| | |
|---|---|
| Title | Data-Driven HRI : Reproducing interactive social behaviors with a conversational robot |
| Author(s) | Liu, Chun Chia |
| Citation | 大阪大学, 2017, 博士論文 |
| Version Type | VoR |
| URL | https://doi.org/10.18910/61827 |
| rights | |
| Note | |

*Osaka University Knowledge Archive : OUKA*

https://ir.library.osaka-u.ac.jp/

Osaka University

# Data-Driven HRI:
# Reproducing interactive social behaviors with a conversational robot

CHUN CHIA LIU

MARCH 2017

# Data-Driven HRI:
# Reproducing interactive social behaviors with a conversational robot

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN ENGINEERING

BY

CHUN CHIA LIU

MARCH 2017

# Abstract

As robots become more prevalent in the modern era, the field of human robot interaction (HRI) provides the promise of integrating robots socially into everyday human life. In many situations, a robot needs to be able to perform several tasks defined for its role. For example, a shop assistance robot needs to be able to greet customers, answer questions, and give recommendations. The question of how the interaction logic and contents should be developed, as well as how interactive robot behaviors can be generated effectively, remains a core challenge.

My proposed solution is to use a data-driven approach - breaking human-robot interactions down into sequences of repeatable behaviors (e.g. proxemics formation) which can be reproduced in a robot using generative HRI models. This simplified representation opens up the possibility of learning top-down multimodal interaction logic directly from data, which is an entirely new approach in the HRI field. Learning directly from data has the potential to be much lower-effort than manual design of interaction logic or hand-crafted interaction contents, and it has the potential to leverage "big data". To demonstrate this approach, I have conducted three studies.

In the **first study**, I applied a data-driven approach to autonomously generate robot behaviors for an entire interaction. To that end, my system enabled a robot to learn an entire social interaction based solely on imitations of completely free-form human-human interactions observed in a real, physical environment. This was made possible through a combination of abstractions: the empirical identification of the typical speech and motion behaviors in the training data, combined with a set of generalizable HRI models specifying spatial formations. The effectiveness of the system was demonstrated through a user evaluation, and was also proven to be robust to speech recognition errors.

For the **second study**, I extended the system to enable a robot not only to respond to human-initiated inputs, but also to reproduce proactive behaviors. The extensions included: (1) introducing a concept of human "yield" behaviors, to predict opportunities for the robot to take proactive action; (2) using interaction history as an input for predicting context-dependent behaviors; and (3) incorporating an attention mechanism to learn which parts of the interaction history are important for predicting robot behaviors. This system was trained from human-human interactions, and its ability to generate proactive robot behavior was validated through offline analysis and a user study.

In the **third study**, I developed a model enabling a robot to autonomously generate multimodal deictic behaviors towards people, such that this model can be used as a fundamental construct in future data-driven applications. The parameters were calibrated empirically to attain a balance between understandability and social appropriateness, based on observed human deictic behaviors. A system evaluation showed that the robot's deictic behavior was perceived as more polite, more natural, and better overall when using my model, as compared with a model considering understandability alone.

With today's trends towards big data, I believe the demonstration of these three studies provides an argument that a data-driven approach shows great promise as a method for collecting, modeling, and learning interactive, multimodal social behavior for robot applications.

# Table of Contents

**1 Introduction 1**

1.1 Objective . . . 1
1.2 Our Approach . . . 2
1.3 Background . . . 4
  1.3.1 Manual Approach . . . 5
  1.3.2 Data-Driven Approaches . . . 5
1.4 Research Challenges . . . 7
  1.4.1 Variations in human behavior . . . 7
  1.4.2 Learning from unlabeled data . . . 8
  1.4.3 Learning mixed-initiative interaction . . . 9
  1.4.4 Social behaviors . . . 10
1.5 Proposal . . . 11
1.6 Organization of Sections . . . 11

**2 Learning Interactive Behaviors 12**

2.1 Introduction . . . 13
2.2 Related Work . . . 14
  2.2.1 Creating Interaction Content . . . 14
  2.2.2 Learning from Data . . . 14
  2.2.3 Using the crowd for learning . . . 15
2.3 Data Collection . . . 16
  2.3.1 Sensor Environment . . . 16
  2.3.2 Training Interactions . . . 16
  2.3.3 Example of human-human interaction . . . 18
2.4 Proposed Technique . . . 18
  2.4.1 Overview . . . 18
  2.4.2 Abstraction . . . 20
  2.4.3 Vectorization . . . 27
  2.4.4 Learning and execution of interactive behaviors . . . 32
  2.4.5 Example of Behavior Execution . . . 35
2.5 Evaluation Experiment . . . 35
  2.5.1 Comparison system . . . 36
  2.5.2 Hypotheses . . . 38

2.5.3 Experiment Setup . . . . . . . . . . . . . . . . . . . . . . . . 38
2.5.4 Measurement . . . . . . . . . . . . . . . . . . . . . . . . 39
2.5.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . 42
2.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
2.6.1 Contribution . . . . . . . . . . . . . . . . . . . . . . . . 50
2.6.2 Validation of the Model . . . . . . . . . . . . . . . . . . . . 50
2.6.3 Assumptions . . . . . . . . . . . . . . . . . . . . . . . . 51
2.6.4 Generalizability and Scalability . . . . . . . . . . . . . . . . . 51
2.6.5 Tradeoff between Variation and Robustness . . . . . . . . . . . 52
2.6.6 Embodiment of the robot . . . . . . . . . . . . . . . . . . . 53
2.6.7 Justification for Comparison System . . . . . . . . . . . . . . 54
2.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
2.8 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

**3 Learning Proactive Behaviors 56**
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
3.2 Related work . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
3.2.1 Learning social behaviors from data . . . . . . . . . . . . . . . 58
3.2.2 Proactive robot behaviors . . . . . . . . . . . . . . . . . . . 59
3.2.3 Learning from history . . . . . . . . . . . . . . . . . . . . 59
3.3 Data Collection . . . . . . . . . . . . . . . . . . . . . . . . . . 60
3.3.1 Scenario . . . . . . . . . . . . . . . . . . . . . . . . . . 60
3.3.2 Sensors . . . . . . . . . . . . . . . . . . . . . . . . . . 60
3.3.3 Participants . . . . . . . . . . . . . . . . . . . . . . . . . 61
3.3.4 Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . 61
3.3.5 Observed Behavior . . . . . . . . . . . . . . . . . . . . . . 62
3.4 Proposed technique . . . . . . . . . . . . . . . . . . . . . . . . 62
3.4.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . 62
3.4.2 Abstraction of typical behavior patterns . . . . . . . . . . . . . 65
3.4.3 Definition of yield actions . . . . . . . . . . . . . . . . . . . 66
3.4.4 Incorporating interaction history . . . . . . . . . . . . . . . . 67
3.4.5 Learning to attend to history . . . . . . . . . . . . . . . . . . 68
3.4.6 Examples of using the Attention Mechanism . . . . . . . . . . 72
3.5 Offline Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . 74
3.5.1 Evaluation procedure . . . . . . . . . . . . . . . . . . . . . 74
3.5.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . 75
3.6 User Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
3.6.1 Comparison System . . . . . . . . . . . . . . . . . . . . . 76
3.6.2 Hypothesis and Prediction . . . . . . . . . . . . . . . . . . . 76
3.6.3 Experiment Setup . . . . . . . . . . . . . . . . . . . . . . . 77
3.6.4 Measurement . . . . . . . . . . . . . . . . . . . . . . . . 78
3.6.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . 78
3.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

3.7.1 Contribution . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
3.7.2 Identifying yield actions in turn-taking . . . . . . . . . . . . . . 82
3.7.3 History Representation . . . . . . . . . . . . . . . . . . . . . . . 82
3.7.4 Generalizability and Scalability . . . . . . . . . . . . . . . . . . 83
3.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
3.8.1 Action Discretization . . . . . . . . . . . . . . . . . . . . . . . . 84
3.8.2 Defining input features . . . . . . . . . . . . . . . . . . . . . . . 84
3.8.3 Defining Robot Actions . . . . . . . . . . . . . . . . . . . . . . 85

**4 An HRI model for Deictic Behaviors 87**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
4.2 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.2.1 Studies of Human Pointing Behavior . . . . . . . . . . . . . . . 89
4.2.2 Human-Robot Interaction . . . . . . . . . . . . . . . . . . . . . 89
4.3 Data Collection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
4.3.1 Objective . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
4.3.2 Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
4.3.3 Categorization of Pointing Types . . . . . . . . . . . . . . . . . . 91
4.3.4 Categorization of Descriptive Terms . . . . . . . . . . . . . . . 91
4.3.5 Results and Analysis . . . . . . . . . . . . . . . . . . . . . . . . 94
4.4 Generative Model for Robot Behavior . . . . . . . . . . . . . . . . . 95
4.4.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.4.2 Understandability . . . . . . . . . . . . . . . . . . . . . . . . . . 96
4.4.3 Social Utility . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
4.4.4 Calibration of Our Model . . . . . . . . . . . . . . . . . . . . . . 100
4.4.5 Examples of Using Our Model . . . . . . . . . . . . . . . . . . . 100
4.4.6 Model validation . . . . . . . . . . . . . . . . . . . . . . . . . . 101
4.5 System Elements . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
4.5.1 Robot Platform . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
4.6 Evaluation with a Robot . . . . . . . . . . . . . . . . . . . . . . . . 105
4.6.1 Comparison System . . . . . . . . . . . . . . . . . . . . . . . . 105
4.6.2 Hypotheses . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
4.6.3 Experiment Setup . . . . . . . . . . . . . . . . . . . . . . . . . . 106
4.7 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
4.7.1 Verification of Hypothesis 1(Referent) . . . . . . . . . . . . . . 107
4.7.2 Verification of Hypothesis 2 (Listener) . . . . . . . . . . . . . . 108
4.7.3 Analysis of understandability and social utility on the behavior-selection process . . . . . . . . . . . . . . . . . . . . . . . . . 108
4.8 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
4.8.1 Interview results from the participants . . . . . . . . . . . . . . 112
4.8.2 Politeness in Pointing . . . . . . . . . . . . . . . . . . . . . . . . 112
4.8.3 Limitations and future work . . . . . . . . . . . . . . . . . . . . 112
4.9 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113

**5 Discussion and Conclusions 115**
5.1 Summary and Achievement . . . . . . . . . . . . . . . . . . . . . . . . . . 115
5.2 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
    5.2.1 Future Direction . . . . . . . . . . . . . . . . . . . . . . . . . . 116
    5.2.2 Behavior modeling in HRI . . . . . . . . . . . . . . . . . . . . . . 117
    5.2.3 Representation of new semantics . . . . . . . . . . . . . . . . . . 117
    5.2.4 Generalizability . . . . . . . . . . . . . . . . . . . . . . . . . . 118
    5.2.5 How much control should an interaction designer have? . . . . 119
    5.2.6 Robots as customer service representatives . . . . . . . . . . . 119
    5.2.7 The potential of data-driven approach . . . . . . . . . . . . . . . 120
5.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120

**Acknowledgments 122**

**Works Cited 124**

**List of Publications 137**

# Chapter 1

# Introduction

## 1.1 Objective

We are living in an unprecedented era of technology development and progress, and today's endless array of technology toys, sensors, and gigantic volumes of data would almost make it seem like we are living something out of a Star Trek movie. We see the volume of data available to us growing at an extraordinary rate. More data has been created in the past two years than in the entire previous history of the human race. Some estimate that by the year 2020, our accumulated digital universe of data will grow from 4.4 zettabytes today to around 44 zettabytes – that's 44 *trillion* gigabytes. That's big by anyone's standard! Surely with today's rate of data acquisition, an interactive robot that can communicate and talk with us is just around the corner.

Recent years have seen great technological progress in robotics in general, and we are accustomed to the idea of robots performing heavy loads of continuous work in factories, power plants, and manufacturing facilities. However, these robots mostly perform tasks that require little, if any, interaction with casual users. Yet, in recent years, robots are becoming more common in public spaces such as restaurants, shops, or hotels, where they often need to communicate and talk with casual users in a natural and intuitive way.

The field of Human-Robot Interaction (HRI) has demonstrated many promising results that can endow robots with natural and socially-appropriate behaviors. Thanks to studies in HRI, we have models for generating robot behaviors, such as how to approach a person [113], turn its head and gaze [146], and understand when it is its turn to speak [15].

While we are making great progress towards developing better HRI models for social robot behaviors, most models are developed in a way that is quite labor-intensive and time-consuming. Often, a programmer is required to explicitly program the behaviors the robot should execute, and interaction logic is heavily reliant on the programmer's ability to imagine a variety of social situations (for example, anticipating all of the questions people may ask the robot) and use their intuition to specify social behaviors and execution rules for the robot.

So, how can we can leverage today's trends towards "big data" to ease the development process for robot behavior in natural human-robot interaction?

My objective in this dissertation is to address several key challenges that are holding back the use of data for generating behaviors and interactions in the HRI field. I advocate a "Data-driven HRI" approach, in which social interaction can be broken down into common, repeatable HRI patterns. This representation allows us to use sensor data captured from human-human interaction in a real, physical environment to: (1) abstract typical action primitives from humans' behaviors, (2) learn interaction logic to reproduce social behaviors in human-robot interaction, and (3) develop generative models for robot behavior.

## 1.2 Our Approach

As robots become more available in our lives, they need to fulfill a set of defined tasks for their roles. For example, a shop assistance robot needs to be able to greet customers, answer questions, give recommendations, guide to various products, and assist the customers in various situations. The successful completion of these tasks is paramount for seamless integration of robots in our society.

We can consider these tasks as consisting of sequences of social behaviors. For example, greeting can be characterized as *approaching* an individual, followed by *gesturing* to an individual an indication to initiate interaction, and then followed by a *verbal* acknowledgement. Likewise, other behaviors have also been identified in form of conversations [132], collaborations [22], and interviews [78].

Assuming that social behaviors are the fundamental constructs for social interactions, what approach should be used for generating social behaviors for robots? The underlying elements in social behaviors are *models* of natural behavior. For example, models for gazing [86], approaching [112], pointing [128], and guiding [121] have been implemented for robots to engage in interaction.

As the repertoire of behavior models grows, the robots' ability to interact naturally with people will also improve. However, research studies typically focus on modeling only small portions of a whole interaction, raising the question: how should the models be integrated to enable an entire social interaction?

One approach for designing interaction logic for a robot is to explicitly program the behaviors the robot should execute, the expected inputs from the environment, and the execution rules it should follow. However, this can be a difficult process, heavily dependent on the designer's ability to imagine a variety of social situations (for example, anticipating all of the questions people may ask the robot) and use their intuition to specify social behaviors and execution rules for the robot, which may be difficult to articulate. This process can be very labor intensive, and it becomes even more difficult to create robust interactions when natural variations of human behavior and errors due to sensor noise are considered.

As a solution to many of these problems, if we can consider certain social interactions as consisting of repeatable task elements, then a data-driven approach can be applied to learn the overall social interaction logic. Consider a museum guide who repeatedly explains about the artist of an exhibit, or a travel agent who repeatedly provides information and answers questions about tour packages. In order to complete their “job", there is some functional requirement which each person needs to complete in a similar way. If the behaviors required to complete the tasks are repeatable, then the robot can learn them.

Data-driven approaches fundamentally provide clear advantages in many ways over hand-crafted interaction logic. Our work does not attempt to prove this fact because it is well-understood. Many works have investigated such data-driven approaches, although many are limited to capturing data in a simulated environments [8, 93], require human annotations [72], only deal with nonverbal behaviors [1, 149], or react only to human inputs and are unable to take initiative [19, 148]. Our focus here is to address the question of how a fully end-to-end data-driven approach can be applied to enable a conversational robot to learn an entire mixed-initiative social interaction based solely on observations of human-human interaction in a real, physical environment.

In Chapter 2, we first demonstrate how HRI behavior modules developed by other reseachers can be used as building blocks, so that a large number of real *in-situ* human-human interactions can be captured. As a result, it is possible to easily and automatically collect a set of multimodal behaviors (e.g. speech, locomotion, proxemics formation) and interaction logic that can be used in a robot. This would reduce the difficulty and effort of interaction design, and it could enable more robust interaction logic, since sensor errors and variation of behavior would be implicitly considered. The evaluation compares other state-of-the-art data-driven techniques (e.g. nearest-neighbor learning) to our proposed system, which shows our proposed system outperforms the state-of-the-art data-driven techniques and is also robust to sensor noise.

Chapter 3 focuses on how additional data-driven techniques can be applied, such that mixed-initiative human-human interaction data can be used to train a robot to both respond to human-initiated inputs and initiate its own action. This extension is important for applications of robots in the real world, as we are seeing a growing market in the service industry for robots which interact with customers. In such situations, the robots may need to proactively engage with their customers. Our proposed system in Chapter 3 is evaluated against the system from Chapter 2, and demonstrated that the proposed system indeed learns the behaviors of a proactive shopkeeper.

Chapter 4 focuses on extending this work to new modalities by developing a gesture model, specifically a behavior model for deictic behaviors towards people. An important element for social interaction is gesture, which can play a crucial role in the processes of interaction and communication. Therefore, we believe it is necessary to include the modality of gesture in future data-driven HRI applications. However, designing behavior models (i.e. gestures) from first principles and logic are often difficult, since the principles that govern our behaviors are part of our implicit knowledge and therefore difficult to articulate. Thus, rather than using a purely data-driven approach, it is useful to observe how humans manipulate the world, and interpret the data to reason out the mechanisms behind the model. Our proposed HRI model for pointing, consisting of a "understandability" factor and a "social utility" factor is evaluated against a pointing model that only considers resolving ambiguities in the environment, similar to the ones proposed by [47, 129].

The objective of this work is to provide a proof-of-concept of a fully data-driven interaction design methodology and to provide observations and suggest directions for future development of this powerful concept, as shown in Fig. 1.1.

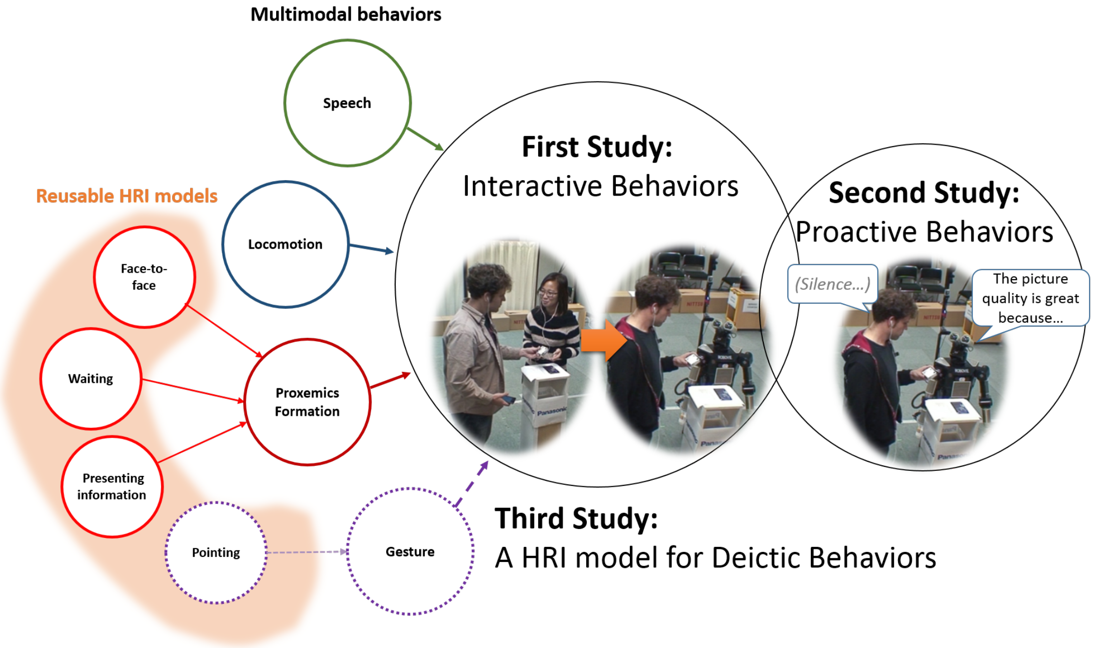


Figure 1.1: Some important topics for generating social interaction logic for a conversational robot: (1) a data-driven approach to learning interactive behaviors, (2) extending data-driven techniques for learning proactive behaviors for mixed-initiative interaction, and (3) a HRI model for generating deictic behavior towards people for future data-driven application.

## 1.3 Background

In this section, we will survey two approaches to interaction design: (1) a manual approach, where a human designer is involved in the integration of behavior models as well as development of interaction content (i.e. utterances), and (2) a data-driven approach, where example interaction data may be used partially or completely, in the integration of behavior models or the development of interaction content.

### 1.3.1 Manual Approach

As mentioned in the above section, it is useful for humans to design behavior models manually to fit the data because we are intelligent, and thus can apply our reasoning abilities to understand the underlying mechanisms. Therefore, it makes more sense that models for social behaviors are developed using a manual approach. First, using our own background knowledge of social constructs, we can study how humans interact with one another in different situations. Once we understand the basic principles that govern our behaviors, we can create an engineered solution and follow the same approach to develop generative models that are appropriate for social robots. These manual approaches for behavior models are developed for locomotive behaviors [119], gestural behaviors for pointing to objects [50, 110, 111, 116, 129] and pointing to space [47, 126].

Another aspect of social robots that involves a manual approach is designing interaction logic. Tools, such as graphical user interfaces [24, 26, 36] or example templates [20] for authoring interaction contents are often used, which help ease the development process for an interaction designer. In Glas et al.'s work, they presented a system that allows the designer to integrate "behavior modules"(e.g. approaching) defined by programmers such that behavior instances (e.g. approaching a waiting person, catching up to a person) can be created.

Manual approaches have the benefits of fine-tuning behavior contents and interaction strategies, which may be important for certain domain. For example, robots as performers may require a human designer to architect the robot's overall image or personality, and fine-tuning its dialogue, movement, facial expression, or gestures. This is seen for robots in theatre performances [21], poetry-reciting agents [91], and news broadcasters [75].

While manual approach provides the designer to fine-tune robot behaviors, this can be a time-consuming and labor-intensive process, heavily dependent on the designerâĂŹs ability to imagine a variety of social situations and often require several iterations to fine-tune the robot's behaviors. Thus, there is a need to make the process for designing interaction easier.

### 1.3.2 Data-Driven Approaches

Considering that social behavior is contextual and involves verbal dialog and nonverbal behaviors at a high level, then social interaction itself can be seen as the accumulation of these communicative patterns. These interaction patterns are not only observable in human-human interaction, but also in human-robot interaction. Kahn et al. identified some patterns such as "initial interaction" and "didactic communication" by observing how children interacted with robots [55]. Peltason and Wrede also defined generic interaction patterns (i.e. "information request" and "clarification") that combine abstract task states with robot dialog acts [99].

Repeatable interaction patterns can also be observed in different scenarios. For example, a museum guide explains factual information about an exhibit, a travel agent provides pricing information about tour packages, or a receptionist explains the whereabouts of the washroom. Over the course of a day, the service provider may need to re-iterate the same information to different customers. While there may be variations in the interaction style among each service provider, there is some functional requirement which each person needs to complete in a similar way. When enough repeatable behaviors have been accumulated, it can be used as training data for generative robot behaviors.

Interaction data has been used to achieve partial autonomy for some robot behaviors, for example, using a teleoperation system. Knox et al. collected interaction data through a teleoperation system, and used hierarchical model using binary logistic regression to learn some robot actions in an educational domain. Their robot was mostly successfully in learning to emulate the demonstrated interaction heuristics [65]. Similarly, Magyar et al. also proposed iterative reinforcement learning from data collected from teleoperation to increase the robot's level of autonomy [77]. Their preliminary result was tested in simulation, in which they plan to learn an interactive lecture using a NAO robot in the future.

Some studies collect interaction data from games [8, 93] or through the web. Breazeal's group developed a gamed called Mars Escape in which players role-played as either a robot or an astronaut to solve collaborative task together. The game data was logged and used to train a robot in the real world. While crowdsourcing data has the potential to easily collect large amounts of training data, the difficulty of attaining immersion still applies, as the operator sits behind a computer screen and types out interaction commands. Thus, the learnt behavior models may not entirely represent what a person would have done in real life.

On the other hand, some works have demonstrated that robot behaviors can be learnt from manually labeled data collected from real human-human interaction. For example, the JAMES robot [62] used a dialog model along with human interaction data to acquire the skills of a bartender. Admoni and Scassellati trained a learning model to predict robot nonverbal actions from labeled interaction data [1]. Similarly, in the work of Leite *et al.*, they took the approach of a semi-situated learning to train an agent with its own multimodal language behavior [72]. Their solution was to crowdsource multiple phrase variations of a line of dialogue. The crowd worker would author and edit a single line of character dialogue and its manner of expression, which would be later used in a similar moments in conversation. Their case study showed an autonomous robot interacting with 200 users using both meaning content and variety of expression. While learning with labeled data is a valid approach in many data-driven techniques (i.e. "supervised learning"), the process of labeling data inherently requires time and effort.

Our work also presents a data-driven approach, where we aim to reproduce robot behaviors for a specific social role. Different from the above works, we adopt a completely hands-off method, where data is used to generate completely autonomous robot behaviors. The idea of using natural human-human interactions to enable human-robot interaction constitutes a promising approach, and could prove to be a much-needed transition of social robots from manual interaction design towards autonomous development of interaction logic.

## 1.4 Research Challenges

Considering the state of the art in robotics technology today, there are a number of major challenges that make it difficult to apply a data-driven approach to generate natural robot behaviors for human-robot interaction. In this section, I will summarize some of these challenges and introduce how data-driven techniques can be applied to assist social robots in generating appropriate behaviors despite these limitations.

### 1.4.1 Variations in human behavior

People don't walk the same way, or talk the same way. Person by person, situation by situation, they typically express themselves in different ways. Due to our diverse background, upbringing, and cultural influences, we often use different phrases to express even semantically similar ideas. For example, there is already a huge number of variations even for something as simple as the way we respond to 'thank you'; we can respond with 'sure', 'no problem', 'you're welcome', 'don't worry about it', 'it's my pleasure', and so on and so forth.

To mitigate this issue, a robot may be programmed to anticipate the different behavior variations it may encounter. However, behavior variations that may arise from person to person pose major challenges to the development of interaction logic. Even if a robot is programmed to handle a great variety of contingencies in an interaction, there is always a risk that something unexpected will happen, and the robot will have no ability to interact meaningfully. This creates a strong reliance on a designer's ability to anticipate every conceivable problem and generate scripted responses, a time-consuming and error-prone task which requires substantial imagination and expertise. Thus, the effort of manual development of interaction logic for robots poses a significant obstacle to the advancement of research in the field.

Traditionally, a technique which has been used to deal with the difficulties of perception or of decisions based on social understanding is to offload such task to a remote human operator, by using a Wizard-of-Oz (WoZ) technique [39]. This involves a remote human operator, who may be unknown to the test subjects, to make a "high-level" decision of a certain sensor input, based on his social understanding and the scenario at hand. Some studies have used this as a technique for building a database of robot behaviors, to gradually increase the robot's autonomy over time [56, 150]. However, this approach has several drawbacks. First, it is costly and non-scalable, as it requires a robot and teleoperation setup. It is also difficult for a teleoperator to attain a high degree of immersion in a social interaction, making it difficult to generate intuitive and natural behavior. Finally, the delays associated with teleoperation can cause failures in interactions, which would not be acceptable in most real-world business or social scenarios.

Faced with an incoming stream of continuous sensor data from real human-human interaction, the question remains – how should a robot figure out which of its myriad of perceptions are relevant to the task, and how to distill these sensor data into a format that is understandable and interpretable for it to anticipate and respond to? These "conditions" for behavior execution may range from sensor input (i.e. speech recognition, human detection) to implicit knowledge based on social norms. If we can capture the condition through sensors, as well as the behavior that follows, it may be possible to learn the necessary behaviors without explicit modeling.

An equally important question is with the problem of generating appropriate robot behavior given a diverse set of examples – given the variations of demonstrated human behaviors, which of the demonstrated behaviors should the robot generated in a defined action space? This is an important consideration when a robot needs to respond with coherent behavior, given that the search over state space becomes enormous as perceptual abilities and complexity of the environment and social scenario increase. In a shop scenario, we define a robot action to be discrete and multimodal, which includes speaking an utterance and a proxemics formation.

The total dimensionality of the set of possible utterance and movement actions a human could perform at any time during a social interaction is enormous. In practice, however, the variation of human behaviors occupies only a small manifold within this high-dimensional space. Thus, it seems reasonable that by identifying common, repeatable behavior patterns, a data-driven strategy could become feasible.

In Chapter 2, I will propose a data-driven technique to abstract and cluster human behaviors captured by continuous streams of sensor data into common, typical behavior patterns. I demonstrate how typical behavior patterns, both speech and locomotion, can be used to reduce the dimensionality of the learning problem for the robot. Not only that, I demonstrate how typical behavior patterns can complement existing HRI models and be used as abstraction to generate robot behaviors in an interaction.

### 1.4.2 Learning from unlabeled data

In many learning tasks, dealing with noise from the environment is a nontrivial problem. For example, low-level manipulation learning tasks such as teaching a robot to stack block may involve dealing with multiple sources of uncertainty (i.e. camera noise, robot arm noise) [25]. Similarly, learning social behaviors often involves a human demonstrator performing a certain social behavior, which is captured by external sensor systems, using speech recognition system that is susceptible to accent. To sidestep environmental noise, a system developer may manually label captured sensor data, such as what objects were recognized or what was spoken.

This problem of using a human-in-the-loop to manually annotate data for abstraction is a very time consuming and hard task [46]. This often involves familiarizing a human with the characteristics of the sensors. If the domain in which the data is collected is a specialized field (i.e. medical), then employing a domain expert to annotate the data may be needed and can also be expensive. Thus, it takes a lot of knowledge about the characteristics of the sensors and a lot of time to define good concept descriptions, even for experts.

Thus, the challenge is how can noisy sensor data be abstracted and used for learning interaction logic, as well as for generating interaction contents, without the time-consuming task of having a human-in-the-loop. For human-robot interaction, this means data collected from the environment will need to be automatically abstracted to a state such that the residual noise will not hinder the performance of the learning model. Given a noisy human speech input recognized by a speech recognition system, a robot needs to learn the right condition for interaction logic. Likewise, a robot also needs to learn the right action (i.e. utterance and motion) to execute, despite noises from speech recognition and motion capturing systems.

In Chapter 2, I will show a combination of data-driven techniques techniques (i.e. clustering and classification), which enables the robot to learn action policies from unlabeled data. This allows behaviors to be captured directly from sensor data, without the time and effort associated with data annotation.

### 1.4.3 Learning mixed-initiative interaction

Human interactions are complex, and they include rich information exchanges in complex environments. In many scenarios, we often expect that our interaction partner will act in a proactive way. For example, we would expect that a tourist information guide would provide us with details about an attraction, or that a museum guide would provide interesting anecdotes about an exhibit, without the need for us to explicitly ask for this information.

Thus, it is important for a robot to also reproduce proactive behaviors in human-robot interaction. While this is usually an easy task for humans, it is difficult for a robot to know when and what proactive action to take due to the complexity and nuances of a conversation. Proactivity requires autonomous, anticipatory robot-initiated behavior, rather than only reacting to the actions or commands of a person. While there have been some studies that focus on proactive robot behaviors, they are often restricted to non-verbal task-related behaviors such as handover [51] or locomotion [58].

Here, we describe the challenge of using a data-driven approach to learn proactive behaviors for a robot. Whereas reactive behaviors can often be modeled in terms of generating a response to a person's action, the generation of proactive behavior is more likely to be dependent upon interaction history or context. However, depending on what was said by the user, the relevancy of the interaction history may matter. For example, a reactive response to a question only depends on the most recent utterance, whereas a proactive behavior, for example, when the robot decides to take the initiative to do something as a result of the human user yielding his turn, may depend on previous interaction context. However, such contextual sensitivity is difficult to capture, and the naïve injection of context information may introduce unnecessary noise and sparse non-repeatable data to the system, hindering the ability for the robot to learn an appropriate action.

In Chapter 3, I propose a data-driven approach to learn proactive behaviors from mixed-initiative human-human interaction captured from sensor data. I will show how interaction history can be incorporated such that the robot is able to generate not only reactive behaviors, but also proactive behaviors when necessary. Lastly, I will show how the techniques described previously can come together as a system, and demonstrate how a robot will be able to reproduce the behaviors of a proactive human demonstrator.

### 1.4.4 Social behaviors

We humans have many behaviors, and collectively these behavior primitives can serve as a substrate for a basic repertoire of social interactions. Similarly, a social robot with behavior primitives will benefit from simpler interaction design, simpler descriptions of high-level tasks,and easier learning of interaction logic.

One challenge is that background knowledge for social interaction is often implicit, and thus difficult to articulate. We are often not consciously aware of our behavior when we communicate, such as the rules governing where we direct our gaze or body orientation. These implicit rules and conventions establish appropriate and acceptable ways for us to act and respond to each other [101].

As a robotic programmer, one needs to be able to articulate these social conventions in a concrete way in order to develop generative robot behavior. While it is intuitive for us to behave according to social conventions, it is often tough for us to explicitly explain why we do the things we do. Very few programmers are also psychologists or behavior analysts, and thus it will be difficult for them to describe the reasoning that governs social behaviors. For example, we often use social cues such as utterances (“let’s turn here”) when we walk side-by-side, or look away when being asked a difficult question, but we cannot explain off the top of our head why we act this way. Thus, not only do programmers need to be able to articulate the rules that govern our behaviors, they also should be able to translate it into system parameters that are understandable and interpretable by a robot.

Hence, there are two challenges when it comes to applying a data-driven technique for interactive robot behaviors, first, how to generate humanlike robot behaviors, and second, how to respond to human actions. Since many implicit rules that govern our deceptively subtle behavior are not extrinsically observable by sensor data, one common approach is to adopt explicit modeling of both the intrinsic and extrinsic parameters. We need to know what we are looking for and what kind of information to present to the robot when trying to develop a specific social behavior.

In the last section of this work, I demonstrate in Chapter 4 how a generative robot behavior model can be developed from observation data of human behaviors, specifically, deictic behaviors towards other people, so that it can be used as a component in future data-driven applications. The idea is to use collected data of human pointing behaviors to identify typical pointing patterns, and to develop a top-down model of generalizable, multimodal deictic behaviors for a robot, consisting of an utterance and a pointing gesture. I also demonstrate how this model can be used in a real shopping mall environment, while taking into account crowds of real shoppers, to generate deictic behaviors that are socially-appropriate for both the referent and the listener.

## 1.5 Proposal

The main novel achievement of this work is describing how data-driven HRI approaches can be applied in robotic applications. My work demonstrates how data-driven approach can be applied:

- Reproducing multimodal behaviors for a conversational robot
- Extending data-driven techniques for a mixed-initiative social interaction
- Developing a reusable HRI pointing model in order to extend data-driven application to other modalities

The end result of this work is a fully end-to-end approach, enabling a robot to learn both reactive and proactive behavior in an entire multimodal social interaction based solely on unscripted, natural human-human interaction in a real, physical environment, without having a human-in-the-loop. While there are many existing data-driven approaches, many are limited to low-level tasks such as trajectory following [16, 88], joint motion replication [92], or capturing data in a simulated environment [96], requiring human annotations [1], learning from text that is typed rather than spoken, or only dealing with nonverbal behaviors. In contrast, I demonstrate that behaviors can be captured from real, non-annotated human-human interaction that is completely free form and contains large amounts of variation in natural speech among the different people.

Once passive collection of sensor data becomes widely available, I believe such work will not only increase the diversity, and complexity of human-robot interaction, but which could also realistically be applied to real commercial robot applications.

## 1.6 Organization of Sections

Chapter 2 addresses the issue of automatic generation of robot behaviors and behavior rules directly from non-annotated data. This work will cover how data is used for abstracting typical behavior patterns, generating interaction content for a conversational robot, and learning interaction logic directly from data.

Addressing the important challenge of generating proactive behavior, Chapter 3 presents a system I developed which uses data-driven techniques to reproduce context-dependent proactive behavior by automatically learning which elements of interaction history are important for predicting a robot behavior at any given time.

An example of developing a reusable HRI model for generative robot behavior will be presented in Chapter 4. This chapter describes how data can be used to inform the development of generative models for deictic behaviors towards people, and then used to develop and calibrate a model based on observed human behaviors. Once such HRI model is developed, it can be used in future data-driven applications to enable a richer and more complex human interaction.

Finally, the results of this work and its implications for the future are discussed in Chapter 5.

# Chapter 2

# Learning Interactive Behaviors

This chapter begins this exploration of data-driven by presenting how human-robot interaction can be learnt from observations of human-human interaction. This chapter presents a set of data-driven techniques for using observations of human-human interactions to generate this interaction logic in a fully-automated way, without requiring any human annotation or programming of the social behaviors.

One set of challenges posed by this problem involves the processing of continuous and noisy sensor observations, in order to identify discrete and abstracted speech and movement action primitives in the training data. In order to achieve this, I present several new techniques, including the use of unsupervised clustering techniques for abstracting and refining speech actions, spatial and trajectory clustering to discretize locomotion actions, and the identification of abstracted spatial formations based on established HRI proxemics models.

The other set of challenges involves the use of machine learning for generating robot actions in response to human behavior. For this, I propose a technique using a Naïve Bayesian classifier to predict abstracted robot actions based on a vectorization of the human's speech and motion. To generate the actual motor and speech synthesis commands for the robot, I propose techniques for reproducing robot speech and locomotion behaviors based on clustered abstractions, in order to robustly reproduce human behavior despite a great deal of sensor noise and natural variation in human behavior.

Finally, I demonstrate this technique in use, training a robot to play the role of a shop clerk in a simple camera shop scenario, showing through a comparison experiment that these techniques successfully enabled the generation of socially-appropriate speech and locomotion behavior. Notably, the robot's performance in terms of correct behavior selection was higher than the success rate of speech recognition, indicating its robustness to sensor noise and showing great promise for the use of this technique in real-world applications.

## 2.1 Introduction

As robots become more prevalent in the modern era, the field of human robot interaction (HRI) provides the promise of integrating robots into everyday human life. These service robots are gaining presence in museums [5, 11, 57, 90], offices [38, 131], elder care [71, 109], shopping malls [33, 40], and healthcare facilities [85]. The ability of the robots to socially integrate into those environments will be essential. For example, a shop assistance robot needs to be able to greet customers, answer questions, give recommendations, guide to various products, and assist the customers in various situations.

One approach for designing interaction logic for a robot is to explicitly program the behaviors the robot should execute, the expected inputs from the environment, and the execution rules it should follow. However, this can be a difficult process, heavily dependent on the designer's ability to imagine a variety of social situations (for example, anticipating all of the questions people may ask the robot) and use their intuition to specify social behaviors and execution rules for the robot, which may be difficult to articulate. This process can be very labor intensive, and it becomes even more difficult to create robust interactions when natural variations of human behavior and errors due to sensor noise are considered.

We believe that a data-driven approach to interaction design could provide solutions to many of these problems. By directly capturing behavior elements such as utterances, social situations, and transition rules from a large number of real, *in-situ* human-human interactions, it may be possible to easily and automatically collect a set of behaviors and interaction logic that can be used in a robot. This would reduce the difficulty and effort of interaction design, and it could enable more robust interaction logic, since sensor errors and variation of behavior would be implicitly considered.

Thanks to recent advances in sensor technology, this idea of data-driven interaction design based on real-world interactions could soon become a realistic possibility. High-precision tracking systems are being deployed in public spaces, enabling passive collection of natural human interaction data [10], and technologies such as microphone arrays may soon provide usable sound source localization and speech recognition in noisy real-world environments [69]. Such technologies could allow enormous amounts of human behavior data to be collected effortlessly. For example, deploying sensor networks in a chain of retail stores could provide hundreds of thousands of example interactions in a matter of few months, which could be used to train a robot to perform the role of a shop clerk.

The possibility of effortless collection of large amounts of interaction data is what gives importance to this idea of data-driven interaction design. HRI researchers have recently begun to take advantage of the scalability of the web to train robots based on collected interaction data from the crowd [8, 135]. We believe that capturing human-human interactive behavior through sensor networks will prove to be another powerful and scalable way to leverage the wisdom of the crowd to create interactive robots.

Our objective in this study is to provide a proof-of-concept of such a data-driven interaction design methodology and to provide observations and suggest directions for future development of this powerful concept. We present a fully-autonomous method for training a socially-interactive robot from observed examples of human-human interaction, wherein behavior contents and interaction logic are extracted directly from noisy sensor data without human intervention.

Included in this work are techniques for (a) identifying typical action elements from a set of example interactions, (b) reproducing observed human behaviors in a robot despite high amounts of sensor noise, and (c) robustly selecting context-appropriate behaviors for the robot to execute in live social interactions.

## 2.2 Related Work

As mentioned above, our goal is to utilize the crowd, by capturing people's movement and speech during live human-human interaction and automatically generating interaction logic for reproducing the observed behaviors based on the set of passively-collected data. Such ideas of learning from data and using the crowd for learning have been explored in a number of different areas within the field of social robotics.

### 2.2.1 Creating Interaction Content

In designing interaction flows for social robots, several custom frameworks have been developed to explicitly break down interaction into subcomponents, such as state (input) and behavior actuation (output) components, and specify transition logic to direct the execution flow based on data from sensor inputs [35, 124]. Teleoperation interfaces have also been used to iteratively build interaction content over a period of time [56, 150]. In this work, we use sensors to capture interaction content directly from human interactions.

### 2.2.2 Learning from Data

In robotic tasks like manipulation, machine learning approaches such as *learning by demonstration* are often utilized to learn from a dataset of examples in order to reproduce a demonstrated task, as it is easier for humans, including non-robotic-experts, to input poses by moving an arm manually, than to explicitly specify them numerically. Some examples include trajectory following [16, 88] or joint motion replication [92]. Typically this is seen as a way to input sensory-motor patterns, but not cognitive and decision-making skills.

In social robotics, machine learning has been used to teach low-level behaviors, for example, to mimic gestures and movements [114], and to learn how to direct gaze in response to gestural cues [87]. In one example, pointing and gaze behaviors were recognized in an imitative game using a hidden Markov model [13]. The challenge in using a data-driven approach to learn an entire social interaction is the level of complexity that goes into decision-making process. The ways we act are often influenced by our intentions, and it is still an open question to how we can extract intentions from only observed behaviors.

Data-driven dialogue systems have been demonstrated in robots which infer meanings from spoken utterances. Rybski *et al.* developed an algorithm which allowed a human to interact with a robot with a subset of spoken English language in order to train the robot on a new task [108]. Meena *et al.* used a data-driven chunking parser for automatic interpretation of spoken route directions for robot navigation [79].

Unlike other works, we focus on training examples based on real human-human interaction, with natural spoken dialogue.

### 2.2.3 Using the crowd for learning

With the advancement of high-precision tracking systems able to monitor real social environments [10, 147], it is becoming possible to collect large amounts of detailed interaction data with little effort. This suggests the possibility of using a "crowdsourcing" approach, like the distributed techniques used over the web to solve complex problems, e.g. users on Amazon's Mechanical Turk helping to annotate images for grasp planning [125].

The use of real human interaction data collected from sensors for learning interactive behaviors has been investigated in numerous works. The robot JAMES was developed to serve drinks in a bar setting, in which a number of supervised (i.e. dialog management) and unsupervised learning techniques (i.e. clustering of social states) have been applied to learn social interaction [31]. In contrast, we propose a completely unsupervised approach for both abstraction and clustering of social states as well as for robot behavior generation

In Young et al.'s work [148, 149], a person provides an example of an interactive locomotion style, which is used to teach the robot to generate interactive locomotive behaviors in real time according to that style. We also propose to use real human interaction to train the robot, but our focus is not only the robot's motion, but its speech as well.

Connectivity to the web has also changed the way interaction data can be collected. The Robot Management System framework was developed to make learning of manipulation and navigation tasks easier by collecting demonstrations from remote users through a browser as a game [135]. The Restaurant Game used annotated crowdsourced data to generate abstracted representation of data to automate game characters [95]. The Mars Escape online game used crowdsourcing to learn robot behaviors [8, 18, 19]. The idea was to use a data-driven approach to develop HRI behaviors from players of an online collaborative game to provide large amounts of training data and reproduce behaviors in a real autonomous robot.

Our work complements these approaches by considering a crowd-based data collection from sensors in a physical environment, where some new challenges include resolving recognition ambiguities due to sensor noise and natural variation of human behavior.

## 2.3 Data Collection

### 2.3.1 Sensor Environment

To collect human-human interaction data for our learning study, we prepared a data collection environment with a sensor network, including a human position tracking system and a set of handheld mobile phones to use for speech recognition, to capture participants' motion and speech.

The position tracking system consists of 16 ceiling-mounted Microsoft Kinect RGB-D sensors, arranged in rows. Particle filters are used to estimate the position and body orientation of each person in the room based on point cloud data [10].

Ideally, we would like to collect people's speech passively. However, modern speech recognition technology is still not robust enough to use with ambient microphones when background noise exists in the environment [30, 123]. For that reason, we developed a smartphone application to capture speech directly from a hands-free headset, and use the Android speech recognition API to recognize utterances, sending the text to a server via Wi-Fi. The user wears a hands-free headset and touches anywhere on the mobile screen to indicate the beginning and end of their speech, so no visual attention is required, making it possible to conduct natural face-to-face interactions without breaking eye contact.

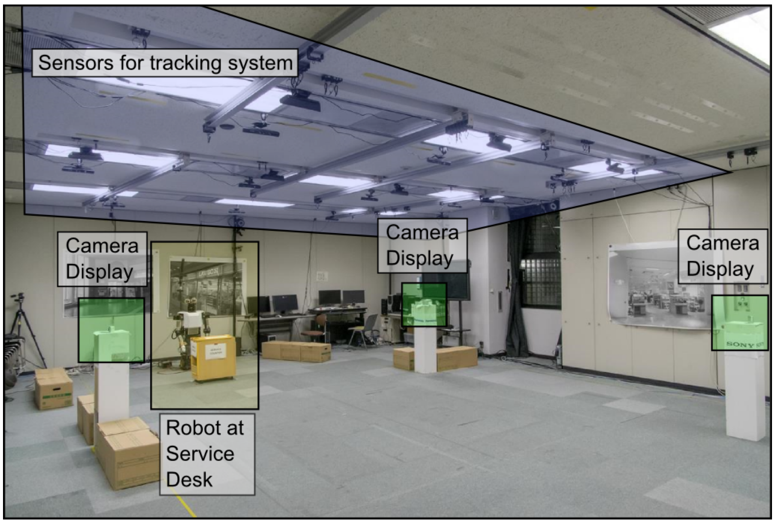
Although the study was conducted in Japan, we found a greater variety of tools available for analysis of English text, so the interactions in this study were carried out in English.

### 2.3.2 Training Interactions

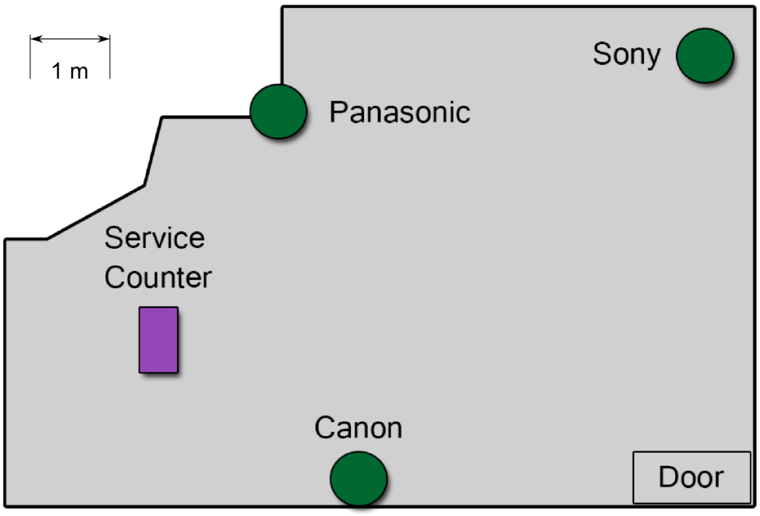
We chose a shopping scenario in a camera shop setting, where we asked one person to role-play as a shopkeeper and one person as a customer. To create a set of training interactions, we set up three product displays, representing different digital camera models, in an 8m x 11m experiment space, shown in Fig. 2.1 (a) and (b). Each product display had a feature sheet with a short list of the camera's relevant features, such as "optical zoom" or "megapixels". We also set up a service counter, where we instructed the shopkeeper to stand at the start of each interaction.

Participants were members of our laboratory and interacted with each other in English. Four fluent English speakers role-played as the shopkeeper. 10 participants, including 7 fluent English speakers, played the role of customer. Each customer took part in 10-20 interactions, for a total of 178 trials.

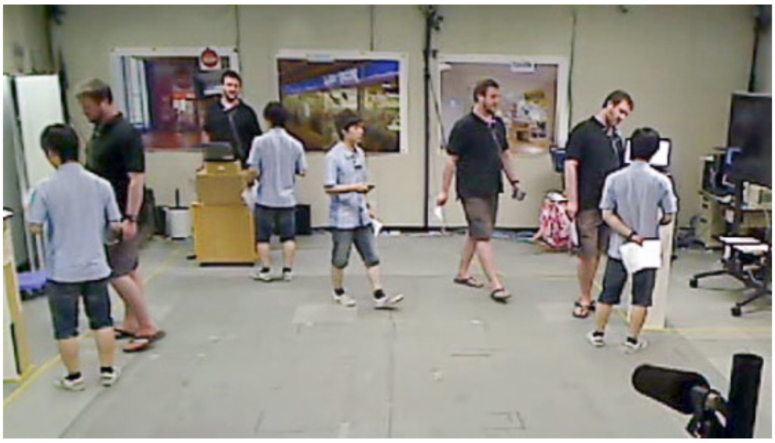
In each trial, the customer was instructed to role-play in one of the following scenarios: (1) a need-based customer, who is looking for a camera with a specific feature (4 trials), (2) a curious customer, who is interested in multiple cameras (4 trials), or (3) a window-shopping customer, who prefers to browse around alone (2 trials). In order to help the participant to naturally role-play as a specific type of customer, we gave the customer a different feature to look for each time. The shopkeeper was not informed



(a) Environment for our data collection



(b) Map of the room



(c) Interaction between a shopkeeper and a customer

Figure 2.1: The laboratory environment setup for this study.

of the chosen scenario, and was instructed to allow the customer to browse, to answer any questions the customer had, and to gently introduce products when appropriate, as shown in Fig. 2.1 (c).

Before the experiment, the participants were trained to use the Android phone and given a list of camera features to ask about. The shopkeeper was given a reference sheet containing a set of feature specifications for each camera. The practice trials were designed to help the participants became accustomed to using the Android phone and to illustrate the differences between the interaction scenarios.

The goal of the data collection was to capture repeatable interactions, so we restricted the scope of the scenario to focus on providing information about the cameras. For this reason, we asked the participants to keep the interactions simple by avoiding other topics, such as negotiating the price of the camera (e.g. “can you give me a better deal?”).

Furthermore, we found it necessary to remind participants not to make up new information that did not exist in our scenario. For example, if a shopkeeper participant was asked “what kind of warranty policy do you have?”, which was not defined in the scenario, they would have had to improvise an answer. These improvised responses would not be useful for learning because of inconsistency over time (in pre-trials, one shopkeeper participant said the store had a 1-year warranty policy on one occasion, but later said it was a 5-year warranty).

### 2.3.3 Example of human-human interaction

Within the defined scenario, the participants interacted in a free-form way, using natural conversational language, and a reasonable degree of variation in people’s phrasing and terminology was observed. Table 2.1 illustrates this variety with transcripts from two example trials by the same participant: (1) a need-based customer looking for a camera with large memory storage, and (2) a curious customer interested in cameras with good battery life.

## 2.4 Proposed Technique

### 2.4.1 Overview

We implemented a fully unsupervised data-driven strategy to enable a service robot to reproduce human behaviors using only captured data from human-human interaction. Our approach represents interaction data via several abstractions, as follows:

1. Customer **speech** is vectorized using Latent Semantic Analysis (LSA) and other text processing techniques (Sec. 2.4.2.1).

2. Shopkeeper **speech** is similarly vectorized, and it is then categorized into speech clusters representing lexically-similar, discrete utterances (Sec. 2.4.2.1).

Table 2.1: Examples from human-human interaction

| **Example of a need-based customer** |
|---|
| S: (Approaches customer) Hi are you looking for anything in particular today? |
| C: Yes I would like to… I am looking for a camera with good storage memory. |
| S: (Guides to Canon) Ok the Canon Rebel XTi can hold 10000 photos. |
| C: Ok, that is very good. What about the price? |
| S: This camera is $400. |
| C: I see. Is it heavy? |
| S: Yes, very heavy. |
| C: How much? |
| S: Like, a kilogram. |
| C: I see, that is very heavy. Well I will think about it. Thank you. (Leaves shop) |
| S: Sure, no problem. |
| **Example of a curious customer** |
| C: (Goes to Sony) Excuse me. |
| S: (Approaches customer) Yes sir how can I help you? |
| C: I am looking for a camera that I can use for a long time without changing the battery. |
| S: (Guides to Canon) Ok we have a couple of options for that; over here is the Canon Rebel XTi. It has a 7 hour battery life. |
| C: I see, and other possibilities? |
| S: (Guides to Panasonic) Other possibilities for long battery life are the Panasonic Lumix… this can run for 9 hours on standby. |
| C: So this is longer. What's the difference between these two? |
| S: This one is far worse in photo quality and it doesn't have a replaceable lens. |
| C: I see, so probably I am more interested in the other model. I will think a little bit about it. Thank you very much. (Leaves shop) |
| S: No problem sir. |

3. Customer and shopkeeper **trajectories** are segmented into stopped and moving segments, which are then clustered to identify typical stopping locations and typical motion trajectories (Sec. 2.4.2.2).

4. An **interaction state** is defined based on the relative positions of the customer and shopkeeper, based on a set of two-person spatial formations taken from other HRI and proxemics work (Sec. 2.4.2.3).

We then analyze the training data to identify discrete **actions**, comprised of speech and/or movement of the customer or shopkeeper (Sec. 2.4.3.2), and we train a machine learning classifier to predict the appropriate shopkeeper action output which follows an observed customer action input.

The **input** (Sec. 2.4.3.3) to the classifier is the processed training data – a vector consisting of the customer's speech vector, spatial states for the customer and shopkeeper (Sec. 2.4.3.1), and the current interaction state of the customer and shopkeeper.

The **output** (Sec. 2.4.3.4) is a discretized shopkeeper action comprised of a speech cluster combined with a target interaction state.

The top part of Fig. 2.2 shows an overview of how the training data is processed to generate an input vector (”input”) and the corresponding shopkeeper action vector (”label”) for training the machine-learning classifier (Sec. 2.4.4.1 - 2.4.4.3).

During real-time operation, the sensor data are processed in the same way as they were during training – a vector is built by combining the LSA vectorization of the customer utterance with the spatial and interaction states abstracted from motion data. This vector is input to the trained classifier whenever a customer action is detected. A shopkeeper action is then predicted, and the speech and spatial formation of the predicted action are executed by the robot (Sec. 2.4.4.4).

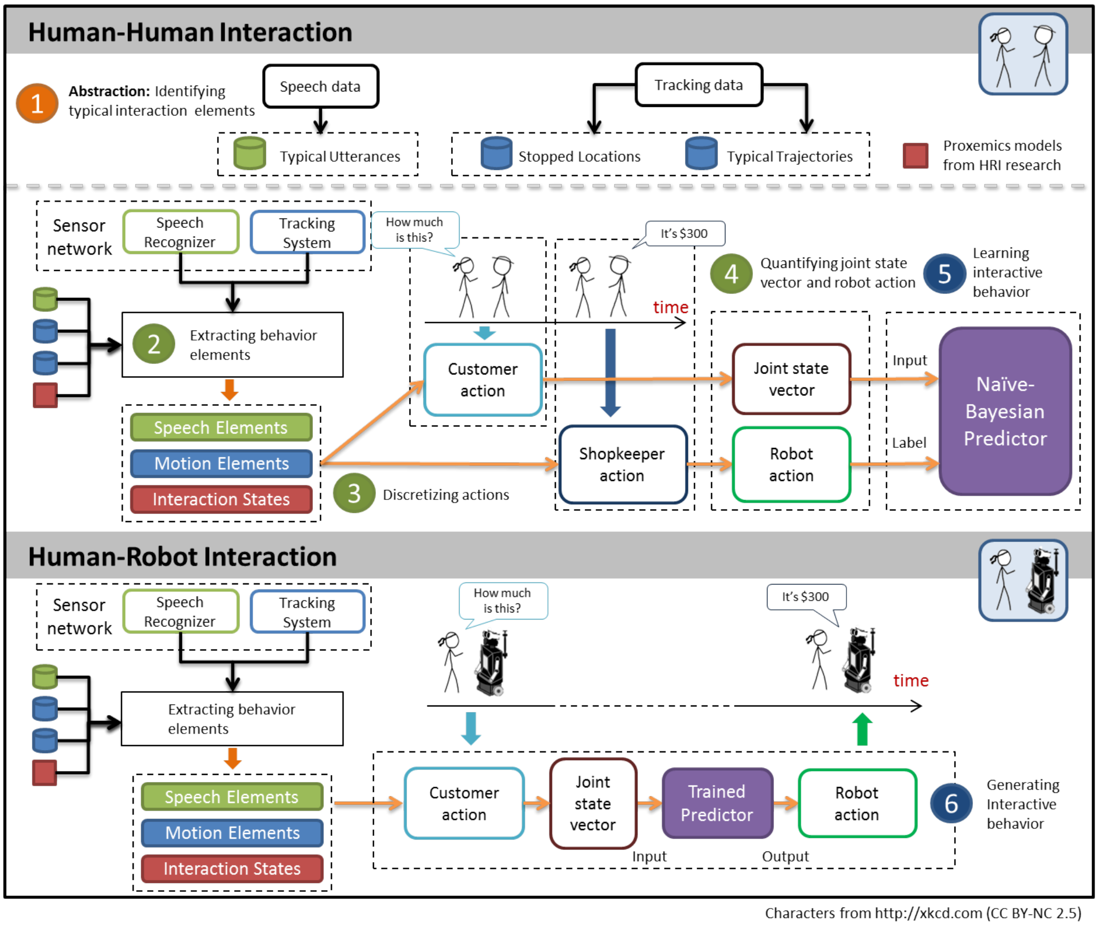
The bottom part of Fig. 2.2 illustrates the processing of the sensor data as an input to generate robot behavior in real-time.

The following subsections will explain the details of these abstraction and vectorization processes, as well as the setup of the learning algorithm itself.

### 2.4.2 Abstraction

One challenge of using a data-driven approach to learn from human-human interaction is that human behavior occupies a very high-dimensional feature space, even considering only speech and locomotion (social behaviors such as gaze, gesture, and facial expression are not considered in the current study). In practice, however, the variation of human behavior occupies only a small manifold within this high-dimensional space – people usually perform actions in predictable ways and follow common patterns. We introduce here a number of abstraction techniques designed to capture these patterns, in order to reduce the dimensionality of the learning problem and diminish the effects of sensor noise.

First, we perform unsupervised **clustering** to identify sets of typical actions in the training data. Clustering is performed for speech data to deal with the large amounts of



Characters from http://xkcd.com (CC BY-NC 2.5)

Figure 2.2: Overall procedure for human-human interaction (data collection) and human-robot interaction (online)

noise associated with speech recognition (Sec. 2.4.2.1), and also for motion trajectories observed by the tracking system, in order to identify typical stopping locations and motion paths in the environment (Sec. 2.4.2.2).

Next, we model each interaction as consisting of a sequence of stable **interaction states**, which last for several turns in a dialogue, recognizable by distinct spatial formations such as talking face-to-face or presenting a product. The modeling of interaction states helps to generate locomotion in a stable way, to specify robot proxemics behavior at a detailed level, and to provide context for more robust behavior prediction.

#### 2.4.2.1 Speech Clustering

A great deal of variation was present in the speech captured in our training data, including alternative phrasings, *e.g* "what is the price" versus "how much does it cost," as well as speech recognition errors, *e.g.* "how much does the scammer cost" rather than "how much does this camera cost?" The challenge of speech processing is to represent these utterances in a way that preserves the similarity between phrases with similar semantic meaning.

The strategy for processing speech elements is shown in Fig. 2.3. As soon as an utterance was captured, **speech recognition** was performed. We then **extracted keywords** using a cloud-based service and created a vectorized representation of the speech results and keywords using **Latent Semantic Analysis** (LSA).

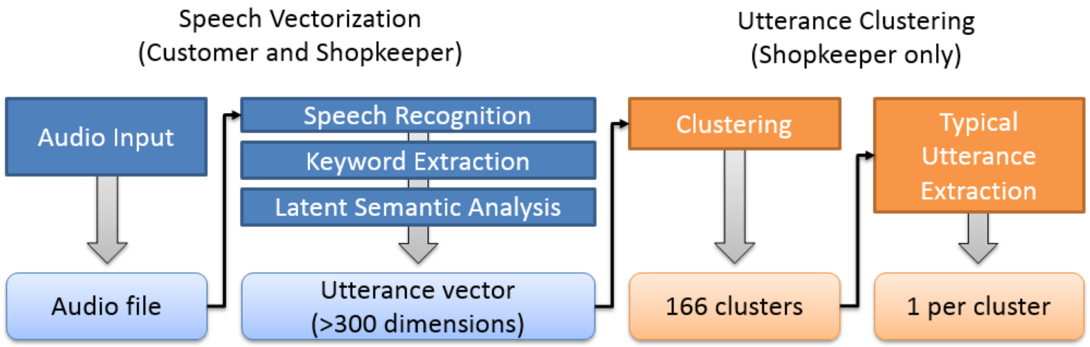


Figure 2.3: The abstraction of speech elements into typical utterances

Further processing was applied to shopkeeper's utterances only, with the goal of minimizing errors so that they could be used for generating robot speech. After vectorization of the utterances, we used unsupervised **clustering** to group them into clusters of similar utterances, and a **typical utterance** was then chosen from each cluster, to be used as content for synthesized speech output. Clustering was not applied to customer utterances, so that the information in the utterance vector could be kept for the purpose of prediction.

**Speech recognition:** For automatic speech recognition (ASR), we used the Google Speech API. An analysis of 400 utterances from the training interactions showed that

Table 2.2: Dimensions of utterance vectors

| | TF-IDF Dimension | LSA Dimension | Instances |
|---|---|---|---|
| **Customer Speech** | 7289 | 346 | 1194 |
| **Shopkeeper Speech** | 9181 | 353 | 1233 |

53% were correctly recognized, 30% had minor errors, e.g., "can it should video" rather than "can it shoot video," and 17% were complete nonsense, *e.g.* "is the lens include North Florida."

**Keyword extraction:** Phrases like "I am looking for a camera with large memory size" and "I am looking for a camera with large LCD size," have different meanings despite lexical similarity. To capture keywords in the phrases, we used AlchemyAPI [1], a cloud-based service for text analysis based on deep learning.

**Latent Semantic Analysis:** We created a vector to represent each utterance using Latent Semantic Analysis (LSA), a technique commonly used for classifying document similarity in text mining applications [70]. To achieve this, we performed several steps which are standard in text processing: we removed stop words, applied a Porter stemmer [102] to remove conjugations, enumerated n-grams (up to N=3), computed a term frequency – inverse document frequency (TF-IDF) matrix, and computed the singular-value decomposition of the TF-IDF matrix, truncating it to reduce the dimensionality of the space. The list of keywords returned for each utterance was separately processed using LSA, and those columns were added to the feature vector.

We chose the dimensionality for the truncated LSA matrix to achieve a 50% "share" (percentage of cumulated singular values) as described in [142]. The numbers of dimensions and instances for each group are presented in Table 2.2.

**Clustering of shopkeeper utterances**: We used dynamic hierarchical clustering [71] to group the observed shopkeeper utterances into clusters representing unique speech elements. 166 clusters were obtained.

**Typical utterance extraction:** From each shopkeeper speech cluster, one utterance was selected for use in behavior generation. We found that simply choosing the utterance closest to the centroid of the cluster was often problematic – sometimes this vector was not actually lexically similar to other utterances in the cluster and contained many errors, as shown in Fig. 2.4.

We instead choose the utterance with the highest level of lexical similarity to the most other utterances in the cluster, as this utterance would be the least likely to contain random errors. For each utterance, we compute the cosine similarity of its term frequency vector with every other utterance in the same cluster, and we sum these similarity values. The utterance with the highest similarity sum is chosen as the typical utterance.

---

[1] http://www.alchemyapi.com

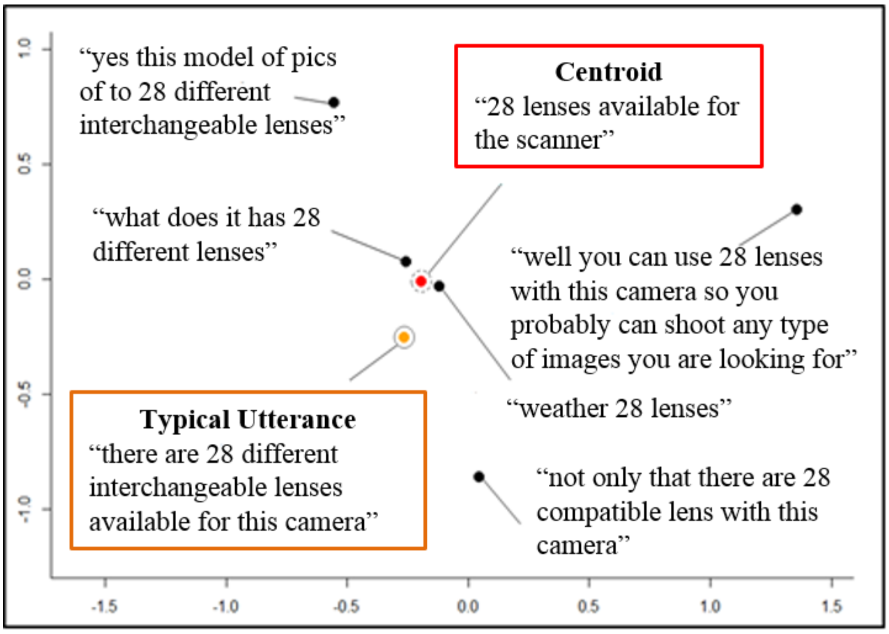


Figure 2.4: An example of typical utterance selection from a shopkeeper speech cluster (ID 292). The utterance vectors have been collapsed to two dimensions using multidimensional scaling (MDS) for visualization. The closest utterance to the centroid and the typical utterance chosen using our technique are shown.

#### 2.4.2.2 Motion Clustering

In the abstraction of motion elements, our primary objectives are (1) to identify common stopping locations in the social space, so that we can discretize our representations of people's motion in the joint state vector, and (2) to identify typical trajectory shapes so that we can estimate people's motion targets. We do so by analyzing and clustering the motion data to characterize the overall sets of stopping locations and motion trajectories that exist in the data.

Using the approach described by Guéguen [43], we analyzed the distribution of trajectories in the data set and selected 0.55 m/s as a threshold speed for trajectory segmentation. We then segmented all observed trajectories in the training data into "stopped" and "moving" segments, and clustered those segments to identify the typical **stopped locations** and **motion trajectories** present in the data set, as illustrated in Fig. 2.5.

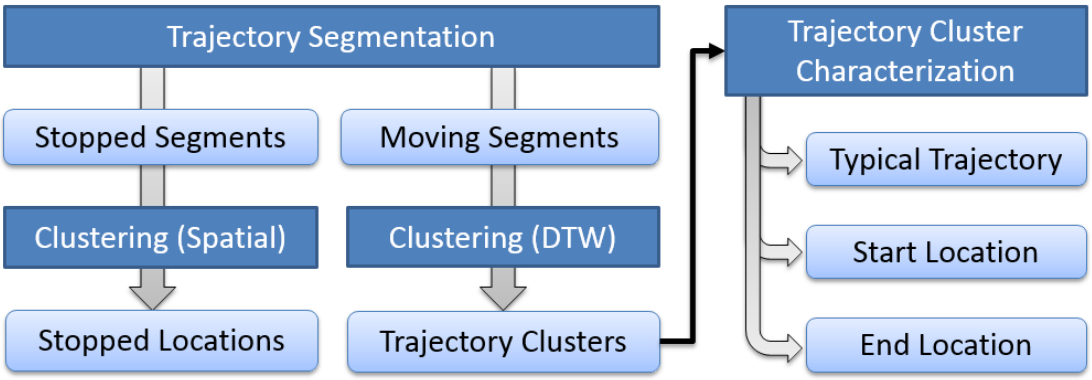


Figure 2.5: The abstraction of motion elements into stopped locations and trajectory clusters

**Stopped location:** The "stopped" segments were clustered spatially with k-means clustering to identify typical stopping locations, six for the customer and five for the shopkeeper. The centroid of each cluster was defined as a "stopped location". Usually, these points corresponded to significant locations such as the cameras or service counter, so for ease of explanation we will refer to these points by the names shown in Fig. 2.6.

**Trajectory clusters:** We clustered the moving segments into 50 trajectory clusters, separately for shopkeeper and customer, using k-medoid clustering based on distances computed between trajectories using dynamic time warping (DTW).

The medoid trajectory for each cluster was designated as its "typical trajectory", and the nearest stopped locations to the start and end points of that typical trajectory were identified. Fig. 2.7 shows some examples of the trajectory clusters.

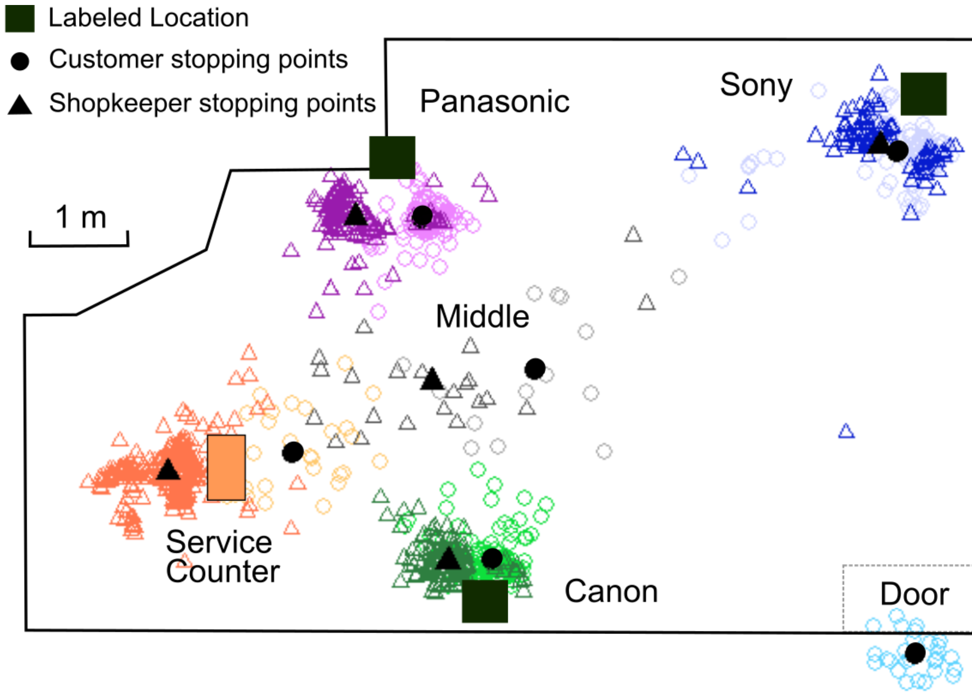


Figure 2.6: Customer and shopkeeper stopping locations. Empty circles and triangles represent individual customer and shopkeeper stopping locations in the raw dataset. Solid markers show the cluster centroids which define the abstracted “stopped locations”.

#### 2.4.2.3 Interaction States

We observed that the participants spent the majority of their time in a few static spatial formations, such as talking face-to-face or standing together at a camera. To capture this aspect of spatial behavior, we model each interaction as consisting of a series of interaction states characterized by common proxemic formations, such as talking face-to-face or presenting a product. The overall movement of the customer and shopkeeper can be seen as primarily serving as a means for transitioning between these interaction states. Fig. 2.8 presents example interaction state sequences observed in the training data.

HRI models have been developed for generating appropriate proxemics behavior in specific social situations such as initiating conversation [120] or presenting an object [145]. These models are useful abstractions, as they enable interaction states to be used not only to describe target destinations for movement, but also to specify proxemics constraints and other behavior at a detailed level for a robot.

In this work, we use three interaction states related to existing HRI models: *present object*, based on [145], *face-to-face*, based on interpersonal distance defined by Hall [44], and *waiting*, inspired by the modeling of socially-appropriate waiting behavior in [64]. Examples of these states are shown in Fig. 2.9.

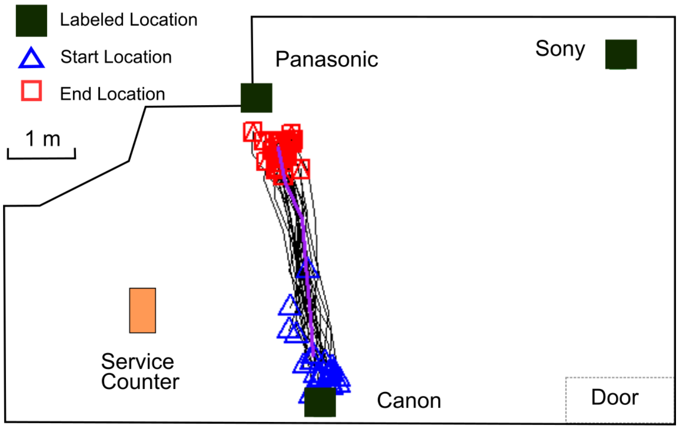
We created rules for identifying each of these interaction states, based on the distance between the interactants and their locations. If both interactants were at stopping locations corresponding to the same camera, the interaction state was categorized as present object. If they were within 1.5m of each other but not at a camera, it was modeled as face to face, and if the shopkeeper was at the service counter while the customer was not, the interaction state was defined as waiting.

In addition, we also identified the current target for a particular interaction state. The *state target* for "present object" can be either Sony, Panasonic, or Canon, whereas the *state target* for the interaction states "face-to-face" and "waiting" is 'none'.

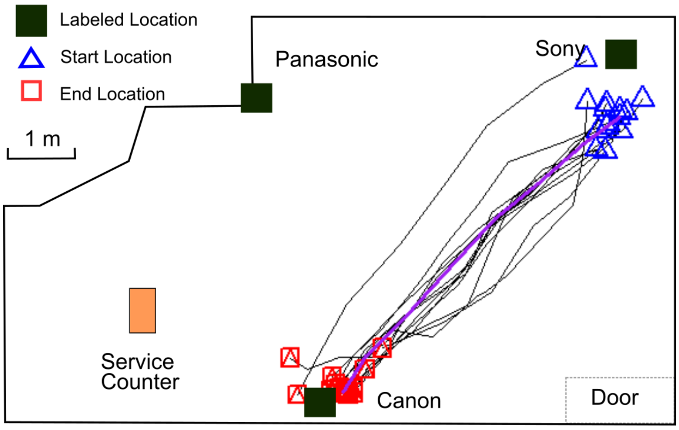
### 2.4.3 Vectorization

When processing time-series sensor data for offline training or online interaction, these abstractions are used for creating vectorized representations of discrete customer and shopkeeper actions, as shown in Fig. 2.10. First, **motion analysis** is performed based on a comparison with typical trajectories. It is then possible to **discretize actions** based on detections of movement and speech. Each customer action is represented by a **joint state vector** describing the abstracted state of both participants at the time of that action, and each shopkeeper action is represented by a **robot action vector** containing the necessary information for a robot to reproduce that action later.

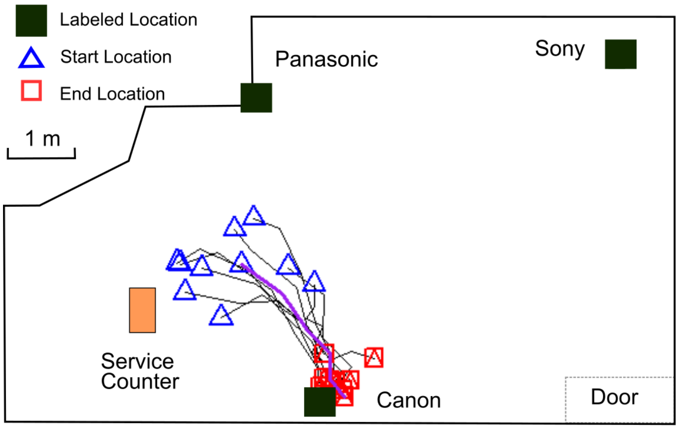
For all processes presented here, the sensor data is sampled at a constant rate of 1 Hz. Except where noted, the same techniques were applied to both the recorded training data and the live data from the online system.



(a) Cluster 24 (Moving from Canon to Panasonic)



(b) Cluster 27 (Moving from Sony to Canon)



(c) Cluster 35 (Moving from Service Counter to Canon)

Figure 2.7: Examples of customer trajectory clusters (from a total of 50 trajectory clusters). The medoid trajectories are highlighted in purple.

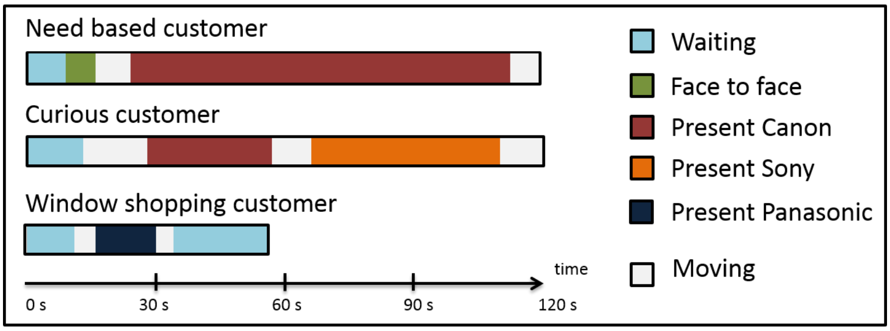


Figure 2.8: Examples of sequences of interaction states from training data for the 3 customer scenarios: curious, need-based, and window-shopping

#### 2.4.3.1 Motion Analysis

We characterize a person's motion using a vector containing three parameters: *current location*, *motion origin*, and *motion target*, corresponding to stopping locations from the clustering.

We identify whether a person is moving or stopped by applying the same speed threshold used in the offline trajectory analysis (Sec. 2.4.2.2). For stopped trajectories, *current location* is set to the nearest stopping location, and *motion origin* and *motion target* are "none".

For moving trajectories, *current location* is "none" and *motion origin* is set to the most recent *current location*. For the customer, the *motion target* field must be estimated, although as we will explain, estimation is unnecessary for the shopkeeper.

**Customer motion target:** To estimate the customer's motion target, we examine the similarity of the customer's trajectory to the typical trajectories identified in clustering, similar to the approach used in [56]. We compute the spatiotemporal distance between the customer's trajectory and each of the typical trajectories from the training data using DTW. The distance calculated for each trajectory cluster is then weighted according to the number of instances in that cluster, and probabilities are summed for trajectories that terminate at the same end location. The motion target is output once the probability of some result is above 50%, usually attained within 2-3 seconds.

**Shopkeeper motion target:** Estimation of the motion target through sensor data is unnecessary for the shopkeeper. Since we always know the robot's target destination with certainty, based on the commands sent to the robot, the shopkeeper's motion target in the training data should also reflect this knowledge of the intended motion target. In order to do so for the training data, we can determine the shopkeeper's actual motion target at any time by looking ahead in time to observe their eventual destination, rather than relying on estimation from the sensor data. By doing so, the shopkeeper motion target from the training data and from real-time data will be consistent.

#### 2.4.3.2 Discretizing Actions

Discrete "customer actions" and "shopkeeper actions" are defined when one of the participants speaks and/or begins moving to a new location. Speech actions are defined at the moment the speech recognition result is received, and motion actions are defined at the moment a motion target is determined. Customer and shopkeeper events are received within the same 1-second interval are classified as two separate events, so no event can contain both customer and shopkeeper speech.

#### 2.4.3.3 Joint state vector (Input)

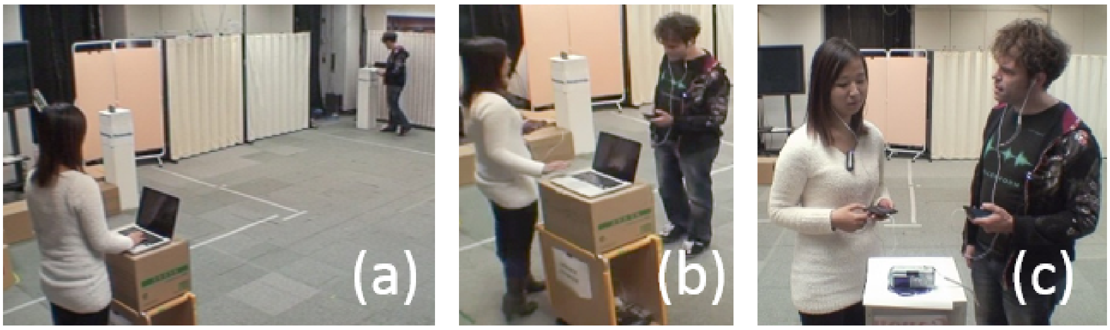


Figure 2.9: Typical interaction states: (a) Waiting: one person is at a designated waiting area and interactants are not near each other, (b) Face to face: both people near and facing each other, but not near an object, (c) Present object: both people stopped near an object.

When a customer action is detected, the state of both interactants is recorded in a *joint state vector*. This vector will be used for training the predictor to identify the most appropriate robot action to perform. The features in the joint state vector are shown in Fig. 2.10 (a). It includes the customer speech vector (including LSA vectors for both the utterance and keywords, 346 dimensions in total), customer and shopkeeper spatial states (each consisting of *current location*, *motion origin*, and *motion target*), and interaction state (*spatial formation* and *state target*).

#### 2.4.3.4 Robot action vector (Output)

When a shopkeeper action is detected, it is represented in a *robot action vector*, which can be translated later into commands for the robot. In our case we are concerned with reproducing only speech and locomotion, so the robot action vector contains two properties: speech (consisting of a *speech cluster*) and interaction state (*spatial formation* and *state target*), as shown in Fig. 2.10 (b).

**Robot Speech:** This field contains information to enable the robot to reproduce a shopkeeper utterance. It is only populated if the shopkeeper action contains a speech component; otherwise, it is left blank.

*Definition:* Directly using the raw text output from speech recognition is not appropriate for generating robot speech, because often it contains speech recognition errors. For this reason, we record the ID of the shopkeeper speech cluster containing the detected speech. For example, if the recognized utterance is "what does it has 28 different lenses", cluster ID 292 would be chosen as the representative shopkeeper speech cluster, as illustrated in Fig. 2.4.

*Generating robot behavior:* As described in Sec. 2.4.2.1, a typical utterance is extracted from each shopkeeper speech cluster, which is expected to contain fewer random errors than a typical instance of recognized speech. To generate a robot speech behavior from a cluster ID, we use this typical utterance as the text to be sent to the robot's speech synthesizer. In the above example, the chosen robot speech would be "there are 28 different interchangeable lenses available for this camera".

**Target Interaction State:** Recall that the interaction state described in Sec. 2.4.2.3 encapsulates the proxemic formation of the two interactants at a given time. We can use this information to generate robot motion by recording the "target interaction state" of the shopkeeper.

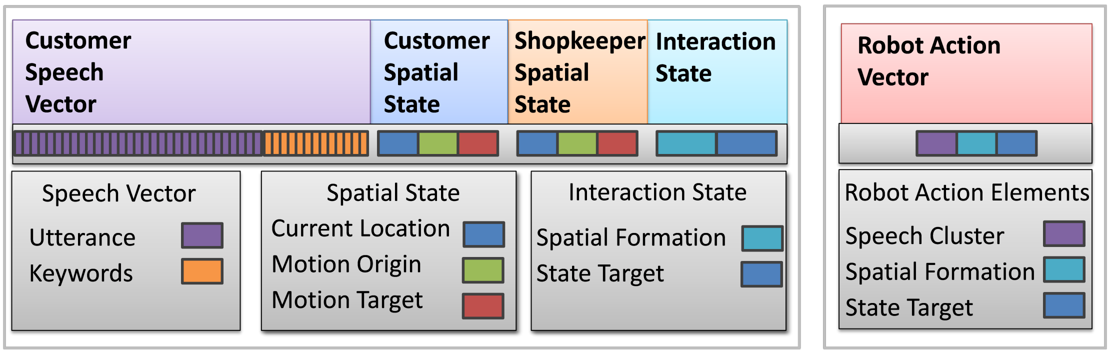


Figure 2.10: (Left) Features in joint state vector. (Right) Features in robot action vector.

*Definition:* If the shopkeeper is not moving at the time the action is detected, then the shopkeeper's current interaction state is recorded. If the shopkeeper is moving, then we look ahead in time to determine the shopkeeper's destination as described in Sec. 2.4.3.1. We then determine the "target interaction state" by evaluating the interactants' spatial formation at the time when the shopkeeper arrives at the destination.

The interaction state is identified in the same way as described in Sec. 2.4.3.3, except that to accommodate the case where the shopkeeper is leading the customer and arrives first, we classify the target state as "present object" if either the customer's *current location* or the customer's *motion target* are the same object as the shopkeeper's *current location*.

*Generating robot behavior:* Then, to generate a robot behavior in the online system we can simply compare the robot's current location with the location necessary to achieve the target interaction state, and command it to move if necessary. For *waiting*, this target location will be the service counter; for *present object*, the target location will be the object of interest; and for *face-to-face*, the target location will not be a fixed location but rather a point in front of the customer. If the robot is not already at the target location, we command the robot to drive to a point near that location. The precise $x, y$ position near the target location is determined by using the HRI proxemics model associated with the target interaction state.

## 2.4.4 Learning and execution of interactive behaviors

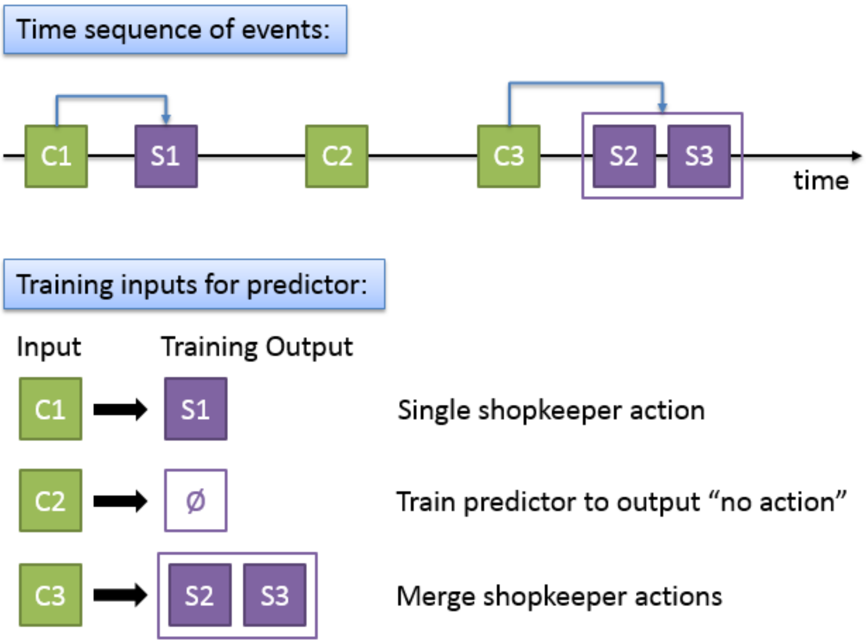


Figure 2.11: Example time sequence of customer and shopkeeper actions.

To use machine learning to determine which robot behaviors should be performed in response to which human actions, we examine the discretized actions to **identify action pairs**, that is, sequential pairs of customer and shopkeeper actions, in the training data. For each action pair, we **train a predictor** using the joint state vector and robot action vector corresponding to the customer and shopkeeper actions. Finally, this predictor is used in the online phase to **generate robot behaviors** in response to detected customer actions.

### 2.4.4.1 Identifying Action Pairs

By examining the time sequence of detected actions (see Sec. 2.4.3.2), we identify correspondences between customer actions and subsequent shopkeeper actions. However, social interactions are not always cleanly divided into action-response pairs, e.g., when two customer actions or two shopkeeper actions occur in a row. Consecutive shopkeeper actions are combined according to a set of rules, and customer actions that are not followed by a shopkeeper action are associated with “no action” for purposes of training the predictor.

Fig. 2.11 shows an example time sequence of customer and shopkeeper actions. The first two, C1 and S1, illustrate the usual case of a customer action followed by a shopkeeper action, and these are paired as training inputs and outputs for the predictor. Customer action C2 is not followed by a shopkeeper action, so it is paired with “no action”. The third customer action is followed by two shopkeeper actions, which are then merged to produce a single shopkeeper action.

Recall that each robot action is comprised of an utterance (166 possibilities) and a target interaction state (5 possibilities). After merging shopkeeper actions, we translate each of the shopkeeper actions into a robot action vector, as described in Sec. 2.4.3.4. The final list of robot action vectors for our data set contained 467 distinct combinations of utterance and interaction state.

### 2.4.4.2 Modeling Delay

There is a natural delay time between customer actions and shopkeeper responses, and if the robot responds too quickly or too slowly, it is unnatural. To reproduce the delay time between customer actions and responses from the shopkeeper, we calculated the average time delay between customer and shopkeeper actions from the training data corresponding to each robot action, and we constructed a lookup table mapping robot actions to average delay times.

For most robot actions, such as answering direct questions, the delay time was usually in the range of 0 - 2.5 seconds. For some behaviors longer pauses were observed. For example, when a customer entered and moved directly to the Sony camera while saying nothing, the system predicted that the robot should approach and offer assistance, after a delay of 17 seconds. If the customer performed another action during this time, the robot responded to that action. In this way, the robot was able to respond to long pauses which occurred, *e.g.*, in the “window-shopping” scenarios.

### 2.4.4.3 Training the Predictor

Once all action pairs in the training data have been identified, we train a naïve Bayesian classifier, using the joint state vector for each customer action as a training input and the subsequent robot action vector corresponding to the shopkeeper action as its training class.

The naïve-Bayesian classifier is a generative classification technique, which uses the formula below to classify an instance that consists of a set of feature-value pairs.

$$a_{NB}=\underset{a_j\in C}{\operatorname{argmax}} P\left(a_j\right)\prod_i P\left(f_i=v_i|\,a_j\right) \tag{2.4.1}$$

$a_j$, denotes a robot action, and $f_i$ denotes a feature in the joint state vector. The naïve-Bayesian classifier picks a robot action, $a_{NB}$, that maximizes the probability of being classified to the robot action given the value $v_i$ for each feature $f$.

Each feature $f_i$ in the joint state vector is multidimensional, consisting of a set of terms $t_{ik}$. For example, the customer speech vector has 346 dimensions, whereas the customer spatial state only has 21 dimensions. Thus, we can rewrite the classifier equation to consider the partial matches between the values for each feature, as in Eq. 2.4.2, where the conditional probability of each term of each feature, given a robot action $a_j$, is computed in the training phase:

$$\begin{aligned} v_i &= \{t_{i1}, t_{i2}, \ldots, t_{im}\} \\ a_{NB} &= \operatorname*{argmax}_{a_j \in C} P\left(a_j\right) \prod_i \left( \prod_k P\left(t_{ik} \; appears \; in \; f_i \,\middle|\, a_j\right) \right)^{w_i} \end{aligned} \tag{2.4.2}$$

We would like to give higher priority to values in the features that are more discriminative in classifying robot action. Gain ratio tells us how important a given feature in the joint state vector is. Therefore, $w_i$, calculated from the gain ratio of each feature, is added as the weighting factor for the classifier.

#### 2.4.4.4 Generating Robot Behaviors

During live interaction between a human customer and the robot shopkeeper, the sensor network records the customer's motion and speech at one-second intervals.

When a customer action is detected, we query the trained naïve Bayesian predictor, passing in the joint state vector corresponding to the social state at that time. The predictor will then output either the ID of one of the 467 robot actions, or it will predict "no action". If a robot action is specified, the system waits for the time specified in the delay table corresponding to that action, and then commands are sent to the robot to move to a destination or speak an utterance.

When the robot action includes an interaction state of "*present object*" or "*face-to-face*", the precise target position is computed according to that formation's proxemics model. While in motion, the robot projects the future position of the customer and recalculates a target location according to the proxemics model every second until it arrives. Some examples of this calculation are illustrated in Fig. 2.12.

In this example, the first target interaction state is "Present Camera 1", shown in Fig. 2.12(a). The robot projects the customer's destination to be X1, so it computes a target destination to point O1. The next target interaction state is "Present Camera 2". In Fig. 2.12 (b), the robot first projects the customer to be moving towards X2, so it begins moving towards point O2. However, in Fig. 2.12 (c), the customer chooses to move to a different location than predicted. The robot dynamically updates its path to move to point O3.

### 2.4.5 Example of Behavior Execution

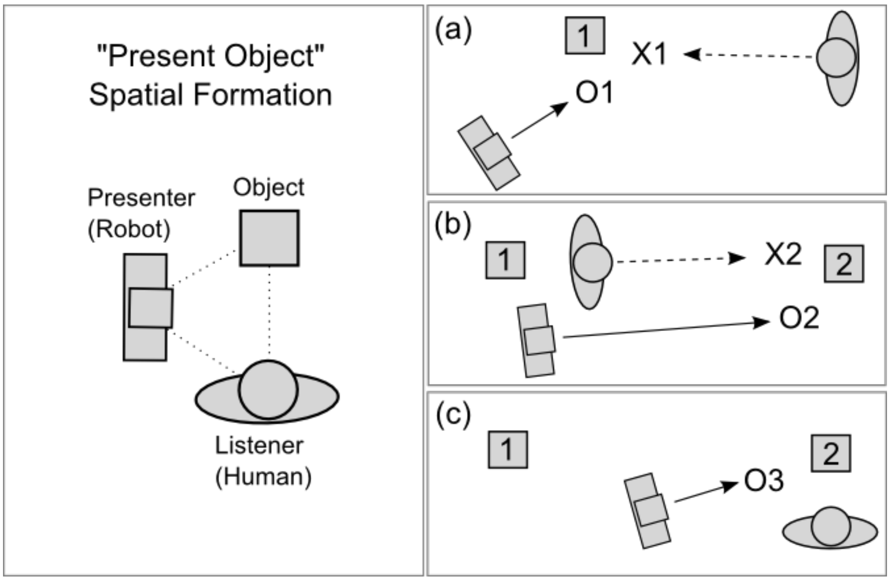


Figure 2.12: Left: Illustration of the target spatial formation corresponding to the “present object” interaction state. Right: Examples of dynamic path planning to achieve the “present object" formation. “X” represents the projected future position of the customer, and “O" represents the calculated target position of the robot in response.

Fig. 2.13 shows an example of a prediction from a live interaction with a robot. In this example, the customer approaching the shopkeeper at the service counter is detected as a customer action, and the predictor is queried with the joint state vector shown in the figure. The predicted robot action consists of an utterance with cluster ID 170 paired with an interaction state, “Present Sony”. The recorded delay time corresponding to “170-Present Sony” action is 2.75 seconds, so the system waits for that duration before executing an action. Because the current interaction state is “waiting” and the target interaction state is “Present Sony”, the robot starts moving to Sony. A speech command is sent to the robot containing the typical utterance from the selected speech cluster, which in this case causes the robot to speak, “over here we have my favorite which is the Sony NEX 5 which is a mini SLR and has 28 replaceable lens”.

## 2.5 Evaluation Experiment

We conducted a comparison experiment to evaluate the quality of the robot’s behavior in live interactions. Because we consider the proposed abstraction technique to be the main contribution which makes it possible to learn interactive behaviors despite high sensor noise, we compared two conditions: (a) *proposed*, using the abstraction techniques including clustering and interaction states described in Sec. 2.4, and (b)

Table 2.3: Differences between proposed system and without-abstraction system

| | **Proposed system** | **Without-abstraction system** |
|---|---|---|
| **Clustering** | | |
| | Cluster shopkeeper speech | No clustering |
| | Cluster motion data | |
| **Vectorization** | | |
| | Motion target prediction based on trajectory clusters | Motion target prediction based on mean motion direction |
| | Abstracted locations | Raw position data |
| | Interaction states used | No interaction states |
| **Predictor** | | |
| | Naïve-Bayesian predictor to select an abstracted action | Nearest neighbor-predictor to select an instance to reproduce |
| **Robot action generation** | | |
| | Motions generated based on target interaction state | Motion generated directly from a shopkeeper motion instance |
| | Utterances generated from shopkeeper speech clusters | Utterances generated directly from a shopkeeper speech instance |

*without-abstraction,* a similar technique we developed that does not use our abstraction techniques.

### 2.5.1 Comparison system

We designed the *without-abstraction* system to be similar to other state-of-the-art data-driven techniques for generating interactive robot behaviors. For example, Admoni et al. [1] developed a system that matches observed data in real-time to the nearest example from human-human training data to select a robot behavior, following the idea that people learn to communicate by mimicking observed behavior in a given situation.

Thus, we created a modified version of our system which also uses the observed sensor data in real-time to find the most similar example from the training data. If our data were not susceptible to noise, the behavior generated by the *without-abstraction* system would have represented exactly what a human shopkeeper had done in a similar situation. The differences between the *proposed* and *without-abstraction* systems are described here and summarized in Table 2.3.

**Speech elements:** Speech is captured and processed using the same standard text processing techniques in both systems. However, no clustering is performed on the shopkeeper's speech in the *without-abstraction* system, so shopkeeper utterances must be generated directly from the raw speech recognition results captured in the training data. Keyword extraction is also not used in the *without-abstraction* system, because its purpose is to assist with clustering of shopkeeper speech.

**Motion elements:** Our proposed technique uses the results from trajectory clustering to define stopping locations and to anticipate a person's motion target. For the *without-abstraction* system, a person's stopping location is represented by their raw $x, y$ position, rather than the nearest stopping point cluster. When moving, a person's motion target is estimated based on their motion direction, rather than using our technique of comparison to the trajectory clusters. Finally, the set of possible motion targets is defined manually for the *without-abstraction* system, rather than using clustering results (we defined five points: the three cameras, the door, and the service counter).

To estimate a person's motion target, the person's mean motion direction $\theta_{motion\_dir}$ is calculated over the last 3 seconds, and the *motion target* is calculated as the $x, y$ position of the nearest object to the mean motion direction from their position in the environment.

$$\begin{aligned}\text{motion target} = \ & \arg\min(\theta_{motion_{dir}} - \theta_{obj_n} \\ & : obj_n \in all\ objects\ in\ environment)\end{aligned} \tag{2.5.1}$$

**Feature vector:** The stopping locations identified in the clustering phase are not available in the *without-abstraction* system, so feature vectors include the following 5 features: the customer's and shopkeeper's current $x, y$ coordinates, the customer's projected $x, y$ motion target, the shopkeeper's actual $x, y$ motion target, and the LSA vector representation of the customer's speech. Interaction state was not included in the feature vector for the *without-abstraction* system.

**Prediction:** The predictor from the *proposed* system cannot be used in the *without-abstraction* system – since shopkeeper utterances are not clustered, there is no set of discrete robot actions to be trained. Instead, we created a "nearest-neighbor predictor" – whenever a customer action is detected, the current raw feature vector is compared to the feature vectors from all customer actions in the training data. The best match is identified, and the subsequent shopkeeper action from the training data is returned as a robot action. Robot actions in this case have two properties: motion target (if moving), and utterance text (if speaking). A lookup table for delay time between the customer and shopkeeper actions was also created in the same way as the proposed system.

For our dataset, the set of customer action vectors consisted of 1636 entries in 330 dimensions, so a k-d tree [6], was used to speed up the nearest-neighbor comparisons.

**Robot behavior generation**: Robot behaviors are generated directly from the specific instance of shopkeeper behavior output by the nearest-neighbor predictor. For movement, the robot moves directly towards the $x, y$ position where the shopkeeper had moved to in the matched instance, instead of using interaction state to generate the target. For speech, the robot speaks the exact phrase captured by speech recognition in the matched instance.

### 2.5.2 Hypotheses

In the comparison experiment, we made the following hypotheses about the effects of our abstraction techniques (clustering and modeling of interaction states) in the *proposed* system, compared with the *without-abstraction* system:

**Speech clustering:** Clustering of shopkeeper utterances will produce more correct utterance behaviors in the robot, because the act of clustering and our technique for typical utterance extraction will reduce the effect of noise in the captured utterances.

**Stopping point clustering:** Representing spatial locations based on abstracted stopping point clusters, rather than as raw positions, will lead to more efficient learning through abstraction. This will also be more robust to sensor noise, since the influence of noise is incorporated in the clustering step.

**Trajectory clustering:** Estimation of motion target will be more accurate when similarity to clustered trajectories is used, compared with raw extrapolation of velocity. This will lead to more appropriate responses to customer motion from the robot.

**Interaction states:** The modeling of movement in terms of transitions between long-term-stable interaction states will result in more reliable locomotion behaviors than reproducing individual movement events.

Based on these hypotheses, we chose to test the following predictions for the comparison between the *proposed* system and the *without-abstraction* system:

1. **Correctness of wording:** The robot will produce more correct wording in the *proposed* system.
2. **Consistency between speech and movement:** The robot's speech and movement will be more consistent with each other in the *proposed* system.
3. **Appropriateness of robot actions:** The robot will respond more appropriately to the customer's actions in the *proposed* system.
4. **Social-appropriateness:** The robot's behaviors will be more socially-appropriate for its role as the shopkeeper in the *proposed* system.
5. **Overall evaluation:** The overall evaluation of the robot's behaviors will be better in the *proposed* system.
6. **Robustness:** The *proposed* system will be more effective at generating appropriate robot behaviors even when recognition errors occur.

### 2.5.3 Experiment Setup

#### 2.5.3.1 Participation

A total of 17 paid participants (11 male and 6 female, average age 34.42, s.d. 13.30) played the role of customer in the experiments. All of them were fluent English speakers (9 North and South Americans, 7 Europeans, 1 Russian).

#### 2.5.3.2 Environment

The experiment was conducted in the same camera shop setting used for the data collection, with three digital cameras displayed in an 8m x 11m experiment space. The same sensor network was used for tracking, and the participants communicated with the robot using an Android phone.

#### 2.5.3.3 Robot Platform

For this experiment, we used Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, a wheeled base, and a speaker that can output synthesized utterances.

Robovie is capable of moving at a speed of 0.7 m/s. For its motion planning, the dynamic window approach (DWA) was implemented to avoid obstacles [32].

Implicit behaviors were implemented into the robot, where the robot makes small arm and head movements while idling, speaking, and moving [120]. Automatic face-tracking of robot's interaction partner was also implemented, and the robot followed the customer with its gaze during all interactions.

#### 2.5.3.4 Procedure

We compared the robot's performance between two conditions: *proposed* and *without-abstraction*, and each participant was asked to role-play for 8 trials in each condition. As in our data collection, participants played each of the following roles: a need-based customer (3 trials), a curious customer (3 trials), and a window-shopping customer (2 trials). The order of the conditions was counterbalanced and the order of the trials within each condition was randomized.

As in our data collection, participants were asked to pretend to be a first-time customer in the camera shop for every trial and the participants performed scripted interactions before the experiment to become familiar with the Android phone interface and confirm their understanding of the instructions.

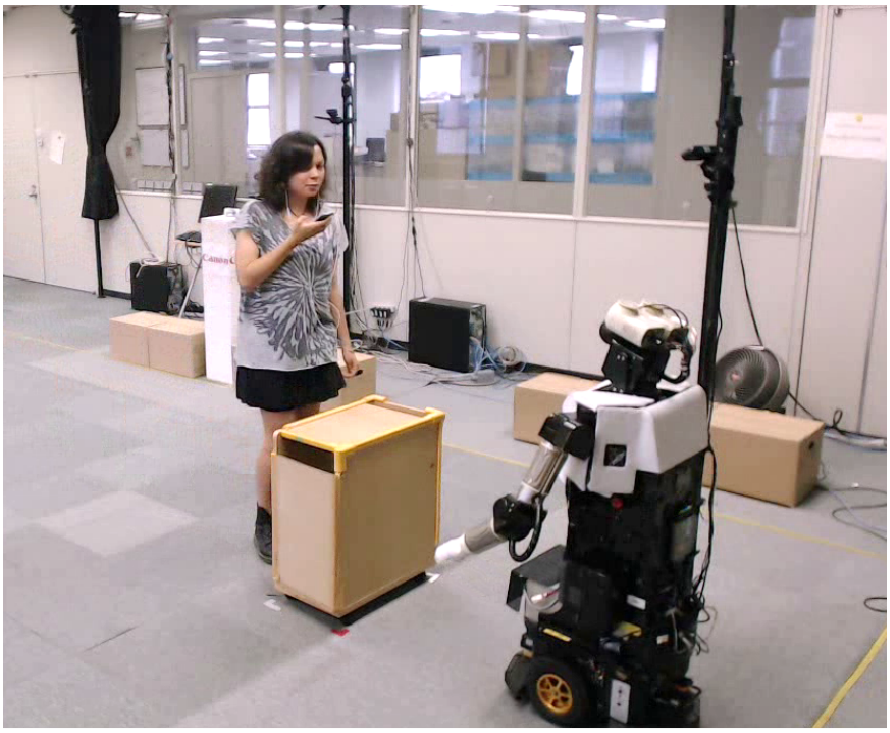
After the 8 trials in one condition were completed, the participant answered a questionnaire. The procedure was repeated with the remaining condition (*without-abstraction* or *proposed*). At the end of the experiment, the participants were interviewed to gain a deeper understanding of their opinions.

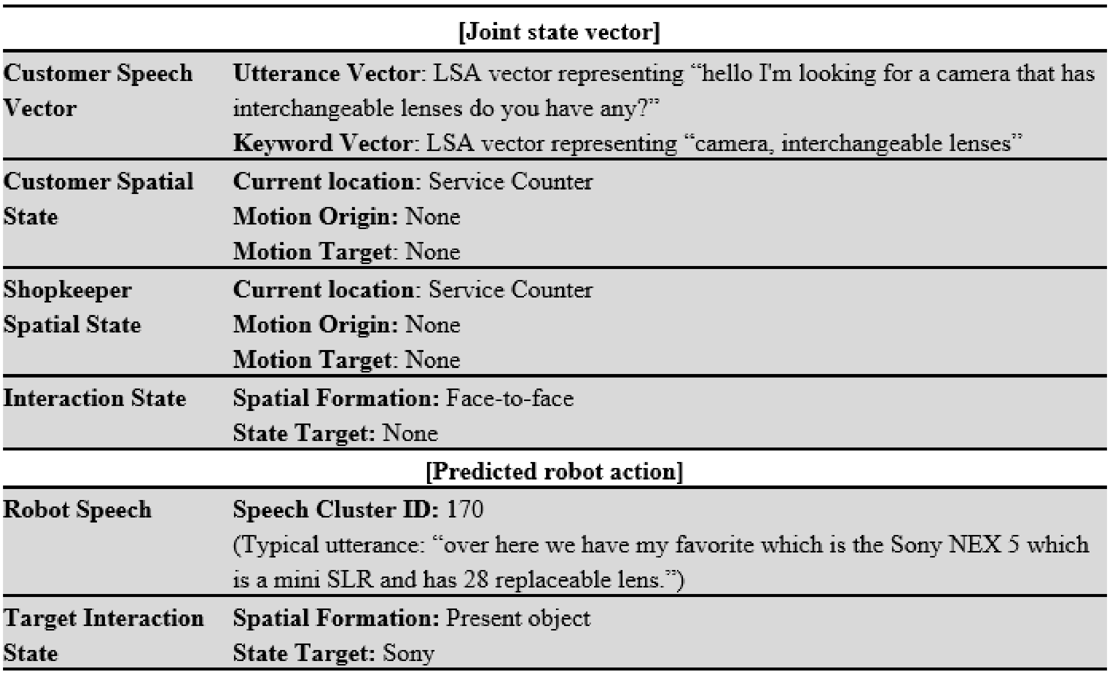
Examples of interactions from the experiment using the *proposed* system can be seen in the video attachment.

### 2.5.4 Measurement

#### 2.5.4.1 Questionnaire

The participant rated the following items on a 1-7 scale (1 being very negative and 7 being very positive for the respective items) in a written questionnaire:





| | |
|---|---|
| **[Joint state vector]** | |
| **Customer Speech Vector** | **Utterance Vector**: LSA vector representing "hello I'm looking for a camera that has interchangeable lenses do you have any?"<br>**Keyword Vector**: LSA vector representing "camera, interchangeable lenses" |
| **Customer Spatial State** | **Current location**: Service Counter<br>**Motion Origin:** None<br>**Motion Target**: None |
| **Shopkeeper Spatial State** | **Current location**: Service Counter<br>**Motion Origin:** None<br>**Motion Target**: None |
| **Interaction State** | **Spatial Formation:** Face-to-face<br>**State Target:** None |
| **[Predicted robot action]** | |
| **Robot Speech** | **Speech Cluster ID:** 170<br>(Typical utterance: "over here we have my favorite which is the Sony NEX 5 which is a mini SLR and has 28 replaceable lens.") |
| **Target Interaction State** | **Spatial Formation:** Present object<br>**State Target:** Sony |

Figure 2.13: Example of a prediction in a live interaction.

1. Correctness of the wording of the robot's utterance:
2. Consistency of the robot's speech and movement
3. Appropriateness of the robot's response to the participant's action
4. Social appropriateness of the robot's behaviors as its role as the shopkeeper
5. Overall evaluation

In the experiment, the robot may give an answer to the customer's question that makes sense, but may not necessarily be accurate. For example, if the customer asks "how much is this camera", the robot may respond with "$600" instead of the correct answer, "$300". Because knowledge of these errors could affect the participant's evaluation of the robot, we informed participants about any informational errors the robot made before they filled out the questionnaire in each condition.

#### 2.5.4.2 Interaction analysis

In our scenario, the human shopkeeper greeted the customer, answered customer's question, and said farewell to the customer when he left the shop. Similarly, the robot also learns to greet the customer, answers customer's questions (correctly), and says goodbye to the customer at the end of the interaction. Thus, the robot aims to imitate the decisions made by people in similar situations in the training data. We believe that by imitating observed actions, we can generate behavior that would be equivalent to a humanâĂŹs, without knowing the purpose or semantic meaning of those behaviors at all.

Thus, to evaluate the performance of the robot, we conducted a detailed action-by-action analysis of the robot's behavior by asking a coder, blind to the experimental conditions, to examine each action (speech or movement) made by the participant, and to judge whether the robot's response to that action was appropriate. The coder was shown examples of acceptable and unacceptable behavior in order to calibrate expectations. Examples of unacceptable behavior included answering a question incorrectly, or failing to guide a customer to a camera when asked to do so. From this evaluation, we calculated*behavior correctness* for each condition, for each participant.

A separate evaluator examined all of the customer speech events in each trial and recorded the number of correct and incorrect speech recognition results. We defined *ASR correctness* by whether the sentence-level meaning of the ASR result was understandable or not. Though some ASR results contained word errors, they were judged as "correct" if the utterance itself was still understandable on a sentence-level. For example, given that the customer said "thanks a lot", the ASR result "thanks a lots" would be considered correct, whereas "insulet" would be considered as incorrect. Further analysis of the speech recognition accuracy can be found in the Appendix.

The *ASR correctness* was then compared with the *behavior correctness* to evaluate the robustness of the behavior generation technique to recognition noise.

## 2.5.5 Results

### 2.5.5.1 Observations

It was quite fun for us to watch the robot acting autonomously – since the learned rules created some interesting variations of behavior, we never knew exactly what how the robot would respond to any situation. Most of the robot's behaviors were executed well - the robot was able to move with the customer to appropriate locations and answer most questions correctly. Although it did make some errors, it was often able to recover and continue the interaction. Many of the participants commented that they really enjoyed the interactions. Table 2.4 shows an interaction example from the experiment.

If the customer was looking for a particular camera feature (*e.g.* interchangeable lens), the robot usually responded correctly, guiding them to a camera with that feature and introducing the camera. The robot also answered most questions about camera features correctly, even though the customers asked in different ways. For example, one customer asked, "this one comes in red, right?", and another customer asked "what color do you have for this?", and the robot was able to answer appropriately by saying "we have red and silver available" to both customers. Likewise, the robot correctly gave the weight of the camera in response to "how much does this camera weigh?" and "excuse me is this camera heavy?"

Sometimes the robot responded correctly despite speech recognition errors. The robot gave correct answers to questions such as the following (correct phrasing in brackets): "Amanda um, and uh,] how much does it weigh?", "I'm sorry does this camera have optimism [optical zoom]?" "How many car what color is coming? [how many, er, what colors does this come in?]", "Skewes me [excuse me] what color does a scammer [this camera] come in?", "how much does a camel [this camera] weigh?", and "I say in the is it a popular vote [I see, and uh, is it a popular model?]". Many recognition errors were fairly common, such as "scammer" or "camel" for "camera", and "OCD" for "LCD", and the system appears to have learned to treat these words as synonyms.

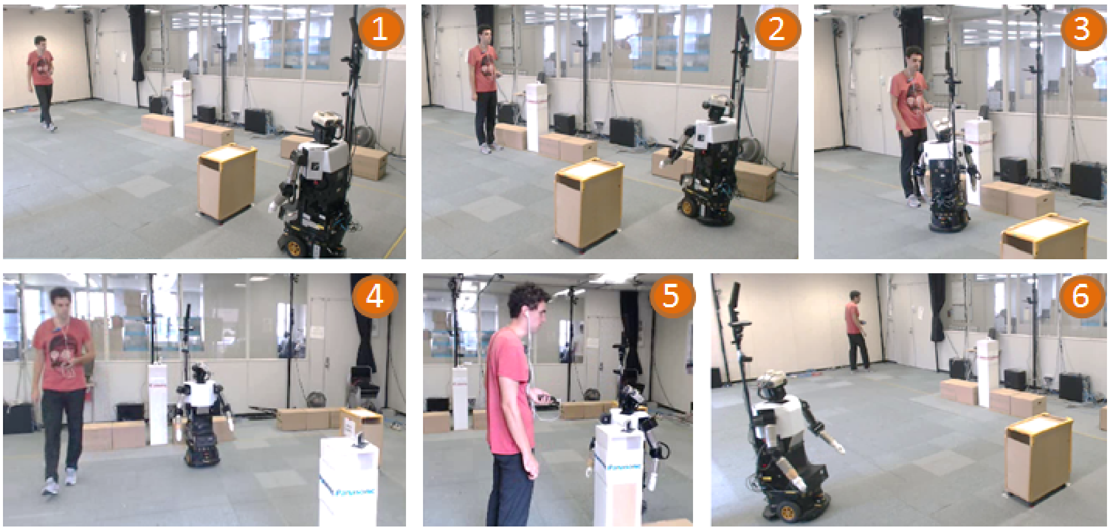
Sometimes it failed to respond correctly due to speech recognition errors. For example, when a customer asked "could you tell me how much this Lumix costs?", the word "Lumix" was recognized as "LINE X", and the robot responded, "yes, sir." Then the customer rephrased his question, "could you tell me how much this is?" and the robot answered correctly. When a customer repeated or rephrased their question, the robot usually responded correctly the second time.

The robot's utterances sometimes contained minor errors, as can be seen in the example in Table 2.4, although some of these mistakes sounded phonetically correct. For example, the robot sometimes said "my I help you?" when the customer entered the shop, yet none of the customers noticed the mistake.

The robot was also able to respond to the customer's motion – when a customer entered and immediately approached the service desk, the robot would greet them immediately, whereas if they walked to one of the cameras first, it would often let them browse for a while before speaking.

When the customer thanked the shopkeeper and left, the robot would respond with

Table 2.4: Example of the robot interacting with a customer



| | |
|---|---|
| (1) | **Customer** walks into the shop<br>**Robot**: "hi can I help you with anything" (at service counter) |
| (2) | **Customer** stops at Canon and says "yes I'm looking for a camera with large memory storage"<br>**Robot** approaches customer, saying "yes we have Canon Rebel XTi I over here this camera has a very large storage memory it can store about 10000 photos" |
| (3) | **Customer**: "how much is it?"<br>**Robot**: "this is $400" |
| (4) | **Customer**: "and what about the battery life?"<br>**Robot**: "7 hours"<br>*The robot answers more questions about Canon (e.g. color, weight)* |
| (5) | **Customer** walks to Panasonic<br>**Robot** follows the customer to Panasonic |
| (6) | **Customer**: "what is the LCD size?"<br>**Robot**: "a 3 inch touch screen"<br>**Customer**: "that sounds nice. I like it."<br>**Robot**: "also this is very light only weighs 150 grams so you can fit right in your pocket"<br>*The robot answers more questions about Panasonic (e.g. color, optical zoom)* |
| (7) | **Customer**:"Thank you for your help. I will think about it." then leaves the shop<br>*Robot returns to the service counter while saying "no problem"* |

phrases such as "no problem" or "you are welcome" and returned back to the service counter. In cases when the customer left the shop without talking to the shopkeeper (*i.e.* a window-shopping customer), the robot thanked the customer for visiting the shop. Three participants commented that the robot was polite in greeting and saying goodbye.

The robot was usually able to move together with the customer or follow them to a camera, and two participants responded that they liked the fact that the robot followed them to different cameras. Occasionally it misinterpreted a person's motion and moved to the wrong camera, but in such cases it usually corrected itself in the following action. If the customer asked a question about a camera while the robot was in another place, it usually moved to the customer's location while answering the question, in order to reconstruct the target interaction state learned during training.

#### 2.5.5.2 Environment

The use of proxemics models based on the interaction state to control the robot's positioning relative to the customer also seemed to work effectively. Two participants commented that the positioning of the robot was very good, and the robot had a good idea of personal space.

#### 2.5.5.3 Questionnaire

Fig. 2.14 shows questionnaire results from the participants. To compare each rating between the *proposed* condition and the *without-abstraction* condition, we conducted a repeated-measures ANOVA for each of the five questions.

This analysis found significant differences between the conditions for all ratings: "Correctness of wording" ($F(1,16)=9.660$, $p=.007$), "Consistency of robot's speech and motion" ($F(1,16)=26.947$, $p<.001$), "Appropriateness of responses" ($F(1,16)=20.564$, $p<.001$), "Social appropriateness in role" ($F(1,16)=14.222$, $p=.002$), and "Overall evaluation" ($F(1,16)=48.944$, $p<.001$).

These results support our hypothesis that the participant would perceive the overall behavior to be better with our proposed system. The results also support our predictions for the correctness of the wording, consistency in the robot's speech and motion, appropriateness of responses to the customer's actions, and the social appropriateness of the robot in its role as the shopkeeper.

#### 2.5.5.4 Interaction Analysis

The results of the interaction analysis are shown in Fig. 2.15. We conducted a repeated-measures ANOVA comparing *behavior correctness* between the *proposed* and *without-abstraction* conditions. The results showed *behavior correctness* to be significantly higher in the *proposed* condition ($F(1,16)=97.507$, $p<.001$). This result further supports our hypothesis regarding appropriateness of responses to the customer's actions.

As some of the appropriateness judgments are subjective, we confirmed the consistency of the coder's evaluations by asking a second coder to independently rate 10% of

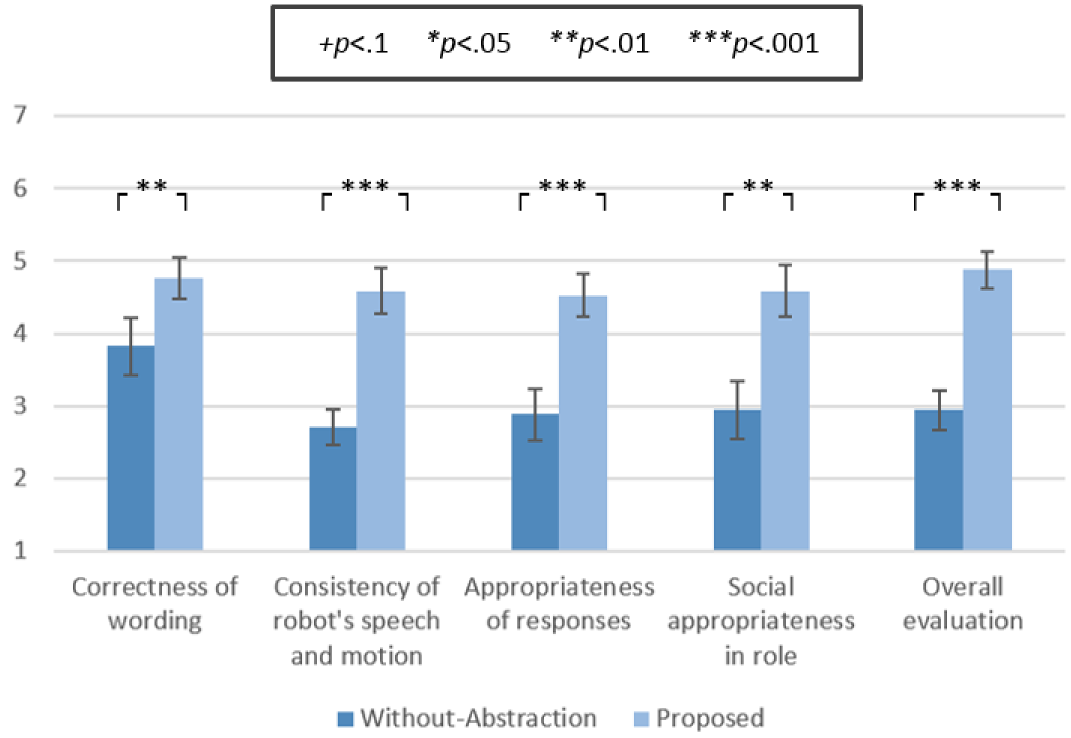


Figure 2.14: Evaluation results of robot behaviors between conditions.

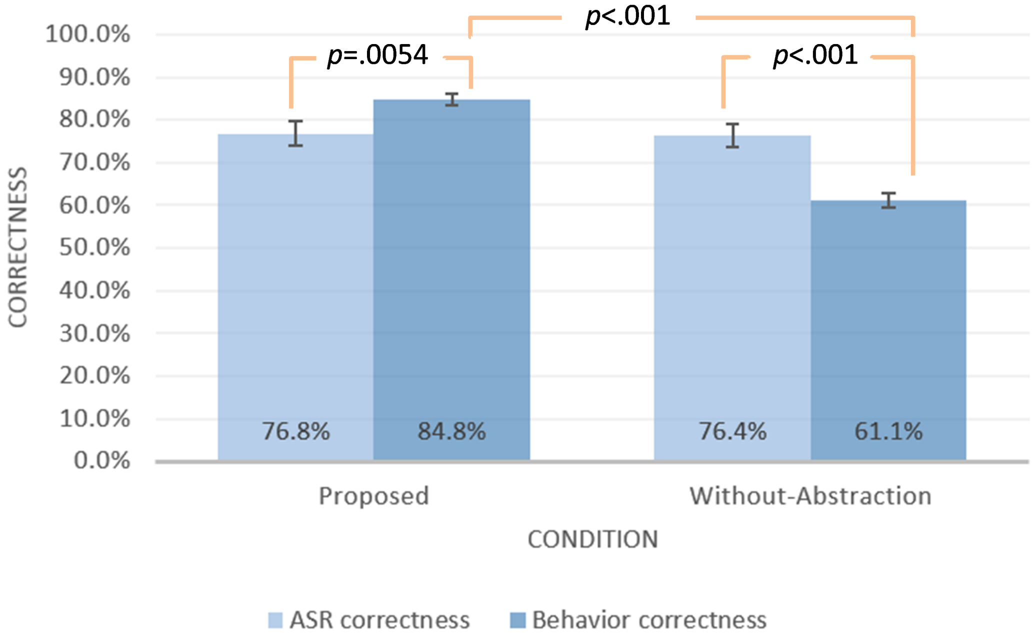


Figure 2.15: Comparison of ASR correctness and robot behavior correctness.

the same interactions. Their results were compared, and a Cohen's Kappa value of 0.76 was calculated, indicating good interrater reliability, so we consider the coder's ratings to have consistency. Next, we compared *behavior correctness* and *ASR correctness* for each condition with a repeated-measures ANOVA. In the *proposed* condition, the *behavior correctness* was significantly higher than *ASR correctness* ($F(1,16)=10.669$, $p=.0054$). In the *without-abstraction* condition, the *behavior correctness* was significantly lower than *ASR correctness* ($F(1,16)=30.356$, $p<.001$). Incidentally, no significant difference was found in ASR correctness between conditions ($F(1,16)=.035$, $p=.854$).

These results confirm our hypothesis that behavior generation in the *proposed* condition is more robust to recognition errors than in the *without-abstraction* condition. We consider this to be an important result, as recognition errors and sensor noise constitute some of the major challenges to data-driven interaction design.

#### 2.5.5.5 Qualitative Analysis

To better understand the nature of our system's performance, we investigated the specific causes of behavior incorrectness. Thus, of the total 1281 robot behaviors observed in the *proposed* system, we analyzed the 201 robot behaviors that were judged as incorrect by the coder. In Table 2.5, we present a qualitative analysis of the errors observed in our *proposed* system, including the possible causes for socially-inappropriate robot behaviors, examples of these errors, and their frequency of occurrence. The results are derived from open-coding and observation from video data and participant feedback in the evaluation experiment. The possible causes are:

**Lack of repeatability**: Some customer behaviors in the human-human interaction were either only observed once or not observed at all in the training data, thus it was difficult for the robot to learn to behave well. Questions such as comparison between two cameras did not often occur in our training data. For this reason, we could not collect enough examples to train the robot well to answer such questions. The robot sometimes answered these questions correctly, but it was usually a pleasant surprise when it did.

We believe that the performance of the system will improve if more data can be collected, and would help the robot answer questions such as comparison between two cameras.

**Error in ASR**: Certain ASR errors would trigger the robot to behave inappropriately, e.g. when an entire sentence gets misrecognized (i.e. "it's expensive" as "sixpence") or when a word about the camera feature gets misrecognized (i.e. "yes how many colors does this camera come in" as "how many calories does this camera come in").

Since our system was trained with real ASR data, it was usually robust to ASR errors. However, some ASR errors were more frequent than others. For example, ASR misrecognized "color" as "kara" on several occasions, but only misrecognized "color" as "calories" on one occasion. In this case, when the customer asked about the camera's color, the robot would respond correctly to the misrecognized word "kara", but not to the misrecognized word "calories".

Table 2.5: Common causes for robot behavior incorrectness with the proposed system

| **Causes** | **Examples** | **Freq.** |
|---|---|---|
| Lack of repeatability | A customer compares one camera with another camera (e.g. "so is this one better than the Sony camera?") | 54 |
| Error in ASR | Misrecognized customer utterance (e.g. "how many colors does this camera come in" misrecognized as "how many calories does this camera come in") | 44 |
| Lack of history representation | A customer already indicated wanting to be left alone, yet Robovie sometimes offered to help several times in a row.<br>A customer says "that's great", and Robovie repeats an utterance that had already been said previously | 23 |
| Error in motion target estimation | Robovie mistakenly estimates the customer to be leaving the shop, when the customer is not planning to leave yet | 17 |
| Error in "farewell" behavior | When a customer leaves the shop, the robot returns to service counter without saying farewell (i.e. does not say "thank you for coming") | 10 |
| Ambiguous shopkeeper behavior | Robovie addresses the customer with "yes sir" | 9 |
| Embodiment | A customer asks the robot a question from across the room (in the training data, most customers first said "excuse me" to the human shopkeeper in order to call them over, before asking a product-related question) | 7 |
| Error in timing | A customer says something new before waiting to hear Robovie's response | 7 |
| Unexpected customer behavior | A customer asks about something outside the scenario scope, to which the robot has not learned a response. | 6 |
| Miscommunication | A customer asks for clarification, such as "Can you repeat that please?" or "did you say 10000 photos" | 5 |
| Missing information | A customer asks "Yes how much is this camera over there?" while pointing to or gazing to another camera | 4 |
| Error due to speech clustering | Robovie responds with "yeah mazzy's ball night hours 515" | 1 |
| Other | The robot failed to respond due to hardware or operational errors (e.g.. network failures, customer forgets to press button on smartphone after speaking) | 14 |

**Lack of history representation**: The lack of interaction history modeling sometimes caused Robovie to repeat himself. Sometimes, when a window-shopping customer asked to be left alone, Robovie would respond with "no problem" and continue letting the customer browse, but since the system contained no long-term history, Robovie sometimes offered to help several times in a row. Though such cases were observed quite a few times (i.e. 15 times), participants did not seem to be mind at all. In fact, one participant thought the robot was being a very eager shopkeeper.

In another example illustrating lack of history representation, a customer asked "What about the Canon camera?" just after asking about the color of the Sony camera. Robovie could not answer correctly, since such question is implicitly referring to the previous question. This exchange is quite complicated but only happened once in the evaluation.

**Error in motion target estimation**: Sometimes the system would misrecognize the motion target of the customer. When the robot misrecognized the customer's motion target as the door (i.e. leaving the shop), the robot might say "thank you for coming" even when the customer was not planning to leave the shop yet. Sometimes when the robot misinterpreted the customer's motion, it would move to the wrong camera, but in such cases it usually corrected itself in the following action.

Sometimes it may be difficult for the robot to estimate the customer's motion target. For example, as the customer enters the shop, it is unclear whether the customer is going to Canon or to the service counter. However, regardless of whether the robot is able to correctly estimate the customer's motion target, it would still greet the customer appropriately since it learned the customer's motion origin is more important than its motion target.

**Error in "farewell" behavior**: In most interactions, the robot acknowledged the customer leaving the shop, e.g. by saying, "thank you for coming." However, in a small percentage of cases, the robot said nothing when the customer left. We speculate that such behavior was learnt from a variety of situations where the human shopkeeper did not verbally acknowledge the customer, e.g., the shopkeeper had already said goodbye, but the customer continued to browse around before leaving; the shopkeeper smiled or nodded to the leaving customer instead of a verbal farewell; or the shopkeeper recognized that a window-shopping customer wanted to be left alone and thus did not verbally acknowledge the leaving customer. As a result, sometimes the robot would not acknowledge or say anything to a leaving customer, but would just return to the service counter.

**Ambiguous shopkeeper behavior**: There were few instances that it was ambiguous whether the robot behavior was actually right. For example, some human shopkeepers would use phrases like "yes sir". Since we did not track the gender of the customer, the robot would learn such phrases, despite whether the customer was female or male.

**Embodiment:** An interesting phenomenon is that customers sometimes acted differently towards the robot than they did towards the human shopkeeper. In the training data, when the human shopkeeper was waiting by the service counter, the customer would usually say "excuse me" first to call the shopkeeper over before asking a question. In the evaluation, customers often asked a question to the robot directly, even from across the room (perhaps because they were speaking to it through the smartphone). These combinations of spatial state and utterance were not observed in the human-human interaction data, so the robot sometimes did not always respond in an acceptable way. For example, it often approached the customer, but did not answer the respective question.

**Error in timing**: Turn-taking is a notoriously difficult problem, and sometimes the customer and robot would speak at the same time. Once a customer action (utterance) is detected, the robot may be triggered to take an action. If the customer speaks again without waiting for the robot to respond, the robot sometimes interrupts the customer while he is speaking.

**Unexpected customer behavior**: A customer may ask a question outside the scenario scope, such as a feature that has not been defined for that camera. Since there are no training examples to handle these questions, the classifier would usually choose the most talked-about feature of that camera as the output behavior for the robot. In our scenario, the robot usually responded with the price of the camera if the customer asked about a non-existent feature.

**Miscommunication**: There were some situations where a customer asked the robot to repeat its utterance, and the robot was unable to do so. Most of the time, the robot spoke understandable and correct utterances, but some customers just wanted a confirmation. In some instances, Robovie would synthesize its speech in a very robotic way (i.e. “10000 photos” synthesized as “one zero zero zero zero photos”), and some customers wanted the robot to repeat for clarification, a situation for which no examples existed in the training data.

**Missing information**: Sometimes the customer may stand at one camera and ask about a feature of a different camera (e.g. “what about the price of that camera?”), while gazing or pointing to the referred camera. Since the robot does not know where “that” is, it would often answer with the price of the camera at the customer’s current location. If reliable sensing of gaze direction and pointing gestures were available, it might be possible to address this problem by representing that multimodal information in the feature vector.

**Error due to speech clustering**: Some clusters were too noisy to produce sensible speech. For example, the speech cluster ID 179 contains 3 shopkeeper utterances, which are all very dissimilar from each other and nonsensical. As a result of this bad cluster, the typical utterance chosen was “yeah Mazzy s ball night hours 515”. However, such instances were rare, and we only found one instance where such a cluster was chosen.

**Other**: Sometimes the robot may fail to respond appropriately or not respond at all due to errors in any of these problems: network connectivity between Google Speech Recognition engine and our system, hardware, software bugs, or the participant forgets to press the button on the Android phone to signal the robot that they started or stopped talking.

## 2.6 Discussion

### 2.6.1 Contribution

In this study, we showed a proof of concept that a purely data-driven approach can be used to reproduce social interactive behaviors with a robot based on example human-human interactions. We demonstrated that by collecting interaction data including natural variation in human behaviors and typical recognition errors, the clustering of the participants' motion and speech, enabled the robot to respond in a natural way to such variations. We saw the robot respond appropriately when people with different speech styles or accents interacted with the robot. This could be an advantage of our approach over grammar-based speech systems, which would have difficulty extracting the meaning from speech recognition results containing errors.

By learning from natural human behaviors, the robot learnt lifelike variation in its behaviors. Explicitly programming multiple phrasings of utterances requires time and effort, but our system implicitly learned to use a variety of synonymous phrases, such as "yes it's very good in low light" and "and if you like to shoot in the dark this is really good", which can help keep interactions interesting and lifelike.

Another merit is that our system naturally learned when speech was location-specific or generalizable to different locations. For example, "Show me a camera with good optical zoom" has the same meaning regardless of where it is spoken, whereas "How much does this cost?" is highly dependent upon the current interaction state target, as each camera is a different price. The robot was able to derive probabilistically how to handle these situations correctly.

The robot learned to mimic the interaction styles of the shopkeepers, such as the casual nature of their speech. We noticed one human shopkeeper in our training interactions spoke quite casually (*e.g.* "okay find me if you want") and used slang words (*e.g.* "600 bucks") at times. As a result, the robot learned to mimic that casual speech for some interactions. Likewise, we asked the human shopkeeper to appear busy and only approach the customer when appropriate. As a result, the robot adapted to a more passive interaction behavior, and waited at the service counter when the customer entered the shop. It could be interesting to explore further how the differences in personality, interaction style, and other personal traits can be modeled and captured from data.

### 2.6.2 Validation of the Model

We believe evaluating how appropriate the robot's action was (i.e. *behavior correctness*) was more important than evaluating how accurately the model was able to exactly replicate a specific example from the training data. Nevertheless, as a reference to understanding the nature of the system, we evaluated the accuracy of our predictor with a 10-fold cross-validation, in which the model predicted a robot action vector out of 467 possible actions from the training examples, and the predicted robot vectors were compared with the actual state vectors of the shopkeeper actions from the training data.

The average accuracy was 26.0%.

Even though the predictor indicates a low accuracy, it often predicts socially-appropriate behaviors. One reason for this is that, as a result from clustering, similar shopkeeper's actions can be clustered into different groups even when they have the same meaning or are interchangeable. For example, shopkeeper behaviors at the Panasonic camera saying "5X optical zoom" and "it has 5 times optical zoom" had the same meaning, but they were respectively clustered into cluster ID 253 and cluster ID 183. When a customer asked "how much optical zoom does this have" the predictor would output 253, while a customer asking "can you tell me about the optical zoom?" predicted cluster 183, even though either cluster would be a correct and socially-appropriate response to either question.

### 2.6.3 Assumptions

There are a number of assumptions implicit in our system design. For example, we assumed that this is a one-on-one interaction where each customer action is followed (optionally) by a shopkeeper's action. We also specified some parameters for our scenario (i.e. number of speech clusters, location of products, number of discretized states), which are needed to tune machine-learning techniques. These problems are not unique to our scenario, as thresholds must be chosen for clustering to work in any problem space, and a finite number of states must be specified to discretize continuous sensor data. We have not yet discovered a good mechanism for choosing these parameters in an automated way for our technique.

We used spatial formations to define 'interaction states' for our scenario. We believe the concept of spatial formation is generalizable, and can be applied to other domains as well. The spatial formations we used are common proxemics formations that characterize the relative positioning between different entities, which have also been adapted into existing HRI models.

Currently, we assume that products do not move in our shop, and thus it is not necessary to track the locations of the products as well. However, in a realistic scenario, it may be possible that products may change locations. To extend the system to accommodate for location changes, a possible solution is add an âĂIJenvironment modelâĂİ to translate raw sensor observations to abstracted states, and again in translating abstracted actions to concrete motion behaviors. Thus, if such low-level position data is abstracted away (e.g. Canon = X4016,Y3228), then we only need to update the mapping when product is moved.

### 2.6.4 Generalizability and Scalability

We believe that this data-driven approach is capable of covering a wide domain of tasks. We can expect our technique to work well with domains that share similar characteristics with ours, i.e. where a limited number of typical, repeatable interactions can be anticipated between the service provider and the visitor. For example, a museum guide moves around to different exhibits and answers a visitor's questions about an exhibit; or an information booth clerk answers the visitor's questions about a department store. For other scenarios where interaction is multimodal and speech or spatial data is not sufficient, we may need to adapt the system to include data from different modalities. For example, we can imagine incorporating skeleton tracking data from a Kinect sensor into our system to train an exercise coach robot.

Our current approach was demonstrated to work well for a scenario where robot behaviors can be trained with a limited amount of data. We collected 178 interactions in the first study, which resulted in 2427 utterances, including both the customers and shopkeepers. This was enough to generate appropriate behaviors for the robot shopkeeper (e.g. answering questions correctly, guiding to correct camera, greeting and farewell).

In terms of scalability with our proposed system, we believe that it will be able to scale up to some degree, for instance, when the number of cameras on display increases. In general, the number of training examples needed is directly proportional to the number of shopkeeper actions that are to be learned. To put it in other words, for the model to learn a single shopkeeper behavior, some constant $C$ number of training examples will have to be captured. Thus, for $n$ behaviors are to be trained, the model will need $Cn$ training in order to train the corresponding shopkeeperâĂŹs actions. In our case, $C$ is approximately equal to 5 training examples. We consider that the amount of training data required is dependent on the number of social behaviors that need to be reproduced, the variability of the customer actions, and the reliability of sensing, thus training effort would scale linearly with the number of behaviors to be learned, such as when the number of cameras on display increases.

The one-step lookahead approach we use might be sufficient for scenarios with highly repeatable interactions that focus on simple questions and answers, such as an information-booth robot or a museum guide, but for more involved interactions it will inevitably become necessary to structure interactions in a more complex way. Extending our current system to include interaction history would seem to be an important consideration for future work. Modeling and remembering different attributes of a person may also be important in an interaction, including everything from name, age, and gender (the robot occasionally said "thank you, sir", to female participants) to dynamic variables like emotional and psychological state, attention target, and goals. In some cases it might be sufficient simply to add these states to the joint state vector to improve prediction, but in many cases it will be important to introduce new behavior models, for example, treating the occurrence of a person's name in speech data in a special way, in order to enable more complex interactive behavior.

### 2.6.5 Tradeoff between Variation and Robustness

There is an inherent trade-off between the variation of the shopkeeper responses and the robustness to sensor noise afforded by clustering similar behaviors. That is, choosing a large number of robot action clusters will lead to more variation in its behaviors, but will increase the likelihood of noise corrupting those behaviors. With our data, we found that 166 clusters preserved a fair amount of variation in the shopkeeper utterances, while providing reasonable robustness to noise. For example, multiple clusters with the same general meaning represent different ways the robot can explain the color of Canon (e.g. "well the also comes in grey red and brown so you have a choice of color is this" and "intense grey red and brown colors"). In high-noise situations, it might make sense to reduce the number of clusters in order to make it easier to reject utterances corrupted by noise. In that case some of these variations would be lost, and the robot might only be able to describe the camera's color in one way. Conversely, in a situation where a greater amount of training data was available, we could choose a higher number of clusters, thus capturing even more natural variations of the spoken utterances while still rejecting noise.

It could also be possible to consider sampling more than one typical utterance from a cluster to use for robot speech. This could lead to a greater degree of lifelike variation in the robot's speech, but it would also increase the risk of ASR errors corrupting the spoken utterances.

### 2.6.6 Embodiment of the robot

One question to be considered in this work is how well the translation of experience from human-human to human-robot interaction can be achieved, given that the robot is embodied as a robot, rather than a human. After all, one could argue that learning to be a human is not necessarily the same as learning to be a robot. Regarding this point, we did observe a few cases where the human-robot interaction differed in some qualitative ways from human-human interaction. For example, one participant talked to the robot in keywords rather than sentences, as if it were a search engine. Some people seemed to treat the robot like a machine and never made eye contact with it. Several participants asked the robot to repeat itself when its speech synthesis was hard to understand. These differences resulted in situations that differed slightly from the training data – e.g., the humans never had difficulty pronouncing their speech, so the system never learned how to repeat and clarify statements.

In most cases, even when differences were observed, such as people not making eye contact with the robot, the difference did not cause any communication problems. The only real problem we observed regarding the dialog flow was the robot's failure to repeat its utterances when asked. We believe specific cases like these are due to a few known issues, e.g. low-quality speech synthesis or speech recognition errors. Such problems are limited and can be expected to decrease as the associated technologies improve. A possible way to handle miscommunication such as a clarification request as an extension to our current system could be to encode the customer's clarification request to a special behavior pattern. Without changing other parts of system, this special behavior pattern could trigger the robot to repeat its previous utterance when it detects the customer asks for clarification. While it is important to keep such differences in mind, we believe this work has demonstrated that the use of human-human interactions holds great potential as a source for generating realistic social behaviors in robots.

### 2.6.7 Justification for Comparison System

We believe our baseline system (i.e. nearest-neighbor) was a reasonable choice for comparison, as it is a state-of-the art technique for generating interactive robot behaviors. Similar to the work of Admoni et al. [1] and Young [149], we also used nearest-neighbor learning technique to match new data in real-time with the nearest example in the training data, which is used to select an appropriate robot behavior. This follows the idea that people learn to communicate by mimicking observed behavior in a given situation.

In some situations, the techniques presented in the baseline system (e.g. nearest-neighbor) provided somewhat reasonable performance, though at some times its performance was poor due to sensor noise. If our data had not been susceptible to noise, the behavior generated by the baseline system would have represented exactly what a human shopkeeper had done in a similar situation.

## 2.7 Conclusion

We have presented a fully-autonomous method that enabled a robot to reproduce socially interactive behavior solely from examples of human-human interactions. Both behavior contents and execution logic are derived directly from observed data captured by a sensor network. We believe this is the first work in the field of social robotics to address this difficult problem. As such, our focus was not on any particular element of the system, but rather on demonstrating the effectiveness of our proposed system as a whole. Our evaluation shows that the robot's behavior using our *proposed* system was rated more highly in a variety of measures than a version of the system that did not use clustering or interaction states. Furthermore, the proposed system showed robustness to sensor noise, achieving an 84.8% behavior correctness rate despite a speech recognition accuracy rate of only 76.8%.

This study has provided a proof-of-concept that interaction can be performed in a data-driven way, directly from observations of human-human interactions. This was made possible through a combination of abstractions: the empirical identification of the typical behavior patterns in the training data, combined with a set of generalizable HRI models specifying spatial formations. Although the interaction scenario we used was somewhat simple, we have suggested many directions in which this work could be extended to capture more complex elements of interactions, and we believe many of the techniques for interpreting sensor data, applying HRI proxemics models, and reproducing human behaviors in a robot despite large amounts of sensor noise will be applicable to other scenarios. This study highlights the importance of behavior modeling in HRI to provide structures useful in interpreting collected sensor data and generating robot behaviors.

Perhaps most importantly, the scalability of this approach gives it the potential to

Table 2.6: Word accuracy and utterance accuracy

| | Data Collection | | Evaluation |
|---|---|---|---|
| | Customer<br>119 utterances | Shopkeeper<br>123 utterances | Customer<br>461 utterances |
| **Word Accuracy** | 79.81% | 76.62% | 87.31% |
| **Utterance Accuracy** | 37.82% | 30.89% | 64.43% |

transform the way social behavior design is conducted in HRI. Once passive collection of interaction data becomes practical, even a single sensor network installation could provide enormous amounts of example interaction data over time, an invaluable resource for the collection and modeling of social behavior. We believe that with today's trends towards big-data systems and cloud robotics, techniques like this will become essential methods for generating robot behaviors in the future.

## 2.8 Appendix

To complement our evaluation of ASR correctness, we also evaluated the output quality of the ASR system based on common metrics of word and utterance accuracy. We used measurements of accuracy rather than the error rate, in order to enable easier comparisons with our other metrics, *ASR correctness* and *behavior correctness*. Word Accuracy is defined as

$$Word\ Accuracy = 1 - \frac{S+D+I}{N} \tag{2.8.1}$$

where S is the number of incorrect words substituted, D is the number of words deleted, I is the number of extra words inserted, and N is the number of words in the correct transcript. Utterance Accuracy is defined as

$$Utterance\ Accuracy = 1 - \frac{N_e}{N_T} \tag{2.8.2}$$

where $N_e$ is the number of utterances containing any errors and $N_T$ is the total number of utterances.

The results are shown in Table 2.6. We speculate the reason why customer's utterance accuracy was much higher during evaluation than during data collection is because customer participants spoke much more clearly to the robot than to the human shopkeeper.

# Chapter 3

# Learning Proactive Behaviors

The previous chapter presented several techniques by which a robot can learn motion and speech behaviors from non-annotated human-human interaction data. However, that technique has the limitation that it is inherently reactive in nature. While it can enable a robot to respond to human-initiated inputs, it cannot enable a robot to proactively initiate behavior on its own. In this chapter, I propose an extension to the previous method, which enables the learning of both human-initiated and robot-initiated behavior for a social robot from human-human example interactions.

This was achieved by extending the technique proposed in Chapter 2 in three ways: (1) extending the turn-taking model by introducing a concept of a "customer yield action", (2) incorporating several steps of interaction history as inputs to the behavior predictor, and (3) using a deep neural network classifier featuring an "attention mechanism" that models the relative importance of each step of the interaction history for generating robot behaviors.

I implemented this new technique in a camera shop scenario, similar to that used in the previous study, and conducted two evaluations of the system's effectiveness. First, I present an offline cross-validation analysis based on human-human data, showing the improved performance of the proposed classifier. Second, I demonstrate the system's effectiveness in live human-robot interactions through a user study, in which participants evaluated the robot's behavior with the proposed proactive system to be significantly better and more proactive than behaviors generated using the system presented in the previous chapter.

We then complement this with a user study of real human-robot interaction in a camera shop scenario, and the results show that the robot was indeed perceived by participants as being more proactive and better overall in comparison, with a robot that only reacts to human-initiated action.

## 3.1 Introduction

The vision of humanoid robots providing service through natural conversational interaction, once a dream of science fiction, is now closer than ever to becoming a reality.

With the arrival of commercial humanoid robot platforms like Pepper, social robots have begun to appear in commercial and public spaces. However, the problem of how to develop social interaction logic for conversational robots, including interactive dialog and interactive motion planning, is still a relatively young and unexplored research domain.

Some works in HRI have already demonstrated techniques for learning speech and motion behavior by imitation from human behavior captured from live interactions [73] and online games [8, 93]. These studies applied data-driven techniques to learn application logic through imitation of human behavior, as opposed to using a more traditional approach of manually designing interaction logic. As the availability of machine power for learning and the availability of large data sets increase, we propose that for situations where large amounts of example human-human interaction data is available, such data-driven approaches could produce more reliable interaction logic and require less effort than manual programming.

A typical approach to designing interaction logic for robots is to specify the robot's behavior in terms of responses to human actions or commands [8, 73, 94]. Such approaches result in fundamentally passive systems, in which the robot only responds to explicit commands or actions from the human. However, many real social situations are mixed-initiative, and it is important for a robot not only to react to a person's actions, but to proactively take initiative as well. For example, a good museum guide not only answers questions about an exhibit, but should also ask questions back and provide interesting anecdotes about the exhibit to the visitor. Likewise, in a shopping scenario, a proactive shopkeeper would take the initiative to explain different product features to a customer.

Nevertheless, learning proactive behaviors in a data-driven way without hand-crafted rules or an explicit model of user's intention [97, 117] can be difficult, as rules for generating reactive versus proactive behavior can have different requirements. For example, in a shopping scenario, a reactive response to a customer's question may depend primarily on the customer's question itself, whereas a proactive behavior, in which the shopkeeper decides to take the initiative to do something (e.g. introducing a new product) as a result of the customer yielding his turn, may depend more strongly on interaction history or context. However, such contextual sensitivity is difficult to capture, and the naive injection of context information may introduce unnecessary noise, making the data too sparse and non-repeatable for the robot to learn an appropriate action. The question remains open as to how a robot can simultaneously and effectively learn the rules for generating both user-initiative and self-initiated actions.

In this work, we will address the question of how to learn both reactive and proactive robot behaviors from human interaction data. While the techniques developed in our previous work from Chapter 2 were sufficient to learn robot actions, those robot actions could only be generated in response to a customer action, and thus that system is unable to generate proactive behavior, e.g. in the case that the customer yields his turn and does nothing.

Thus, we propose three extensions to our previous work. First, we introduce a concept of a **"yield action"** enabling the robot to identify opportunities for a proactive action to be generated. Second, since proactive behaviors are often sensitive to the context of the interaction, we propose to incorporate **interaction history** as a training input. Third, we use an **attention mechanism** in our learning system**,** which has the ability to "attend" and learn which parts of the interaction history are important when predicting robot behaviors. In this work we will present this proposed architecture and demonstrate through offline analysis and live interactions with users that the proposed system can effectively reproduce proactive behavior learned from human interaction data.

## 3.2 Related work

### 3.2.1 Learning social behaviors from data

Several data-driven approaches have been applied to learning interactive behaviors for social robots. For example, Young et al. used learning from demonstration to generate real-time interactive paths for an animated characters and robots to match the style of interactive motion behaviors, based on a pattern-matching algorithm [148, 149].

Frameworks focused on crowdsourcing have been developed to enable learning of overall interaction logic from data collected from simulated environments, such as The Robot Management System framework [135] and The Mars Escape online game [8, 18]. Remote users can interact collaboratively either in an online game, or through the web, and the interaction data are logged and used to develop HRI behaviors in a real autonomous robot. Our work complements these approaches by considering crowd-based data collected directly from human-human interaction using sensors in a physical environment, which presents unique challenges regarding resolving noise from sensor data, abstracting natural variations of human behavior, and discretizing actions for a robot to reproduce.

The use of real human interaction data collected from sensors for learning interactive behaviors has been investigated in numerous works. The robot JAMES was developed to serve drinks in a bar setting, in which a number of supervised (i.e. dialog management) and unsupervised learning techniques (i.e. clustering of social states) were applied to learn social interaction [31]. Admoni and Scassellati proposed a model that uses empirical data from annotated human-human interactions to generate nonverbal robot behaviors in a tutoring applications. The model can simultaneously predict the context of a newly observed set of nonverbal behaviors, and generate a set of nonverbal behaviors given a context of communication [1]. Similar to these works, we use data from human-human interaction for learning robot behaviors, but we adopt a completely hands-off approach, with no human annotation needed for abstraction of social states or for robot behavior generation.

### 3.2.2 Proactive robot behaviors

Strategies for generating proactive robot behavior have been investigated in many works. Huang et al. investigated proactive, reactive, and adaptive collaboration strategies between a human and robot manipulator for a handover task (i.e. unloading a dish rack) [51]. Other works focus on recognition of human intention in order to make autonomous decisions on what task to execute [2, 82, 115, 117]. Our work also focuses on generating proactivity, but we do not focus on a specific task, but rather generating proactive behaviors for an entire social interaction.

In the context of domestic assistive robots, Cesta et al. evaluated a "proactive" interaction modality (where the system takes the initiative) and an "on-demand" interaction (in which the user explicitly requests a service) between an elderly user and an assistive robotic agent [14]. These studies have reported positive user evaluation of proactive robot behaviors, where the human was the main beneficiary of the interaction process. Likewise, we expect that the ability of a robot service provider to generate proactive behavior may improve the user's experience.

### 3.2.3 Learning from history

There are some techniques that have been developed for learning robot behaviors from history, such as goal-directed and habitual robot behaviors through a Bayesian dynamic working memory system [140], or incorporating history in learning for mobile robots [80, 83]. Even though we also want to learn from history, we believe our work is a bit closer to the field of language or dialog learning, where dialogue is a major part of the interaction.

In the context of learning from history for dialog in particular, many techniques involving deep neural networks have been developed recently for handling language-related tasks, which are inherently sequential and require some level of history or memory. Recurrent neural networks (RNN) [81] are often used for tasks like language processing, and Long Short-Term Memory (LSTM) [52] recurrent neural network techniques are often used for tasks such as word-by-word machine reading, where the meaning of a sentence can only be understood when interpreted in the context of previously encountered words [17]. A related technique is supplementing a neural network with an attention mechanism, which learns which part of an input sequence is important for predicting a response [3, 49, 130]. While there are algorithms proposed for learning from history, it is unclear how these algorithms can be applied for learning human-robot multimodal interaction, which is an objective in our work that we hope to demonstrate.

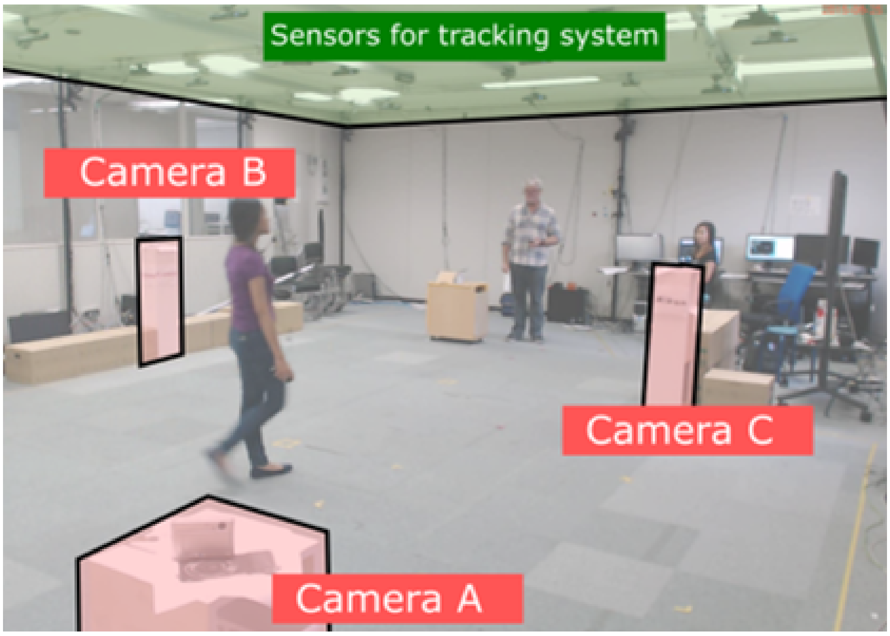


Figure 3.1: Environment setup for our study, featuring three camera displays. Sensors on the ceiling were used for tracking human position, and smartphones carried by the participants were used to capture speech.

## 3.3 Data Collection

### 3.3.1 Scenario

We chose a camera shop scenario for this study as a typical example of the kind of repeatable interaction for which this technique would be most useful. We set up a simulated camera shop environment in our laboratory with three camera models on display, each at a different location (Fig. 3.1), and we asked a participant to role-play a proactive shopkeeper. The shopkeeper interacted with participants role-playing customers, walking with the customers to different cameras in the shop, answering questions about camera features, and proactively introducing new cameras or features when the customers had no specific questions. We recorded the speech and motion data of both the shopkeeper and the customers during these interactions.

### 3.3.2 Sensors

To capture the participants' motion and speech data, we used a human position tracking system to record people's positions in the room, and we used a set of handheld smartphones for speech recognition.

The position tracking system used data from 20 Microsoft Kinect 1 sensors, arranged in rows on the ceiling. Particle filters were used to estimate the position and body orientation of each person in the room based on point cloud data [10].

Speech was captured via a smartphone with a hands-free headset, using the Android speech recognition API to recognize utterances and sending the text to a server via Wi-Fi. Users were required to touch the mobile screen to indicate the beginning and end of their speech. Although it would be ideal to passively collect speech data from microphones in the environment and automatically detect the start and stop of speech activity, reliable technologies to do this are not yet widely available.

Location data for the shopkeeper and the customer were recorded at a rate of 20 Hz. Speech data were recorded at the start and end of each speech event, as signaled by participants tapping on their Android phones.

### 3.3.3 Participants

The customer participants had varied levels of knowledge about cameras and were not selected according to any specific criteria aside from English-speaking ability (due to the use of speech recognition in the study). We employed a total of 9 customer participants (8 male, 1 female, average age 34.1, s.d. 3.9).

To select a participant to play the role of a proactive shopkeeper, we interviewed participants and observed some trial interactions. We asked the customer participants to provide feedback on various shopkeepers in terms of how proactive, helpful, and interested the shopkeeper was. We selected one shopkeeper participant (male, age 54) with a naturally outgoing personality and a great interest in cameras based on our interview with him, as well as the feedback from the customers. He played the shopkeeper in all interactions.

### 3.3.4 Procedure

For this data collection, the shopkeeper was encouraged not only to answer any questions the customer had, but also to take initiative in assisting the customer, either by introducing new camera features or presenting a different camera.

The customer participants were instructed to browse as much or as little as they liked, and told that they could ask questions about cameras or simply listen to the shopkeeper's recommendations.

To create variation in the interactions, customer participants were asked to role-play in different trials as advanced or novice camera users, and to ask questions that would be appropriate for their role. Some camera features were chosen to be more interesting for novice users (color, weight, etc.) and others were more advanced (High-ISO performance, sensor size, etc.), although they were not explicitly labeled as such.

Customer participants were not given a specific target feature or goal for the interaction, as we were mostly interested in capturing the shopkeeper's proactive sales behavior. All participants were instructed to focus their discussion on the features listed on the camera spec sheet, ranging from 8 to 10 features for each camera, to minimize the amount of "off-topic" discussion.

Customer participants performed 24 interactions each (12 as advanced and 12 as novice) for a total of 216 interactions. 17 interactions were removed due to technical failures of the data capture system and one participant who did not follow instructions. The final data set consisted of 199 interactions, including a total of 2568 shopkeeper utterances and 2299 customer utterances.

### 3.3.5 Observed Behavior

Overall, the shopkeeper participant followed our suggestions and acted in a very proactive way. He often spoke in long, descriptive utterances and volunteered extra information when answering questions. In cases where a customer was silent or not asking questions, he frequently provided additional information about a camera or guided the customer to a new camera, so we considered his behavior to be fairly proactive and thus appropriate for this study.

This interaction data differed from that of the previous study in a few ways. First, the shopkeeper's utterances tended to be much longer and more complex, sometimes talking about 2 or 3 topics in one sentence. Second, the shopkeeper often proactively spoke if some silence had elapsed after his last utterance. Third, the customers demonstrated more "backchannel" utterances. For example, a customer might say, "oh, ok," after listening to an explanation, but not ask a follow-up question. In such situations, the shopkeeper in this study often performed proactive behaviors, such as volunteering more information about the current camera or continued his previous explanation.

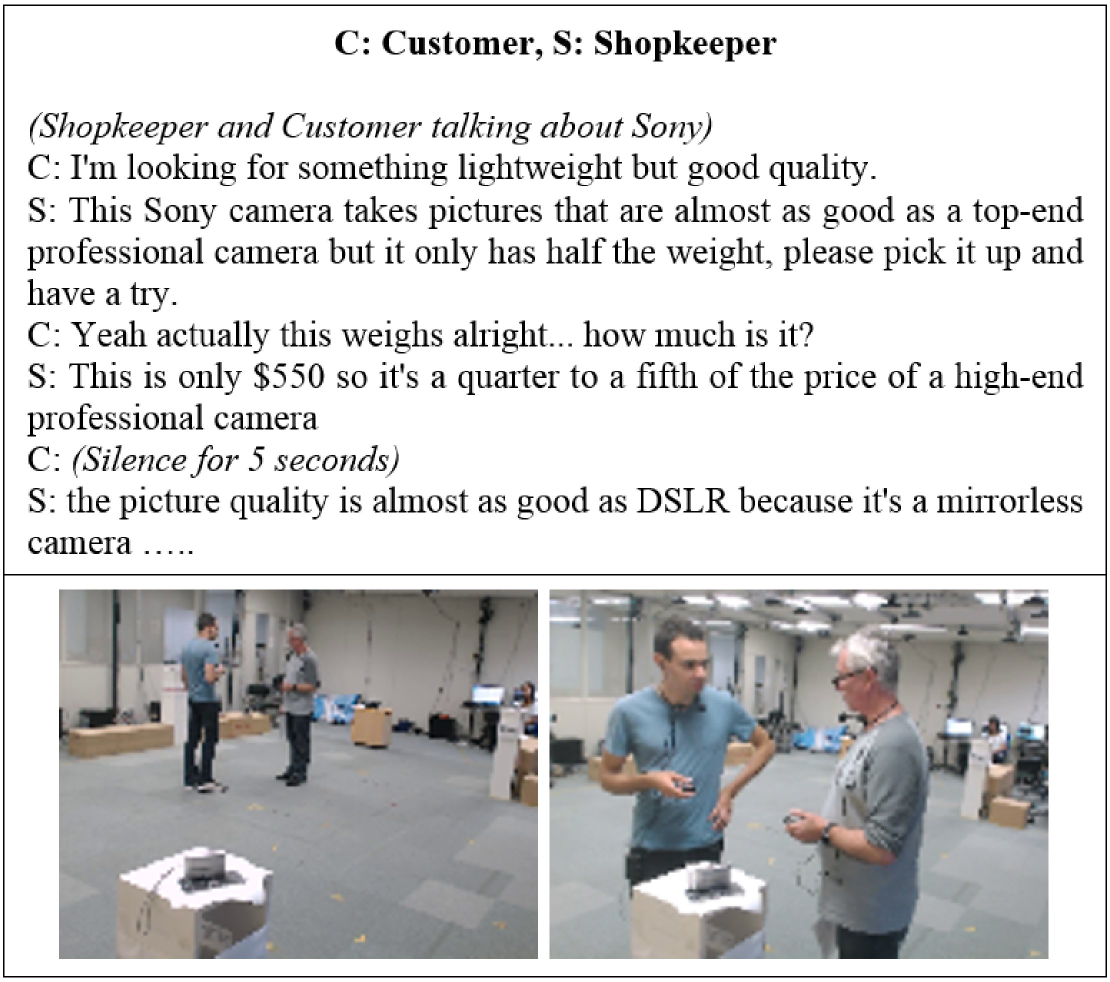
We performed an analysis of the customer utterances to identify whether an utterance required a response (such as a question or a request) or did not require a response (such as a backchannel utterance). We found that 527 (22.8%) of the customer's 2299 utterances did not seek a response from the shopkeeper. There were also 209 instances when the customer was occupied with playing with the camera, reading the spec sheet, or just decided not to do anything, and thus did not speak or move for some time. In these situations, the shopkeeper took the initiative to perform some proactive behavior.

Figure 3.2 illustrates an example interaction from the new data collection. The customer first asks about a lightweight camera, prompting the shopkeeper to show the customer to the Sony camera. The shopkeeper then answers the customer's question about the price. Next, after several seconds of silence, the shopkeeper proactively presents more information about a different feature.

## 3.4 Proposed technique

### 3.4.1 Overview

In order to reproduce both reactive and proactive behaviors for a robot, we used a sequence of techniques that enable behavior contents and interaction logic to be directly learned from noisy sensor data without human intervention. An overview of the techniques is shown in Fig. 3.3, which illustrate how behaviors are learnt from



**C: Customer, S: Shopkeeper**

*(Shopkeeper and Customer talking about Sony)*
C: I'm looking for something lightweight but good quality.
S: This Sony camera takes pictures that are almost as good as a top-end professional camera but it only has half the weight, please pick it up and have a try.
C: Yeah actually this weighs alright... how much is it?
S: This is only $550 so it's a quarter to a fifth of the price of a high-end professional camera
C: *(Silence for 5 seconds)*
S: the picture quality is almost as good as DSLR because it's a mirrorless camera …..

Figure 3.2: An example interaction from the data collection

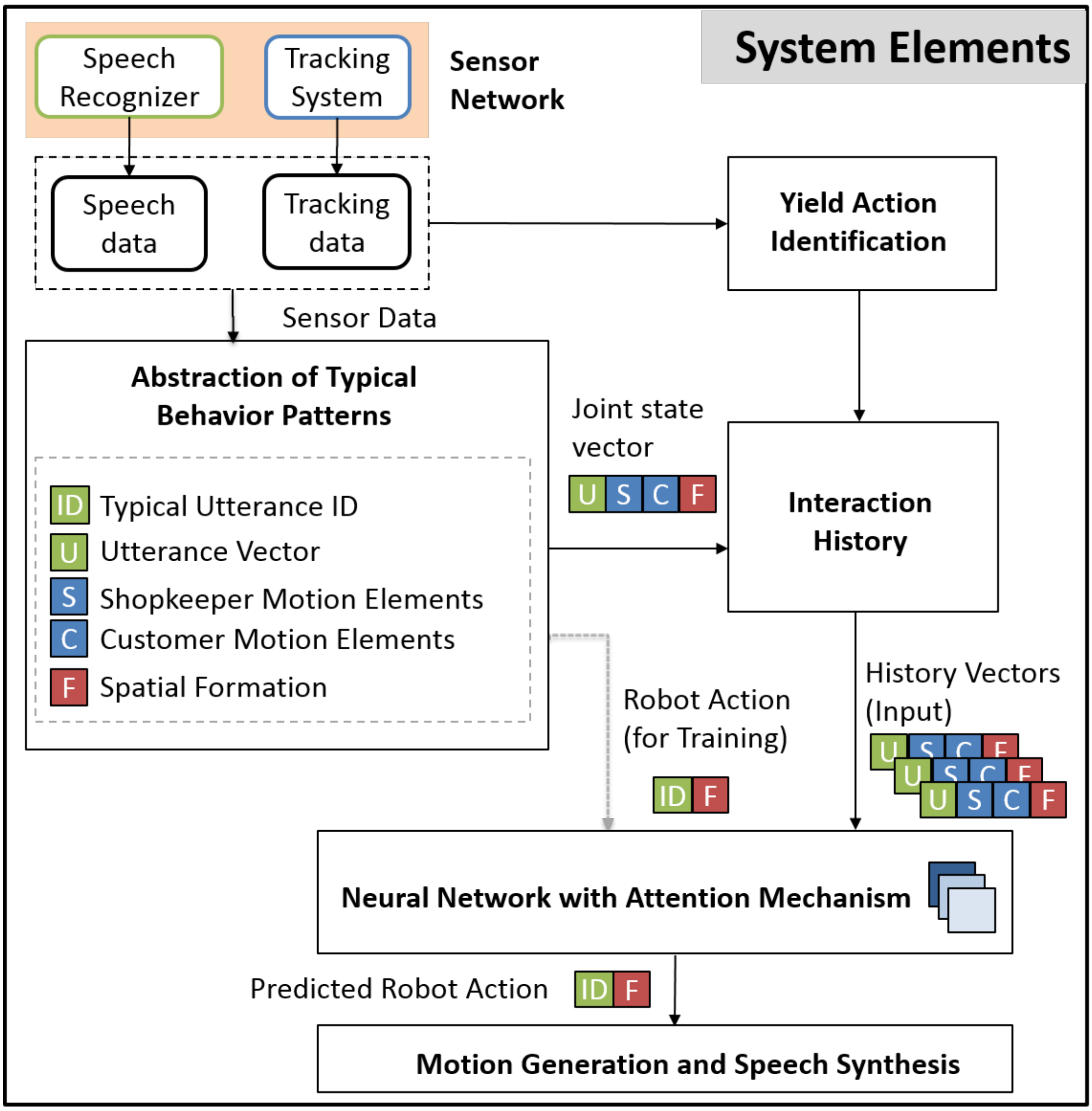


Figure 3.3: Overview of the proposed system elements.

human-human interaction and generated in human-robot interaction. The key steps of the techniques are listed here:

1. **Abstraction of typical behavior patterns** (Sec. 3.4.2): Continuous streams of interaction data captured from sensors are abstracted into typical behavior patterns, and the corresponding *joint state vector* and *robot action* are defined.

2. **Defining yield actions** (Sec. 3.4.3): To enable the robot to generate proactive behavior, we introduce the concept of a *yield action*, which represents the moment when an interactant yields his turn and does nothing, allowing the robot to take initiative.

3. **Incorporating interaction history** (Sec. 3.4.4): We introduce interaction history by concatenating the last $k$ joint state vectors to provide contextual information for generating proactive behavior.

4. **Learning to attend to history** (Sec. 3.4.5)**:** To improve the efficiency of learning, we propose the use of an "attention" mechanism which ascribes weights to the relative importance of various steps of interaction history as inputs to learn appropriate behaviors.

In this work, we used the techniques presented in Chapter 2 for Step 1, while Steps 2-4 constitute the novel contributions of this work which enable proactive behavior generation.

### 3.4.2 Abstraction of typical behavior patterns

In order to learn effectively despite the large variation of natural human behaviors and noisy inputs from the sensor system, the continuous stream of captured sensor data needs to be discretized by time into behavior events, and then abstracted into common behavior patterns. Here we briefly describe our techniques:

1. To find common, typical behavior patterns in the training data, we used unsupervised clustering and abstraction to identify utterance vectors, typical utterances, stopping locations, motion paths, and spatial formations of both participants in the environment.

2. An interaction is discretized into a sequence of actions, which are defined whenever: (1) a participant speaks and/or (2) a participant begins moving to a new location.

3. For each action detected, the abstracted state of both participants at the time is represented as a *joint state vector*, with features consisting of their abstracted motion state, the utterance vector of the current spoken utterance, and their spatial formation.

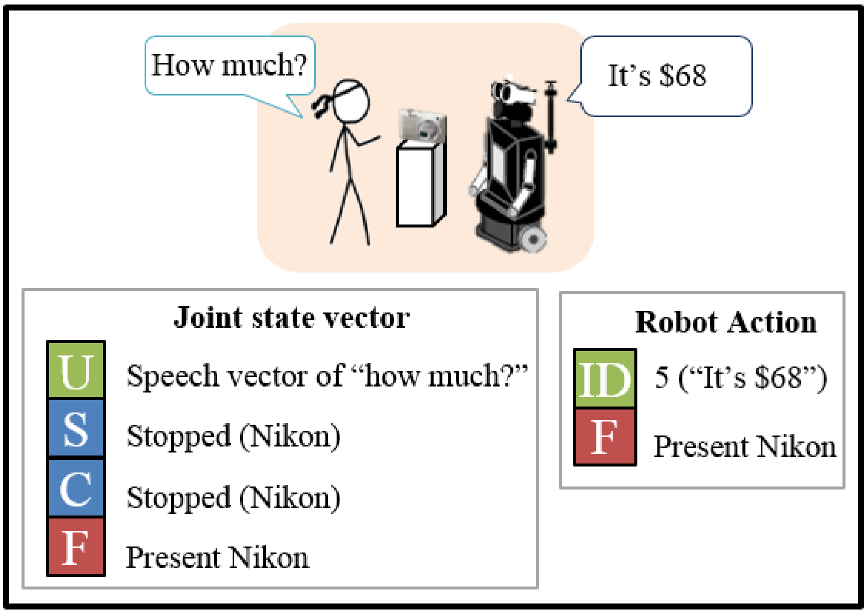


Figure 3.4: Example of abstraction for joint state vector and robot action.

4. For each observed shopkeeper action, we define a corresponding executable robot action, consisting of a typical utterance, represented by an utterance ID, and a target spatial formation. For example, a robot action can consist of utterance ID 5 along with the target formation of *present Nikon*. This triggers the robot to execute the typical utterance "It's $68" associated with utterance ID 5 and execute a motion to attain the formation of *present Nikon*. Fig. 3.4 shows an example of how *joint state vector* and robot action are abstracted from the sensor data.

These data processing and abstraction techniques closely follow the procedure followed in Chapter 2, and additional details are presented in the Appendix.

### 3.4.3 Definition of yield actions

To enable the robot to predict the timing when a proactive action should be generated, we define a *yield action*. A *yield action* represents a moment when the customer is yielding the floor, providing an opportunity for a proactive behavior to be executed. This can be observed in natural human-human interaction, where the interactive partners engage in various phases of turn-taking dynamic such as seizing, holding, and yielding the floor [27, 28]. In our observations from the human-human interaction, we noticed that the customer was sometimes occupied with playing with the camera or reading the spec sheet, or sometimes just decided not to do anything, and thus did not speak or move for some time, indicating that the customer may have relinquished his turn. As observed in 209 instances from our training examples, the shopkeeper often seized the opportunity to do something proactive, either by proactively talking about another feature or introducing a new camera.

Here, we elucidate how *yield actions* can be identified from the time-series data. Using the same deep-rooted turn-taking principles that govern human social behavior, an action is identified whenever: (1) a participant speaks an utterance (end of speech), and/or (2) a participant changes their moving target, and/or (3) an interactant yields his turn of taking an action.

Since we already have full knowledge of the entire sequence of actions for an interaction in the training data, we can assume that the customer has yielded his turn whenever we observe two consecutive occurrences of shopkeeper actions, based on the findings presented by Duncan [27] and our observation that the shopkeeper proactively performed another action after his previous action. For example, after a detection of a shopkeeper speech action (e.g. answering a question), if the subsequent observed action is another shopkeeper speech action (e.g. talking about a camera feature), we can assume that a customer *yield action* has occurred between the two shopkeeper actions. Likewise, this strategy can be applied for the detection of a shopkeeper *yield action*.

Since we do not have the knowledge of future action events in real-time human-robot interaction, we need to detect the customer *yield action* by determining the exact moment when the customer yields his turn. Studies in HRI have analyzed the timing in turn-taking interaction, where the response delay of an user yielding the floor and another user seizing the floor was investigated [15, 134].

Turn-taking is a complicated problem, involving gaze, prosodic, linguistic, and gestural signals as well as timing, but for the current study we make the simplifying assumption that we can detect a *yield action* using a timing threshold. This assumption has been made in other spoken dialogue systems as well [105]. To determine a time threshold for identifying yield actions, we computed the average amount of time elapsed between two consecutively observed shopkeeper actions in the training data. This value was calculated to be 3.52 seconds.

In our system, we thus defined a customer *yield action* to occur if the customer did not begin speaking or moving within 3.52 seconds after the end of the previous robot action.

### 3.4.4 Incorporating interaction history

Although single-step prediction might be sufficient for answering questions, there are many situations where context is important. For example, an answer to a customer's question such as, "how much does this cost," can be generated based on the most recent customer utterance and spatial location – information from interaction history is not necessary. However, some statements or backchannel utterances from a customer, such as "Okay," or "I see", do not contain information which uniquely determines a robot response. In cases like these, as well as in cases where the customer has yielded the turn by silence, an appropriate proactive shopkeeper action will depend to some degree on the previous interaction context.

For example, if a customer yields a turn after the robot presents a camera feature, it might be appropriate for the robot to elaborate in more detail on the feature it just presented, which is directly dependent on the history of the robot's previous utterance. At other times, it might be appropriate for the robot to present a new feature, in which case interaction history is necessary to choose a feature which has not been previously discussed. Sometimes there is an inherent sequence to robot behaviors. For example, the robot might first move to a new camera to introduce it, and subsequently offer for the customer to pick it up and try the camera, and these behaviors cannot be executed in the reverse order. There are other situations where the robot asks for confirmation about something the customer said, and the customer responds by saying "yes," in which case the robot must take an action which is based on the customer's previous, rather than current, utterance.

To address these cases, we propose the use of *interaction history* to provide enough information for the robot to determine an appropriate action for a given context. From our observations of the training data, we found that the shopkeeper's proactive actions are typically dependent on only a few steps of history, such as responding to the customer's previous statement or elaborating on his own previous statement or explanation. There is a tradeoff in which including a longer history increases the system's ability to learn based on history context, but it also increases the dimensionality of the input vector, requiring more training data for stable learning. For the amount of training data available in our study, 3 steps of history seemed to be a good balance, and adequate enough to demonstrate a proof-of-concept to enable the robot to learn certain proactive behaviors, such as presenting a new camera feature.

Thus, we chose to include the three most recent discrete actions as inputs to the classifier. Once an action is detected, a *joint state vector*, describing the state of both interactants at the time, is appended to the *interaction history*, which is kept at a fixed size of 3 steps.

To illustrate the concept of *interaction history* in our system, Fig. 3.5 shows an example of how customer and shopkeeper actions from the training data are segmented into sets of 3 action vectors $(action_{t-3}, action_{t-2}, action_{t-1})$. These action vectors are used as inputs for training the behavior predictor. The subsequent shopkeeper action is represented as a robot action vector, and it is used as the training output for the predictor. In this way, interaction history segments are used to train the robot to predict an appropriate action.

### 3.4.5 Learning to attend to history

While using interaction history may provide valuable context for predicting proactive behavior, the increased complexity and noise of back-and-forth dialog history also introduces irrelevant information, and thus considerably slows the rate of learning task [23]. The inclusion of irrelevant information may thus hinder the robot's ability to learn correct behaviors.

To help the system learn more effectively, we can exploit the fact that some behaviors

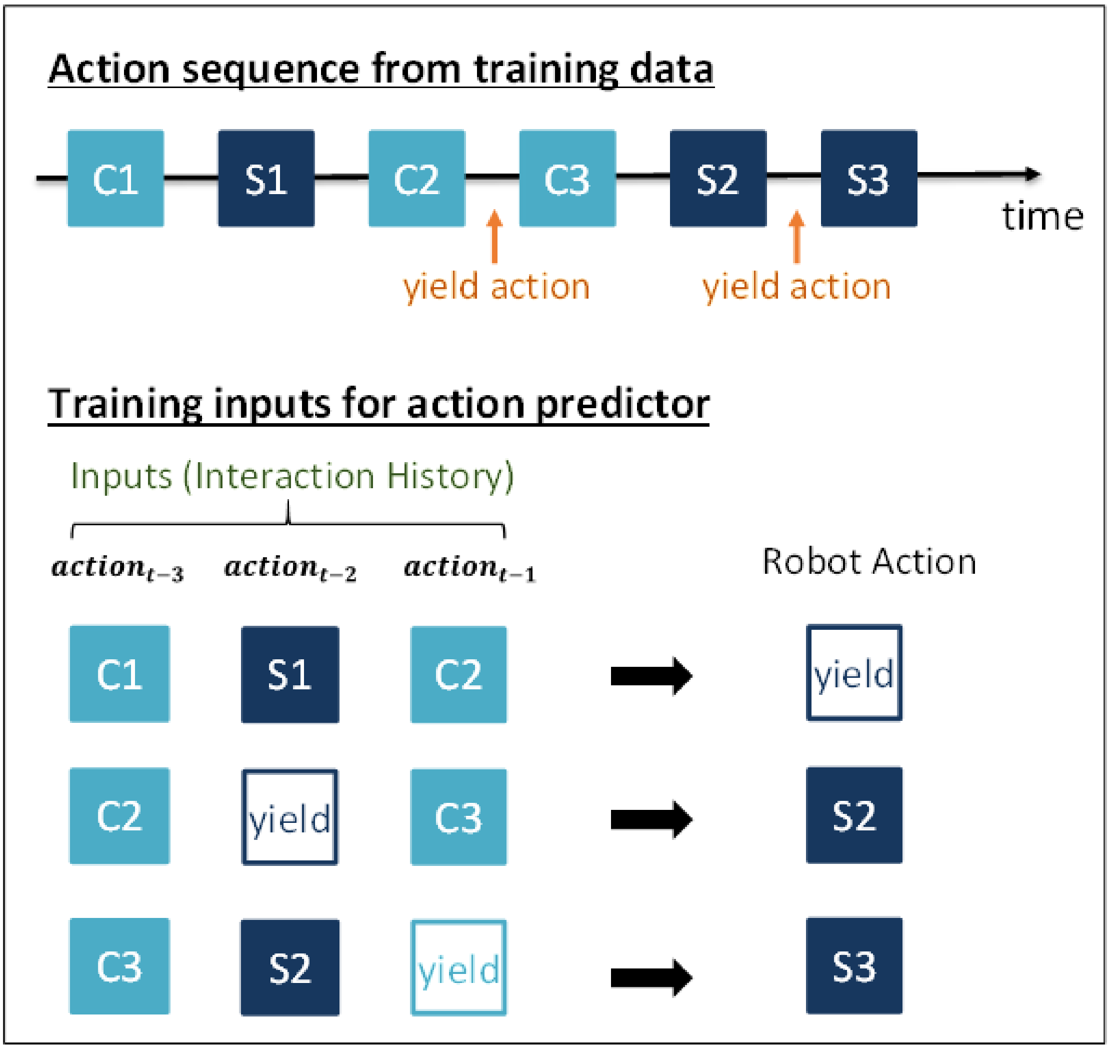


Figure 3.5: Example of how actions are identified in the training data. A yield action is identified whenever two consecutive actions are detected from the same participant without any action detected in between.

are more dependent upon specific steps of history than others. For example, answering a customer's direct question about a camera feature is primarily dependent only on the customer's most recent utterance, that is, $action_{t-1}$. On the other hand, when a customer yields the turn and the robot generates a proactive behavior, the decision is more likely to be dependent upon the robot's own previous action, $action_{t-2}$, and possibly also the customer's previous action, $action_{t-3}$. In the case where the customer says "yes" when the robot asks for confirmation, the decision may depend most heavily on $action_{t-3}$. If the predictor can be trained to focus only on the most relevant steps of history, it may be possible to improve the efficiency of learning.

To achieve this, we applied a recently introduced architecture in the deep learning field, a feed-forward deep neural network with an *attention mechanism* proposed by Raffel and Ellis [104]. For each possible training label, the *attention mechanism* takes each input in the sequence and learns an adaptive weighted average based on each input. This value can be thought as the "relevance" of the inputs, according to the context. Thus, this method has the capability to learn which part of interaction history is relevant for generating a robot action, and also the advantage of visualizing into the neural network to see which part of the history the network is attending to.

Fig. 3.6 shows the schematic of the neural network, where the training input is the interaction history, consisting of an input sequence of the three most recent *joint state vectors*, $X = \left\{jsv_{t-3}, jsv_{t-2}, jsv_{t-1}\right\}$. The hidden layer $H = \{h_{t-3}, h_{t-2}, h_{t-1}\}$ is generated by a forward pass through a regular deep neural network (DNN). The attention mechanism, $a(h_t)$, is computed using a single layer perceptron and then a softmax operation to normalize the values between zero and one, as expressed in Eq. 3.4.1.

$$\begin{aligned} \gamma &= \tanh\left(W_a h_t + b_a\right) \\ a(h_t) &= softmax(\gamma) \end{aligned} \tag{3.4.1}$$

where $W_a$, $b_a$ are parameters optimized by the network using the backpropagation algorithm. Thus, we can model each conditional probability of the robot actions as:

$$p(robot\,action \mid X) = g(Ha) \tag{3.4.2}$$

where $g$ is a nonlinear, multilayered neural network that outputs the probability of a robot action.

The value of $a(h_t)$ is a weight learnt by the DNN, which describes how much of each step in the *interaction history* should be considered for each robot action. So if $a_{t-1}$ is a large number, this would mean that the DNN pays the most attention to the most recent step of the *interaction history*, and thus is important for predicting the robot action.

Fig. 3.7 depicts an example interaction segmented into actions during online operation of the system. Customer and shopkeeper actions are detected when they speak, whereas the customer *yield action* is identified after a specified time has elapsed since the last shopkeeper action. When an action is detected, the interaction history, consisting of a sequence of *joint state vectors,* is sent as a query to the trained DNN, which updates an attention value for each input. The neural network then predicts the probability

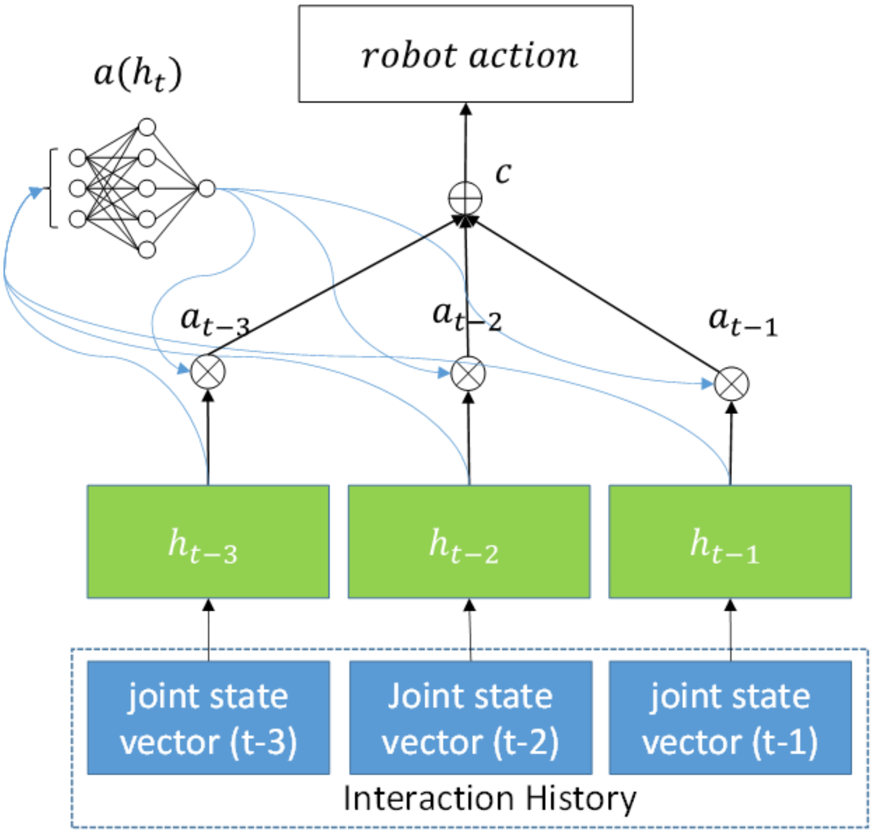


Figure 3.6: Schematic of the deep neural network with an attention mechanism.

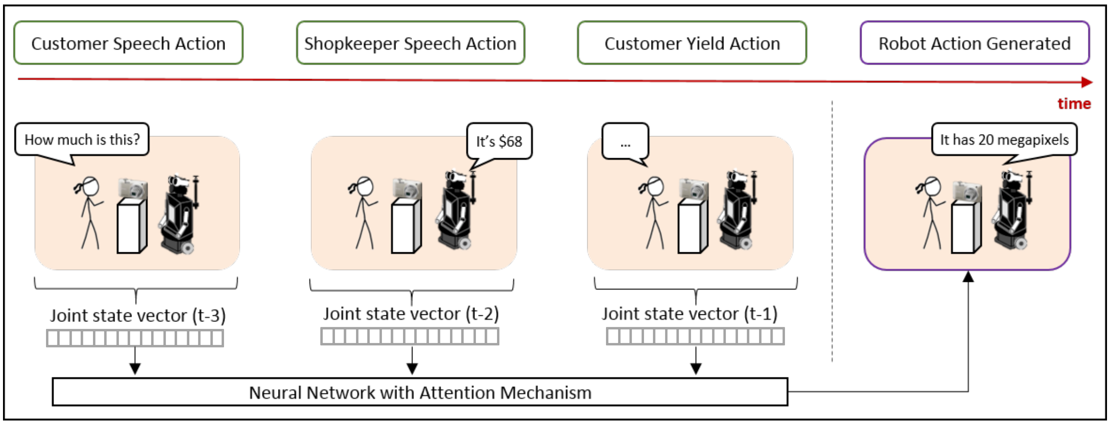


Figure 3.7: Example of how actions are discretized and represented as joint state vectors in the interaction history during online operation of the system. A customer yield action is generated when no action has been detected for 3.52 seconds since the last robot action.

condition for each robot action and selects the robot action with the highest probability condition for generation.

### 3.4.6 Examples of using the Attention Mechanism

Here, we would like to illustrate some examples of our system with the attention mechanism. One feature of the attention mechanism is that the value of $a(h_t)$ provides us with a way to visualize which step of the input sequences the neural network is attending to. The higher the value of $a(h_t)$ for a certain step in the *interaction history*, the more it is considered for predicting a robot action.

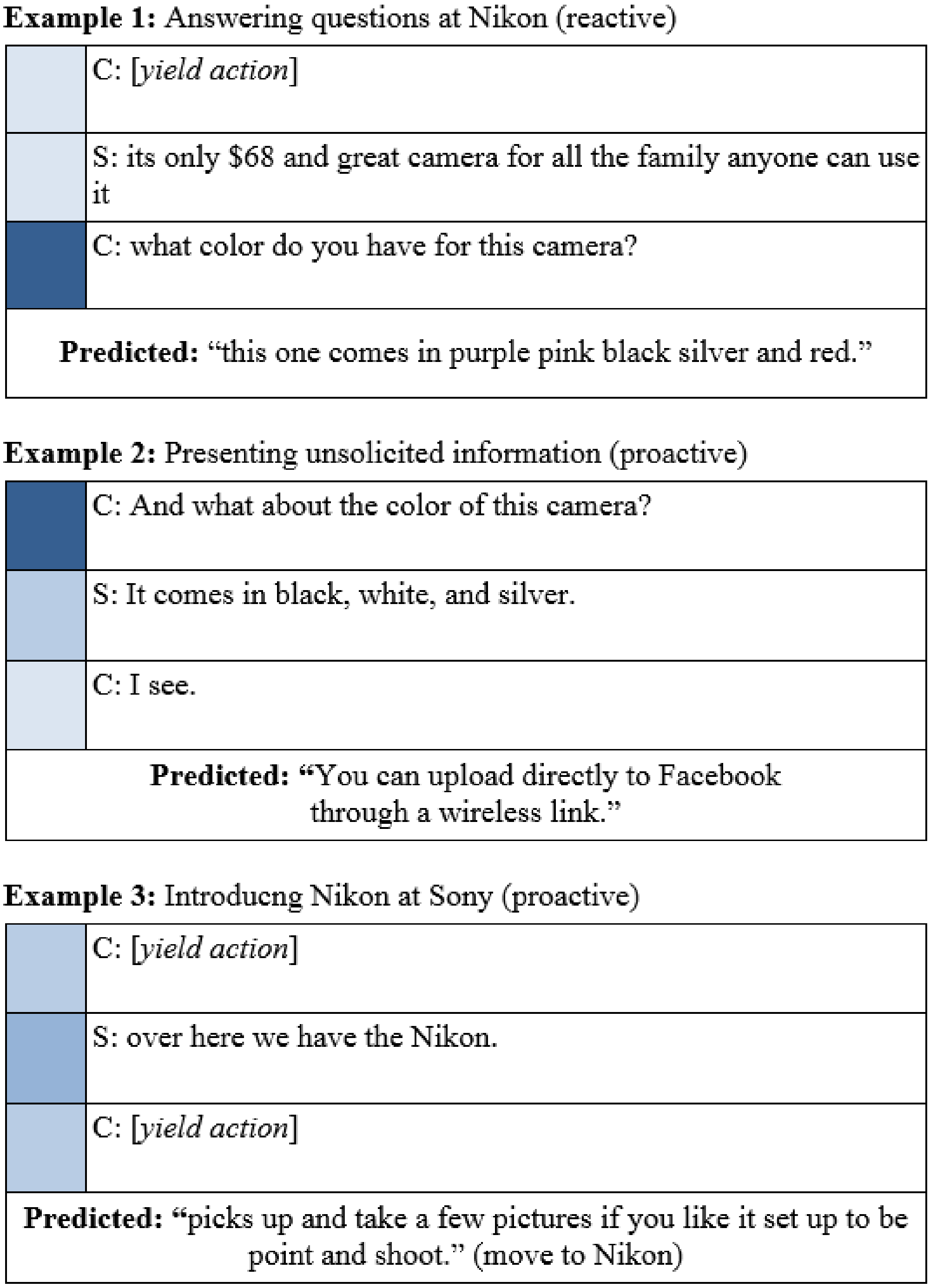
Fig. 3.8 shows some examples of predictions made by our system, in which darker shades of blue represent higher attention weights. For simplicity of presentation, only utterances are shown, but our system uses spatial data as well. These examples were generated by taking a sequence of three actions from the training data (customer – shopkeeper – customer) and feeding them into a trained deep neural network with an attention mechanism to predict an output shopkeeper utterance.

**Example 1** shows a typical exchange in which a customer asks a question about features of a camera. In this case, the predictor has correctly predicted the answer, and the attention model selects the most recent customer utterance, that is, the question, as the most important factor for predicting the robot's answer.

**Example 2** shows a typical example of a situation where the shopkeeper must generate proactive behavior which is not answering a question. In this case, the attention model chooses the customer's *previous* utterance as the most relevant. We hypothesize that this is because the customer's previous question helps to define the set of proactive behaviors which would be appropriate in this context. In this case, the system chooses to present a different feature of the same camera.

**Example 3** shows a typical example of a situation where the shopkeeper must generate proactive behavior which is not answering a question. In this case, the attention model detects a customer *yield action*, and chooses the shopkeeper's previous utterance as the most relevant input. In this case, the system chooses to move to introduce a new camera. We observed that the robot was able to learn the appropriate behavior due to interaction history, which would not have been possible if the robot was only to predict based on the most recent customer action, that is, the customer *yield action*.

These examples show some successful predictions, but we are not claiming that the attention mechanism will work for all situations. Sometimes, the predictor fails using our current approach. These examples were chosen because they illustrate that an attention model such as this could be a useful tool for visualizing a black-box system like a DNN.



**Example 1:** Answering questions at Nikon (reactive)

C: [*yield action*]

S: its only $68 and great camera for all the family anyone can use it

C: what color do you have for this camera?

**Predicted:** “this one comes in purple pink black silver and red.”

**Example 2:** Presenting unsolicited information (proactive)

C: And what about the color of this camera?

S: It comes in black, white, and silver.

C: I see.

**Predicted:** “You can upload directly to Facebook through a wireless link.”

**Example 3:** Introducng Nikon at Sony (proactive)

C: [*yield action*]

S: over here we have the Nikon.

C: [*yield action*]

**Predicted:** “picks up and take a few pictures if you like it set up to be point and shoot.” (move to Nikon)

Figure 3.8: Examples of successful predictions using our attention mechanism technique for a history length of three. Shaded boxes show the relative weight of $a(h_t)$ from the DNN assigned to each action, indicating its importance in predicting the final prediction. Darker shading indicates higher weight.

## 3.5 Offline Evaluation

Before evaluating our system with a live robot, we performed an offline evaluation of the behavior predictor through cross-validation with the training data, in order to confirm the effectiveness of the proposed inclusion of history and attention in the learning mechanism.

### 3.5.1 Evaluation procedure

A cross-validation data set was generated by randomly selecting 500 customer-shopkeeper-customer behavior sequences from the dataset, together with the following shopkeeper behavior which was to be predicted. The remainder of the training data, excluding the selected sequences, was used for training the predictors.

Five predictor variants were evaluated. All evaluations included the proposed detection of *yield action*, and the conditions differed by the type of classifier, the inclusion of history, and the use of the attention model.

1. *NB-1*: A Naïve Bayesian classifier trained on the most recent single customer action. This was the classifier from Chapter 2, so we designated it as the baseline for comparison.
2. *NB-3*: A Naïve Bayesian classifier trained with history (i.e. the most recent three steps of actions: customer–shopkeeper–customer).
3. *DNN-1*: A DNN trained on the single most recent customer action.
4. *DNN-3*: A DNN trained with history (i.e. the most recent three steps of actions: customer–shopkeeper–customer).
5. *DNN-3-AM*: A DNN trained with history, which also incorporated an attention mechanism, as described above.

Normalized initiation, described by [53], was used to initialize the DNN in (3) – (5). The networks were trained to minimize the cross entropy loss for 10000 epochs between the target output and the observed output for the entire training set.

To perform this comparison, we evaluated the "social appropriateness" of the predicted behaviors, rather than simple prediction accuracy, because many synonymous and equally acceptable utterance behaviors exist in the data set. For example, "$2000", "it's only $2000", and "the camera body is only $2000", could all be considered equally valid answers to the question of the price of one of the cameras. This approach is similar to the procedure used in Chapter 2 for evaluating the appropriateness of robot behaviors.

We asked a human coder, naïve to the experimental conditions, to manually rate the acceptability of each prediction as "acceptable" or "unacceptable". Unacceptable behaviors included factually incorrect responses, failures to answer a question, strange behaviors like moving away to a new camera while a person was waiting for a response,

Table 3.1: Results of manually-coded cross-validation comparison. The result of DNN-3-AM showed a signfiant difference when compared with the baseline system

| Classifier | Behavior Correctness | Significance (vs. NB-1) |
|---|---|---|
| **NB-1 (baseline)** | 56.2% | – |
| **NB-3** | 39.0% | $p < .001$ |
| **DNN-1** | 60.2% | N.S. |
| **DNN-3** | 61.8% | N.S. |
| **DNN-3-AM** | 62.4% | $p < .05$ |

and repetition of the previous shopkeeper behavior if not appropriate to do so. To evaluate the appropriateness of the classifiers, the predicted behavior should exhibit similiar traits as that of a proactive human shopkeeper. In our scenario, the human shopkeeper proactively introduce new camera features or a new camera when the customer is silent or says a “backchannel” utterance, which may be dependent on the interaction history. Similarly, the classifer should also predict utterances that are introduce new camera features or a new camera when the customer input is silent or says a “backchannel” utterance.

As the behavior appropriateness ratings require subjective judgment, we confirmed the consistency of the coder’s evaluations by asking a second coder to independently rate the same data set. Their results were compared, and a Cohen’s Kappa value of 0.80 was calculated, indicating very good interrater reliability, so we consider the coder’s ratings to be reliable.

### 3.5.2 Results

To evaluate statistical significance of differences between the conditions, a chi-squared test was performed, comparing each of the classifiers against the NB-1 (baseline) classifier. The results of the cross-validation comparison are shown in Table 3.1.

For the *NB-3* classifier, the chi-squared test showed significance, $\chi^2$ ((1, $N$ =500) = 28.63, $p$<.001) indicating that simply adding history to the Naïve Bayes classifier resulted in significantly worse performance than simple single-step prediction. For the *DNN-1* classifier, a chi-squared test did not show statistical significance, $\chi^2$ ((1, $N$ =500) = 1.46, $p$ = .227). The performance of the *DNN-3* classifier again did not show a significant difference from the baseline in a chi-squared test, $\chi^2$ ((1, $N$ =500) = 2.75, $p$ = .097). The proposed *DNN-3-AM* classifier provided the highest performance, and a chi-squared test showed a significant difference from the baseline, $\chi^2$ ((1, $N$ =500) = 4.45, $p$ = .035).

This evaluation shows that simply adding history as inputs to the original *NB-1* classifier resulted in significantly worse performance, whereas the proposed *DNN-3-AM* technique incorporating both history and the attention model, performed significantly better than the baseline predictor.

## 3.6 User Study

We designed the *without-abstraction* system to be similar to other state-of-the-art data-driven techniques for generating interactive robot behaviors. For example, Admoni et al. [1] developed a system that matches observed data in real-time to the nearest example from human-human training data to select a robot behavior, following the idea that people learn to communicate by mimicking observed behavior in a given situation.

Thus, we created a modified version of our system which also uses the observed sensor data in real-time to find the most similar example from the training data. If our data were not susceptible to noise, the behavior generated by the *without-abstraction* system would have represented exactly what a human shopkeeper had done in a similar situation. The differences between the *proposed* and *without-abstraction* systems are described here and summarized in Table 2.3.

### 3.6.1 Comparison System

We designed the baseline system, NB-1, to be the system used in Chapter 2. This baseline system is the state-of-the-art approach that aims to generate an entire HRI by directly learning from observations of human-human examples in the physical world, which is consistent with the goal we want to achieve in this Chapter.

To observe the effect of the new proposed features in live interaction, we conducted a user-study to compare the two conditions: (a) *proposed*, using customer *yield actions* and the DNN-3-AM classifier, and (b) *baseline*, a system using the NB-1 classifier and not using customer *yield actions*.

### 3.6.2 Hypothesis and Prediction

In the evaluation experiment, we made the following hypotheses about the effects of our proposed techniques:

1. Identifying customer *yield actions* will lead to more proactive robot behaviors in the *proposed* system, since the robot is able to identify the moment when it should take an action.

2. Using DNN-3-AM classifier will enable the robot to generate behaviors that are context-sensitive in the *proposed* system, thus the robot will behave in a more socially-appropriate way.

3. Overall, this will lead to better interaction using our *proposed* system, since proactive behavior is often desirable in a good service interaction.

### 3.6.3 Experiment Setup

#### 3.6.3.1 Participants

A total of 15 paid participants (11 male and 4 female, average age 31.3, s.d. 2.37) played the role of customer in the experiments. All of them were fluent English speakers.

#### 3.6.3.2 Environment

The experiment was conducted in the same camera shop setting used for the data collection, with three digital cameras displayed in an 8m x 11m experiment space. The same sensor network was used for tracking, and the participants communicated with the robot using an Android phone for speech recognition.

#### 3.6.3.3 Robot Platform

For this experiment, we used Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, and a wheeled base. We implemented the proposed techniques in the robot, enabling the robot to autonomously generate behaviors based on inputs from the sensor network and speech recognition results. For its motion behavior, Robovie is capable of moving at a speed of 0.7 m/s. For its motion planning, the dynamic window approach (DWA) was implemented to avoid obstacles [32]. For its utterances, we used the Ximera speech synthesis system [60] to output synthesized utterances.

To make the interaction process more natural, we implemented idling behavior in the robot for both conditions, in which the robot makes small arm and head movements while idling, speaking, and moving [120]. Automatic head-tracking of the robot's interaction partner was also implemented, and the robot followed the customer with its gaze during all interactions.

#### 3.6.3.4 Procedure

We compared the robot's performance between two conditions: *proposed* and *baseline*. For each condition, we asked participants to role-play for 4 trials, to reflect a variety of social situations in a camera shop. To evaluate how well the robot reproduced the behavior of the human shopkeeper in a variety of situations, the participants were asked to role-play as: (1) a need-based customer (2 trials): who was looking for features as either someone familiar or unfamiliar with cameras, and (2) a quiet customer (2 trials): who was not looking for anything in particular and didn't have much to say, and was encouraged to read the spec sheets or play with the cameras. For both customer types, they were encouraged to walk around the shop and show an interest in learning about camera features. The order of the conditions was counterbalanced and the order of the trials within each condition was randomized.

As in our data collection, participants were asked to pretend to be a first-time customer in the camera shop for every trial and the participants performed 2 sample interactions before the experiment to become familiar with the Android phone interface and confirm their understanding of the instructions.

After the 4 trials in one condition were completed, the participant answered a questionnaire. The procedure was repeated with the remaining condition: (*baseline* or *proposed*).

### 3.6.4 Measurement

Before the experiment, we explained to each participant that the goal of this project was to create a proactive robot shopkeeper which could assist customers in a camera shop, and that we would like them to evaluate how well the robot was able to demonstrate that proactivity. After the experiment, we had each participant fill out a written questionnaire, rating the following items on a 1-7 scale (1 being very negative and 7 being very positive for the respective items):

1. How proactive was the robot's behavior?
2. How socially appropriate were the robot's behaviors?
3. Overall evaluation

After the questionnaire was completed, the participants were interviewed to gain a deeper understanding of their opinions of the robot's behavior.

### 3.6.5 Results

#### 3.6.5.1 Questionnaire Results

Fig. 3.9 shows questionnaire results from the participants. To compare each rating between the *proposed* condition and the *baseline* condition, we conducted a repeated-measures ANOVA for each of the three questions.

We verified that all of our predictions were supported, as this analysis found significant differences between the conditions for all ratings: "Proactivity" ($F(1,14)=28.332$, $p<.001$), "Social Appropriateness" ($F(1,14)=5.250$, $p=.038$), and "Overall evaluation" ($F(1,14)=7.875$, $p=.014$).

1. The results support our hypothesis that the participants would perceive the robot to be more proactive using the *proposed* system than the *baseline* system.
2. The results support our hypothesis that participants would perceive the robot to be more socially appropriate with our *proposed* system than the *baseline* system.
3. The results supported our hypothesis that the *proposed* system would lead to a better overall interaction than with a *baseline* system.

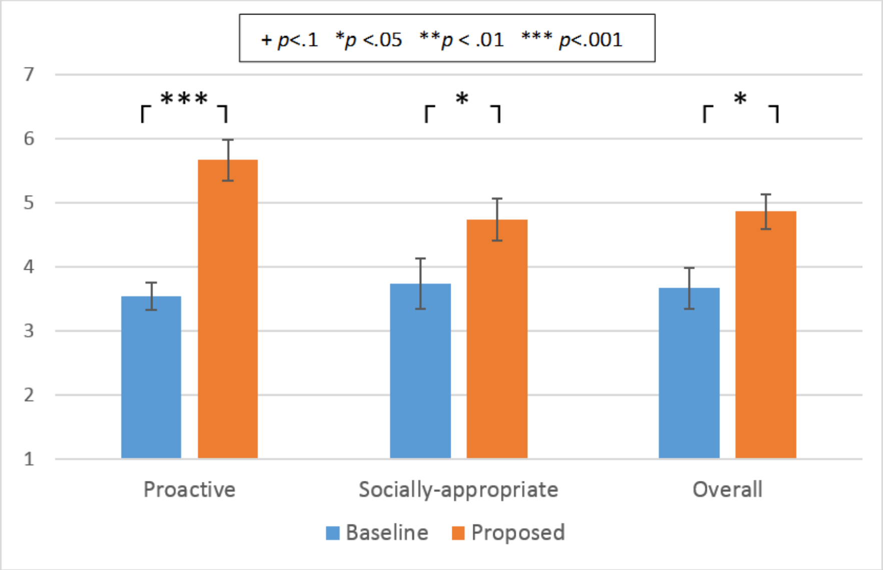


Figure 3.9: Questionnaire results from the user study evaluation.

#### 3.6.5.2 Qualitative Observations

We observed that there was qualitative difference between the behaviors of the *proposed* robot and the *baseline* robot. When the robot was waiting by the service counter and noticed the customer was playing with the camera, the *proposed* robot would approach the customer and start explaining about the features of that particular camera. In contrast, the *baseline* robot would not take any initiative to approach the customer, but rather just stayed at the service counter until the customer asked a question.

When both the customer and the *proposed* robot were at the same camera, the robot would proactively explain camera features to the customer, even before the customer asked anything about the camera. For example, when the customer was looking at the Nikon camera, the *proposed* robot would say: “pick it up see how light it is it is only 120 grams”. If the customer continued to play with the camera and thus did not say anything, the *proposed* robot would continue to introduce a few other features, e.g. the optical zoom or the price.

On occasion, the *proposed* robot would also proactively offer information about a different camera to the customer, without the customer asking about it. For example, while at Sony, the robot would sometimes introduce a different camera to the customer, “over here we have the Nikon”. In contrast, the *baseline* robot would just answer questions, but not take any initiative to talk about camera features or introduce new cameras. Rather, it stood silently by the customer when the customer had nothing to say to the robot.

We also observed that the *proposed* robot was able to generate behaviors appropriate to the interaction context, even when what the customer just said contained little information for what the robot should do next. For example, in one case the *proposed* robot asked a customer who was looking to take travel pictures,“so you need a camera you can take anywhere use easily”. With the customer’s response of “yes yes I need that”, the robot then introduced a small lightweight camera, Nikon, to the customer. Because interaction history was implemented in the *proposed* robot, the robot was able to answer appropriately according to the interaction context. In contrast, the *baseline* robot would not have been able to respond to such customer utterance properly (i.e. “yes yes I need that”), since it only uses the most recent customer’s utterance for prediction.

An example interaction with the *proposed* robot interacting with a quiet customer is shown in Table 3.2. At first, the robot answers the customer’s question about the Canon camera. Then the customer walks to Nikon camera, for which the robot proactively introduces a couple of the camera features (similar to the human shopkeeper, the robot would sometimes introduce 2 or 3 features in a single utterance). The customer then walks to Sony, which triggers the robot to also move to Sony and talk about a feature of the Sony camera. The customer thanks the robot and leaves the shop, at which point the robot also thanks the customer for coming to the shop. As observed, this interaction demonstrates that the *proposed* robot was able to answer questions (i.e. reactive behavior) and proactively explain new features (i.e. proactive behavior).

#### 3.6.5.3 Interview results

From our interview result, many participants thought both *proposed* and *baseline* robots were friendly, and liked it when the robot greeted them when they entered the shop. Many participants commented that they felt more engaged with the *proposed* robot than with the *baseline* robot. The *proposed* robot asked them questions (e.g. “what sort of pictures do you take?”) and talked about camera features, when the customer was just simply playing with the camera. One participant also said that he liked when the *proposed* robot initiated conversation, since he had no idea what he would say to a robot in a shop. Many participants also commented that the *proposed* robot seemed more approachable, more attentive, and more aware.

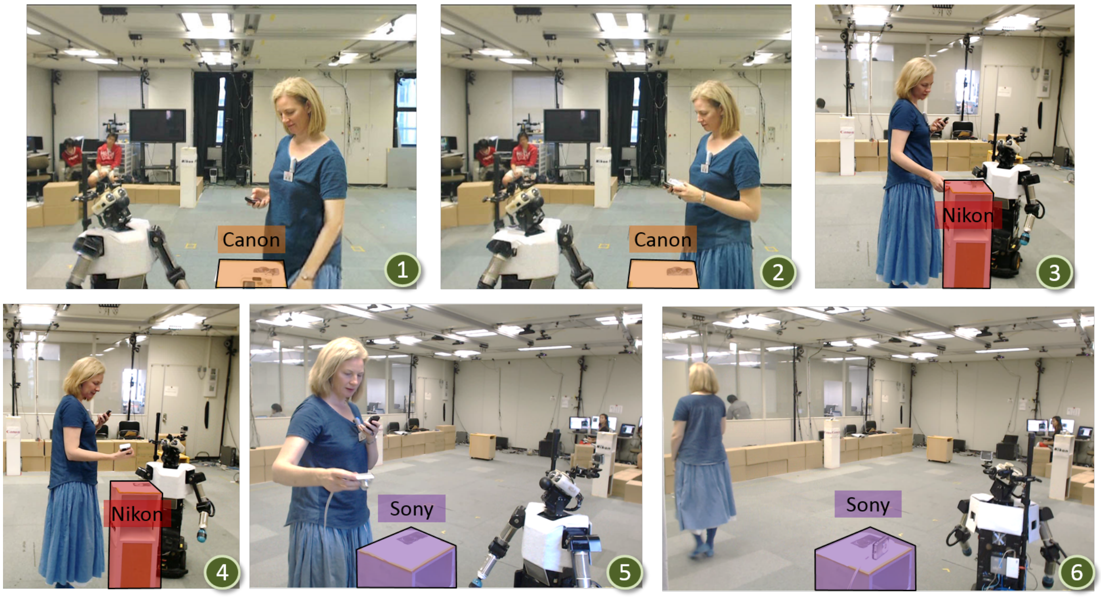
It is interesting to note that some participants preferred the interaction style of the *proposed* robot much more than the *baseline* robot. One participant said the *baseline* robot reminded her of a surveillance system, where the robot is watching to see if she has damaged any goods. Another participant felt annoyed by the *baseline* robot, as it followed him around the shop, but did not say anything to him when he was just looking at the cameras.

## 3.7 Discussion

### 3.7.1 Contribution

In this study, we demonstrated that the robot was able to generate both reactive and proactive behaviors from examples of human-human interaction. We found the robot

Table 3.2: An example of the robot interacting with a customer



| | |
|---|---|
| (1) | *(Customer and Robot at Canon)*<br>**Customer**: “Can you tell me a little bit about this camera?”<br>**Robot**: “You have full creative control it has every possible manual setting” |
| (2) | **Customer** *(picks up the camera and plays with it)*<br>**Robot** “We have one set up here with a zoom lens if you would like to try it" |
| (3) | **Customer**: “Ok thank you” *(walks to Nikon)*<br>**Robot**: *(moves to Nikon)*<br>**Customer**: *(reads the spec sheet)*<br>**Robot**: *(goes to Nikon)* “Would you like to take a couple pictures with at first” |
| (4) | **Customer**: *(plays with the camera)*<br>**Robot**: “Here is the optical zoom you can see the effect it has on your picture”<br>**Customer**: *(continues playing with the camera)*<br>**Robot**: “The two most important things with this kind of cameras that the pictures are great quality because everybody’s going to look at them and I can respond very quickly because it’s all about being out with friends and family”<br>**Customer**: “Yeah that’s true”<br>**Robot**: “It comes in a range of colors it takes fantastic pictures it’s really easy to use so you can focus on the photograph instead of all the camera settings” |
| (5) | **Customer**: “Oh thanks” *(walks to Sony)*<br>**Robot**: *(moves to Sony)* “Good afternoon how can I help”<br>**Customer**: *(plays with the camera)*<br>**Robot**: “It’s an excellent camera that takes the same quality pictures as a top-end camera without the top and price” |
| (6) | **Customer**: “Okay well thanks so much for the information” *leaves the shop*<br>**Robot**: *(returning to service counter)* “No problem have a good afternoon” |

was able to not only answer customer questions, but also proactively assist the customer by introducing new features or a new camera. At times, the robot was also able to respond based on interaction context, even when what the customer just said contained very little information (e.g. “yes please”). Through an offline evaluation and a user-study evaluation, we demonstrated that the robot was perceived as more proactive, more socially-appropriate, and better overall with our proposed techniques, as compared to a baseline system that did not use our techniques.

While the robot was able to generate proactive behaviors, we recognize that there were some behavior elements such as gestures missing, which made the interaction a bit confusing. For example, while at one camera, the robot would sometimes introduce a camera at a different location (e.g. “over here we have the Nikon”). This sometimes confused the customer, since the robot did not point to the camera it was referring to. In the data collection, the human shopkeeper would point to the camera he was referring to, but this abstraction of deictic behaviors was not part of our learning system, thus was not reproducible by the robot. In future work, considering deictic behavior as part of the learning system would be worth exploring as an extension to the current system.

### 3.7.2 Identifying yield actions in turn-taking

In this study, we demonstrated that proactive behavior can be generated by identifying *yield actions* based on a timing threshold. This was validated by a user-study evaluation, where the participants perceived the robot to be more proactive with our proposed technique for identifying *yield actions*.

While identifying *yield actions* based on a timing threshold worked well in our situation, we believe that this technique can be improved by including other ways of identifying *yield actions*. For example, we noticed that a few customers signaled the robot to continue speaking by making eye contact with the robot or nodding to the robot. These nonverbal behaviors as signals for the robot to continue taking initiative are explored in both psychological [28, 42] and HRI studies [86, 106]. Thus, the detection of non-verbal feedback for a more natural turn-taking behavior in a robot could be interesting to explore in future work.

### 3.7.3 History Representation

In our scenario, we demonstrated that the robot was able to reproduce the behaviors of a proactive shopkeeper with a fixed-length of three history steps with our *proposed* system. While the choice of three history steps was enough for our scenario, we wonder whether perhaps increasing the length of history would allow the robot to behave in a more complex way. For example, sometimes the customer would state their goal at the beginning of an interaction, “I am looking for a camera that is easy to carry around”. Since only the immediate history was used for training and generating robot behavior, the robot may forget to talk about features that are related to a lightweight camera after a while.

Although it may be desirable to include the entire interaction history as training inputs, there is a trade-off between learning the appropriate proactive behaviors and learning question-answer behaviors. When only a short history length is used, it is likely the robot may learn question-answer well, but not learn context-dependent proactive behaviors well. In contrast, if the interaction history is too long, there is a possibility the robot may learn some context-dependent behavior well, but fail to answer some questions correctly. Nevertheless, we demonstrated that including just the immediate history reproduced reasonable proactive behaviors for the dataset we have, whereas how absolute history can be represented for a more complex interaction can be explored in future work.

### 3.7.4 Generalizability and Scalability

We believe that this data-driven approach can be applied for domains where repeatable interactions can be captured, and where the proactive behaviors demonstrated by the human follow a formulaic pattern and are context-dependent. Thus, this data-driven approach can also apply to several domains, for instance, a museum tour guide robot. The task of an art museum tour guide robot not only includes answering questions about a particular artwork (e.g. facts about the artist), but also includes proactively explaining about other interesting anecdotes about that particular artwork (e.g. the medium used or time period completed). We can also imagine a tourist center robot, where its tasks could include both answering questions about a tourist attraction (e.g. operating hours) and expatiating about other details (e.g. admission cost).

There may be some domains to which our approach cannot be generalized. These domains might require proactive behaviors that are dependent on subtle social cues or establishing a knowledge base about the user. One example might be an educational robot that proactively teaches a language, where the lesson is tailored to the student's comprehension of that language. We imagine such domain would be difficult to learn with our current approach, since such framework containing the knowledge about a user (i.e. level of comprehension) is not represented in our system.

## 3.8 Conclusion

In this work we have successfully demonstrated a system designed to reproduce not only reactive behaviors for a robot (e.g. answering questions), but also proactive behaviors (e.g. providing unsolicited information) that are learned from human-human interactions. This was accomplished through three proposed techniques, including detection of yield actions, incorporating interaction history, and using an attention mechanism to learn which history steps are important for predicting the robot behavior. First, we demonstrated that our proposed technique was rated the highest in terms of behavior correctness among five different methods for predicting robot behaviors. Then, we validated our approach in a comparison user-study. Our evaluation showed that the proposed system enabled the robot to generate more proactive, more socially-appropriate, and better behaviors overall, as compared with a version of the system that did not use these techniques.

Social robots are now appearing in the real world, and we are seeing a growing market in the service industry for robots which interact with customers. In such situations, proactive behavior may prove to be necessary to enable the robots to proactively engage with their customers and users. In this work we have successfully demonstrated one way in which a data-driven approach from our previous work can be extended to reproduce proactive behaviors from a human shopkeeper, and we believe that data-driven techniques like these will become a valuable tool for building real-world interaction logic for social robots.

# Appendix

Here we describe the data abstraction techniques we used that enable the learning of high-level interaction logic in human-robot interaction to be achieved in an entirely data-driven way, that is, without any kind of manual annotation or cleanup of the sensor data. This follows the work presented in Chapter 2.

## 3.8.1 Action Discretization

We can represent an interaction as a sequence of actions, which are defined when one of the participants speaks and/or begins moving to a new location. Speech actions are defined whenever a participant speaks an utterance (end of speech), and motion actions are defined at the moment when the motion target changes.

## 3.8.2 Defining input features

Here, we describe the features used in the *joint state vector*, including the abstraction of motion (consisting of *current location*, *motion origin*, and *motion target* of both participants, and a *spatial formation*), and an *utterance vector* of the current spoken utterance. The total dimensionality of the input features was 1244.

### 3.8.2.1 Motion Abstraction

The purpose of the motion abstraction step is to characterize a set of stopping locations, motion trajectories, and spatial formations which can be used to describe the motion of the customer or shopkeeper as a combination of discrete state variables rather than raw position or velocity data.

To begin the analysis, we segmented all trajectories in the training data into moving and stopped trajectories, based on a velocity thresholding technique presented in [43]. We spatially clustered these trajectory segments to identify a discrete set of typical **stopping locations** and **motion trajectories** for each role (customer and shopkeeper).

For stopping locations, we used k-means clustering, identifying five stopping locations for the customer (i.e. the locations of the 3 cameras, the middle, and the door) and five for the shopkeeper (i.e. the locations of the 3 cameras, the middle, and the service counter).

For moving trajectories we used k-medoid clustering based on spatiotemporal matching using dynamic time warping.

We created rules for identifying a predetermined set of common **spatial formations** based on the distance between the interactants and their locations. The rules for spatial formations are similar to three existing HRI proxemics models: (1) *present object* [145]: both interactants were at stopping locations corresponding to the same camera, (2) *face-to-face* [44]: both interactants are within 1.5m of each other but not at a camera, and (3) *waiting* [64]: if the shopkeeper was at the service counter while the customer was not.

In addition, we also identified the current spatial target for a particular spatial formation. The *formation target* for "present object" can be either Sony, Nikon, or Canon, whereas the *formation target* for the spatial formation "face-to-face" and "waiting" is "none".

#### 3.8.2.2 Utterance Vectorization

We performed utterance vectorization for the customer and shopkeeper using common text-processing techniques. Specifically, we removed stop words, applied a Porter stemmer, enumerated n-grams up to 3, and performed Latent Semantic Analysis [70] to reduce the dimensionality to 1000. To emphasize important keywords, we also used the AlchemyAPI cloud-based service[1] to automatically extract keywords from each utterance and represented the keywords separately in the vector (200 dimensions). By using this procedure, we were able to take any input utterance and represent it using a 1200 dimensional vector. Vectorization of customer and shopkeeper utterances were performed independently.

### 3.8.3 Defining Robot Actions

In our system, each observed shopkeeper action must correspond to a discrete robot action. A robot action consists of an utterance (represented by an ID number) with a corresponding target formation.

**Shopkeeper Utterance**: In order to reproduce shopkeeper speech with a robot, it is necessary to define a set of discrete utterance actions. Common utterances are frequently repeated in the training data (for example, variants of "How may I help you?" occur 188 times), but these instances often include slight differences due to speech recognition errors or individual variation. Thus, we used bottom-up hierarchical clustering based on lexical cosine similarity to group these repeated and similar utterances into clusters corresponding to discrete robot speech actions.

---

[1] http://www.alchemyapi.com

From each shopkeeper utterance cluster, one utterance was selected for use in behavior generation. We choose the utterance with the highest level of lexical similarity to the most other utterances in the cluster, as this utterance would be the least likely to contain random errors. For each utterance, we compute the cosine similarity of its term frequency vector with every other utterance in the same cluster, and we sum these similarity values. The utterance with the highest similarity sum is chosen as the typical utterance. A total of 761 typical utterances was extracted from the shopkeeper utterance clusters, which can be used to generate robot speech. Notice the typical utterance can also be "none", which means that the robot does not output an utterance.

**Target Formation**: We use the same abstraction rule described in Sec. 3.8.2.1 to represent a target spatial formation for the robot (i.e. *present product*, *face-to*-face, *waiting*, or *none*). This allows the robot to precisely calculate its target position and facing direction defined by the specefic HRI model, in accordance with its estimation of the customer's destination.

For example, if the predicted target formation is different from the robot's current formation, the robot moves to attain the new target formation. Specifically, if the predicted formation is *face-to-face*, the robot approaches the customer; if the predicted formation is *waiting*, it returns to the service counter; if the predicted formation is *present-object*, the robot approaches the target object; and if the predicted formation is *none*, the robot stays where it is.

# Chapter 4

# An HRI model for Deictic Behaviors

This chapter illustrates how a generative behavior model can be developed, such that it can be used as a building block in future data-driven application. In previous chapters, we noticed the robot was able autonomously interact with customers in a mixed-initiative interaction, but ambiguities still rise when the robot tries to introduce a new camera to the customer. This motivates our current study of developing a deictic model, so that it can applied in future data-driven application to enable more natural and humanlike robot behaviors.

This chapter presents a study illustrating the process of developing a model for generating deictic (reference) behaviors in a robot. The behaviors include pointing motions and spoken utterances, and the model is developed based on observations of human interactions. Specifically, this model addresses the behavioral differences between referring to objects and referring to people, and it incorporates social information about the openness of the interaction.

Here I first present an empirical study in which a set of natural deictic behaviors were observed in a variety of social situations. I then propose a model explaining the differences between these behaviors in terms of a balance between understandability and social appropriateness. Calibrating this proposed model based on empirical human behavior, I developed a system able to autonomously select among six deictic behaviors and execute them on a humanoid robot. Finally, I present an evaluation of the system in an experiment in a shopping mall, and the results show that the robot's deictic behavior was perceived by both the listener and the referent as more polite, more natural, and better overall when using the proposed model, as compared with a model considering understandability alone.

## 4.1 Introduction

The importance of natural and humanlike human-robot interaction is gaining more attention as robots gain presence in museums [5, 89, 122], classrooms [57], and elderly care facilities [7, 109]. In order to facilitate natural and intuitive communication, humanlike spoken, locomotive [119], and gestural behaviors are being developed for robots, and one important area of focus is in deictic gestures, such as pointing. Several studies in human-robot interaction have focused on generating human-like multimodal referring acts using both speech and gesture for objects [50, 110, 111, 116, 129] and space [47, 126].

Our study focuses on a method for generating behaviors for a robot to point to a person. There are important differences in the way someone gestures towards objects and the way someone gestures towards a fellow person. When pointing to people, it is often considered more appropriate to gesture casually to them rather than using a very obvious pointing gesture, i.e. with an extended index finger. However, in most situations there would be no reason not to use a clear and precise pointing gesture when identifying an object.

As social human-robot interactions become more complex, it will be important to consider the social appropriateness of a pointing gesture within the context of the conversation. For example, if an elder-care provider is consulting with another practitioner about the health condition of a particular senior person, he would probably discreetly point out that person, using a subtle pointing gesture, in order to reduce the risk of the referent becoming aware and avoid causing anxiety to the referent. In such a scenario, if a robot directly singled out the individual when discussing a sensitive topic (i.e. a "closed" conversation), the robot would probably be perceived as socially-inappropriate. It would be more appropriate for the robot to discreetly identify the referent, even if it meant being less clear to its listener about the referent's identity. However, if the conversation was not of a sensitive nature, and the topic being discussed is neutral or positive (i.e. an "open" conversation), the social consequences would be less severe, and it might be acceptable for the robot to be more obvious about identifying the referent.

Existing models for generating deictic behaviors in robots are typically designed for referring to objects, and thus do not consider this element of social appropriateness. In this study, we present a model for generating socially-appropriate deictic behaviors for pointing to people.

First, we present an empirical study of human pointing behavior, in which we confirm that people usually do not use precise pointing gestures, that is, they typically do not use the index finger to directly point towards another person, and that this phenomenon becomes even more pronounced in the case of private, or "closed," conversation.

We then propose a generative model for deictic behaviors, based on the idea of a balance between understandability and social appropriateness: more precise pointing gestures can increase understandability, but they can also be socially inappropriate. Based on this concept and the data from our human behavior observations, we have developed a model enabling a robot to reproduce human deictic behavior towards people.

Finally, we describe our implementation of this model in a real robot system and present results from an experiment conducted with a robot in a shopping mall, showing that people evaluated the robot's behaviors as more natural and polite when social appropriateness was considered in behavior selection.

## 4.2 Related Work

### 4.2.1 Studies of Human Pointing Behavior

According to Kendon, the intention of precise pointing is to single out an object which is to be attended to as a particular individual object [63]. He categorized this type of pointing as the Index Finger Extended, for which not only the index finger, but almost any extensible body part or held object can be used. The idea that index finger pointing singles out a particular entity is a well-established idea in human science literature, and it provides a useful basis for our categorization.

Some studies have examined the use of reference terms for people. In such studies, the focus was mainly on generating a referring expression (e.g. “This is the coach”) to single out someone as an individual person [48, 98, 139]. Accordingly, we also consider verbal descriptive terms as part of our model for generating deictic behavior.

### 4.2.2 Human-Robot Interaction

Various generative robot behaviors first look at how humans behave as the basis of behavior design. For example, Semwal et al. developed and verified a control system for humanoid bipedal locomotion that was biologically based on human gait cycles [119]. However, the mechanism that drives us to act a certain way may not be obvious to us. Hence, various studies use data-driven methods to extract the underlying mechanisms that govern our behaviors, such as recognizing our emotional states through ECG data [138], or identifying features that uniquely define us through EEG data [68]. In our work, we first observe human deictic behaviors through data collection, and then we incorporate the main factors that were identified in our analysis into our model.

Similar to Kendon’s work of index finger pointing to single out an object, studies have attempted to model the idea of pointing as a way to resolve ambiguity. Bangester et al. focused on the use of full pointing (arm fully extended) and partial pointing (elbow bent) by varying the number of pictures in an array to manipulate the ambiguity of a reference [4]. We will combine this idea of resolving ambiguity with an additional politeness factor that applies when pointing to people.

Some studies in human-robot interaction have focused on generating human-like multimodal referring acts using both speech and gesture for objects [50, 110, 116, 129], and space [47, 126]. Brooks and Breazeal [9] describe a framework for multimodally referring to objects using a combination of deictic gesture, speech, and spatial knowledge. Schultz et al. focused on spatial reference for a robot using perspective taking [118]. In these studies, the robot points to a static object in the environment and produces an appropriate deictic behavior that indicates where the target is. We will also study multimodal behaviors in human-robot interaction, but with a focus on the social aspects of pointing to people.

## 4.3 Data Collection

### 4.3.1 Objective

We collected data from observations of real human deictic behavior so we could generate a model enabling a robot to point naturally to people. Since pointing to objects has been explored extensively in other research, we chose to focus on ways in which pointing behaviors vary when pointing to people. In particular, we were interested in examining three factors:

**Object vs. person:** As discussed in the introduction, we expected that people would point precisely to objects but less precisely to people.

**Open vs. closed:** We expected that people would use less obvious gestures in "closed" conversation, e.g. talking about someone in a negative way, than in "open" conversation.

**Known vs. unknown:** We wondered whether people's behavior would be different if they already knew the referent, such as in the case where saying their name would be enough to identify the referent without ambiguity.

### 4.3.2 Procedure

We conducted the data collection in a shopping mall, as shown in Fig. 4.1 (a), with 17 participants (11 female, 6 male, average 23.7 years old), who were paid. We asked the participants to role-play as customers in the shopping mall. An experimenter asked the participant's opinions about other products or visitors in the mall, and the participant freely answered using deictic behaviors. The participants were not explicitly instructed to use deictic behaviors, but rather instructed to "indicate" who the referent was.

We measured the behavior of the participants under 5 scenarios, chosen to measure the factors described above. The scenarios were defined as follows:

1. **Object**: Referring to a product in the shopping mall that does not belong to either the participant or the confederate (e.g. "Which of these cellphones do you think looks better?").

2. **Open/Known**: Referring to a mutual friend (one of two other acquaintances) in an open conversation. (e.g. "With which of our friends did you take the same bus to the mall?")

3. **Open/Unknown**: Referring to a random, unknown customer in an open conversation (e.g. "Which person did you see at the train station yesterday?")

4. **Closed/Known**: Referring to a mutual friend (one of two other acquaintances) in a closed conversation, such as gossiping negatively. (e.g. "Which of our friends do you think has no fashion sense?")

5. **Closed/Unknown**: Referring to a random, unknown customer in a closed conversation (e.g. "Which person do you think looks unfriendly?")

Each scenario consisted of 6 pre-determined questions, which were counter-balanced. Before the experiment, we had a short ice-breaker session to familiarize the participant with two additional experimenters, who were role-playing as the acquaintances in the "known" scenarios. The two acquaintances stood at different locations for each question. In the "unknown" scenarios, the participants were instructed to refer to actual customers in the shopping mall. Video of each participant's behaviors was recorded, and as we expected that positions of surrounding people might affect the speaker's deictic behavior (i.e., identifying a referent among many customers would be more difficult than when only a few customers were present), we used a human tracking system based on 2D laser range finders (LRF) [34] to capture the positions of the people in the environment. Fig. 4.1 (b) shows the map of the environment in which the data collection was conducted.

The degree of crowding could not be explicitly controlled since the experiment was conducted in a shopping mall. However, all trials were conducted under similar conditions during weekday mornings and afternoons, with an average of 10.46 people present in the environment across all trials.

For each question, the speaker's pointing type and use of a verbal descriptive term were coded and categorized from the recorded videos, as explained below.

### 4.3.3 Categorization of Pointing Types

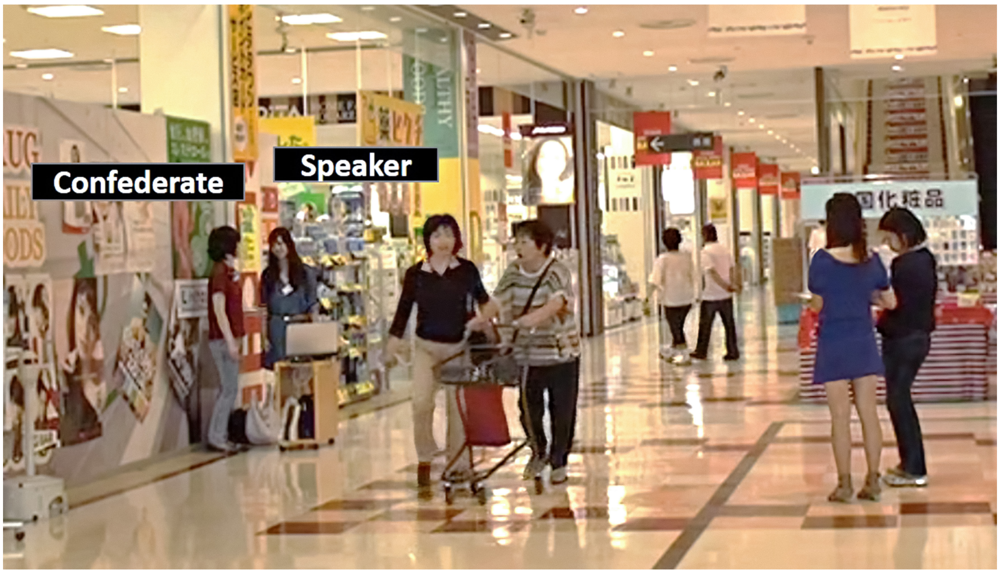
We classified pointing gestures into three categories (see Fig. 4.2): "gaze only", "casual pointing", and "precise pointing". "gaze only" was defined as when the speaker only gazes in the direction of the referent, without the use of any other pointing gestures. "casual pointing" was coded as when the arm was only partially extended. These also corresponded with the "Open Hand Neutral", "Open Hand Prone", and "Open Hand Oblique" pointing gestures as defined by Kendon [63]. "precise pointing" was defined as when the speaker's arm and index finger were fully extended, based on Kendon's definition.

There was a range of variation in the amount of extension of the upper arm and the forearm among participants, so for simplicity, we categorized the pointing type as precise pointing only when the arm and the index finger were fully extended. All other pointing was coded as casual pointing.

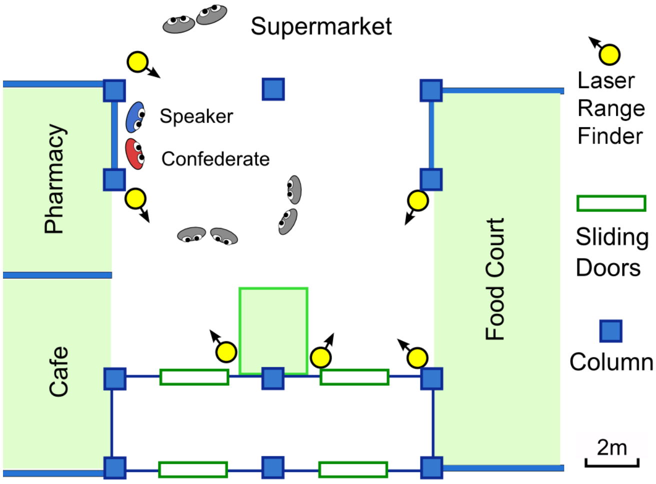
### 4.3.4 Categorization of Descriptive Terms

We analyzed the video to identify whether people used a verbal descriptive term. Here, a "descriptive term" is defined as an utterance aside from the referent's name that uniquely singles out the referent from other people, e.g. based on relative location ("the person in front of the coffee shop") or a visible feature ("the person in the blue shirt").

If only the referent's name was used, it was classified as "name only". If the participant used only a general deictic reference term ("that person"), it was classified as "no descriptive term", since terms like "this" or "that" may not uniquely single out the referent among surrounding people [7].



(a) A photo of the data collection environment.



(b) Map of the data collection environment.

Figure 4.1: The shopping mall environment where the data collection was performed.

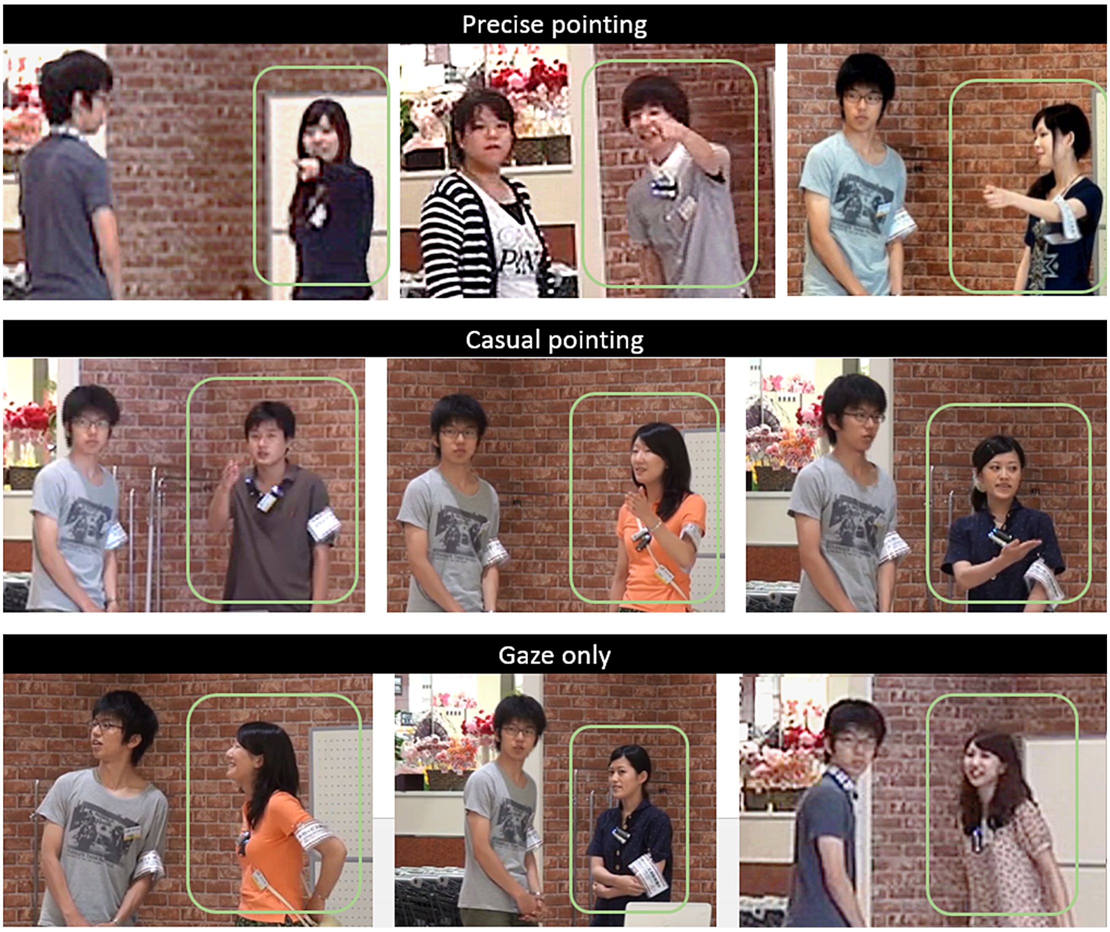


Figure 4.2: Categorization of different pointing types

Table 4.1: Ratio of behaviors performed from data collection

| | Pointing Style | | | Descriptive Term | | |
|---|---|---|---|---|---|---|
| Scenario | Gaze Only | Casual | Precise | Desc. Term | Name Only | None |
| Open/Known | .206 | **.706** | .088 | .402 | **.461** | .137 |
| Open/Unknown | .265 | **.637** | .098 | **.922** | 0 | .078 |
| Closed/Known | **.814** | .167 | .020 | .245 | **.588** | .167 |
| Closed/Unknown | **.559** | .373 | .069 | **.951** | 0 | .049 |
| Object | .049 | .333 | **.618** | **.980** | 0 | .020 |

### 4.3.5 Results and Analysis

For each of the 5 scenarios, a total of 102 reference behaviors were observed (6 questions for each of the 17 participants). Using the recorded videos, an experimenter annotated the pointing behaviors and whether descriptive terms were used by the participants in each trial. This was used for the tabulation of Table 4.1. The experimenter also noted down the referent's position at the time when the speaker made the reference behavior, as well as how long it took for the speaker to make the reference behavior. We noticed that in addition to the use of deictic pointing behaviors to describe the referent, some speakers also used other techniques of representation, such as using gesture to act out putting on a jacket to describe a referent wearing a jacket. These types of gestures were only observed a few times among the participants, and were not a universal phenomenon. In this paper, we avoid these special cases and focus only on deictic language and referential gestures.

The relative frequencies of behaviors for each scenario are shown in Table 4.1, with the most frequently used behaviors in each scenario highlighted in bold.

**Object vs. person:** Participants rarely used precise pointing when referring to people (precise pointing: $<$10.0% for all cases), compared with referring to objects (precise pointing: 61.8%). This suggests there is a social factor that causes the speaker not to want to point precisely, in which he might risk singling someone out.

**Open vs. closed:** In closed conversations, "gaze only" was most common, whereas in open conversations, "casual pointing" was most common. Our interpretation is that as pointing precision increases, the noticeability of the gesture also increases, hence increasing the likelihood of the referent becoming aware of the conversation. This suggests that in closed conversation, the speaker is more concerned about whether the referent becomes aware of the conversation than in open conversation.

In the closed scenario, we also observed that the speaker would often lean closer to the confederate when trying to identify the referent. This phenomenon was not observed in the open scenario. This was more evident when the referent was nearby in closed conversations. Studies have indicated that the forward body lean conveys a sense of intimacy, attraction, and trust [12, 137]. Due to the sensitive information that was being exchanged in the “closed” conversation, we speculate that the participants exhibited such behaviors due to feeling a greater sense of trust or affiliation with the confederate.

Interestingly, in closed conversation, some speakers would also giggle or nervously laugh when they were describing someone negatively (*e.g.* “I think that person with the shopping cart has no fashion sense at all.”). We did not observe speakers laughing or giggling nervously in the open conversation, suggesting that the speakers had higher level of discomfort when describing the referent in the closed conversation than the open conversation [29].

**Known vs. unknown:** Interestingly, we did not see much difference in the use of gesture depending on whether the referent was known or unknown. However, the speaker used more descriptive terms when the referent was unknown to the listener than when the referent was known (e.g. for the Open/Unknown case, 92.2% used descriptive terms, while for the Open/Known case, only 40.2% used descriptive terms).

In general, we found that the speaker took more time to identify an unknown referent. When the referent was unknown to the confederate, the speaker would often repeat or elaborate on describing the referent. For example, the speaker saying, “the person wearing the blue jacket is the person I saw on the bus today,” would sometimes be followed by the confederate confirming, “you mean that person in blue?” The speaker would then describe the referent in further detail such as, “he is also wearing glasses.” On average, the speaker spent 6.25 seconds describing an unknown referent, and 4.41 seconds describing a known referent. Some speakers still used pointing behavior even when using the referent’s name (e.g. in the Open/Known case, casual pointing with name was used 32.4% of the time), even though the name would be enough to unambiguously identify the referent. Perhaps this was to make it easier for the listener to understand the reference, or to share the speaker’s area of spatial attention.

## 4.4 Generative Model for Robot Behavior

### 4.4.1 Overview

Previous studies have modeled pointing as a way to resolve ambiguity when referring to an object. We thus include **understandability** as the first factor in our model, which we define to encompass both resolution of ambiguity and ease of understanding. For example, a crowded environment where a lot of effort is required to identify a person will lower the ease of understanding for the listener.

We then define an additional factor of **social utility**, which reflects the desire of the speaker to be polite by not singling the referent out (see Fig. 4.3). We believe that social utility is the main reason for the variations in deictic behavior between referring to people and referring to objects.

We propose a model to generate humanlike deictic behaviors in a robot by combining these factors of understandability and social utility into a behavior utility function. There is an inherent trade-off between these two factors. For example, pointing precisely at a particular individual may easily identify that person (high understandability), but the

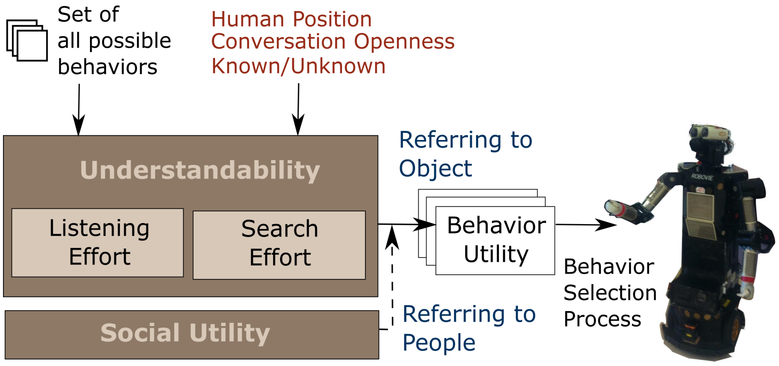


Figure 4.3: Overview of generative model for robot behavior

speaker may have made that person feel singled out and uncomfortable (low social utility).

To select a deictic behavior for a robot, the behavior utility function is evaluated for each of the potential deictic behaviors the robot can perform. We consider six behavior possibilities in our model: one of three pointing behaviors (gaze only, casual pointing, or precise pointing) combined with either the use or the non-use of a descriptive term.

## 4.4.2 Understandability

### 4.4.2.1 Overview

Regarding understandability, we generally assume that with some effort, the listener will eventually identify the target, but pointing makes it easier to search for the referent since the listener can focus their search to a specific region that was pointed to. In this sense, pointing has reduced the listener's time and effort in searching for the referent. The speaker's use of a descriptive term about the referent can also help the listener reduce search effort, since providing cues can help to quickly distinguish the referent among other people or objects. We introduce this concept of "search effort" as one component of understandability. The more search effort is required, the less understandability the listener will have.

Although the use of a descriptive term may help decrease search effort, it also imparts extra cognitive load on the listener to interpret the descriptive term, and hence decreases their ease of understanding. We designate this component of understandability as "listening effort". We modeled the understandability as a function which decreases as the sum of these two effort factors. We assumed perfect understanding if no effort is required.

$$Understandabilty = 1 - (Search\ Effort + Listening\ Effort) \tag{4.4.1}$$

Eq. (4.4.1) does not include explicit weighting factors for these two terms because, as we will explain below, our definitions of Search Effort and Listening Effort implicitly include parameters which can be tuned to adjust their relative weights in contributing to understandability.

#### 4.4.2.2 Search Effort

**Modeling Based on Search Time** We modeled "search effort" based on the concept of a visual search task [143], in which an observer is searching for a target among a variable number of distractors (other people or features in the environment). Longer visual search times roughly equate to higher search effort. Hence, we approximate the search effort as linearly proportional by a factor $w_1$, with visual search time ($t_{search}$), as shown in Eq. (4.4.2). $w_1$ is a parameter which will be tuned.

$$Search\,Effort = w_1 \times t_{search} \tag{4.4.2}$$

The variable number of distractors, or the total amount of distraction $D_T$, is the sum of both the number of human distractors and the environmental distraction. To search for a target among distractions, the listener spends attention and time, $t_{reaction}$, from item to item until the target is found or until all items have been checked [127, 136]. The visual search time for such a task is computed as the average reaction time, $t_{reaction}$, spent on each distraction, times the total amount of distraction ($D_T$), as shown in Eq. (4.4.3). The modeling of $t_{reaction}$ will be explained in the following subsubsections.

$$\mathrm{t_{search}}{=}\mathrm{t_{reaction}} \times D_T \tag{4.4.3}$$

**The Effect of Pointing Precision on Distraction** Pointing singles out a spatial area, but not necessarily a single entity in the world. Other studies have modeled pointing as a cone representing the angular resolution of the pointing gesture [66], which is centered along a beam originating from the pointing finger to the intended target, and has the angular width of a given resolution angle on either side of the beam. Previous findings indicate a resolution angle of a precise pointing cone of about 12 to 24 degrees [67]. We approximated the pointing cone's resolution angle $\theta_{pointing\ precision}$ to be 15 degrees to either side for precise pointing and 60 degrees to either side for casual pointing. For gaze only, we used an angle of 90 degrees, based on the human's forward-facing horizontal field of view.

Recall that our visual search time model is based on searching for a target among a number of distractions, $D_T$. Even when there is only one person in the environment, it will still take some time to find the referent, particularly when the speaker points casually to a referent located far away.

The **number of human distractors**, $D_h$, is defined as the number of people who could potentially be the referent and within the resolution angle of the pointing cone, that is, $\theta_{\text{pointing precision}}$.

Since the environmental distraction is not discrete, we expect it to increase linearly with the pointing angular width. We model $D_e$, the **environmental distraction**, as a constant noise factor $\tau$ per unit angular resolution, integrated over the residual angular

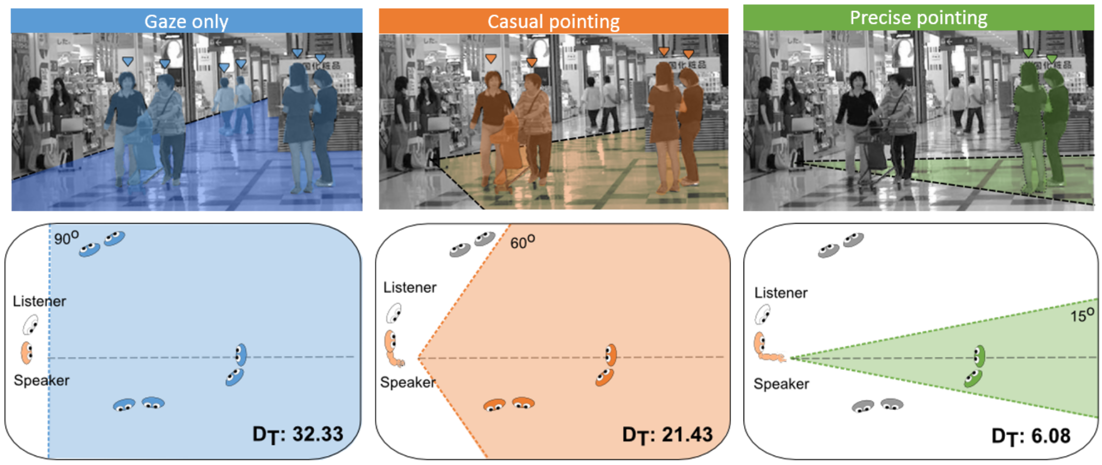


Figure 4.4: An example of $D_T$ for each pointing type in an environment with a total of 6 people. The highlighted people fall within the resolution angle of the pointing cone and are used for calculating $D_T$. Using "gaze only" leads to the highest $D_T$.

resolution of the pointing cone, excluding the angle $\theta_{referent}$ occupied by the referent, as shown in Eq. (4.4.4). The value of τ will be larger for more cluttered environments.

$$D_e = \tau(2 \bullet \theta_{pointing\ precision} - \theta_{referent}) \tag{4.4.4}$$

Recall in the previous section that the total amount of distraction $D_T$ is the sum of both the number of human distractors, $D_h$ and the environmental distraction, $D_e$. Thus, $D_T$ will be directly influenced by the pointing gesture used by the speaker. An example from our data collection, shown in Fig. 4.4, illustrates how $D_T$ is affected by the different sizes of the pointing cones. In this example, 6 people are present in the speaker's forward horizontal field-of-view of 180 degrees in our shopping mall environment. Using gaze only, all 6 people within the speaker's view will be included as human distractors, whereas casual pointing reduces $D_h$ to 4 people, and precise pointing reduces $D_h$ to 2 people. Likewise, $D_e$ is affected by the pointing type according to equation (4.4.4), in this case, 26.33 for gaze only, 17.43 for casual pointing, and 4.08 for precise pointing.

**The Effect of Descriptive Term on Reaction Time** To distinguish the referent from other people, a speaker may use a unique description term in addition to pointing. Previous studies have shown that providing a cue [144] or being familiar with the target [141] can reduce the uncertainty of the target and consequently reduce the reaction time. If the referent is known to the listener, the speaker will use the referent's name to describe him in all cases (e.g. it will be unnatural to describe a mutual friend as "the man in blue shirt" rather than "Jack"). Thus, we model the **reaction time** $\mathrm{t}_{\mathrm{reaction}}$ to be shortest when the referent is known (see Eq. (4.4.5)). When the referent is unknown to the listener, search time will be longer. However, use of a descriptive term will reduce $\mathrm{t}_{\mathrm{reaction}}$ compared with not using a descriptive term.

$$\mathrm{t_{reaction}} = \begin{cases} \mathrm{t_k}, & \text{if known} + \text{ using name} \\ \mathrm{t_{ud}}, & \text{if unknown} + \text{ using descriptive term} \\ \mathrm{t_u}, & \text{if unknown} + \text{ no descriptive term} \end{cases} \tag{4.4.5}$$

#### 4.4.2.3 Listening Effort

The second factor in the *Understandability* equation is listening effort, representing the effort associated with the time required to listen to a descriptive term. For simplicity, we assign one of two discrete values to the **listening effort**: $c_{desc}$ if a descriptive term is used, or $c_{no\ desc}$ otherwise in our model, as shown in Eq. (4.4.6). Since listening to a name or reference term requires less time, therefore less effort, than a descriptive term, we expect $c_{desc} > c_{no\ desc}$.

$$\text{Listening Effort} = \begin{cases} \mathrm{c_{no\ desc}}, & \text{no descriptive term} \\ \mathrm{c_{desc}}, & \text{using descriptive term} \end{cases} \tag{4.4.6}$$

### 4.4.3 Social Utility

We model social utility as a quantity that will decrease if the speaker makes the referent feel uncomfortable or singled out. The loss in social utility is especially high in “closed” cases, when the content of closed conversation is leaked to the referent (e.g. the referent hears bad comments about him). To quantify this phenomenon, we consider the risk of the referent becoming aware of the conversation ($R_{awareness}$), multiplied by the cost to social utility ($C_{social}$) if the referent becomes aware, as shown in Eq. (4.4.7).

$$\text{Social Utility} = -(\mathrm{R_{awareness}} \times \mathrm{C_{social}}) \tag{4.4.7}$$

Recall that in our previous section we model precise pointing to have the effect of ruling out distraction. The presence of many distractors within the pointing cone, e.g. due to a less precise pointing gesture, makes it less clear whether the speaker is actually pointing to the referent, whereas a precise gesture with few distractors leaves little room for doubt. Thus we approximate the **awareness risk** ($\mathrm{R_{awareness}}$) as the inverse of the total amount of distraction:

$$\mathrm{R_{awareness}} = \frac{1}{(\mathrm{D_h} + \mathrm{D_e})} \tag{4.4.8}$$

The **cost to social utility** is dependent upon the openness of the conversation. As explained above, the penalty to social utility due to the referent becoming aware of the conversation is much more severe in closed conversation than in open conversation. Thus, we model the cost to have one of two discrete values, based on the openness of the conversation, where $\beta_{\mathrm{closed}} > \beta_{\mathrm{open}}$.

$$\mathrm{C_{social}} = \begin{cases} \beta_{closed}, & \text{if conversation is closed} \\ \beta_{open}, & \text{if conversation is open} \end{cases} \tag{4.4.9}$$

Table 4.2: Ratio of predicted behaviors from data collection using calibrated parameters

| | Pointing Style | | | Descriptive Term | | |
|---|---|---|---|---|---|---|
| Scenario | Gaze Only | Casual | Precise | Desc. Term | Name Only | None |
| Open/Known | .196 | **.804** | 0 | 0 | **1** | 0 |
| Open/Unknown | 0 | **.804** | .196 | **.99** | .001 | 0 |
| Closed/Known | **1** | 0 | 0 | .001 | **.99** | 0 |
| Closed/Unknown | **1** | 0 | 0 | **1** | 0 | 0 |
| Object | 0 | 0 | **1** | **.833** | 0 | .167 |

Table 4.3: Calibrated model parameters

| Search Effort | | | | | Social Utility | | | Listening Effort | |
|---|---|---|---|---|---|---|---|---|---|
| $\omega_1$ | $t_k$ | $t_{ud}$ | $t_u$ | $\tau$ | $\beta_{open}$ | $\beta_{closed}$ | $w_{ref}$ | $c_{desc}$ | $c_{nodesc}$ |
| .013 | .03 | .07 | .3 | 8.5 | .273 | 30 | 25[cm] | .011 | 0 |

### 4.4.4 Calibration of Our Model

We manually calibrated our model based on the results of our data collection by adjusting parameters for our model until the correspondence between the most frequently predicted behaviors for each scenario (highlighted in bold in Table 4.2) and the most frequently used behaviors in that scenario from the data collection (highlighted in bold in Table 4.1) were maximized. Table 4.3 shows the calibrated parameters.

### 4.4.5 Examples of Using Our Model

The examples in Figures 4.4 and 4.5 illustrate situations where our model chooses different behaviors based on the amount of distraction and the scenario. The figure shows each person's position in the environment. The resolution angles for each of the three pointing cones (90° for gaze only, 60° for casual pointing, and 15° for precise pointing) are drawn as different shades of red dashed lines radiating out from the speaker.

Figure 4.5 shows examples in the Open/Unknown scenario. The most common behavior in this scenario is casual pointing. However, precise pointing is sometimes used in crowded environments, where it is harder to identify the referent. This is due to the distraction effect, as modeled previously.

Fig. 4.5 (a) is a case where the participant used precise pointing to identify the referent. In this crowded environment, there were 8 people within the region of casual pointing; thus, casual pointing would yield low understandability. However, precise pointing reduces the number of human distractors to 2, providing much higher understandability. Fig. 4.5 (b) illustrates a less crowded example. Here, due to the

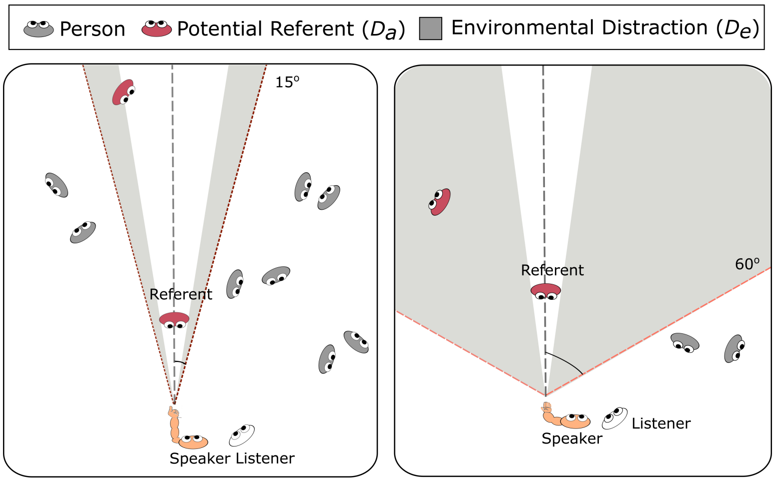


Figure 4.5: Open/Unknown scenario: examples showing the influence of distractors on behavior selection. (a) Precise pointing is chosen. (b) Casual Pointing is chosen.

smaller number of distractors, the model chooses casual pointing, which yields enough understandability while yielding higher social utility.

Fig. 4.6 shows two examples in the Open/Known scenario. As in the unknown scenario, the most common gesture is casual pointing. However, since the referent is already known to the listener, less ambiguity needs to be resolved. Fig. 4.6 (a) shows a crowded environment, but here casual pointing is enough to yield enough understandability. When the environment becomes less crowded, as in Fig. 4.6 (b), using gaze only would be enough for understandability, while yielding high social utility.

### 4.4.6 Model validation

The goal of our model is to generate a reasonable policy for producing socially-appropriate behaviors, rather than exactly replicating individual people's deictic behaviors. It is often difficult for a system to replicate exactly what humans do due to natural variation or randomness that arises among individuals. For instance, in the "Open/Unknown" scenario, there were 5 trials where 5 human distractors were tracked in the environment. Of the 5 trials, 1 participant used "gaze only", 3 participants used "casual pointing", and 1 participant used "precise pointing". This suggests that some deictic behaviors may be used interchangeably in some situations or dependent on the personality or culture of individuals. For this reason, we aimed to generate robot behaviors based on the dominant behavior trends observed from the data collection.

Table 4.4 shows the confusion matrix of the predicted behavior using our model, based on the observed behavior from our data collection. The overall prediction accuracy was 81.3% for the "Closed/Known" scenario, 55.8% for the "Closed/Unknown" scenario, 42.1% for the "Open/Known" scenario, and 52.0% for the "Open/Unknown" scenario.

As a result of our calibrated parameters, our model tends to perform on the side of caution (*i.e.* the robot chooses deictic gestures that are less socially awkward). In both

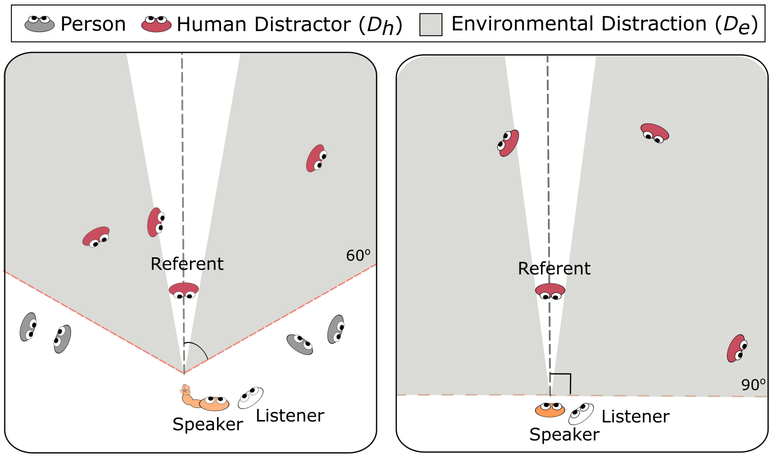


Figure 4.6: Open/Known scenario: examples showing the influence of distractors on behavior selection. Casual pointing is chosen. (b) Gaze only is chosen.

"Closed/Known" and "Closed/Unknown" scenarios, our model always selects "gaze only." This is consistent with the human behaviors observed in the data collection, where "gaze only" is the most frequently observed human behavior. Furthermore, in the data collection, people avoided using precise pointing for both "Closed" scenarios, and our model also behaves in the same way - the specificity (true negative rate) for precise pointing in "Closed/Known" was 98.0% and for "Closed/Unknown" was 93.1%.

In the "Open" scenarios, casual pointing constituted the majority of observed deictic behaviors (70.6% for "Open/Known" and 63.7% for "Open/Unknown"), and our model similarly predicted casual pointing the majority of the time (80.4% in both scenarios). The model was less successful in reproducing the other pointing behaviors, and we believe this variability could be due to individual preferences, or possibly related to unmodeled factors such as the precision of the descriptive terms used. It is also possible that gaze only and casual pointing can be used interchangeably in some situations, in which case multiple behaviors might be socially appropriate.

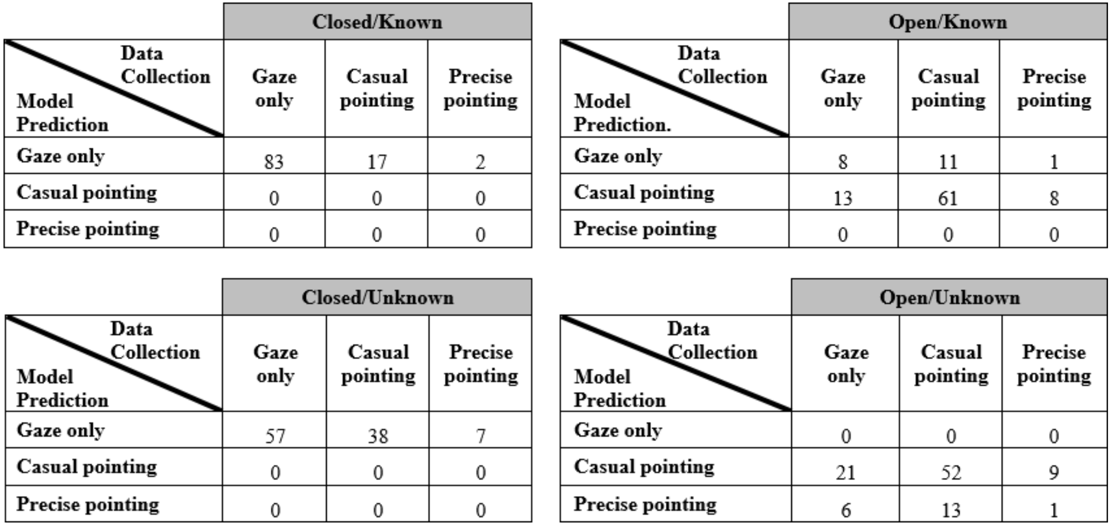
## 4.5 System Elements

Fig. 4.7 illustrates the system architecture for autonomously generating the robot's pointing behavior and utterances. We set up the **human tracking system** in the entrance hall of a shopping mall, covering an area of approximately 15m by 15m, as shown in Fig. 4.1. Pedestrian tracking was performed using the ATRacker[1] human tracking system presented in [34], utilizing 6 laser range finders (LRF's) mounted in portable poles placed around the environment. This system combines range data from multiple sensors to track the trajectories of potential distractors in the environment using particle filters, and can provide position data within 6 cm error at a data rate of 37 Hz.

---

[1] ATRacker is a product of ATR Promotions: http://www.atr-p.com/products/HumanTracker.html

Table 4.4: Confusion matrix for observed behavior from data collection and model prediction



| | Closed/Known | | |
|---|---|---|---|
| Data Collection / Model Prediction | Gaze only | Casual pointing | Precise pointing |
| Gaze only | 83 | 17 | 2 |
| Casual pointing | 0 | 0 | 0 |
| Precise pointing | 0 | 0 | 0 |

| | Open/Known | | |
|---|---|---|---|
| Data Collection / Model Prediction. | Gaze only | Casual pointing | Precise pointing |
| Gaze only | 8 | 11 | 1 |
| Casual pointing | 13 | 61 | 8 |
| Precise pointing | 0 | 0 | 0 |

| | Closed/Unknown | | |
|---|---|---|---|
| Data Collection / Model Prediction | Gaze only | Casual pointing | Precise pointing |
| Gaze only | 57 | 38 | 7 |
| Casual pointing | 0 | 0 | 0 |
| Precise pointing | 0 | 0 | 0 |

| | Open/Unknown | | |
|---|---|---|---|
| Data Collection / Model Prediction | Gaze only | Casual pointing | Precise pointing |
| Gaze only | 0 | 0 | 0 |
| Casual pointing | 21 | 52 | 9 |
| Precise pointing | 6 | 13 | 1 |

A **dialogue generator** able to produce utterances for the robot to speak was also implemented in the robot platform. This was used for producing questions to start each trial, such as, "who did you see at the bus stop yesterday?", as well as for generating deictic utterances based on the openness of the conversation and the familiarity of the referent. When necessary, the content for descriptive terms was automatically generated based on information entered before the experiment by a human experimenter (i.e., the person's name and their badge color).

With current speech recognition technology, it is difficult to accurately understand a person's speech in a noisy shopping mall. This noisy environment may risk the results of the experiments not making sense (e.g. if the robot misrecognized the name of the referent chosen by the listener). To mitigate such risk, a human operator acts as a **speech recognizer** by listening to the listener's utterance transmitted through a GUI. Upon hearing the listener's response for the chosen referent, the operator manually tags the referent among the set of people detected by the human tracking system, and clicks "start" to trigger the calculation of the most appropriate deictic behavior in the **generative model**, which was implemented in the robot using all the equations with calibrated parameters. Through its **speech synthesizer** and **actuator**, the robot autonomously executes the selected deictic behavior based on the output of the model.

### 4.5.1 Robot Platform

The robot platform we used was Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, a wheeled base, and a speech synthesizer. We implemented motion behaviors for the three pointing behaviors: gaze only, casual pointing, and precise pointing (Fig. 4.8), and we implemented utterance behaviors incorporating the use or non-use of a descriptive term. Robovie's pointing gestures

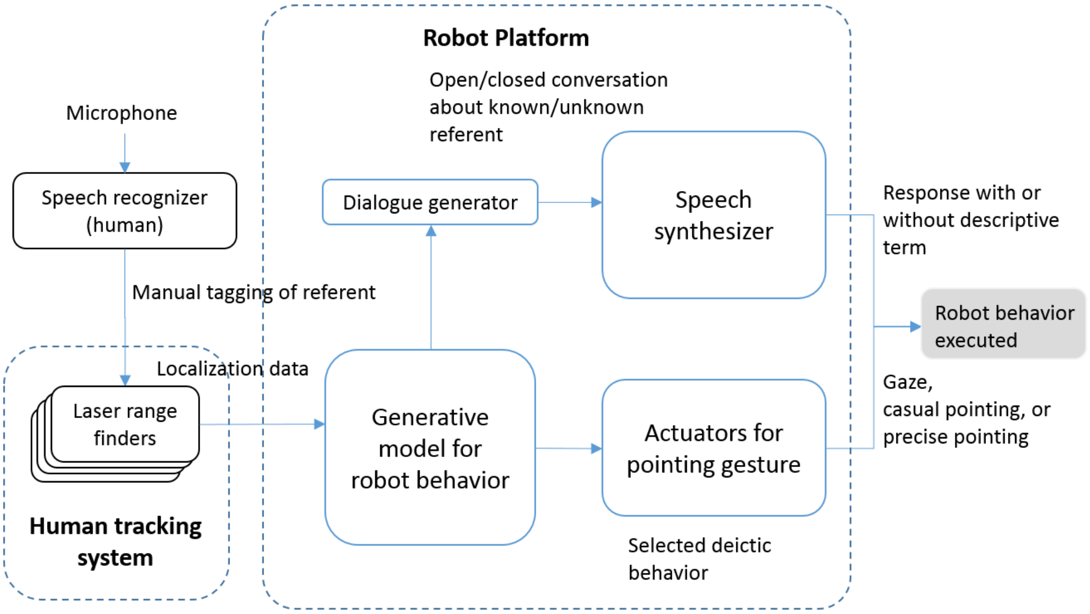
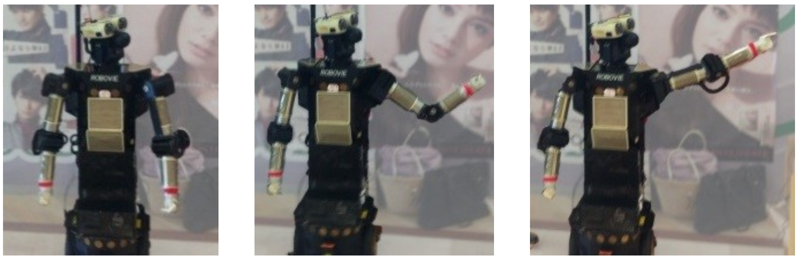

Robot Platform
Open/closed conversation about known/unknown referent
Microphone
Speech recognizer (human)
Dialogue generator
Speech synthesizer
Response with or without descriptive term
Manual tagging of referent
Robot behavior executed
Localization data
Laser range finders
Generative model for robot behavior
Actuators for pointing gesture
Gaze, casual pointing, or precise pointing
Human tracking system
Selected deictic behavior

Figure 4.7: System architecture for person-reference model: Inputs from speech recognizer and human tracking system are fed into the generative model, which then automatically calculates the appropriate deictic behaviors. The robot then responds verbally through its speech synthesizer and generates gaze and pointing gestures with its actuators.



(a) Gaze Only (b) Casual Pointing (c) Precise Pointing

Figure 4.8: Examples of Robovie performing the three pointing behaviors.

were implemented to best resemble what was commonly observed among the human participants.

## 4.6 Evaluation with a Robot

### 4.6.1 Comparison System

The purpose of pointing is to provide information, so it is reasonable to assume that a baseline pointing model would be optimized for understandability. In a field experiment, we compared the performance of our model against a model that considers only understandability but not social utility. This comparison model was chosen because it represents a typical state-of-the-art approach to generate deictic behaviors for referring to objects, such as demonstrated in the work of [50, 110, 116, 129]. In the following section, the baseline model will be referred to as the "object-reference model".

### 4.6.2 Hypotheses

We made the following hypotheses for the referent and listener:

**Predictions for referent evaluations**

1. The referent will perceive the robot's behavior as *more polite*. Since the robot's pointing will be less precise, the referent is less likely to feel singled out.

2. *Understandability* will be*lower* with the person-reference model, as the intention of social utility is to reduce the risk of the referent's awareness of conversation.

3. The referent will perceive the robot's behavior to be *more natural* because the person-reference model is calibrated after observations of real human behavior.

4. Politeness will be more important than understandability, since the referent is not directly involved in the conversation. Thus the referent will evaluate the proposed model as *better overall* than the object-reference model.

**Predictions for listener evaluations**

1. Listeners will rate the robot as *more polite* with the person-reference model, due to sympathy with the referent, and because the listener will feel uncomfortable if information is leaked to the referent in closed conversations.

2. *Understandability will be sufficient* with the person-reference model. Although there is a tradeoff between understandability and social utility, the model will provide enough understandability for the listener.

3. The robot's behavior will be rated *more natural* because the person-reference model is calibrated after observations of real human behavior.

4. As the person-reference model determines an appropriate balance between understandability and politeness, listeners will rate it *better overall* than the object-reference model.

### 4.6.3 Experiment Setup

We implemented our model in a communication robot and hired participants to evaluate the robot's behavior in a series of short interactions. The experiment used a within-participants design and was counterbalanced between two conditions: *person-reference model* and *object-reference model.*

#### 4.6.3.1 Procedure

We compared two conditions: the *person-reference-model* condition (our proposed model, including understandability and social utility) and the *object-reference model* condition (including understandability, but not social utility).

One participant acted as a listener and conducted short question-and-answer interactions with Robovie in a shopping mall. The other participant and a confederate acted as other customers. For each condition, Robovie and the listener asked each other a series of 8 questions: 2 questions each for four scenarios: Open/Known, Open/Unknown, Closed/Known, and Closed/Unknown, and each time Robovie made a reference to either the second participant or the confederate.

To prepare for the "known" scenarios, the participants and the confederate were asked to introduce themselves. This self-introduction was also intended to make the participants feel more invested in the conversation so they would become embarrassed if information were leaked in "Closed" scenarios.

Participants' names were entered into the system before each trial, so the robot could refer to the referent by name in "known" scenarios. To standardize the descriptive terms for the "unknown" cases, the human distractors wore different colored badges so Robovie could refer to them by their badge color.

For "Open" scenarios, the listener asked Robovie two pre-determined "neutral" questions. For the "Closed" scenarios, Robovie asked the listener two pre-determined "sensitive" questions, e.g., "Which person do you think has bad fashion sense?" The listener answered by selecting either the second participant or the confederate. Because we believed that the listener might feel embarrassed by Robovie's impoliteness, Robovie then repeated the opinion stated by the listener while performing the selected deictic behavior, e.g. pointing while saying, "So you think Tanaka-san has poor fashion sense?"

Since the volume of the robot's voice may affect evaluations, we adjusted the volume of the robot's voice to be louder in the "Open" scenarios. For the "Closed" scenarios, the volume was adjusted to a level that only the listener could hear.

After the four scenarios in one condition were completed, both participants answered questionnaires. The procedure was repeated with the remaining condition (*person-reference model* or *object-reference model).* The conditions were counter-balanced. At the end of the experiment, the participants were interviewed to gain a deeper understanding of their opinions.

#### 4.6.3.2 Environment

All trials were conducted on weekdays in the same shopping mall location as the data collection. As the other people in the environment were shopping mall customers, we could not explicitly control the degree of crowding. However, we believe that the distribution of people in the environment was fair between conditions. On average, in the *person-reference model* condition, 6.61 people (s.d. 3.75) were present in the environment, compared with 6.53 people (s.d. 3.93) in the *object-reference model* condition.

#### 4.6.3.3 Measurement

Both the listener and the referent rated the following items on a 1-7 scale (1 being very negative and 7 being positive for the respective items) in a written questionnaire:

1. *Naturalness* of the robot's deictic behavior.
2. *Understandability* of the robot's deictic behavior
3. *Perceived politeness* of the robot's deictic behavior
4. *Overall goodness* of the robot's deictic behavior

Because there were variations in the operator's speed and level of ambient noise, participants were asked not to consider timing or volume of the robot's utterances in their evaluations.

#### 4.6.3.4 Participation

A total of 26 trials were conducted. 33 participants were hired (19 male, 14 female, average age 23 years old). 19 participants played the roles of listener and referent in different trials, but no participant played either role twice.

## 4.7 Results

### 4.7.1 Verification of Hypothesis 1(Referent)

Figure 4.9(a) shows the questionnaire results from the referents. A one-way repeated-measures analysis of variance (ANOVA) was conducted with one within-participants factor, *model*, in two levels: *object-reference model* and *person-reference model*, for all measurements. The analysis revealed significant differences in *overall evaluation* ($F(1,25)=21.763$, $p<.001$, $\eta^2=.465$), *politeness* ($F(1,25)=15.391$, $p=.001$, $\eta^2=.381$), and *naturalness* ($F(1,25)=7.335$, $p=.012$, $\eta^2=.227$), and there was an almost-significant difference in *understandability* ($F(1,25)=3.362$, $p=.079$, $\eta^2=.119$).

These results support our hypothesis that the referents would perceive the overall behavior to be better with the person-reference model. The result also supports our predictions for *politeness* and *naturalness*, but not our prediction for *understandability*.

### 4.7.2 Verification of Hypothesis 2 (Listener)

Figure 4.9(b) shows the questionnaire results from the listeners. A one-way repeated-measures ANOVA was conducted for all measurements. There were significant differences in *overall evaluation* ($F(1,25)=10.192$, $p=.004$, $\eta^2=.290$), *politeness* ($F(1,25)=25.0$, $p<.001$, $\eta^2=.500$), and *naturalness* ($F(1,25)=4.972$, $p=.035$, $\eta^2=.166$), but no significant difference in *understandability* ($F(1,25)=2.235$, $p=.147$, $\eta^2=.082$).

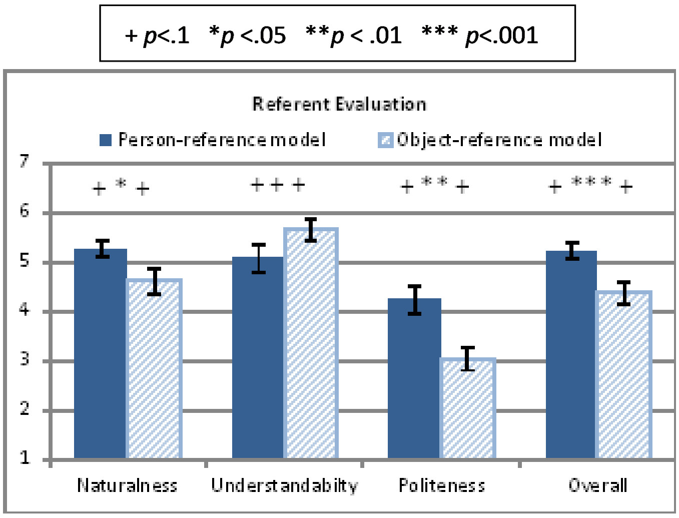
These results support our prediction that listeners would rate the person-reference model better in *overall evaluation*, as well as our predictions for *politeness* and *naturalness*.

### 4.7.3 Analysis of understandability and social utility on the behavior-selection process

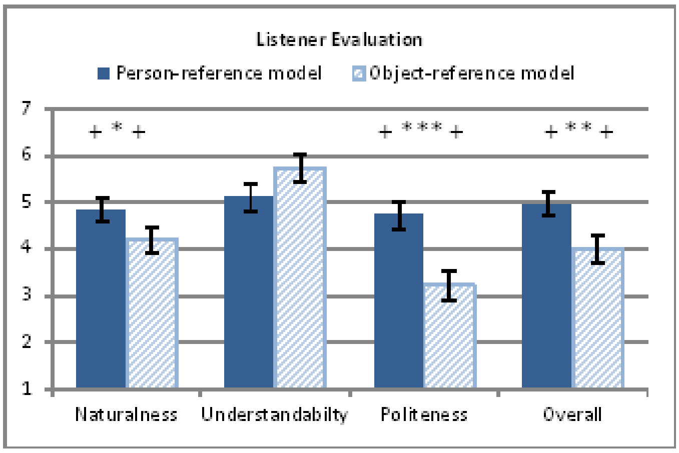
Our comparison experiment demonstrates how our *person-reference model* that considers both understandability and social utility can be used to improve the overall robot's deictic behavior, as compared with a model considering understandability alone. We provide a numerical analysis on the interactions observed in our experiment, in order to demonstrate how the values of understandability and social utility contributes to the behavior-selection process of our *person-reference model* under different scenarios.

To illustrate the tradeoff between understandability and social utility, Fig. 4.10 shows plots of the numerical values of understandability, social utility, and total behavior utility as a function of the amount of distraction in the environment (based on a "gaze only" pointing cone) for all of the "Open/Known" trials in our experiment. In this scenario, the robot's verbal behavior defaults to using the referent's name (i.e. without the use of a descriptive term), thus only 3 deictic behaviors are possible. Based on equations (4.4.1) – (4.4.3), we expect understandability to decrease linearly with $D_T$, as observed in Fig. 4.10 (a). Precise pointing, with the smallest $\theta_{\text{pointing precision}}$, results in the highest value in understandability, followed by casual pointing and gaze only. From equations (4.4.7) and (4.4.8), we expect the negative effect of social utility to become weaker as $D_T$ increases, as seen in Fig. 4.10 (b), since the greater amount of crowding reduces the feeling of being singled out. Note that in Fig. 4.10 (b), many of the data points for precise pointing fall below the bottom of the graph, since precise pointing leads to the lowest social utility.

Finally, our *person-reference model* considers both understandability and social utility. As shown in Fig. 4.10 (c), the behavior utility of "gaze only" decreases as the environment becomes more crowded, whereas the behavior utility of "casual pointing"

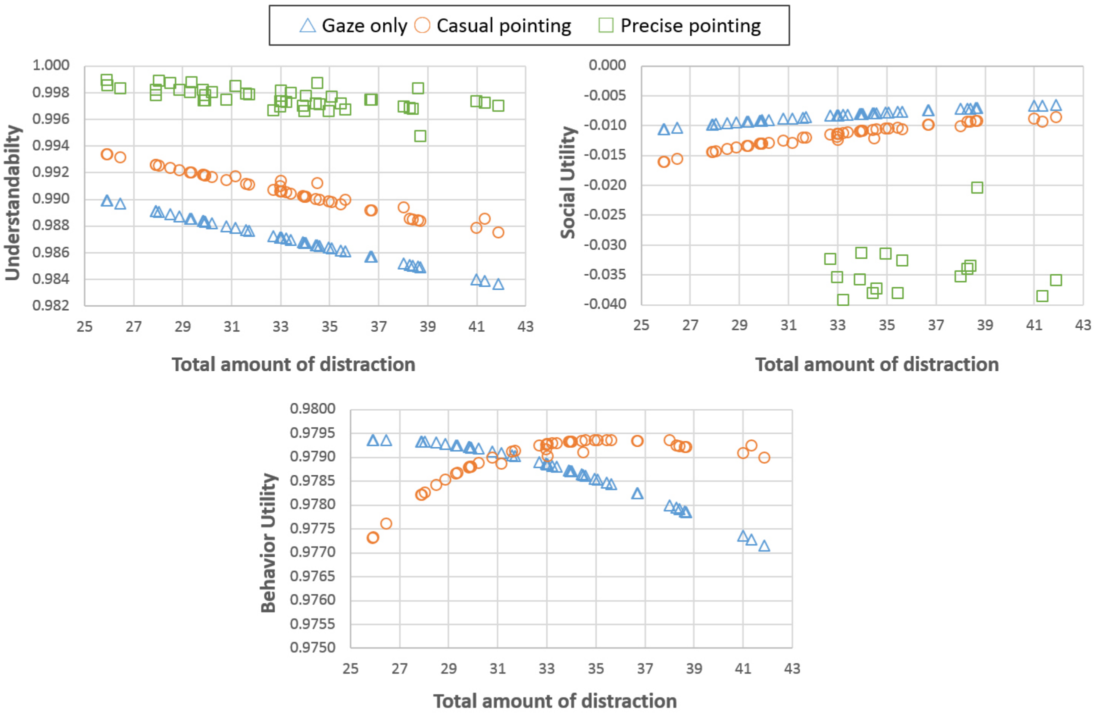


(a) Referent evaluation between conditions



(b) Listener evaluation between conditions

Figure 4.9: Evaluation results of Robovie's behaviors between conditions



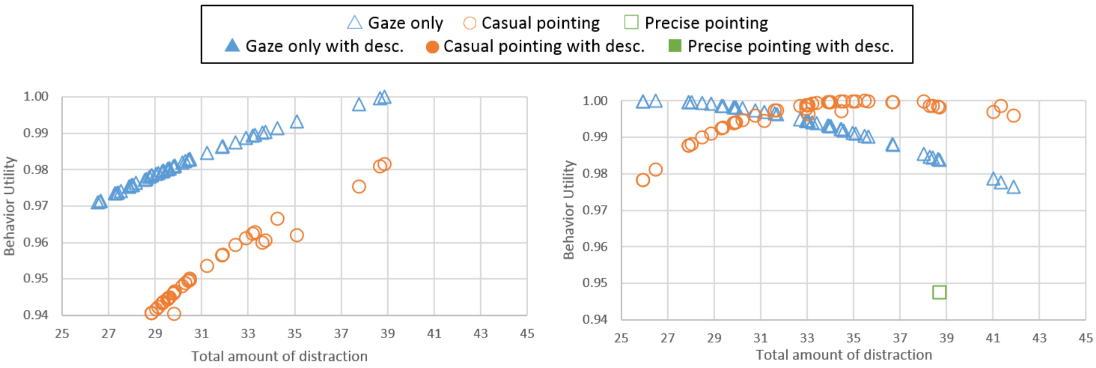
,

Figure 4.10: The x-axis is the total amount of distraction, $D_T$, in the environment observed by the robot's forward facing horizontal field-of-view of 180 degrees at behavior execution time. For the three pointing behaviors in "Open/Known" scenario, the values of: (a) "Understandability" is negatively linear proportional to $D_T$, (b) "Social Utility" is inversely proportional to $D_T$, (c) "Behavior Utility", gaze only has the highest behavior utility when $D_T$ is low, and casual pointing has the highest behavior utility as $D_T$ increases. "Precise pointing" is too low to be shown on this figure, although it might be selected in extremely crowded situations.

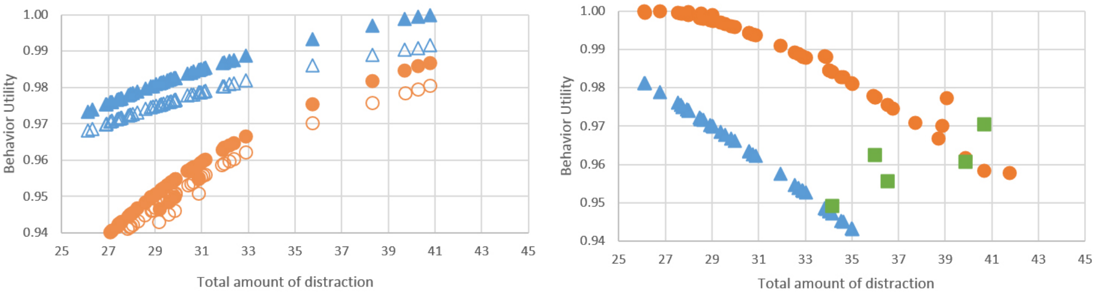
has an increasing trend as $D_T$ increases. The model selects the behavior with the highest behavior utility, resulting in "gaze only" when $D_T$ is low, and "casual pointing" as $D_T$ increases. The behavior utility of "precise pointing" is too low to be shown on the figure, although it might be selected in extremely crowded situations. By contrast, the *object-reference model* would choose "precise pointing" in all cases, to maximize understandability regardless of social factors.

Next, we provide an analysis of the behavior-selection process of our *person-reference model*, or the behavior utility values, for all scenarios. When the referent is unknown, a verbal description can also be used to resolve ambiguity of the referent. Verbal description increases understandability, but has no effect on social utility. Therefore, the robot will use verbal description along with a pointing gesture for all unknown referents, as can be seen in Fig. 4.11 (c).

In "Closed" conversation, the robot always chooses to use "gaze only," as shown in 4.11 (a) and (c). The high value of $C_{social}$ leads to a greater influence of social utility



**(a) Closed/Known** **(b) Open/Known**



**(c) Closed/Unknown** **(d) Open/Unknown**

Figure 4.11: Behavior Utility (normalized by scenario) of the three pointing behaviors calculated by the model for all experimental trials. The deictic behavior with the highest behavior utility is chosen by the robot. (a) "Closed/Known": gaze is always selected, (b) "Open/Known": Gaze is chosen when $D_T$ is low, but casual pointing is chosen as $D_T$ increases, (c) "Closed/Unknown": Gaze-only with description is always chosen, (d) "Open/Unknown": Casual pointing is mostly chosen, but when the total amount of distraction is high, precise pointing with description is chosen.

than understandability in the values of behavior utility. As a result, the behavior utility follows the same trend of its social utility.

In "Open" conversation, a tradeoff between understandability and social utility is observed in the behavior-selection process. In the "Open/Known" case, the robot uses "gaze only" when $D_T$ is low. As $D_T$ increases and requires a more precise pointing gesture to resolve ambiguity, the robot uses casual pointing (Fig. 4.11 (b)). In the "Open/Unknown" case, the robot mostly used casual pointing together with a verbal description in our experimental environment. However, an increasing trend of using precise pointing together with a verbal description can be observed when there is a large amount of distraction (4.11 (d)).

## 4.8 Discussion

### 4.8.1 Interview results from the participants

Many participants said that they rated our proposed model better because the robot behaved more politely. For listeners, it was particularly embarrassing when the robot repeated his/her negative comment about the referent together with precise pointing. One participant commented she was worried the referent might get angry if he overheard the negative comments about him. It is interesting to note that one participant perceived the robot to be more "child-like" in the *object-reference model*, since the participant associated impoliteness in the robot's pointing behavior with the behavior of a child.

No significant difference was found for understandability. One possible reason is that the referents were asked to watch and evaluate the robot, so they were inevitably more aware of the conversation than a typical bystander would be.

### 4.8.2 Politeness in Pointing

When we first tried to set up a preliminary observation of people's pointing behavior, we set up a scenario where we asked the participants to imagine role-playing as a store clerk who was trying to indicate a store manager to a customer. In this scenario, we found that often the participants role-playing as the clerk were reluctant to use the index-finger pointing to identify the store manager in almost all scenarios, even when it was ambiguous who the store manager was due to the crowds. Instead, the participants used the more polite form of pointing, often with their palm up and hand open, to show the customer the whereabouts of the store manager. This was categorized as "Open Hand Supine" by Kendon, which may be semantically described as presenting or being ready to receive [63, 84]. It is possible that the participants were using this gesture to present the manager (i.e. the referent) to the customer.

Keeping in mind of these observations, it is important to consider the role and purpose of the robot when designing social behaviors. As described, a person may use more polite pointing gestures when interacting with a professional relation like a superior or customer, as compared to interacting with a friend. Thus, for example, a customer service robot might need to give priority more polite pointing behaviors, whereas a personal companion or non-humanoid robot might choose to maximize the understandability in its pointing behavior.

### 4.8.3 Limitations and future work

In this study, we developed a model for choosing deictic gesture and utterance behaviors that balance the issues of being polite and being easy to comprehend. While our study used a general categorization of gestures such as casual and precise pointing, there are many details which could be investigated in future work regarding the implementation and details of those gestures. For example, Kendon contrasted the semantic implications for different orientations of open-hand pointing – when a person introduces another person, they usually use an open hand, palm up gesture as an implication for offering [63], whereas when a person makes a critical remark, they use an oblique open-handed pointing. Incorporating the semantic meanings of the pointing gestures into our models may extend the robot's role and the scope of its interaction. For instance, a robot shop assistant presenting the manger or an educator making a critical remark may require the robot to adapt different subtle pointing hand orientations. Our study also examined the effect of the use or non-use of descriptive terms, but future research could investigate the relative effects of different kinds of descriptive terms or different levels of specificity. It may also be possible for models to be developed to quantify the degree of precision of a given pointing gesture, enabling more precise estimation of the pointing cone.

We understand that the use of kinesics or deictic behaviors may vary among cultures. The participants of this study were all Japanese, in which using body language may be remarkably restrained away from their in-group [54]. It is also worth noting that Japanese people may refrain from making hand gestures when the third-person referent is present, possibly to reduce the opportunity for offending anyone present and help sustain contextual harmony [107]. Imaginably, if this study was conducted in another culture, we might observe participants using more precise pointing. However, we believe our model does represent a universal phenomenon of how people point toward others, and hence, a valid model for robot deictic behavior as well.

In our experiment, a human operator was employed for two tasks: (4.4.1) to input the participant's name and the color of their name tag into the system before each trial, for later use in generating descriptive terms; and (4.4.2) to act as a speech recognizer in real time and tell the robot which person the chosen referent was. If this technique were to be used in a real social robot application, we expect that the referent's name would already be known, and the other functions could potentially be automated, e.g. using computer vision for identifying clothing color, and using gesture recognition and speech recognition to understand who the referent is. While implementing these functions robustly is not trivial, we expect that with improvements in sensor technology the technique could be employed in an autonomous way.

While there are several possible directions for future work and refinement of the techniques presented here, we believe that this study has provided a successful demonstration of a practical technique for reproducing an important phenomenon which occurs in real human deictic behavior.

## 4.9 Conclusion

In this work, we have presented a model enabling robots to generate socially-appropriate deictic behaviors for referring to people, based on the openness of the conversation and familiarity with the referent, as well as the positions of people in the environment. In an empirical data collection, we observed that people's behavior varied both in terms of their pointing behaviors ("gaze only", "casual pointing", and "precise pointing") and their use of descriptive terms. We confirmed that people's deictic behaviors towards another person differed from their deictic behaviors towards objects, and we observed variation according to social context and presence of other people in the environment. From this data we developed a model enabling a robot to select socially-appropriate deictic behaviors towards humans based on a balance between understandability and social appropriateness for a given scenario.

Finally, we evaluated our model using a real robot in a shopping mall in an experimental comparison between our proposed model and a simpler model based only on understandability. The results showed significant differences for perception of the robot's deictic behaviors, in which the robot's behaviors were perceived to be more natural ($p < 0.05$ for both the referent and listener), polite ($p < 0.01$ for the referent and $p < 0.001$ for the listener) , and better overall ($p < 0.001$ for the referent and $p < 0.01$ for the listener) when using our proposed model. These results confirm that by considering social appropriateness in the model we were able to generate better social behavior for the robot.

# Chapter 5

# Discussion and Conclusions

## 5.1 Summary and Achievement

The theme of this thesis was to explore, research, and implement how a data-driven approach can be applied to human-robot interaction in real-world environments. The core contribution of this work is presented as follows:

- Reproducing multimodal behaviors for a conversational robot
- Extending data-driven techniques for a mixed-initiative social interaction
- Developing a reusable HRI pointing model in order to extend data-driven application to other modalities

The contribution in each chapter described above were integrated into a Mixed-Initiative Interaction Learning Framework, to enable a seamless end-to-end framework for preprocessing data, abstracting to common behavior patterns, complementing with existing HRI models, extracting typical generative behaviors, and learning interaction logic. The work in this thesis have been tested through many field experiments or demonstrations, where participants were able to interact with a humanoid robot using natural spoken dialogue and motion.

Many roboticists may propose other methodology to be more apt when it comes to designing human-robot interaction. Certainly, many interaction development approaches such as coding, designing interactions through a graphical interfaces, or teaching through teleoperation are also valid and well-established methodologies in the HRI field. The objective of this work is not to dispute the validity of these methodologies, but rather to help expand the available set of tools and techniques available in a new conceptual direction, and to demonstrate this approach to be worthy of consideration. What I want to accomplish is to provide some first insights into its potential, as well as its limitations. This knowledge is very important to the ongoing discussion about data-driven techniques in the HRI community.

## 5.2 Discussion

### 5.2.1 Future Direction

While I proposed a data-driven approach that required non-annotated data for learning interaction logic and generating robot behaviors, there are many possibilities for future work to make it more generalizable to other domains.

#### 5.2.1.1 Learning semantics of the conversation

One limitation of the current work is that while the interaction is multimodal (i.e. speech and motion), the learning is superficial, based only on the lexical form of the text, and has no understanding of the deeper semantics of the conversation.

While the current approach works for our scenario and may be generalizable to domains such as a museum tour guide robot, it relies on certain assumptions we made for our system. For example, there are the underlying assumptions that each product stays at a fixed location and that information regarding a specific product does not change. Thus it is as yet unclear to what degree this approach will work, such as how to update the robot when the price of a camera changes. It will be interesting to explore additional techniques in the future that does not require new data to be collected when product information has changed.

Models of the structure of conversation would also be useful tools for extending this work into more complex domains. Some work has explored the use of generic dialogue patterns in HRI [55, 99], and it is plausible that some kind of templates could be used to help structure data-driven HRI, especially if utterances could be analyzed semantically.

#### 5.2.1.2 Reinforcement learning

Another possible extension is to complement this approach with corrective guidance or a reward function to improve the robot's performance. Thomaz and Breazeal employed a reinforcement-learning technique to their learning framework, where the result of people giving rewards to a virtual robot in a simulated game improved their system's performance, including a significant drop in the number of failed trails encountered during learning [133]. Thus, it would be interesting to see if such a technique could be used in conjunction with my approach to dynamically improve human-robot interaction in real-time.

#### 5.2.1.3 Hybrid approach

One constraint to my learning approach is that rules for interaction logic are learnt probabilistically, and often these learnt rules are only comprehensible by the machine and not by a human, making the task of a human debugging a trained system difficult.

One potential direction is to investigate data-driven techniques that can automatically extract human-readable rules for interaction logic, which would open up the potential for a human designer to improve the performance of the trained system by editing specific rules. A flowchart approach could then be used to supplement the data-driven approach, as editing rules from a flowchart interface is intuitive for a human designer.

### 5.2.2 Behavior modeling in HRI

The current study used existing HRI proxemics models in order to create generalizable behavior templates that could be recognized and reproduced, such as the present-object formation. These models provide generalizable structural elements which can be helpful in learning complicated interactive behaviors. In the future, it may be beneficial to add new modalities to our learning system. To do so, the system would need to (1) recognize human actions using the new modality (e.g. pointing, grasping), (2) represent those actions in the feature vector, and (3) generate those actions again by the robot, in order to mimic the humans.

In our scenario, the robot would benefit greatly if it were able to perceive and actuate pointing or grasping. For example, a few participants pointed to a camera at a different location and said, “what about that one?”, but the robot was unable to figure out which camera the participants were pointing to. The robot’s lack of pointing capability also confused some participants, especially when the robot was first trying to guide the participants to another camera. It would also be beneficial if the robot could perceive or actuate grasping, so that it could assist the customer when he or she picked up a camera, or so the robot could pick up a camera to show to the customer. Incorporating existing models for pointing [47, 74] and grasping [45, 59] would help with the learning system.

Interestingly, there may be some situations in which it would be useful for the robot to be capable of perceiving the new modality, even if it were not able to to generate behaviors in that modality itself. For example, the ability to recognize a customer’s emotions based on body language or facial expressions might help a robot like Robovie provide better service by improving its interaction logic to be responsive to the customer’s emotions. This would be helpful even if the robot did not have the capacity to mimic the body language or facial expressions of the shopkeeper.

### 5.2.3 Representation of new semantics

It would be great if we could extend our current approach to incorporate new semantics. By representing semantic features, such as the properties relating to the products or the customers, in the feature vector, it might be possible to learn generalizable patterns. For example, if different models of the cameras (e.g. DSLR versus point-and-shoot) were represented as new semantics, then the robot might learn some DSLR-specific (but not model-specific) utterances that could be used at multiple cameras. To do so, it might be possible to employ existing object recognition systems [76, 100], or use a cloud-based service [61] to recognize different camera models.

Another possible semantic representation is the knowledge about the customer. Currently, the robot has no knowledge about the customers’ identities, such as their age, gender, or if they are a repeated visitor. Often a salesperson may change his sales pitch or way of presentation depending on who the customer is. We can imagine a salesperson recommending a simple point-and-shoot camera to an elderly woman. If we can capture such information about the customer and represent it as a semantic, the robot may be able to provide a better and more personalized interaction [37].

### 5.2.4 Generalizability

The core principle of data-driven HRI is using generalizable models to learn repeatable social tasks, as well as learning âĂIJconditionsâĂİ for these social tasks from real, noisy interaction examples from physical environment. The social tasks may be domain-dependent and/or role-dependent. For each domain, there are certain tasks a service provider must fulfill, for example, a museum guide should explain about the artist of an exhibit, while a travel agent should provide information and answer questions about tour packages.

In all these scenarios, the service providers perform the same interactions, in the same way, many times. Because the shopkeeper may say similar utterances over and over again, we can extract typical utterances through clustering and learn to generate robot behaviors from repeatable patterns in the interaction data. Thus, this approach could probably be most effective in domains where such repeatable characteristics exist, such as for a museum guide, a tourist information guide, or a receptionist.

Another element of data-driven approach is to use generalizable behavior models (i.e. proxemics formations) to ease the difficulty of learning from high-dimensional, unconstrained, real human-human interaction. Because we used behavior models for proxemics formations, low-level sensor data related to certain tasks does not need to be learned, but rather can be generalized and abstracted into high-level semantics. For example, in our scenario, a task of the shopkeeper is to approach the customer. Rather than learning every possible position of where the robot shopkeeper could be in relation to the customer, we used the âĂIJface-to-faceâĂİ behavior model to abstract the interaction data and to generate reliable robot movements.

Because interaction data are captured in real environment, we also can learn the “conditions” for these social tasks in a real environment. Rather than programming what “keywords” the robot should respond to, noisy speech recognition are captured. For example, the word âĂIJkaraâĂİ is often recognized instead of âĂIJcolorâĂİ, which would not have been intuitive for a programmer. Our robot would still behave appropriately under such conditions (i.e. when the customerâĂŸs utterance is misrecognized as âĂIJkaraâĂİ), whereas a hand-crafted robot may have made a mistake.

There may be some domains to which our approach cannot easily be applied. These domains could consist of non-repeatable tasks, which cannot be learnt through our data-driven approach (e.g. a social worker, an educator). One example might be an educational robot that proactively teaches a language, where the lesson is tailored to the studentâĂŹs comprehension of that language. We imagine such domains would be difficult to learn with our current approach, since lessons are tailored to the knowledge of the student (i.e. level of comprehension).

There may well be other domains where manual programming the robot or teaching through teleoperation might be considered more practical than learning from data. For example, robots as performers may require a human designer to architect the robot's overall image or personality, and fine-tune its dialogue, movement, facial expression, or gestures. This approach has been used for robots in theatre performances [21], poetry-reciting agents [91], and news broadcaster [75].

### 5.2.5 How much control should an interaction designer have?

One of the consequences of using a data-driven approach is not all behaviors generated by the robots are as predictable as those from a scripted interaction. In my work, I had a pretty good grasp of which aspect of the interaction a robot would succeed in (e.g. answering questions) and which aspect of the interaction the robot would be prone to failure (e.g. when the user does something not observed in the training data). How well a robot learns a behavior depends on the quality of the data, which heavily relies on how well the demonstrator is able to provide repeatable training examples that are within the realm of the scenario.

Thus, one potential risk of a trained system is the quality of the training data really matters. For example, it took less than 24 hours for Twitter to corrupt an innocent Microsoft AI chatbot, Tay. As people started to engage with this "conversational understanding" chatbot, it seems like Tay assimilated many of the internet's worst tendencies into its tweets, and it learned all sorts of misogynistic, racist, and bigoted remarks, proving correct the old programming adage: flaming garbage pile in, flaming garbage pile out.

Thus, I argue that a certain degree of control is necessary when it comes to applying a data-driven approach, either by screening out poor demonstrators or implementing a mechanism to filter out unwanted training data. This is important, especially given the trend of crowdsourcing data, allowing people from the web, or even anonymous users, to pilot interactive behaviors. We should also consider when data are being obtained in a real shop, when some employees are actually not very good at their job. It could be helpful to filter these people out before training.

### 5.2.6 Robots as customer service representatives

The role of the robot as a customer service provider (i.e. shopkeeper) was the chosen domain for this work, which I believe will be the first frontier of commercial "robots in the wild" in human-robot interaction. Businesses have already started to use AI to figure out what customers want, and some AI chatbots has been deployed as service providers: such as Echo for the e-commerce giant Amazon, a customized Watson for the outdoor attire company North Face, and even TacoBot for fast-food chain Taco Bell. I believe that some aspects of customer service, where precision and efficiency is valued (e.g. answering questions, presenting products) is where some of the greatest breakthroughs in AI will be seen, and developing a platform that is capable of sophisticated decision-making processes will be essential for business to stay competitive.

Despite the concerns that robots will replace people in jobs like customer service, I believe it simply means that we can apply our talents in fresh ways and new jobs are born. Customer service employees often experience frustration and negativity in their job [41, 103], and have to face daily drudgery of answering the same queries. These repetitive and basic tasks can be outsourced to a learning robot. I imagine a scenario in the future where the robots will handle simple and mundane tasks, while the humans will transition to supervisory roles, and customer service will become more efficient, intelligent, and have more empathy. Additionally, we may see a growth in jobs specializing in maintaining social robots in commercial spaces, or perhaps even specialized actors to provide example interactions for a robot with personality.

### 5.2.7 The potential of data-driven approach

The question of what tasks are learnable still remains. There may be certain tasks, be it our appreciation for fine art or our desire for spiritual fulfillment, which may just not be possible to learn no matter how many resources or input data are presented to the system, and no learning machine can learn all possible tasks efficiently. It is a humbling experience when we realize how specialized and adaptable our brain can be, and that we have the capability to learn an enormous amount of knowledge and skills, and also are able to grasp abstract concepts and ideas. Even if we build machines with super-human intelligence, they will have limited abilities to outsmart us in the real world. That is why even though they may beat us at chess or go, they might not be able to truly appreciate something as simple as a beautiful sunset.

Finally, using a data-driven approach may expedite the deployment of robots in society. Domain-specific human-robot interaction can be quickly learnt once passive observation of human-human interaction from sensors becomes an acceptable practice for data collection. Once robots become a common fixture in our society, people will likely to be open and receptive to assistance from robots. Thus, I see this as a positive outcome, and the reality of humans and robots to coexist in our daily life may just be around the corner.

## 5.3 Conclusion

The contribution of this work is highly unique and novel, since it is the first work in the field of social robotics that provided a fully end-to-end approach to enable a robot to learn an entire mixed-initiative social interaction based solely on human-human interaction in a real, physical environment. While there are many existing data-driven approaches, many are limited to capturing data in a simulated environment, requiring human annotations, learning from text that is typed rather than spoken, only dealing with nonverbal behaviors, or reacting only to human inputs and not be able to take initiative on its own.

The techniques presented in this work have elucidated some of the key research challenges of learning from unlabeled data from unscripted human-human interactions in a real, physical environment. Unforeseen difficulties uncovered in each individual study, such as how individual HRI behavior models can be utilized to enable an entire interaction, or the necessity of proactive behaviors in interactions, provided motivation for further studies and improvements to the system.

Thus, in Chapter 2, we demonstrated that behaviors can be captured from real, non-annotated human-human interaction âĂŞ- that were completely free form and contained large amounts of variation in natural speech among the different participants – and we developed various techniques to enable this data-driven approach (i.e. clustering and utilizing existing behavior models in HRI for proxemics formations). This system was evaluated against state-of-the-art data-driven techniques, and was demonstrated to generate robust robot behaviors, despite a high amount of sensor noise.

In Chapter 3, we demonstrated that it is possible to extend our data-driven approach so that it covers proactive behavior, that is, be able to not only react to the customerâĂŹs actions, but also proactively take actions by introducing new camera features or a new camera. A comparison user-study of a baseline reactive system (i.e. system from first study) versus a proposed system (i.e. new data-driven techniques added to system from first study) demonstrated that this system was successful in generating socially-appropriate robot behaviors in a mixed-initiative social interaction.

To extend our data-driven approach to additional modalities and more complex social scenarios, we need to develop parameterized, generalizable behavior models, so we demonstrated example of how such a model could be developed in Chapter 4. The purpose of the evaluation in this case was to show the success in creating such a model, and indeed it showed that the model was meaningful and effective.

Through these studies, we have demonstrated the practical application of data-driven learning of high-level interaction logic in real human-robot interaction in a camera shop scenario, where participants were given a fair amount of freedom to walk around and use natural spoken language to interact with the robot. The robot shopkeeper was able to autonomously answer customer's questions, greet them, answer their questions, and proactively offer them information about the products.

Thus, this work provides a successful, concrete example of a coherent collection of data-driven techniques which can be used for abstraction of typical behaviors, learning of interaction logic, generation of interaction contents, and reproduction of proactive behaviors, which together enable the practical deployment of conversational social robot to provide services in real-world environments.

# Acknowledgments

I would like to thank my advisor who supervised my research during my Doctor's Course. First, Dr. Hiroshi Ishiguro, for inspiring me to push the boundary of robotic research and its role in society. I would like to thank Dr. Takayuki Kanda for his guidance and insight of what makes a good research. He gave me the flexibility and freedom to explore new research and technology challenges, while also teaching me the importance of presenting research problems with clear logical flow and methodology and providing valuable feedbacks when defending my idea. I would also like to thank Dr.Kensuke Harada and Dr.Koh Hosoda for contributing their time as readers for my thesis.

Next, I would like to thank the various talented researchers and engineers that I worked with at Advanced Telecommunication Research Institute International (ATR). First, I would like to thank Dr. Norihiro Hagita, as the lab director for providing the vision to the lab. Next, this work would not have been possible without the senior researchers in the lab, Dr. Masahiro Shiomi, Dr. Satoru Satake, Dr. Dylan F. Glas, who have contributed enormously to the technology that enabled the success of my experiments. I would like to thanks to Dr. Takashi Minato and Dr. Satoshi Koizumi for their help in providing resources, as well as negotiating and coordinating my experiments.

I would like to thank Dr. Dylan F. Glas, for his guidance and support during my graduate study. We have countless discussion regarding the implementation of a particular system, research, new research ideas, and the future direction of robotics. While there are things that we may disagree on, his input and advice has been an invaluable resource to help me think critically about a problem, as well as tackle a research problem from different viewpoints.

I first started this journey as a Canadian intern during my undergraduate study. Without all the interns, engineers, researchers, colleagues, friends, and administration staff at ATR over the years, I would not have been able to accomplish my study, specifically I would like to thank: Yunchen Fu, Kanae Wada, Benoit Toulmé, Kuanhao Zheng, Florent Ferreri, Thomas Kaczmarek, Greg Cole, Peace Cho, Gyuhee Lee, Clement Congard, Nick Mulligan, Sayaka Taniguchi, Kanako Tomita, Deneth Karunathethne, Nobuyuki Iwasaki, and all other staff at ATR. I would also like to thank all the other students and staff at Dr. Hiroshi Ishiguro's laboratory in Osaka University, for providing their support, help, and friendship.

Lastly, I would like to thank my parents, for their unconditional love and support to me. For it is my parents, whom showed me that there is no bad situation in life but only bad attitude, and whom gave me the freedom to pursue my goals, and believing in me

that it is possible to accomplish anything as long as you put effort and work into it. I would also like to thank my brother, Dr. Tony Liu, for his steadfast support throughout my study, and showing me the importance of work ethics.

The work presented in Chapter 4 was supported in part by the Ministry of Internal Affairs and Communications of Japan and in part by JSPS KAKENHI Grant Number 25240042. The work in Chapter 2 is supported in part by JSPS KAKENHI Grant Number 25240042 and in part by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project. The work in Chapter 3 is supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

The work presented in Chapter 4 has been granted permission by Springer and Copyright Clearance Center. It was originally published in International Journal of Social Robotics, “A Model for Generating Socially-Appropriate Deictic Behaviors Towards People”, Vol. TBA,2016,pp.TBA,Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro. The work presented in Chapter 2 has been granted permission by 2016 IEEE. Reprinted, with permission, from Phoebe Liu, Dylan F. Glas, Takayuki Kanda, and Hiroshi Ishiguro, “Data-Driven HRI: Learning Social Behaviors by Example from Human-Human Interaction”, in IEEE Transactions on Robotics, Vol. 32, No. 4, pp. 988-1008, 2016. All other materials ©2016 Phoebe Liu.

# Works Cited

[1] H. Admoni and B. Scassellati. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 196–199. ACM, 2014.

[2] M. Awais and D. Henrich. Proactive premature intention estimation for intuitive human-robot collaboration. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4098–4103. IEEE, 2012.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] A. Bangerter and E. Chevalley. Pointing and describing in referential communication: When are pointing gestures used to communicate? In *MOG 2007 Workshop on Multimodal Output Generation*, page 17, 2007.

[5] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Towards a humanoid museum guide robot that interacts with multiple persons. In *Humanoid Robots, 5th IEEE-RAS International Conference on*, pages 418–423, 2005.

[6] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[7] K. Berns and S. A. Mehdi. Use of an autonomous mobile robot for elderly care. In *Advanced Technologies for Enhancing Quality of Life (AT-EQUAL)*, pages 121–126, 2010.

[8] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013.

[9] A. G. Brooks and C. Breazeal. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 297–304. ACM, 2006.

[10] D. Bršćić, T. Kanda, T. Ikeda, and T. Miyashita. Person tracking in large public spaces using 3-d range sensors. *Human-Machine Systems, IEEE Transactions on*, 43(6):522–534, 2013.

[11] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1):3–55, 1999.

[12] J. K. Burgoon, D. B. Buller, J. L. Hale, and M. A. Turck. Relational messages associated with nonverbal behaviors. *Human communication research*, 10(3):351–378, 1984.

[13] S. Calinon and A. Billard. Teaching a humanoid robot to recognize and reproduce social cues. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pages 346–351. IEEE, 2006.

[14] A. Cesta, G. Cortellessa, V. Giuliani, F. Pecora, R. Rasconi, M. Scopelliti, and L. Tiberio. Proactive assistive technology: An empirical study. In *IFIP Conference on Human-Computer Interaction*, pages 255–268. Springer, 2007.

[15] C. Chao and A. L. Thomaz. Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human-Robot Interaction*, 1(1):1–16, 2011.

[16] S. P. Chatzis and Y. Demiris. Nonparametric mixtures of gaussian processes with power-law behavior. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(12):1862–1871, 2012.

[17] J. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

[18] S. Chernova, N. DePalma, E. Morant, and C. Breazeal. Crowdsourcing human-robot interaction: Application from virtual to physical worlds. In *RO-MAN, 2011 IEEE*, pages 21–26, 2011.

[19] S. Chernova, J. Orkin, and C. Breazeal. Crowdsourcing hri through online multiplayer games. In *AAAI Fall Symposium Series*, 2010.

[20] D. Choi, J. Ha, M. Jung, M. Park, and H. Park. Development of robot scenario script language and tool for non-expert. *Journal of Automation and Control Engineering Vol*, 3(6), 2015.

[21] D. Choi, D.-W. Lee, D. Y. Lee, H. S. Ahn, and H. Lee. Design of an android robot head for stage performances. *Artificial Life and Robotics*, 16(3):315–317, 2011.

[22] H. H. Clark. *Using Language*. Cambridge University Press, 1996.

[23] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[24] C. Datta, C. Jayawardena, and B. MacDonald. Robostudio: A visual programming environment for rapid authoring and customization of complex services on a personal service robot. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2352–2357. IEEE, 2012.

[25] M. P. Deisenroth, C. E. Rasmussen, and D. Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. 2011.

[26] J. P. Diprose, B. Plimmer, B. A. MacDonald, and J. G. Hosking. A human-centric api for programming socially interactive robots. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 121–128. IEEE, 2014.

[27] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[28] S. Duncan. On the structure of speaker-auditor interaction during speaking turns. *Language in society*, 3(02):161–180, 1974.

[29] R. J. Edelmann. The effect of embarrassed reactions upon others. *Australian Journal of Psychology*, 34(3):359–367, 1982.

[30] M. Forsberg. Why is speech recognition difficult. *Chalmers University of Technology*, 2003.

[31] M. E. Foster, S. Keizer, Z. Wang, and O. Lemon. Machine learning of social states and skills for multi-party human-robot interaction. In *Proceedings of the workshop on Machine Learning for Interactive Systems (MLIS 2012)*, page 9, 2012.

[32] D. Fox, W. Burgard, and S. Thrun. The dynamic window approach to collision avoidance. *Robotics & Automation Magazine, IEEE*, 4(1):23–33, 1997.

[33] C. P. Gharpure and V. A. Kulyukin. Robot-assisted shopping for the blind: issues in spatial cognition and product selection. *Intelligent Service Robotics*, 1(3):237–251, 2008.

[34] D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita. Laser-based tracking of human position and orientation using parametric shape modeling. *Advanced Robotics*, 23(4):405–428, 2009.

[35] D. F. Glas, S. Satake, T. Kanda, and N. Hagita. An interaction design framework for social robots. In *Proceedings of Robotics: Science and Systems*, 2011.

[36] D. F. Glas, S. Satake, T. Kanda, and N. Hagita. An interaction design framework for social robots. In *Robotics: Science and Systems*, volume 7, page 89, 2012.

[37] D. F. Glas, K. Wada, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita. Personal greetings: Personalizing robot utterances based on novelty of observed behavior. *International Journal of Social Robotics*, pages 1–18, 2016.

[38] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1338–1343. IEEE, 2005.

[39] A. Green, H. Huttenrauch, and K. S. Eklundh. Applying the wizard-of-oz framework to cooperative service discovery and configuration. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 575–580, 2004.

[40] H.-M. Gross, H. Boehme, C. Schroeter, S. Müller, A. Koenig, E. Einhorn, C. Martin, M. Merten, and A. Bley. Toomas: interactive shopping guide robots in everyday use-final implementation and experiences from long-term field trials. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2005–2012. IEEE, 2009.

[41] M. Groth and A. Grandey. From bad to worse: Negative exchange spirals in employee-customer service interactions. *Organizational Psychology Review*, 2(3):208–233, 2012.

[42] E. Gu and N. I. Badler. Visual attention and eye gaze during multiparty conversations with distractions. In *International Workshop on Intelligent Virtual Agents*, pages 193–204. Springer, 2006.

[43] L. Guéguen. *Segmentation by Maximal Predictive Partitioning According to Composition Biases*, volume 2066 of *Lecture Notes in Computer Science*, chapter 4, pages 32–44. Springer Berlin Heidelberg, 2001.

[44] E. Hall. Handbook for proxemic research. *Anthropology News*, 36(2):40–40, 1995.

[45] K. Harada, S. Kajita, K. Kaneko, and H. Hirukawa. Dynamics and balance of a humanoid robot during manipulation tasks. *IEEE Transactions on Robotics*, 22(3):568–575, 2006.

[46] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[47] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita. Pointing to space: modeling of deictic interaction referring to regions. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 301–308. IEEE, 2010.

[48] S. L. Haywood, M. J. Pickering, and H. P. Branigan. Do speakers avoid ambiguities during dialogue? *Psychological Science*, 16(5):362–366, 2005.

[49] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[50] R. M. Holladay, A. D. Dragan, and S. S. Srinivasa. Legible robot pointing. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 217–223. IEEE, 2014.

[51] C.-M. Huang, M. Cakmak, and B. Mutlu. Adaptive coordination strategies for human-robot handovers. In *Proceedings of Robotics: Science and Systems*, 2015.

[52] C. Hulme, S. Maughan, and G. D. Brown. Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of memory and language*, 30(6):685–701, 1991.

[53] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[54] S. Ishii. Characteristics of japanese nonverbal communicative behavior. *Communication (Journal of the Communication Association of the Pacific)*, 2:43–60, 1973.

[55] P. H. Kahn, N. G. Freier, T. Kanda, H. Ishiguro, J. H. Ruckert, R. L. Severson, and S. K. Kane. Design patterns for sociality in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 97–104, kahn08, 2008. ACM.

[56] T. Kanda, D. F. Glas, M. Shiomi, and N. Hagita. Abstracting people's trajectories for social robots to proactively approach customers. *Robotics, IEEE Transactions on*, 25(6):1382–1396, 2009.

[57] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro. A two-month field trial in an elementary school for long-term human-robot interaction. *Robotics, IEEE Transactions on*, 23(5):962–971, 2007.

[58] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita. An affective guide robot in a shopping mall. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 173–180. ACM, 2009.

[59] K. Kaneko, K. Harada, and F. Kanehiro. Development of multi-fingered hand for life-size humanoid robots. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 913–920. IEEE, 2007.

[60] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda. Ximera: A new tts from atr based on corpus-based technologies. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[61] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg. Cloud-based robot grasping with the google object recognition engine. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 4263–4270. IEEE, 2013.

[62] S. Keizer, M. Ellen Foster, Z. Wang, and O. Lemon. Machine learning for social multiparty human–robot interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(3):14, 2014.

[63] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, 2004.

[64] T. Kitade, S. Satake, T. Kanda, and M. Imai. Understanding suitable locations for waiting. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 57–64. IEEE Press, 2013.

[65] W. B. Knox, S. Spaulding, and C. Breazeal. Learning from the wizard: Programming social interaction through teleoperated demonstrations. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pages 1309–1310. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[66] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. *Deixis: How to determine demonstrated objects using a pointing cone*, pages 300–311. Springer, 2006.

[67] P. Kühnlein and J. Stegmann. Empirical issues in deictic gesture: referring to objects in simple identification tasks. *Report 2003/3, SFB*, 360, 2003.

[68] P. Kumari and A. Vaish. Information-theoretic measures on intrinsic mode function for the individual identification using eeg sensors. *IEEE Sensors Journal*, 15(9):4950–4960, 2015.

[69] K. Kumatani, J. McDonough, and B. Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *Signal Processing Magazine, IEEE*, 29(6):127–140, 2012.

[70] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[71] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

[72] I. Leite, A. Pereira, A. Funkhouser, B. Li, and J. F. Lehman. Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 13–20. ACM, 2016.

[73] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro. Data-driven hri: Learning social behaviors by example from human-human interaction. *IEEE Transactions on Robotics*, 32(4):988–1008, 2016.

[74] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita. It's not polite to point: generating socially-appropriate deictic behaviors towards people. In *8th ACM/IEEE International Conference on Human-Robot Interaction*, pages 267–274. IEEE Press, 2013.

[75] R. Livshits. *Does Ishiguro Dream of Electric Sheep? Androids as a distinctive emergent phenomenon in Japan*. PhD thesis, University of Victoria, 2016.

[76] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.

[77] G. Magyar, M. Virčíková, and P. Sinčák. Increasing the robotâĂŹs level of autonomy in social human-robot interaction through interactive reinforcement learning. In *New Friends 2016: 2nd International Conference on Social Robots in Therapy and Education*, 2016.

[78] J. D. Matarazzo and A. N. Wiens. *The interview: Research on its anatomy and structure*. Transaction Publishers, 1972.

[79] R. Meena, G. Skantze, and J. Gustafson. A data-driven approach to understanding spoken route directions in human-robot dialogue. In *INTERSPEECH*, 2012.

[80] F. Michaud and M. J. Matarić. Learning from history for behavior-based mobile robots in non-stationary conditions. *Machine Learning*, 31(1-3):141–167, 1998.

[81] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

[82] N. Mirnig, A. Weiss, and M. Tscheligi. A communication structure for human-robot itinerary requests. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 205–206. ACM, 2011.

[83] Y. Mohammad and T. Nishdia. Self-initiated imitation learning. discovering what to imitate. In *Control, Automation and Systems (ICCAS), 2012 12th International Conference on*, pages 726–732. IEEE, 2012.

[84] C. Müller. Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256, 2004.

[85] J. Mumm and B. Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 331–338, mumm11, 2011.

[86] 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 *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68, 1514109, 2009. ACM.

[87] Y. Nagai. Learning to comprehend deictic gestures in robots and human infants. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 217–222. IEEE, 2005.

[88] M. Nicolescu and M. J. Mataric. Task learning through imitation and human-robot interaction. *Models and mechanisms of imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions*, 2005.

[89] M. Nieuwenhuisen and S. Behnke. Human-like interaction skills for the mobile communication robot robotinho. *International Journal of Social Robotics*, 5(4):549–561, 2013.

[90] I. Nourbakhsh, C. Kunz, and T. Willeke. The mobot museum robot installations: a five year experiment. In *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 4, pages 3636–3641 vol.3, 2003.

[91] K. Ogawa, K. Taura, and H. Ishiguro. Possibilities of androids as poetry-reciting agent. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 565–570. IEEE, 2012.

[92] M. Ogino, H. Toichi, Y. Yoshikawa, and M. Asada. Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping. *Robotics and Autonomous Systems*, 54(5):414–418, 2006.

[93] J. Orkin and D. Roy. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, 3(1):39–60, 2007.

[94] J. Orkin and D. Roy. Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 385–392. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[95] J. Orkin and D. K. Roy. Understanding speech in interactive narratives with crowdsourced data. In *Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE)*, 2012.

[96] J. D. Orkin. *Collective artificial intelligence: simulated role-playing from crowd-sourced data.* PhD thesis, Massachusetts Institute of Technology, 2013.

[97] A. K. Pandey, M. Ali, and R. Alami. Towards a task-aware proactive sociable robot based on multi-state perspective-taking. *International Journal of Social Robotics*, 5(2):215–236, 2013.

[98] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, 2007.

[99] J. Peltason and B. Wrede. Modeling human-robot interaction based on generic interaction patterns. In *AAAI Fall Symposium: Dialog with Robots*, pages 80–85, 2010.

[100] S. Pillai and J. Leonard. Monocular slam supported object recognition. *arXiv preprint arXiv:1506.01732*, 2015.

[101] M. Polanyi. *Personal knowledge: Towards a post-critical philosophy.* Psychology Press, 1962.

[102] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[103] M. G. Pratt and L. Doucet. *Ambivalent feelings in organizational relationships*, chapter 11, pages 204–226. SAGE Publications Ltd, 2000.

[104] C. Raffel and D. P. Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.

[105] A. Raux and M. Eskenazi. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 1–10. Association for Computational Linguistics, 2008.

[106] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE, 2010.

[107] D. Richie. *A lateral view: Essays on culture and style in contemporary Japan.* Stone Bridge Press, 1998.

[108] P. E. Rybski, J. Stolarz, K. Yoon, and M. Veloso. Using dialog and human observations to dictate tasks to a learning robot assistant. *Intelligent Service Robotics*, 1(2):159–167, 2008.

[109] A. M. Sabelli, T. Kanda, and N. Hagita. A conversational robot in an elderly care center: An ethnographic study. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 37–44, 2011.

[110] S. Sakurai, E. Sato, and T. Yamaguchi. Recognizing pointing behavior using image processing for human-robot interaction. In *Advanced intelligent mechatronics, 2007 IEEE/ASME international conference on*, pages 1–6. IEEE, 2007.

[111] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[112] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans?: Strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 109–116. ACM, 2009.

[113] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. A robot that approaches pedestrians. *IEEE Trans. Robotics*, 2012.

[114] B. M. Scassellati. *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, Massachusetts Institute of Technology, 2001.

[115] A. J. Schmid, O. Weede, and H. Worn. Proactive robot task selection given a human intention estimate. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 726–731. IEEE, 2007.

[116] J. Schmidt, N. Hofemann, A. Haasch, J. Fritsch, and G. Sagerer. Interacting with a mobile robot: Evaluating gestural object references. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3804–3809. IEEE, 2008.

[117] O. C. Schrempf, U. D. Hanebeck, A. J. Schmid, and H. Worn. A novel approach to proactive human-robot cooperation. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 555–560. IEEE, 2005.

[118] A. C. Schultz and J. G. Trafton. Towards collaboration with robots in shared space: spatial perspective and frames of reference. *Interactions*, 12(2):22–24, 2005.

[119] V. B. Semwal, S. A. Katiyar, R. Chakraborty, and G. Nandi. Biologically-inspired push recovery capable bipedal locomotion modeling through hybrid automata. *Robotics and Autonomous Systems*, 70:181–190, 2015.

[120] C. Shi, T. Kanda, M. Shimada, F. Yamaoka, H. Ishiguro, and N. Hagita. Easy development of communicative behaviors in social robots. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 5302–5309, shi10, 2010.

[121] M. Shiomi, T. Kanda, D. F. Glas, S. Satake, H. Ishiguro, and N. Hagita. A network robot system for cooperative guide service in a shopping mall. *Journal of the Robotics Society of Japan*, 29(6):544–553, 2011.

[122] M. Shiomi, F. Zanlungo, K. Hayashi, and T. Kanda. *A framework with a pedestrian simulator for deploying robots into a real environment*, pages 185–196. Springer, 2012.

[123] B. Shneiderman. The limits of speech recognition. *Communications of the ACM*, 43(9):63–65, 2000.

[124] G. Skantze and S. Al Moubayed. Iristk: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 69–76. ACM, 2012.

[125] A. Sorokin, D. Berenson, S. S. Srinivasa, and M. Hebert. People helping robots helping people: Crowdsourcing for grasping novel objects. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2117–2122. IEEE, 2010.

[126] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Krose. Biron, where are you? enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 934–940. IEEE, 2006.

[127] S. Sternberg. High-speed scanning in human memory. *Science*, 153(3736):652–654, 1966.

[128] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. Natural deictic communication with humanoid robots. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1441–1448. IEEE, 2007.

[129] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, and Y. Anzai. Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model. *Connection Science*, 18(4):379–402, 2006.

[130] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *28th International Conference on Neural Information Processing Systems*, pages 2440–2448, 2969512, 2015. MIT Press.

[131] L. Takayama, E. Marder-Eppstein, H. Harris, and J. M. Beer. Assisted driving of a mobile remote presence system: System design and controlled user evaluation. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1883–1889, takayama11, 2011.

[132] D. Tannen. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press, 1989.

[133] A. L. Thomaz and C. Breazeal. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI*, volume 6, pages 1000–1005, 2006.

[134] A. L. Thomaz and C. Chao. Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine*, 32(4):53–63, 2011.

[135] R. Toris, D. Kent, and S. Chernova. The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing. *Journal of Human-Robot Interaction*, 3(2):25–49, 2014.

[136] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[137] D. L. Trout and H. M. Rosenfeld. The effect of postural lean and body congruence on the judgment of psychotherapeutic rapport. *Journal of Nonverbal Behavior*, 4(3):176–190, 1980.

[138] A. Vaish and P. Kumari. A comparative study on machine learning algorithms in emotion state recognition using ecg. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 1467–1476. Springer, 2014.

[139] I. Van Der Sluis and E. Krahmer. Generating referring expressions in a multimodal context: An empirically oriented approach. *Language and Computers*, 37(1):158–176, 2001.

[140] G. Viejo, M. Khamassi, A. Brovelli, and B. Girard. Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in behavioral neuroscience*, 9, 2015.

[141] Q. Wang, P. Cavanagh, and M. Green. Familiarity and pop-out in visual search. *Perception and Psychophysics*, 56(5):495–500, 1994.

[142] F. Wild, C. Stahl, G. Stermsek, and G. Neumann. Parameters driving effectiveness of automated essay scoring with lsa. In *Proceedings of the 9th CAA Conference*, pages 485–494, 2005.

[143] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin and review*, 1(2):202–238, 1994.

[144] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan. How fast can you change your mind? the speed of top-down guidance in visual search. *Vision research*, 44(12):1411–1426, 2004.

[145] F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita. How close?: model of proximity control for information-presenting robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 137–144. ACM, 2008.

[146] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka. Precision timing in human-robot interaction: coordination of head movement and utterance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 131–140, 1357077, 2008. ACM.

[147] W. Yan and D. A. Forsyth. Learning the behavior of users in a public space through video tracking. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 370–377. IEEE, 2005.

[148] J. E. Young, T. Igarashi, E. Sharlin, D. Sakamoto, and J. Allen. Design and evaluation techniques for authoring interactive and stylistic behaviors. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(4):23, 2014.

[149] J. E. Young, E. Sharlin, and T. Igarashi. Teaching robots style: designing and evaluating style-by-demonstration for interactive robotic locomotion. *Human-Computer Interaction*, 28(5):379–416, 2013.

[150] K. Zheng, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita. Designing and implementing a human-robot team for social interactions. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2012.

# List of Publications

## International Journal Publications

1. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, and Hiroshi Ishiguro, "Data-Driven HRI: Learning Social Behaviors by Example from Human-Human Interaction", *IEEE Transactions on Robotics*, Vol. 32, No. 4, pp. 988-1008, 2016.

2. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, and Hiroshi Ishiguro, "Learning proactive behavior for interactive social robots", *Autonomous Robots, Special Issue on Learning for Human-Robot Collaboration*.
   (submitted for review)

3. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita, "A Model for Generating Socially-Appropriate Deictic Behaviors Towards People", *International Journal of Social Robotics*, Vol. 9, No. 1, pp. 33-49, 2017.

## International Conference Publications

1. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita, "It's Not Polite to Point – Generating Socially Appropriate Deictic Behaviors Towards People", *ACM/IEEE 8th Annual Conference on Human-Robot Interaction (HRI 2013)*, pp. 267-274, Tokyo, Japan.
   **Nominated for Best Paper**

2. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita, "How to Train Your Robot - Teaching service robots to reproduce human social behavior", *23rd International Symposium on Robot and Human Interactive Communication (RO-MAN 2014)*, pp. 961-968, Edinburgh, Scotland.
   **Nominated for KROS / RSJ Distinguished Interdisciplinary Research Prize**

## Workshop and Symposium Papers

1. Dylan F. Glas, Phoebe Liu, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita, "To Point or Not to Point: How Cloud Intelligence Can Help Make Robots

More Polite”, *9th IEICE Technical Committee on Cloud Network Robotics*, Keio University, Yokohama, Japan, June 2013.

2. Dylan F. Glas, Phoebe Liu, Takayuki Kanda, and Hiroshi Ishiguro, “Can a social robot train itself just by observing human interactions?”, Workshop on Machine Learning for Social Robotics, *IEEE International Conference on Robotics and Automation (ICRA 2015)*, Seattle, WA, USA, May 2015.
3. Phoebe Liu, Dylan F. Glas, Takayuki Kanda, and Hiroshi Ishiguro, “Learning Interactive Behavior for Service Robots – the Challenge of Mixed-Initiative Interaction”, Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics (BAILAR), *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016)*, New York, NY, USA, August 2016.
4. Dylan F. Glas, Malcolm Doering, Phoebe Liu, Takayuki Kanda, and Hiroshi Ishiguro, “Robot’s Delight - A lyrical exposition on learning by imitation from human-human interaction”, *2017 Conference on Human-Robot Interaction, Video Session (HRI 2017)*, Vienna, Austria, March 2017.

# Patent

1. Phoebe Liu, Dylan F. Glas, and Takayuki Kanda, “A method for learning human-robot interaction logic for locomotion and dialogue by imitation from example human-human interactions”, Applied on 2016/12/8 (Patent No.2016-238910)