Are you talking to Me? Dialogue Systems Supporting Mixed Teams of Humans and Robots

John Dowding and William J. Clancey and Jeffrey Graham

ABSTRACT

This position paper describes an approach to building spoken dialogue systems for environments containing multiple human speakers and hearers, and multiple robotic speakers and hearers. We address the issue, for robotic hearers, of whether the speech they hear is intended for them, or more likely to be intended for some other hearer. We will describe data collected during a series of experiments involving teams of multiple human and robots (and other software participants), and some preliminary results for distinguishing robot-directed speech from human-directed speech. The domain of these experiments is Mars-analogue planetary exploration. These Mars-analogue field studies involve two subjects in simulated planetary space suits doing geological exploration with the help of 1-2 robots, supporting software agents, a habitat communicator and links to a remote science team. The two subjects are performing a task (geological exploration) which requires them to speak with each other while also speaking with their assistants. The technique used here is to use a probabilistic context-free grammar language model in the speech recognizer that is trained on prior robot-directed speech. Intuitively, the recognizer will give higher confidence to an utterance if it is similar to utterances that have been directed to the robot in the past.

1. INTRODUCTION

Mars-analogue field studies are an important component in the planning for human exploration of the Moon or Mars, to evaluate new technologies and work practices for planetary exploration[3]. In these studies, two subjects play the role of astronauts performing Extra-Vehicular Activities (EVAs), together with one or two robotic assistants, while a third subject plays the role of Habitat Command (HabCom), overseeing the EVA from the habitat. The astronaut subjects wear simulated space suits that provide weight, sensory and mobility limitations, and a portable computing platform containing a laptop computer, wireless network receiver, and GPS and bio-medical sensors. The astronauts communicate with each other, with HabCom, and with the field-test support staff using radios.

The mobile autonomous robots used in the field have been provided by NASA/Johnson Space Center's Automation, Robotics and Simulation Division [10]. The robots are equipped with high-mobility components, cameras and other sensors (including GPS for localization), and on-board laptops running the software systems. The robots are named "Boudreaux" and "Thibodeaux", and each has his own capabilities and uniquely distinguishable synthesized voice.

Planetary exploration is just one of a growing number of challenges that are likely to require cooperating teams of humans and robots, including applications to search and rescue[19], and the military[12]. Spoken dialogue interfaces are a natural choice for many human-robot applications[25]. Spoken dialogue is especially critical in the domain of planetary exploration, due to the mobility limitations imposed by pressurized space suits and gloves[28, 2].

The vast majority of prior work in spoken dialogue systems, both commercial applications and research, have primarily been concerned with situations in which a single user is speaking to a dialogue system, and where all speech from the user is understood to be directed towards that dialogue system[1, 5, 17, 24, 27]. Given the inevitability of speech recognition errors, these systems emphasis robustness[11, 29, 18]. Any speech that is not responded to indicates a communication failure between the user and system.

In contrast to these applications, communication in mixed human-robot teams will take place both between the robots and the humans, and amongst the humans themselves. Data from our field tests (see Section 3) indicates that approximately 60% of the speech data is between the human participants, making human-human speech the dominant condition. In this condition, a failure to respond can be completely felicitous.

The primary technical challenge in this paper then is to distinguish between the human-human speech that the dialogue system hears, and the human-robot speech, with the dialogue system hopefully only responding to the human-robot turns. We have considered two alternatives to addressing this classification problem. The first is to require the user to send a physical signal (button-push or key-stroke) to indicate to the robot when they are talking to it. The disadvantages of this alternative are that it requires extra effort from the user, especially given the limited hand mobility in the space suit glove, and that the geological survey and rock sampling tasks frequently require that the hands remain free. The second alternative we considered is to require the user to speak a special phrase, such as *Computer* on Star Trek, to indicate to the computer when it is being spoken to. We did adopt this approach in part, since the robots are sometimes addressed by name, *Boudreaux* or *Thibodeaux*) (see Section 2). But, requiring such an address before each human-robot turn seems likely to increase the cognitive burden on the user, and

Planning
Start walking to way point one activity.
What is my current activity?
What is my next activity?
How much time is left?
What time is it?
Change duration to twenty minutes.
Navagation
Where is way point one?
Where is Thibodeaux?
Name this location Work Site Two.
What locations are near me?
Track my location every twenty seconds.
Science Data
Create a new sample.
Download an image from my camera.
Label it image one.
Associate it with the sample bag.
Record a voice note.
List sample bags.
Associate the voice note with sample bag one.
Play the voice note associated with sample bag one.
Robot Commands
Boudreaux come here.
Thibodeaux move to Work Site Two.
Follow me.
Watch Astronaut Two.
Halt.
Do you still have network connectivity?
Take a picture of Astronaut Two.
Take a panorama and label it image at Work Site Two.
Print a curation label for sample bag one.
Thibodeaux team with Astronaut Two.
Other
Say that again.
Can you hear me?
Increase volume.
Shut up.

Table 1: Dialogue System Examples

force the user to repeat themselves when they inevitably forget to begin the turn with the required phrase.

2. DIALOGUE SYSTEM

The dialogue system acts as a front-end to a software architecture for supporting surface EVA operations[4]. The capabilities provided by this architecture include monitoring the astronauts physical status through bio-medical sensors, monitoring the progress through the scheduled activities that comprise the EVA, monitoring the locations of all participants using GPS sensors, assisting with navigation, assisting the astronaut in collecting, describing, and logging science data (rock samples, voice annotations, digital images, etc.) with time and GPS stamps, and commanding the robotic assistants (Boudreaux and Thibodeaux). The dialogue system provides access to about 90 functions supporting these activities. Some examples of language supporting these functions are given in Table 1. Initiative in this system usually stays with the human user, although the dialogue system may ask yes/no questions for confirmation, and a robot may ask a yes/no question if it is trying to balance conflicting goals (e.g. when it has been asked to follow the astronaut, and also asked to maintain network connectivity, and observes that network throughput is decreasing).

Date	Speech Data
Sept. 2-13, 2002	3,695
March 31-April 11, 2003	4,659
April 26-May 7, 2004	11,079
April 4-15, 2005	5,933

Table 2: Amount of Speech Data Collected

The dialogue system uses an architecture that has been used for a number of deployed spoken dialogue systems[17, 22, 14]. This architecture combines independent software components using the Open Agent Architecture[15]. The components included in this application include the Nuance speech recognizer[20], the Festival speech synthesizer[8], the Gemini language interpretation and generation system[6], the Brahms agent modeling and simulation environment[26], and a dialogue manager. The custom software components required to build this spoken dialogue system are the grammar and lexicon (approximately 1250 words) developed in Gemini and compiled into a Nuance grammar, and the dialogue manager.

The language model used in the Nuance speech recognizer for this application was developed using the techniques of [7] in Gemini. Initially, a grammar is developed in Gemini in a typed unification-grammar (TUG) representation, which is then compiled into a Context-Free Grammar (CFG) in Nuance's Grammar Specification Language (GSL) form. This is in-turn converted to a Probabilistic Context-Free Grammar (PCFG) using training data and Nuance's COMPUTE-GRAMMAR-PROBS tool. Using the more compact TUG formalism allows us to represent a large CFG using a small number of rules. In this application, the grammar in TUG formalism comprised 75 Gemini rules, while the resulting CFG contained 4,368 rules. More details about how the language model is trained will be given in Section 4.

3. MARS-ANALOGUE FIELD TESTS

The spoken dialogue system we are developing has been used in 4 Mars-analogue field tests over the past 4 years. Each field test is two-weeks long, and takes place in harsh environments on Earth to simulate harsh environments on the moon or Mars. These field tests have been conducted as part of the Mobile Agents project at NASA's Ames Research Center (Principle Investigator: William Clancey), and in cooperation with the Mars Society. The most recent field tests have been conducted at the Mars Society Desert Research Station[16] in south-eastern Utah. The speech recognition experiments described in Section 4 were carried out using speech data from the 2003 and 2004 field tests. The 2002 data was omitted since it was collected in a single speaker scenario[1] A set of 14,103 utterances will be used for these experiments, of which 8,389 utterances (60%) is human-human speech, and 5,714 utterances is human-robot speech (40%). Figure 1 shows a frequency distribution of utterance length (number of words) for the human-human subset of the corpus, with a mean utterance length of 6 words. Figure 2 shows a similar frequency distribution of the human-robot utterances, mean length of 3.8 words.

The speech data has a few characteristics that distinguish it from

[1]The 2005 data, and some of the 2004 data, is still in the process of being transcribed.

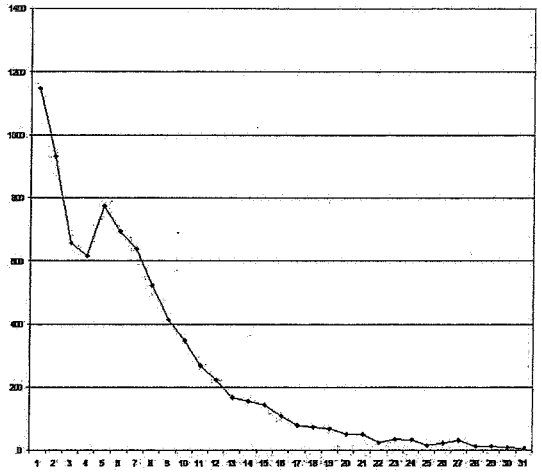

Figure 1: Word-length Distribution of Human-Human Utterances

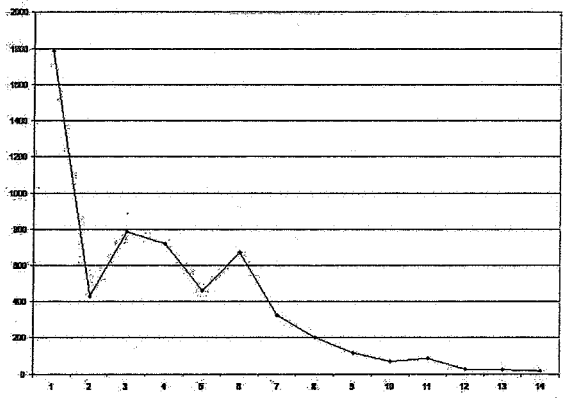

Figure 2: Word-length Distribution of Human-Robot Utterances

most existing speech corpora. Although the speech data includes data from 10 different speakers, the vast majority of the data comes from just 3 speakers, the subjects primarily playing the roles of Astronaut 1, Astronaut 2, and HabCom. Since these same 3 speakers have participated in 3 of the field tests, they are now expert users of both the spoken dialogue system, and the underlying surface EVA operations intelligent agents. These speakers are also typically able to hear each other interacting with their dialogue systems, which may bias them towards using the same language. These characteristics may tend to reduce the perplexity of the collected data.

	Train	Test
# Human-Human	4,194	4,195
# Human-Robot	2,854	2,860
Word Error	6.5%	6.5%
False Accept	8.8%	9.8%
False Reject	9.7%	9.6%
xCPU RT (Human-Human)	0.63	0.62
xCPU RT (Human-Robot)	0.39	0.37

Table 3: Recognition Results

4. RECOGNIZER PERFORMANCE RESULTS

The speech recognition experiments proceeded by first splitting the available data into training and test sets, approximately 50-50. To access recognizer performance, we use three accuracy metrics:

- False Accept rate – The percentage of human-human utterances that are inappropriately responded to.
- False Reject rate – The percentage of human-robot utterances not responded to.
- Word Error rate – The percentage of word recognition errors on the non-rejected human-robot utterances.

In addition to the error metrics, we also provide performance metrics to give an indication of the computational intensity of this task [2]. The xCPU RT (percentage of CPU real time) is an indicator of how long it took the CPU to recognize each utterance, compared to how long (in seconds) the utterance was. Any number less than 1 indicates that the recognizer is recognizing faster than the speaker is speaking, and indicates real-time performance.

The recognition results are summarized in Table 3, showing that False-Accept and False-Reject rates of under 10% were achieved, in both the training and test conditions. The word error rate is somewhat higher than experienced on tasks of similar vocabulary size, above a target range of 3-5%[17]. While the xCPU RT metric shows that performance on both human-human and human-robot speech was faster than real-time, the performance on human-human speech was nearly twice slower.

[2]Experiments were conducted on a 3.2GHz Pentium 4 computer with 2GB memory.

4.1 Dialogue Context

The dialogue manager can act as a secondary filter on the true false accept rate of the system, rejecting any utterances that are not pragmatically interpretable in the current discourse context. Figure 3 shows the distribution on word-length of the utterances that get falsely accepted by the speech recognizer in the test data, showing that the likelihood of being falsely accepted drops quickly as utterance length increases. Intuitively, shorter utterances are harder to classify correctly. This appears to be due to two factors: first that shorter utterances are more likely to accidentally match in-domain utterances (e.g. misrecognitions of "what" as "halt"), second that correctly-recognized short utterances like "yes" can be equally well addressed to either the robot or another human. Of these, many of the short utterances can be rejected by the dialogue manager when they are misrecognitions of affirmative or negative responses ("yes", "no", "right", "okay", etc.), and when the dialogue manager has not recently asked a yes/no question. The dialogue manager can also reject longer utterances when they contain pronouns or definite descriptions that cannot be resolved in context.

Unlike the language model, the dialogue manager cannot be stati-

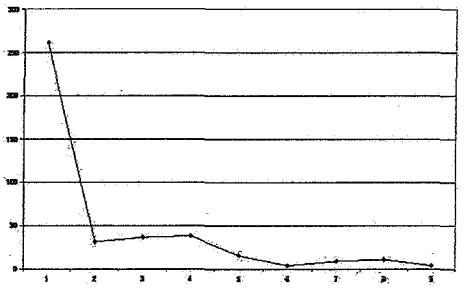

Figure 3: Word-length Distribution of False Accept Utterances

cally trained on training data and tested on a held-out test set. In a real-life field test, the speech recognition performance will affect how the dialogue proceeds, and in-turn effect dialogue context at each turn. Since the dialogue manager is built by-hand for this domain, it is essentially held constant before the field test each year. In order to evaluate the effectiveness of the dialogue manager as a secondary filter, we analyzed all of the falsely accepted utterances in the '04 data (both training and test). We found that, of the 516 human-human utterances that were falsely accepted by the speech recognizer, only 248 (48%) resulted in actions being taken by the dialogue manager. Taken as a percentage of all the human-human utterances in the 04 data, the combined false accept rate experienced during the 04 field test was 248/5074 (4.9%). It is important to note that no robot moved incorrectly during any of the speech commanding.

5. RELATED WORK

Paek, Horvitz and Ringger[21] address the issue of continuous listening in a multi-human environment containing a mix of computer-directed and human-directed speech. Their target application is a speech-based interface to a system that can control a Microsoft PowerPoint presentation, moving backwards and forwards through the slides by voice command. They report promising results on a small dataset (N=16), showing that their system rarely responded to a human-directed utterance. In this experiment, there was only a single speaker, but multiple listeners.

Two recent papers address the issue of "cross-talk", when a single-user spoken dialogue system is unintentionally exposed to speech that is directed to another hearer, frequently the person running the experiment. Gabsdil and Lemon[9] report a dialogue system that is sensitive to the possibility that a certain percentage of the speech that it hears may be other-directed speech (or non-speech sounds), and the dialogue system should decide whether to accept, reject, or ignore a user utterance. They report that a combination of recognition confidence features and pragmatic plausibility features (computed from dialogue context) improves the ability of the system to reject utterances that are out-of-grammar or crosstalk. Renders, Rayner, and Hockey[23] report on a experiment going beyond the utterance-based recognition confidence scores used in this paper to using word-based confidence scores. Their intuition is that not all word recognition errors are equally likely, and that some errors are more likely to indicate cross-talk than others. They train a Support Vector Machine based on these word confidence scores to classify cross-talk. They also experiment with an alternative cost function, allowing for the possibility of treating a falsely-accepted utterance as having a higher cost than a falsely-rejected utterance.

This work differs from the above work in several respects. We are considering contexts in which two speakers are engaged in a task requiring them to speak with each other, so the human-directed speech is not an anomalous condition, but is the dominant condition. The subjects are also engaged in a real-world task, not an academic study, so we have some confidence that the distribution of human-human versus human-robot speech may be representative of other real-world tasks.

6. CONCLUSIONS AND FUTURE WORK

We have described a spoken dialogue system that operates in a mixed human-human/human-robot environment, where the human-human speech is dominant, about 60% of all speech. We have described some preliminary speech recognition results showing word-error rates in the vicinity of 6.5%, with false accept and false reject error rates below 10%.

It is clear, both from this work, and the work of Gabsdil and Lemmon [9] that dialogue context provides valuable information for distinguishing between human-directed and robot-directed speech, and likely of addressee recognition more generally [13]. We plan to explore this, and the impact of non-dialogue forms of context, by building targeted context-specific language models for different contexts. We plan to address three contexts in the near term: the context in which the system has just asked the user a yes/no question, the context in which one or more of the robots is in motion, and the context in which the astronaut is themselves in a traveling task. The first context is motivated by the observation that many of the current false accepts are misrecognitions of affirmative or negative responses. The second context is motivated by the observation that another class of frequent false accepts are misrecognitions of the short commands "halt" and "stop". The third context is motivated by the observation that the astronaut-subjects appear to stay more on-task during science-data collecting activities, but are more likely to go off-task during traveling activities, and that this is likely to impact the distribution of human-directed and robot-directed speech.

7. REFERENCES

[1] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. Towards conversational human-computer interaction. *AI Magazine*, 2001.

[2] J. Braden and D. Akin. Development and testing of a space suit analogue for neutral buoyancy eva research. In *32nd International Conference on Environmental Systems*, San Antonion, TX, July 2002.

[3] W. Clancey, P. Lee, and M. Sierhuis. Empirical requirements analysis for mars surface operations using the flashline mars artic research station. In *Proceedings of the Fourteenth International FLAIRS Conference*. AAAI Press, 2001.

[4] W. Clancey, M. Sierhuis, R. Alena, D. Berrios, J. Dowding, J. Graham, K. Tyree, R. Hirsh, W. Garry, A. Semple, S. Buckingham Shum, N. Shadbolt, and S. Rupert.

Automating capcom using mobile agents and robotic assistants. In *American Institute of Aeronautics and Astronautics 1st Space Exploration Conference*, Orlando, FL, 2005.

[5] D. Dahl, M. Bates, M. Brown, K. Hunicke-Smith, D. Pallet, C. Pao, A. Rudnicky, and E. Shriberg. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ, 1994.

[6] J. Dowding, M. Gawron, D. Appelt, L. Cherny, R. Moore, and D. Moran. Gemini: A natural language system for spoken language understanding. In *Proceedings of the Thirty-First Annual Meeting of the Association for Computational Linguistics*, 1993.

[7] J. Dowding, B. A. Hockey, C. Culy, and J. M. Gawron. Practical issues in compiling typed unification grammars for speech recognition. In *Proceedings of the Thirty-Ninth Annual Meeting of the Association for Computational Linguistics*, 2001.

[8] Festival. *The Festival Speech Synthesis Systems*. http://www.cstr.ed.ac.uk/projects/festival, 2005.

[9] M. Gabsdil and O. Lemon. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of Forty-Second Annual Meeting of the Association for Computational Linguistics*, 2004.

[10] R. Hirsh and e. a. Jeffrey Graham. Intelligence for human-assistant planetary surface robots. In *Intelligence for Space Robotics*, May 2006.

[11] E. Jackson, D. Appelt, J. Bear, R. Moore, and A. Podlozny. A template matcher for robust nl interpretation. In *Human Language Technology Workshop*, pages 37–42, 1993.

[12] F. Jentsch, A. Evans, M. Feldman, R. Hoeft, S. Rehfeld, and M. Curtis. A scale mout facility for studying human-robot interactions. In *24th Army Science Conference Procedings*, 2004.

[13] N. Jovanovic and R. den Akker. Towards automatic addressee identification in multi-party dialogues. In *SIGDIAL*, 2004.

[14] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters. Multimodal dialogues with intelligent agents in dynamic environments: The WITAS conversational interface. In *Proceedings of 2nd Meeting of the North American Association for Computational Linguistics*, 2001.

[15] D. Martin, A. Cheyer, and D. Moran. Building distributed software systems with the open agent architecture. In *Proceedings of the Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, Blackpool, Lancashire, UK, 1998.

[16] MDRS. http://www.marssociety.org/mdrs, 2005.

[17] R. Moore, J. Dowding, H. Bratt, J. Gawron, Y. Gorfu, and A. Cheyer. CommandTalk: A spoken-language interface for battlefield simulations. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 1–7, 1997.

[18] G. V. Noord, G. Bouma, R. Koeling, and M. Nederhof. Robust grammatical analysis for spoken dialogue systems. (1):45–94, 1999.

[19] I. Nourbakhsh, K. Sycara, M. Koes, M. Young, M. Lewis, and S. Burion. Human-robot teaming for search and rescue. *IEEE Pervasive Computing*, (1):72–78, January-March 2005.

[20] Nuance. http://www.nuance.com, 2005.

[21] T. Paek, E. Horvitz, and E. Ringger. Continuous listening for unconstrained spoken dialog. In *Proceedings of ICSLP-2000*, Beijing, 2000.

[22] M. Rayner, B. Hockey, and F. James. A compact architecture for dialogue management based on scripts and meta-outputs. In *Proceedings of ANLP 2000*, 2000.

[23] J. Renders, M. Rayner, and B. Hockey. Kernel methods for identification of cross-talk and misrecognition. In *(under submission)*, 2005.

[24] S. Seneff, E. Hurley, C. Pao, P. Schmid, and V. Zue. Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[25] C. Sidner, C. Kidd, C. Lee, and N. Lesh. Where to look: a study of human-robot engagement. In *ACM International Conference on Intelligent User Interfaces (IUI)*, 2004.

[26] M. Sierhuis, W. Clancey, and R. Van Hoof. Brahms: a multiagent modeling and environment for simulating social phenomena. In *First Conference of the European Social Simulation Association (SIMSOC VI)*, Groningen, The Netherlands, 2003.

[27] W. Ward and S. Issar. The cmu atis system. In *Spoken Language System Technology Workshop*, pages 249–251, 1995.

[28] M. Welsh and D. Akin. The effects of extravehicular activity gloves on human hand performance. In *31st International Conference on Environmental Systems*, Orlando, FL, July 2001.

[29] K. Worm. A model for robust processing of spontaneous speech by integrating viable fragments. In *17th International Conference on Computational Linguistics*, pages 1403–1407, 1998.