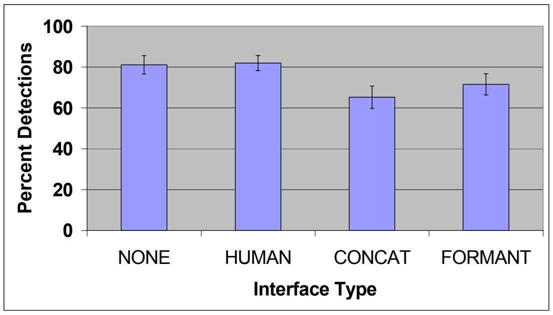

Figure 1. Percent Detections for Speech Types

Message difficulty, shown in Figure 2, had a marginal effect on event detection ($F_{(2,22)}$ = 3.22, p=.059). Listening to either the easy (75.93) or difficult (69.91) messages resulted in fewer detections compared to the baseline (81.11), but this reduction was significant only in the case of the difficult messages. The interaction between interface type and message difficulty was not significant ($F_{(2,22)=}$ 1.46, p>.05).

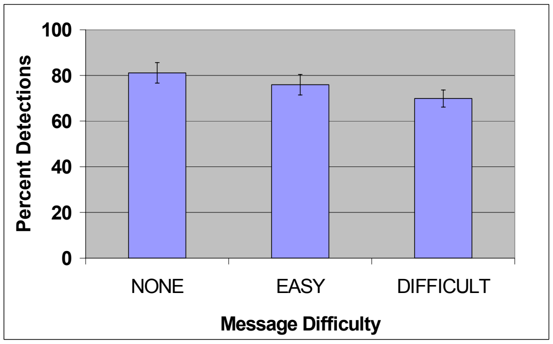

Figure 2. Percent Detections for Message Difficulty

Event Detection Latency. Interface type did not affect the event detection latencies which ranged from 1.65 to 1.73s across the four conditions ($F_{(3,33)}$ <1, p>.05). Message difficulty, however, did affect detection latency. Drivers detected the events more quickly during easy message presentations (1.60) than during difficult message presentations (1.81; $F_{(2,22)}$ = 6.68, p<.05). The performance for the no message (1.70) condition fell in between those values. The interaction was not significant ($F_{(2,22)}$ =1.47, p>.05).

Driving Performance

Comparisons of driving performance were made across the comparable 30s driving segments in the various conditions. A number of measures of vehicle control were obtained while drivers drove in the various conditions (see Table 3). However, no significant differences as a function of interface type or message difficulty were observed in the data collected.

Table 3. Vehicle Control Measures

Driving Performance Measure	Range of Values Obtained Across Conditions
Longitudinal Speed (mean)	22.32 – 22.73 m/s
Longitudinal Speed (SD)	.45 - .57 m/s
Lateral Lane Position (SD)	.41 - .45 m
Steering Wheel Angle	10.48 – 10.76 degrees

Performance on Message Tasks

Message Response Accuracy. The type of speech used to present the messages influenced the drivers' accuracy in responding to the messages. As shown in Figure 3, drivers responded most accurately to messages presented in the human voice (89.58; $F_{(2,22)} = 3.37$, p=.05). Responses were less accurate for messages presented in both the concatenative (82.64) and the formant synthesized voices (85.42); significantly so in the case of the concatenative voice. Message difficulty did not affect drivers' response accuracy to the messages ($F_{(2,22)}$ =1.79, p>.05). The interaction of interface type and message difficulty for message performance was not significant ($F_{(2,22)}$ =1.24, p>.05).

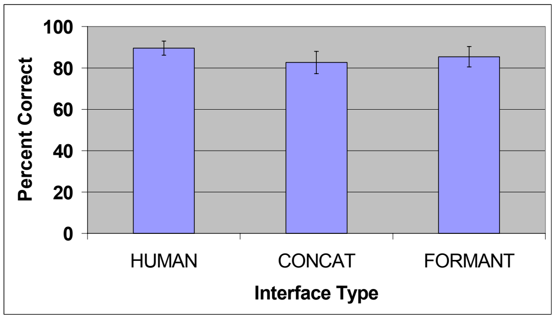

Figure 3. Message Response Accuracy as a Function of Interface Type

Message Response Latency. There was a marginal main effect for interface type on message response latency ($F_{(2,22)} = 2.96$, p=.07). Drivers tended to respond more quickly to the messages presented in the human voice (.78) compared with the concatenative (1.07) and the formant (1.37) voices. Post hoc tests indicated the difference between human and formant was marginally significant at p = .09. Neither the effect for message difficulty ($F_{(1,11)} = 1.19$, p>.05) nor the interaction ($F_{(2,22)}$ =1.49, p>.05) was significant.

Subjective Ratings of Mental Effort and Preference

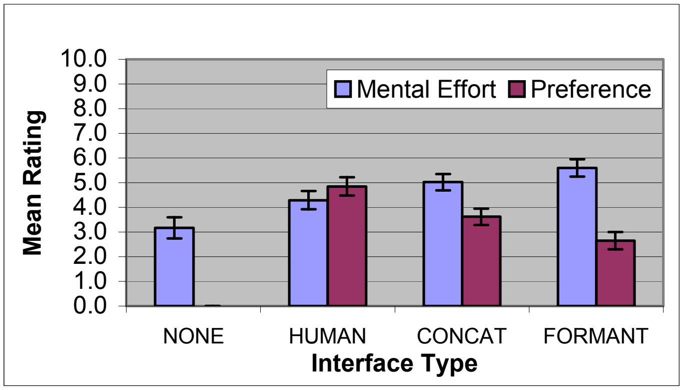

Figure 4. Mean Ratings of Mental Effort and Preference

As shown in Figure 4, drivers indicated increasing amounts of mental effort associated with the no task, human voice, concatenative speech and formant speech conditions, respectively. The lowest level of mental effort was found in the no-task condition, which differed significantly from all other conditions ($F_{(3,33)}$ =12.63, p<.0001). Although the ratings for mental effort increased numerically for the human, concatenative and formant conditions, only the difference between the formant and human conditions was significant. When the impact of message difficulty on drivers' mental effort ratings was examined, significantly higher ratings were given when doing any task, easy or difficult, compared to driving with no task ($F_{(2,22)}$ =18.27, p<.0001; ps<.05). There was no significant difference between easy and difficult tasks. The interaction was not significant ($F_{(2,22)}$ =1.83, p>.05). Preference ratings for the interfaces decreased as mental effort increased, as shown in Figure 4. All differences among the means differed statistically for the preference ratings ($F_{(2,22)}$ =12.04, p<.0002).

DISCUSSION

Speech-based interfaces are proposed as safer alternatives for in-vehicle systems where reading large amounts of information from a screen would take drivers' eyes away from the road for extended periods of time. In this study, we investigated the impact of using a speech-based e-mail system on driver performance. Three types of speech output systems (recorded human voice and two types of synthetic speech) were used to read e-mail messages. Drivers were assessed on visual event detection, vehicle control and performance on the message tasks. Ratings of mental effort and interface preference were collected. The results indicated that using a speech-based system while driving can negatively impact driver performance. Drivers were most accurate at detecting the visual events when they were not engaged in any e-mail task. When using the e-mail system, the type of speech system mattered. Visual event detection was poorest when drivers were using the synthetic concatenative speech system. Detections were also reduced during the presentation of the difficult messages.

Interestingly, none of the vehicle control measures recorded in this study proved sensitive to any of the interface type or message difficulty manipulations. When no effects are found, there is always concern that the measures were not sensitive enough or that the correct measures were not chosen. Other researchers, however, investigating the distraction potential of speech systems have reported a similar lack of effect on driving measures. Tsimhoni et al. (2001) found that neither the speech type used nor the message type had a significant effect on the basic driving measures they used, standard deviation of lane position and steering wheel angle. Seppelt and Wickens (2003) also reported that drivers who were involved with in-vehicle tasks preserved their lane-keeping performance but their hazard response suffered.

An additional question of interest, although not directly related to driving, concerned drivers' performance on the messages themselves. Responses were more accurate to the e-mail messages when they were presented in human speech rather than concatenative speech. There was also a marginal tendency for drivers to respond more quickly when messages were presented in the human voice. These findings are consistent with the idea that that it is easier to listen and respond to human speech compared with synthetic speech (Luce et al, 1983). Drivers reported increased workload associated with the synthetic speech systems and this was also reflected in their reduced preference ratings for those systems.

A number of factors that were not addressed in this study provide interesting directions for future work. Greater demands could be placed on the drivers by manipulating the driving environment and task complexity. Driver factors such as experience with speech systems and driver age could also be examined.

As the use of speech-based systems increases for automotive use, it is important that the safety impacts of both text-to-speech and voice recognition systems (e.g., Schreiner, Blanco & Hankey, 2004) are addressed. While it is likely that speech-based interfaces may be safer than visual-manual interfaces in some applications, significant cognitive demands can be associated with their use. In the present study, drivers did not have to look away from the road to complete the tasks, yet they were less likely to detect visual changes in the driving environment under the synthetic speech conditions. Not all speech systems are created equally; there are differences among types of speech systems in their impact on driver behaviour.

This research demonstrates the importance of using additional measures beyond vehicle control, such as event detection, when investigating the safety impact of speech-based in-vehicle technologies. Performing an auditory task produced a decrement in visual detection that would have been missed if only driving performance measures had been examined. This study is part of a larger research program investigating the safety and human factors issues involved with speech-based in-vehicle devices. The ultimate goal of this research program is to produce guidelines for in-vehicle tasks that are compatible with the task of driving.

REFERENCES

Jamson, A.H., Westerman, S.J., Hockey, G.R., & Carsten, O.M.J. (2004). Speech-based email and driver behaviour: Effects of an in-vehicle message system interface. *Human Factors, 46:* 625-639.

Lai, J.L., Wood, D., & Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech. ACM SIGCHI CHI 2000 Proceedings: 321-328.

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*: 631-640.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors, 25*: 17-32.

Ranney, T.A., Harbluk, J.L., & Noy, I. (In press). The effects of voice technology on test track driving performance: Implications for driver distraction. *Human Factors*.

Schalk, T. (2002). AVIOS Column: The evolution of global speech technology. http://www.speechtechmag.com

Schreiner, C., Blanco, M., & Hankey, J. (2004). Investigating the effect of performing voice recognition tasks on the detection of forward and peripheral events. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.*

Seppelt, B. & Wickens, C.D. (2003). In-vehicle tasks: Effects of modality, driving relevance, and redundancy. Aviation Human Factors Division Institute of Aviation; University of Illinois.Technical Report AHFD-03-16/GM-03-2 prepared for General Motors Corporation.

Tsimhoni, O., Green, P., & Lai, J. (2001). Listening to natural and synthesized speech while driving: Effects on user performance. *International Journal of Speech Technology, 4:* 155-169.