



Listening to Natural and Synthesized Speech while Driving: Effects on User Performance

OMER TSIMHONI AND PAUL GREEN

University of Michigan Transportation Research Institute, Ann Arbor, MI 48109-2150, USA

JENNIFER LAI

IBM Corporation/T.J. Watson Research Center, Hawthorne, NY 10598, USA

Received ; Revised March 23, 2001

Abstract. The effects of message type (navigation, E-mail, news story), voice type (text-to-speech, natural human speech), and earcon cueing (present, absent) on message comprehension and driving performance were examined. Twenty-four licensed drivers (12 under 30, 12 over 65, both equally divided by gender) participated in the experiment. They drove the UMTRI driving simulator on a road consisting of straight sections and constant radius curves, thus yielding two levels of low driving-workload. In addition, as a control condition, data were collected while participants were parked. In all conditions, participants were presented with three types of messages. Each message was immediately followed by a series of questions to assess comprehension. Navigation messages were about 4 seconds long (about 9 words). E-mail messages were about 40 seconds long (about 100 words) and news messages were about 80 seconds long (about 225 words). For all message types, comprehension of text-to-speech messages, as determined by accuracy of response to questions, and by subjective ratings, was significantly worse than comprehension of natural speech (79 versus 83 percent correct answers; 7.7/10 versus 8.6/10 subjective rating). Driving workload did not affect comprehension. Interestingly, neither the speech used (synthesized or natural) nor the message type (navigation, E-mail, news) had a significant effect on basic driving performance measured by the standard deviations of lateral lane position and steering wheel angle.

Keywords: speech synthesis, text-to-speech, comprehension, driving, driver distraction

Introduction

The term “mobile worker” no longer describes a transient state for a person, but represents a category of employees who work from a multitude of locations. Employees are being encouraged to “work smarter” since it often does not seem possible to work harder. As a result, system designers are stepping up to the challenge of giving users access to the same functionality in a mobile setting that they would have if seated in their offices.

Motor vehicle manufacturers and computer companies alike view the time workers spend in a vehicle as an opportunity to deliver E-mail and other individu-

alized messages. The in-vehicle information systems market is projected to exceed \$5 billion by the year 2003 (Bruno, 1999) as a result of a huge increase in the overall electronics content of motor vehicles (Richardson and Green, 2000). Increases will occur in both the feature content of conventional systems (entertainment and climate control) and in the addition of telematics applications (navigation, traffic information, as well as E-mail and Internet access). The changes in motor vehicles resulting from this increase should enhance the comfort and convenience of driving and expand the range of activities drivers can perform, possibly improving how efficiently people utilize their time. There is significant pressure to incorporate

all sorts of new applications because the technology is feasible. However, there are significant concerns that these systems could overwhelm drivers with information, especially if the information is presented visually in the vehicle, drawing the drivers' eyes away from the road (the eyes-off-the-road problem). In addition, these systems may create cognitive demands, causing the driver not to completely evaluate the current driving situation (the mind-off-the-road problem). Finally, some systems create immediate demands for attention, such as a ringing phone (the immediacy problem). (See Green (2000a) for additional information.)

One potential solution to the eyes-off-the-road problem is to present information auditorally. Auditory information, often speech, must be audible, intelligible, and comprehensible. Much of the prior automotive research on speech intelligibility has concerned listening to speech in trucks (e.g., Morrison and Casali, 1994, 1995). Intelligibility for speech messages in the context of truck cab noise, as determined by a Modified Rhyme Test (Logan et al., 1989), was similar to that for pink noise of the same signal-to-noise ratio. The intelligibility results obtained from participants could be predicted reasonably well from calculations using the Articulation Index (Kryter, 1972).

If information is presented auditorally, the immediacy problem can become even more significant than for visual presentations. One potential solution to the immediacy problem is to precede auditory messages with earcons (auditory cues), providing drivers with an opportunity to prepare for the following message (Belz et al., 1997).

How speech is presented has direct bearing on application usefulness, usability, and safety. Synthesized speech is generally not as well liked as human speech and takes longer for people to become accustomed to it (Francis and Nusbaum, 1999). However, given the difficulties and the high cost of delivering large amounts of dynamic data with pre-recorded human speech, for the near term many in-vehicle applications will use synthesized speech (also referred to as synthetic speech or text-to-speech) for the reading of E-mail messages or profile-specific news stories. So far, the impact of synthetic speech, relative to recorded human speech, on attention and driving performance has not been studied.

In addition to specific implementation issues, there is concern that auditory messages, by their nature, will be intrusive. When radios were first added to cars, they were heavily criticized as distracting to drivers. However, research suggests that it is not the listening

per se that disrupts driving (Brown, 1965), but rather the additional tasks such as putting in tapes or changing stations, that cause driving performance to deteriorate (Jaencke et al., 1994).

More recently, there has been considerable interest in another auditory source, cellular phones. Crash data from Japan (Green, 2000b) and other evidence (Goodman et al., 1997; sources available on the car-talk web site [<http://cartalk.cars.com/About/Drive-Now/scientific-evidence.html>]) suggest that answering the phone is the leading cause of phone-related crashes, followed by conversation and dialing. The greater the complexity of the phone conversation, the greater the interference with driving. Talking on a phone is quite different from talking to a passenger. Passengers, especially those in the front seat, often serve as co-drivers, searching for hazards and moderating the flow of conversation based on the driving situation. A caller knows nothing of the driving situation and behaves accordingly. In fact, if someone says something over the phone, it is rude not to respond. At the present time, cellular phone use while driving is banned in several countries and in a few cities in the U.S. There are no state laws banning cellular phone use while driving, but there are proposals for such laws (<http://www.ncsl.org/programs/esnr/cellphone.pdf>).

Data on the specific matter of interest here, comprehension of in-vehicle messages, is much more limited. Fleming et al. (1998) had 32 licensed drivers drive an instrumented car on an expressway. While driving, pre-recorded traffic messages containing six to 14 items were presented (e.g., "I-94 eastbound at Southfield freeway, continuing construction, right lane blocked, three mile backup"). Participants identified if the messages were relevant to a particular route and recalled them. Consistent with the limits of human short term memory, typically only four items were reliably recalled regardless of message length, with the road and crossroad being most common. Degradation in audio quality (not quantified) decreased recall and recognition performance. Drivers believed the simulated traffic message system was safe and easy to use while driving. However, listening to a message did increase speed variance, and subsequently responding to questions from an experimenter led to further degradation. Other driving performance measures were generally unaffected by drivers' listening to traffic messages. In general, as drivers become more loaded, they tend to slow down, and speed variability may increase. With further loading, lane variability

increases, and finally, so does the frequency of lane excursions.

Concurrent with the present experiment, Lee et al. (submitted) had 24 young people (ages 18–24) drive in a simulator while verbally interacting with a speech-based E-mail system. Their participants showed a 30 percent increase in response time (from 1010 ms to 1310 ms) to a periodically braking lead vehicle when interacting with the E-mail system. The drivers reported generally that the speech-based interaction caused a greater cognitive load for them, and they perceived it to be more distracting.

Given the lack of data on the effect of in-vehicle auditory messages on driving, the current experiment was conducted. (See Tsimhoni et al., 2000; for complete details.) In planning this experiment, considerable emphasis was placed on building upon the sponsor's prior research and replicating previous test conditions in a new context, resolving key product decisions the sponsor had (e.g., the merits of a particular text-to-speech implementation), and exploring the range of design options. Furthermore, the design was for a single experiment, not a series of experiments. Thus, for example, a variety of message types were explored rather than focussing on a single message type, namely E-mail. In this initial effort, three message types were examined, with the lengths examined being those most likely to occur naturally. Subsequent work should focus on E-mail, the most information-rich and personal of the message types of interest.

The present experiment was motivated by findings in Lai et al. (2000), which suggested that perception of synthesized speech requires greater cognitive resources than perception of human speech. Such demand for cognitive resources could divert attention away and impact performance from concurrent tasks negatively, such as driving. The underlying hypothesis of the present experiment was that an increase in cognitive workload would lead to a decrease in both the comprehension of messages and the driving performance. In this experiment, cognitive workload increased with the use of synthesized speech, the increased complexity of the messages presented to the driver, and the increased difficulty of the driving course. All of these factors were predicted to negatively impact comprehension and driving performance. Therefore, the overall goal of this experiment was to examine some implications of listening to text messages on driving safety and the comprehension of those messages while driving. The interaction of task performance with driving

workload, ranging from low to moderate, was also examined. Specifically:

1. How does listening to messages while driving affect comprehension of auditory messages as a function of voice type (synthesized versus natural speech), earcon cueing (present, absent), message type (navigation, E-mail, news story) and the visual demand of driving (parked, straight, curve)?
2. How does driving performance (lane variance, steering wheel angle variance) vary as a function of these factors? Steering related measures served as the focus because of their tight connection to safety and crash risk and the ease with which such measures could be collected. There are many candidate measures of driving safety (Green, 1993), but the focus of this experiment was on a limited set feasible within the bounds of this project.
3. What are the relative subjective preferences for synthesized and natural speech?
4. Are there differences between drivers associated with their age and gender?

Method

Overview

Participants drove a simulator on roads consisting of straight sections and constant radius curves while listening to three types of messages: navigation, E-mail, and news stories. The messages either were spoken by a text-to-speech generator or via pre-recorded human speech. Half of the navigation messages were preceded by an earcon (a 1.5-second duration chime), and the remainder were presented without earcons. After participants listened to each message, the car went into automatic driving mode and several multiple-choice questions were displayed on an in-vehicle display. The participants responded to the questions verbally by saying the number corresponding to the correct answer. The participants then rated their understanding of each message on a 10-point scale.

Experimental Design

All the major manipulations in this experiment were performed within subjects. Age and gender were between-subject factors. Participants were presented with messages of two voice types (synthesized, human), three driving workload levels (parked, straight,

curve), and four message types (navigation, navigation without earcon, E-mail, news). Since the same message could not be repeated more than once to each participant, different (but presumably equally difficult) messages were presented in each trial. Hence, each participant heard each message only once. To compensate for differences between messages, presentation order and voice type were counterbalanced across participants.

Messages were preceded by an earcon. To test the effect of earcon on performance, a set of navigation messages without earcons was added to the full-factorial design. Due to experiment time limitations, manipulation of earcon (absent, present) was only applied to navigation messages. The effect of earcon was treated as a fourth level of message type—navigation without earcon.

Test Participants

Twenty-four licensed drivers participated in this experiment in March 2000. There were 12 younger participants (21–28 years old, mean of 25) and 12 older (65–71 years old, mean of 68). In each age bracket there were six men and six women. The two age groups were chosen to represent two extreme segments of the driving population in terms of performance, with the older segment growing as a percentage of the population. In prior research of driving while using telematics (Green, 2001) task performance was very sensitive to the effect of age and factors associated with it.

Participants were recruited via an advertisement in the local newspaper and from the University of Michigan Transportation Research Institute (UMTRI) participant database. All participants were paid \$35 for their participation. Volunteers with hearing aids or other self-reported hearing difficulties were screened out of the experiment before recruitment. An audiometer (Maico Hearing Instruments model MA-27) was used to test hearing in both ears. All participants had hearing thresholds of 20 dB or better at frequencies between 500 Hz and 2000 Hz and adequate hearing at other frequencies. Subjectively, all participants reported they could easily hear and understand a test message that was played to them at the beginning of the experiment.

In addition, participants were tested for far visual acuity, near visual acuity, and color vision using a Titmus vision tester (Stereo Optical Vision model OV-7M). Twenty-three participants had far visual

acuity of 20/40 or better as required by Michigan State law. (One older participant had far visual acuity of 20/70, acceptable for daytime driving.) Five older participants had near vision (80 cm) acuity worse than 20/40. Subjectively, all participants reported they could easily see details on the simulator screen and read text from the in-vehicle LCD monitor.

While the average mileage reported by U.S. drivers is about 13,000 miles per year [<http://www.fhwa.dot.gov/ohim/hs97/nptsdata.htm>], the participants reported driving 3000 to 25,000 miles per year (mean of 11,600).

Less than half of the participants owned cellular phones. Previous exposure to computer-generated speech was rated on a 6-point scale (1—“never heard of it”, 3—“heard it once or twice”, 6—“work with it”). Except for two younger participants, no participants had been highly exposed to synthesized speech, most having heard it once or twice or less.

Test Materials and Equipment

Simulator. Participants drove the UMTRI Driver Interface Research Simulator (version 7.2.2), a medium-fidelity driving simulator based on a network of Macintosh computers (www.umich.edu/~driving/sim.html, Olson and Green, 1997). The simulator (Fig. 1) consists of a mockup of a car, a projection screen, a torque motor connected to the steering wheel, a sound system (to simulate wind noise as well as noise from the engine, drive train, and tires), a sub-bass sound system (to provide vibration), a computer system to project images of an instrument panel, and other hardware. The projection screen, offering a horizontal field of view of 33 degrees and a vertical field of view of 23 degrees, was 6 m (20 ft) in front of the driver, effectively at optical infinity.

Simulated Roads. The simulated roads were designed to impose two levels of constant workload by manipulating road curvature (straight sections and sharp curves of nine degrees of curvature). These two levels of driving workload were chosen based on a previous experiment (Tsimhoni and Green, 1999), in which the visual demand of curves in the simulator was quantified, and the relation between visual demand and the reciprocal of curve radius was found to be linear. They provided a reasonable range of visual demand that drivers would experience on real roads.

The simulated two-lane roads were 3.66 m (12 feet) wide. Oncoming vehicles were positioned statically in

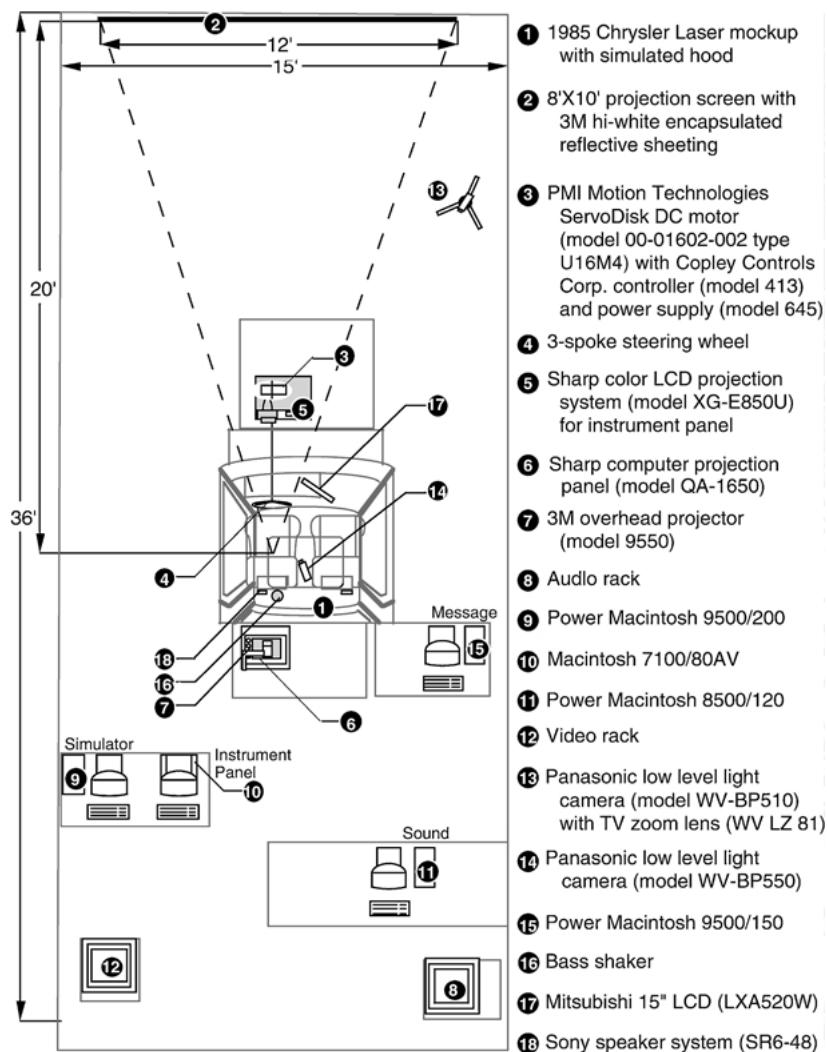


Figure 1. Plan view of UMTRI's Driver Interface Research simulator.

the left lane to provide a more realistic feeling of a busy road. A lead vehicle drove in front of the simulator vehicle at a fixed speed of 72.5 km/hr (45 mi/hr). A sample of the road is shown in Fig. 2.

Task. Every participant heard three types of messages in both synthesized and recorded human speech: navigation messages, E-mail messages, and news messages. Navigation messages were four seconds long (9 ± 2 words) and consisted of direction of turn, name of road to turn to, and distance, for example, “turn right in 1.5 miles onto Radiance Drive.” These three pieces of information appeared in random order to reduce practice effects. The participants were then presented with

two questions, which randomly asked for two of the three pieces of information.

E-mail messages were 40 seconds long (106 ± 7 words). (This length was selected based on an informal survey of 15 engineering students at the University of Michigan that showed mean length of the body of their E-mail messages to be 100 words.) The E-mail messages in this experiment concerned a variety of topics appropriate for corporate recipients. The participants were asked three questions about each message, the first of which was a general question about the purpose of the message or the tone in which it was spoken. The other two questions probed for comprehension of details contained in the message. Table 1 shows a



Figure 2. Simulator road scene (actual scene was in color).

sample of an E-mail message and its corresponding questions.

News messages were 80 seconds long (223 ± 8 words) and represented stories that might appear on an Internet news site. In fact, they were originally taken from websites such as *cnn.com* and shortened to fit the word limit. The participants were asked four questions about each news message, the first of which was a general question about the main topic of the message. Table 2 summarizes some message characteristics of the news and E-mail messages presented in the experiment. The Flesch–Kincaid readability statistics were computed using Microsoft Word 98 for the Mac.

The IBM Embedded ViaVoice TTS engine (http://www-4.ibm.com/software/speech/enterprise/ms_1.html) was used to generate the synthesized speech messages for the experiment. This engine is a high quality product, and a fair representation of the current state-of-the-art in mobile text-to-speech synthesis. The messages were recorded as WAV format sound files (11.025 kHz, 16 bit, Mono), converted to AIFF format,

and transferred to compact disc (CD) (one message per track) using a Macintosh 9500/120. The engine speed setting used was 0.6. Measurement of the rate of speech for the synthesized messages showed the number of words per minute to range from 147 to 158 depending on the message.

The human-speech messages were recorded by a semi-professional announcer using a Shure SM58 microphone and a Shure M267 microphone mixer. The recording software was Felt Tip Sound Studio (version 1.1.4 for Macintosh) in AIFF format (44.1 kHz sample rate, 16 bit, Mono). After recording, the files were re-sampled at 11.025 kHz to simulate lower sound quality and then transferred to a CD. The sound quality of the messages was reasonably similar to the expected quality of future message systems in production motor vehicles. Given the multitude of experimental factors in the current experiment, the sound quality of the messages was not manipulated. Sample sound files are available for download from www.umich.edu/~driving/sounds.

Table 1. Sample E-mail message and matching questions.

Colleagues,

Safe driving at our site is everyone's responsibility. In the past few months, security has received a number of complaints about speeding cars in the parking areas. Excessive speed, combined with decreased visibility and wet roadways during the winter months, increases the risk of accidents. In addition, a number of employees have been parking in the roadway next to the grassy islands, and therefore obstructing the flow of traffic. All employees are asked to observe the following guidelines:

- Observe the posted speed limits;
- Observe traffic signs and pavement markings;
- Respect the two-hour limit on visitor parking;
- Do not park in areas marked "No Parking" or "Fire Zone."

The tone of this message is best described as:

- 1) Friendly;
- 2) Informative;
- 3) Persuasive;
- 4) Encouraging;

In this message which factor was NOT attributed to increased risk of accidents:

- 1) Irresponsible drivers;
- 2) Excessive speed;
- 3) Low visibility;
- 4) Wet roads;

A custom SuperCard program (version 3.6, IncWell Digital Media Group), running on a Power Macintosh 9500/150, was used to play the messages from the CD player of the computer at exact segments of the roads and to display the questions on an in-vehicle display (Mitsubishi 15 in flat panel LCD monitor, model LXA520W). A finger-mounted switch was used by the participants to advance to the next question. The time from question presentation to switch-press for the next question was recorded to the nearest 33 ms. These times were written to a data file with their corresponding message number and question number.

Test Activities and Sequence

The participants began by completing a biographical form and a consent form. Next, they completed a vision test, a hearing test, and then moved to the driving simulator. The experimenter played a sample message and verified that the participant found the sound level comfortable and could understand what was spoken. Then the participant read a few sentences out loud from the in-vehicle display, as a proof that he or she would be able to read questions and answers.

The experiment started with a five-minute practice of driving the simulator on a road that consisted of

Table 2. Message characteristics.

	Message title	Words	Duration (s)		Flesch-Kincaid	
			Human speech	TTS	Ease 0 = easy 100 = hard	Grade
E-mail	Safe driving at our site	107	42	45	53	9.8
	Expanded curriculum	114	41	44	42	12.0
	Corporate travel guidelines	100	37	40	48	11.5
	Building a petaflop computer	106	44	45	29	12.0
	Visitors' schedule	112	40	43	65	7.5
	Task force	97	36	39	60	8.7
	Mean ± stdev	106 ± 7	40 ± 3	43 ± 3	50 ± 13	10.3 ± 1.9
News	Women in science	231	86	88	40	12.0
	Priceline	229	81	91	57	10.5
	Portugal	221	81	89	47	11.6
	Killer whale	215	76	80	56	9.9
	Women firefighters	212	75	78	42	12.0
	Frenchman	231	77	84	66	8.4
	Mean ± stdev	223 ± 8	79 ± 4	85 ± 5	51 ± 10	10.7 ± 1.4

straight sections and several curves. The participant then drove on a similar road for baseline driving-data collection. After the participant was comfortable with how to operate the driving simulator, a series of practice messages was presented in both the synthesized and human speech types with the simulator running but parked. After listening to each message, the participant was presented with several questions with multiple choice answers on the in-car display. Questions were answered by speaking the selection number (1, 2, 3 or 4) out loud. Each question was presented one at a time, with the participant controlling advancement to the next question by means of a finger switch fitted to the right index finger.

A similar practice session was conducted with the participant driving in both straight sections and curves. The participant was first instructed to "drive safely on the right lane, as you would normally." No specific instructions were given on how much attention to allocate to driving versus listening to the messages, but the experimenter made sure that the participant followed the instructions to drive safely. After the participant pulled away from the side of the road, the car automatically accelerated to a preset cruise control speed of 72.5 km/hr (45 mi/hr). Cruise-controlled driving was selected for this experiment to simulate a likely mode in which drivers might use an E-mail system on a highway and to eliminate the variability caused by speed reductions when the participant listened to the messages. In order for the participant to read and answer the questions, the car went into an "auto-pilot" mode after the message was played.

A fixed duration was given for answering all of the questions for each message. The duration chosen provided some time pressure, but all questions could be easily answered within the time limit. The participant was instructed to answer the questions as accurately as possible within the time limit. After answering all the questions, participants rated their understanding on a 10-point scale with 10 being "perfect understanding." After the fixed duration, the auto-pilot was turned off and the driver resumed control.

Evaluation Measures. Comprehension accuracy was measured by scoring the number of multiple choice questions that were answered correctly. Time on task was measured for each question from when the question was presented until the participant clicked the switch to advance to the next question. Driving performance measures included standard deviation of the steering

wheel angle, standard deviation of the lateral lane position in the lane, and the number of lane excursions. In general, as driving performance decreases, the values of these measures increase.

Random Guessing Baseline Experiment—Could Answers to Questions be Correctly Guessed from the Question?

In a small-scale experiment, all questions from the main experiment were presented to participants *without* the messages, to check whether the participants could guess correct answers based on information presented in the questions or in the multiple-choice answers. Ten participants, four men and six women, three from UMTRI and seven from IBM, participated. The participants were presented with a paper version of all the news and E-mail questions presented in the experiment (including practice questions) while sitting in a quiet office. The participants were instructed to answer the questions, in the order presented, based on their common sense, prior knowledge of the subject of matter, or hints and clues in the structure of the questions and answers. As in the main experiment, they said out loud the number of the answer that seem most likely to be correct.

As there were four possible answers in each multiple-choice question, random guessing would have produced approximately 25 percent accuracy. In fact, 23 percent of the questions were answered accurately (Fig. 3), confirming that there were no clues in the questions or answers to aid subjects in choosing a correct answer.

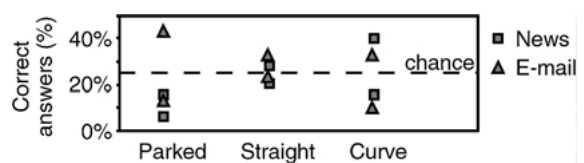


Figure 3. Correct answers in the guessing baseline experiment categorized by the workload level at which each message was presented in the main experiment. Each data point represents a single message.

Stationary Baseline Experiment—Were the Questions Associated with Each Type of Message Equally Difficult?

In the stationary baseline experiment, the main experiment was replicated without driving, that is, zero driving workload. Twelve young participants responded to messages from the main experiment, while seated in the simulator room. The apparatus for the stationary baseline was similar to that used for the main

driving experiment. The same computer program, CD player, speakers, and monitor for displaying questions were used. The purpose of this experiment was to complement the results of the main experiment by providing a baseline for the difficulty of messages without the effect of driving workload. In the main experiment, each message was always presented at the same workload level, and the combination of message and workload was not counterbalanced across participants.

Statistical Analyses. Repeated measures Analysis of Variance (ANOVA) tests were performed on all dependent variables. Tests of 3-level factors were followed by pairwise comparisons, with Bonferonni corrections made to the alpha levels to account for multiple comparisons.

Results

Overview

Consistent with the hypotheses mentioned above, participants were less likely to understand the content of the messages read by synthesized speech than the human speech, and they rated the synthesized speech more difficult to understand than the human speech. However, voice type had no effect on driving performance.

Contrary to expectations that comprehension accuracy would degrade as the messages got longer and more detailed, E-mail messages had the lowest comprehension scores even though they were shorter than news messages.

At first, driving workload seemed to improve comprehension accuracy rather than worsen it. However, careful comparison with the stationary baseline experiment revealed that the workload effect could be attributed to differences in the difficulty of messages presented in the different workload conditions.

No participants committed any lane excursions. Therefore, no results are reported for that measure of driving performance.

Comprehension Data

The comprehension level of synthesized speech messages (79 percent) was lower than the level of recorded human messages (83 percent) ($F(1, 20) = 5.5, p < 0.05$). This effect was consistent across message types (Fig. 4). Decreased comprehension level of synthesized

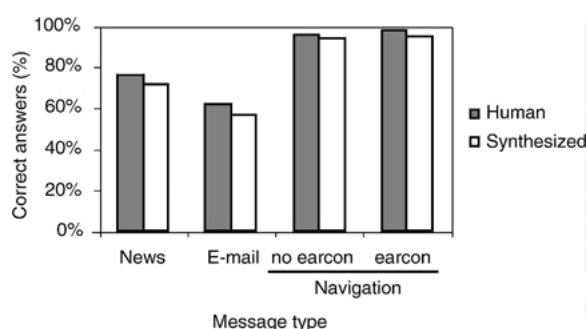


Figure 4. Comprehension by voice type and message type.

messages was in agreement with the findings of Lai et al. (2000).

There was a significant difference in comprehension of the different types of messages ($F(3, 60) = 174, p < 0.0001$). While almost all questions about navigation messages were answered correctly (95.5 ± 14 percent), only 74 ± 24 percent of questions about news messages were correct. Furthermore, only 59 ± 30 percent of questions about E-mail messages were answered correctly.

Whether a navigation message was preceded by an earcon did not significantly affect comprehension. The level of comprehension dropped from 96 percent with an earcon to 95 percent without it but the difference was not significant. It is not surprising that navigation messages had the highest levels of comprehension, as they were short and simple. The difference between news and E-mail messages, however, was not predicted. Although news messages were longer, and had more details than did E-mail messages, they resulted in higher comprehension levels. Perhaps the structure of the news messages was easier to follow and was more interesting. In contrast, the context of the E-mail messages might have taken more time to grasp, and the writing style was more "to the point."

Overall, younger participants performed better than older participants ($F(1, 20) = 13.8, p < 0.005$). While younger participants averaged 84 ± 25 percent, older participants averaged 77 ± 29 percent. There was no significant difference between the genders; nor was there a significant interaction between age and gender.

Due to limitations in the experiment design, driving workload in this experiment was confounded with specific messages (i.e., each message was always played at the same level of driving workload). Since comprehension level was very sensitive to which message was played, but not sensitive to the driving workload, it was impossible to measure the effect of workload

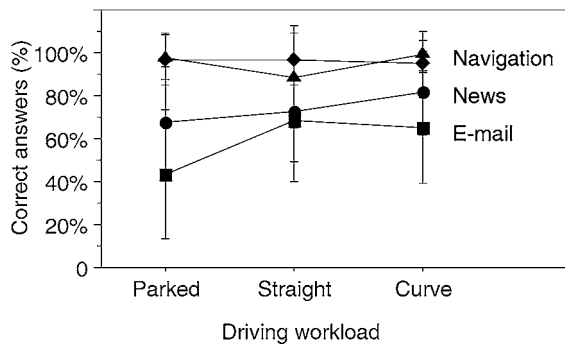


Figure 5. Comprehension by message type and driving workload—before adjustment.

directly. Figure 5 shows the results before adjusting for difficulty of individual messages. Comprehension levels while parked ($76 \pm 31\%$) were lower than while driving ($85 \pm 22\%$, $F(2, 40) = 7.7$, $p < 0.005$).

To investigate the effect of workload, a comparison was made between the results of the main experiment and the results from the stationary baseline experiment. Figure 6 compares comprehension of messages in the stationary baseline experiment and the main experiment. Each point in the figure represents the mean comprehension level for 12 young participants for one message in either of the experiments. While large differences in comprehension were found between the different messages, no consistent difference was found between the stationary baseline and the main experiment.

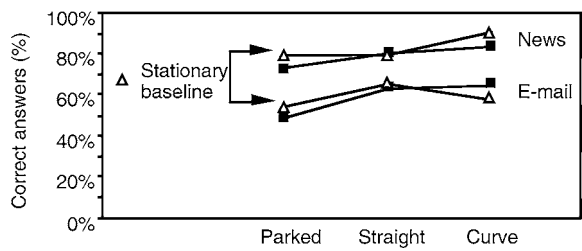


Figure 6. Comprehension of messages: stationary vs. driving. Each data point represents a single message.

In fact, the comprehension results of the baseline and the main experiment were extremely similar, which suggests that there was no driving workload effect on the comprehension of messages.

Time to Answer Questions. The time to answer questions incorrectly was longer than the time to answer correctly. Moreover, the time to answer questions was highly correlated with the percent of correct answers ($r = -0.58$, $p < 0.0001$). Figure 7 presents the distributions of response times categorized by message type and by whether the answer was correct or incorrect. Indeed, the medians of the distributions of correct answers (top row) were up to three seconds less than the medians of the distributions of incorrect answers (bottom row). Thus, on average, time to answer questions could be used as an indicator of the difficulty of understanding messages.

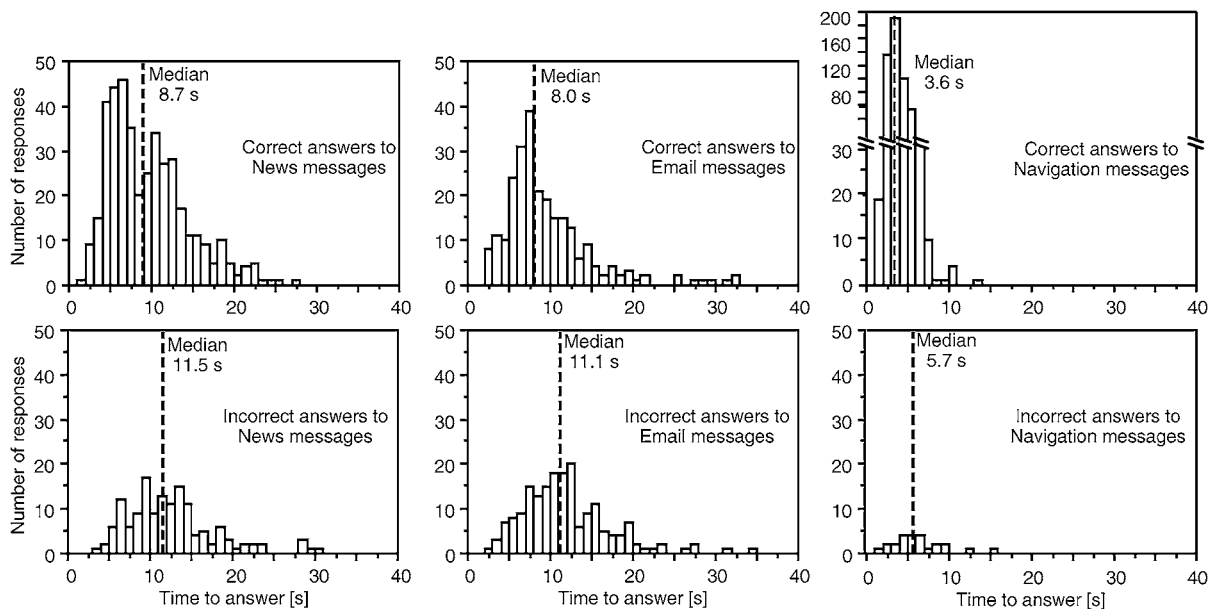


Figure 7. Time to answer questions by message type and whether the answer was correct.

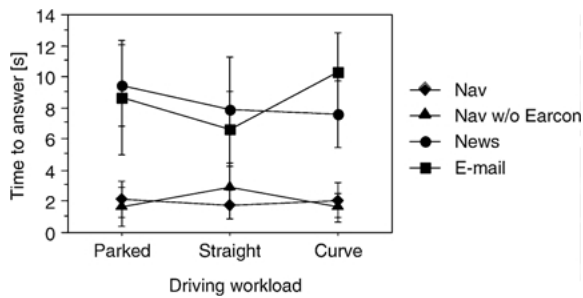


Figure 8. Response time to questions by message type and driving workload.

Overall, the effect of voice type on the time to answer questions was very small and only near significant ($F(1, 20) = 3.3, p < 0.10$). Questions that were presented in the human speech condition took 7.1 s to answer, while in the synthesized speech condition they took 7.3 s.

Participants responded to questions about navigation messages much faster than to questions about E-mail or news messages ($F(3, 60) = 359, p < 0.0001$) because the navigation questions were shorter and of a similar nature from message to message. While questions about navigation messages were answered within 4.0 s, the mean time to answer questions about E-mail messages (10.5 s) was not significantly different than the time to answer news messages (10.3 s) (Fig. 8). As in the comprehension results discussed earlier, the interaction between message type and driving workload was significant ($F(6, 120) = 35.7, p < 0.0001$).

Driving Performance. The standard deviation of lateral lane position and the standard deviation of steering wheel angle were calculated over 5-second inter-

Table 3. The relation between the variance in steering wheel angle and lateral lane position.

		SD of lateral lane position	
		Low	High
SD of steering wheel angle	Low	Better driving	Inattention
	High	Over correction	Worse driving

vals within each condition. In general, large standard deviations of both measures indicated worse driving performance. However, low standard deviation of steering wheel angle with high standard deviation of lateral position might suggest inattention to the road, as shown in Table 3.

When driving without listening to messages, the standard deviation of steering wheel angle for curves (0.05 radians) was three times as large as it was for straight sections (0.016 radians) (Fig. 9). Listening to messages in either synthetic or human voice did not affect these values. Lateral lane position showed a similar trend. The standard deviation of lateral lane position (0.11 m) was twice as large for curves than for straight sections (0.05 m) but it was not affected by listening to messages in either of the voice types. To put these findings into perspective, in a different study that involved reading from maps while driving the simulator (Tsimhoni et al., 1999), the above measures increased by 80% when reading from maps relative to just driving.

On straight sections, there were no differences in the standard deviation of steering wheel angle due to message type. In curves, however, the standard deviation of steering wheel angle in navigation messages (0.065 radians) was larger than in news messages (0.052 radians) and E-mail messages (0.045

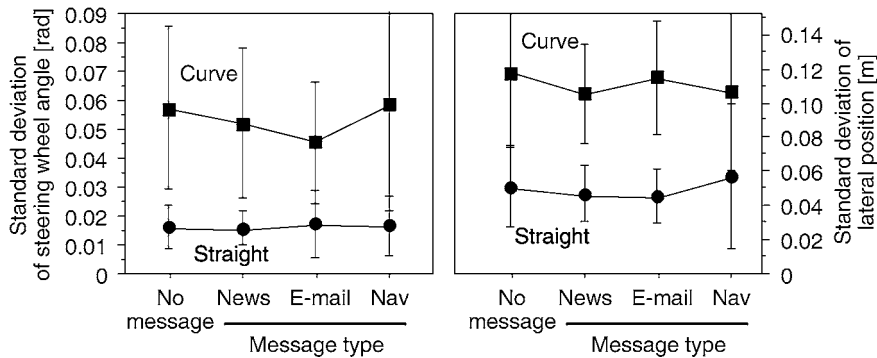


Figure 9. Lateral driving performance by message type and driving workload.

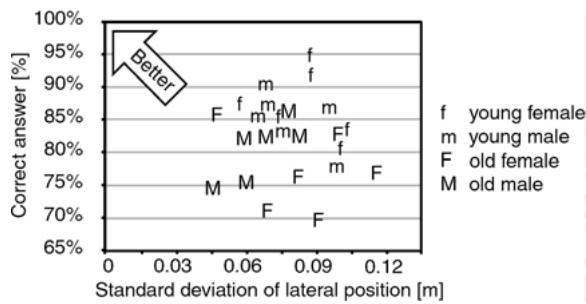


Figure 10. Tradeoff between comprehension and driving performance.

radians, ($F(2,40) = 10.5, p < 0.0005$). This might be attributed to the sampling method of the data, which only allowed two samples for the navigation messages in contrast to eight for the news and E-mail messages due to the shorter duration of the navigation messages. However, it is also possible that at the beginning of messages the standard deviations were larger and therefore driving performance suffered more in short messages. The standard deviation of lateral lane position did not differ significantly with types of messages.

Although the differences between age and gender groups were not significant at the commonly cited $p < 0.05$ level, some interesting trends were found that achieved engineering significance ($p < 0.10$). The standard deviation of steering wheel angle was slightly larger for older participants (0.031 vs. 0.038; ($F(1,20) = 2.98, p < 0.10$), and slightly larger for women (0.030 vs. 0.039; ($F(1,20) = 3.56, p < 0.10$). The standard deviation of lateral lane position had a different pattern. Older men had somewhat better driving performance than the other groups (0.21 vs. 0.28; ($F(1,20) = 3, p < 0.10$). In light of the hypothesized

interaction mentioned in Table 3, older participants most likely paid more attention to the driving task than did younger participants although they had all received the same instructions.

Between-Task Tradeoffs. Figure 10 presents comprehension as a function of driving performance by age and gender. The age effects discussed above clearly can be seen. On the vertical axis, older participants had lower comprehension levels than younger participants. On the horizontal axis, older men had the best driving performance. Although there is no absolute method to analyze the tradeoff between comprehension and driving, a qualitative assessment can be made. The top left corner of the figure represents overall better performance, while the bottom right corner represents overall worse performance. On the opposite diagonal, the top right corner of the figure represents better comprehension with worse driving performance, while the bottom left corner represents worse comprehension with better driving performance. Older participants had slightly better driving performance but slightly worse comprehension than younger participants.

Subjective Ratings of Understanding. Participants rated their understanding of each of the messages on a scale of 1 (understood nothing) to 10 (understood the entire message perfectly) (Fig. 11). Synthesized messages were rated 7.7 ± 1.4 , significantly lower (more difficult to understand) than human messages, which were rated 8.6 ± 1.3 ($F(1,20) = 36.7, p < 0.0001$). E-mail messages were rated lower than news messages (7.9 and 8.4, respectively; ($F(1,20) = 17.3, p < 0.0005$), a pattern similar to that of the objective measure of comprehension data (Fig. 4).

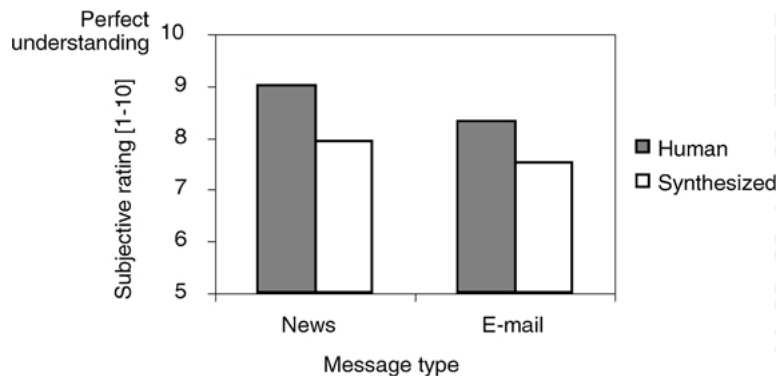


Figure 11. Subjective rating by voice type and message type.

The effect of curvature on subjective rating of understanding was similar to that found in the objective measures. Messages that were played while the vehicle was parked were rated most difficult to understand, followed by those for straight sections, and then curves ($F(2,40) = 13.2, p < 0.0001$).

Conclusions

Some people hold the opinion that telematics distract drivers to the extent that whoever uses them while driving is doomed to death. Others believe they can do anything they want while they drive with no adverse effects to their safety or the safety of others. The current experiment was a first step to provide quantifiable data on this controversy.

The underlying hypothesis of the experiment was that an increase in cognitive workload (when using synthesized speech, listening to longer messages, or driving at higher driving workload levels) would lead to a decrease in both the comprehension of messages and driving performance. As expected, comprehension of messages decreased when listening to text-to-speech. However, comprehension was not affected by driving workload and did not fully follow the prediction that long messages with more details would decrease comprehension. Driving performance was not affected at all by the type of voice or the type of message, but it was very sensitive to the level of driving workload.

Comprehension of text-to-speech messages as determined by the accuracy of response to questions, was significantly worse than comprehension of natural speech. This occurred for all message types. In addition, participants clearly rated the human speech as significantly more understandable than the synthesized speech, which confirms previous research in the area.

Perhaps one of the most interesting findings was that comprehension accuracy for synthesized speech did not degrade linearly with the length of the message, as it had in an earlier in-lab experiment (Lai et al., 2000). When driving, the navigational messages were easiest to understand, followed by the news stories, with E-mail messages coming in last even though the news stories were longer and contained a greater level of detail. This may be because drivers have a well-established model for driving and listening to news on the radio. News stories are general and usually do not affect or involve drivers directly. In contrast, E-mail messages are usually addressed directly at drivers,

involve their lives and often require them to take an action (i.e., attend a meeting, place a phone call, or get in touch with a person).

The three levels of driving workload did not affect comprehension. This result is probably due to demand-related motivation (the curves were more difficult, so participants concentrated more).

Earcon cueing did not significantly affect comprehension or driving performance. Due to the nature of the experiment, participants were expecting the messages and, therefore, they were not taken by surprise when a navigation message was not preceded by an earcon.

Interestingly, neither the voice type used (synthesized or human) nor the message type (navigation, E-mail, news) had a significant effect on the standard deviation of lateral lane position, a measure of driving performance. It was expected that the less understandable voice (synthesized) and more complex messages would degrade driving performance. Quite likely, there were no significant effects because the driving conditions explored were of low workload and the driving task loaded the visual and manual channels, while the listening task loaded the auditory and cognitive channels. The fact that basic driving performance in this study did not degrade is an important finding, especially in light of the magnitude of degradation when reading from in-vehicle displays. (For example, compare no lane excursions in this experiment to approximately 1 lane excursion per minute while reading from a map on the same road curvature.)

Older participants drove slightly better than younger participants but scored lower on message comprehension. Older participants were expected to perform slightly worse overall because of cognitive deterioration due to age, which normally leads to increasing limitations on processing capacity. These limitations affected their performance on the listening task but not their performance on the driving task. Experience and strategy may have caused this. Older participants were far more experienced drivers than younger participants. Their driving performance was therefore less likely to deteriorate. On the other hand, it is possible that older participants made a strategic decision not to let the listening task affect their driving. Analysis of steering wheel angle and lateral position suggests that younger participants paid less attention to the driving task than older participants although they had received similar instructions to give priority to the driving task. Furthermore, the authors acknowledge that these two

age groups differed not only in age but also in other factors such as driving experience, risk averseness, and allocation of priorities. Age and gender were used because they were consistently found to be relevant in previous research (Green, 2001), are two of the most commonly cited factors in crash statistics, and were likely to make the design more sensitive.

Findings from this experiment showed no significant change in driving performance while listening to messages produced with either synthesized or human speech. The authors caution that this result should not be generalized to higher levels of driving workload until they are examined. As one of the first few studies in this field, the authors wanted to avoid overwhelming subjects or creating an experiment that was so complex that it would not be manageable. Thus, driving conditions were uncomplicated: cruise control, a lead vehicle that never changed speed or braked, and very predictable two-lane roads with minimal traffic and no intersections or unexpected events. This is perhaps one of the most easy driving scenarios in which drivers may find themselves, but is a possible real-world scenario for drivers to listen to email and other information. Future work should involve measurements of the effect of listening to E-mail messages with greater levels of complexity in the driving conditions, such as random unexpected events (e.g., appearance of road debris on the road) and more traffic.

The authors believe that this is an important topic in need of further study. Future work not only needs to address higher levels of driving workload; it needs to address more complex levels of interaction with in-car speech systems. In this experiment, drivers merely listened to messages; with a real system drivers would be required to interact with the system to a greater extent. This would involve both trying to remember the correct phrasing or command for the system and keeping the context of the interaction in mind. These additional complexities would further tax the cognitive resources of the driver. Analyses of higher driving and interaction workloads are expected to reveal a negative impact on driving performance.

Acknowledgments

This work was conducted jointly by the University of Michigan Transportation Research Institute (UMTRI) and IBM Corporation, T.J. Watson Research Center. The authors would like to thank IBM for funding this research.

References

- Belz, S.M., Winters, J.J., Robinson, G.S., and Casali, J.G. (1997). Auditory icons: A new class of transportation subsystem (SAE paper 973185), Warrendale, PA: Society of Automotive Engineers.
- Brown, I.D. (1965). Effect of a car radio on driving in traffic. *Ergonomics*, 8(4):475–479.
- Bruno, A. (1999). Auto industry drives telematic services. *RCR Radio Communications Report*, 18(37):40–41.
- Fleming, J., Green, P., and Katz, S. (1998). *Driver performance and memory for traffic messages: Effects of the number of messages, audio quality, and relevance* (Technical Report UMTRI-98-22). Ann Arbor, MI: The University of Michigan Transportation Research Institute.
- Francis, A.L. and Nusbaum, H.C. (1999). Evaluating the quality of synthetic speech. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston, MA: Kluwer Academic Publishers, pp. 63–97.
- Goodman, M., Bents, F.D., Tijerina, L., Wierwille, W., Lerner, N., and Benel, D. (1997). *An investigation of the safety implications of wireless communication in vehicles* (Technical Report DOT HS 808 635). Washington, D.C.: U.S. Department of Transportation (<http://www.nhtsa.dot.gov/people/injury/research/wireless/>).
- Green, P. (1993). *Measures and methods used to assess the safety and usability of driver information systems* (Technical Report UMTRI-93-12). Ann Arbor, MI: The University of Michigan Transportation Research Institute (also published as FHWA-RD-94-088, McLean, VA: U.S. Department of Transportation, Federal Highway Administration, August, 1995).
- Green, P. (2000a). Dealing with potential distractions from driver information systems. (SAE paper 2000-01-C008) Paper presented at the *Convergence 2000 Conference*, Dearborn, Michigan, October 16–18, 2000.
- Green, P. (2000b). The human interface for ITS display and control systems: Developing international standards to promote safety and usability. Invited paper presented at the *International Workshop on ITS Human Interface in Japan*, Utsu, Japan, June 8, 2000.
- Green, P. (2001). Variations in task performance between younger and older drivers: UMTRI research on telematics. Paper presented at the *Association for the Advancement of Automotive Medicine Conference on Aging and Driving*, Southfield, Michigan, February 19, 20, 2001.
- Jaencke, L., Musial, F., Wogt, J., and Kalveram, K.T. (1994). Monitoring radio programs and time of day affect simulated car-driving performance. *Perceptual & Motor Skills*, 79(2):484–486.
- Kryter, K.D. (1972). Speech communication. In H.P. Van Cott and R.G. Kinkade (Eds.), *Human Engineering Guide to Equipment Design*, Washington, DC: US Government Printing Office, pp. 161–226.
- Lai, J., Wood, D., and Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech, *ACM SIGCHI CHI 2000 Proceedings*, pp. 321–328.
- Lee, J.D., Caven, B., Haake, S., and Brown, T.L. (submitted to Human Factors), Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers attention to the roadway. [<http://www-nrd.nhtsa.dot.gov/driver-distraction/PDF/27.PDF>].
- Logan, J.S., Greene, B.G., and Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86(2):566–581.

- Morrison, H.B. and Casali, J.G. (1994). Intelligibility of synthesized voice messages in commercial truck cab noise for normal-hearing and hearing-impaired listeners. *Proceedings of the 1994 Human Factors and Ergonomics Society 38th Annual Conference*. Santa Monica, CA: Human Factors and Ergonomics Society, pp. 801–805.
- Morrison, H.B. and Casali, J.G. (1995). Interior noise levels and intelligibility of synthesized speech messages in 1993-vintage commercial truck cabs *Proceedings of the 20th Annual National Hearing Conservation Conference III/XX*. Cincinnati, OH: The National Hearing Conservation Association, pp. 150–156.
- Olson, A. and Green, P. (1997). A description of the UMTRI driving simulator architecture and alternatives (Technical Report UMTRI-97-15). Ann Arbor, MI: The University of Michigan Transportation Research Institute.
- Richardson, B. and Green, P. (2000). Trends in North American intelligent transportation systems: A year 2000 appraisal (Technical Report 2000-9). Ann Arbor, MI: The University of Michigan Transportation Research Institute.
- Tsimhoni, O. and Green, P. (1999). *Visual Demand of Driving Curves Determined by Visual Occlusion*. Paper presented at the Vision in Vehicles 8 Conference. Boston, MA: August 22–25.
- Tsimhoni, O., Yoo, H., and Green, P. (1999). *Effects of workload and task complexity on driving and task performance for in-vehicle displays as assessed by visual occlusion* (Technical Report UMTRI-99-37). Ann Arbor, MI: University of Michigan Transportation Research Institute.
- Tsimhoni, O., Green, P., and Lai, J. (2000). Listening to synthetic and natural speech while driving: Effects on user performance (Technical Report UMTRI 2000-31). Ann Arbor, MI: The University of Michigan Transportation Research Institute.