

# A Multi-Modal Approach for Natural Human-Robot Interaction

Thomas Kollar, Anu Vedantham, Corey Sobel,  
Cory Chang, Vittorio Perera and Manuela Veloso

Carnegie Mellon University  
5000 Forbes Ave.  
Pittsburgh, PA 15213

**Abstract.** We present a robot that is able to interact with people in a natural, multi-modal way by using both speech and gesture. The robot is able to track people, process speech and understand language. To track people and recognize gestures, the robot uses an RGB-D sensor (e.g., a Microsoft Kinect). To recognize speech, the robot uses a cloud-based service. To understand language, the robot uses a probabilistic graphical model to infer the meaning of a natural language query. We have evaluated our system in two domains. The first domain is a robot receptionist (roboceptionist); we show that the roboceptionist is able to interact successfully with people 77% of the time when people are primed with the capabilities of the robot compared to 57% when people are not primed with its capabilities. The second domain is a mobile service robot, which is able to interact with people via natural language.

## 1 Introduction

Our aim is to make robots that can interact with people in a natural and intuitive way and can intelligently understand and respond to the intentions of humans. In this paper, we develop an approach for human-robot interaction that uses speech and gesture to create a rich space of verbal and non-verbal meaning. Our aim is to craft a set of gestures that are intuitive for a human while at the same time understandable by a robot, and which can be combined with speech input to improve the experience of people interacting with social robots. This is a challenging problem because understanding the intentions of a human may require reasoning over the combination of a large space of gestures and natural language queries.

In this paper we address this challenge by developing a robot that is able to combine multi-modal input to understand gestures and task-constrained speech. Specifically, the robot is able to recognize a hand-raising gesture, detect when people are approaching and determine when the nearest person is facing the robot. If the person is not close enough for the interaction to be successful, it will guide the person to move to more successful states. By guiding the people to these states, the robot is able to better understand task-constrained natural language queries that involve the robot's status as well as the location of people in a building. This is achieved by learning a probabilistic graphical model that connects the natural language to the type of question being asked.

We evaluate our approach in two domains. The first domain is a roboceptionist that has been placed in the entryway of a building. The second domain is a mobile service robot that can perform tasks such as delivering objects and escorting people in indoor environments [1]. In the roboceptionist domain, ten different people interacted with the robotic system and asked it questions; we evaluated the results in two conditions. In the first condition, people were not primed about the capabilities of the robot and could interact with it in any way that felt natural to them. In the second condition, people were primed by telling them what the robot could do. In all conditions, we found that the roboceptionist was able to recognize the desired gestures and track people effectively nearly all of the time. In addition, we found that the roboceptionist was able to understand greetings, questions about its capabilities and questions about the location of people's offices 77% of the time when people were primed about the robot's capabilities. Interactions with un-primed people were only successful 57% of the time, indicating that priming (or long-term interaction) is an important consideration for robotic systems. We also evaluated the performance of the mobile service robot at understanding commands and found that it had successful interactions with people 73% of the time.

## 2 Related Work

Others have designed robots for long-term social human-robot interaction. Gockley et al. [2] designed a robot receptionist named Valerie to investigate human-robot social interaction and found that after nine months of deployment, people still interacted with the robot on a daily basis. This robot now interacts with people in multiple languages [3] and is aware of the presence of a person, but is not able to recognize gestures. Salichs et al. [4] developed a robot to cover a wide spectrum of human-robot social interaction and was able to engage in both physical and speech driven events.

Work in human-robot interaction will often use gestures. Bohus and Horvitz [5] developed a system capable of differentiating speakers in a turn-based speaking environment. The system was able to determine who was speaking to whom by evaluating hand gestures and other cues. Wu and Huang [6] describe static gesture recognition of hand pose as well as and temporal gesture modeling. Eisenstein [7] shows how the use of both hand gesture and speech can be used to resolve ambiguous noun phrases, the segmentation of speech into topics and the production of keyframe summaries of spoken language. Christoudias et al. [8] investigates whether reliable audio-visual classifiers can be learned through co-training. Scassellati [9] describes a model of joint attention in the context of recognizing social cues, showing the role of imitation and gesture in terms of head movements. The interface for the roboceptionist takes inspiration from other social robots, such as Kismet [10] and interaction paradigms are related to Mutlu et al. [11], who describe how a robot can establish roles by using gaze cues.

### 3 Approach

In this paper, we present a robot that interacts with people in a natural way by understanding multi-modal input. The approach uses speech, gesture, and an animated character to communicate with people and understand their intentions.

#### 3.1 Gesture and Event Recognition

Gestures are a natural interface for people that want to interact with robots. To recognize gestures and track people, the robot uses an RGB-D camera. The information extracted from the camera is used to interact with people and determine when to ignore them if they are uninterested in the robot.

**Interactions via Gesture** Since many people walk by the robot, it needs a way to differentiate people showing interest from those that are uninterested in it. To address this challenge, we developed an approach to recognize a hand-raise gesture to indicate the start of an interaction. When the person’s hand is up and the person is far away, the robot will ask people to come closer. When the person’s hand is up and the person is nearby, then the robot will initiate an interaction.

The robot uses the skeleton provided by the RGB-D camera<sup>1</sup>, as in Figure 1. The position of the hip  $\hat{h}$ , head  $\hat{d}$  and the hand  $\hat{n}$  are used to determine whether the hand is raised. If we define the hip/head vector to be  $\hat{v}_{h \rightarrow d} = \hat{d} - \hat{h}$  and the hip/hand vector to be  $\hat{v}_{h \rightarrow n} = \hat{n} - \hat{h}$ , then the projection of the hip/hand vector onto the hip/head vector is:

$$p = \frac{\hat{v}_{h \rightarrow n} \cdot \hat{v}_{h \rightarrow d}}{|\hat{v}_{h \rightarrow d}|} \frac{\hat{v}_{h \rightarrow d}}{|\hat{v}_{h \rightarrow d}|} \quad (1)$$

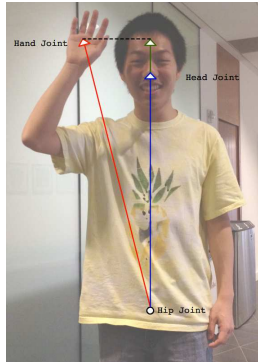
If the norm of  $p$  is greater than the norm of  $\hat{v}_{h \rightarrow d}$ , then the length of the projected vector is longer than the hip/head vector and the robot will detect the person as raising their hand. Figure 1 shows these vectors. Because this approach uses the relationship between the hip/hand vector and the hip/head vector, it is able to recognize when a hand is raised even when the orientation of the body or the RGB-D sensor not vertical.

**Interactions via Proximity** Another good indicator that a person would like to speak with the robot is whether they are approaching the robot or not. If a person is coming toward the robot or is very near to the robot, then the interaction is started. In qualitative experiments, we found that most of the people uninterested in the robot walk away from it. To know if someone is walking towards the robot, the four closest people are detected and if the change in the position of any person indicates that they are moving toward the robot, then the robot will start interacting with them.

Concretely, we use two conditions to guarantee that a person is approaching the robot. The first condition is that the person is moving along a line with

---

<sup>1</sup> The Microsoft Kinect API was used.



**Fig. 1.** An illustration of the vectors used to compute when a hand is raised.

a constant angle with respect to the robot. If  $\theta = \arctan 2(x, z)$  (where the  $x$  and  $z$  coordinates correspond to the ground plane) is approximately the same between two different timesteps, then a person is moving toward the robot. The second criteria is that the person is moving closer to the robot. This can easily be verified by checking that  $dist(x_2, z_2) < dist(x_1, z_1)$ .

**Interactions via Orientation** Examining if a person is facing the robot also acts as a valid indicator of a desire to speak with the robot. In order to determine if a person is facing the robot, we create two vectors, one from the robot to the center of the person's chest, and another from one shoulder to the other. We know that the person is facing the robot if the dot product of these two vectors is approximately zero, since this would indicate that the two vectors are orthogonal.

If  $\hat{l}$  is the position of the left shoulder,  $\hat{r}$  is the position of the right shoulder,  $\hat{k}$  is the position of the robot, and  $\hat{c}$  is the position of the chest of the person, then we can compute the shoulder vector as  $\hat{s} = \hat{r} - \hat{l}$  and the vector from the robot to the person as  $\hat{r} = \hat{c} - \hat{k}$ . If the dot product is almost 0 ( $\hat{r} \cdot \hat{s} < \text{delta}$ ), then the person is facing the robot.

**Interactions when Multiple People are Present** Using the above components, we can synthesize an algorithm that decides who to speak with and when to speak with them. First, the robot finds the four closest people, chooses people who are facing the robot and responds to the closest person with their hand raised. If a person is raising his or her hand, but is not close enough to speak with the robot, it will ask the person to move closer. Otherwise, the robot will begin processing speech.

### 3.2 Dialog

The robot is able to have social dialog as well as provide navigational assistance in the Gates-Hillman Center. The dialog with people is centered around

inferring a question type from the natural language query, and responding with an appropriate answer. We use the speech recognition provided by an Android tablet (the primary interface to the robot), which does not use a fixed grammar. Queries that the robot is able to understand include:

1. Basic Greetings (e.g., “Hello, how are you?”)
2. Queries about itself (e.g., “What can you do?”, “What is your name?”)
3. Location of people in the CS department (e.g., “Where is Joydeep Biswas’ office?”)
4. Location of landmarks (e.g., “Where is the cafe?”, “Any nearby cluster?”, “How about the nearest bathroom?”)

To understand natural language questions, we take the top 10 results from the speech recognizer and infer the most likely interpretation using a probabilistic graphical model. If  $q$  is a multinomial distribution over the question type and  $s$  is the input sentence, then the goal is to compute:

$$\arg \max_q p(q|s) \quad (2)$$

By breaking down a sentence into words  $w_i$  and using Bayes rule, we can rewrite this probability distribution as:

$$p(q|w_1, \dots, w_n) = \frac{[\prod_{i=1}^n p(w_i|q)] p(q)}{p(w_1, \dots, w_n)} \quad (3)$$

The words are assumed to be independent of each other since our goal is to retrieve the overall meaning for a sentence. Thus, we can compute the above terms as:

- (1) Computing  $p(w_j|q)$

If  $w_j$  is a word and  $c_j$  is the count of word  $j$  for the category  $q$  which has  $|w|$  total word occurrences, then this term is computed as:

$$p(w_j|q) = \frac{c_j + 1}{[\sum_j c_j] + |w|} \quad (4)$$

- (2) Computing  $p(q)$

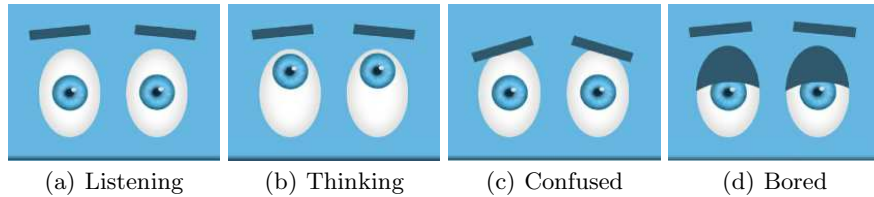
If  $c_q$  is the number of times question  $i$  was asked and  $|q|$  is the number of question types, then this term is computed as:

$$p(q) = \frac{c_q + 1}{[\sum_i c_i] + |q|} \quad (5)$$

In order to handle cases where there exists a word in the input that was not present in the training corpus, we have employed additive (Laplace) smoothing and add one occurrence of all words and questions. The resulting model is a Naive Bayes Classifier with additive smoothing [12]. To train the model, we collected query/category pairs by asking people what questions they would ask



**Fig. 2.** The interface has eyes that follow the closest tracked person.



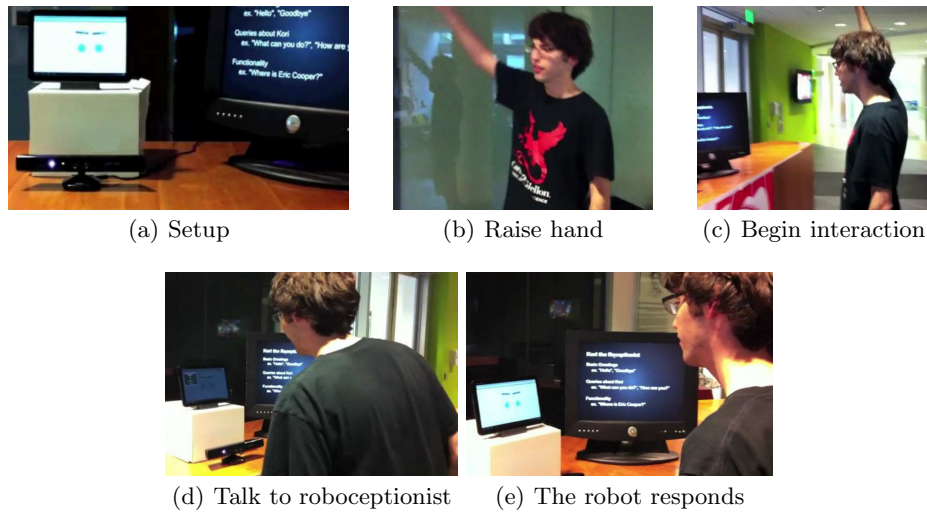
**Fig. 3.** Different expressions modeled for the roboceptionist

the receptionist. We formulated seven different situations (e.g., greeting, status, capabilities, directions) and asked 26 people to write down how they would interact with the robot via natural language, resulting in 182 training sentences. These sentences were as short as one word (e.g., “hello”) or as long as 14 words. Each sentence was annotated with a corresponding query type.

When a query type is a greeting, a question about the robot’s capabilities or status, or goodbye, the robot responds to the person using a set of fixed responses. When the query type is about the location of a person, the robot will look up the location of this person by accessing a database of people in the School of Computer Science, including name, email and office number. The language understanding system matches names from the natural language query to this database by string matching. If any of the names exactly match a full name in the database, then the robot responds with the office location (e.g., “Joydeep Biswas is in room 7707.”). If not, the system tries to match all the last names to the query. If there are no matches to the last name, then it will try to match first names. The speech recognition typically finds the requested person in one of the top two results. When there are multiple people that a name could refer to, the robot informs the person of the top candidate office locations.

### 3.3 Robot Interface

As a part of the roboceptionist, we have created an approachable interface that displays some realistic characteristics and an easily recognizable appearance. Specifically, the interface consists of two eyes that are able to track the closest human to the robot. Figure 2 shows the eyes tracking a person moving from left to right. The eyes take on a cartoonish appearance to emphasize approachability, recognizing that face-to-face contact between the robot and humans is a key



**Fig. 4.** A typical interaction. (a) contains the basic setup of the roboceptionist, including a Microsoft Kinect and an Android tablet. (b) shows an example interaction where the person raises their hand to catch the attention of roboceptionist. If user is too far away, the roboceptionist will ask the person to come closer. In (c), when user is close enough, he/she can raise their hand to begin talking to roboceptionist. In (d) the person talks to roboceptionist. In (e), the roboceptionist responds. . We would like to thank Lydia Utkin and Feiyang Feng for their help designing the user interface.

factor in creating natural human-robot interaction. Some example expressions, which depend on the state of the robot, can be seen in Figure 3.

### 3.4 System Overview

The overall interaction with the roboceptionist can be seen in Figure 4. In Figure 4(a), the roboceptionist finds and tracks the person. The eyes will follow the person as they move to the right and left of the robot. In Figure 4(c), the robot responds to a person if he or she is facing the robot and raising his or her hand to demonstrate a desire to initiate a conversation. In Figure 4(d), the roboceptionist will listen for input from a person, compare the speech to the current set of questions and then respond accordingly. Free-form speech recognition is performed using a cloud-based service on the Android tablet. The user indicates a desire to leave by saying some phrase equivalent to “good bye.” The interaction with a mobile service robot proceeds in the same way but does not use gestures as a part of the interaction.

## 4 Illustrative Results

We have performed an analysis of both the roboceptionist and the mobile service robot. Specifically, the roboceptionist’s performance at recognizing gestures and

understanding language was evaluated when a single person was present. To test the recognition of gestures on the roboceptionist, we recorded 76 interactions from 10 different individuals. These people varied in height, ranging from 5 feet to 6 feet tall. We observed these interactions in two separate modes. In the first mode, people initiated interactions with the roboceptionist in whatever way was natural for them. In this mode, the roboceptionist would fail to recognize gestures only when it failed to detect the person. The system failed to detect a person 12.5% of the time, primarily when a person was closer than 1 meter to the RGB-D sensor. When testing the RGB-D sensor with people approaching the robot between 4 to 6 feet away, it failed to detect a person only when they were partially out of view of the camera (this happened very infrequently). The primary failure mode was to miss a hand-raising gesture. Another failure mode happened when a person raised their hand, but would speak before the roboceptionist started listening. Since people would repeat their earlier request, this failure mode would generally resolve itself.

Second, we evaluated the roboceptionist by having 10 people interact with it. For example, the interaction in Figure 4 consists of the following turns:

- *User approaches the roboceptionist and raises hand.*
- Roboceptionist: You must come closer.
- *User walks closer to the roboceptionist and raises hand again.*
- Roboceptionist: Hello, speak to the Kinect.
- User: Hello.
- Roboceptionist: Hi, How are you today?
- User: Fine.
- Roboceptionist: That's good. Ask a question.
- User: Where is Stephanie Rosenthal?
- Roboceptionist: Stephanie Rosenthal has an office in room 7505.
- User: How about Joydeep Biswas?
- Roboceptionist: Not too sure what you said, please repeat.
- User: Joydeep Biswas.
- Roboceptionist: Joydeep Biswas has an office in room 7707.
- User: Goodbye
- Roboceptionist: Bye Bye
- *User walks away.*

The interactions were evaluated in two different modes. In one, people were primed by telling them they could use one of 10 question types (e.g., directions to a person's office). In a second mode people were able to interact as they wanted. In both of these modes we set no restrictions on how the queries were phrased to determine how robustly the robot understood its meaning. We performed trials with 10 different people for a total of 76 queries (e.g., "hello" in the above conversation); 54 of the queries occurred in the non-primed mode and 22 occurred in the primed mode. A successful query was one where the robot inferred the correct meaning or correctly inferred that it wasn't able to answer the query. We found that priming significantly helped the performance of the roboceptionist, as can be seen in Table 1. When it did fail, the speech recognition generally did not interpret the user's request correctly or the user's request was not yet supported.



	w/o priming	with priming
Performance	56.86%	77.27%

**Table 1.** Performance of the roboceptionist as measured by the percentage of interactions that were successful from start to end.

In order to further evaluate our work we also deployed the interface on CoBot [1], an indoor service robot. The model from Section 3.2 was trained on a corpus collected during interactions with ten faculty and staff members in our building. During those meetings, a person queried the robot and a researcher responded with answers to their queries. Each entry in the log was labeled with one of ten different labels: *hello*, *greeting*, *what can you do*, *task information*, *task physical*, *person information*, *yes*, *no*, *thanks* or *goodbye*.

The performance of the model was evaluated using three held-out logs consisting of 37 queries. We found that 72.97% of the test queries were categorized correctly. Table 2 shows the mistakes that the system makes during evaluation.

	capabilities	no	person	thanks	yes	task
capabilities	4	0	0	0	0	0
no	0	3	1	0	1	0
person	0	0	3	0	0	0
thanks	1	0	0	3	0	0
yes	0	0	0	0	5	0
task	1	0	0	0	0	5

**Table 2.** A subset of the test queries for the service robot. The rows correspond to different ground-truth categories and the columns are the predicted categories. For example, one query that should have been “thanks” was mis-classified as “capabilities.” *Capabilities* queries are questions about the robot’s capabilities, *yes/no* are responses to a yes/no question, *person* are queries for a person in the building and *thanks* usually involves thanking the robot for doing something.

## 5 Conclusions and Future Work

In this paper, we have presented a multi-modal approach to human-robot interaction. We have demonstrated our approach on a robot receptionist and a mobile service robot. The roboceptionist is able to recognize and respond to certain gestures and natural language questions in a flexible and intuitive way. Using an RGB-D sensor, the robot is able to detect when a hand is raised to initiate a conversation, detect when a person is moving toward the robot and detect when a person is facing the robot. Our approach uses a probabilistic graphical model to connect language to an underlying meaning. We have shown that the roboceptionist interacts appropriately with people 77% of the time and that the service robot performs interacts appropriately 73% of the time.

In future work, we anticipate performing a deeper semantic analysis of a natural language query to understand multi-modal input involving speech and gesture and jointly handling uncertainty in the speech recognition, the parser, and the gesture recognition. We also anticipate multi-robot coordination to enable our mobile service robot to work in unison with the robot receptionist.

## Acknowledgments

We would like to thank Lydia Utkin and Feiyang Feng for their help designing the user interface. This work was partly supported by the Center for Computational Thinking at Carnegie Mellon University, which is sponsored by Microsoft. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

## References

1. S. Rosenthal, J. Biswas, and M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *International Conference on Autonomous Agents and Multiagent Systems*, 2010, pp. 915–922.
2. R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. Schultz, and J. Wang, "Designing robots for long-term social interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2005, pp. 1338–1343.
3. M. Makatchev, I. Fanaswala, A. Abdulsalam, B. Browning, W. Ghazzawi, M. Sakr, and R. Simmons, "Dialogue patterns of an arabic robot receptionist," in *Proceedings of Human-Robot Interaction*, 2010, pp. 167–168.
4. M. Salichs, R. Barber, A. Khamis, M. Malfaz, J. Gorostiza, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and D. Garcia, "Maggie: A robotic platform for human-robot social interaction," in *IEEE Conference on Robotics, Automation and Mechatronics*, june 2006, pp. 1–7.
5. D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *International Conference on Multimodal Interfaces*, 2010, pp. 5:1–5:8.
6. Y. Wu and T. S. Huang, "Vision-based gesture recognition: A review," in *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, 1999, pp. 103–115.
7. J. Eisenstein, "Gesture in automatic discourse processing," Ph.D. dissertation, 2008.
8. C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, "Co-adaptation of audio-visual speech and gesture classifiers," in *International Conference on Multimodal Interfaces*, 2006, pp. 84–91.
9. B. Scassellati, "Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot," 1999, pp. 176–195.
10. C. L. Breazeal, "Sociable machines: expressive social exchange between humans and robots," Ph.D. dissertation, 2000.
11. B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *International Conference on Human-Robot Interaction*, 2009, pp. 61–68.
12. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Nédellec and C. Rouveirol, Eds. Springer Berlin / Heidelberg, 1998, vol. 1398, pp. 4–15.