Using Speech to Reply to SMS Messages While Driving: An In-Car Simulator User Study

Yun-Cheng Ju, Tim Paek
Microsoft Research
Redmond, WA USA
{yuncj|timpaek}@microsoft.com

Abstract

Speech recognition affords automobile drivers a hands-free, eyes-free method of replying to Short Message Service (SMS) text messages. Although a voice search approach based on template matching has been shown to be more robust to the challenging acoustic environment of automobiles than using dictation, users may have difficulties verifying whether SMS response templates match their intended meaning, especially while driving. Using a high-fidelity driving simulator, we compared dictation for SMS replies versus voice search in increasingly difficult driving conditions. Although the two approaches did not differ in terms of driving performance measures, users made about six times more errors on average using dictation than voice search.

1 Introduction

Users love Short Message Service (SMS) text messaging; so much so that 3 trillion SMS messages are expected to have been sent in 2009 alone (Stross, 2008). Because research has shown that SMS messaging while driving results in 35% slower reaction time than being intoxicated (Reed & Robbins, 2008), campaigns have been launched by states, governments and even cell phone carriers to discourage and ban SMS messaging while driving (DOT, 2009). Yet, automobile manufacturers have started to offer infotainment systems, such as the Ford Sync, which feature the ability to listen to incoming SMS messages using text-to-speech (TTS). Automatic speech recognition (ASR) affords users a hands-free, eyes-free method of replying to SMS messages. However, to date, manufacturers have not established a safe and reliable method of leveraging ASR, though some researchers have begun to explore techniques. In previous research (Ju & Paek, 2009), we examined three ASR approaches to replying to SMS messages: dictation using a language model trained on SMS responses, canned responses using a probabilistic context-free grammar (PCFG), and a "voice search" approach based on template matching. Voice search proceeds in two steps (Natarajan et al., 2002): an utterance is first converted into text, which is then used as a search query to match the most similar items of an index using IR techniques (Yu et al., 2007). For SMS replies, we created an index of SMS response templates, with slots for semantic concepts such as time and place, from a large SMS corpus. After convolving recorded SMS replies so that the audio would exhibit the acoustic characteristics of in-car recognition, they compared how the three approaches handled the convolved audio with respect to the top n-best reply candidates. The voice search approach consistently outperformed dictation and canned responses, achieving as high as 89.7% task completion with respect to the top 5 reply candidates.

Even if the voice search approach may be more robust to in-car noise, this does not guarantee that it will be more usable. Indeed, because voice search can only match semantic concepts contained in the templates (which may or may not utilize the same wording as the reply), users must verify that a retrieved template matches the semantics of their intended reply. For example, suppose a user replies to the SMS message "*how about lunch*" with "*can't right now running errands*". Voice search may find "*nope, got errands to run*" as the closest template match, in which case, users will have to decide whether this response has the same meaning as their reply. This of course entails cognitive effort, which is very limited in the context of driving. On the other hand, a dictation approach to replying to SMS messages may be far worse due to misrecognitions. For example, dictation may interpret "*can't right now running errands*" as "*can right*

now fun in errands". We posited that voice search has the advantage because it always generates intelligible SMS replies (since response templates are manually filtered), as opposed to dictation, which can sometimes result in unpredictable and nonsensical misrecognitions. However, this advantage has not been empirically demonstrated in a user study. This paper presents a user study investigating how the two approaches compare when users are actually driving – that is, when usability matters most.

2 Driving Simulator Study

Although ASR affords users hands-free, eyes-free interaction, the benefits of leveraging speech can be forfeit if users are expending cognitive effort judging whether the speech interface correctly interpreted their utterances. Indeed, research has shown that the cognitive demands of dialogue seem to play a more important role in distracting drivers than physically handling cell phones (Nunes & Recarte, 2002; Strayer & Johnston, 2001). Furthermore, Kun et al. (2007) have found that when in-car speech interfaces encounter recognition problems, users tend to drive more dangerously as they attempt to figure out why their utterances are failing. Hence, any approach to replying to SMS messages in automobiles must avoid distracting drivers with errors and be highly usable while users are engaged in their primary task, driving.

2.1 Method

To assess the usability and performance of both the voice search approach and dictation, we conducted a controlled experiment using the STISIM Drive™ simulator. Our simulation setup consisted of a central console with a steering wheel and two turn signals, surrounded by three 47'' flat panels placed at a 45° angle to immerse the driver. Figure 1 displays the setup.

We recruited 16 participants (9 males, 7 females) through an email sent to employees of our organization. The mean age was 38.8. All participants had a driver's license and were compensated for their time.

We examined two independent variables: *SMS Reply Approach*, consisting of *voice search* and *dictation*, and *Driving Condition*, consisting of *no driving*, *easy driving* and *difficult driving*. We included *Driving Condition* as a way of increasing cognitive demand (see next section). Overall, we conducted a 2 (*SMS Reply Approach*) × 3 (*Driving Condition*) repeated measures, within-subjects design experiment in which the order of *SMS Reply* for each *Driving Condition* was counter-balanced. Because our primary variable of interest was *SMS Reply*, we had users experience both *voice search* and *dictation* with *no driving* first, then *easy driving*, followed by *difficult driving*. This gave users a chance to adjust themselves to increasingly difficult road conditions.

Figure 1. Driving simulator setup.

Driving Task: As the primary task, users were asked to drive two courses we developed with *easy driving* and *difficult driving* conditions while obeying all rules of the road, as they would in real driving and not in a videogame. With speed limits ranging from 25 mph to 55 mph, both courses contained five sequential sections which took about 15-20 minutes to complete: a residential area, a country highway, and a small city with a downtown area as well as a business/industrial park. Although both courses were almost identical in the number of turns, curves, stops, and traffic lights, the easy course consisted mostly of simple road segments with relatively no traffic, whereas the difficult course had four times as many vehicles, cyclists, and pedestrians. The difficult course also included a foggy road section, a few busy construction sites, and many unexpected events, such as a car in front suddenly breaking, a parked car merging into traffic, and a pedestrian jaywalking. In short, the difficult course was designed to fully engage the attention and cognitive resources of drivers.

SMS Reply Task: As the secondary task, we asked users to listen to an incoming SMS message together with a formulated reply, such as:

(1) *Message Received: "Are you lost?" Your Reply: "No, never with my GPS*"

The users were asked to repeat the reply back to the system. For Example (1) above, users would have to utter "*No, never with my GPS*". Users

could also say "*Repeat*" if they had any difficulties understanding the TTS rendering or if they experienced lapses in attention. For each course, users engaged in 10 SMS reply tasks. SMS messages were cued every 3000 feet, roughly every 90 seconds, which provided enough time to complete each SMS dialogue. Once users uttered the formulated reply, they received a list of 4 possible reply candidates (each labeled as "*One*", "*Two*", etc.), from which they were asked to either pick the correct reply (by stating its number at any time) or reject them all (by stating "*All wrong*"). We did not provide any feedback about whether the replies they picked were correct or incorrect in order to avoid priming users to pay more or less attention in subsequent messages. Users did not have to finish listening to the entire list before making their selection.

Stimuli: Because we were interested in examining which was worse, verifying whether SMS response templates matched the meaning of an intended reply, or deciphering the sometimes nonsensical misrecognitions of dictation, we decided to experimentally control both the SMS reply uttered by the user as well as the 4-best list generated by the system. However, all SMS replies and 4-best lists were derived from the logs of an actual SMS Reply interface which implemented the dictation and the voice search approaches (see Ju & Paek, 2009). For each course, 5 of the SMS replies were short (with 3 or fewer words) and 5 were long (with 4 to 7 words). The mean length of the replies was 3.5 words (17.3 chars). The order of the short and long replies was randomized.

We selected 4-best lists where the correct answer was in each of four possible positions (1-4) or All Wrong; that is, there were as many 4-best lists with the first choice correct as there were with the second choice correct, and so forth. We then randomly ordered the presentation of different 4-best lists. Although one might argue that the four positions are not equally likely and that the top item of a 4-best list is most often the correct answer, we decided to experimentally control the position for two reasons: first, our previous research (Ju & Paek, 2009) had already demonstrated the superiority of the voice search approach with respect to the top position (i.e., 1-best), and second, our experimental design sought to identify whether the voice search approach was more usable than the dictation approach even when the ASR accuracy of the two approaches was the same.

In the *dictation* condition, the correct answer was not always an exact copy of the reply in 0-2 of the 10 SMS messages. For instance, a correct dictation answer for Example (1) above was "*no I'm never with my GPS*". On the other hand, the *voice search* condition had more cases (2-4 messages) in which the correct answer was not an exact copy (e.g., "*no I have GPS*") due to the nature of the template approach. To some degree, this could be seen as handicapping the *voice search* condition, though the results did not reflect the disadvantage, as we discuss later.

Measures: Performance for both the driving task and the SMS reply tasks were recorded. For the driving task, we measured the numbers of collisions, speeding (exceeding 10 mph above the limit), traffic light and stop sign violations, and missed or incorrect turns. For the SMS reply task, we measured duration (i.e., time elapsed between the beginning of the 4-best list and when users ultimately provided their answer) and the number of times users correctly identified which of the 4 reply candidates contained the correct answer.

Originally, we had an independent rater verify the position of the correct answer in all 4-best lists, however, we considered that some participants might be choosing replies that are semantically sufficient, even if they are not exactly correct. For example, a 4-best list generated by the dictation approach for Example (1) had: "*One: no I'm never want my GPS. Two: no I'm never with my GPS. Three: no I'm never when my GPS. Or Four: no no I'm in my GPS.*" Although the rater identified the second reply as being "correct", a participant might view the first or third replies as sufficient. In order to avoid ambiguity about correctness, after the study, we showed the same 16 participants the SMS messages and replies as well as the 4-best lists they received during the study and asked them to select, for each SMS reply, any 4-best list items they felt sufficiently conveyed the same meaning, even if the items were ungrammatical. Participants were explicitly told that they could select multiple items from the 4-best list. We did not indicate which item they selected during the experiment and because this selection task occurred months after the experiment, it was unlikely that they would remember anyway. Participants were compensated with a cafeteria voucher.

In computing the number of "correct" answers, for each SMS reply, we counted an an-

swer to be correct if it was included among the participants' set of semantically sufficient 4-best list items. Hence, we calculated the number of correct items in a personalized fashion for every participant.

2.2 Results

We conducted a series of repeated measures ANOVAs on all driving task and SMS reply task measures. For the driving task, we did not find any statistically significant differences between the *voice search* and *dictation* conditions. In other words, we could not reject the null hypothesis that the two approaches were the same in terms of their influence on driving performance. However, for the SMS reply task, we did find a main effect for *SMS Reply Approach* ($F_{1,47} = 81.28$, $p < .001$, $\mu_{Dictation} = 2.13$ (.19), $\mu_{VoiceSearch} = .38$ (.10)). As shown in Figure 2, the average number of errors per driving course for *dictation* is roughly 6 times that for *voice search*. We also found a main effect for total duration ($F_{1,47} = 11.94$, $p < .01$, $\mu_{Dictation} = 113.75$ sec (3.54) or 11.4 sec/reply, $\mu_{VoiceSearch} = 125.32$ sec (3.37) or 12.5 sec/reply). We discuss our explanation for the shorter duration below. For both errors and duration, we did not find any interaction effects with *Driving Conditions*.

3 Discussion

We conducted a simulator study in order to examine which was worse while driving: verifying whether SMS response templates matched the meaning of an intended reply, or deciphering the sometimes nonsensical misrecognitions of dictation. Our results suggest that deciphering dictation results under the duress of driving leads to more errors. In conducting a post-hoc error analysis, we noticed that participants tended to err when the 4-best lists generated by the dictation approach contained phonetically similar candidate replies. Because it is not atypical for the dictation approach to have n-best list candidates differing from each other in this way, we recommend not utilizing this approach in speech-only user interfaces, unless the n-best list candidates can be made as distinct from each other as possible, phonetically, syntactically and most importantly, semantically. The voice search approach circumvents this problem in two ways: 1) templates were real responses and manually selected and cleaned up during the development phase so there were no grammatical mistakes, and 2) semantically redundant templates can be

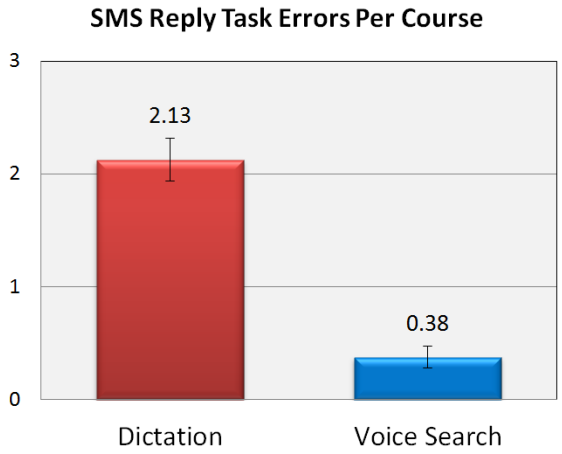

Figure 2. Mean number of errors for the dictation and voice search approaches. Error bars represent standard errors about the mean.

further discarded to only present the distinct concepts at the rendering time using the paraphrase detection algorithms reported in (Wu et al., 2010).

Given that users committed more errors in the *dictation* condition, we initially expected that *dictation* would exhibit higher duration than *voice search* since users might be spending more time figuring out the differences between the similar 4-best list candidates generated by the dictation approach. However, in our error analysis we observed that most likely users did not discover the misrecognitions, and prematurely selected a reply candidate, resulting in shorter durations. The slightly higher duration for the voice search approach does not constitute a problem if users are listening to all of their choices and correctly selecting their intended SMS reply. Note that the duration did not bring about any significant driving performance differences.

Although we did not find any significant driving performance differences, users experienced more difficulties confirming whether the dictation approach correctly interpreted their utterances than they did with the voice search approach. As such, if a user deems it absolutely necessary to respond to SMS messages while driving, our simulator study suggests that the most reliable (i.e., least error-prone) way to respond may just well be the voice search approach.

References

Distracted Driving Summit. 2009. Department of Transportation. Retrieved Dec. 1: http://www.rita.dot.gov/distracted_driving_summit

Y.C. Ju & T. Paek. 2009. A Voice Search Approach to Replying to SMS Messages in Automobiles. In *Proc. of Interspeech*.

A. Kun, T. Paek & Z. Medenica. 2007. The Effect of Speech Interface Accuracy on Driving Performance, In *Proc. of Interspeech*.

P. Natarajan, R. Prasad, R. Schwartz, & J. Makhoul. 2002. A Scalable Architecture for Directory Assistance Automation. In *Proc. of ICASSP*, pp. 21-24.

L. Nunes & M. Recarte. 2002. Cognitive Deamnds of Hands-Free-Phone Conversation While Driving. *Transportation Research Part F*, 5: 133-144.

N. Reed & R. Robbins. 2008. The Effect of Text Messaging on Driver Behaviour: A Simulator Study. Transport Research Lab Report, PPR 367.

D. Strayer & W. Johnston. 2001. Driven to Distraction: Dual-task Studies of Simulated Driving and Conversing on a Cellular Phone. *Psychological Science*, 12: 462-466.

R. Stross. 2008. "What carriers aren't eager to tell you about texting", New York Times, Dec. 26, 2008: http://www.nytimes.com/2008/12/28/business/28digi.html?_r=3

D. Yu, Y.C. Ju, Y.-Y. Wang, G. Zweig, & A. Acero. 2007. Automated Directory Assistance System: From Theory to Practice. In *Proc. of Interspeech*.

Wei Wu, Yun-Cheng Ju, Xiao Li, and Ye-Yi Wang, Paraphrase Detection on SMS Messages in Automobiles, in ICASSP, IEEE, March 2010