The HRI-CMU Corpus of Situated In-Car Interactions

David Cohen, Akshay Chandrashekaran, Ian Lane, and Antoine Raux

Abstract This paper introduces the HRI-CMU Corpus of Situated In-Car Interactions, 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 in the vehicle, including, near-field and far-field microphones, internal and external cameras, GPS, IMU, and OBD-II devices. The corpus is unique in that it not only contains transcribed speech, annotation of dialog acts and gestures, but also includes grounded object references and detailed discourse structure for the navigation task. We present the corpus and provide an early analysis of the data contained within. The initial analysis indicates that discourse behavior has strong variation across participants, and that general trends relate physical situation and multi-tasking to grounding behavior.

David Cohen
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA, 15213, e-mail: david.cohen@sv.cmu.edu

Akshay Chandrashekaran
Carnegie Mellon University, NASA Research Park #23, Moffett Field CA, 94043, e-mail: akshay.chandrashekaran@sv.cmu.edu

Ian Lane
Carnegie Mellon University, NASA Research Park #23, Moffett Field CA, 94043, e-mail: ian.lane@cs.cmu.edu

Antoine Raux
Honda Research Institute USA, 425 National Ave. #100, Mountain View CA, 94035 e-mail: araux@lenovo.com (The author is now at Lenovo Labs)

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 most research to date has focused on speech-only interaction over the telephone [25, 10, 22, 28], recently there has been increased interest in spoken dialog systems that can operate in physically situated environments. Examples include the Mission Rehearsal exercise described in [5], the Microsoft Receptionist [3], the CoSy project [7] and the AIDAS [12] and Townsurfer [9] systems.

A broad array of research challenges exist within the area of situated interaction, all of which need to be considered to realize robust and natural interaction. Challenges include monitoring and understanding situational context [15, 6, 8], understanding situated [18, 14] and spatial language [17, 24], grounding of object references in situated dialog [2], and co-reference resolution [27, 13, 20]. Additionally, in multimodal interaction, gaze, gestures and user actions [4, 11, 21] must all be understood in relation to the physical environment in which they occur.

While corpora exist to develop and evaluate the performance of component technologies within spoken dialog systems, there is limited data available on situated tasks in the real-world. Existing corpora focus on simple dialog over the telephone [29], with robots [1], smart homes [19], and in cars [23].

In this paper we introduce a multimodal corpus of situated in-car interactions that we collected to both analyze situated human-human interaction and to develop core technologies to support research in situated interaction. The corpus consists of interactions between a driver and passenger performing information retrieval, navigation, scheduling and messaging tasks. Data collection was performed using a highly sensored data collection platform that synchronously captured data across a wide range of audio, visual, and vehicular sensors. The resulting corpus contains synchronized data streams, time aligned transcriptions of driver and passenger interactions, as well as annotations of discourse domain, dialog acts, gestures and grounded references to physical objects and actions. In Section 2 of this paper we describe the data collection procedure and platform. Section 3 details the annotation performed and Section 4 presents an initial analysis of the corpus.

2 Data Collection Procedure

Data collection was performed at and around the Carnegie Mellon University campus at NASA AMES Research Park, Moffett Field CA. Collection was performed in a highly sensored vehicle as described in Section 2.2 below. The collection procedure was designed to elicit spontaneous, situated dialog between the driver and passenger, where the passenger's role was of a co-pilot, who supported the driver to complete the set of assigned tasks. Drivers were external participants that were recruited and compensated for their participation. They had no prior knowledge of the geography of the area or the tasks they were to perform. The co-pilot was one of six

lab assistants who were familiar with the geography of Moffett Field and the tasks to complete. No instructions about how to interact were given to either participant.

2.1 Scenario Tasks

Each driver completed five tasks of increasing complexity. After hearing a brief explanation of the experiment, the driver and co-pilot negotiated the first trip, and started driving. After completing a task, the co-pilot provided the driver with the instructions for the next one. A short description of each task is given in Table 1. In tasks 2 through 4, the path or goal needed to be altered due to an unforeseen event. The co-pilot simulated these events by providing new information to the driver (in the form of traffic updates, text messages, etc.) at an appropriate time or location within a task. Tasks 4 and 5 were designed to involve receiving and responding to text messages while simultaneously performing a navigation task. Subjects were able to achieve all tasks, though with varying degrees of efficiency.

Table 1 The 5 task scenarios and planned interruptions used for data collection.

Task	Instructions Given	Interruptions
1	Pick a sight-seeing destination on Moffett Field and navigate to it	None
2	Go to the post office, the gym, then McDonald's	Trip to the gas station to refill
3	Drop off colleague at their meeting, then go to your meeting	Unplanned detour to avoid traffic
4	Go to your second meeting	Invited to friend's house, then and asked to return to destination in 3 to drop off documents forgotten by colleague
5	Return to hotel	None

2.2 Collection Platform

Data collection was performed using CESAR [16], the Car Environment Sensor Adjustable Rig. The CESAR platform was developed specifically to capture synchronized recordings across a large number of audio, visual and vehicular sensors, and could be moved between vehicles. Within this corpus a total of ten vehicles were used during the data collection. The rig consists of three main components, a data collection PC, which resided in the trunk of the vehicle, a roof rack, on which external sensors (external cameras, GPS antenna and IMU) were mounted, and a set of internal sensors (internal cameras, microphones and OBD-II connector) which were mounted in the cabin of the vehicle.

Table 2 Sensors and capture settings during data collection

Sensor	Location	Description	Rate (Hz)	Sample Size
Stereo Camera pair	External	Two cameras mounted 100 cm apart	30	640X480 (x2)
GPS	External	SF2050 GPS unit	50	128 Bytes
IMU	External		120	32 Bytes
Driver Camera	Internal	Logitech C910 USB Camera	30	640X480
Kinect	Internal	Microsoft Kinect Sensor	30	640X480
Headset Mic	Internal	Countryman e6 microphones	48000	16bits
OBD-II	Internal		10	256 Bytes

Table 2 lists the sensors used in the data collection. 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 microphones to capture the driver and copilot interaction. A CAN-BUS device was used to capture the car's On-Board Diagnostic (ODB-II) information.

3 Annotation and Corpus Overview

36 runs were transcribed, and more detailed annotation has been performed on 15 of them. The same detailed annotation is planned for 5 more runs, and additional annotators will be used to quantify annotator agreement for our annotation scheme. The detailed annotation performed includes another round of speech transcription validation, domain annotation, grounded object references including gestures, and navigation discourse annotation.

3.1 Speech Transcription

Speech transcription was performed by Appen Butler Hill, then researchers in our group gave another pass on the 15 runs that were being annotated in more depth. The Kaldi speech decoder was used to align word boundaries. Table 3 summarizes the speech data in this corpus.

Table 3 Amount of annotated speech data in the corpus (hours).

	Driver	Copilot	Total Speech	Total Audio
Transcribed	4.58	6.53	11.11	22.56
Fully Annotated	1.88	2.65	4.53	9.17

3.2 Domain

Each word is labeled with the domains it is relevant to. This allows later annotators and researchers to quickly extract the sections of the data that are of interest to them. The Alerts / Messaging domain relates to any alerts the copilot delivers to the driver or any messages that the copilot is relaying between the driver and his contacts. The Navigation domain includes any discussion about where the driver is going and how to get there. The Scheduling domain covers discussion pertaining to when different people will be performing high-level tasks such as going to meetings. The Experiment-OOD domain is dialog where the participants break character or indicate that they are taking part in a controlled experiment. The domains are not mutually exclusive, so a word can belong to several domains. Table 4 breaks down the amount of speech data by speaker and domain.

Table 4 Percentage of speech by domain, broken down by speaker.

Speaker	Navigation	Business Search / Local Guide	Alerts / Messaging	Scheduling	OOD	Experiment-OOD	Total(Hrs)
Copilot	47.9	3.93	12.5	3.03	18.62	14.03	2.69
Driver	25.7	2.31	12.8	6.39	32.0	20.8	1.97
All	38.5	3.25	12.7	4.41	24.3	16.9	4.66

3.3 Object References

Groups of words that refer to a specific object or set of objects are labeled and grounded to one of over 800 geo-located objects on Moffett Field. We also labeled references to objects which are not stationary, and objects which are not in the immediate situation, such as the driver's fictional colleague and friend. Also, the presence of gesture to help ground a reference was annotated as yes / no. Table 5 summarizes the results.

Table 5 Break down number of object references by referent class and speaker. Number of references accompanied by gesture are in parentheses. This table excludes references to the driver, copilot and the car they are driving in.

	Building or Public Space	Person or Vehicle	Road or Driveway	Traffic Signal	Other	Total
Copilot	1083 (122)	654 (15)	599 (224)	186 (46)	234 (66)	2756 (473)
Driver	777 (63)	540 (5)	187 (55)	31 (5)	136 (26)	1671 (154)
Total	1860 (185)	1194 (20)	786 (279)	217 (51)	370 (92)	4427 (627)

3.4 Navigation Discussion Units

Navigation discussion units (NDUs) are sections of discourse that contain the initial presentation and grounding dialog of a single navigation action. This is a domain-specific example of a grounding discourse unit [26]. The choice of the NDU was based on the idea that low-level navigation actions are the primary pieces of information that needs to be grounded. This intuition proved useful, as 63% of Navigation domain words could be annotated as belonging to an NDU grounding one of the main primitive actions: Go To, Leave, or Stop At. Another 5% of Navigation words were part of an NDU describing some other navigation action, and 17% were discussion about setting a destination. Detailed analysis of the remaining navigation dialog has not been done, but there were several cases of pointing out landmarks and announcing task completion or other reflections on the task. Each of the three main NDU types is grounded to the specific section of drive-able area where the action under discussion is to take place. For example, an NDU where the copilot tells the driver to 'turn left here' would usually be marked as Go To, with the grounded parameter set to the section of road to the left of the upcoming intersection. This is to enable our later analysis to examine the interaction between dialog and execution, which can be traced from GPS data.

More detail of the structure of NDUs in the corpus is in the following section, but Table 6 shows an example of one of the most typical NDUs in the corpus, and Table 7 shows a more interesting example where two NDUs are interleaved.

Table 6 The most common type of NDU in the data contains only a single utterance - a direction from the copilot.

Speaker	Transcript
Copilot	go straight here

Table 7 A more interesting situation in which a previously grounded NDU is re-presented as a reminder by the copilot. There were 24 pairs of overlapping NDUs in the corpus (48 total), making up 3% of the total annotated NDUs.

NDU ID	Speaker	Transcript
1	Copilot	and a right at the next stop sign
1	Driver	alright
2	Driver	so we go in here or not
2	Copilot	yes
2	Driver	we do
1	Copilot	let's turn right here
1	Driver	okay

3.5 Dialog Acts

Within each NDU, words were broken up into dialog acts to analyze the discourse structure in more depth. The set of dialog acts contains a standard mix, with the addition of the domain-specific "Request Direction"; Direct, Offer, Request Direction, Ask Clarification, Give Clarification, Reject, Acknowledge / Confirm, Other. Along with the label, we also recorded weather or not a gesture was present and contributed to the meaning of the dialog act. Since all our annotations were at the word level, this annotation missed dialog acts that were purely gesture with no spoken component. Table 8 shows the most common DA sequences composing an NDU. In the majority of cases, only a small amount of grounding discussion is required.

Table 8 The most common sequences of dialog acts composing an NDU

Number of Samples	Sequence
530	Direct
300	Direct, Ack
65	Direct, Give Clarification
48	Direct, Give Clarification, Ack
29	Offer, Ack
27	Offer, Give Clarification
319	Remaining, 207 other DA sequences

4 Analysis

In this section, we show some early analysis of the corpus. We attempt to shed light on the relationship between the physical situation and dialog behavior. We also investigate the effects of multi-tasking, and differences across copilots and drivers.

4.1 Copilot and Driver Differences

Our initial investigations show that there are drastic differences in dialog form across copilots and across drivers given a single copilot. Table 9 shows the words per dialog act and DAs per NDU for each copilot. Plotting the distributions of these variables and others across runs reveals wide differences not just in scale but in shape. Further investigation of these differences is upcoming work, but in the next several sections we describe some general trends that have emerged.

Table 9 Words per Dialog Act by Copilot

Copilot	A	B	C	D	E	F
Number of DAs	1330	670	114	620	260	234
Avg. words per DA	4.97	5.42	4.39	4.01	3.67	4.29
Std. words per DA	3.11	3.65	3.30	3.02	2.81	3.18
Number of NDUs	760	349	71	257	116	121
Avg. DAs per NDU	1.76	1.92	1.63	2.42	2.26	1.91
Std. DAs per NDU	1.45	1.68	1.41	1.75	1.70	1.72

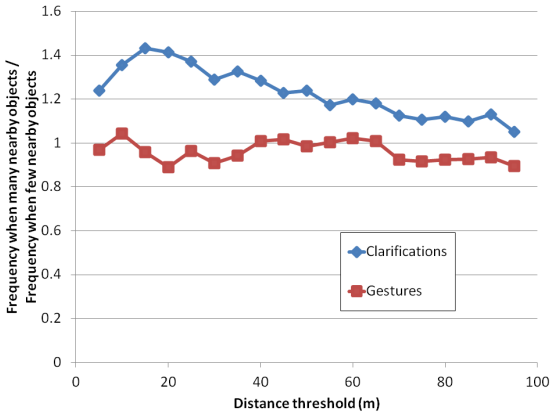

Fig. 1 The ratio of clarification/gesture frequency when many objects vs few objects are near the car during an NDU (see text for details).

4.2 Navigation Dialog and Situation Ambiguity

This corpus allows us to investigate the relationship between physical situation and dialog behavior. Here, we consider two key attributes of an NDU, whether it contains a clarification or clarification request, and whether it is accompanied by gesture. Our hypotheses related to these attributes are that physical situations that are more complex or ambiguous (such as an intersection with many roads or a location with many buildings) entail the need for more clarification and gesturing to assist in the disambiguation process.

To perform a quantitative analysis, we used GPS data and our manually annotated map of over 800 situated objects on Moffett Field to determine the complexity of a physical situation. We counted the number of objects that are within a certain radius of the vehicle (hereafter "nearby objects") at the time of a given NDU. To measure of correlation, we split the whole set of NDUs into two subsets of equal size: NDUs with fewer nearby objects than the median (high ambiguity situations), and NDUs with more nearby objects than the median (low ambiguity situations). For each subset, we compute the proportion of NDUs containing a clarification, and the proportion of NDUs containing gesture. One empirical question is, what distance threshold should we use to classify objects as "nearby"? To answer this, we

computed clarification and gesture frequency while varying the radius in 5 meter increments from 5 to 95 meters. Figure 1 shows the ratio between clarification frequency in high ambiguity situations vs low ambiguity situations. A value of 1.5 on the vertical axis indicates that clarifications are 1.5 times more likely to happen in high ambiguity situations than in low ambiguity situations. We performed a similar analysis for gestures (also on Figure 1). For clarifications, results indicate that high ambiguity situations consistently yield higher clarification rates (Y-axis value > 1) for all thresholds. This is consistent with our first hypothesis. The curve has a maximum at 15 meters, showing that, in this corpus, the density of objects within a 15-meter radius around the car is a good measure of situation ambiguity. The difference between the distributions of number of nearby objects (with a threshold of 15 meters) for NDUs with and without clarification is highly statistically significant ($p < 0.001$, using the Mann-Whitney test). No such result holds for gestures where there seem to be little correlation between our measure of situation ambiguity and gesture frequency.

4.3 Task / Dialog Interaction

To gain some insight into how the task state relates to dialog behavior, we defined a binary function $multitasking(t)$. The participants are multi-tasking according to this function if within 5 seconds of t, there are words annotated with at least two of the following task-oriented domains - Navigation, Business Search / Local Guide, Alerts / Messaging, or Scheduling. An NDU is multi-tasking if any point within the NDU is multi-tasking. In this section we compare dialog behavior between multi-tasking and non-multi-tasking situations. Table 10 shows standard statistics comparing word length, words per dialog act, and dialog acts per NDU while multi-tasking vs not multi-tasking. All of these measures of communication efficiency are lower while multi-tasking. Figure 2 shows the side-by-side histograms of how many DAs are used per NDU when multi-tasking vs. not mult-tasking. One stark difference we can observe from this is that while multi-tasking, NDUs are nearly three times more likely to last only one dialog act.

5 Conclusion and Future Work

This corpus provides a unique opportunity to do multi-modal task-based interaction in a dynamic in-car situation. Our initial annotation and analysis shows interesting trends relating physical situation to dialog behavior. Upcoming work will try to better quantify the differences and similarities between users and co-pilots, annotate 5 more runs, and gather inter-annotator agreement numbers to better understand the sources of variation.

Table 10 Word lengths, words per DA, and DAs per NDU while multi-tasking vs. not multi-tasking

	Multi-tasking	Not Multi-tasking
Number of words	4672	15218
Avg. word length (s)	.263	.284
Std. word length (s)	.307	.351
Number of DAs	543	2672
Avg. words per DA	4.64	4.72
Std words per DA	3.21	3.26
Number of NDUs	357	1317
Avg. DAs per NDU	1.53	2.04
Std. DAs per NDU (s)	1.62	1.59

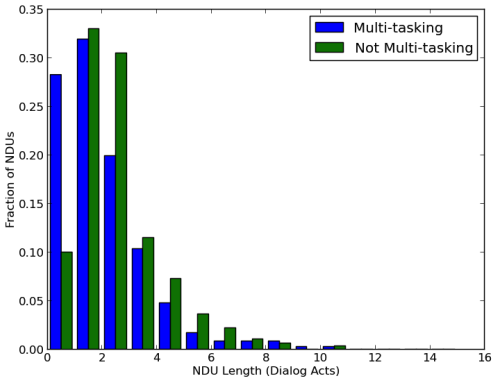

Fig. 2 The number of dialog acts per NDU while multi-tasking vs. not multi-tasking.

6 Acknowledgments

This research was performed at CMU under the sponsored research agreements 26660 and 29831 with the Honda Research Institute, USA. We would like to thank Teruhisa Misu, Rakesh Gupta and Victor Ng-Thow-Hing from HRI-USA, for their useful feedback when designing, collecting and annotating this corpus.

References

1. Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong. "you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*, 2004.
2. M. Crocker. Grounding spoken interaction with real-time gaze in dynamic virtual environments. In *International Conference on Computational Linguistics*, 2012.

3. Eric Horovitz Dan Bohus. Dialog in the open-world: Platform and applications. In *Proc. ICMI*, 2009.
4. Pattie Maes David Merrill. Augmenting looking, pointing and reaching gestures to enhance the searching and browsing of physical objects. In *Pervasive Computing*, 2007.
5. Jeff Rickel David Traum. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proc. AAMAS*, 2002.
6. Anind K. Dey. Situated interaction and context-aware computing. In *Personal and Ubiquitous Computing*, 2001.
7. G. J. M. Kruijff et. al. Situated dialogue processing for human-robot interaction. In *Cognitive Systems*, 2010.
8. Stephanie Seneff et. al. Exploiting context information in spoken dialog interaction with mobile devices. In *Proc. Intl. Workshop on Improved Mobile User Experience*, 2007.
9. Teruhisa Misu et. al. Situated multi-modal dialog system in vehicles. In *Proc. ICMI*, 2013.
10. Milica Gasic. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *Proc. Automatic Speech Recognition and Understanding*, 2011.
11. et al. H. Zender. An integrated robotic system for spatial understanding and situated interaction in indoor environments. In *Proc. AAAI*, 2007.
12. Antoine Raux Ian Lane, Yi Ma. Immersive interaction within vehicles. In *Proc. Spoken Language Technology Workshop*, 2012.
13. James Allen Joel Tetreault. Semantics, dialogue, and reference resolution. Technical report, Rochester University Dept. of Computer Science, 2006.
14. Zahar Prasov Joyce Chai. Fusing eye gaze with speech recognition hypotheses to resolve exophoric reference in situated dialogue. In *Proc. EMNLP*, 2010.
15. Jun Rekimoto Katashi Nagao. Ubiquitous talker: spoken language interaction with real world objects. In *arXiv preprint cmp-lg/9505038*, 1995.
16. Ian Lane. Cesar: The car environment sensor adjustable rig. Technical report, Carnegie Mellon University, 2012.
17. Yi Ma, Antoine Raux, Deepak Ramachandran, and Rakesh Gupta. Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 169–178. Association for Computational Linguistics, 2012.
18. Amy Isard Manuel Giuliani, Mary Ellen Foster. Situated reference in a hybrid human-robot interaction system. In *Proc. INLG*, 2010.
19. Sebastian Möller, Florian Gödde, and Maria Wolters. A corpus analysis of spoken smart-home interactions with older users. 2008.
20. C. Muller. *Fully Automatic Resolution of It, This and That in Unrestricted Multi-Party Dialog*. PhD thesis, 2008.
21. Gabriel Skantze Samer Al Moubayed. Turn-taking control using gaze in multiparty human-computer dialogue: Effects of 2d and 3d displays. In *Proc. Intl. Conf. on Auditory-Visual Speech Processing*, 2011.
22. Maxine Eskenazi Sungjin Lee. Pomdp-based let's go system for spoken dialog challenge. In *Spoken Language Technology Workshop*, 2012.
23. Masahiko Tateishi, Katsushi Asami, Ichiro Akahori, Scott Judy, Yasunari Obuchi, Teruko Mitamura, Eric Nyberg, and Nobuo Hataoka. A spoken dialog corpus for car telematics services. In *DSP for In-Vehicle and Mobile Systems*, pages 47–64. Springer, 2005.
24. Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.
25. Blaise Thomson. Training a real-world pomdp-based dialogue system. In *Proc. Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, 2007.
26. David R Traum. *A Computational Theory of Grounding in Natural Language Conversation.* PhD thesis, University of Rochester, 1994.
27. D. Lim W. M. Soon, H. T. Ng. A machine learning approach to coreference resolution of noun phrases. In *Computational Linguistics*, 2001.

28. Jason Williams. A belief tracking challenge task for spoken dialog systems. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, 2012.
29. Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France, August 2013. Association for Computational Linguistics.