Using Driver's Speech to Detect Cognitive Workload

Chip Wood (1), Kari Torkkola (1), Snehal Kundalkar (2)

(1) Motorola, Center for Human Interaction, Tempe, AZ, USA
chip.wood@motorola.com, kari.torkkola@motorola.com
(2) Arizona State University, Department of Computer Science, Tempe, AZ, USA
snehal_sk@hotmail.com

Abstract

In a recent driving simulator study [1], drivers were asked to engage in spontaneous conversations with the remote experimenter. While driving a rather complicated world, they used a hands-free cell phone. Each driver engaged in four “Neutral” and four “Intense” conversations of approximately three minutes each. The objective driving performance and the subjective “workload” showed significant differences between topic types. In the current study, an analysis of the speech during the spontaneous conversations was undertaken to see if there was any correlation between speech patterns and the conversation topic type. This paper discusses the successful results and their implications.

1. Introduction

Modern automobiles contain many infotainment devices designed for driver interaction. Navigation modules, entertainment devices, real-time information systems (such as stock prices or sports scores), and communication equipment are increasingly available for use by drivers. In addition to interacting with on-board systems, drivers are also choosing to carry in mobile devices such as cell phones to increase productivity while driving. Because technology is increasingly available to allow people to stay connected, informed, and entertained while in a vehicle, many drivers feel compelled to use these devices and services in order to multitask while driving.

This increased use of electronic devices along with typical personal tasks such as eating, shaving, putting on makeup, can cause the driver to become physically and/or cognitively overloaded and devote less than optimal attention to the driving task. This can increase risk of injury to the driver, passengers, surrounding traffic and nearby objects.

Detection of driver “workload” could be used in intelligent driver’s aid systems to manage electronic devices [1] or redirect the driver's attention to critical driving tasks [2]. Driver's advisor systems need to make real-time judgments about the state of the driver to be able to present appropriate information (warnings/advice) in an appropriate manner at appropriate times. The workload of the driver is one such key judgment. Identifying increased workload and the consequences of such workload on performance will help to target the appropriate times for managing access.

2. Remote Conversations Increase Driver Workload

Research [3], [4] suggests that cell phone conversations can affect driving performance. Other research [5], [6] show other cognitive distracters produce deficits in driving performance. For example, using a car-following paradigm, when the lead vehicle suddenly decelerates, drivers performing a cognitive distraction task take longer to release the gas pedal.

Increases in driver workload, possibly degrading driving performance, from cell phone conversations may arise from three sources:

- Conversing with a remote talker
- Processing information
- Timely responding to a query from the remote talker

To display the effects of higher workload from a cell phone conversation, it is not sufficient that the driver be simply listening. For example, evidence suggests that driver response times to unexpected events do not increase when the drivers are listening to the radio or to a book on tape [3]. There is also evidence that passengers conversing with the driver modulate the pace of the conversation in response to events in the traffic environment [7]. However, the remote talker will be unaware of the driving conditions and cannot adjust the conversation accordingly. To maintain the cell phone conversation, the driver must continue processing
information from the remote talker, even during critical driving maneuvers. The driver needs to be actively engaged in the conversation, processing information from the remote talker, and then producing verbal responses [3].

2.1. Estimating Driver’s Workload

There are two approaches to estimating driver workload. One is to assess the driver's real time behavior using measures such as avoidance responses to unexpected hazards. However, it is difficult to find reliable and valid workload measures based on driver behavior. Another approach is to evaluate the cockpit, road, and traffic situation confronting the driver. For this latter approach, the assumption is that these environmental variables directly influence the driver's workload.

In order to accurately time the delivery of an alert, it would be better to use the driver's behavior to infer the severity of demands on that driver than to use the complexity of the environment to draw these inferences. Given the fact that there are significant individual differences in distractibility [8], some drivers will be less capable than others at driving in complex traffic environments while talking on a cell phone. Using behavior to estimate demands imposed on a driver allows the alert to be targeted more directly at drivers who are most sensitive to these demands, and the alert can be activated only when those drivers need it.

2.2. Vehicle Control Measures of Workload

While using behavioral measures to estimate demands on the driver sounds like an ideal method, one challenge is to find measures that accurately reflect that workload. One type of workload measure is the driver's control over the vehicle heading, because a driver under high workload may not perform timely steering wheel corrections to maintain an agreement between the heading of the vehicle and the road. This is relatively easy and inexpensive to obtain, being based on steering wheel movements and lane positioning.

A previous study evaluated workload measures derived from vehicle control metrics under two different kinds of distraction, visual and auditory, and under two different road configurations, curvy and straight [4]. The distraction task that most closely resembled a cell phone call, the auditory task, produced no significant effects on any of the performance measures. This result was in line with outcomes from previous studies on the effects of cell phone calls on driver performance [9]; [10].

Thus, the workload measures in this study were not reliable enough to use as indicators

2.3. Hazard Avoidance Measures of Workload

Another behavioral approach is to measure driver responses to unexpected hazards that occasionally occur in the course of driving. A common task used in evaluating this approach is car following [11]; [6]. In this task, a driver follows another vehicle and has to decelerate when that lead vehicle decelerates. Typical performance measures for this task include the time to lift the foot from the gas pedal, the time to press the brake, the rate of deceleration and the minimum speed.

The studies with this type of task have demonstrated its sensitivity to auditory/cognitive distracters. Drivers generally take longer to initiate the deceleration when they are simultaneously performing an auditory/cognitive distraction task. Such tasks include making logical judgments about sentences [5] and talking on a cell phone [12]. Furthermore, the effects of a cell phone conversation on the drivers' response times are greater than those of other baseline tasks, such as listening to the radio. This result demonstrates that the genesis of the delayed response is the active engagement of the driver in the conversation [3].

2.4. Proactive and Reactive Workload Measures

While the time to react to unexpected events appears to be a reliable indicator of increased driver workload, it is also a reactive measure. That is, a greater delay in a driver's response to avoid a potential hazard is a single event that directly and immediately increases the risk to the driver. If the driver does not make the appropriate response quickly enough, then there will be a collision. However, the fact that the driver is under increased workload is not evident until there is a delay in the response, at which point it is likely to be too late to effectively take any action, intelligent or otherwise.

By contrast, variability in lane keeping is a proactive measure. The tendency to wander in the lane is an ongoing behavior that does not necessarily directly translate into immediate risk to the driver. However, if such wandering were a valid measure of increased workload, then it could predict the likelihood that the driver will be in danger in the near future. In this case, the evidence of increased workload is detected before the events that lead to an accident.

Thus, the two measures of workload, vehicle control and driver response to unexpected events, are each problematic as indicators of increased driver workload.

If, however, one could reliably estimate driver workload from non-driving behavior it could lead to a proactive measure.
2.5. Conversation Levels Affect Driver Workload Differently

During a full-scale simulator, car following study [1], 20 drivers were asked to engage in spontaneous conversations with the remote experimenter. The drivers used a hands-free cell phone. The experiment had two cognitive levels of conversation types, 1) neutral and 2) intense. Each driver engaged in four different “neutral” and four different “intense” conversations of approximately three minutes each during their drives. The total conversation time was 20 Drivers x 4 Conversations x 3 minutes = 480 min or 8 hours. Both driving performance and subjective “workload” burden were measured. The performance was significantly degraded and “workload” was significantly higher during the intense vs. the neutral conversations, indicating different cognitive workloads.

With the above evidence that different levels of conversation affect both the driver’s driving performance and his subjective “workload”, the question becomes “Can we find variables in the driver’s conversational speech that would indicate this difference of cognitive workload and be useful as a proactive measure for real-time feedback during cell phone calls?” An analysis of the spontaneous conversations was undertaken to find statistical indicators for the conversation type (neutral/intense) directly from the speech signal. The statistical variables studied were pitch and intensity of driver's speech, and distribution of durations of pause and speech segments within driver's turn in the dialogue between the driver and the experimenter. Standard deviations of the same four indicators were also calculated.

3. Experimental setup

3.1. Driving Simulator

Using scenario driven research, a Driver Advocate™ (DA) [2] system is being designed to advise the driver about potentially unsafe situations based on information from environmental sensors [13] DA is an intelligent, dynamic system that monitors, senses, prioritizes, personalizes, and sends alerts to the driver appropriate to the moment. This has the potential to sharply decrease driver distraction and inattention.

To support the realization of DA, a DA Authoring Tool (DAAT) has been developed to coordinate with a DriveSafety driving simulator and allow the merging of the simulated driving performance, the environmental sensors, and the intelligent use of audio, visual, and tactile feedback to alert the driver to potential danger and unsafe driving behavior.

The simulator consists of a complete (minus drive train) 1997 Saturn automobile positioned within life-size projected 3-channel front & 1-channel rear video screens (Figure 1). It is equipped with high fidelity audio and sub-woofers, full force feedback in steering wheel and pedals, and climate controlled cab, but no hydraulic chassis motion. All driver controls such as steering wheel, brake, gas pedal are monitored and affect the motion through the virtual world in real-time. Various hydraulics, air pressure, and motors provide realistic force feed back to driver controls to mimic actual driving. The simulator setup has been enhanced by adding several video cameras, microphones, and eye tracking infrared sensors to record all driver actions during the drive.

The simulator hardware and software are combined with Motorola proprietary experimental capabilities that allow complex intelligent experimental capability through the DA Authoring Tool and complex interaction with the virtual simulation environment through the Motorola DASP interface which is a Motorola designed and supported protocol.

![Figure 1: The driving simulator](image)

3.2. Data Collection

The collection system captures data from three different applications- 1) Driver Advocate, 2) the DriveSafety simulator, and 3) the SeeingMachines FaceLab eye/head tracker. These databases collect approximately 425 variables either at 60 frames/sec or as event markers. In addition, cabin audio and video of four different views of the driver and environment is digitally captured in MPEG2. All this data and audio/video is time synced, annotated, and analyzed by Motorola proprietary software. The combined databases
and video produce approximately 400Mb of data for each 10 minutes of drive time.

The data collected contains almost the entire scope of the driving virtual world – the auto, the driver and his performance, the environment, and traffic. This allows a complete analysis of virtually everything that occurs during a drive, especially the driver’s performance given a specific set of traffic, road, and environmental conditions.

These objective measures do not however allow for any assessment of the driver’s cognitive load or intentions.

3.3. Experiment Design

The simulator authoring tool, HyperDrive, was used to create the driving scenario for the experiment. The drive simulated a square with curved corners, 2800 meters on a side. The roadway was a typical suburban 2-lane with 1-lane each way. All drives used daytime dry pavement driving conditions with good visibility. Drivers drove five legs around for a total of 14 km. (Fig 2).

![Fig 2 Driving world](image)

The drivers were instructed to drive around the track, maintaining a speed of between 45 and 65 mph (72.4 and 104.6 kph). The driver’s vehicle started at the point marked “Start” in Fig 2. As the subject vehicle rounded the curve to start Leg 1, another vehicle, the “lead vehicle”, pulled onto the track in front of it. This vehicle was programmed to travel at a speed that allowed it to maintain a distance of 2 seconds with the subject vehicle. It maintained this distance as long as the subject vehicle was traveling at least 45 mph. Otherwise, it traveled at 45 mph.

Occasionally the lead vehicle would decelerate. During these events, which are shown as red squares in Fig 2, the lead vehicle no longer maintained a 2-second distance with the subject vehicle. Instead, it decelerated at a rate of 10 m/s^2 until it reached a speed of 25 mph. It then accelerated again until it was traveling at least 45 mph, at which point it once again automatically maintained the 2-second distance with the subject vehicle. The total time between the start of the deceleration until the start of the acceleration was 7 seconds.

After two orientation and practice drives, we collected data while drivers drove about 15 minutes in the simulated world. Drivers were instructed to follow all normal traffic laws, maintain the vehicle close to the speed limits (55 mph, 88.5 kph), and to not pass the lead car.

3.4. Subjects

There were 20 subjects, 10 males and 10 females. They had a mean age of 32.7 years (s.d. 12.3) and a median age of 27 years. The minimum age was 20 years and the maximum was 65 years. The distribution of ages was slightly skewed toward the minimum age, with 12 of the subjects being less than 30.

The subjects had been licensed an average of 16.3 years (s.d. 12.5). The median number of years with a license was 11, the minimum was 3 and the maximum was 50. Most of the subjects, 13, had their license less than 13 years.

Eight of the subjects reported driving less than 200 miles per week, 10 reported driving at least 200, and 2 subjects did not respond to this question. Fifteen subjects reported having at most 1 accident, with the remaining 5 reporting 2 to 4 accidents. Twelve reported having had at most 1 moving violation, and the remaining 8 reported receiving between 2 and 4 violations.

Fourteen of the subjects reported getting at least 7 hours of sleep the night before and three reported getting less than 6 hours. Six subjects reported getting 1 or 2 hours less sleep than usual the previous night. Twelve subjects reported being moderately or very rested and 3 reported being moderately tired.
3.5. Conversations

Periodically, as the subjects drove around the track, they engaged in a conversation over a simulated hands-free mobile phone in the vehicle. The other participant in this conversation was the lead experimenter sitting in the control room of the simulator lab. There were two such conversations during the course of each trip, an intense conversation and a neutral conversation. There was also a control condition during which there was no conversation. The locations at which these conversations started and ended are shown in Figure 2.

There were two types of conversations: “intense” and “neutral”. The topics for intense conversations focused on controversial political and social issues, whereas the topics for neutral conversations focused on preferences for leisure activities and technology (Table 1).

<table>
<thead>
<tr>
<th>Neutral</th>
<th>Intense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars</td>
<td>The Death Penalty</td>
</tr>
<tr>
<td>Computers</td>
<td>The Environment</td>
</tr>
<tr>
<td>Movies</td>
<td>Terrorism</td>
</tr>
<tr>
<td>Television</td>
<td>Taxation</td>
</tr>
<tr>
<td>Reading</td>
<td>The Middle East</td>
</tr>
<tr>
<td>Music</td>
<td>Conflict</td>
</tr>
<tr>
<td>Sports</td>
<td>Education</td>
</tr>
<tr>
<td>Investment</td>
<td>Gun Control</td>
</tr>
</tbody>
</table>

Table 1. Conversation topics.

The conversations began with a general question (e.g., "What do you think of the “topic”?"). When the subject had finished his or her answer, the experimenter continued the conversation with a series of non-directive questions derived from the "Eliza" simulation [14] (Table 2).

<table>
<thead>
<tr>
<th>What is it you really want to know?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are such questions much on your mind?</td>
</tr>
<tr>
<td>What answer would please you most?</td>
</tr>
<tr>
<td>What do you think?</td>
</tr>
<tr>
<td>What comes to mind when you ask that?</td>
</tr>
<tr>
<td>Have you asked such questions before?</td>
</tr>
<tr>
<td>Have you asked anyone else?</td>
</tr>
<tr>
<td>Because</td>
</tr>
<tr>
<td>Is that the real reason?</td>
</tr>
<tr>
<td>Don't any other reasons come to mind?</td>
</tr>
<tr>
<td>Does that seem to explain anything else?</td>
</tr>
<tr>
<td>What other reasons might there be?</td>
</tr>
<tr>
<td>Perhaps</td>
</tr>
<tr>
<td>You don't seem quite certain.</td>
</tr>
<tr>
<td>Why the uncertain tone?</td>
</tr>
<tr>
<td>Can't you be more positive?</td>
</tr>
<tr>
<td>You aren't sure?</td>
</tr>
<tr>
<td>Don't you know?</td>
</tr>
</tbody>
</table>

Table 2. Examples of questions from "Eliza".

3.6. Topic Choices

Before the drives began the drivers were asked to rank which 4 topics from each category they preferred, either because of interest, knowledge, or to avoid personally sensitive ones. The conversation topics were randomly selected from those ranked by any particular driver.

Among the intense topics for the conversations, most subjects, 80%, ranked education among their top choices. Half ranked guns and half ranked terror among their top choices, 40% ranked cloning and 40% ranked taxation highly, 35% gave high rankings to the environment, 30% to the death penalty, 30% to the Middle East, and 20% to euthanasia.

Among the neutral topics, the most highly ranked topics were music (75%), movies (60%), cars (45%), computers (45%) and investing (45%). Reading, sports and television were chosen by 35% of the subjects.
4. Earlier study Results

4.1. Driving Performance

The analysis of car following focused on the driver's performance during the course of the deceleration events. The time during each event was broken down into 1-second bins, and average values for the performance variables were computed for each bin. This analysis was performed for 5 variables: 1) forward velocity, 2) gas pedal pressure, 3) brake pressure, 4) headway distance, and 5) temporal headway.

All 5 variables showed statistical differences between the driving performances during intense vs. neutral conversations.

4.2. Subjective Workload

The results of NASA TLX data were analyzed to assess the impact of the conversation type on the subjects' subjective workload. The hypothesis was that intense conversations would produce higher subjective workloads than neutral conversations. An ANOVA was performed for each of the NASA TLX items: 1) Mental Demand, 2) Physical Demand, 3) Temporal Demand, 4) Performance, 5) Effort, and 6) Frustration. The results showed a significant effect of Conversation Type on subjective judgments of mental demand, temporal demand, performance, and frustration. There was also a trend toward an effect of Conversation Type on subjective judgments of effort. The overall result summarizing all of these effects is that drivers indicated that they perceived lower demand, higher performance and lower frustration when they engaged in neutral conversations, as opposed to intense conversations.

5. Applying Data Mining for Workload Detection

The research question to be answered is as follows. Can we find indicators in the conversational speech of the driver that would point out the degree as to how much of the driver's workload is in the conversation, how much in the driving? Since driver's workload cannot be accessed directly, but the speech patterns are available, the aim of the project is to find statistical indicators for the workload directly from the speech signal. The variables used are averages and standard deviations of pitch, intensity, durations of pauses, and durations of continuous speech periods.

4.1. Data Preparation

The audio data was extracted from the driving database which had both the driving video and the audio data. The audio data was stored in 16 bit, at 11025 Hz, .wav format to enable further and flexible speech processing and analysis.

4.1. Speaker Verification

The audio track of the experiment is a conversation between the driver and the experimenter. The simulator cockpit microphones picked up both the live driver's voice, and because a hands free speaker phone was used, also the remote experimenter's voice through the speakers. Since we are only interested in the driver's voice, an automatic method of extracting only his voice is required. In a unique approach, a speaker verification system was used to automatically differentiate the experimenter's voice from the driver's. The system we used, CipherVox, is a speaker verification system developed at Motorola. It is based on the Motorola Polynomial Classifier (MPC) [15].

A set of utterances was collected manually from the experimenter's speech, and was used to generate the experimenter's voice model used for training CipherVox. This voice model was used in all subsequent verification attempts. Any new exemplar is compared to the experimenter's threshold to generate a Boolean verification decision. The verification scenario used in this project was text independent verification which allows any phrase, or set of phrases, to be used for the verification utterances. No specific vocabulary is required. Technologies that support this type of verification do not require a distinct verification prompt. Hence, they can be used passively to verify individuals based on any speech that is available to the application.

The experimenter's voice model generated from the enrollment process was used as input for the verification process along with the speech data from any given conversation. The verification function returned scores, which were stored in a text file. Verification was done at a rate of 60 Hz using a speech window size of one second.

Our first attempt was to train the verification system only with the experimenter's voice and then use the resulting scores to determine locations of driver's speech, but that turned out not to be reliable enough. Thus we had to train a verification system separately for the experimenter, the driver, and the background driving noise. The final score for the driver's voice was then determined by the following equation,
$S_F = 2S_D - S_E - S_N$, 

where $S_D$, $S_E$, and $S_N$ denote the individual verification scores for the driver, the experimenter, and the background noise. Figure 3 plots an example of the resulting verification score. This measure turned out to be a good indicator for the presence of driver's speech once the threshold was manually determined.

## 5. Results

Given the locations of driver's speech segments within conversations, the averages of the indicators along with their standard deviations are calculated. These values are compared between two scenarios where the driver is involved in a neutral conversation or in an intense conversation.

From the results, it was found out later that an automatic gain control was used during the data collection process. This rendered the intensity indicator useless. The other indicators produced results listed in Table 3.

### 5.1. Observations

These results show that the standard deviation of pitch over the driver's speech segments is always larger in an intense conversation as compared to the neutral conversation. Pitch average is lower in intense conversation as compared to neutral conversation but by an insignificant amount. Speech segment duration did not show any significant differences, in neither the average, nor the standard deviation. However, durations of pauses showed interesting differences. Although average durations of pauses are similar in both conditions, the pause durations are more variable in intense conversations. As we measured the durations of 1839 pause segments in intense conversations, and 1741 pause segments in neutral conversations, the difference between the two pause duration standard deviations is significant beyond all confidence levels using the F-test. Same applies also to the difference in pitch standard deviations in the two conditions.

The above indicates that, if not a reliable indicator, at least further evidence for increase in driver's cognitive workload could be gathered by observing the variance of the pitch and the variance of pause segment durations in driver's speech. This evidence could be combined with other possible evidence of the workload, such as driving performance.

<table>
<thead>
<tr>
<th></th>
<th>Neutral Conversation</th>
<th>Intense Conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Average (Hz)</td>
<td>153.8</td>
<td>150.2</td>
</tr>
<tr>
<td>Pitch STD (Hz)</td>
<td>30.6</td>
<td>36.0</td>
</tr>
<tr>
<td>Pause segment duration average (sec)</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>Pause segment duration STD (sec)</td>
<td>1.30</td>
<td>1.67</td>
</tr>
<tr>
<td>Speech segment duration average (sec)</td>
<td>1.691</td>
<td>1.686</td>
</tr>
<tr>
<td>Speech segment duration STD (sec)</td>
<td>1.239</td>
<td>1.202</td>
</tr>
</tbody>
</table>

Table 3: Averages and standard deviations of pitch, durations of pauses, and durations of continuous segments of driver’s speech both during conversations on neutral topics and conversations on intense topics.

## 6. Conclusions

Investigations of the driver’s spontaneous speech patterns showed that the “Intense” conversation could be differentiated from the “Neutral” conversation by speech parameters only. Drivers showed differences in driving performance and subjective workload between the conversation types in an earlier experiment [1]. If these two measures do indeed indicate the degree of cognitive workload on the driver, then examining the driver’s speech patterns while in a cell phone conversation could be a proactive workload measure. Since the driver must speak during the call, this research shows a possible new
tool in the arsenal of workload measures especially while driving.

7. References


