The Effect of Speech Interface Accuracy on Driving Performance

Andrew Kun [1], *Tim Paek* [2], *Zeljko Medenica* [1]

[1] University of New Hampshire, ECE Department, Kingsbury Hall, Durham, NH 03824, USA
[2] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

andrew.kun@unh.edu, timpaek@microsoft.com, zeljko.medenica@unh.edu

10.21437/Interspeech.2007-406

Abstract

With the proliferation of cell phones around the world, governments have been enacting legislation prohibiting the use of cell phones during driving without a "hands-free" kit, bringing automotive speech recognition to the forefront of public safety. At the same time, the trend in cell phone hardware has been to create smaller and thinner devices with greater computational power and functional complexity, making speech the most viable modality for user input. Given the important role that automotive speech recognition is likely to play in consumer lives, we explore how the accuracy of the speech engine, the use of the push-to-talk button, and the type of dialog repair employed by the interface influences driving performance. In experiments conducted with a driving simulator, we found that the accuracy of the speech engine and its interaction with the use of the push-to-talk button does impact driving performance significantly, but the type of dialog repair employed does not. We discuss the implications of these findings on the design of automotive speech recognition systems.

Index Terms: automotive speech recognition

1. Introduction

With the proliferation of cell phones around the world, and the distraction posed by their use on driving, many governments have been enacting legislation to prohibit their use during driving without a "hands-free" kit, bringing automotive speech recognition to the forefront of public safety. For example, in the U.S., the Governors Highway Safety Association [8] reports that 6 states (NY, CA, CT, DC, NH, NJ) have already adopted hands-free laws which affect about 23% of the general U.S. population. Because of these laws, many mobile operators now require device manufacturers to ship at least voice-dialing.

Furthermore, with voice-dialing as a feature on almost every cell phone, mobile speech recognition has successfully reached the wider consumer market, albeit sometimes with rudimentary pattern recognition. However, as mobile devices have increased in memory capacity and computational power, it has become possible to not only support client-side recognition using a speech engine but also a whole range of multimedia functionalities (such as music, navigation and email) that can be controlled with speech. In fact, the trend for cell phones has been to create smaller devices with greater functional complexity. Because speech is an input modality that can scale to smaller form factors than manual and visual interfaces [6], it is likely to be a key modality for interaction.

Although research studies have been conducted on the use of automotive speech interfaces on driving performance, to date, no study has explored how the accuracy of the speech engine, the use of the push-to-talk (PTT) button, and the type of dialog repair employed by the speech interface affect driving. Given the important role that automotive speech recognition is likely to play in consumer lives, this paper endeavors to fill this research gap. The paper is organized as follows. After discussing related research in Section 2, we describe and present the results of the experiment we conducted using a driving simulator in Section 3. In Section 4, we discuss the implications of these results on the design of automotive speech recognition systems. Finally, in Section 5, we outline directions for future research.

2. Related Research

Because the vast majority of drivers (60 to 70%) report using their cell phones at least sometimes during driving, many research studies have examined how cell phone use can cause driver distraction, and how driver distraction can then lead to accidents (see [2] for literature review). Distraction comes in two forms: first, *physical distraction* occurs when drivers have to simultaneously operate their phones while controlling their vehicles; and second, *cognitive distraction* occurs when drivers have to divert at least part of their attention to the phone conversation at hand.

The hands-free laws that have been enacted are geared toward handling physical distraction. Research studies examining how speech interfaces influence driving performance indicate that people generally drive at least as well, if not better when using speech interfaces than manual interfaces in tasks such as music selection, destination entry [3] and operating police radios [5]. However, to date no research study has investigated how recognition accuracy and other aspects of a speech interface affect driving, and in fact, a recent literature review [2] highlights the need for such research. After all, many "hands-free" kits still involve pressing a push-to-talk button on either the handset itself or mounted somewhere, and if recognition problems occur, it is at least possible that drivers will physically move closer to the device or be cognitively distracted by frustration.

Apart from the above studies, researchers have explored a cognitive architecture for predicting the effect of automotive interfaces, such as voice-dialing, on driving performance [7]. Although we could have used the proposed cognitive architecture to predict performance, we would still have needed to validate the predictions with an empirical experiment, which we now describe.

3. Experiment

In order to assess the effect of recognition accuracy on driving performance, we conducted a factorial design experiment using three factors:

Recognition Accuracy: For accuracy, we compared two levels: *High* at 89% and *Low* at 44%. In order to absolutely fix accuracy, we did not utilize the recognizer but instead instructed participants to issue *pre-selected* voice commands and had the system respond either correctly or incorrectly 89% of the time or 44% of the time.

PTT: In order to rule out any driving effect due to the manual act of employing the PTT button, we compared two input methods: using a PTT button that was mounted on the center console to initiate interaction or *ambient recognition* where the participant issues a command and it is automatically recognized. We placed the PTT button on the center console where it was only reachable by moving a hand from the wheel, and not on the steering wheel where a participant could have operated the PTT button without moving a hand.

Dialog Repair: Failure to recognize a voice command can be manifested by the system in several ways. For our experiment, we had the system respond with either a *misunderstanding* (incorrect recognition) or no understanding at all.

Dialog Repair was treated as a between-subjects variable, and Recognition Accuracy and PTT as a within-subjects variable. In summary, participants either had all misunderstandings or non-understandings, and within each run of the simulator, they used either the PTT or nothing to initiate recognition. Regardless of what participants uttered, the system responded at either the *High* or *Low* Recognition Accuracy levels.

3.1. Driving environment

We conducted our experiments in a high-fidelity driving simulator with a 180º field of view and a motion base (Figure 1). The simulation presented a two-lane, 3.6 m wide, curvy road in daylight. Participants saw a leading vehicle that traveled at 97 km/h (60 mph). No other traffic was present.

Figure 1 Driving simulator.

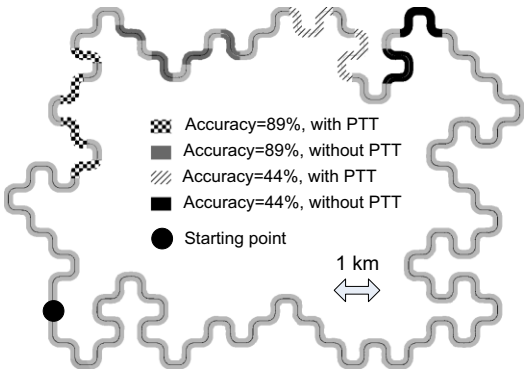

Figure 2 The simulated road. The shaded segments are where one representative subject performed the four interaction sets.

Figure 2 shows the simulated road used in our experiment. It also shows the road segments where one representative subject performed the four interaction sets. Because the simulated road had many curves, participants spent approximately 65-70% of their interaction times driving in the curves. We selected this layout because driving in curves forces participants to pay attention to the driving task and not just to the spoken task.

3.2. Procedure

Participants were given an overview of the simulator, and were trained in the spoken and driving tasks. The driving task consisted of following the leading vehicle at a constant distance without departing from the right lane. Participants did not receive reminders or warnings related to their performance on the driving task. The spoken task was to change channels, and initiate the transmission of messages, on a police radio. This task was related to the ongoing work at the University of New Hampshire on the Project54 system. The Project54 system integrates devices in police cruisers and provides a speech user interface to these devices [4]. In our experiment, participants were told they would be testing speech recognizers to be used for an in-car system for patching radio messages from police headquarters. Messages include instructions about who the message should be retransmitted to, which the system gives to the participant verbally. The participant's task is to select a radio to use to retransmit the message (using the "Zone" keyword), select a channel on that radio (using the "Channel" keyword) and initiate retransmission (using the "Retransmit" keyword). After retransmission is confirmed by the system, the participant has to return to the Troop A channel of Zone Troop A. A successful return to this channel completes the task. The example interaction below between the system (S) and a participant (P) demonstrates a successful task completion:

1 ***S**: Message received from Troop A Adam. Retransmit message to channel Bedford in zone B Boston.*
2 ***S**: Go to zone B Boston.*
3 **P**: Zone B Boston. *
4 ***S**: Zone B Boston.*
5 ***S**: Go to channel Bedford.*
6 **P**: Channel Bedford. *
7 ***S**: Channel Bedford.*
8 **P**: Retransmit.
9 ***S**: Retransmit.*
10 ***S**: Go to zone Troop A Adam.*
11 **P**: Zone Troop A Adam. *
12 ***S**: Zone Troop A Adam.*
13 ***S**: Go to channel Troop A Adam.*
14 **P**: Channel Troop A Adam. *
15 ***S**: Channel Troop A Adam.*
16 ***S**: Listening.*

Note that lines 2, 5, 10 and 13, uttered by the system, are reminders to the participant of what to say next. Thus, the participants were not required to memorize the command grammar. As such, although participants were told to use the proper command grammar during the interactions, this was not enforced.

In addition, participants were told to initiate interaction in two modes of operation, one requiring the use of a PTT button, the other utilizing ambient recognition. They heard verbal system announcements signaling changes between

recognizers and between modes of operation. Participants were not informed about the accuracy of the individual recognizers.

For Dialog Repair, to recover from a misrecognition, participants were instructed to say "Cancel" and then reissue the original command. For non-understandings, they were told to simply reissue the original command. For the two global commands "Cancel" and "Retransmit," recognition was always perfect.

The experiment was completed by 20 participants between 20 to 41 years of age (the average participant age was around 26 years, 60% were male and none were police officers). As discussed earlier, ten participants encountered misunderstandings and ten non-understandings. Each participant performed four sets of interactions, one for each of the four possible combinations of the Recognition Accuracy and PTT factors of our experiment. Within each set, participants issued a total of 18 zone and channel change commands (in the example interaction above there are four such commands, each denoted with a *). Thus, each participant issued a total of 4 x 18 = 72 zone/channel commands. They also issued several "Cancel" and "Retransmit" commands.

3.3. Data

We recorded three measures of driving performance: lane position, steering wheel angle and the velocity of participants' cars. These values were provided by the simulator and they were sampled at a 10 Hz rate. We analyzed the variances of these three variables, which represent measures of driving performance. In each case, a higher variance represents worse driving performance. We calculated the variances by taking into account the values of the variables only during spoken interactions between the system and the participant. We hand-coded the duration of spoken interactions. For each of the participants, the variances were calculated for each task and then they were averaged for each combination of the Recognition Accuracy and PTT design factors. Thus, for each participant, we calculated a total of four average variances, one for each set of tasks performed for a given combination of Accuracy and PTT usage.

Lane position represents the position of the center of the participant's simulated car, in meters. Clearly, large variances in lane position are the most serious sign of poor driving performance, since they indicate that the participant weaved in his/her lane or even departed from the lane.

Steering wheel angle is measured in degrees. In the case of curvy roads, large steering wheel angle variance is not in itself a sign of poor driving performance. After all, just following a curvy road requires varying the steering wheel angle constantly. However, steering wheel angle variance can be used as a relative measure of driving performance when comparing the performance of multiple participants on roads that represent similar driving difficulty. A higher variance is an indication of increased effort expended by a driver to remain in his/her lane.

Finally, the simulated car's velocity is measured in meters/second. A relatively large variance in the velocity of a car does not necessarily indicate unsafe driving. However, drivers often reduce speed when they are concerned about safety or when they are distracted. For example, a driver may slow down on a narrow road or when talking to a passenger.

3.4. Results

Given that lane position, steering wheel angle and velocity may be correlated with each other, we first performed a mixed model, multivariate analysis of variance (MANOVA) to determine the effect of Recognition Accuracy, PTT and Dialog Repair on driving performance. A significant main effect was found for Recognition Accuracy (Wilks's Λ=0.38, $F(3,16)=8.6$, $p=.001$), but not for PTT or Dialog Repair, the between-subjects variable. On the other hand, a significant interaction effect between levels of Recognition Accuracy and PTT was observed (Λ=0.51, $F(3,16)=5.1$, $p=.01$). No significant deviations from model assumptions were found.

In order to follow up on the significant effects, we performed a univariate ANOVA on each of the driving performance measures. We found that Recognition Accuracy affected steering wheel angle variance very significantly ($F(1,18,)=28.2$, $p<.001$), but not lane position or velocity. The mean of this variance is shown in Figure 3. The *Low* Recognition Accuracy resulted in a higher variance. This indicates worse driving performance. When the speech Recognition Accuracy was *Low* participants expended more effort on steering than when the Recognition Accuracy was *High*.

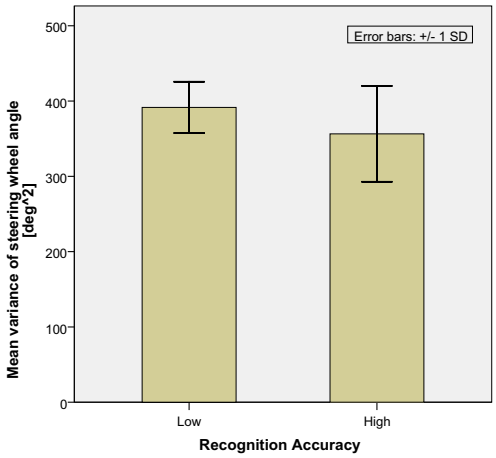

Figure 3 Steering wheel angle variance is affected by Recognition Accuracy.

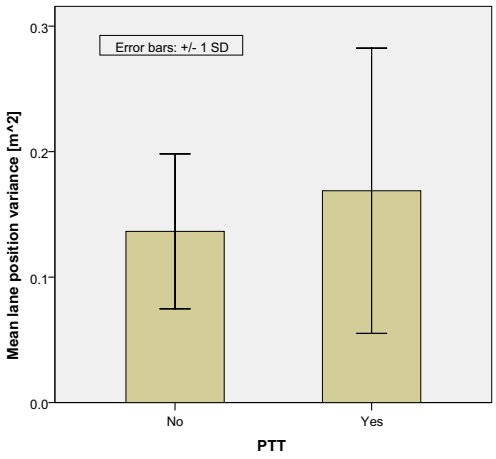

Figure 4 When Recognition Accuracy is *Low*, lane position variance is affected by PTT usage.

We also found that the interaction between Recognition Accuracy and PTT had a significant effect on lane position variance ($F(1,18)=7.1$, $p<.05$), but not steering wheel angle or

velocity. The means of the lane position variance, depending on PTT usage, are shown in Figure 4 for *Low* accuracy. When recognition accuracy was *High*, mean lane position variance did not depend on the use of the PTT button. However, when the recognition accuracy was *Low*, using the PTT button resulted in a significantly higher mean variance than not using the PTT button. In other words using the PTT button when the Recognition Accuracy was *Low* interfered with participants' ability to keep their cars in a steady position in the right lane.

4. Discussion

In our results, we found that the level of recognition accuracy significantly influenced variance in steering wheel angle, which in and of itself is not particularly serious. However, we also found a significant interaction effect between accuracy and the use of PTT on lane position, which is a serious sign of poor driving. Apparently, when recognition accuracy is high enough, operating the PTT button is not very distracting, even though the button is placed on the center console and requires the driver to release the steering wheel. However, when recognition accuracy is very low (44%), the added effort of operating the PTT button is distracting and results in worse driving performance. One reasonable explanation for this is that as users experience poor speech recognition accuracy, they become frustrated and manifest their anger by vigorously depressing the PTT button when this button is available. In effect, user frustration acts as a hidden variable.

Although follow-up studies are needed, these preliminary results have implications for the design of automotive speech recognition systems. Clearly, the ideal system would not rely on a PTT button. However, because speech recognizers do not work well without a PTT button in noisy environments, such as a car, PTT buttons are likely to be part of an in-car speech interaction system for some time. For this case, our results imply that the placement of the PTT button should be carefully chosen if recognition performance is less than stellar. It seems intuitive that placing the PTT button on the steering wheel within close reach of the hand would interfere with driving performance the least (e.g., this is what is done in most of the police cruisers that use the Project54 system). However, this intuition still needs to be validated through experiments. Furthermore, in our results, we also found that when recognition accuracy is high, operating the PTT button does not significantly interfere with driving performance, at least with the button on the center console. This implies that having a robust and accurate speech recognizer provides some flexibility for how the other elements of the in-car speech interaction system can be designed.

5. Conclusions & Future Direction

In this paper, we describe an experimental investigation of the influence of three factors of automotive speech recognition on driving performance: recognition accuracy, PTT use and dialog repair type. The results we present indicate that recognition accuracy can indeed influence driving. However, further studies need to be conducted to determine if the effect on driving poses a safety hazard to both passengers and other people on the road. We hypothesize that the answer to this question will depend on the difficulty of the driving task. As such, we are in the process of creating quantitative measures of driving task difficulty so that in follow-on studies we will be able to use multiple types of roads with varying degrees of driving difficulty.

Our results also show that PTT use can influence driving when the recognition rate is low. Once again, further studies are needed to determine how low the recognition rate needs to be for PTT use to become a problem. In our experiment we set the low recognition rate at 44%, which corresponds to unrealistically poor recognizer performance. This value was used only as a first attempt to show that low recognition rates can influence driving.

Further studies are also needed to determine how PTT button location influences driving. We are currently in the process of expanding the study to explore the effect of the PTT button location on driving performance. We expect to test using a button on the steering wheel as well as a foot pedal PTT button.

Finally, we need to explore how user frustration may result from interactions with in-car speech recognizers and how this may influence driving. In our experiment participants filled out a questionnaire with answers on a five point Likert scale, coded from 0 (not at all) to 4 (yes). When asked if they were frustrated with the speech interaction, the most frequent response code was 3 (somewhat) and 1 (not quite) for the *Low* and *High* recognition rates, respectively. Our hypothesis is that frustration affects driving performance negatively. We expect that detecting user frustration (e.g. using prosodic cues as in [1]) may be necessary in order to monitor spoken interactions and eliminate the risk posed by speech recognizers in the car.

6. Acknowledgement

Work at the University of New Hampshire was supported by the US Department of Justice under grants 2005CKWX0426 and 2006DDBXK099.

7. References

[1] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. & Stolcke, A. (2002). "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", in Proc. of ICSLP.

[2] Baron, A. & Green, P. (2006). "Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review", University of Michigan Transportation Research Institute. Technical Report UMTRI 2006-5.

[3] Dragutinovic, N. & Twisk, D. (2005). "Use of Mobile Phones While Driving – Effects on Road Safety: A Literature Review", SWOV Institute for Road Safety Research, 2005, Report R-2005-12.

[4] Kun, A., Miller, W. T. & Lenharth, W. (2004). "Computers in police cruisers," *IEEE Pervasive Computing*, 3(4): 34 – 41

[5] Medenica, Z, Kun, A. (2007). "Comparing the Influence of Two User Interfaces for Mobile Radios on Driving Performance", to appear in Proc. of Driving Assessment.

[6] Rosenfeld, R., Olsen, D., & Rudnicky, A. (2001). "Universal Speech Interfaces", *Interactions*, 8(6): 34-44.

[7] Salvucci, D. (2001) "Predicting the Effects of In-Car Interfaces on Driver Behavior Using a Cognitive Architecture", in Proc. CHI, pp. 120-127.

[8] Governors Highway Safety Association (2007). "Cell Phone Laws", retrieved February 2007, from GHSA website: http://www.ghsa.org/html/stateinfo/laws/cellphone_laws.html