



Getting Start with UTDrive: Driver-Behavior Modeling and Assessment of Distraction for In-Vehicle Speech Systems

*Pongtep Angkitittrakul, DongGu Kwak, SangJo Choi,
JeongHee Kim, Anh PhucPhan, Amardeep Sathyanarayana, John H.L. Hansen*

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, TX

{angkitit,dgk061000,sjc063000,jxk069000,app04100,axs063000,john.hansen}@utdallas.edu

Abstract

This paper describes our first step for advances in human-machine interactive systems for in-vehicle environments of the UTDrive project. UTDrive is part of an on-going international collaboration to collect and research rich multi-modal data recorded for modeling behavior while the driver is interacting with speech-activated systems or performing other secondary tasks. A simultaneous second goal is to better understand speech characteristics of the driver undergoing additional cognitive load since dialog systems are generally not formulated for high task-stress environment (e.g., driving a vehicle). The corpus consists of audio, video, brake/gas pedal pressure, forward distance, GPS information, and CAN-Bus information. The resulting corpus, analysis, and modeling will contribute to more effective speech systems which are able to sense driver cognitive distraction/stress and adapt itself to the driver's cognitive capacity and driving situations for improved safety while driving.

Index Terms: in-vehicle speech system, driver distraction, multi-modal resource, driver behavior modeling, safety driving

1. Introduction

There has been significant interest in development of effective human-machine interactive systems in diverse environmental conditions. One application which has received much attention is speech interactive systems in the car to allow the driver to stay focused on the road. Several studies have shown that drivers can achieve better and safer driving performance while using speech interactive systems to operate an in-vehicle system compared to manual interfaces [2, 5]. Although better interfaces can be incorporated, operating a speech interactive system will still divert a driver's attention from the primary driving task with varying degrees of distraction. Ideally, drivers should pay primary attention to driving, rather than any secondary tasks. With current life styles and advanced in-vehicle technology, it is inevitable that drivers will perform secondary tasks, or operate driver assistance and entertainment systems while driving. In general, the common tasks of operating the speech interactive systems in a driving environment includes (1) cell-phone dialing, (2) navigation/destination interaction, (3) e-mail processing, (4) music retrieval, and (5) generic command and control or in-vehicle telematics system. If such secondary tasks or distractions lie within the limit of the amount of spare cognitive load for the driver, he or she can still focus on driving. There-

fore, the design of safe speech interactive systems for in-vehicle environments should take into account factors from the driver's cognitive capacity, driving skills, and their degree of proficiency for the cognitive application load. With knowledge of such factors, an effective driver behavior model with real-time driving information, can be integrated into a smart vehicle to support or control driver assistance systems to manage driver distractions (e.g., suspend applications in a situation of heavy driving workload).

Another aspect presents in a car environment is the a variety of background noises that effect the quality of the input acoustic signal for the speech interface. More importantly, drivers have to modify their vocal effort to overcome perceived noise levels, namely the Lombard effect [9]. Such effects on speech production (e.g., speech under stress) can degrade the performance of automatic speech recognition (ASR) system more than the ambient noise itself [6]. At a higher level, interacting with an ASR system when focused on driving may result in a speaker missing audio prompts, using incomplete grammar, adding extra pauses or fillers, or extended time delays in a dialog system. Desirable dialog management should be able to employ multi-modal information to handle errors and adapt its context depending on the driving situations.

Building effective driver behavior recognition frameworks requires a thorough understanding of human behavior and the construction of a mathematical model capable of both explaining and predicting the drivers' behavioral characteristics. In recent studies, several researchers have defined different performance measures to understand driving characteristics and to evaluate their studies. Such measures include driving performance, driver behavior, task performance, etc. Driving performance measures consist of driver inputs to the vehicle or measures of how well the vehicle was driven along its intended path [1]. Driving performance measures can be defined by longitudinal velocity and acceleration, standard deviation of steering-wheel angle and its velocity, standard deviation of the vehicle's lateral position (lane keeping), mean following distance (or head distance), response time to brake, etc. Driver behavior measures can be defined by glance time, number of glances, awareness of drivers, etc. Task performance measures can be defined by the time to complete a task and the quality of the completed task (e.g., do drivers acquire information they need from cell-phone calling). Therefore, multi-modal data acquisition is very important to these studies.

UTDrive is part of three-year NEDO-supported international collaboration between universities in Japan, Italy, Singapore, Turkey, and USA. The UTDdrive (USA) project has been designed specifically to: (i) collect rich multi-modal data

This work was sponsored by grants from NEDO (Japan) and the University of Texas at Dallas under project EMMITT.

recorded in a car environment (i.e., audio, video, gas/brake pedal pressures, forward distance, GPS information, and CAN-Bus information including vehicle speed, steering angle, pedal status), (ii) assess the effect of speech interactive system on driver behavior, (iii) formulate better algorithms to increase accuracy for in-vehicle ASR systems, (iv) design dialog management which is capable of adapting itself to support a driver's cognitive capacity, and (v) develop a framework for smart inter-vehicle communications.

The results of this project will help to develop a framework for building effective models of driver behavior and driver-to-machine interactions for safe driving. In real driving situations, even a small improvement in cognitive driver load management can improve and reduce accidents.

2. Multi-Modal Data Acquisition and Hardware Setup

In this section, we discuss the data acquisition and integration of the hardware components in the data-collection vehicle.

2.1. Audio

A custom designed adjustable five microphone array, based on an earlier fixed array design [7] with omni-directional microphones was installed on top of the windshield next to the sun-light visors to capture audio signals inside the vehicle. Since there are various kinds of car background noise (e.g., A/C, engine, turn-signals, music, vehicles passing) present in driving environments, the microphone array configuration will allow application of beam-forming algorithms to enhance the quality of input speech signals [7, 8, 13]. In our setup, each microphone was mounted in a small movable box individually attached to an optical rail, as shown in Fig. 1. This particular design allows the spacing between each array microphone to be adjusted across the width of the windshield (e.g., linear scale, logarithmic scale, etc.) In addition, the driver speech signal is also captured by a close-talk microphone (Shure Beta-54). This microphone provides the reference speech of the speaker, and allows the driver to move their head freely while they are driving the vehicle.



Figure 1: Custom-designed adjustable-spaced microphone array.

2.2. Video

Two Firewire cameras are used to capture visual information of driver's face region and front-view of the vehicle, as shown

in Fig. 2. Real-time computer vision is an important component in understanding driver behavior (e.g., face and eyes detection to measure driver glances). In addition, studies have shown that combining audio and visual information of driver can improve ASR accuracy of low-SNR speech [3, 14]. Integrating both visual and audio content allows for rejecting of unintended speech prior to speech recognition and significantly improves in-vehicle dialog system performance [14] (e.g., determining the movement of the driver's mouth, body, and head positions).

2.3. CAN-Bus Information

As automotive electronics advance and government required standards evolve, control devices that meet these requirements have been embracing modern vehicle design resulting in the deployment of a number of electronic control systems. The Controller Area Network (CAN) is a serial, asynchronous, multi-master communications protocol suited for networking vehicle's electronic control systems, sensors, and actuators. The CAN-Bus signals contain real-time vehicle information in the form of messages integrating many modules, which interact with the environment and process high and low speed information. In the UTDrive project, we obtain CAN signals from the OBD-2 port through the 16 points J1962. Messages captured from CAN while the driver is operating the vehicle (e.g., steering wheel angle, brake and gas pedals, vehicle speed, engine speed, and vehicle acceleration) are desired to study driver behavior. Studies have shown that driver behavior can be modeled and predicted by the patterns of driver's control of steering angle, steering velocity, car velocity, and car acceleration [11], as well as driver identity itself [4, 12].

2.4. Transducers and Extensive Components

In addition, the following transducers and sensors are included into the UTDrive framework:

- Brake and gas pedal pressure sensors: provides continuous measurement of pressure driver puts on the pedals.
- Distance sensor: provides the forward head distance to the next vehicle.
- GPS: provides standard time and position of vehicle.
- Hands-free car kit: provides safety during data collection and audio data of both audio channels to be recorded for wireless dialog interaction.
- Biometrics: heart-rate and blood pressure measurement, physiological microphone.

2.5. Data Acquisition Unit (DAC)

The key component of multi-modal data collection is data synchronization. In our setup, we use a fully integrated commercial data acquisition unit (DAC). With a very high sampling rate of 100 MHz, the DAC is capable of synchronously recording multi-range input data (i.e., 16 analog inputs, 2 CAN-Bus interfaces, 8 digital inputs, 2 encoders, and 2 video cameras), and yet allows sampling rate for each data to be set individually. The DAC can also export all recording data as a video clip in one output screen, or individual data in its proper format (e.g., .wav, .avi, .txt, .mat, etc.) with synchronous time stamps. The output video stream can be encoded to reduce its size, and then transcribed and segmented with an annotation tool. Fig. 2 shows a snapshot of a recording video clip with all data displayed on the



Figure 2: A snapshot from a recording screen.

screen (e.g., audio channels on the top, two camera screens in the middle, sensors and CAN-Bus messages on the left bottom, and GPS information on the right bottom).

In order to avoid signal interference, the power cables and the signal cables were wired separately on both sides of the car. The data acquisition unit is mounted on a customized platform on the backseat behind the driver. The power inverter and supplier units are designed to be housed in the trunk space.

Table 1: Data Acquisition of Driving Signals.

Signals	Acquisition Rate (Hz)
Audio: 5-channel microphone array	25,000
Audio: close-talk Microphone	25,000
Sensor: brake/gas pedal pressures	100
CAN: steering wheel	50
CAN: brake	40
CAN: engine RPM	32
CAN: vehicle speed	2
GPS information	1
Video: two Firewire cameras	640x480, 15 fps

3. Data Collection Procedure

For data collection, we recruit subjects from UTD's students, staff, and faculty. Each participant will drive the data-collection vehicle following two different routes in neighborhood areas of the UTD campus (Richardson-Dallas, TX); the first route represents a residential area environment and the second represents a business-district environment. Each route take 10-15 minutes to complete. For each round, participants will drive the vehicle with different cognitive loads (e.g., free-style driving, driving with assigned tasks) for approximately one hour. Due to safety concerns, the assigned secondary tasks are common tasks with mild to moderate degrees of cognitive load. The participants are encouraged to drive the vehicle for three sessions with at least one week separation between sessions. Unless volunteering, participants will receive \$100 for completion of three driving sessions. The main secondary tasks are to: (i) Interact with commercial ASR dialog systems using hands-free car kit. The driver call an airline's flight connection system to check the de-

parture/arrival gates of particular flights, and call a voice portal to obtain information depending on personal interest (e.g., weather forecast at arrival city of their trip.), (ii) Read signs, street names, license plate numbers, etc., (iii) Tune radio; Insert a CD, Select CD track, (iv) Have general conversation with passenger, (v) Report driving activities, and (vi) Change/Keep lanes. Currently, the UTDrive plan will include 100-300 drives over the next six months. A present 50 sessions from 30 drivers have been obtained.

4. Driver Distraction

While it is clear that driving behavior deteriorates for drunk drivers (the major cause of fatal accidents in U.S.), the use of cell-phone technology can impact cognitive load, yet be a valuable resource in requesting emergency assistance when accident occurs. However, communication technologies require a driver to cognitively interact with the device for a long period of time compared to CD players or activities such as eating and drinking. According to the National Highway Traffic Safety Administration (NHTSA), there are four distinct types of driver distraction: visual, auditory, biomechanical (physical), and cognitive distractions. Although these four modes of distraction are separately classified, they are not mutually exclusive. For example, interacting with in-vehicle spoken dialog system may include all four forms of distraction: dialing the phone (physical distraction), looking at the phone (visual distraction), holding a conversation (auditory distraction), and focusing on the conversation topic (cognitive distraction). Fig. 3 shows sample data plots of (a) accelerator in RPM, (b) vehicle speed in km/h, and (c) normalized steering-wheel angle of a driver for the same route twice. The free-style or normal driving (neutral) is shown on the top of each plot, and driving while interacting with a spoken dialog system is shown below each neutral. Here, the driver maintains a smoother driving patterns when not interacting with the system, and a driver needs to make excessive steering-wheel maneuver for lane keeping (the vertical line in plot (c) illustrates the sharp movement of steering wheel between left and right). The collection of multi-modal data for in-vehicle applications is critical in transitioning speech technology to real-world environments. Dialog system developers do not take into account the cognitive load a user may experience while performing driving and dialog tasks. Speech recognition research has made important strides in addressing stress and noise for ASR [6], however better corpora are needed by the multi-modal research if speech technology is to have an impact in real every-day environment.

5. Discussion

This paper has described an overview of the UTDrive project and vehicle setup for real-time multi-modal data acquisition in a real-driving scenario. The objective of our project is to develop mathematical models that are able to predict driver behavior and performance while using speech interactive systems, as well as improve speech interactive systems to accomplish reduced distraction/improved safety for in-vehicle systems.

5.1. Transcription Conventions

The major challenge facing our efforts and the community, in utilizing rich multi-modal data is a unified transcription protocol. Besides speech transcription, which is well-defined in the speech community, multi-layer transcription must be addressed.

For example:

- Audio: different types of background noise inside and outside the car, passenger's speech, radio sound, and ring-tone occur.
- Driving Environment: type of roads (number of lanes, curve or straight, highway or local, speed limit), traffic (traffic situation, traffic light, surrounding vehicles), road condition.
- Driver Activity: look away from the road, talk to passenger, dial a phone, talk to the phone, look at rear mirror, look at control panel, sleepy, day-dreaming, etc.
- Vehicle Mode: left or right turn, left or right lane changing, U-turn, stop and go, stop, etc.

5.2. Research Directions

Our next step is to research for the optimal feature extraction from driving signals and mathematical modeling techniques that can explain and predict driver behavior. Recent research has found that cepstral features of driving signals and Gaussian Mixture Model (GMM) efficiently model driver behavior and achieve good driver identification performance [10]. We are presenting researching multi-modal driver distraction detection.

6. References

- [1] A. Baron and P. Green, "Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review," Tech. Report UMTRI-2006-5, Feb. 2006.
- [2] C. Carter and R. Graham, "Experimental Comparison of Manual and Voice Controls for the Operation of In-Vehicle Systems," in *Proc. of the IEA2000/HFES2000 Congress*, Santa Monica, CA
- [3] T. Chen, "Audio-visual speech processing," *IEEE Sig. Proc. Magazine*, vol. 18, no. 1, pp 9–21, 2001.
- [4] H. Erdogan, A. Ercil, H.K. Ekenel, S.Y. Bilgin, I. Eden, M. Kirisci, H. Abut, "Multi-modal person recognition for vehicular applications," N.C. Oza et al. (Eds.): MCS-2005, LNCS-3541, pp. 366–375, Monterey, CA, Jun. 2005.
- [5] C. Forlines, B. Schmidt-Nielsen, B. Raj, P. Wittenburg, and P. Wolf, "Comparison between Spoken Queries and Menu-based interfaces for In-Car Digital Music Selection," *TR2005-020*, Cambridge, MA: Mitsubishi Electric Research Laboratories.
- [6] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communications*, Special Issue on Speech Under Stress, vol. 20(2), pp. 151-170, Nov. 1996.
- [7] J. H.L. Hansen, J. Plucienkowski, S. Gallant, R. Gallant, B. Pellom, and W. Ward, "CU-Move: Robust speech processing for in-vehicle speech systems," in *ICSLP*, pp. 524–527, 2000.
- [8] T.B. Hughes, H.-S. Kim, J.H. DiBiase, and H.F. Silverman, "Performance of an HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 3, pp. 346–349, 1999.
- [9] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.
- [10] C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura, "Driver Modeling Based on Driving Behavior and Its Evaluation on Driver Identification," *Proc. of the IEEE*, vol. 95, no. 2, pp. 427–437, Feb. 2007.
- [11] A. Pentland and A. Liu, "Modeling and Prediction of Human Behavior," *Neural Computation*, vol. 11, pp. 229–242, 1999.
- [12] A. Wahab, T.-C. Keong, H. Abut, and K. Takeda, "Driver recognition system using FNN and statistical methods," Chapter 3 in *Advances for in-vehicle and mobile systems*, Abut, Hansen, Takeda (Ed.s.), Springer, New York, 2007.
- [13] X.-X Zhang and J.H.L. Hansen, "CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Environments," *IEEE Trans. Speech & Audio Proc.*, vol. 11, no. 6, pp 733–745, Nov. 2003.
- [14] X.-X. Zhang, K. Takeda, J.H.L. Hansen, and T. Maeno, "Audio-Visual Speaker Localization for Car Navigation Systems," in *INTERSPEECH-2004*, Jeju Island, Korea, 2004.

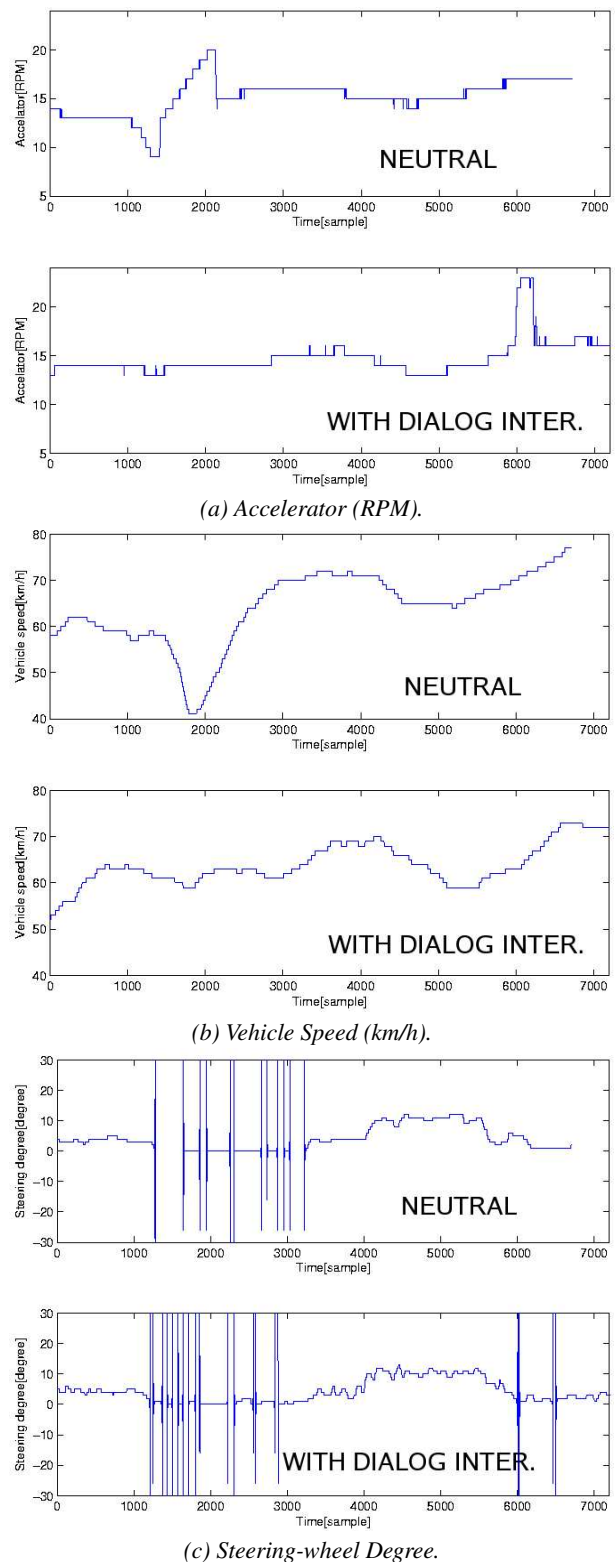


Figure 3: Comparison of driving signals (Accelerator, Vehicle Speed, and Steering-Wheel Movement) of a driver for the same road. Top: No distraction, Bottom: Interact with a voice portal.