Proceedings of the
2007 IEEE Intelligent Vehicles Symposium
Istanbul, Turkey, June 13-15, 2007



# UTDrive: Driver Behavior and Speech Interactive Systems for In-Vehicle Environments

Pongtep Angkititrakul, Matteo Petracca*, 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, USA
{angkitit,axs063000,John.Hansen}@utdallas.edu, matteo.petracca@polito.it

*Abstract*— **This paper describes an overview of the UTDrive project. UTDrive is part of an on-going international collaboration to collect and research rich multi-modal data recorded for modeling driver behavior for in-vehicle environments. The objective of the UTDrive project is to analyze behavior while the driver is interacting with speech-activated systems or performing common secondary tasks, as well as to better understand speech characteristics of the driver undergoing additional cognitive load. The corpus consists of audio, video, gas/brake pedal pressure, forward distance, GPS information, and CAN-Bus information. The resulting corpus, analysis, and modeling will contribute to more effective speech interactive systems with are less distractive and adjustable to the driver's cognitive capacity and driving situations.**

## I. Introduction

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], [6]. Although providing better interfaces, operating a speech interactive system will still divert a driver's attention away from his or her 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 a speech interactive system 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. Therefore, the design of safe speech interactive systems for in-vehicle environments should take into account the factors from the driver's cognitive capacity, driving skills, and the proficiency degree of the cognitive load of the applications. 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 a variety of background noise effects the quality of the input acoustic signal of the speech interface. More importantly, drivers have to modify their vocal effort to overcome noise levels in their ears, namely the Lombard effect [11]. 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 [7], [5]. At a higher level, interacting with an automatic speech recognition (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 the dialog system. Desirable dialog management should be able to employ multi-modal information to handle errors and adapt its context depend 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 [12]. 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, 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 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 UTDrive (USA) project has been designed specifically to:

- collect rich multi-modal data recorded in a car environment (i.e., audio, video, gas/brake pedal pressures, forward distance, GPS information, and CAN-Bus in-

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

* M. Petracca is with the Dipartimento di Automatica e Informatica, Torino, Italy

1-4244-1068-1/07/$25.00 ©2007 IEEE.

formation including vehicle speed, steering angle, pedal status),
- assess the effect of speech interactive system on driver behavior,
- formulate better algorithms to increase accuracy for in-vehicle ASR systems,
- design dialog management which is capable of adapting itself to support a driver's cognitive capacity,
- 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.

## II. UTDrive Data and Equipment

In this section, we present an overview of the hardware setup and collection protocol for UTDrive.

### A. Audio

A custom designed five microphone array with omni-directional microphones was installed on top of the windshield next to the sunlight visors to capture audio signals inside the vehicle. Since there are various kinds of car background noise (e.g., A/C, engine, turn-signals, vehicles passing) presented in driving environments, the microphone array configuration will allow us to apply beam-forming algorithms to enhance the quality of input speech signals [8], [9], [15]. 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 microphone of the array to be adjustable across the width of the windshield (e.g., linear, logarithmic, etc.). One of our preliminary studies showed that logarithmic scale outperformed linear scale in terms of SNR improvement for some noise conditions, with a basic delay-and-sum beam-forming processing. The optimization of array configuration and beam-former processing is another challenge which is being considered.

In addition, the driver speech signal is also captured by a close-talk microphone (Shure Beta-54) which was connected to a phantom power supply. This microphone provides the reference speech of the speaker, and allows the driver to move their head freely while they are driving the vehicle.

### B. 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 to understand 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], [16]. Integrating both visual and audio contents allows us to reject unintended speech prior to speech recognition and significantly improve in-vehicle human-machine dialog system performance [16] (e.g., determining the movement of the driver's mouth, body, and head positions).

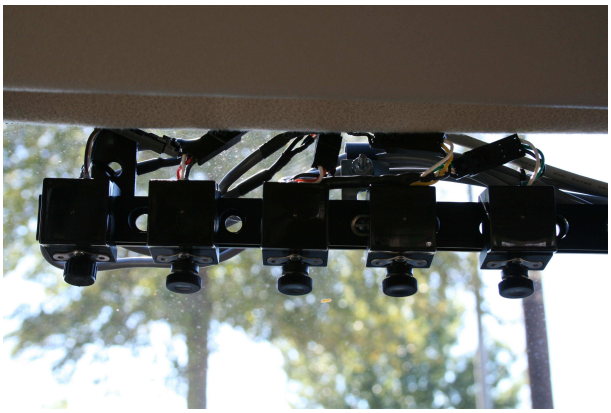

Fig. 1. Custom-designed adjustable-spaced microphone array.

Fig. 2. Sample of two synchronous video streams (left: front view, right: driver).

### C. 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 these 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 the 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 [13], as well as driver identity itself [4], [14].

### D. 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, compensating the binary information (on/off) obtained from CAN-Bus.
- Distance sensor: provides the forward 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.
- Biometrics: heart-rate and blood pressure measurement.

### *E. Data Acquisition Unit (DAC)*

The key component of multi-modal data collection is synchronization of all data. In our data collection, we use a fully integrated commercial data acquisition unit (DAC). With a very high sampling rate of 100 MHz, 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. DAC can 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. 3 shows the UTDrive vehicle and its core components.

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.

## III. Data Collection Process

For data collection, each participant will drive the vehicle using two different routes in the neighborhood areas of the UTD campus (Richardson-Dallas, TX); one route represents a residential area environment and the second route represents a business district environment. Each route takes 10-15 minutes to complete one round. The participant drives the vehicle two rounds for each route, the first round is normal driving and second round is driving and performing some secondary tasks. Due to safety concerns, the assigned secondary tasks are common tasks with mild to moderate degrees of cognitive load. Consequently, drivers use a hands-free car kit when they interact with dialog systems on the cell-phone. Participants can refuse to perform any tasks which they do not feel comfortable in using during driving.

The main secondary tasks are to:

- Interact with commercial ASR dialog systems. The drivers call an airline's flight connection system to check the departure/arrival gates of particular flights, and call a voice portal to obtain information depending on personal interests (e.g., forecast weather at arrival city of their trips).
- Read signs, street names, license plate numbers, etc.
- Tune radio, Insert a CD, Select CD track.
- Have general conversation with passenger.
- Report driving activities.

In order to acquire session-to-session variability, each subject is encouraged to participate in the driving for two more times with at least one week separation between sessions. Currently, the UTDrive plan will include 100–300 drivers completing 1-2 routes over the next six months. Results will be presented from these subjects. Sample resulting data streams from the integrated collection platform are shown in Fig. 3.

## IV. Inter-Vehicle Communications

Another aspect of the UTDrive project is the communications between vehicles. Data collected in the vehicle can be exchanged with the other vehicles, by means of a wireless network, in order to support safety applications. Inter-vehicle communications of driving behavior and vehicle status can avoid collision between vehicles at the intersections and parking lots (e.g., restricting back off distance if another car will not stop), or allow vehicles to send requests for help or alert other vehicles about high risk of accidents using voice over IP (VoIP) communications.

Inter-vehicle communications are challenging due to the high variability of the wireless channel conditions and the topology of the network. The wireless channel, in fact, is affected by noise, refraction, reflection, and attenuation of the electromagnetic waves, which can generate packet losses during transmission. The variation of the speed and the route of the vehicles, instead, produces changes in the topology of the network, and requires new routing protocols.

In order to develop safe applications based on VoIP communications, we evaluated the performance of one-hop inter-vehicle networks transmitting speech over an IEEE802.11b network. In particular, we studied the scenario in which a car is parked near the roadside and is sending a VoIP help request to another car driving in close proximity.

In our experiments, we used two Cisco3200 wireless routers with a 6 dBi gain dipole antenna, a transmission power of 100 mW and a CSMA/CA data retry limit of 64. The parked car was at the edge of a parking lot, while the second car was driven on the UTD campus between parking spaces and buildings. Other interfering wireless networks were present on campus. In order to send data on the network, we used a UDP-based traffic generator sending 5-minutes VoIP flows at the typical bit rates of GSM AMR, iLBC, G729, and MELP speech coders with concurrent video flows at 500 kb/s. The results of the collected data, in terms of Packet Loss Rate (PLR), are shown in Table I. Our experimental results show that the packet loss rates for the video and speech flows are comparable, while the packet loss rates among all the speech flows decrease when the dimension of the packets decreases and the speech frame length increases.

## V. Conclusion and Future Work

This paper has described an overview of the UTDrive project and vehicle setup for real-time multi-modal data

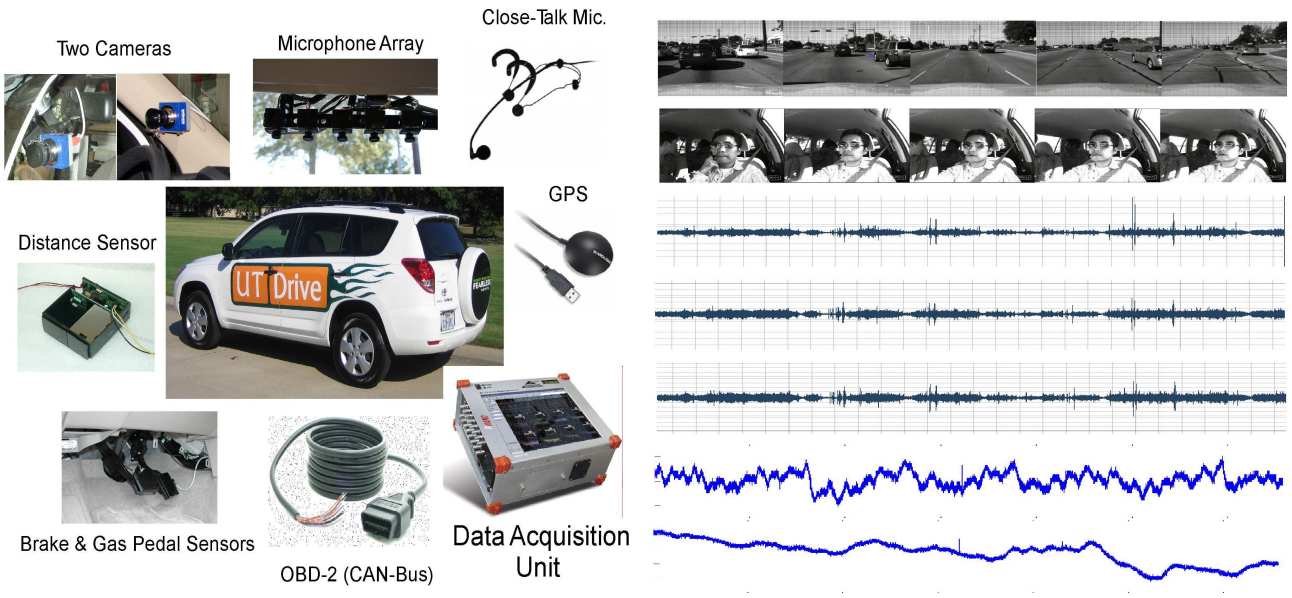


Fig. 3. UTDrive setup and sample synchronous data (two video streams, three audio streams, gas and brake pedal pressure).

TABLE I
PACKET LOSS RATES FROM DIFFERENT SPEECH CODERS.

| VoIP Bit Rate | Speech PLR | Video PLR | Speech Frame Length (ms) |
|---|---|---|---|
| iLBC 13.33 kb/s | 14.23% | 14.01% | 30 |
| iLBC 15.20 dB/s | 20.70% | 20.84% | 20 |
| GSM AMR 12.2 kb/s | 16.80% | 17.62% | 20 |
| GSM AMR 7.4 kb/s | 14.27% | 14.84% | 20 |
| GSM AMR 4.75 kb/s | 15.86% | 16.38% | 20 |
| G.729 8 kb/s | 20.34% | 20.53% | 10 |
| MELP 2.4 kb/s | 13.78% | 13.62% | 22.5 |

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. An up-to-date summary of data collection and driver modeling will be presented at the meeting with further details and discussion on international collaboration, exchange, and transcription standard.

## VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of each student assistant in our team–Levi Tarvis Noecker, Jeremy Hayes, Anh Phuc Phan, Tyler Creek, Peter Moreno, Vitali Ruder, Paul Grein, Wayne Lanham, and Gustavo Litovsky.

## REFERENCES

[1] A. Baron and P. Green, "Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review," Technical 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 *Proceedings 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] R. Fernandez and R.W. Picard, "Modeling driver's speech under stress," *Speech Communications*, vol. 40(1-2), pp. 145–159, 2003.

[6] 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.

[7] 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, November 1996.

[8] 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.

[9] 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.

[10] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Lie, T. Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," in

[11] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[12] N. Oliver and A.P. Pentland, "Graphical Models for Driver Behavior Recognition in a Smartcar," *Intelligent Vehicles 2000*, Detroit, Michigan, Oct. 2000.

[13] A. Pentland and A Liu, "Modeling and Prediction of Human Behavior," *Neural Computation*, vol. 11, pp. 229–242, 1999.

[14] 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 (Edts.), Springer, New York, 2007.

[15] X.-X Zhang and J.H.L. Hansen, "CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Environment," *IEEE Trans. Speech & Audio Proc.*, vol. 11, no. 6,pp 733–745, Nov. 2003.

[16] 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.

[17] http://www.cisco.com/application/pdf/en/us/guest/products/ps272/c1650/cdccont_0900aecd800fe971.pdf