A Multimodal Corpus of Situated Interactions within Vehicles

|  |  |  |
| --- | --- | --- |
| Ian Lane  Carnegie Mellon University  NASA Ames Research Park  Moffett Field, CA, USA  lane@cs.cmu.edu | David Cohen  Carnegie Mellon University  NASA Ames Research Park  Moffett Field, CA, USA  david.cohen@sv.cmu.edu | Antoine Raux  Honda Research Institute  Affiliation / Address line  Affiliation / Address line  ARaux@hra.com |

Abstract

In this paper, we introduce a multimodal corpus of human-human interactions collected within highly sensored vehicles. The corpus consists of interactions between a driver and copilot performing tasks including navigation, scheduling and messaging. Data was captured synchronously across a wide range of sensors, including, near-field and far-field microphones, internal and external cameras, GPS, IMU, and OBD-II devices. The corpus was collected to investigate situated interactions within vehicles and is unique in that it not only contains transcribed speech, dialog acts and annotated gestures, but also grounds object references and navigation discussions to physical objects and actions. An initial analysis of the corpus indicates that a large variety of objects are used as reference points for collaboratively performing navigation within the vehicle and gestures are pervasive within this task.

1. Introduction

Developing intelligent agents that can understand and interact with users in dynamic, physically situated environments remains a grand challenge for spoken dialog research. While there has been a wealth of research in interactive spoken dialog agents over the past decade [1,2,3] only recently have researchers begun to explore spoken language understanding [4,5] and interaction [6] in physically situated environments.

Many challenges lay ahead in order to develop the technologies required to enable spoken dialog agents to understand the environment in which they operate and to interact with users in these environments. Challenges include monitoring and understanding situational context, understanding spoken language, gestures and user actions in a physical environment, physically grounding object references and actions, and incorporating environmental cues in turn taking.

To address these challenges, rich multimodal corpora of situated human-human interactions are necessary. While some such corpora exist for robots [], smart homes [], and cars [], there is, to our knowledge, no available corpus of in-car human-human interaction with rich sensor information from both inside and outside the car.In addition, because these interactions are grounded in the physical world, it is not sufficient to merely transcribe the participants' speech, but one must also annotate the situational context in detail.

In this paper we describe a multimodal corpus of in-car human-human interactions that we collected to both analyze situated human-human interaction and to develop core technologies to support future intelligent agents, that can understand and interact in complex and dynamic real-world environments. The corpus consists of interactions between a driver and copilot performing tasks including navigation, scheduling and messaging.

Data collection was performed using a highly sensored data collection rig that synchronously captured data across a wide range of sensors placed on the vehicle, including; audio from headset and far-field microphones, video from internal and external cameras, as well as, GPS, IMU, and OBD-II data. The resulting corpus contains synchronized raw sensor data, time aligned manual transcriptions of driver and copilot speech, as well as annotations of discourse domain, dialog acts, gestures and grounded references to physical objects and actions.

In this paper we introduce the corpus, data collection rig, collection procedure and annotation scheme used in this project. For the navigation domain we also provide an initial analysis of object references and presence of gestures in this domain.

|  |  |
| --- | --- |
|  |  |
| (a) External sensors mounted on roof rack | (b) Internal sensor mounted inside of vehicle |

Figure 1: Sensor placement on data collection vehicles

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sensor** | **Location** | **Description** | **Sampling Rate** | **Sample size** |
| Stereo Camera-pair | External | 2 Pointgrey Firewire cameras mounted 100cm apart on | 30fps | 640x480 RGB (x2) |
| 360 Camera | “ | Sony Bloggie with 360o lens | 30fps | 1920x1080 |
| GPS | “ | SF2050 GPS unit | 50Hz | 128 Bytes |
| IMU | “ |  | 120Hz | 32 Bytes |
|  |  |  |  |  |
| Driver Camera | Internal | Logitech C910 USB camera | 30fps | 640x480 |
| Kinect | “ | Microsoft Kinect Sensor | 30fps | 640x480 |
| Headset Microphone | “ | Countryman e6 microphones with stereo USB audio | 48kHz | 16bit |
| Microphone Array | “ | 8 x C2070 Senhieer microphones with USB audio | 16kHz | 16bit |
| OBD-II Information | “ |  | 10Hz | 256 Bytes |

Table 1: Sensors and capture settings during data collection

1. The CESAR Data Collection Platform

Data collection was performed using **CESAR**, the “**C**ar **E**nvironment **S**ensor **A**djustable **R**ig” which was developed at CMU specifically to capture synchronized recordings across audio, visual and vehicular sensors. Additionally the platform was designed to be easily shifted between vehicles. In the data collection described in this paper we collected data in 10 different vehicles across the 43 runs.

The Rig consists of three main components, a data collection PC which resides in the trunk and operates off the 12V power supply of the vehicle, a roof rack on which the external sensors are mounted (Figure 1-a), and a set of internal sensors which are mounted in the cabin of the vehicle (Figure 1-c). Table 1 lists the sensors used in the data collection described below.

External sensors included external cameras to capture the driver’s field-of-view, a high-precision GPS, and an IMU for car orientation and chassis vibration. Internal sensors consisted of a USB camera, a Kinect and a microphone-array which were setup to capture the driver and copilot interaction as shown in (Figure 1-b). A CAN-BUS device was used to capture the car's On-Board Diagnostic (ODB-II) information

The data collection PC consists of a standard desktop PC with a 3.2GHz i7 Intel processor and 32GB of memory, running Windows 7. An 8-way SSD raid was used to capture uncompressed sensor data. One 40 minute data collection run, typically produced 420GB of data which was first captured onto the SSD raid and then copied to an external disk drive for storage. The data collection software was written in C++ and C#, and optimized to limit frame loss.

|  |  |  |
| --- | --- | --- |
|  | **Initial instructions** | **Unannounced tasks / interruptions** |
| **1** | Pick a sight-seeing destination and visit it. | None. |
| **2** | Go to the post office, the gym, McDonald's. | Trip to the gas station to refill. |
| **3** | Drop off colleague at their meeting, go to meeting. | Unplanned detour to avoid traffic. |
| **4** | Go to a second meeting. | Text message exchange with a friend, plan a visit. Return to destination 3 to drop documents forgotten by colleague. |
| **5** | Go back to the hotel. | Negotiate with two friends via dictated text messages to arrange dinner at a restaurant. |

Table 2. The 5 scenarios as initially described to the participants, and the associated interruptions.

1. Data Collection and Annotation
   1. Data collection Procedures

We collected driver/co-pilot interactions while performing various tasks involving navigation inside Moffett Federal Airfield. Each session involved a driver, which was recruited for the experiment and had no knowledge of our research or of the geography of Moffett Field, and a co-pilot who was familiar with Moffett Field and the tasks to complete. No instructions about how to interact with each other was given to either participant.

After hearing a brief explanation of the experiment, the driver and co-pilot negotiated the first trip, and started driving. Once each scenario was completed, the co-pilot provided the driver with the instructions for the next scenario and they performed its tasks. At the end of the experiment, they returned to their starting point and the driver received compensation for their participation.

* 1. Data Collection Scenarios

In the course of one run, the participants had to complete five scenarios of increasing complexity. A short description of each scenario is given in Table 2.

All scenarios involve driving between locations on Moffett Field. In scenarios 2 through 5, the path or goal needs to be changed dynamically while driving due to some unforeseen event. Such events were simulated by having the co-pilot pretend that they just received new information (either in the form of traffic updates, text messages, etc) and informing the driver (who was unaware that the event would occur). These interruptions make this corpus a valuable resource to investigate planning and negotiation dialogs between driver and co-pilot occurring while driving. Also, we designed scenarios 4 and

|  |  |
| --- | --- |
| **Domain** | **Description** |
| Navigation | Driving directions and clarification dialog |
| Scheduling | Dialog discussing times and places for scheduling events |
| Alerts & Messaging | Alerts such as gas and traffic alerts, as well as any discussion related to mediating messages be-tween the driver and their contacts |
| Business Search & Local Guide | Recommendation and business directory lookups, as well as discussion about the various sight-seeing destinations |
| Out-of-Domain | Discussion out-side of the above domains, which might naturally occur in a real scenario |
| Experiment OOD | Discussion relating to the conditions of the experiments |

Table 3: Domain labels used during annotation

5 to involve receiving and responding to text messages in order to analyze the way humans would deal with such tasks in a safe way, as opposed to traditional methods such as reading from and typing on a smart phone.

We found that all subjects were able to achieve the tasks, though with varying degrees of collaboration and efficiency.

* 1. Annotations

The corpus was transcribed, and a multi-pass detailed annotation was performed on the 20 runs with the best data. The first pass seeks to facilitate more detailed annotation by annotating each domain of interest. Then we focused on quantifying some expected characteristics of the situated navigation domain. Specifically, 5 types of annotation were performed:

**(1) Transcription of speech:** Speech from the driver and copilot was manually segmented and transcribed.

**(2) Domain annotation:** Each utterance was tagged with all the domains it pertains to as listed in Table 3.

|  |  |
| --- | --- |
| **Domain** | **Time spent in domain (percentage of speech)** |
| Navigation | 53 mins (51%) |
| Out-of-Domain | 18 mins (17%) |
| Experiment-OOD | 16 mins (15%) |
| Alerts & Messaging | 15 mins (14%) |
| Scheduling | 8 mins (8%) |
| Business Search … | 5 mins (5%) |
| Total speech time | 104 mins |
| Total interaction time | 181 mins |

Table 4: Total time spent in each discourse domain over subset of 5 data collection runs

**(3) Annotation of Grounded Navigation Discourse units:** Navigation discourse units are defined as segments of dialog which ground a physical navigation action. Based off initial data viewing, we think that the majority of navigation actions being discussed can be categorized as Go-to-region, Leave-region, or Stop-at-region. All of these action categories are parameterized by a single spatial region, which could be a parking space, road, etc. For example, a direction to turn left at an intersection and the ensuing grounding dialog are interpreted to be a Go-to-region navigation discourse unit, parameterized by the section of street which lies to the left of the intersection.

**(4) Dialog act annotation with gestures:** Within each navigation discourse unit, grounding dialog acts is segmented and labeled. The trial set of dialog acts includes a typical set of dialog acts: direct, offer, acknowledge, etc. Each dialog act annotation is also given a boolean parameter signifying whether or not the speaker uses non-verbal communication to contribute to the dialog act's meaning.

**(5) Grounded object reference annotation with gestures:** Within the navigation domain utterances, references to objects within the immediate vicinity of the speaker are labeled and parameterized with the object being referred to. These objects may be permanent structures such as traffic signs, buildings, or streets, in which case the annotation will point to those objects in a detailed map of Moffett Field. In other cases, the objects will be mobile, in which case a new object will be created for the purpose of grounding the reference.

Transcription, domain annotation, and object reference annotation with gestures have been completed for a 5-run subset of the corpus. Table 4 shows the amount of speech time (in minutes) broken down by domain.

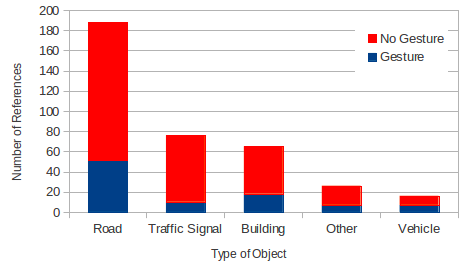


Figure 3: Co-occurrence of object reference and gesture within navigation discourse

Figure 3 shows object references broken down by the type of object being referred to and by the presence of gesture along with the utterance. The categories of object considered are: building or public space, person or vehicle, road or driveway, traffic signal, or other.

Some examples from the other category that appeared in these runs are trees, a yellow fence, and a sign that says 'swimming pool'. These initial numbers suggest that gesture plays a role in a significant percentage of situated object references, and that a interesting variety of objects are useful for discussing navigation information.

1. **Conclusion**

In this paper, we introduced a multimodal corpus of human-human interactions collected within vehicles. The corpus was collected specifically to investigate situated interactions within vehicles in order to both analyze situated human-human interaction and to develop core technologies to support future intelligent agents. It is unique in that is not only contains transcribed speech, dialog acts and annotated gestures, but also grounds object references and navigation discussions to the physical objects and actions that are performed in the real world. An initial analysis of object references in a subset of the corpus indicates that a large variety of objects are useful as reference points for collaboratively performing navigation tasks. Analysis of the presence of gestures in such references indicates the importance and pervasiveness of non-verbal communication in this task.

**Acknowledgments:** This research was performed at CMU under the sponsored research agreement 260-456-112 with the Honda-Research-Institute, USA

References

Al

Kawaguchi, N., Matsubara, S., Takeda, K., & Itakura, F. (2005). CIAIR In-Car Speech Corpus--Influence of Driving Status--. IEICE transactions on information and systems, 88(3), 578-582.