**Abstract**

# Robots for social skills therapy in autism: evidence and designs toward clinical utility

**Elizabeth Seon-wha Kim**
2013

Given evidence that some individuals with autism spectrum disorders (ASD) have greater interest or facility in interacting with mechanical than social elements of everyday life, there has been much interest in using robots as facilitators, scaffolds, or catalysts for social behavior within interventions. This dissertation presents evidence toward the clinical utility of interaction with robots for communication and social skills therapies for children with ASD. Specifically, we present novel, group-based, well-controlled observations of social behaviors produced by populations with ASD and with typical development (TD), during brief interactions with social robots. Importantly, we present evidence that a robot can elicit greater social interaction with an interventionist than can an asocial engaging technology, or another adult, suggesting that the appeal of a technology cannot alone mediate or elicit social behavior in children with ASD; rather, sociality must be entwined with interaction with the technology. In addition, we present evidence validating novel technologies and interaction designs that support the application of social robots to the specific domain of speech prosody therapy. Finally, this dissertation suggests systematic design guidelines promoting clinically effective collaborations between human-robot interaction scientists and clinical researchers and providers who support individuals with ASD.

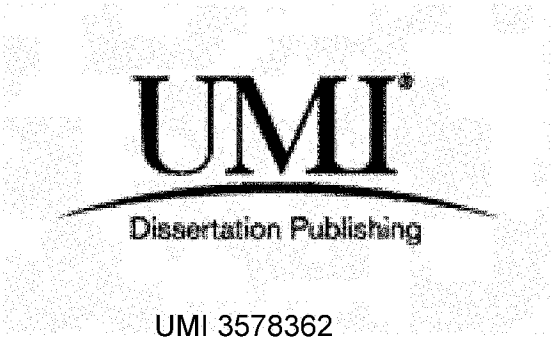
# Robots for social skills therapy in autism: evidence and designs toward clinical utility

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Elizabeth Seon-wha Kim

Advisor: Brian Scassellati

Readers: Cynthia Breazeal (MIT)
Holly Rushmeier
Brian Scassellati
Steven Zucker







ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as dissertation for the degree of Doctor of Philosophy.

_____

(typed name) Principal Advisor

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as dissertation for the degree of Doctor of Philosophy.

_____

(typed name)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as dissertation for the degree of Doctor of Philosophy.

_____

(typed name)

Approved for the University Committee on Graduate Studies

_____

# Dedication

May the merit of this work be apportioned to all beings. May all beings be free from suffering.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgments

Thanks to Mom, Dad, Catho, Anne, 외할머님, 이모, and 우리가족, for a lifetime of love and understanding; I've always known you've got my back. I feel lucky to have you. Thanks to Mary, Rich, Karla, and Ben Lawlor Adelt, and Carolyn, Bernard Shaak, and Jamie, Kim, Leslie, and your families, for teaching me again and again how to give and receive love.

I feel profound and humble gratitude to Scaz, who made all of the work in this dissertation possible. All my learning at Yale was made possible by you. Thank you.

I am also grateful to Steve Zucker, Holly Rushmeier, and Cynthia Breazeal for their thoughtful responses and suggestions, guidance, and support during my years at Yale. And for reading! And to Dana Angluin and Fred Shic for their energetic and morally supportive mentorship. I owe tremendous gratitude to Rhea Paul, Fred Volkmar, Ami Klin, and Kathy Koenig, our clinical collaborators, for deepening my understanding of autism, and for making our clinical experiments possible.

My deep thanks, too, to the Social Robotics Lab, including Justin, Chris, Ganghua, Henny, Brad, Alex, Aditi, Sam, Danny, Miriam, and Francesca; the Department of Computer Science, the Laboratory for Developmental Communication Disorders (LDCD), the Technology and Innovation Lab (TIL), and the Child Study Center.

I have continuously learned from the thoughtful perspectives of my colleagues Dan Leyzberg, Kevin Gold, Liz Simmons, Megan Lyons, Lauren Berkovits, Becca Loomis, Julie Wolf, Cindy Bethel, Emily Bernier, Hilary Barr, Erica Newland, Justin Hart, Wilma Bainbridge, Elaine Short, Taylor Brown, Kate Tsui, Jonathan Tirrell, Michelle Levine, Maysa Akbar, Stephanie Maynard, Tammy Babitz, Erin MacDonnell, and Heidi Tsapelas. Thank you for the fun you have brought to this work, for your helpful ideas, the conscientious work and many hours you have devoted to our studies. You have made invaluable contributions to this dissertation.

Thanks and love to my chosen family Talya, Victoria, Nikhil, Dan, Ulli, Ana, Melanie, Liz, Nic, Carolee, Ksenia, Mariana, Anastasia, Jenny L., and Lucy. You've colored my hours and ushered in the seasons. And I'd fear to think of my days without David, Lynn, Zoe, Edie, Jenny A., Gabi, Amir, Annie, Jordan, Junlin, Derrick, Juhi, the FFF, Sachiko, Antonis, Argyro, Raj, Evan, Michelle, Jaime, Sierra, Yan and Shelby. Thank you.

My humble gratitude to Sandy Wells, Jud Brewer, Angela 이선생님, Pat Ryan, Lois G., Laverne M., Kathryn B., again, to Carolyn, Dana, Fred, and Scaz, and to the memories of Patricia Zander, Doris Yeingst, and Norm Liden. Well before I knew how I might want to move through this life, you generously opened my imagination and gently fed me clues. Thank you for feeding me as many as I could stomach at any moment.

# Chapter 1

# Introduction

Difficulties with social interaction and communication are among the core deficits in autism spectrum disorders (ASD; American Psychiatric Association, 2000). This dissertation establishes the feasibility of using social robots for communication and social skills therapy for children with ASD, by demonstrating spontaneous social engagement with robots, by sample populations of children with ASD, in a clinical setting.

## 1.1 Autism

Autism spectrum disorders are complex disorders of brain development, characterized by difficulties in social interaction, communication, and repetitive behaviors (American Psychiatric Association, 1994, 2000).[1]

Among individuals with ASD, deficits manifest heterogeneously (Volkmar & Klin, 2005). Social deficits are among the primary characteristics associated with autism spectrum disorders (American Psychiatric Association, 2000; Carter, Davis, Klin, & Volkmar, 2005; Joseph & Tager-Flusberg, 1997; Kanner, 1943; Mundy, Sigman, & Dawson, 1989). As many

[1] In May 2013, the American Psychiatric Association released the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-5), changing the definition of autism by eliminating subclassifications such as Asperger's Syndrome, and by collapsing the communication and social interaction categories of behavior

as half of all affected children do not develop functional speech (Tager-Flusberg, Paul, & Lord, 2005). Even for individuals who develop strong mechanical language skills, difficulties persist in conversational interactions (Mesibov, 1992; Paul, 2008; Tager-Flusberg et al., 2005). For example, common deficits in conversational skills include difficulty managing turn-taking and topics of discourse, using inappropriate style of speech to fit conversation partners and settings, and trouble inferring what information is relevant or interesting to others (Paul, 2008). Production and perception of affective expressions, as well as eye contact and other nonverbal attentional cues, can also be inappropriate, unconventional or deficient in individuals with autism (Mundy, Sigman, Ungerer, & Sherman, 1986). Restricted and repetitive behaviors can also be extremely disruptive in some individuals (Turner-Brown, Lam, Holtzclaw, Dichter, & Bodfish, 2011).

Deficits associated with ASD impact individuals' abilities to function independently in social, occupational and other important areas of life (American Psychiatric Association, 2000). Many individuals need high levels of support and care throughout their lives, and early intervention is considered critical (American Psychiatric Association, 2000; Klin, Lang, Cicchetti, & Volkmar, 2000; Mullen, 1995; Sparrow, Cicchetti, & Balla, 2005; Volkmar, Lord, Bailey, Schultz, & Klin, 2004).

### 1.1.1 Social skills and communication therapies

Diverse intervention approaches seek to ameliorate functional difficulties brought on by repetitive and restricted behaviors, and also seek to improve social and communication skills in children with ASD (reviewed, for example, in Paul, 2008; and Volkmar et al., 2004).

With the goal of improving individuals' ability to function, communication and social skills therapies range correspondingly with the heterogeneity impairments and abilities

exhibited by individuals with ASD. Interventions tend to vary with respect to (1) the behaviors targeted; (2) the extent to which targeted behaviors are naturalistically elicited (e.g., by tempting a child to ask for help by keeping a toy out of reach), or are explicitly elicited through instruction (i.e., through highly structured repetition); and (3) the setting in which training or reinforcement takes place—anywhere from highly controlled clinical settings to naturalistic environments such as the child's home or classroom.

For pre-verbal children, didactic therapies are highly structured, passive for the children and initiated by the clinician, and rely on external reinforcers like food. More naturalistic, contemporary adaptive behavioral analysis techniques, like milieu therapy, reward children by giving them access to intrinsic reinforcers when they initiate requests to interact with these reinforcers.

Among verbal individuals with ASD, their speech prosody (that is, their tone of voice, or the "melody" and rhythm of their speech) often sounds odd (Diehl & Paul, 2011; Mesibov, 1992; Paul, Augustyn, Klin, & Volkmar, 2005; Shriberg et al., 2001), for example atypically flat or sing-song (Nadig & Shaw, 2012; Tager-Flusberg & Caronna, 2007). Atypical prosody production is a primary marker by which neurotypical individuals single out those with ASD as different (Mesibov, 1992; Van Bourgondien & Woods, 1992).

Problems with pragmatic language use are pervasive, including "use of irrelevant detail; inappropriate topic shifts; topic preoccupation/perseveration; unresponsiveness to partner cues; lack of reciprocal exchange; inadequate clarification; vague references; scripted, stereotyped discourse; excessively formal style (for speakers with Asperger syndrome only)" (Paul, Orlovski, Marcinko, & Volkmar, 2009).

Interventions for children with ASD target foundational social skills like joint attention, eye contact, and cooperative play and problem-solving; verbal production and understanding, and pragmatic language use and appropriate prosody production; and the reduction of maladaptive behaviors, including repetitive and repeated behaviors and circumscribed interests.

Many interventions are based on an applied behavior analysis paradigm, in which the provider or interventionist entices children to practice targeted behaviors, and then reinforces the performed behaviors (Paul, 2008). Reinforcement can be manifested in the delivery of preferred edibles, providing access to preferred toys or objects, or by allowing the child to engage in a preferred activity (for instance, watching a favorite television program).

For even slight improvements in any targeted behavior, therapy tends to require regular, repeated lessons and practice, over the course of months or years. For instance, Handleman (1979) taught four boys (with chronological ages of 6 to 7 years) "described as 'autistic' by various agencies" who "could produce and imitate sounds and words but did not use language as a spontaneous interpersonal communication skill" to respond to simple questions (e.g., *--What do you smell with? –Nose*, or, *--What do you sit on? --Chair.*). Participants trained three times a day, on five specific questions, four times each, with their mother or a tutor. The tutor or mother would ask one of the five questions, and if the child responded correctly within 10 s, would reinforce the response with praise and food; if the child responded incorrectly, the tutor or mother would say, "no," and then speak the correct response. Treatment for each child continued until that child answered correctly 75% of the time on two training sessions in a row. Treatment took between 100 and 1740 trials (25-435 sessions over 9-145 days). A survey (Goldstein, 2002) of a wide variety of communication

treatment studies revealed broadly ranging treatment durations, most ranging over 24-36 sessions, at intervals ranging from daily to weekly, with one intervention requiring weekly home visits for 6 months, followed by bi-monthly home visits for an additional 12 months (Howlin & Rutter, 1989). It is notoriously difficult for children with ASD to generalize the skills they develop through intervention to environments, social partners, and scenarios different from those on which they specifically train (Prelock, Paul, & Allen, 2011).

## 1.2 Robot applications for autism

### 1.2.1 Special interests in technology motivate HRI applications to autism intervention

Evidence suggests that in children with ASD, more so than in children with typical development (TD), circumscribed interests are frequently focused on nonsocial objects, activities, and phenomena (Turner-Brown et al., 2011). Such nonsocial objects include devices, mechanical items, and physical systems; telephone pole insulators, lawn sprinklers, cranes, trains, and Legos are common objects of circumscribed interests (South, Ozonoff, & McMahon, 2005; Turner-Brown et al., 2011). One parent observed:

> Our son Tom, who has a diagnosis of autism, developed an intense interest in trains at around age 2. This intense level of interest caused some friction in our family. For example, whenever we visited a local zoo, he was only interested in riding on the zoo trains. At first, this "obsession" seemed typical until we found ourselves planning all of our family vacations around opportunities to see or ride on trains. At the age of 3, he had only 14 vocabulary words and "train" was one of them. Any conversation we held revolved around trains, and this interfered with his relationships with peers, and even his sister. He wouldn't play with or talk about anything else.

Eye-tracking studies also show a perceptual bias for looking at physical objects, rather than social actions or people, in ASD more so than in TD (Sasson, Elison, Turner-Brown, Dichter, & Bodfish, 2011).

Before nonsocial interests had been established as common among children with ASD, anecdotal evidence of circumscribed interest in devices and mechanical objects motivated parents and researchers to explore robots as potential supplements to therapy and education for children with ASD. For example, in 2000, Sun Microsystem sponsored a contest to hack Tiger Electronics' Furby toy, after the mother of a child with ASD contacted them, saying that her four-year-old son's interactions with the toy taught him to speak about himself and expanded his vocabulary by at least 6 words (Fleishman, 2000; Kahney, 1999).

### 1.2.2 Scaffolding theory

The goal of HRI research for children with autism is to use robots to improve therapies, most often those targeting communication and social skills. There tend to be two theories for how robots can contribute to therapy. The first is to view robots as customizable scaffolds (as in Robins, Dautenhahn, te Boekhorst, & Billard, 2005): Robots can serve as social practice partners for children with ASD. Therapists can scale down the complexity or diversity of the robot's behavior to facilitate practice for children who have high social anxiety or who tend to be overwhelmed by more sophisticated or unpredictable interactions. As a child improves in skills, the interventionist can gradually increase the complexity of the interactions, and begin to incorporate practice with people, to generalize the child's gains to ecologically functional (that is, real world) interactions. This thrust within HRI for autism is motivated by the pedagogical theory scaffolding, which suggests that educators should structure lessons to support (that is, provide scaffolding) for underdeveloped skills on which

the targeted skill depends. This allows the student to focus on the targeted skill at hand (Vygotsky, 1978; Wood & Middleton, 1975).

### 1.2.3 Embedded reinforcers: a novel theory

This dissertation suggests a second, novel theory of robots' utility in autism intervention: as reinforcers that catalyze social interaction with other people. There is evidence that the use of child-preferred, or *intrinsic*, reinforcers leads to improvements in social engagement (reviewed in Paul, 2008). Furthermore, *embedding* social interaction into the delivery of a child's preferred reinforcer (for example, singing a child's favorite song, rather than playing a video recording of the song) elicits greater social initiation, increased non-verbal (bodily) orientation to face an interaction partner, and more positive affect (L. K. Koegel, R. L. Koegel, Harrower, & Carter, 1999; R. L. Koegel, Dyer, & Bell, 1987; R. L. Koegel, Vernon, & L.K. Koegel, 2009).

The long-term aim of our research is to evaluate and fulfill the potential of social robots as *embedded reinforcers*, which elicit and reward social behavior in interventions for children with autism. Although there is ample evidence that children with ASD (as well as children and adults with typical development) will engage socially with robots, our long-term aim focuses on the ways social robots may support therapies that improve social interaction with other people (Duquette, Michaud, & Mercier, 2008; Feil-Seifer & Matarić, 2009; Kozima, Michalowski, & Nakagawa, 2009; Robins et al., 2005; Stanton, Kahn Jr., Severson, Ruckert, & Gill, 2008). *Social robots* are robots that are designed to evoke social behaviors and perceptions in the people with whom they interact. There is promising case study evidence that robots, both socially evocative and not (discussed below), can elicit social engagement

between children and the robots themselves, and can mediate social engagement between children and adults. Whereas Koegel et al. (2009) have shown that embedding social interaction, *within the delivery* of preferred reinforcers, increases production of target behaviors, we are interested in further embedding social interaction, *within the reinforcer or motivator itself.* It is our eventual hope that social robots can translate children's interest in novel technologies into increased motivation for participating in social interactions and social partnerships with people. Such an approach, if effective, could provide new methods to facilitate and augment behavioral, communicative, and social therapies that improve interactions between individuals with ASD and with other people (Scassellati, 1996).

Both scaffolding and embedded reinforcement theories suggest that some children with ASD may find robots especially motivating. Over the last 10 years or so, numerous studies have observed that children with autism seem to enjoy interacting with robots which imitate their movements (Dautenhahn & Billard, 2002); robots that bounce to express positive affect in response to children's gaze, or turn their heads back and forth to make eye contact with a child and a caregiver (Kozima, Nakagawa, & Yasuda, 2005); and robots that blow bubbles when a child pushes a button on the robot (Feil-Seifer & Matarić, 2009).

### 1.2.4 A brief history of HRI explorations in autism intervention

Socially assistive robotics, a recently emergent field, has, over the last decade or so, investigated social robots as assistive tools for elderly, or physically or cognitively impaired individuals (Scassellati, Admoni, & Matarić, 2012; Tapus, Matarić, & Scassellati, 2007), and as supportive tools for social and communication skills therapy in children with ASD (Duquette et al., 2008; Feil-Seifer & Matarić, 2009; Kozima et al., 2009, 2005; Robins et al.,

2005; Scassellati, 2005; Stanton et al., 2008; Werry & Dautenhahn, 1999). Multiple studies have shown that children with ASD will interact with robots using social behaviors, e.g., by directing speech to the robot (Duquette et al., 2008; Feil-Seifer & Matarić, 2009; Kozima et al., 2009; Robins et al., 2005; Stanton et al., 2008). Several of these studies have further demonstrated that children with ASD will interact with a parent, caregiver, or another human while engaged with a robot partner (Feil-Seifer & Matarić, 2009; Kozima et al., 2009; Robins et al., 2005), for instance, by expressing excitement to a robot, and then turning to express this excitement to a parent (Kozima et al., 2009).

Previous demonstrations of the benefits of robotic interaction on social behaviors were demonstrated over case studies of three or four individual children. However, there have been few demonstrations over larger samples (Diehl, Schmitt, Villano, & Crowell, 2012; Scassellati et al., 2012). It thus had remained an open question whether the beneficial effects of social robots extend more broadly across the autism spectrum. This dissertation presents the first large, group studies ($N$ = 24, Chapter 3; $n$ = 18, Chapter 4) of human-robot interaction in children with ASD. We specifically focus on interventions targeting atypical prosody, a domain of atypical social functioning that has been identified as functionally impactful and an important target for intervention (Paul, Shriberg, et al., 2005).

## 1.3 Making robots useful for interventions for ASD

Here we describe the technologies and investigations needed to support a robotic intervention for speech prosody in children with ASD. Many of the technologies apply broadly to diverse human-robot interactions. Many of the investigations we describe here also apply broadly to HRI, as well as to interventions for ASD.

In general, the engineering scientific discipline of human-robot interaction (HRI) seeks to understand people's behaviors, and to build technologies that can recognize and respond to them, in ways that support the goals of the application in question. HRI for a special population requires models of human behavior specific to that population, and technologies tailored to that population's idiosyncrasies. Autism is characterized by social and communicative impairments, and therefore observations and descriptions of typically developing populations' social behaviors will not generalize wholesale to children with ASD. However, it is more feasible and perhaps ethically responsible to begin HRI investigations with typically developing adult populations because children's participation in experiments requires parental support, greater ethical consideration, and support specific to developmental vulnerabilities, especially for participants with ASD, who themselves and whose families may be under exceptional strain. From studies with typically developing adults, we can develop approximations from which we can prepare to investigate an ASD population's behaviors.

A handful of empirical questions and corresponding automating technologies inform and support our pursuit of clinically helpful robots. We suggest that beneficial interactions require our understanding four aspects of human-robot interaction:

**Question 1: How engaged and motivated are participants?**

**Question 2: How do participants behave during and after human-robot interaction?**

**Question 3: What design elements support interaction?**

**Question 4: How should a robot adapt to maintain a long-term relationship?**

These questions and this dissertation's contribution to each will be discussed in greater detail below.

In this dissertation, we begin in Chapter 2 with an examination of the typically developing adults' use of affectively expressive speech prosody, while teaching robotic learners a toy (i.e., easy) learning task. The study described in Chapter 2 demonstrates that neither prior training nor prior knowledge of the learners' capabilities, (1) adults easily and readily use speech to instruct robotic learners; (2) infuse their speech with intensely affectively expressive speech prosody; (3) provide spoken instruction before the learner completes a trial action, not only in response to completed actions, suggesting that traditional interactive machine learning algorithms, which are designed to learn from passive environmental state changes, do not adequately model the rich, social input sources provided by human instructors; (4) human instructors vary the amount of instructive feedback they provide, depending on the learners' path or history, contradicting an assumption made by classic interactive machine learning algorithms, which model reward functions as path- and history-independent, that is, depending only on the learner's present state within the learning task (E. S. Kim, Leyzberg, Tsui, & Scassellati, 2009). This study demonstrates the feasibility of eliciting natural social behaviors from typically developing adults, within human-robot interaction, suggesting the possibility of eliciting natural social behaviors from children with ASD. Our findings also suggest the need to reconsider the application of classic machine learning algorithms to human social behavioral input.

Chapter 3 describes the most impactful and innovative contributions of this dissertation: (1) the unique ability of social robots to facilitate spoken interaction between children with high-functioning ASD[2] (HFA) and an adult clinician or educator, and (2) evidence

[2] High functioning with ASD is typically defined as full-scale IQ above 70. More detailed characterization of participants is provided in Chapter 3.

supporting a theory that social robots uniquely embed sociality into reinforcing interactions. The chapter details a study in which children with HFA briefly interacted with an adult, robot, and computer game (E. S. Kim et al., 2013). All interactions were semi-structured and guided by a single confederate, who was constant for all participants and interactions, and who fulfilled the role of an *interventionist* (that is, a therapist or educator). In between interactions, participants completed brief interview and play sessions, with each participant interacting with a constant interviewer. Adult and robot interactions were designed to emulate communication or social skills intervention sessions, and were highly parallel in structure to each other, to establish a contrast between an optimally capable social interaction partner (the adult) and a social robot. The computer game was designed to be similar to the robot interaction with respect to spatial and physical challenges, but included limited structured social interaction. Our intent was to contrast the robot's and computer game's asocial and technological appeal. In short, the social robot was designed to present an intersection of the computer game's mechanical and technological appeal, with the adult interaction partner's social engagement. Participants spoke more while interacting with the robot than while with the adult or computer game interaction partners. They directed the same amount of speech to the robot and adult (and little to the computer game). Most interestingly, participants directed more speech to the confederate during the robot interaction than during the other two. This finding indicates a social robot's unique ability to motivate and mediate social interaction between children with HFA and other people. It also supports a view of social robots as socially embedding reinforcers.

Chapter 4 further establishes the feasibility of using social robots to elicit and reinforce engagement in tasks like those used in communication therapies. The study described in this

chapter specifically targets production of affective expression in speech prosody, in school-aged children with HFA (E. S. Kim, Paul, Shic, & Scassellati, 2012). This study also provides opportunities to compare responses with those from an age- and IQ-matched sample of children with TD. The study shows that (1) both children with ASD and those with TD enjoy and engage with a robot-directed, repetitive prosody production task; (2) children with ASD, more so than those with TD, increase their social engagement with a clinician after interacting with the robot; (3) that children with ASD appear to more motivated to interact with the robot than those with TD; and (4) that motivation for post-intervention human and robot interaction correlate positively with verbal ability and enjoyment of the therapeutic robot interaction, and inversely with severity of autistic symptoms. In sum, Chapter 4 further establishes the feasibility of using a robot in a behavioral communication intervention, indicates greater reinforcement in children with ASD than in children with TD, and begins to explore neuropsychological characteristics of participants who might especially benefit from social robot interaction.

In Chapter 5 we present two novel prototypes of automated systems that recognize the expression of affective (E. S. Kim & Scassellati, 2007) and shared-belief-modifying (E. S. Kim, Gold, & Scassellati, 2008) intents in speech prosody. The first system also uses the output of its affective prosody classifier as feedback for an interactive learning system, which is validated over for a single instructor, on a toy, trivial learning problem. Despite the methodological limitations of the affective classification and learning system, and the modest accuracy of the shared-belief-modifying cues recognition system, these two systems offer proof of concept that social behavioral recognition can be automated and applied to relatively unconstrained contexts. Automatic recognition and response within flexible

interaction conditions will make social robots increasingly affordable and reliable, both of which will be essential to the uptake of robots therapeutic applications.

Chapter 6 describes interdisciplinary collaborative challenges and solutions that we have observed and developed, respectively, over the course of the studies presented in this dissertation (previously published in E. S. Kim et al., 2012). While there are research fields dedicated to the study of interdisciplinary scientific research collaboration, this chapter discusses challenges and strategies unique to robotics and autism research, with the goal of improving human-robot interaction scientific methodology to the influential and highly rigorous scientific standards of evidence-based medical and psychological intervention.

The descriptive, technological, theoretical and methodological contributions offered by these chapters advance our understanding along all four of the research questions we have outlined above and now detail below.

**Question 1: How engaged and motivated are participants?**

Do participants want to interact with the robot? How motivated are they, and how well do they focus, on completing tasks during the interaction? Readers familiar with human-computer interaction methodology will recognize this as a question of acceptability and usability. Clinically, motivation to engage in teaching interactions has been described as fundamentally underlying the learning and improvement of other communication, social, or academic skills (R. L. Koegel, L. K. Koegel, & McNerney, 2001). Finally, motivation to engage in an interaction indicates reinforcement. Reinforcers are used both to entice participants into performing a target behavior, and to reward them after they have performed it. Therefore, the question of the extent to which participants are motivated to engage is triply fundamental to socially assistive robotics applied to autism intervention,

informing us of whether participants will use a particular technology, the general pedagogical potential of a particular interaction with the technology, and the specific reinforcing potential of that interaction.

The question of engagement directly bears on embedded reinforcement theory. Engagement can be measured in terms of compliance and affect.

In this dissertation, we describe three original studies of human-robot interaction, which were the first large group studies to establish that untrained adults with TD (Chapter 2), children with ASD (Chapters 3 and 4), and children with TD (Chapter 4), will readily use speech to interact with a robot. Furthermore, In Chapter 3, we find they speak more to an interventionist while interacting the robot than while with another person. This suggests that in comparison with another adult person, the robot better motivates participants to interact with a clinician, educator, or healthcare provider (henceforth, we will sometimes refer to a person fulfilling the therapeutic role as an *interventionist*).

We also quantitatively established that children with ASD and TD enjoy these interactions, establishing our robot interactions as intrinsically reinforcing (Chapter 4).

**Question 2: How do participants behave during and after human-robot interaction?**

What communicative and social behaviors do participants perform during or following interaction with a robot? Does the quality or frequency improve as a result of interaction with the robot? Do participants perform behaviors that have been identified as important targets for improvement? Can these behaviors form the basis for therapeutic interactions? Are they performing adaptive or maladaptive behaviors? (This question checks that engagement in a robot interaction does not reinforce behaviors that will actually diminish

participants' functioning. For example, if a participant becomes addicted to a video game, this may actually further isolate him socially, rather than improving his daily functioning.)

Exploration of this question provides us with information about which behaviors may be reasonable to target using robot interaction, as well as what kinds of interactions (e.g., spoken, non-verbal) we can expect participants to engage in. On a scaffolding theory view, they tell us what behaviors participants lack, which a robot or interaction design may be able to compensate for, to facilitate the practice of a target behavior. On an embedded reinforcement view, this question asks whether participants will engage socially with the reinforcer, and whether the types of behaviors participants exhibit will be therapeutically beneficial.

In addition, our understanding of participants' behaviors during human-robot interaction can inform and also benefit from the development of automatic behavior-recognition technologies. Characterizing participants' behaviors and motivations, and helpful robotic behaviors and adaptations, allows us to design useful human-robot interactions. These also inform the design of automatic recognition of participants' behaviors, and control of robotic behaviors. The more accurately we can automate the above four recognition and action processes, the more we can automate control of the interaction, reducing some of the enormous human labor involved in autism therapy. In addition, automating perception of human behaviors and control of robotic responses deepens and expands our understanding of these processes.

We suggest that automation will play a crucial role in the successful deployment of socially assistive robots. The labor required to manually operate a robot can overwhelm limited resources available to healthcare providers, educators, and interventionists. In

addition to saving labor, automatic perception of social behavior reciprocally utilizes and deepens our psychological understanding of these behaviors. Data collected through neuropsychological or therapeutic interactions can be used to train automatic perception systems; and in return the statistical models that underlie automatic perception systems can reveal or deepen or understanding of commonalities and differences among presentations of behavior (Campbell, Shic, Macari, Chang, & Chawarska, 2012). Automated robotic perception dovetails with improved psychological models.

In addition, the more we can automate these aspects of interaction, the more flexibly we can deploy them (for instance, for practice outside the clinic and in children's homes), and the more we can relieve therapists', educators', and clinicians' labor over the course of what tend to be time- and labor-intensive interventions.

In Chapter 3 we present observations that children with ASD verbalize more toward an interventionist while with the robot than while with another person, and as much toward the robot as toward a person. This finding supports the feasibility of using a robot in a speech-based clinical interaction. We also found that participants speak more to an interventionist while interacting with a robot than they do while interacting with another person or an engaging computer game. This suggests that social robots may make uniquely helpful reinforcers.

In Chapter 4 we contrasted IQ- and age-matched TD and ASD children's social behaviors toward a robot and their reactions following immediately afterward. We observed that in comparison with the TD group, the ASD group spent longer in free playtime with the robot, following semi-structured interaction and interviews, and increased more the amount of time during which they participated in post-interaction interviews. This confirms

our intuition that behaviors observed in children with TD do not strictly hold for children with ASD.

In Chapter 2 we found adults spontaneously use affectively expressive speech prosody to communicate with a robot. Surprisingly, we also found that they do so in ways that classical machine learning models do not sufficiently use. This suggested that children with ASD behave similarly, motivating us to do the experiment described in Chapter 4. This also provides us with data from which to begin to try to automate perception of robot-directed speech prosody, which may be useful in general, but also in prosody interventions for children with ASD; for even if their prosody is different from typical adults', from automating perception of TD adult robot-directed prosody, we may develop some technologies and models which will may help us eventually automatically evaluate children's production of target behaviors in prosody interventions.

In Chapter 5, we present novel technologies that automate perception of certain aspects of speech prosody. First, we present an automatic recognition system and the learning system that uses the output of recognition as feedback. Although this system is demonstrated only for a single speaker (myself), learning the admittedly trivial task of identifying a goal state within a 9-state finite state machine, nevertheless our demonstration establishes a proof of concept that affective prosody recognition can be performed automatically, in real time and used to drive robotic behavior selection. This supports future research in automatic perception of intervention target behaviors. Our other system, which identifies prosodic shared belief cues suggests, also gives proof of concept of automatic perception of these communicative intents. Our validation study of this system sheds new light on the presence of such cues in infant-directed speech.

**Question 3: What design elements support interaction?**

What aspects of an interaction and a robot's form or behavior elicit, motivate, or reinforce targeted behavior and sustain interaction? What kinds of responses are legible and reinforcing to children with ASD? To what extent, and for whom, are robots better than other reinforcers? This is a design question.

In Chapter 2, we found that robot-directed prosody changes with robot behavior, exposing one way by which a robot's behavior may elicit different kinds of social behavior. Based on pilot observations, we chose to emphasize real-time, Wizard-of-Oz-controlled responses to participants' communications to the robot, as well as a seamless familiarization protocol, allowing us to examine participants' naïve, intuitive selection of ways to communicate to the robot in the face of minimal explanation of the robot's perceptive and productive capabilities, and highlighting typically developing adults' spontaneous anthropomorphization of the robot's behaviors. In addition, we carefully designed transparent (that is, easily legible) communication of the robot Pleo's behaviors (Thomaz & Breazeal, 2006b). This last design consideration was motivated by Chapter 5's study of prosodic input to instruct the humanoid robot Nico how to refine its waving movements, in which we discovered that it was difficult, even for the experimenter, to distinguish between the slight adjustments the robot made from one trial to the next.

Chronologically, the experiment that followed is the one described below in Chapter 4, in which we observed that the Pleo robot's slow locomotion bored participants, who waited sometimes 10 seconds at a time for the robot to approach each next challenge. We also observed that several participants were eager to pet and touch Pleo. In the experiment that chronologically followed (described in Chapter 3) we thus removed any need for the robot's

locomotion. We also found in the study in Chapter 4 that several participants were eager to touch Pleo, and so we added an opportunity for participants to touch the robot in the experiment in Chapter 3, and opportunities for participants to do as they pleased, after the completion of semi-structured protocols in both studies. Just as we emphasized transparent communication of the Pleo robot's intentions and affective states in Chapter 2, we sought to make all behaviors even more obviously comprehensible for use with children with ASD in Chapter 3 and 4. However, knowing that individuals with ASD sometimes struggle in perception of others' mental states, and because of participants' relatively earlier developmental states, we also carefully designed opportunities for the confederate to seamlessly repair misunderstandings of the robot's behaviors. Finally, we were careful in both Chapters 3 and 4 to design errorless interactions, so that participants were always rewarded for engaging, whether or not their answers or actions were correct; that is, they were neither chastised nor penalized for incorrect responses throughout the interactions.

**Question 4: How should a robot adapt to maintain a long-term relationship?**

How should the design of an interaction with a robot, and the robot itself, adapt in order to sustain participants' engagement and maintain reinforcement potency for targeted behaviors, over the course of a therapeutically effective timescale (generally, months of weekly or more frequent visits)? Like Question 3, Question 4 investigates design, though on a much longer time scale.

In Chapter 5 for various communicative intents, we acoustically described prosodic behaviors and automatically recognizing them. First, we built online recognition (similar to a previous online recognition system described by Breazeal & Aryananda, 2002) and added *online learning* from prosodic input from humans. We proved that it *could* be done, but that

this wasn't sustainable: the pace is too slow to maintain a human teacher's engagement. Also, the grain of adaptation was too fine for a human to reliably judge and give feedback on.

In Chapter 2 we learned that people's feedback depends on history, and that human teachers provided input of different types than post-action feedback. We see that machine learning from human input has a ways to go to understand a human teacher's strategies, expectations of a learner. These are structured at a higher level than just positive/negative feedback. They are more nuanced and organized according to a longer-time structure.

In Chapter 2 we found that robot-directed prosody changes with robot behavior exposes changing expectations of a learner, which informs the way a robot should adapt throughout an interaction, and perhaps beyond a single interaction. This motivates further exploration into people's expectations of adaptation throughout a relationship of repeated interactions. In Chapter 5, we develop a system that refines a physical behavior depending on affective prosodic input. The goal of socially guided machine learning systems (SGML) is to understand how robots can adapt to people automatically. Chapters 2 and 5 explore the ways that a machine learner can learn use social inputs from typically developing adults, a basis for further exploration into social inputs from children with ASD.

The state of the art may be years from systems that efficiently use human input to learn. This indicates the value that socially assistive robots can contribute to machine learning, by establishing the social behaviors humans use to interact with robots. In the mean time, while we actively develop adaptive systems, we can continue to deploy socially assistive robots, operating them manually.

Clearly, while the application of socially assistive robots to autism intervention engenders its own particular research questions, investigations for this particular application tend to be

mutually supportive of research in human-robot interaction more generally, of machine learning from human input, of developmental psychology. Knowledge travels between socially assistive robots for autism and these other lines of inquiry, in both directions.

# Chapter 2

# How People Talk When Teaching a Robot

The study presented in this chapter was the first group demonstration of untrained people's spontaneous use of affectively expressive speech prosody directed toward a social robot (E. S. Kim et al., 2009). The primary motivations for this work were two-fold, (1) to describe the (a) way naïve people with TD use affectively expressive prosody when teaching, and to describe in detail (b) whether and (c) how they modulate these expressions depending on the learner's performance; (2) to establish the feasibility of using robots to elicit affectively expressive prosody in a special population, such as children with ASD. We created a teaching interaction for adults with TD, in which they were instructed to help a small dinosaur robot, walking down a miniature street, learn which toy buildings should be knocked down. We examined affective vocalizations provided by untrained human teachers to robotic learners. In unscripted one-on-one interactions, adult participants with typical development provided vocal input to a robotic dinosaur as the robot selected toy buildings to knock down. We found that (1) people vary their vocal input depending on the learner's performance history, (2) people do not wait until a robotic learner completes an action

before they provide input and (3) people naïvely and spontaneously use intensely affective vocalizations. We established that a statistical sample of adults easily and spontaneously use affectively expressive speech prosody to instruct robotic learners, suggesting the feasibility of testing robots' ability to elicit affectively expressive prosody from children with ASD. Our results also suggest that machine learning from human social input may require the modification of classical machine learning algorithms. Our findings suggest modifications may be needed to traditional machine learning models to utilize human interactive input in ways that are richer, more accurate, and more acceptable to users.

## 2.1 Motivation

We wanted to see if untrained, typical adults would use affectively expressive prosody when talking with a socially engaging robot, in a teaching task. We wanted to first establish the acceptability and ease of socially interacting, particularly with speech prosody, with a robot, and the technical feasibility of producing an interaction with the robot over a large group. Though our ultimate aim is to develop socially assistive robots for children with ASD, we chose to work first with typically developing adults, whose participation entails fewer ethical, developmental and logistical constraints than a special needs population of children. Our tests of acceptability, usability, and feasibility over a large group were fundamentally important, because prior to this study, there existed evidence only of trained individuals' production of affective prosody with robots.

At the same time we were interested in learning what kinds of spoken input people give to a robotic learner. This study was largely exploratory in nature: if untrained, neurotypical

adults wouldn't use expressive prosody while interacting with a robot, there would be no sense in trying to get kids with ASD to do it.

We also recognized the opportunity to explore how we might computationally model human social behaviors as inputs to a machine learning system. Robotic and computational systems that learn from human input have included those that learn from demonstration or by imitating a human instructor (Argall, Chernova, Veloso, & Browning, 2009; Breazeal & Scassellati, 2002; Schaal, 1997); methods that learn the salience of an object or task based on observations of a human's affective expressions or attention (Breazeal & Thomaz, 2008; Lockerd & Breazeal, 2004; Nicolescu & Mataric, 2001; Thomaz, Berlin, & Breazeal, 2005); and systems by which a human shapes behavior using methods inspired by operant conditioning (Blumberg et al., 2002; Kaplan, Oudeyer, Kubinyi, & Miklósi, 2002; Knox, Glass, Love, Maddox, & Stone, 2012; Knox, Stone, & Breazeal, 2013; Thomaz & Breazeal, 2006b). Among these systems are efforts to learn from untrained behavior performed by non-expert humans. Generally, this emergent field could be described as socially guided machine learning (SGML), including both behaviors that people might typically use to interact with other people—like speech—as well as behaviors that people might use to interact with anything they consider to have a mind—such as clicker training with dogs. Explorations into interactive machine learning have observed the need to modify classical machine learning models to fit natural human teaching preferences (e.g., Thomaz & Breazeal, 2006b). Likewise, we in this study, we shaped our experiment design in order to facilitate exploration of human interactive behaviors with respect to commonly used interactive machine learning algorithms.

Specifically, we were motivated to explore the applicability of human prosodic input to reinforcement learning algorithms, as these algorithms have been popularly used in learning from human interaction (Blumberg et al., 2002; Broekens, 2007; Isbell, Shelton, Kearns, Singh, & Stone, 2001; Kaplan et al., 2002; A. Stern, Frank, & Resner, 1998; Thomaz & Breazeal, 2006a). Reinforcement learning has classically been viewed in two senses, (1) as a broad framework of the problem of learning from rewards in an environment, and (2) as a collection of specific techniques (Sutton & Barto, 1998). Classical reinforcement learning techniques are based on Markov Decision Processes, such that the feedback or rewards a learner will encounter depend only on the current and past (or perhaps a small number of recently past) states in the learner's recent history (Russell & Norvig, 2003, pp. 763–784). Otherwise (and generally) reinforcement learning algorithms assume that the feedback or rewards a learner receives are independent of the learner's history. Intuitively, however, this contradicts human expectations that learner's won't make the same mistakes multiple times in a row, and human emotional responses to progress or apparent mastery in learning. We designed our interactive study to explore these expected contradictions.

## 2.2 Hypotheses

### Hypothesis 1

Naïve instructors will provide affective guidance and feedback. Explorations into interactive machine learning have observed the need to modify classical machine learning models to fit natural human teaching preferences (e.g., Thomaz & Breazeal, 2006b). We hypothesized that naïve speakers, given the opportunity to comment on a robotic learner's intended action,

would vocalize about the intended action before the action is completed. We expected less intensity in affect expressions voiced before an action is completed then those voiced after.

### Hypothesis 2

Naïve instructors will use affective vocalizations without explicit instruction. In previous studies of naïve speakers talking to robots, affective prosody has been elicited only when participants were explicitly "instructed to express each communicative intent (approval, attention, prohibition, and soothing) and signal when they felt that they had communicated it to the robot" (Breazeal & Aryananda, 2002); they were instructed to act as though talking to a child or pet. Based on our own anecdotal observations of naïve people interacting with Pleo robots, we hypothesized that naïve people would use affective prosody when talking to a robot without explicit instruction.

### Hypothesis 3

Vocalizations will vary with respect to the history of a robotic learner's performance. Our predictions fell into two categories. (a) We expected differences in comparing a consistently successful learner with one that initially struggled. Specifically, we hypothesized that naïve speakers would produce more intensely positive prosody in response to a robotic learner's correct choice if it followed a series of wrong choices, then if it followed a series of correct choices. (b) We expected that people would speak less, and with weaker affective prosody, as a robotic learner consistently succeeded. We hypothesized that, for a robot that made a series of correct choices, both the amount of vocalization and the strength of affect in prosody would fall as its successful streak continued.

## 2.3 Methods

To investigate these hypotheses, we designed an experiment in which participants were asked to help two robot dinosaurs pick the right buildings to demolish as they walked through a model city. Unbeknownst to the participant, the dinosaur robots were being controlled by a remote operator; this experiment model is called "Wizard of Oz" (WOz; Dahlbäck, Jönsson, & Ahrenberg, 1993; Riek, 2012; Steinfeld, Jenkins, & Scassellati, 2009).

### 2.3.1 Participants

We recruited 27 participants, 9 male and 16 female, 18 years of age and above, from the Yale University and New Haven communities. Our exclusion criteria were based on English proficiency and previous research or coursework experience in artificial intelligence. We excluded data from three participants from post-experiment data analysis, due to technical failure of the robot or recording devices for two participants, and gross non-compliance of protocol by one participant (he deliberately instructed the robot to make mistakes).

### 2.3.2 Experiment design and procedures

#### 2.3.2.1 Interaction protocol

A testing session lasted approximately 30 minutes. Participants gave informed consent to be recorded. The participant was brought into the room containing the two dinosaurs and the demolition training course. The participant stood at the edge of a table and clipped a lapel microphone to his/her shirt collar. The two robots, which we introduced as "Fred" and "Kevin," stood in front of the demolition training course, close to and facing the participant.

The participant was told the following:

> These are our dinosaurs; their names are Kevin and Fred. Kevin is the one with the red hat with the "K" on it. Fred is the one wearing a bandanna. Today they're going to train to join a demolition crew. They'll be knocking over buildings with their heads. Behind them is the training course that they'll be running today. They'll go one at a time: Fred will be first and I'll take Kevin and leave the room. When Fred's done, then it'll be Kevin's turn. (The ordering of the dinosaurs varied per participant.)
>
> You are going to help them pick the red "X"-marked buildings in the training course to demolish. In the training course, you'll see there are three pairs of colored buildings standing across from one another – the purple pair at the far end, the silver pair in the middle, and the orange pair closest to us. The robots will do the training course sequentially, starting at the purple buildings and walking towards us. For each pair, you'll see that one is marked with an "X." Kevin and Fred can see the "X"s too. For each pair of buildings it's important that they knock down the building with the "X" and that they don't knock down the unmarked building.
>
> They already know how to knock down buildings. We want you to help them understand that they should only knock down the buildings with the red "X"s and all of the ones with the "X"s. You're going to help them by talking with them. We encourage you not to make any assumptions about how this might work. Just act naturally and do what feels comfortable. Please stay in this area [demarcated by caution tape]. The training is complete when an orange building falls.

The experimenter then engaged the participant by asking him/her to say hello to the robots and explain to them the task, in his/her own words. The robots returned the greetings with growls and acknowledged the receipt of instructions by looking and vocalizing at the participant in time with his/her words. The experimenter then solicited questions or provided additional clarification for the task from the participant.

Once the participant was comfortable with the task, the experimenter placed one of the robots at the start position, between the first pair of buildings, facing the participant. The experimenter left the room with the other robot. Then the participant guided the first robot in training. The first robot gave a "Charge!" vocalization indicating the start of the trial. The robot slowly (over 4 seconds) communicated its intent to topple one building in the first pair, by slowly turning its head towards it while vocalizing a slowly increasing growl. If the

participant did not vocalize negatively towards the robot, it continued to push the building over, realizing its communicated intent. Otherwise, the robot discontinued its communication of intent-to-knock-down. It then turned its head towards the other building in the pair, communicating its intent to knock down it down. After the building fell, the robot walked forward to the next pair of buildings and repeated the sequence for the second and third pairs of buildings. The experimenter returned to the training room when the participant verbalized (to the robot) that it had completed the training course, or after a period of time elapsed (approximately 30 seconds) after a building in the final pair collapsed, whichever came first.

When the training was complete, the participant was given a few minutes' break while the experimenter reset the demolition training course (set the buildings upright). The participant then engaged in a training session with the other dinosaur while the experimenter and the first robot waited outside. The second training session proceeded identically to the first, except in the robotic learner's predetermined sequence of communicated intents. One robotic learner, Fred, initially chose the correct building for every pair, whereas the other robotic learner, Kevin, communicated intent to topple the wrong building in the first and second pairs, and then communicated intent to topple the correct building in its third trial.

Once the second training session was complete, the participant completed a survey. Then the experimenter debriefed the participant by showing him/her the WOz control room, explaining the technology, explaining the purposes of the study, and answering any questions.

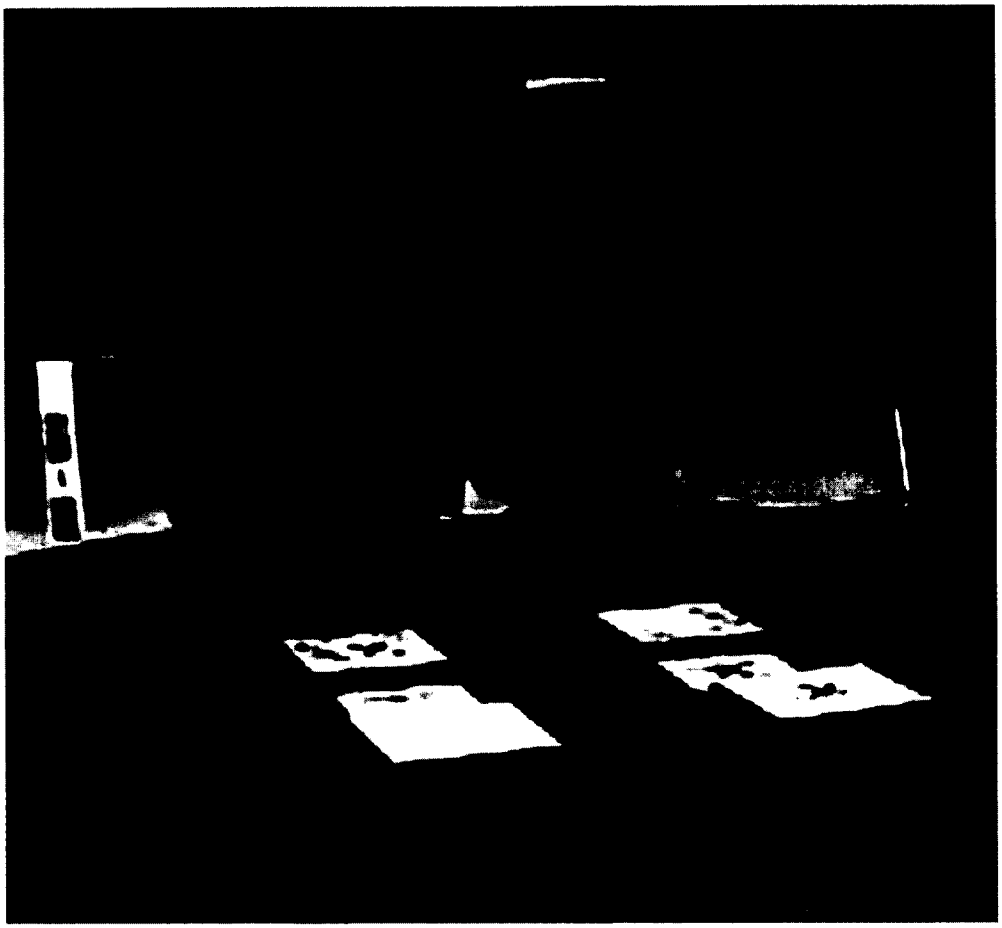


Figure 2.1 A participant talks to one of two robotic learners, as it completes the demolition training course. (Best viewed in color.)

#### 2.3.2.2 Interaction environment

Pleo is an 8-inch tall, 21-inch long dinosaur robot, sold commercially as a toy by UGOBE Life Forms (UGOBE Life Forms, 2008). In this experiment, we endowed our robotic learners Kevin and Fred with distinct recorded "voices" designed to distinguish them as individual social actors (Nass, Steuer, & Tauber, 1994). Although we intended to exactly counterbalance the order in which participants interacted with the two robotic learners, Fred completed the training course first in 33% of the testing sessions. This imbalance was due to

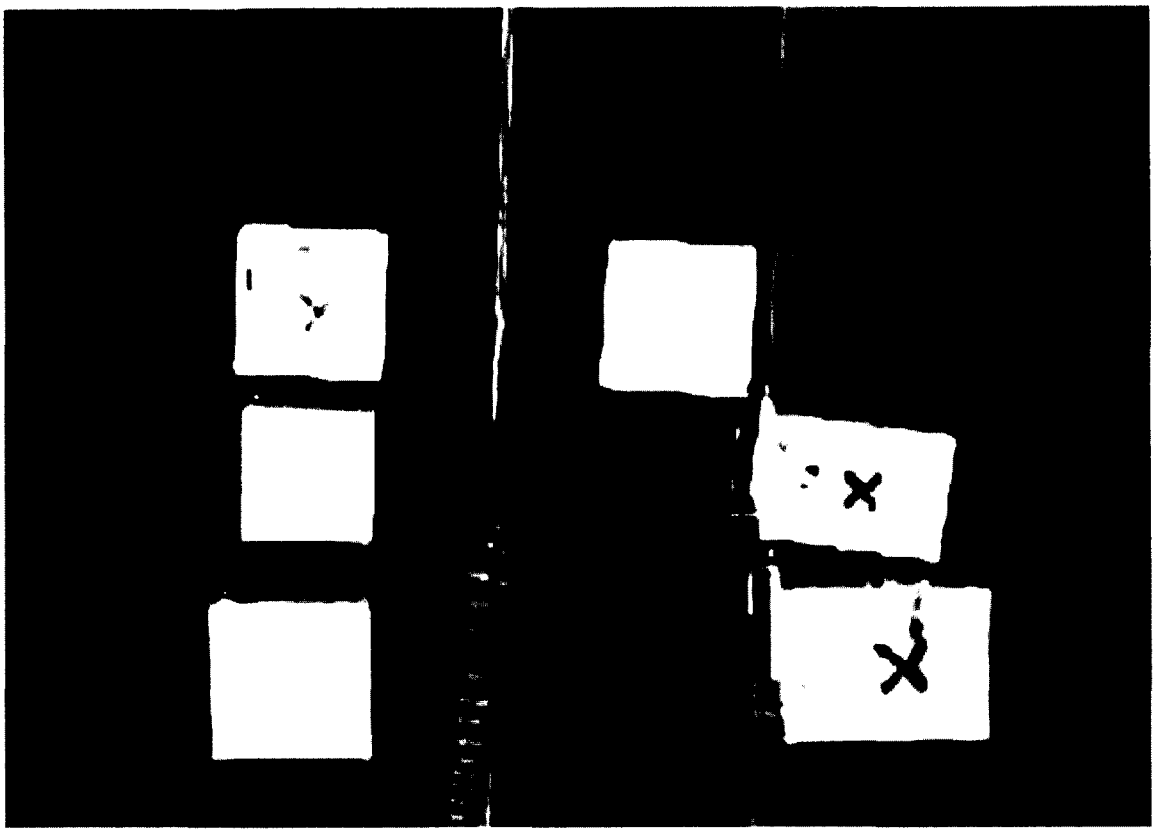


Figure 2.2 The overhead view used for Wizard of Oz control of the robot's locomotion. North of this frame, a participant is standing at the end of the table. Pairs of identically painted and sized buildings lined the road, down which the robotic learner walked. Each building was placed on the opposite side of the road from its pair. One building within each pair was marked with red "X"s, indicating it should be toppled; the other building was unmarked. For each pair, the robot first walked forward until its head was between the two buildings. It then communicated its intent to knock down one of the two buildings, and then fulfilled its intent or corrected itself, depending on the human tutor's communications to it. After toppling a building in a pair, the robot walked forward until its head was between the next pair. The three pairs of buildings were separated from each other along the road by spaces of 3 inches. From the robot's perspective, the "X"-marked buildings were right, right, and left buildings, in the successive pairs.

the exclusion of 3 participants' data from analysis, as well as an accounting error during data collection.

The robot's "demolition training course" sat on a large table top, 3 feet off the ground, comprised of 8- to 12-inch tall toy, cardboard buildings (shown in Figure 2.2). The buildings

were arranged in a series of identical pairs, each straddling a double-yellow-lined "road" (that is, a raised, 30-inch-long track along which the robot walked). One building in each pair was marked with red "X"s on all sides, indicating that it should be toppled. The distribution of marked buildings was controlled to be constant for all participants.

Participants stood about three feet from the nearest point of the training course throughout their entire interactions with Kevin and Fred, to prevent physical manipulation of the robot or training course. Each robot and the training course's road faced the participant, so that as the robot walked through the series of buildings, it approached the participant.

#### 2.3.2.3 Robot control

The WOz design was necessary to ensure a real time interaction between the participant and the robot dinosaurs. Autonomous robot control may not have provided a fluid interaction because we did not construct narrowly defined expectations of participants' interactions with the robot. Thus, WOz control allowed us to simulate the application domain. Each robot was alone with the participant, so the participant spoke directly to the robot, and not to the person operating the robot. The deception entailed by WOz was approved by Yale's Institutional Review Board. None of the participants guessed, in the survey or debriefing, that a human had secretly controlled the robots.

For WOz supervision, we used an overhead webcam for accurate estimation of the robot's range to strike buildings (shown in Figure 2.2) and a video camera aimed at the participant for viewing facial expression (shown in Figure 2.1). The wizard was also able to hear the participant through the clip-on lapel microphone.

The robot dinosaurs were controlled using infrared (IR) signals. The IR receiver was located in the dinosaur's nose. IR signals were sent from long distance IR beacons through an IguanaIR USB-IR transceiver (IguanaWorks, 2008), controlled in Linux using LIRC (Linux Infrared Remote Control) software (Bartelmus & Scheibler, 2008).

The wizard controlled the robot dinosaur's motions and vocalizations using a combination of scripted behaviors, which were mapped to inputs on a USB handheld gaming pad. For example, pressing the joystick forward caused the robot dinosaur to walk forward, and pressing to the left or right caused the robot to move his head in the respective direction. These scripts were created and modified using UGOBE's software development kit and MySkit (DogsBody & Ratchet Software, 2008).

To appear autonomous and life-like, the robot dinosaurs were programmed with idling behaviors. Affective vocal and motor responses provided a heightened sense of communication. For example, the dinosaur would put his head down and make a sad "oh" sound when reprimanded. To ensure responsiveness, the walk and idle behaviors were short. Also, the intention script (dinosaur moving head towards a building and roaring) was interruptible, in the event that the participant reprimanded or corrected the robot.

### 2.3.3 Analysis of vocal input

#### 2.3.3.1 Three types of vocalizations

Each participant's interaction was both video and audio recorded. The resultant audio recordings were segmented and analyzed. We noted participants' vocalizations fell into three cycling phases based on the robot's progress in each trial. All three phases occurred for each trial: direction, occurring before the dinosaur picked a building; guidance, occurring while the dinosaur swung its head to knock over a building; and feedback, occurring after the

building fell or the dinosaur abandoned his effort. We segmented our audio data along this dimension.

For each robot, the first of the three trials began with the robot placed between the first pair of buildings, where he would indicate his readiness by vocalizing. He would then signal his intent, lasting a few seconds. Then, if he was not reprimanded or corrected, he knocked over the building he intended to. In this case, the first sentence describes the direction phase, the second describes guidance and the last is feedback.

In this manner, phases cycled from direction to guidance to feedback, then back to direction. By design (and expectation of participants' behaviors) each trial included one or two cycles: either the robot arrived between the pair of buildings, motioned towards the correct building, and knocked it down; or after arriving at the pair, the robot motioned towards the wrong building (and received reprimand, which we categorized as guidance), then replied to the reprimand (we categorized speech produced during this action as direction), then motioned towards the correct building (guidance), and knocked it down (feedback).

We performed the audio segmentation according to these guidelines and exported them to our raters. The segmentation was performed by recognizing the dinosaur sounds we heard on the recording that uniquely identified the phases of each trial. The only phases for which that rule did not apply were between trials: separating the last phase of one trial (feedback) and the first of the next trial (direction). We waited for a two-second pause in our participants' vocalizations, and if there was none, we divided based on the transcription of the words used such that once they stopped using disparaging words (e.g. "no," "stop"), that moment divided the trials.

#### 2.3.3.2 Annotating affect

For each audio clip which was segmented by phase, we analyzed word counts, prosody ratings, prosodic intensity ratings, and post-experiment survey responses. Our raters were not informed of the content of the audio nor the experimental design. We randomized our audio files split by phase (average length approximately 20 seconds) and asked two raters separately to rate each audio clip's prosodic affect as either positive, negative, or neither. Positively affective prosody was described to the raters as sounding "encouraging," "approving," or "pleasant," whereas negative affect was described sounding "discouraging," "prohibiting," or "disappointing." We also asked the raters to rate the intensity of the affect on a differential semantic scale (Osgood, Suci, & Tannenbaum, 1967) from 0 (neutral or no affect) to 2 (very strong affect), and their respective confidences for each judgment on a differential semantic scale from 0 (not sure) to 2 (quite sure). Word count was also extracted from the audio clips.

The ratings of two naïve raters of affective prosody (prosody ratings and prosody intensity ratings) showed high agreement ($K$= 0.84 using Cohen's quadratically-weighted (1968), normalized test).

## 2.4 Results

Most phase durations were short and contained few words ($M$ = 7.71 words/phase or 1.26 words/sec; $SD$ = 8.43 words/phase or 1.36 words/sec).

### 2.4.1 Instructors vocalize before, during, and after a learner's actions (Figure 2.3)

We hypothesized that participants would provide affective guidance and feedback (Hypothesis 1). We found that participants provided an almost equal number of words while a learner performed an action (guidance) and after the action was completed (feedback). In addition, we found that participants provided a similar number of words well before the learner communicated any intent to act (direction). (See box-and-whisker plots in Figure 2.3.) Over all phases, the frequency of words spoken was on average 1.26 words/sec, with a standard deviation of 1.36 words/sec.

### 2.4.2 Instructors express affect during and after a learner's actions (Figure 2.4)

We hypothesized that naïve instructors would use affective prosody when speaking to a robot (Hypothesis 2). Although we did not specifically instruct participants to use affectively expressive prosody, they vocalized with intensely affective prosody during guidance (affective intensity, $M = 1.28$, $SD = 0.93$) and feedback ($M = 1.89$, $SD = 0.78$; see Figure 2.4). Participants showed no positive or negative affect during direction (affective intensity $M = 0.47$, $SD = 0.68$). One-way ANOVAs and affective intensity were both significant ($p < 0.001$ for both ANOVA tests, $F(2) = 58.2, 19.2$).

### 2.4.3 Instructors say less as a learner continually succeeds (Figure 2.5)

During all phases of interaction with Fred, the robot who consistently moved to topple only correct buildings, participants spoke less from one trial to the next ($p = 0.002$, linear regression, see Figure 2.5). Speech rate dropped during the guidance ($p = 0.018$) and feedback instruction phases ($p = 0.038$), but not the direction phase ($p > 0.1$).

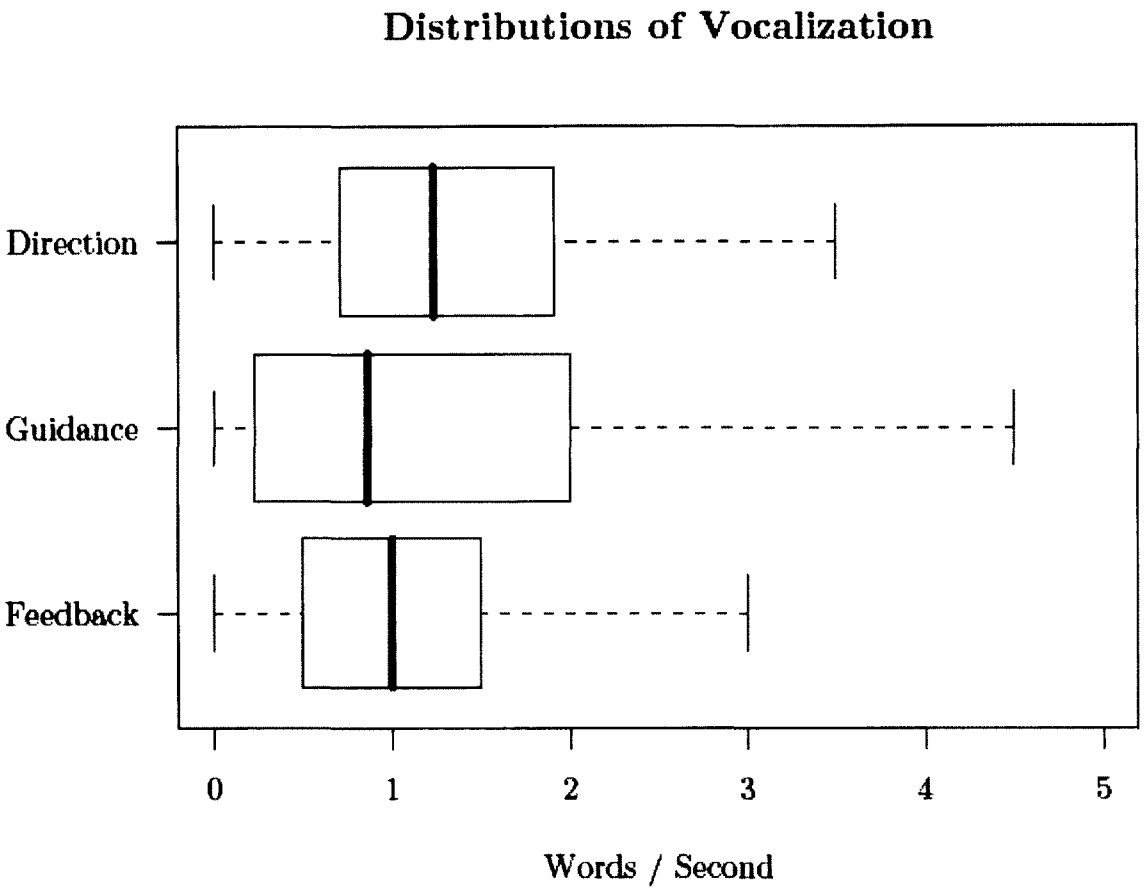



Figure 2.3 Rates of speech (number of words/sec) are similar across all three instruction phases.[3] We verified that these trends could not be explained by the order in which participants interacted with the two robotic learners. In a two-way ANOVA (trial number x learner-order), we found a highly significant main effect for trial number ($p = 0.0018$, $F(1) = 10$) and for learner-order ($p = 0.0004$, $F(1) = 13$), but no effect of interaction ($p = 0.38$, $F(1) = 0.7$). A similar test for Kevin (the learner who in three trials selected wrong, wrong, and finally

[3] Box-and-whisker plots provide a snapshot of a distribution: the bold line marks the median; the left and right edges of the box mark the medians of the lesser and greater halves of the distribution, and also define the lower and upper bounds of the second and third quartiles, respectively; the least and greatest whisker ends (bars) denote the minimum and maximum values in the distribution.

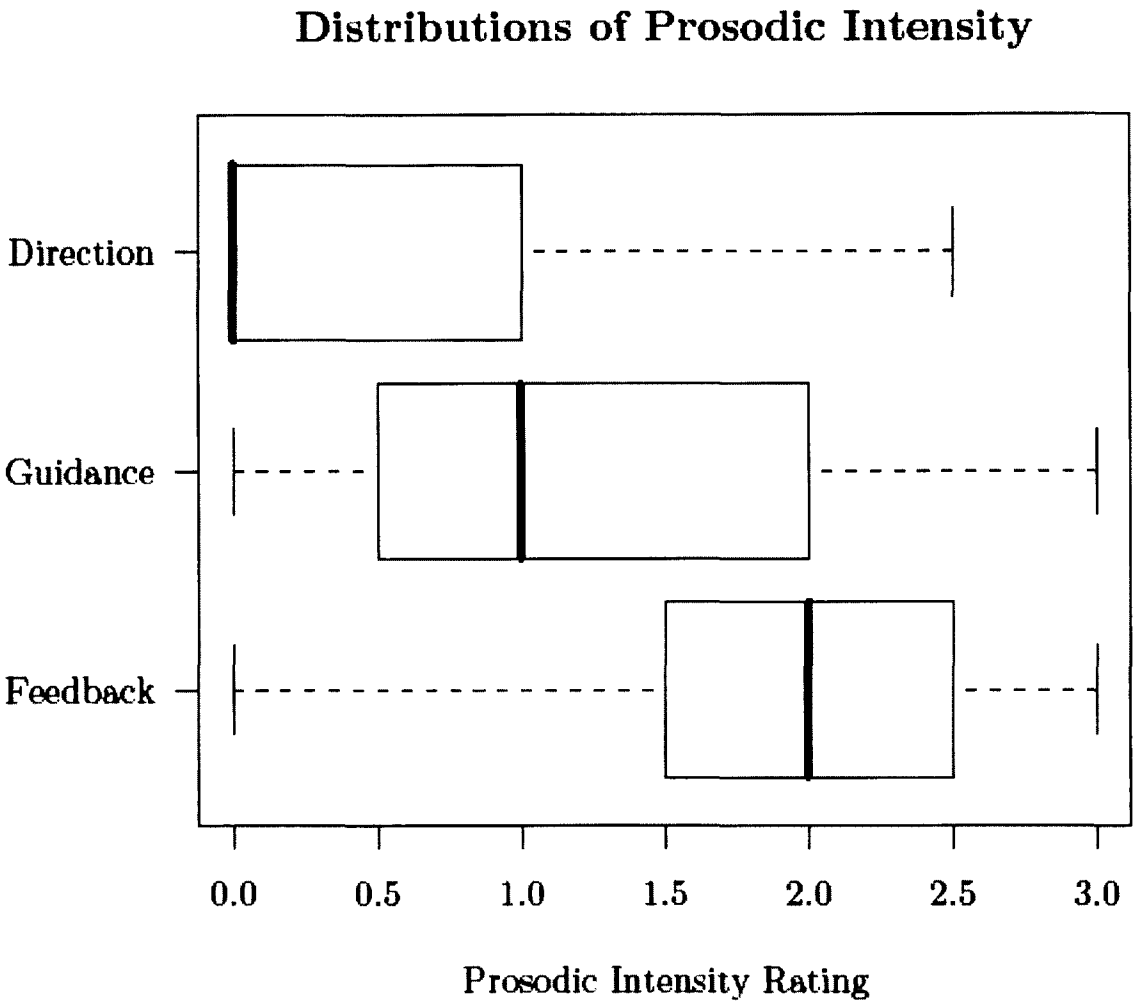



Figure 2.4 The distributions of the intensity of the affective prosody during each phase demonstrate that people use prosodic reinforcement as feedback on an ongoing or previously finished behavior. Affective prosody intensity ratings ranged from 0 (neutral or no affect) to 3 (intense affect).

correct buildings) showed no trend of decreasing word/sec over trials ($p = 0.57$, $F(1) = 0.38$).

### 2.4.4 Instructors say more after a new breakthrough (Figure 2.6)

We compared direction, guidance, and feedback phases during the third trial for Kevin against those for Fred. Recall that in the first two trials, Kevin initially communicated intent

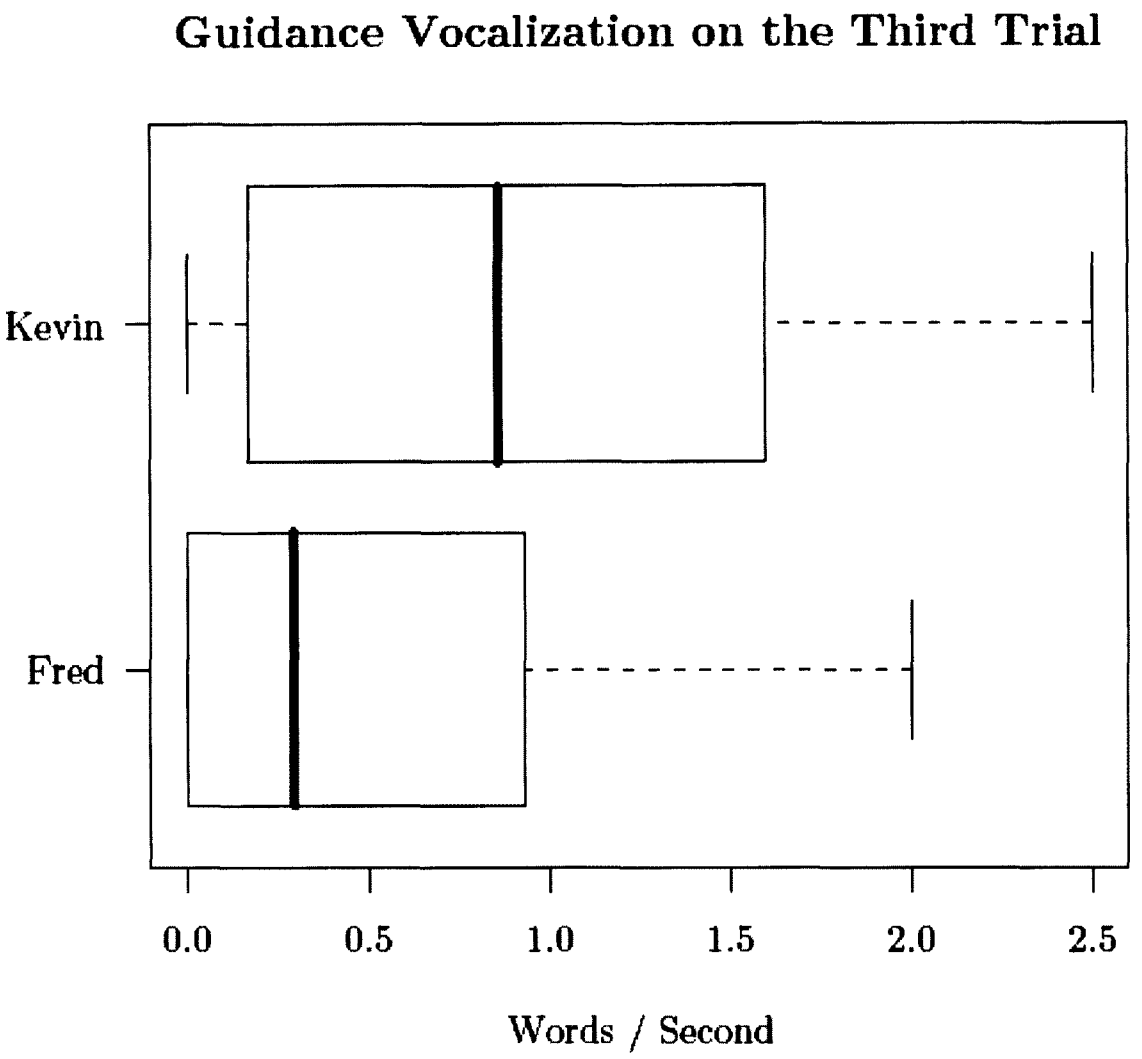



Figure 2.5 Distributions of the number of words spoken per second during the third trial's guidance phase. One robotic learner (Fred) consistently communicated intent to topple only correct buildings, while the other (Kevin) at communicated intent to topple the wrong buildings in its first two trials. In the third trial, shown here, both robotic learners selected the correct building, representing consistently correct behavior in the case of one robot (top, Fred), and an indication of improvement, or progress in learning, in the second robot (bottom, Kevin). During the guidance period (during which the robot communicates its building selection but has not yet toppled it) in the third trial, the improving robot (bottom, Kevin) received more utterances than the consistently correct robot (top, Fred), with marginal significance ($p = 0.051$).

to topple the wrong buildings, while Fred only communicated intent to topple correct buildings in the first two trials. In the third trial, both Kevin and Fred initially communicated intent to topple the correct building.

We hypothesized that prosody would be more intensely positive in response to Kevin's than to Fred's third trial intent (Hypothesis 3), since this would showcase the participants' relative excitement at Kevin's improvement. Considering only guidance and feedback phase audio clips, we found that participants voiced marginally significantly more words/sec to Kevin than to Fred ($p = 0.089$, $F(1) = 3$). We found neither a main effect of learner-order nor an interaction between learning condition with learner-order. Figure 2.6 shows the trend for participants to give more guidance and feedback to Kevin than to Fred. We found no such difference for affect or affective intensity ratings.

## 2.5 Discussion

This study provided the first large group-based evidence that untrained, typically developing adults will spontaneously direct affectively expressive prosody to a robot. We hypothesized that participants would provide affective guidance while a learner was carrying out trial actions, as well as feedback after actions were completed (Hypothesis 1). We found that in addition to providing guidance and feedback, participants provided direction—verbal instructions spoken before the learner communicated any intent to act—and that participants spoke an almost equal amount throughout all three phases (i.e., direction, guidance, and feedback) of the learning trials (see Section 2.4.1).

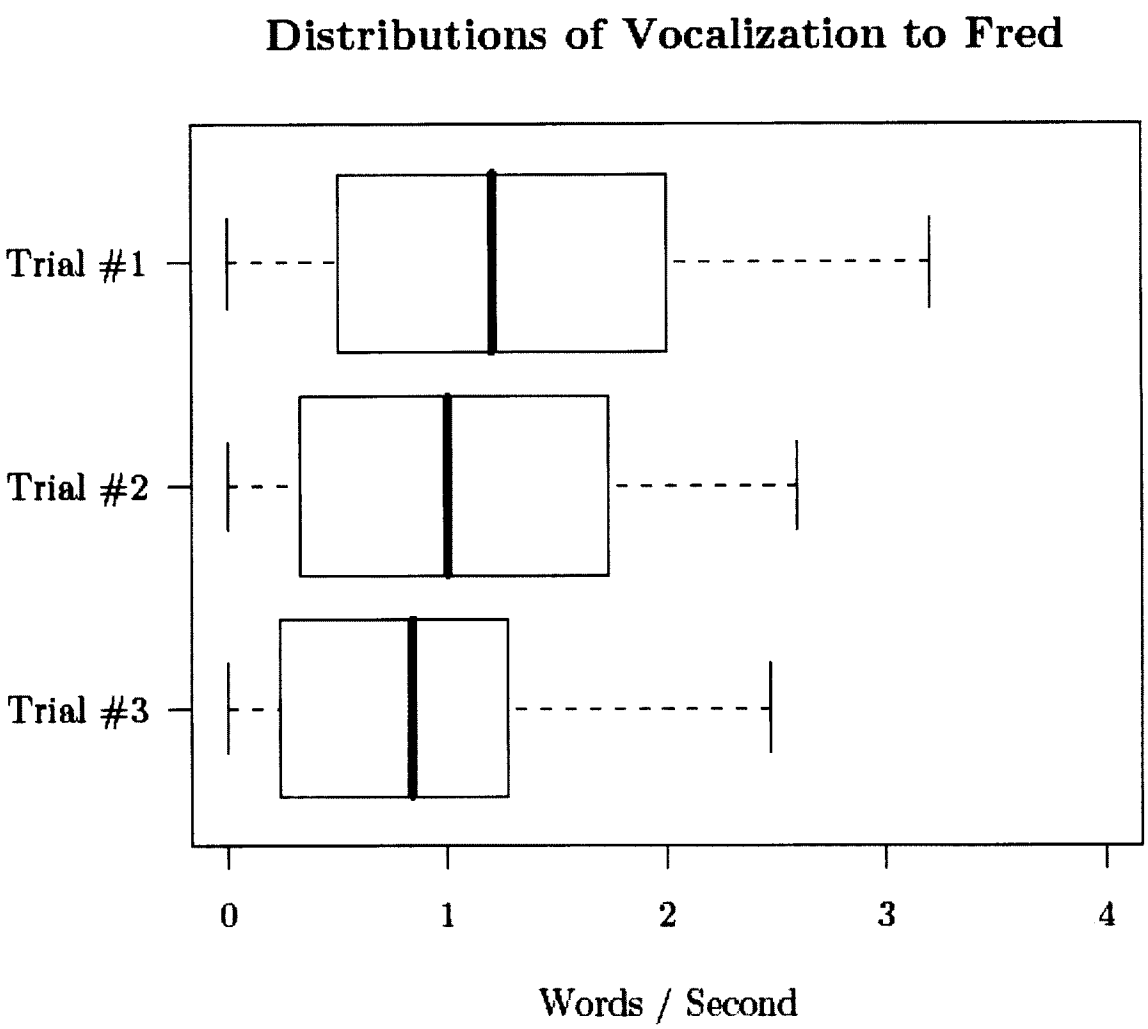



Figure 2.6 These are the distributions of the number of words spoken per second during the third trial's guidance phase. In the first two trials, Fred has consistently intended to topple only correct buildings, while Kevin has intended to topple the wrong buildings. In this third trial, both dinosaurs initially intend to knock down the correct building. In guidance during intent in the third trial, Kevin receives more utterances than Fred, with marginal significance ($p = 0.051$).We also hypothesized that naïve people would use affective prosody when speaking to a robot (Hypothesis 2). Participants used affectively expressive prosody during guidance (while the robot expressed its intent) and feedback (after the robot had completed an action) phases of learning trials but not during direction (before the robot had indicated its selection). These distinct amounts of affect intensity are consistent with the intuition that positive and negative affect are used to provide reinforcement as guidance for an ongoing behavior; or as feedback for a finished action; whereas reinforcement is not given during direction, before a behavior begins.

Participants spoke less, from one trial to the next, for the learner who always made correct selections. This was not true for the improving learner, which initially made two incorrect selections, before finally making a correct selection. This was true, regardless of the order in which they interacted with the two learners. This indicates that a human teacher's spoken inputs to a learner should not be modeled as independent from one trial to the next, that teachers' inputs depends on the learner's performance history.

Finally, we had hypothesized that prosody would be more intensely positive in response to the initially-incorrect learner's final, correct selection, than to the always-correct learner's final correct selection. Considering only utterances produced during guidance and feedback phase, because the direction phase precedes the robot's selection, we found that participants voiced marginally significantly more words/sec to the initially-wrong-but-finally-correct robot than to the always-correct robot's last selection.

### 2.5.1 Implications on machine learning

Our results suggest that spoken inputs on learning trials should not be modeled as path-independent, but rather spoken reward signals depend on the history of intentions shown by the learner, even if the learner's ultimate performance (or path) is identical (all the same buildings were knocked down by both learners; the difference is that the struggling learner twice initially indicated incorrect selections, before being corrected and ultimately demolishing correct targets). This finding contradicts an assumption required by classical reinforcement learning models, that rewards can be considered independent. Algorithms using human social inputs as rewards should take into account the rewards' dependence on the learner's history of communicated intent, or history of dependence on guidance, not only the learner's history of performance.

Specifically, we found that human teachers tailor their feedback to account for the history of the learner's performance. In terms of a machine learning model, we view the affective vocalization reward signal as neither stationary nor path-independent, two assumptions made by standard algorithms. We found this to be true in two ways. First, the robotic learner that performs the correct action in a third trial will receive significantly more guidance and feedback if it previously made wrong choices then if it has been consistently correct. This shows that human feedback to a robotic learner is not path-independent. Second, for a learner who is consistently successful, guidance and feedback wane.

We suggest to HRI researchers interested in implementing machine learning from human vocalization that they model human reinforcement signals as dependent on the progress of the learner. Furthermore, we suggest that machine learning from human teaching should make use of currently neglected vocalizations giving direction to the robot before it acts as well as guidance to the robot as it indicates its intent to act. Direction has traditionally been ignored, and guidance has only recently been explored in machine learning from human input (e.g., Thomaz & Breazeal, 2006a).

Our findings bear on the application of reinforcement learning algorithms to human-robot and human-computer interactions. First, our results suggest that applications based on classical reinforcement learning algorithms (which utilize only feedback arriving after actions are completed) should be extended to take advantage of non-reward inputs arriving before learning-task-specific actions are taken. Such flexibility has been demonstrated in the form of guided action selection, utilizing naïve people's guidance input to a learner which communicates its consideration of action options (Thomaz & Breazeal, 2006a). Further, our results suggest that assumptions of path- and history-independence in Markov-decision-

process-based reinforcement learning algorithms are violated in the contest of rewards supplied by human affective expressions. We suggest that reinforcement learning techniques should be adapted or that human-interactive learning tasks should be modeled differently than they historically have been, in order to account for human instructors' adaptations to their sense of the learner's progress or mastery.

### 2.5.2 Implications on autism research

This study established the feasibility of eliciting affectively expressive speech from typically developing adults, encouraging us to try developing speech-based human-robot interactions for children with ASD. Further, our examination of variations in amount of speech and intensity of prosody, depending on the phase of instruction within each trial, and depending on the history of the learner's communicated intentions, give us insight into how to design machine learning systems that use human input, but also reveal the kinds of robotic behaviors which might better elicit affectively intense prosody or more speech, from typically developing adults and potentially from individuals with ASD. Specifically, we observed that a learning robot that makes mistakes is likely to receive more intensely affective prosodic feedback from typically developing adults; and so we can seek to elicit more intensely affective prosodic feedback from individuals with ASD under similar circumstances. The project of improving a robot's ability to learn from human input may also eventually improve human-robot interactions with individuals with ASD, by allowing us to make robots more adaptive to spoken feedback. Although studies of long-term relationships between robots and humans are limited, evidence suggests that any human-robot relationships, as we might intuit any relationships, will fail if either party fails to remember and adapt to shared knowledge and experiences (Kozima et al., 2009). Finally, this

study provides us with a sample of affectively expressive prosody, from which we may be able to train automatic systems to recognize and classify affect.

### 2.5.3 Limitations

Manual robot control, such as Wizard of Oz presents a potential weakness in experimental validity: if the robot controller is not blind to the hypotheses, she or he may influence the robot's behaviors to disadvantage the null hypothesis. In this case, the wizard could introduce bias by, for instance, making the robot behave with greater uncertainty, or respond to instructors' vocalizations more frequently or emphatically before the buildings collapsed, in order to elicit more vocalizations during these Direction and Guidance phases of interaction. Better experimental control would be achieved by blinding the robot controller to our hypotheses. However, due to limited resources, we chose instead to operate the robot ourselves, though we had designed the experiment. To some extent we ask our readers to trust that we remained faithful to our experimental protocol, which strictly states when the robot should respond and with which behaviors. However, an additional validation step can be taken: a rater, who is blind to the hypotheses, can measure the fidelity robot's behavioral fidelity to the experimental protocol. We did not perform these measurements in this or the other experiments described in this dissertation. This is a limitation to the validity of all the Wizard-of-Oz-controlled interactions described in this dissertation (see also Chapter 3 and Chapter 4). However, we can make video recordings of all of our data samples available to such validation, and we may undertake such fidelity measures ourselves at a later date.

## 2.6 Conclusions

We designed and conducted an experiment in which naïve teachers helped a dinosaur robot learn to topple marked buildings in a demolition training course. Our goal was to investigate how people intuitively talk without explicit instruction when teaching robots. We found that naïve vocalizations during human-teacher/robot-learner interaction appear to segment into three distinct phases, providing different kinds of input to the learner. These three phases are direction (before the learner acts), guidance (as the learner indicates intent) and feedback (after the learner completes a task-action). We observed that naïve human teachers vocalize readily throughout all three phases. Our experiment showed that people are affectively expressive as they direct the robotic learner well before it approaches the learning task, as the learner communicates its intention to act (effectively querying the teacher), and in giving feedback for actions the learner has taken. Thus, we have affirmed an intuition held by human-robot interaction (HRI) researchers that naïve speakers do spontaneously use strongly positive and negative affective prosody when talking to a robot. We have also found that some human teaching behaviors do not fit well within classical machine learning models of interactive learning. Finally, our results are consistent with previous observations of human teachers' behaviors toward fellow-human learners, showing a correlation between children's improving language skills and declines in feedback from their parents (Chouinard & Clark, 2003). As they do for infant learners, our findings suggest that people modify feedback for a robotic learner, depending on the learner's progress. This may suggest that typical adults will teach a robotic learner similarly to the ways they would teach a human learner. This also supports an interesting line of inquiry: in what other ways will people

afford social expectations and behaviors toward a robot similar to those they would afford to another person?

# Chapter 3

# Social robots as embedded reinforcers of social behavior in children with autism

In the previous chapter, our study of a group of adults with TD had established that they spontaneously used affectively expressive prosody when speaking to teach a robot. This encouraged us to extend our test of social acceptability to children with ASD. In addition, we compared social behavior while interacting with the robot against that while interacting with another person and with another attractive device, a touchscreen computer game.

In this chapter we present a study of 4- to 12-year-old children with autism spectrum disorders (ASD; $N$=24) during triadic interactions with an adult confederate and an interaction partner, varying in randomized order, among (1) another adult human, (2) a touchscreen computer game, and (3) a social dinosaur robot (E. S. Kim et al., 2013). Children spoke more in general, and directed more speech to the adult confederate, when the interaction partner was a robot, as compared to a human or computer game interaction partner (E. S. Kim et al., 2013). Children spoke as much to the robot as to the adult interaction partner.

This study provides the largest demonstration of social human-robot interaction in children with ASD to date. We find that of the three interaction partners tested, the robot best motivates or facilitates interaction with another person—not just social interaction with objects. This is strong evidence that robots may be developed into useful tools for social skills and communication therapies, specifically by embedding social interaction into intrinsic reinforcers and motivators. This study also indicates, importantly, that the appeal of a technology cannot alone mediate or elicit social behavior in children with ASD; rather, sociality must be entwined with interaction with the technology.

## 3.1 Methods

We designed a randomized, controlled, crossover experiment to compare the effects of interactions with a social dinosaur robot (Figure 3.1) against the effects of interactions with a human or an asocial novel technology (a touchscreen computer game). Each participant in our study completed a sequence of three 6-minute *interactional conditions*, in random order: one in which the interaction partner was a dinosaur robot, another in which the partner was an adult, and a third in which the partner was a touchscreen computer game. All interactional conditions were guided and facilitated by a human confederate (different from the adult interaction partner) and took place in a standard clinical observation room.

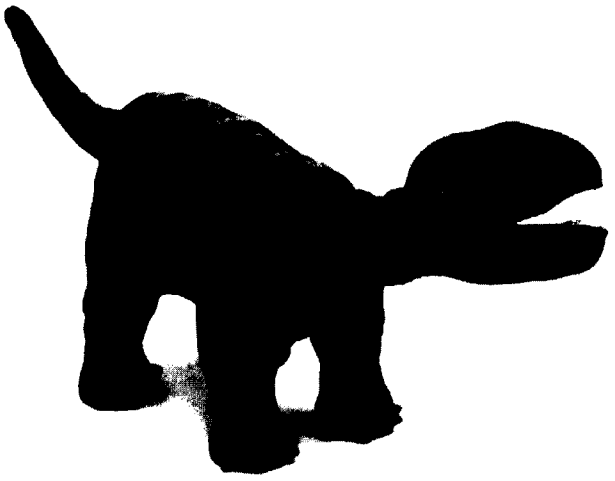


Figure 3.1. The socially expressive robot Pleo. In the robot condition, participants interacted with Pleo, a small, commercially produced, toy dinosaur robot. Pleo is about 21 inches long, 5 inches wide, and 8 inches high, and was designed to express emotions and attention, using body movement and vocalizations that are easily recognizable by people, somewhat like a pet dog. For this study we customized Pleo's movements, synchronized with pseudo-verbal vocalizations, to express interest, disinterest, happiness, disappointment, agreement, and disagreement.

Before the first, after the final, and between interactional conditions, each participant also completed 6-minute, semi-structured *interview-and-play sessions*, which we will also refer to as *interviews*. Interview-and-play sessions gave participants rest from the more structured interactional conditions. They were conducted in another clinical observation room, different from the room where interactional conditions were administered. The interactional conditions and interspersed interviews are described in greater detail below (see Section 3.2).

We expected that children with ASD would find (1) the robot interactional condition social and engaging; (2) the human adult interactional condition social but less engaging; and (3) the computer game interactional condition engaging but not social. Thus we hypothesized that children with ASD would verbalize more while interacting with a social robot than while interacting with either a human adult or a computer game. Given evidence, from case studies (Kozima et al., 2009) and from our own pilot studies, that interaction with

a social robot motivates high levels of curiosity and increases social behaviors such as sharing and excitement with an adult, we also hypothesized that children would direct more speech toward an adult confederate when the interaction partner was a robot rather than when the partner was another adult or a computer game. These hypotheses were intended to support our ultimate goal—to understand the utility of social robots as reinforcers of social interaction with people (as opposed to robots).

### 3.1.1 Participants

Participants were recruited from two ongoing studies at a university-based clinic specializing in assessment, intervention, and educational planning for children with ASD. These included a multi-site comprehensive study of families in which one child is affected by autism, and a longitudinal study of language development in children with ASD. Inclusion criteria included a chronological age of 4- to 12-years and a previous diagnosis of high-functioning ASD (defined as full-scale IQ >= 70 and verbal fluency with utterance production of at least 3 words).

Of the 30 initial volunteers for the study, two were excluded from participation due to below-threshold IQ measurement. Of the remaining 28 participants, four were excluded from analysis: one participant withdrew before completing the procedure; one was excluded for failing to meet ADOS criteria for ASD; and two were excluded due to technical recording problems that precluded speech annotation.

In the 24 participants that ultimately constituted our analytical sample, ages ranged from 4.6 to 12.8 years ($M$ = 9.4, $SD$ = 2.4). IQ eligibility was confirmed within one day of participation in this study using the Differential Abilities Scale (DAS-II: $M$ = 94.2, SD = 11.7, Min = 72, Max = 119; Elliott, 2007). Similarly, within one day of participation in this

study, all participants completed the Autism Diagnostic Observation Schedule—Module 3 (ADOS—Module 3; Lord et al., 2000a) with an experienced clinician and diagnosis was confirmed by a team of clinical experts. Twenty participants met ADOS criteria for autism, and four for autism spectrum disorder. Of the 24 participants for whom analysis is presented in this article, three were female.[4] Twenty participants were white (and not of Hispanic origin), two were black (and not of Hispanic origin), and two were Hispanic or Latino.

## 3.1.2 Materials

### 3.1.2.1 Video recording

All interactional conditions and interviews were recorded using Mini-DV video cameras on stationary tripods from distances of six feet and four feet from participants in the interactional conditions and interviews, respectively.

### 3.1.2.2 Robot, robot behavior, and robot control

The Pleo robot was used in the robot interactional condition because previous investigations have shown that healthy adults (E. S. Kim et al., 2009) as well as children with autism (pilot studies) readily engage socially with this robot. Pleo (Figure 3.1) is an affectively expressive, toy dinosaur robot, recommended for use by children three years and older. It was formerly commercially produced and sold by UGOBE LifeForms; a larger, different model is now produced and sold by Innvo Labs (Innvo Labs, 2012). Pleo measures approximately 21 inches long, 5 inches wide, and 8 inches high. It is untethered, battery-powered, and has 15 degrees of mechanical freedom. We extended UGOBE software to render Pleo controllable by a handheld television remote control, which communicates with Pleo via a built-in infra-

[4] This gender ratio is roughly consistent with reported gender ratios of prevalence of ASD in the United States, of between 4- and 5-to-1 (CDC, 2012).

red receiver on the robot's snout, allowing us to instantaneously play any one of 13 custom, pre-recorded, synchronized motor and sound scripts on the robot. Pleo plays sounds through a loudspeaker embedded in its mouth.

We pre-programmed Pleo with 10 socially expressive behaviors, including a greeting, six affective expressions, and three directional (left, right, center) expressions of interest (to be directed towards nearby objects). All socially expressive behaviors were made up of motor movements synchronized with non-speech vocal recordings. We also pre-programmed three non-social behaviors: a bite (for holding blocks), a drop from the mouth (for letting go of blocks), and a forward walking behavior used when the robot interactional condition called for Pleo to interact with an object that was beyond its reach. Each of these 13 triggered behaviors each endured for less than 2 seconds, and were initiated with the push of Pleo's remote control.

When not executing one of the 13 triggered behaviors, Pleo continuously performed a background behavior designed to maintain the appearance of its animacy. In the background behavior, Pleo periodically shifted its hips, bent and straightened its legs, and slightly nodded its head up and down, or left and right. Robot behaviors, and their carefully matched adult counterparts, are detailed in Table 3-1.

We used hidden, Wizard-of-Oz-style, real-time, human remote control of the robot, a popular design paradigm in human-robot interaction research (Dahlbäck et al., 1993; Riek, 2012; Steinfeld et al., 2009), in order to elicit each participant's belief that Pleo was behaving and responding autonomously. In truth the adult interaction partner, who remained present for all interactional conditions, secretly operated the robot using a television remote control, hidden underneath a clipboard. The Wizard of Oz paradigm affords a robot with the

Table 3-1 Pleo's pre-programmed behaviors. Ten behaviors were socially expressive, including a greeting, six affective expressions, and three directional (left, right, and straight ahead) expressions of attention, and were carefully matched with vague verbalizations in the adult interaction partner. In addition to the ten social behaviors, Pleo had three non-social behaviors (walk, bite, drop), and a "background" behavior to express animacy (i.e., that Pleo takes note of its environment and experiences feelings of boredom or interest). All behaviors were carefully designed to be expressed multi-modally, through vocal prosody, and body and head movement.

| **Social intent expressed, or non-social activity** | **Robot** | | **Adult** | |
|---|---|---|---|---|
| | **Movements** | **Pseudo-verbal vocalization** | **Movements** | **Vaguely verbal vocalization** |
| *Greeting and satisfaction* | Tail wags, head raises. | "Heee!" | Smiles and looks at participant. | An enthusiastic "Hi, <participant's name>!" |
| *Selection of or interest in an object (in one of directions for robot)* | Head lowers toward left, right, or center. | A prolonged, enthusiastic "Ooh!" | Looks in direction of object, points from afar. | "Oooh!" or "That one." |
| *Yes* | Head nods up and down. | "Mm hmm!" | Looks at participant, nods. | "Mm hmm!" or "Yes!" |
| *Enthusiastic Affirmative* | Head raises, tail wags briefly, hips wiggle briefly. | "Woohoo!" | Lifts head slightly or sits moderately upright and smiles moderately at participant. | A slightly moderated "Nice!" or "All right!" |
| *Elation* | A dance: heads raises and moves left and right, hips wiggle, knees bend and straighten. | An extended victory song. | Sits upright energetically, smiles widely at participant, claps hands or puts hands in air. | An extended and exaggerated "Woohoo!" or "Awesome!" or "Fantastic!" |
| *No* | Head shakes side-to-side. | "Unh unh." | Shakes head back and forth and frowns slightly. | "Unh unh" or "No." |
| *Dissatisfaction* | Head and tail lower, mouth opens. | "Ehhh." | Frowns moderately, looks slightly downward, and hangs head slightly. | "Ehhh." |
| *Intense Disappointment* | Head and tail lowers, head shakes slowly from side to side. | A prolonged audible sigh, followed by a whimper. | Slumps in chair or puts chin in hands, hangs head, looks downward. | An audible sigh, followed by an extended, exaggerated "Awwww," or "Oh maan." |
| *Bite* | Head raises, mouth opens for several seconds, then closes. | "Aaaaahhhh...ch omp." | | |
| *Drop from mouth* | Head lowers, mouth opens widely. | <silence>. | | |
| *Walk* | Pleo takes four very short (0.5-inch) steps forward. | "Hup, hup, hup. Hup!" | | |
| *Background Animacy* | Head occasionally moves up and down, and left and right; hips wiggle occasionally; knees bend and straighten occasionally. | <silence>. | | |

appearance of autonomous perception and behavior, with an accuracy and flexibility that currently only humans can produce. Under Wizard of Oz control, the Pleo robot has been shown to successfully impart an appearance of autonomous social interaction, both to adults with typical development (E. S. Kim et al., 2009) and to school-aged children with ASD (pilot testing).[5]

The adult interaction partner was present for all three interactional conditions. In order to obscure the adult interaction partner's manual control of the robot, the confederate explained to participants that the adult partner would remain present for the robot condition, for the purpose of observing the robot's behavior. To maintain consistency with the robot condition, the confederate explained that the adult partner would remain present during the computer game, as well, for the purpose of ensuring that the computer worked. Throughout the robot and computer game conditions, the adult partner stood apart from the participant, confederate, and interaction partner, pretending to read papers on a clipboard and remaining silent unless addressed by the participant (see Figure 3.2). In the robot condition, the adult partner hid the robot's television remote control beneath the clipboard.

It is important to note that most children, including those with typical development, largely or entirely ignored the adult interaction partner during the robot and computer game conditions. Only one participant voiced suspicion that the adult controlled the robot, and subsequently discovered the television remote beneath the clipboard at the end of the robot

[5] Please see our discussion (Section 2.5.3) of challenges to experimental control introduced by our use of manual robot control by Wizard of Oz paradigm.

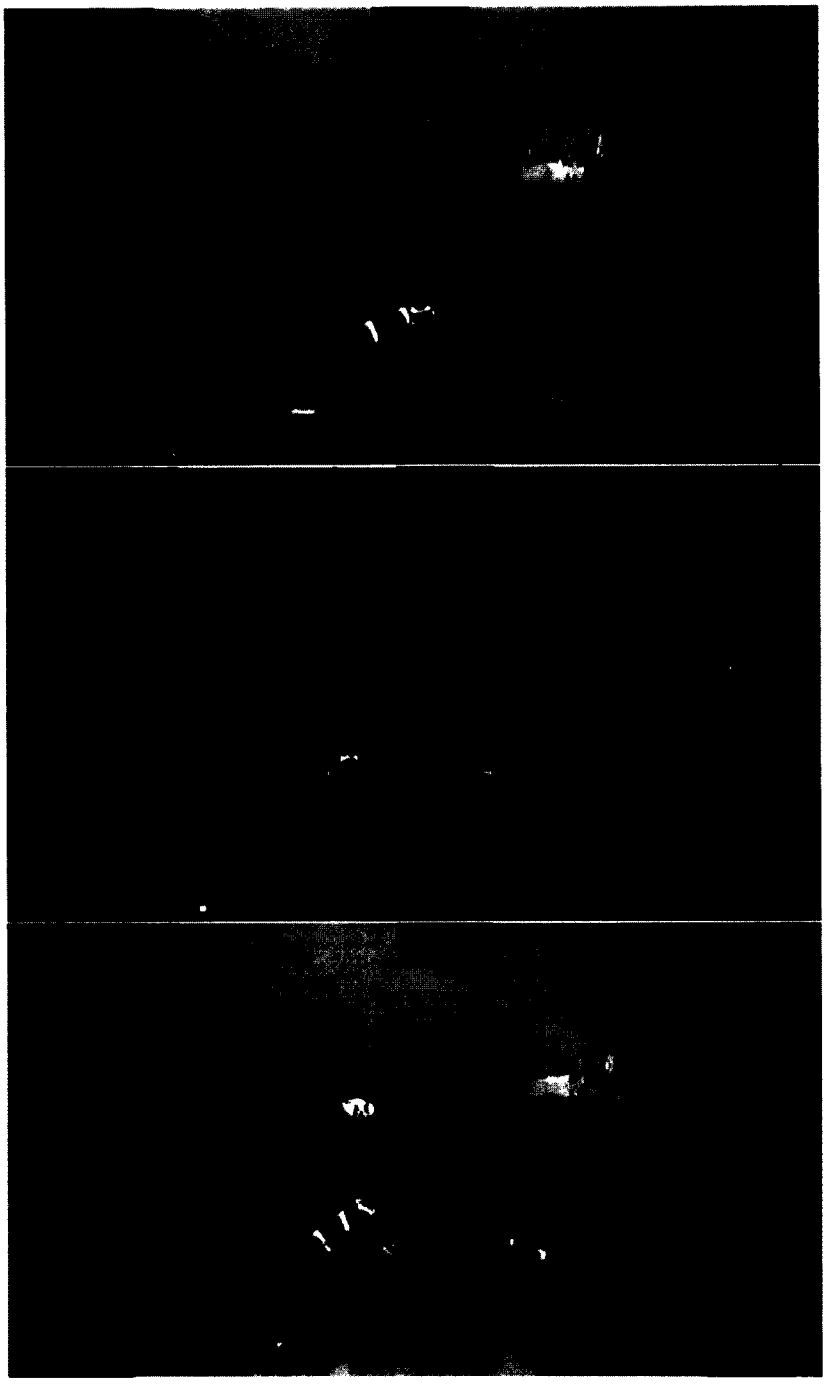


Figure 3.2 Three interactional conditions: adult (top), robot (middle) and touchscreen computer game (bottom). The confederate sits to the participant's right.

interactional condition. We included this participant in analysis nonetheless, because his discovery was made too late to affect his behavior while interacting with the robot.

## 3.2 Procedures

### 3.2.1 Adult and robot interactional conditions

The adult, robot, and computer game interactional conditions were semi-structured and were completed by all participants in randomized orders. Interactional conditions took place on a 3-foot square table, with the participant and confederate sitting at adjacent sides. During the adult condition, the adult interaction partner sat to the other side of the participant, opposite the confederate. For the robot and computer game conditions, the adult's chair was left empty, and the adult stood several feet away from the table with clipboard in hand.

The adult and robot interactional conditions were designed to elicit social interaction, and were semi-structured closely in parallel to each other. The touchscreen computer game interaction was not designed to elicit social interaction, and thus did not match the interactional structure of the adult and robot conditions. In all three conditions, children manipulated blocks: multi-colored, magnetically linking tiles in the robot condition; multi-colored, interlocking blocks in the adult condition; and tangrams, which the participant could move and turn by dragging or tapping the touchscreen with his finger (or a stylus, if preferred) in the computer game condition.

The adult and robot interactions were designed to elicit a host of social perception, reasoning, and interactive behaviors from participants. These included taking turns with the interaction partner; identifying the interaction partner's emotions or expressions of preference for one particular block or another; and shared, imaginative, and tactile play. The confederate's role was to guide the participant through an ordered, standard set of activities and cognitive probes, by subtly directing the adult or robot partner when to deliver pre-scripted cues or affirmations, and by asking increasingly restrictive questions of the

participant. In the robot and adult interactional conditions, one of each of the following probes and activities were completed, in order:

(1 – Probe) The participant presents blocks to the robot or adult interaction partner, and then is asked to identify whether the partner likes or dislikes their colors.

(2 – Activity) The participant assembles the blocks into a structure of his or her own choosing. The participant and partner take turns selecting each next block to add to the structure.

(3 – Probe) During their turns, the adult and robot interaction partners do not manipulate their chosen block directly. Instead, to indicate choice, the adult vaguely points at a block, saying, "That one;" while the robot turns its head to look at a block, saying, "Oooh!" to choose a block. The participant is asked to identify which block the adult or robot has chosen, and then adds the block to the structure.

(4 – Probe) When the structure was completed, the adult or robot interaction partner expressed elation pseudo-verbally ("Woohoo!") and bodily (clapping hands or wagging tail, respectively), as further described in Table 3-1. The participant was asked to identify the partner's emotional state. Next, the confederate removed the blocks from the table, and the adult or robot interaction partner expressed disappointment (as described in Table 3-1). The participant was again asked to identify the partner's emotional state.

(5 – Activity) Pet the robot freely, or invent a secret handshake with the adult partner. In the robot condition, petting was included to give participants an opportunity explore the robot, while in the adult condition the secret handshake game was included to match the robot condition's tactile, interactive, and inventive petting activity. In the secret handshake game, participants were instructed to tap or shake the adult partner's hand in any way they

chose. The adult partner then presented his or her right hand as though to shake hands until the participant made contact, after which he or she exclaimed in delight, and then presented his or her hand open-palmed as if to give a high-five and again expressed delight when the participant made contact a second time. With the robot, participants were offered a chance to guess the robot's favorite spot to be petted. The robot exclaimed in delight after first contact, and participants were then told that the robot had another favorite spot. After being petted a second time, the robot expressed *elation* (happy dance).

Items 1, 3, and 4, above, probed participants' perception and understanding of the robot and adult interaction partners' expressions of affect and preference. Each probe was delivered through a series of increasingly restrictive cues or presses. First the interaction partner would express an emotion or preference (e.g., lowering the head and sighing with prosody expressing disappointment), after which the partner and confederate waited silently for two seconds, giving the participant an opportunity to respond or comment spontaneously. Some participants immediately comforted the robot or adult interaction partner, while others did not respond to the emotional or preferential expression. If participants responded appropriately, the confederate guided the interactional condition to the next activity or probe. Otherwise, the confederate delivered a press, asking the child to interpret the behavior (e.g., "Why do you think Pleo/Taylor said that? How do you think he feels?"). If the participant did not appropriately respond to the confederate's first press, the confederate delivered a second, more restrictive press, offering optional interpretations (e.g., "Do you think he's happy? Do you think he's sad?"). If the participant still did not respond appropriately, the confederate resolved the probe, stating the correct interpretation (e.g., "He seems sad."). Finally, in response to the participant's or confederate's identification of the

interaction partner's emotional or preferential intent, the partner would affirm the correct interpretation by nodding and saying, "Mm hmm!"

The robot and the adult stimuli's social expressions were conveyed using body language, pseudo-verbal or verbal (respectively), and vocal prosodic indications. The adult interaction partner was careful not to explicitly declare his or her communicative intent; for instance, rather than saying, "I feel disappointed," she or he would sigh and say, "Oh, man." (See Table 3-1).

### 3.2.2 Computer game interactional condition

At the time of this study's data collection (Spring through Fall 2010), touchscreen technology was relatively novel, only having recently emerged in consumer products. For instance, the first Apple iPad touchscreen computer was released in April 2010, and by November 2010, there only were an estimated 15.4 million iPhones (all touch-enabled) in use in the United States, out of a total of at least 234 million mobile phones in the U.S. (Dediu, 2011). We structured the computer game condition stimulus to involve little social interaction, in order to evaluate our hypothesis that despite its relatively novel, sophisticated technology intended to match the novelty and sophistication of Pleo's technology, participants would be more socially engaged due to interaction with Pleo than with the touchscreen computer game.

In the computer game condition, the confederate explained the goal of the tangrams game, and showed the participant how to manipulate the tangram objects using his finger, or the touchscreen's stylus if the participant requested, and then stopped initiating interaction, allowing the child to play the game at his or her own initiation and pace. If the participant asked for assistance, the confederate responded verbally or with minimal demonstration to

answer the participant's question. Also, even if the participant did not ask for help but apparently struggled to understand the puzzle, to strategize about a particularly challenging portion of the puzzle, or to manipulate a tile, then the confederate verbally offered assistance. All children were presented with the same three puzzles, in consistent order of increasing difficulty, but were allowed to select alternate puzzles if they requested.

### 3.2.3 Interview-and-play sessions

We interleaved a total of four interviews before, after, and between the interactional conditions, beginning with an interview preceding the first interactional condition. Each participant interacted with a single experimenter, who was different from the adult interaction partner and the confederate, for all four interviews. Interviews maintained consistent, loose structure, and concluded with imaginative play with miniature wooden dolls or with stuffed animal toys, and allowed participants rest from interactional conditions.

### 3.2.4 Dependent variables

We counted the number of utterances participants produced during the interactional conditions, and judged to whom each utterance appeared to be directed. Number of utterances has been shown to be a useful metric in tracking the effects of social and communicative therapies (R. L. Koegel, O'Dell, & L. K. Koegel, 1987; Maione & Mirenda, 2006). An utterance was defined as a verbal production that either expresses a complete proposition (subject + predicate) or is followed by more than 2 seconds of silence. Utterances were transcribed from video recordings by me, and then were confirmed by an independent rater. Following transcription we judged the intended audience or recipient of each utterance to be the confederate, the adult partner, the robot, the computer game, some

combination of the previous, the participant him- or herself, or indeterminable. Judgments of all utterances' recipients were confirmed by an independent rater (agreement was 96%, $\kappa = 0.88, p < 0.0001$).

## 3.3 Results

### 3.3.1 More speech while interacting with robot (Figure 3.3)

A repeated-measures two-factor ANOVA (interactional condition x order) revealed a main effect of interactional condition (robot, adult, or touchscreen computer game) on the total number of utterances produced by each participant within each interaction condition, $F(1.9, 33.4) = 8.13, p < .001$, but no main effect of order of presentation of interactional conditions, $F(5, 18) = 0.46$, and no interaction effect between interactional condition and order, $F(9.3, 33.4) = 1.12$.

One-tailed paired t-tests showed that participants produced more utterances during the robot ($M = 43.0$, $SD = 19.4$) than the adult condition ($M = 36.8$, $SD = 19.2$), $t(23) = 1.97, p < .05$), and more in either the robot ($t(23) = 4.47, p < .001$) or adult conditions ($t(23) = 3.61, p < .001$) than in the touchscreen computer game condition ($M = 25.2$, $SD = 13.4$).

### 3.3.2 More speech directed toward the confederate, when interacting with the robot (Figure 3.4)

The number of utterances directed toward the confederate varied with interactional condition, $F(1.8, 33.0) = 3.46, p < .05$. There was no main effect of order ($F(5, 18) = 0.48$), or of interaction between interactional condition and order ($F(9.2, 33.0) = 0.967$).

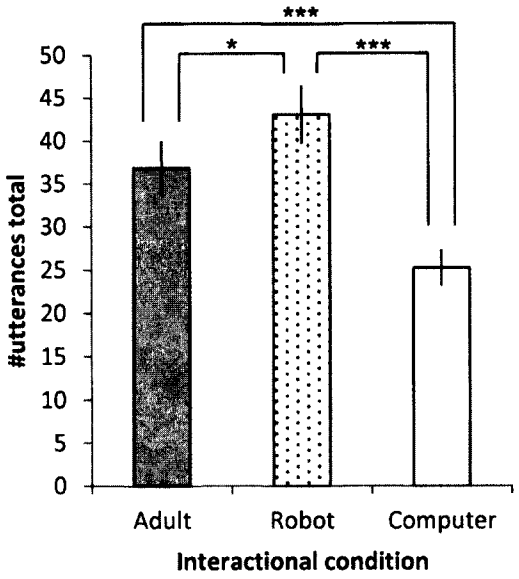



Figure 3.3 Bars show means, over distributions of 24 children with ASD, of total number of utterances produced in the adult (left), robot (center), and computer game (right) conditions. Error bars are ±1 SE. * $p<.05$; ** $p<.01$; *** $p<.001$.

Children with ASD directed a higher number of utterances to the confederate in the robot ($M = 29.5$, $SD = 16.6$) than in the adult condition ($M = 25.5$, $SD = 15.5$), ($t$ (23) = 1.87, $p < 0.05$) and more in both the robot ($t$ (23) = 3.05, $p < .01$) and adult ($t$ (23) = 2.15, $p < .01$) conditions than in the touchscreen computer game condition ($M = 20.5$, $SD = 10.1$).

### 3.3.3 More speech directed to robot and adult than to computer game interaction partner; amount of speech directed to robot comparable to amount directed to adult (Figure 3.5)

A repeated-measures two-factor ANOVA, with interaction partner repeating, revealed that the number of utterances directed toward the interaction partner (robot, adult, or touchscreen computer game) varied with interactional condition ($F$ (1.5, 26.9) = 15.20, $p <$ .001). However, there was no effect of order of condition presentation ($F$ (5, 18) = 0.86, p > .05), or of the interaction between condition and order ($F$ (7.5, 26.9) = 0.50, p > .05).

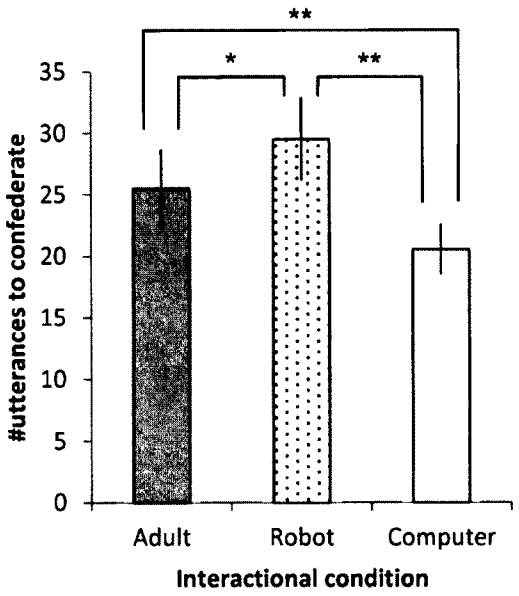



Figure 3.4. Bars show means, over 24 children with ASD of number of utterances directed toward the confederate, in the adult (left), robot (center), and computer game (right) conditions. Error bars are ±1 SE. * $p<.05$; ** $p<.01$; *** $p<.001$.

There were significantly more utterances directed toward the robot ($t$ (23) = 5.40, $p <$ .001; one-sided t-test), and toward the adult ($t$ (23) = 8.22, $p < .001$; one-tailed t-test) than toward the touchscreen computer game ($M$ = 0.5, $SD$ = 0.8). There was no difference ($t$ (23) = 0.02) in the number of utterances directed towards the interaction partner in the robot condition ($M$ = 13.5, $SD$ = 12.0) as compared to the adult condition ($M$ = 13.5, $SD$ = 7.8).

## 3.4 Discussion

We found that children with ASD spoke more, in general, while interacting with a social robot than with another adult or a novel, touchscreen computer game. It should come as no surprise that the robot and adult elicited greater verbal interaction than the computer game, given that the computer game interaction condition was not designed to encourage social

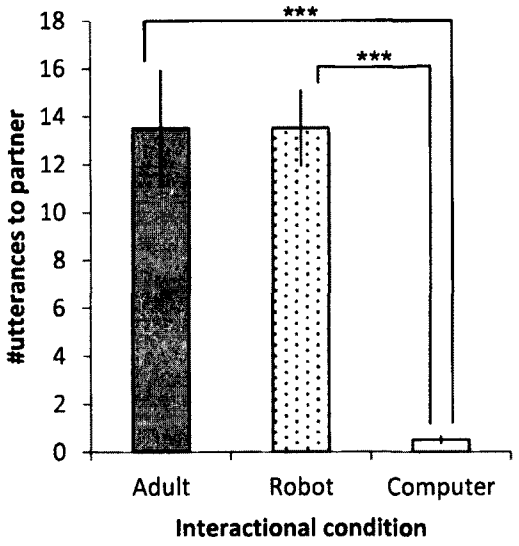



Figure 3.5. Bars show means, over 24 children with ASD, of number of utterances directed toward the adult (left), robot (center), and computer game (right) conditions. Error bars are ±1 SE. * $p<.05$; ** $p<.01$; *** $p<.001$. Participants directed a comparable number of utterances to the adult partner as they did to the robot partner.

interaction. What is most interesting is our finding that a social robot elicits more speech than another human.

Between the adult and robot conditions, we found no difference in the amounts of speech children directed to the adult and robot interaction partners, respectively, and no difference in the number of utterances not directed to anybody. Rather, the increase in total speech found in the robot condition can be attributed to an increase of speech directed toward another adult, the confederate. One possible explanation for the absence of difference in the amount of speech to the robot and to the adult may be that the structure of the associated interactional conditions severely limited the speech the adult was allowed to produce as it was designed to match the limited verbal capabilities of the robot. In this sense, the protocol was designed to support more verbal interaction with the confederate than to the interaction partners.

The robot's greater efficacy in eliciting utterances toward the confederate appears to be due to the excitement and interest (that is, preference) children spontaneously expressed for it over the adult interaction partner. Qualitatively we observed that participants verbalized conjectures and asked questions about how the robot works, whether or not the robot "is real," and what the robot was doing throughout the robot condition. The children also spontaneously asked for permission to, or stated their interest in, touching or playing with the robot. In short, we attribute the robot's greater facilitation of utterances to the participants' greater curiosity about the robot than about the adult interaction partner, during respective conditions.

Heightened verbalization during the robot condition may also reflect the effects of the robot's embedding of social interaction into engagement with it. Our protocol was designed to reinforce interaction with both the confederate and the interaction partner, but as explained previously, the controlled structure of the protocol allowed the confederate greater flexibility in speaking with participants than it did the adult or robot interaction partners. In this sense, this design better reinforced verbal interaction with the confederate than with interaction partners. This may explain why we saw a difference in the amount of speech to the confederate, between adult and robot conditions; and why we did not see any difference in speech to the respective interaction partners.

Our findings suggest potential utility in communication and social skill interventions for children with ASD. The ultimate goal of such interventions for children with ASD is to improve their ability to interact socially. We have shown that interaction with a social robot elicits speech directed socially toward an adult confederate, not just toward the robot itself, and not undirected speech. In other words, of the three interaction partners tested, the robot

best motivates or facilitates an ecologically useful social behavior—interaction with another person—not just social interaction with objects.

This is the first controlled study, over a statistical sample, to demonstrate a social robot's ability to facilitate social interaction with another person. This is also the first study to show this effect for older and higher-functioning children with ASD, whereas previous demonstrations have been presented in small number case studies of younger children with lower functioning (Feil-Seifer & Matarić, 2009; Kozima et al., 2009).

Social robots may draw a comparison with assistive animals, which also elicit social behavior during interaction. It is worth noting that robots have a unique advantage over trained animals in that robots can (1) be highly customizable in form and behavior, (2) therapists and parents can control or (if need be) stop a robot instantly and with ease, and (3) robots can be produced in volume at potentially far smaller cost than that required to train assistive animals.

Previous studies of embedded reinforcers have demonstrated social improvements over the course of lengthy therapy sessions, repeated over several weeks(R. L. Koegel et al., 2009). It is remarkable that the observed increases in verbal interaction afforded by a social robot occur immediately, during interaction with the robot. Further research must be conducted in the long-term durability of social robots' embedded reinforcing effects.

The near absence of participants' speech during the computer game interaction seems obvious and perhaps evokes a straw man, but it demonstrates an important design lesson: an appealing and engaging technology will not elicit social behavior, unless sociality is designed into interaction with that technology.

### 3.4.1 Limitations and future directions

One limitation of this study is that we examined only the quantity, and not the semantic content or communicative function, of utterances under different conditions. A cursory examination suggested that the content of utterances varied across participants. For example, the number of spontaneous comments and questions about the robot (e.g., "Is he real?" "Did you build it?" "Does it have a battery?" "If there was another robot, they would be friends.") ranged from zero to twelve. We plan more sophisticated pragmatic and semantic analyses in the future to better understand the nature of the increases in verbal production that we have observed in the robot over the adult condition.

It is also important to note that short-term effects of interaction with a robot do not necessarily predict long-term effects. This was demonstrated, for instance, in a two-week field trial of school-aged, typically developing children's daily interaction, over with a social robot, in which most children's interactions with the robot declined in the second week (Kanda, Hirano, Eaton, & Ishiguro, 2004). Because any effective therapy requires repeated opportunities to practice target behaviors, our study of short-term effects cannot alone indicate the utility of a social robot a therapeutic tool. Long-term study of motivation, reinforcement, and pedagogical impact are required. While our study cannot speak to long-term effects, our encouraging short-term findings motivate investment in longitudinal studies. We are hopeful that as technology improves, social robots' interactive behaviors will become increasingly complex and adaptable to relationships with individuals. Kanda et al. suggested that children who shared more "common ground" with their robot sustained interaction over time with the robot (Kanda et al., 2004).

Our original intent in comparing a robot with an adult was to compare the robot against an agent operating at the upper limit of social capability. However, the adult interaction partner was unfamiliar to participants. A familiar adult might be considered even more capable socially, with respect to individual participants. Small numbers of children with autism have been observed to prefer interaction with a robot to that with an unfamiliar adult behaving like a robot (Dautenhahn & Werry, 2004; Robins, Dautenhahn, & Dubowski, 2006), and children with autism have also been shown to prefer interacting with their caregivers, to interacting with strangers (Sigman & Mundy, 1989). Our work compared triadic therapy-like interactions with an unfamiliar adult and unfamiliar robot, and with an unfamiliar therapist-like confederate. Our study cannot speak to differences between a robot and a familiar adult, or to triadic interactions with a robot and a familiar therapist. Therefore, the effects of familiarity on interaction with an adult merit future investigation.

We chose the Wizard of Oz robot control paradigm in order to examine responses to a social robot operating at the upper bounds of its social interaction capabilities. We share an aspiration—with many contemporaries in human-robot interaction research—of eventually developing technologies that give social robots truly automatic perception of, and response to, their environments and interaction partners' behaviors. At present, however, Wizard of Oz remains a standard design paradigm, given that state-of-the-art technologies do not yet afford highly reliable automatic speech recognition or other socially important perceptual capabilities, especially not for individuals with widely varying verbal and social abilities and behaviors. Currently, training any automatic perceptual system would be especially difficult, given the vastly heterogeneous presentations of social behaviors we expect to encounter among children with ASD; automatic perception must wait for advances in our

understanding and description of typical and atypical social behaviors (Volkmar & Klin, 2005). Another limitation, as discussed in Section 2.5.3, is that Wizard of Oz control introduces a potential confound if the robot controller is aware of the hypotheses, which was the case in this study. Although not performed, it is possible to validate this methodology, by letting an independent rater examine the robot's behavior for fidelity to the interaction protocol (that is, the script).

Our choice to design the computer game interaction without the social structures built into the adult and robot conditions may evoke a straw man. We chose to contrast the level of sociality, in order to highlight that engagement and motivation to interact with a novel technology cannot in itself support social interactions, unless interaction with the technology specifically delivers sociality.

Previously the benefits had been shown (in small numbers of children with ASD) of using social interaction to deliver a preferred reinforcer (R. L. Koegel et al., 2009). We suggest that social robots may additionally enable a unique type of beneficial embedding, by which social interaction not only *delivers* the preferred reinforcer (e.g., a person presents a child with a robot), but also that the preferred reinforcer *is itself the object and source of social interaction*, not requiring an external social agent to deliver the preferred reinforcer. Social robots may bridge interest in novel technology with motivation for social interaction: if interaction with a social robot itself is rewarding to an individual child, then social interaction more generally may become more rewarding for that child. As technology develops to allow social robots greater and more flexible range of interaction, further research should explore whether they can elicit improved social behavior in children with low social motivation, and can then transfer this behavior to human social partners. Our sample population included

only highly functioning individuals; future research should examine whether social robots offer a unique therapeutic support to children with lower functioning.

Finally, our work is just a first step in the larger goal of providing new tools for clinicians to use in interventions for individuals with ASD, not as alternatives to clinicians or trained peers, but as supplements. The true test of the efficacy of social robotics in facilitating social-communicative improvements in children with ASD will require larger field studies comparing long-term learning and skill generalization in the presence and absence of social robots. These studies are ongoing.

## 3.5 Conclusions

We have demonstrated that a group of school-aged children with ASD are motivated to engage with a social robot, using speech and touch. Further, we have shown that the robot can elicit greater verbalization than a social (but less preferred) interaction partner, an adult human. We have shown that a robot elicits greater verbalization than a preferred but asocial interaction partner, a computer game. More importantly, a social robot increases social interaction with another person, more so than an adult or a computer game. This is the first large group study to show that a robot can mediate social behavior with another adult; previous demonstrations of robot's mediation of interpersonal social interaction have been presented in small-number case studies of younger children with lower functioning (Feil-Seifer & Matarić, 2009; Kozima et al., 2009).

This study's findings suggest that robots may uniquely facilitate social interaction between children with ASD and adults interventionists. They also support the theory that

social robots may provide this support by uniquely embedding sociality into interaction with a powerful reinforcer.

# Chapter 4

# Affective Prosody in Children with ASD and TD toward a Robot

In this chapter we present an original study that demonstrates a quantitatively measured improvement in motivation for human-human social interaction, in a group of school-aged children with ASD, effected by social interaction with a robot, in comparison with a sample of age- and IQ-matched typical controls. We also found that children with ASD and peers with TD enjoy and engage in an interaction structured around the repetitive production of encouraging prosody (E. S. Kim et al., 2012). Our findings show children's motivation to interact vocally with a social robot, suggesting social robots' potential as practice partners in autism interventions. Children with ASD and TD also willingly, repetitively directed affectively expressive speech prosody toward the robot, suggesting that affective prosody may be a viable target behavior in robot-based interventions for children with ASD.

## 4.1 Motivation and research questions

Studies showing successful therapy with visual biofeedback from surface muscular sensors for communication disorders (Andrews, Warner, & Stewart, 1986; Gentil, Aucouturier, Delong, & Sambuis, 1994) initially motivated us to consider using other technology-based

feedback for therapy-like vocal prosody practice. Atypical prosody has frequently been reported as one of the telltale indicators of odd social behavior in individuals with ASD (Paul, Augustyn, et al., 2005). With a goal of determining the viability and utility of incorporating a robot into ongoing experimental interventions and assessments for affective expression in prosody production, we pilot tested a robot interaction in which four school-aged children (two females and twin males, ages ranging from 4.9 to 10.1 years) repeatedly practiced using encouraging prosody to help the Pleo robot (described in greater detail in section 4.1.1) complete a task.

In pilot tests, three participants appeared to exhibit more positive affect during and immediately following interaction with the robot. They also verbally engaged with the robot in repeated trials, producing prosodic and verbal expressions of encouraging affect when interacting with the robot. Two pilot participants also spoke more and engaged in more eye contact with the members of our experimental team, following interaction with the robot. The same two pilot participants also spoke to the robot with heightened variation in prosodic expression of affect. In addition, these participants seemed to make more eye contact and orient themselves to face experimenters more after interaction with the robot. These encouraging social improvements motivated us to examine the statistical stability of such effects, during and immediately after interaction with the robot.

We formulated two hypotheses. First, we expected that children with ASD and those with typical development (TD; that is, a control sample) would equally (a) engage in, and (b) enjoy, interaction with a social robot in a brief, repetitive verbal task. Second, we hypothesized that, more so than controls, children with ASD would show the following improvements in interpersonal social behavior following interaction with the robot: (a)

higher levels of participation in pre-scripted one-on-one interviews, and (b) increased time spent facing the interviewer.

Though our pilot studies suggested improvements in eye contact, we did not measure this behavior due to technical limitations: manual eye-tracking was not possible because of insufficient video recording resolution. Initially we planned a third measure of change in interpersonal social behavior, namely that more so than participants with TD, participants with ASD would increase the variety of types of prosodically expressed affect after interacting with the robot. We have not completed analyses of affective prosody. The differences we observed in pilot testing were remarkable but subtler than can be captured by established five-emotion-category coding. We continue to work to establish stable, reliable measurements to describe the subtler affective variations we initially observed.

To address our two hypotheses, we recruited two comparison groups of school-aged children, a group with ASD and a control group with TD. We designed a three-part protocol, beginning with (1) a semi-structured interview to establish individualized baseline social behaviors, followed by (2) interaction with the robot, and ending with (3) a post-robot interview, used to gauge changes in each participant against his or her own baseline.

Primary dependent variables included Likert ratings of affective valence and engagement with the interviewer or task during robot interaction; and total duration, time spent speaking, and time spent orienting to face the interviewer, in the pre- and post-robot interviews. These measurements are described in greater detail in Section 4.2.4.

## 4.2 Study design and methods

Given our long-term aim to explore robots as intervention supplements for atypical prosody, we designed a robot interaction to provide opportunities for participants to practice encouraging prosody. To test our hypotheses regarding immediate effect following robot interaction, pre- and post-robot interviews were designed to balance natural conversation with controlled parallel structure to allow comparison between the two interviews.

Each participant interacted with the socially expressive robot Pleo (Figure 4.1) for 4 to 8 minutes. Before and after robot interaction, participants completed two brief (3- to 16-minute), parallel, semi-structured interviews.

The interviews and robot interaction all were conducted in a therapy and research examination room, in the presence of an interviewer (a clinically trained research assistant) and another adult who secretly operated the robot (me or another, trained robotics graduate student). Following the final interview, children were offered optional, unstructured time (henceforth, *free play* time) to interact with Pleo.

The interviews and robot interaction were video recorded, and behavioral observations were annotated, following interaction, from these video recordings. When he or she would tolerate it, each participant also wore a lightweight head-mounted boom microphone for future analysis of speech prosody production.

### 4.2.1 Participants

We recruited participants (ages 9 to 14 years) with and without a recent autism spectrum disorders diagnosis, and established two comparative groups of participants, ASD and control, respectively. The ASD group included 18 participants (15 male and three female;

ages ranging from 9.1 to 14.97 years, $M$ = 10.9, $SD$ = 1.7). This gender ratio is roughly consistent with reported gender ratios of prevalence of ASD in the United States, of between 4- and 5-to-1 (CDC, 2012). A 19th participant was excluded from analysis because the robot interaction was interrupted by battery malfunction. The control group (ages ranging from 10.0 to 13.7, $M$ = 11.7, $SD$ = 1.3) included 11 participants (five female) with typical development and one (male) participant with specific language delay but no ASD diagnosis.

Diagnoses of children with a previous ASD diagnosis were confirmed (or ruled out in the case of the participant with specific language delay), using Module 3 of the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000b), by two experienced psychologists at the Yale Child Studies Center, within one day of participating in the present study. Typical development diagnoses were confirmed using clinical judgment the lifetime Social Communications Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003). All participants with typical development scored 8 or lower on the SCQ.

IQ was evaluated for the ASD group using the Differential Abilities Score (DAS; Elliott, 2007) and the Wechsler Intelligence Scale for Children–4th Edition (WISC-IV; Wechsler, 2003); and for the control group, the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). The ASD and control groups were well matched on verbal and cognitive abilities, with all participants having Verbal and Performance (or nonverbal) IQ above 70 (ASD VIQ, $M$ = 102.6, $SD$ = 21.4; ASD PIQ, $M$ = 107.8, $SD$ = 19.5; control VIQ, $M$ = 109.7, $SD$ = 17.6; control PIQ, $M$ = 111.7, $SD$ = 14.2).

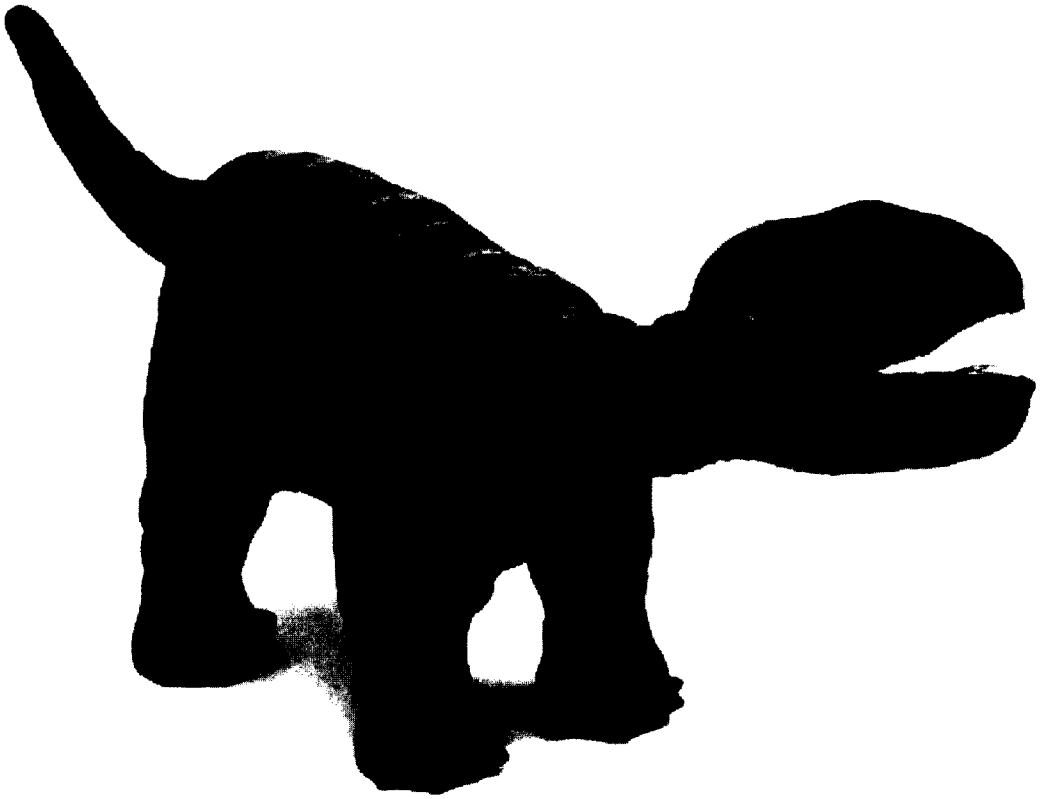


Figure 4.1 In our human-robot interaction study, participants spoke to Pleo, a small, commercially produced, toy dinosaur robot. Pleo was designed to be expressive of emotions and attention.

### 4.2.2 Robot, robot behavior, and robot control

The Pleo robot was used in the robot interaction portion of the study. We were motivated to use the Pleo platform by our past observation that adults with typical development spontaneously use intensely affective prosody when instructed to speak to the Pleo robot (E. S. Kim et al., 2009). Pleo (Figure 4.1) is an affectively expressive, commercially produced, toy dinosaur robot, recommended for use by children ages 3 and up, and measuring approximately 21 inches long by 5 inches wide by 8 inches high. It was formerly produced and sold by UGOBE LifeForms. It is untethered, battery-powered, and has 15 degrees of freedom. We extended third-party software to make Pleo controllable by a handheld

television remote control, through the built-in infra-red receiver on its snout, allowing us to playback any one of 13 custom recorded, synchronized motor and sound scripts. Pleo plays sounds through a loudspeaker embedded in its mouth.

We pre-programmed Pleo with eight socially expressive and three walking behaviors (forward, left, and right). Each behavior included synchronized motor and nonverbal vocal recordings (performed by me). Social behaviors are described in Table 4-1. When Pleo was not executing one of these 11 behaviors, it performed an idling behavior to maintain the appearance of animacy. Pleo's idling behavior included occasional slight hip wiggling, head turning or raising, and subtle tail wagging, all of which were performed randomly in time.

We used Wizard-of-Oz style robot control (Dahlbäck et al., 1993; Steinfeld et al., 2009), allowing participants to believe that Pleo was behaving autonomously, while an investigator secretly manually operated the robot. We chose Wizard-of-Oz style control to fulfill our design objective that the robot should express reliable, contingent social behavior in response to speech. We did not expect that speech recognition technology would be sufficiently reliable with this population to afford highly reliable perception. This is especially the case for the heterogeneous presentations of social behaviors we expected to encounter among children with ASD (Volkmar & Klin, 2005).

In order to obscure the true role of the robot controller to participants, the interviewer instead introduced the robot controller as "Pleo's trainer," who would observe the protocol in order to take care of Pleo and to gauge its progress in overcoming its fear.[6] The robot controller sat in between and about two feet behind the interviewer and the participant.

[6] The role of robot controller was filled by a member of the research team, and was not blind to the hypotheses of the experiment. Please see our discussion (Section 2.5.3) of challenges to experimental control introduced by our use of manual robot control by Wizard of Oz paradigm.

Throughout the protocol, the robot controller sat silently, watching the interview or interaction between the robot and participant, occasionally glancing down at papers on a clipboard. Very infrequently, if the participant or interviewer addressed the robot controller, she or he would respond. During the robot interaction segment, the robot controller used a handheld television remote control, hidden beneath the clipboard, to operate Pleo. The robot controller left her or his seat only to set up or remove Pleo for the robot interaction segment, and freely answered the participant's questions during optional, post-protocol playtime with Pleo.

It is important to note that most children, including children with typical development, entirely or largely ignored the robot controller during the interviews, robot interaction, and optional following playtime. In addition, only one of 31 experimental participants and 5 pilot participants asked whether Pleo was controlled by remote, and neither that participant, nor any other, guessed that Pleo's trainer was in fact controlling the robot.

### 4.2.3 Experimental protocol

We designed our robot interaction protocol to provide opportunities for children to speak to the robot using affectively expressive prosody, with the objective of examining effects on affective prosody toward another person following robot interaction. We were also interested in gauging the effects of social interaction with the robot on other social behaviors that are commonly problematic for speaking children with ASD. These include face-to-face orientation to another person; spontaneous production of topically relevant utterances; indication of interest, or relevant response to, a story told by another person; unusual focus on a topic of special interest; and appropriate expression of emotion using vocal prosody.

Table 4-1 Pleo's eight pre-programmed affectively expressive behaviors. Pleo also was pre-programmed with a forward, left, and right walking behavior, and with an idling behavior to maintain the appearance of animacy.

| **Affect expressed** | **Movements** | **Non-verbal vocalization sounds roughly like...** |
|---|---|---|
| *Greeting or Affirmative* | Tail wags, head raises. | a prolonged, enthusiastic "Hi!" |
| *Fatigue* | Legs bend, head lowers, tail lowers. | an extended, relaxed yawn. |
| *Excitement* | Tail wags vigorously, head rises high, hips wiggle. | "Woohoo!" |
| *Fear and Surprise* | Tail rises rapidly. Then tail lowers, hips quiver rapidly, head lowers. | a high-pitched abrupt "Oh!" followed by a quavering "Ohhh..." |
| *Fear and Uncertainty* | Tail raises, then hips and shoulders quiver, and head lowers. | "Eech!" |
| *Boredom* | Head and tail lower slightly and loll slowly, side-to-side. | a short, aimless, hummed melody. |
| *Enthusiastic Affirmative* | Head raises quickly, tail raises and wags briskly. | "Aye aye!" |
| *Elation* | Head rises, tail raises and wags, hips shake, legs bounce. | a victory song. |

Pre- and post-robot interviews in this protocol were designed to facilitate measurement over these various behaviors. Analysis of these behaviors is ongoing.

### Protocol environment and instructions

Experimental procedures took place in a clinical testing room roughly identical to rooms (or, in some cases, the very same room) in which the participant completed a battery of other assessments and research protocols preceding this experiment. For the entire protocol, including both interviews and the robot interaction, participants sat facing a long table, with the interviewer seated about two feet to the side of the participant (during the robot interaction, the interviewer also served as a confederate, guiding the participant through the interaction.) The robot controller sat between the two, about two feet behind (farther from the table). Throughout the entire protocol, the tabletop was covered with a six-foot-long play-mat, illustrated with "Dino World," a green- and brown-colored jungle scene, striped with a series of four blue rivers. The protocol environment can be viewed in Figure 4.2.

Prior to entering the protocol environment, the interviewer gave participants a brief overview of the protocol's interview-interaction-interview structure. The interviewer also gave detailed instructions for the robot interaction: "After we talk for a few minutes, Pleo will come out. He is a small dinosaur robot. We are training Pleo to get over his fear of water. He will walk across Dino World. But it has rivers, and he is afraid of them. You can help him when he's scared, by talking to him in your encouraging voice. Pleo's trainer will be there, to make sure he's okay and to see how he does." When each participant entered the protocol environment, the interviewer introduced him or her to Pleo's trainer (whose role as the robot controller was kept secret from the participant).

During the interviews, the robot was hidden in an unmarked cardboard carrying case. In pilot testing we observed that the robot's presence distracted children from listening to instructions, suggesting that it would distract them from engaging in interviews as well. We also kept Pleo hidden to control potential effects of familiarization to the robot's presence, between pre- and post-robot interview performance. For the post- robot interview, and if participants asked to play with the robot during the pre-robot interview, the interviewer explained, "Pleo is having a nap now."

**Pre- and post-robot interview protocol**

Interviews were conducted in the same setting as the robot interaction, with the participant seated in front of the play-mat used in the robot interaction. The interviewer sat two feet to the left of the participant, and the robot controller sat between and slightly behind the participant and interviewer.

Pre- and post-robot interviews were semi-structured in the sense that the interview was conversational and allowed the participant to introduce topics of his or her own interest.

However, the interviewer attempted to limit spontaneous discussion, in order to complete a pre-defined series of conversational objectives. As each objective was completed, the interviewer attempted to redirect the conversation to the next.

We designed the pre- and post-robot interviews to be almost entirely parallel in structure to each other, in order to facilitate comparison between the two and to control for confounding variations between the two. Each interview began with an opportunity for the participant to freely talk (for up to three distinct points of new information) about two of three topics suggested by the interviewer (animals, pets, and hobbies in the pre-robot interview, and previous experiences with robots, dinosaurs, and favorite things to learn about in the post-robot interview); a story told by the interviewer about a time when she needed encouragement; and two opportunities for the participant to spontaneously ask what happened next in the interviewer's narrative. The interviewer then asked the participant to discuss a hypothetical or remembered episode in which someone helped, or could help, the participant by encouraging him or her. Finally, the interviewer asked the participant to model or recall (produce) an example encouraging utterance that was, or might be, helpful. Abbreviated examples of prompts delivered by the interviewer, in both the pre- and post-robot interviews are provided in Table 2. The scripts illustrate the parallel structure of the interviews, which was designed to control for conversational content and turn-taking balance when comparing pre- and post-robot social behaviors with the interviewer.

Throughout the participant's conversational turns in the first two interview tasks, the interviewer responded to the participant's utterances. For example, one participant said, "I'm interested in history," to which the interviewer responded, "Yeah, you had World War II history books with you yesterday." To provide opportunities for the participant to show

interest in the interviewer's personal story, the interviewer first paused for three seconds, and if the participant did not comment or ask about the interviewer's story, the interviewer asked, "Do you want to know what happened?"

All interviews were conducted by a research assistant with extensive clinical experience in conducting experimental language and communication protocols with children with ASD. Interviews were 3 to 16 minutes long. The few longer interviews stretched out because of the participant's hesitations to respond or persistent redirection to topics of his or her own interest. A few interviews lasted slightly longer because the participant left his or her seat, at which point the interviewer had to coax the participant to be reseated before resuming the interview.

We controlled for effects of novelty of the first interview and increasing familiarity to the second interview in two ways. First, participants were familiar with the interviewer because our protocol concluded a one- to two-day battery of assessments and experimental protocols, over which she hosted all participants. In addition, the interviewer conducted four of these preceding protocols, including two with experimental protocols featuring brief interview components, the Gray Oral Reading Test (Wiederholt & Bryant, 2001), and an experimental protocol to assess theory of mind.

We controlled for novelty and familiarization to the interview structure by designing our semi-structured interviews to roughly parallel the structure of another longer (30- to 40-minute) experimental protocol, the Yale in vivo Pragmatic Protocol (YIPP; (Paul, 2005)), which all participants completed within one day of, and prior to, our study. The YIPP is designed for children ages 9 to 17 years, and like our interviews, provides opportunities for the child to spontaneously expound on a topic of choice, and to indicate interest in the YIPP

interviewer's stories about herself. Although YIPP interviews were conducted by another clinician (not our protocol's interviewer), the parallel interview structures were intended to control for novelty and familiarity effects of the semi-structured interview format.

**Robot interaction protocol**

The robot interaction protocol was designed to provide the participant with opportunities to direct affectively and verbally encouraging utterances to the robot. The interviewer mediated the robot interaction by providing instructions to the participant, and by reminding the participant to speak, or clarifying Pleo's affective communications to the participant, if he or she hesitated to speak to Pleo.

At the end of the pre-robot interview, the robot controller or interviewer brought Pleo out from its unmarked cardboard carrying case to the start position on the far end of the play-mat, with its face oriented approximately toward the participant. The robot controller or interviewer then sat again. The robot controller remained silent unless the participant or interviewer directly addressed her or him. Participants rarely addressed the robot controller, and the interviewer typically addressed the robot controller only occasionally, when Pleo's feet became caught on the play-mat. While the robot controller placed Pleo at the start of play-mat, the interviewer briefly reiterated instructions to the participant: "Use your encouraging voice when Pleo gets scared of crossing the rivers." At this point, the interviewer introduced the participant to Pleo.

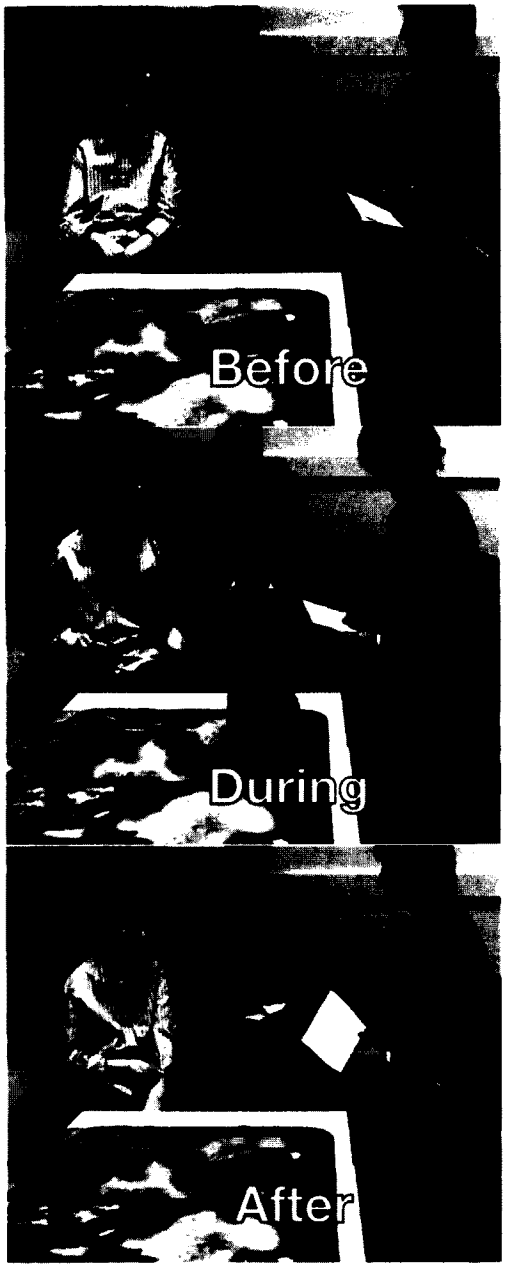



Figure 4.2 Three still images, captured from a video recording of a participant with ASD, showing the pre-robot interview (top), robot interaction (center), and post-robot interview (bottom) within our clinical testing environment. During robot interaction, the Pleo robot walked across the illustrated play mat, toward the participant. Pictured (from left to right) are a participant, the robot controller, and the interviewer. In the post-robot interview, this participant spent 11% more time facing the interviewer than he did in the pre-robot interview. In the ASD group, we found such the size of such increases to be negatively associated with age.

The robot interaction protocol opened with a brief introductory sequence to familiarize the participant with Pleo's communicative capabilities, followed by Pleo's walking across the play-mat toward the participant. For the familiarization sequence, the interviewer guided the participant through two tasks: a greeting to Pleo and a directive to begin crossing the play-mat. The interviewer first instructed the participant to greet Pleo. If after a second prompt the participant would not do so, the interviewer greeted Pleo: "Hi, Pleo!" Pleo responded to the participant's or interviewer's greeting by expressing the behavior *Greeting or Affirmative* in return, raising its head, nonverbally vocalizing a greeting, and wagging its tail (Table 4-1 describes Pleo's eight affectively expressive behaviors in detail). The interviewer then instructed the participant to tell Pleo, "Let's get started!" Again, if after a second prompt the participant would not tell Pleo to begin, the interviewer did so instead. Pleo responded to the participant or interviewer's directive by expressing an *Enthusiastic Affirmative.*

At each blue river painted on the play-mat, Pleo stopped walking and expressed *Fear and Surprise*, to elicit robot-directed speech from the participant. At each river crossing, a series of three increasingly restrictive prompts were delivered by Pleo and the interviewer, to encourage the participant to speak to Pleo. These prompts were structured in the style of errorless teaching, such that any speech toward Pleo was accepted. Following Pleo's initial *Fear and Surprise* expression, after a 3-second pause, if the participant did not speak to Pleo, the robot then expressed *Fear and Uncertainty.* If after a 3-second pause, the participant still did not speak to Pleo, the interviewer told the participant, "I think Pleo is scared. You can help him by talking to him in your encouraging voice." Finally, if after a third 3-second pause, the participant still did not speak to Pleo, the interviewer herself encouraged Pleo, for example, "Don't be scared, Pleo. You can cross the water!" Once the participant or

Table 4-2 Prompts for parallel, semi-structured pre- and post-robot interviews.

| Objective or Task | Pre-Robot Interview Script | Post-Robot Interview Script |
|---|---|---|
| Free exposition (two of three topics) | Do you have any pets at home?<br>Do you have a favorite animal?<br>I used to collect hippo toys. Do you collect anything? | Have you played with a robot before?<br>What do you know about dinosaurs?<br>What do you like learning about? |
| Interest in interviewer 1 | When I was younger, I was afraid to learn to swim, and it caused me trouble. *Pause for 3 seconds.* | I used to love playing video games. One time I got in a bit of trouble because of it. *Pause for 3 seconds.* |
| Interest in interviewer 2pre | My girl scout troupe was planning a canoeing trip, and I was the only one left who hadn't passed the swim test. *Pause for 3 seconds.* | (See 2post below.) |
| No task (interviewer resolves her story) | My dad was really encouraging. He'd say, "Don't worry! You'll be ok. Just give it a try!" That really helped me. | I stayed up past my bedtime one night to beat the game. My brother stayed up with me and encouraged me. He said, "You can do it! Keep going!" I got in trouble for staying up late, but I was happy I beat the game. |
| Interest in interviewer 2post | (See 2pre above.) | I don't really play video games anymore. *Pause for 3 seconds.* |
| No task (interviewer resolves her story) | In the end I got to go canoeing. | My PlayStation broke. |
| Describe encouraging situation | If you were scared to do something, do you think it would help if someone encouraged you? | Do you think it would help if someone encouraged you with something you had trouble with? |
| Model encouraging statement | What kinds of things did/would they tell you to help? | What kinds of things did/would they tell you to help? |

interviewer had spoken to Pleo, whether encouraging or not (e.g., one participant expressed disgust at Pleo's hesitation at the fourth river and said, "Come on, Pleo. It's just water."), Pleo expressed an *Enthusiastic Affirmative*, crossed the river, and then expressed *Excitement.* The interviewer then narrated, with a variation of the phrase, "He did it! I think talking to him helped!"

After crossing the final river, Pleo stepped across a finish line marked with red tape, off the play-mat, and onto the end of the table, inches away from the participant. Pleo expressed *Elation* (a victory song and dance), and the interviewer congratulated the participant on helping Pleo finish his task. The interviewer then explained that Pleo would rest while they spoke (for the post-robot interview), and the robot controller removed Pleo from the table and returned it to its carrying case.

### Free play protocol

Following the post-robot interview, the interviewer asked each participant if he or she would like to play with the robot. Three participants (all with ASD) were not offered free play, due to time constraints. In addition, all but three participants (two with ASD, and one with TD) who were offered free play accepted, and if parents were available, they were allowed to join the free play interaction. Free play was discontinued when participants or parents chose to stop, or when the interviewer determined that the participant was losing interest.

## 4.2.4 Social behavior measurements

The dependent variables in this experiment are measurements of the quality of participants' social behavior. During the robot interaction portion of the protocol we judged ratings of *affective valence*, and of *engagement* in the robot encouragement task (or engagement with the robot or other people). During the pre- and post-robot interviews we annotated, and summed the durations of, brief episodes during which the participant turns his or her head to face the confederate or the robot controller (*face-to-face orientation*); and measured the interviews' durations themselves (*pre- and post-robot interview durations*).

As part of an exploratory analysis, we also measured the duration of the optional free play session, which followed the post-robot interview (*free play duration*).

### Affective valence during the robot interaction

Two raters independently judged the valence of each participant's affect, from video recordings, for 5-second intervals of the robot interaction, judging one out of every four 5-second intervals (or 5 of every 20 seconds). Affective valence was rated on a Likert-type

scale from 0 to 5; where 0 and 1 represented intensely negative and negative affect, respectively; 2 and 3, neutral affect with more negative than positive valence, or vise versa, respectively; and 4 and 5, positive and intensely positive affect, respectively. Inter-rater reliability was measured both as percent agreement and as weighted ***K***, in both cases, allowing raters to disagree by one point. Agreement was 98%, and ***K*** was .78.

### Engagement during the robot interaction

Two raters independently judged video recordings for engagement and compliance of each participant's engagement in the task, or engagement with the robot, the confederate, or the robot controller. Again, these ratings were determined for one out of every four 5-second intervals (i.e., 5 of every 20 seconds) of the robot interaction. Engagement was rated on a Likert-type scale from 0 to 5. Ratings of 0 and 1 represented intense non-compliance and non-compliance, respectively. For example if, during the 5-second interval in question, the participant stood and walked away from the table on which the robot interaction took place, this interval would receive a 0 rating for engagement; or if, in a 5-second interval, the participant hung his head and refused to comply with the interviewer's request to speak to the robot, the interval in question would receive a rating of 1. Ratings of 2 and 3 indicated neither non-compliance nor positive display of interest in the task or with the confederate or robot controller, with more or less reinforcement required on the part of the confederate. For instance, if the participant complied with instructions to speak to the robot, or answered the confederate's questions, but only after several prompts from the confederate, this would warrant a rating of 2; or if the participant required two to three prompts from the confederate before responding or speaking to the robot, even if the reason was interaction with the confederate, this warranted a rating of 3. Ratings of 4 and 5 indicated positive

expressions of engagement with the task or other people. For instance, a rating of 4 was given to intervals in which the participant complied immediately following the confederate's request to speak to the robot or answer a question, or in which no request was made while the robot walked, and the participant maintained their gaze on the robot, or looked at the confederate or robot controller without disrupting the progress of the task of speaking to the robot. A rating of 5 was given if the participant spontaneously engaged the confederate or robot (e.g., created encouraging phrases to the robot which had not been offered as examples by the confederate, or spoke to the robot spontaneously and not only when the confederate had instructed the participant to speak), or changed his or her posture (e.g., leaned forward) to nonverbally interact with the robot. Inter-rater reliability was measured both as percent agreement and as weighted ***K***, in both cases, allowing raters to disagree by one point. Agreement was 95%, and ***K*** was .67.

### Face-to-face orientation during interviews

We did not plan to explore questions about eye contact because we did not expect participants to tolerate wearing automatic head-mounted eye tracking devices, or to remain stationary enough to facilitate automatic table-mounted eye tracking; and because our video recordings were not of sufficient resolution to facilitate manual eye tracking.

Instead, we explored *face-to-face orientation*, the behavior during which a participant turned his or her head to face the interviewer. (Given limited space, the participant's and interviewer's chairs were typically left facing the play mat, roughly parallel to each other, and never arranged such that an angle formed between their front edges would form an angle smaller than 90 degrees). Face-to-face orientation appears to be a novel metric in autism research. We initially considered face-to-face orientation as a surrogate for eye contact, as

these two behaviors overlap in function (it is difficult to make eye contact with someone, without turning to face that person). However, there also appear to be some distinct functions, as well. For example, a listener or speaker may break eye contact in concentration or lower her eyes to reduce the affective intensity of a conversation, but may still orient her face to face the other's. Face-to-face orientation appears to give conversation partners access to one's facial expression, for instance, of affect.

From video recordings, we used VCode software (Hagedorn, Hailpern, & Karahalios, 2008) to mark the beginnings and ends of episodes during which each participant angled his or her head such that he or she was oriented face-to-face with the interviewer. Face-to-face orientation was defined as occurring when the angle between the participant's and the interviewer's faces was smaller than 20 degrees in any direction. More specifically, this angle was defined at the intersection between two vectors, each parallel to the participant or interviewer's line of sight, if the eyes were looking straight ahead. Face-to-face orientation episode markings were verified in VCode, which can synchronize visualizations of episodes and video from which they are annotated. In this paper we analyzed the percent time spent in face-to-face orientation (as a sum over the duration of all episodes, divided by the duration of the video).

Face-to-face orientation was examined for children in the ASD group and a small subset of children in the TD group ($n$=3) in both the pre- and post-interviews. Fewer children were examined in the TD group because of analysis time-constraints and expectations of a wide separation between the ASD and TD group on this measure.

**Interview durations before and after robot interaction**

We annotated the beginning of the pre-robot interview to be the time when the confederate led the participant into the protocol setting (the clinic testing room) and introduced the participant to the robot controller. The post-robot interview began after the robot controller or confederate removed the robot from the table. Both pre- and post-robot interviews ended when the confederate delivered or reminded the participant of instructions for the next segment of the protocol (i.e., robot interaction or free play, following the pre- and post-robot interviews, respectively). The duration of the interview was largely controlled by the participant, and as such provides an easy to calculate surrogate for the child's willingness to continue and elaborate upon the presented interview scenarios. Note that this measure is not without its complexities, a point we will return to in the discussion.

## 4.3 Results

Like typically developing controls, participants with ASD had no difficulties engaging with the robot as indicated by engagement ratings (TD: $M$ = 4.36, $SD$ = .50; ASD: $M$ = 4.27, $SD$ =.62; $t(27)$=.39). Likewise, affective valence during the robot interaction was similar between groups (TD: $M$ = 3.68, $SD$ =.63; ASD: $M$ = 3.60, $SD$ =.69; $t(28)$=.33).

There was also no difference in the amount of time children with ASD spent in the pre- or post-robot interview (pre: TD: $M$ = 267s, $SD$ = 57s; ASD: $M$ = 286s, $SD$ = 108s; $t(27.8)$=.65; post: TD: $M$ = 312s, $SD$ = 72s; ASD: $M$ = 395s, $SD$ = 199s; $t(21.7)$=1.6), nor any between-group differences in additional time spent in the post-robot interview as compared to the pre-robot interview (i.e. $time_{delta} = time_{post} - time_{pre}$) ($time_{delta}$: TD: $M$ = 45s, $SD$ = 85s; ASD: $M$ = 86s, $SD$ = 166s; $t(25.1)$=.85). However, within-subject, paired

comparisons between the time spent in the pre- and post-robot interviews indicated a significant increase of children with ASD ($t(16)=2.13, p=.05$) but not for TD children ($t(10)=1.7, p=.11$). The ASD group also spent significantly longer than the TD group playing with the robot during free play (TD: $M$ = 207s, $SD$ = 49s; ASD: $M$ = 307s, $SD$ = 137s; $t(20.3)=2.7, p=.02$, Cohen's d=0.97).

Children with ASD, as compared to TD children, appeared to face the interviewer less in both the pre- (TD: $M=.76$, $SD=.34$; ASD: $M=.30$, $SD=.26$, d=1.52) and in the post-robot interviews (TD: $M=.85$, $SD=.14$; ASD: $M=.31$, $SD=.23$, d=2.84). Because of the small $N$ of the TD group for this measure, the group difference in face-to-face looking ratio was not significant ($p=.14$); even so the post-robot group difference was highly significant ($p<.01$). Paired t-tests (that is, repeated measures) analysis among individuals in the ASD group showed no change in face-to-face looking ratios from the pre- to the post-robot interview.

An exploratory analysis of the cognitive and behavioral associations with primary outcome variables in ASD indicated trends such that those participants, with smaller increases in pre- to post-robot interview duration ($time_{delta}$), displayed more negative affect during the robot interaction (affect x $time_{delta}$: $r=.48, p=.050$), worse language skills (CELF language standard score x $time_{delta}$: $r=.50, p=.042$), and greater levels of social and behavioral impairments (ADOS total: $r=-.49, p=.055$). In children with TD, the affect relationship was not observed (affect x $time_{delta}$: $r=.14$), but the same direction of the language relationship was suggested (CELF language standard score x $time_{delta}$: $r=.48, p=.14$). No ADOS scores were available for the TD group. For time spent face-to-face in children with ASD, the change from the pre- to post-robot interview was negatively associated with

chronological age ($r = -.66, p < .01$); this relationship with age was not noted for either the pre- ($r = .33, p = .25$) or post-robot interview face-to-face ratios ($r = -.22, p = .45$).

## 4.4 Discussion

### 4.4.1 Summary of results and limitations

The results of this study confirm our first hypothesis, that children with ASD and their typically developing peers engage and enjoy verbal interaction with the robot to similar extents. Our observations only partially support our second set of hypotheses: a) compared to the TD group, children with ASD spent more time in the post-interview process, but b) did not show a greater increase in face-to-face orientation. We will discuss the implications of these findings in turn.

It is clear that children with ASD were motivated to interact with the robot, as indicated by their greater predisposition to spend time with the robot (compared with children in the control group) when given an option to play with the robot freely subsequent to the post-robot interview. Although two of the three participants who opted out of playtime had an ASD, in one case it appears he exhausted himself by talking at great length during the post-interview about the robot with the interviewer. In a second case, the participant was suspected of having comorbid diagnosis with oppositional defiant disorder, and was generally reluctant to participate in all phases of this and other experiments during his visit. In the case of the one child with TD who opted against free play, she simply seemed uninterested in playing with the robot. During free play, some participants in both the control and ASD group showed great ingenuity in understanding the robot's "water"-sensing mechanisms or logic, or in exploring what the robot enjoyed, feared, or understood. For

example, participants frequently hypothesized that the robot was programmed to fear the color blue, which it sensed through the camera on its snout. Several participants tested their hypotheses by finding blue objects in the room and holding them up to Pleo's snout.

The results of our study, which confirm our first hypothesis, add to the mounting evidence that robots may be a highly tolerated, and even enjoyable, component of intervention for children with ASD. It is important to note that during the robot interaction phase, children with ASD showed similar levels of affective enjoyment and engagement as typically developing children. By comparison, in natural social situations and under laboratory testing conditions, children with ASD often exhibit limited affective response (Joseph & Tager-Flusberg, 1997; Kasari, Sigman, Mundy, & Yirmiya, 1990; Yirmiya, Kasari, Sigman, & Mundy, 1989).

The observation of an increased change in time spent in the post-robot interview session for children with ASD as compared to TD children (our second hypothesis) suggests that interaction with the robot may lead to greater verbal elaborations or increased verbal participation by children with ASD. However, it is important to note the limitations of this very coarse measure. First, though the variability of the interview duration is largely controlled by the participant, the interview itself is pre-scripted and pre-planned. For this reason, there is a limit to how much leeway each child can be afforded in terms of true "back-and-forth" verbal exchanges with the interviewer. Second, many of the questions can be answered quite succinctly (e.g. "Do you have any pets at home?"); to spend additional time in these phases of the interview may suggest difficulty understanding the question or may result from the expression of incoherent, meandering, or off-topic responses. Third, though the structure of the pre- and post-robot interaction interviews was designed to be

parallel, we cannot rule out the possibility that specific tasks may be more accessible to participants in one group versus the other. For example, the "encouraging situation" question (Table 4-1) may require access to long-term memory in the pre-interview, but in the post-test an example might be readily accessible from the robot-interaction phase. Of course, the design of conducting two parallel interviews sequentially over a short period of time may in of itself bias results, as participants may become more comfortable with increasing interactions to the interviewers. The measure of "total interview duration" thus coarsely suggests the behavioral changes resulting from robot interaction may be found for individuals with ASD, but does not isolate the mechanism, nor provide an unambiguous description of causal relationships or quality of the responses. Further fine-grained analyses of the video recordings will have to be conducted to decipher the underlying structure responsible for increased changes in pre- to post-robot interview duration; an expanded study, consisting of repeated exposures over multiple sessions, would be necessary to gauge the generalizability and repeatability of our observations.

Along these lines, a close examination of the variability associated with the outcome measures examined in this study suggested a wide heterogeneity of responses in the ASD group. In an effort to decode this variability, we examined correlations between clinical features and performance metrics. The results suggest that those children with fewer social and communicative deficits responded to the robot interaction with greater enthusiasm, as reflected by increased time in post-robot interviews and higher affect ratings while interacting with the robot. This suggests that while, as a group, children with ASD behaved similarly to those with TD for most aspects of the robot interaction, responses to the robot interaction were modulated by the degree of socio-cognitive impairments. However, these

relationships may also suggest that floor effects could exist in the outcome measure of total interview duration. In other words, the positive relationship between language skills and increases in interview duration post-robot interaction suggest that it is the more verbally capable participants with ASD who may be responsible for the observed between-group (TD vs. ASD) differences in increased interview time; conversely, the children with lower verbal ability may be stretched to their capacity in both the pre- and post-robot interviews. Again, these results highlight the need to carefully examine the relationships among hypotheses, outcome measures, and the individual characteristics of participants in interpreting the results of interactions between robots and children with ASD.

It is also clear, however, from our analysis of face-to-face interactions, that interacting with the robot does not generally result in increased orienting towards the face for our participants with ASD. Though children with ASD exhibit the expected decreased orienting to the face before the robot interaction, the frequency of their decreased looking remains virtually unchanged post-robot interaction. Though from a certain point of view this result is disappointing, from another point of view the result is quite understandable. In the state in which this study was conducted (Connecticut, US), the standard of care for individuals with autism is quite high. In fact, a recent study of community and standard care practices in toddlers with ASD suggests that treatment-as-usual now produces results that are competitive with more specialized intervention programs (Steiner, Goldsmith, Snow, & Chawarska, 2012). It may be optimistic to assume that an extremely brief guided interaction with a robot might be able to effect a change on one of the most highly-targeted behaviors for individuals with ASD: eye-contact and natural conversation. However, examination of the relationships between change in face-to-face looking pre- to post-robot interview

showed a prominent negative relationship with age, suggesting that the paradigm as a whole might be more suited, and more effective, for younger children with ASD.

While there are many other aspects of this data that can be examined, especially regarding the degree to which children utilized appropriate prosodic intonation and their production of socially appropriate behavior, here we report only on a subset of possible measures that 1) point towards potential effects that may be due to engagement with the robot, 2) further our understanding of the applicability of the design across the heterogeneity of the autism spectrum, 3) are immediately available and accessible, and 4) illustrate points regarding clinical-HRI partnerships. In the case of this study, the working agreement we have with our clinical partners is to first publish the results of the study here, within a robotics-oriented venue, while preparing for additional analyses that will clarify and solidify our understanding of our data.

### 4.4.2 Summary of contributions

We found that school-aged children with high functioning ASD and with TD engage and enjoy verbal interaction in a repetitive task to produce encouraging prosody. Participants" enjoyment during this task indicates continued exploration of social robotic applications to social skills and communication therapies for children with ASD. In the literature examining social human-robot interactions, this study represents the largest comparison between children with ASD and peers without ASD. This study is also the first group demonstration of HFA children's engagement in prosody-therapeutic interaction with a robot.

Additionally, compared to the control group, children with ASD spent more time engaging with an examiner during a post-interaction interview. This may suggest that interaction with a social robot helps to catalyze or promote immediately following

interaction with other people, in a way that is uniquely heightened among children with ASD.

We also innovated face-to-face orientation as a novel measure of social engagement, which we observed far less in participants with ASD than in their peers with TD. This measure may be useful for gauging and tracking social skills in future studies of robot interaction and general social interaction among individuals with ASD.

# Chapter 5

# Automatic recognition of communicative intentions from speech prosody

In this chapter we describe methods for automatic recognition of communicative intentions in speech prosody, as well as a closed-loop system that uses affective prosody as an input for learning. We have shown in Chapter 2, Chapter 3, and Chapter 4 that a manually operated robot can elicit social behaviors, enjoyment and motivation from children with ASD. The participant's experience of a single, brief interaction with a robot is apparently unaffected by the fact that the robot's perception and responses are not automated. However, automation of such perception and responses would afford several advantages. First, automation can relieve labor burdens, which will eventually be important in the successful deployment of socially assistive robots, particularly in behavioral interventions, which tend to require at-least-weekly, repeated visits over several months. A robot controller's role could be eliminated by automating a robot's perception and selection of behaviors. Automatic perception could support clinicians' and caretakers' assessments and observations of a child, reducing the amount of time they spend assessing the child's behaviors, allowing them, for

instance, to move to successive trials more rapidly, to spend more energy on building rapport than on evaluating behaviors, or to spend less time reviewing intervention sessions to assess the child's performance). Closed-loop automations that uses automatic perception to drive automatic response (i.e., behavior selection) could further reduce labor burdens by making robots available as practice partners outside the clinic or perhaps without little or no supervision. All these labor savings could make robots more accessible to clinicians and caretakers with limited resources. Second, computational modeling and automation tend to be mutually supportive endeavors. Such modeling may deepen our understanding the behaviors in question, contributing theoretical and practical insights. As described in Question 2, such insights may support the selection of target behaviors for intervention or our assessment of behavioral performance. Finally, investigations of systems that learn from social behavioral input may inform us about the adaptations users expect from robotic partners. This knowledge may inform designs for extended human-robot relationships that can support long-duration therapies (see Question 4).

The systems presented in this chapter automatically recognize (1) affect and (2) shared belief management cues from speech prosody. The first system also integrates its prosody percept as an input for machine learning. It first recognizes the affective valence (positivity or negativity) expressed in motherese-like, robot-directed speech, and then uses this to drive a reinforcement learner. This system's efficacy is demonstrated over a toy learning problem, for a single speaker (myself) who is knowledgeable of the experiment's goals and of the prosody recognition and learning systems. Despite the obvious lack of experimental control, and the triviality of the learning task, the system we present here demonstrates a proof of concept of an interface that learns from speech prosody.

The second system we describe in this chapter identifies cues that speakers use to add new information, compare against previously stated information, or suggest questionable information shared between herself and the listener (E. S. Kim et al., 2008). The system uses acoustic rules originally described for adult-directed speech (Beckman & Elam, 1997; Pierrehumbert & Hirschberg, 1990) to identify the same cues in a corpus of infant-directed speech. Blind human ratings of the communicative intent of utterances in the corpus validated the acoustic classification.

Given the challenges of working with special needs populations and children, and the heterogeneity of prosodic deficits found among individuals with ASD (for example, some speaking in a sing-song with others speaking in a monotone; Grossman, Bemis, Plesa Skwerer, & Tager-Flusberg, 2010; Paul, Augustyn, et al., 2005; Tager-Flusberg et al., 2005), the systems we present in this chapter automate prosodic expression recognition in the speech of typically developing adults.

In addition to describing both systems, we present demonstrations of their perceptual performance and of the first system's learning performance, and we discuss implications on interventions for atypical prosodic production in ASD.

## 5.1 System 1: Learning from affective prosody

The first system recognizes affective valence from robot-directed speech prosody in real time. In E. S. Kim & Scassellati (2007) we showed that this affective valence signal can be used as a real-time feedback mechanism to a robotic learning system. Affective expression is among the communicative intents frequently abnormally expressed by some individuals with ASD (Paul, 2005; Tager-Flusberg et al., 2005), and has been targeted for intervention.

We constructed a reinforcement learning system for our humanoid robot Nico (e.g., Doniec, Sun, & Scassellati, 2006; G. Sun & Scassellati, 2005), which uses prosodic feedback to refine the parameters of a social waving behavior. We defined a waving behavior to be an oscillation of Nico's elbow joint, parameterized by amplitude and frequency. Our system explored a space of amplitude and frequency values, using q-learning (Russell & Norvig, 2003, pp. 775–777) to refine the wave behavior until it optimally satisfied a human tutor. To estimate the tutor's affective valence as feedback in real-time, we first segmented speech from ambient noise using a maximum-likelihood voice-activation detector. We then used a k-nearest neighbors classifier (Russell & Norvig, 2003, pp. 733–736), with k=3, over 15 prosodic features, to estimate a binary approval/disapproval feedback signal from segmented utterances. Both our voice-activation detector and prosody classifier were briefly trained on the speech of the tutor. We showed that our system learns the tutor's desired wave, over the course of a sequence of trial-feedback cycles. We demonstrated our learning results for a single speaker on a space of nine distinct waving behaviors.

## 5.1.1 Introduction

### 5.1.1.1 Socially-guided machine learning

Evidence matches intuition that, as with people or pets, long-term relationships with robots are best maintained if the robots adapt to their human counterparts. Even without human-level knowledge or understanding, pets are able to maintain intimate relationships with their human counterparts, probably in large part on the basis of their responsiveness to social information, including touch, speech prosody and spoken words, and body language and nonverbal gestures. Beyond recognizing verbal and non-verbal communications, a robot should respond, adapt to, learn from that which is communicated (it's only so helpful for a

dog to hear that his person is forbidding it; much more helpful if the dog can learn not to dig in the garbage can). Much work has been done to recognize communicative modalities, like speech, nonverbal gestures, gaze direction. And much work has been done in machine learning from physical environments, for example, to map a novel environment or to learn the affordance of a tool or object.

Besides being important for adapting to new information in the environment (e.g., avoiding danger, keeping a secret upon a stranger's entrance to the room), learning in response to human social cues is important for human-robot cooperative tasks.

A recent exploration into human-guided machine learning has revealed that a simulated robot can learn a simple sequential task, such as a cleaning up a virtual kitchen, given feedback from a human tutor. In Sophie's Kitchen, a tutor communicates using a mouse to scroll a feedback meter between extremes of strong approval and strong disapproval (Thomaz, Hoffman, & Breazeal, 2006).

The present work extends the exploration of human-guided machine learning into the physical world, where a robot learns to modify its behavior, given a more naturally social human communication: speech prosody.

#### 5.1.1.2 Communicating prosodic affect to robots and computers

Speech prosody is essentially "tone of voice." It is comprised of the highness or lowness, the scratchiness or smoothness, the loudness or softness, and the quickness or slowness, with which a speaker can alter their pronouncement of an utterance. Functionally, while prosody also communicates syntactical and pragmatic information, in the present work we are concerned with its function as a mode for communicating emotions and attitudes, or affect.

Humans modulate their tone of voice to communicate affect. We raise our voices in frustration, or comfort small children using hushed speech. We use consistent tones of voice to indicate displeasure or joy to our pets. In the last decade, numerous studies have shown that, with varying degrees of constraint and accuracy, affect can be classified automatically from recordings of speech (Breazeal & Aryananda, 2002; Liscombe, Venditti, & Hirschberg, 2003; Robinson-Mosher & Scassellati, 2004; Slaney & McRoberts, 2003).

In the present work, in response to affective prosody, we extend beyond hard-coded expressive postures to using prosodic affect recognition to drive a system, which learns to refine the social behavior of waving.

### 5.1.2 Refining behavior using prosodic feedback

We have implemented our prosody-driven learning system on our humanoid robot Nico, within Nico's lab environment. Our learning system is trained using an interaction loop, shown in

Figure 5.1. For each iteration of the interaction loop, Nico performs a waving behavior, after which it waits a pre-determined amount of time for a possible utterance of feedback. If the tutor utters a response, the affect of the utterance is calculated, producing a binary approving/not-approving result. This binary approval signal is the feedback that drives the q-learning system. Nico iterates through the interactive loop until the q-learner fixates for some pre-selected number of cycles on a single waving behavior, which Nico estimates to be the goal behavior.

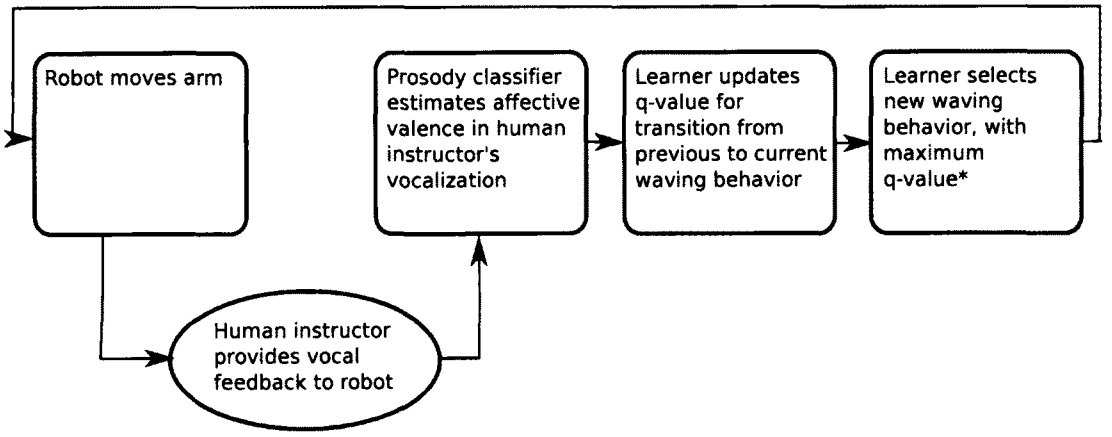



Figure 5.1 Interaction loop flow for prosody-driven learning. This loop iterates until the robot selects the same waving behavior a pre-determined number of cycles in a row, at which point it declares that behavior to be the goal behavior. Nico's estimate of prosodic affect takes the form of a binary approval/not-approval signal.

#### 5.1.2.1 Infant- and robot-directed speech

Nico, shown in Figure 5.2, is an upper-torso robot, built in the proportions of a one-year-old infant. Nico is equipped with a seven degree-of-freedom neck and head assembly, and a six degree-of-freedom arm. Nico wears a fixed smile and infant clothing, encouraging humans to interact with it socially. We make a fundamental assumption regarding human interaction with Nico: we assume that people will interact with Nico as though it is a small child or an infant, speaking to it using exaggerated prosody. This is because it is easier to classify affect in the prosody of infant-directed speech (Liscombe et al., 2003).

Speakers tend to use prosody with exaggerated features when speaking to infants. This speaking style is frequently termed Motherese (Fernald, 1989). Compared with adult-directed

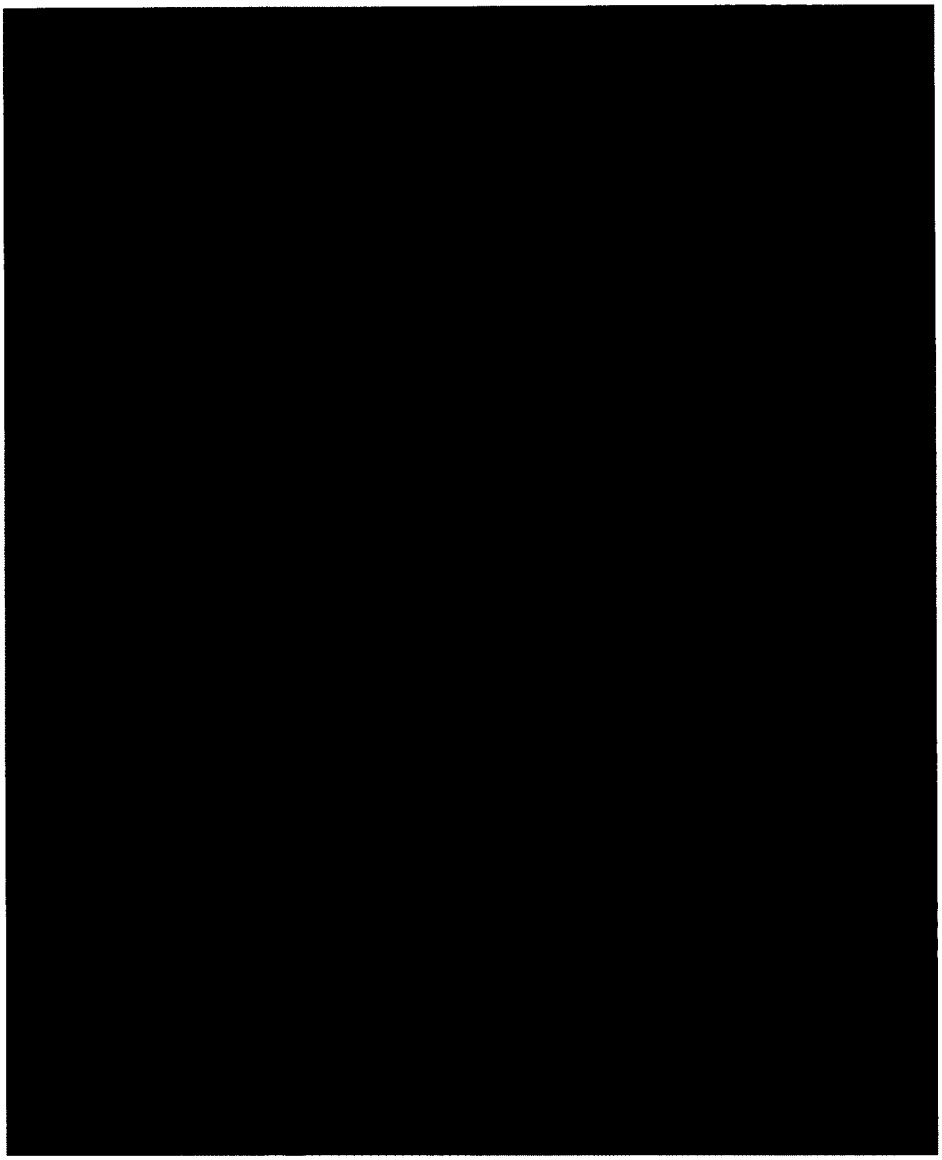


Figure 5.2 Our humanoid robot Nico, waving. Nico has a torso built to the proportions of a 50th-percentile, one-year-old human infant.

prosody, infant-directed prosody features higher pitch and wider pitch range (Fernald & Simon, 1984; Garnica, 1977; Menn & Boyce, 1982; D. N. Stern, Spieker, Barnett, & MacKain, 1983), and longer vowels at phrase (Morgan, 1986) and clause boundaries (Ratner, 1986). Breazeal and Aryananda observed that people tend to extend their use of Motherese

to their robot Kismet, a humanoid with facial features designed to appear childlike (Breazeal & Aryananda, 2002).

#### 5.1.2.2 Interaction environment and audio capture

Nico's tutor's utterances are recorded in a real-time interaction loop, coupled with Nico's actions, within our lab environment. Acoustically, the lab environment is extremely noisy, given the unavoidable proximity of a rack of computers controlling Nico's motor and visual systems. The tutor's speech is recorded at a sampling frequency of $F_S = 22050Hz$, using a mono-input microphone clipped to the tutor's clothing, within six inches of the tutor's mouth, to increase signal energy, given high environmental noise.

Following acknowledgement from the robot's motor controllers that Nico has finished performing its waving behavior, three seconds of audio are recorded, within which time the tutor has presumably responded to Nico's movement.

#### 5.1.2.3 Overview of affective prosody recognition

We estimate prosodic affect within each audio response clip as follows:

1. We first cut the response clip into overlapping, short-time analysis windows, each 25ms long. The start times of neighboring windows are separated by 10ms. Short-time windowing is necessary for spectral analysis of auditory data, in order to employ notions of stationarity in frequency for any temporal segment. These short-time windowing values are standard in speech recognition (Quatieri, 2002).

2. We perform voice-activation detection (VAD), checking each short-time window for speech. We then concatenate consecutive windows to form continuous speech segments, smoothing over brief inconsistencies in VAD output.

3. We estimate the prosodic affect in each speech segment, and send this estimate to the waving behavior learner.

#### 5.1.2.4 Speech segmentation

We use a VAD to segment this three-second response clip to isolate short-time (10 ms-separated, 25 ms-long) analysis windows containing speech. Windows are derived by multiplying each three-second response clip $x[n]$ by a Hamming windowing function:

$$x_l[n] = x[n]w[n,\tau] \tag{5.1}$$

defined as

$$w[n,\tau] = \begin{cases} 0.54 - 0.4\cos\left[\dfrac{2\pi(n-\tau)}{N_w - 1}\right], & \text{for } 0 \le n \le N_w - 1 \\ 0, & otherwise \end{cases} \tag{5.2}$$

where $N_w$ is the number of samples in the window (Quatieri, 2002, p. 62). At a recording sampling frequency of 22050 Hz, $N_w = 551$. For each window, our voice-activation detector conducts maximum-likelihood detection over three features calculated over a short-time window of the acoustic signal $x_l[n]$:

1. total energy over the window

$$energy_l = \sum_{1}^{N} (x_l[n])^2 \tag{5.3}$$

2. variance of the log-magnitude-spectrum

$$vlms_l = \mathrm{var}\left(\log\left(\mathrm{abs}(X(j\omega))\right)\right) \tag{5.4}$$

3. variance of the log-spectral-energy

$$vle_l = \text{var}(\log(X(j\omega)^2)) \tag{5.5}$$

where $X(j\omega)$ is the discrete-time Fourier transform of $x_l[n]$.

The VAD is trained on auditory data recorded from the tutor's voice before the interaction loop begins. For our learning system, we trained on 15 seconds of continuous, ambient noise, and 8 seconds of continuous, uninterrupted speech from the speaker. Voice-activation detection performance is described in Section 5.1.3.

#### 5.1.2.5 Functional, perceptual, and acoustic properties of speech prosody

Prosody is the music of speech. It is manifested in variations of psychoacoustic percepts of pitch, loudness, duration of syllables and pauses, and voice quality (e.g., hoarseness or voiceless whisper). In English prosody is somewhat determined by linguistic considerations, such as stress on syllables within words, and question versus non-question information. Otherwise, English prosody flexibly conveys paralinguistic or nonlinguistic information, such as the speaker's intention or attitude, and mood or affective state (Mixdorf, 2002).

Pitch is the highness or lowness of the voice, sometimes called the tune or melody of an utterance. Pitch is a percept that roughly correlates acoustically with the fundamental resonant frequency *f0* of voiced phonemes, including vowels, nasals (the sounds of the letters "m" and "n"), voiced obstruents (e.g., "b") and approximants (e.g., "l"). Most adult male voices vary in pitch over frequencies from 50 to 300 Hz. The pitch of adult females and children can range from 150 to 1000 Hz (Quatieri, 2002). The physical correlate of pitch is the fundamental frequency (*f0*) of a periodic signal, and the physical correlate of volume is acoustic energy (Quatieri, 2002).

#### 5.1.2.6 Classification of prosody by *k*-nearest neighbors

Our prosody classifier decides whether or not an utterance indicates approval. We choose to map utterances to a simple, binary approving/not-approving signal because such a binary signal can apply to various machine learning contexts, and to simplify affect classification.

This system uses pitch and volume features, as we presume that for the purposes of providing approving and disapproving feedback, the tutor will tend to produce consistently short utterances in a consistent tone of voice. We have designed our classification features based on those used by Robinson-Mosher and Scassellati in the same noisy lab environment. Our 15 features are comprised of statistics derived from estimates of from pitch, energy, and energy-weighted pitch. Each of these measurements is estimated for each short-time window in the speech segment.

We estimate *f0* using a cepstral method (Noll, 1967; Quatieri, 2002). We post-process *f0* estimates by applying a temporal smoothing filter (uniformly average, over three windows), which averages each window's *f0* estimate with those of its immediately preceding and following neighbors.

We estimate energy for each speech segment window according to Eqn. 1. Finally, we derive a new measurement, for each short-time window, of energy-weighted pitch by taking the product of the pitch and energy estimates.

From these three measurements of pitch, energy, and energy-weighted pitch, we calculate the mean, variance, nonzero minimum, maximum, and range (or maximum-minimum) values over the speech segment. This gives us our 15 classification features.

We presume that our binary classes of approval and not-approval will separate well and cluster within each class. Therefore, we use $k$-nearest neighbors, to classify novel utterances. High accuracy in preliminary trials led us to select $k = 3$.

The prosody classifier's training data is acquired from the individual tutor, in an interaction loop similar to the final learning interaction loop. To generate training examples of approving and not-approving prosodic affect, the tutor is given a simple, interactive training game, in a similar style to the interaction sequence used to train Nico. This training game is designed to elicit prosody similar to that elicited during Nico's waving training, and to provide automatic labeling for the prosody classifier's training data. In the prosody classifier's training game, the tutor is told to train a remote robot on how far it must travel from a hazard to reach safety. The tutor is given the threshold of safe distance from a practice hazard. The tutor is allowed only to provide the remote robot with information via tone of voice. Training involves presentation to the tutor of a sequence of distances traveled by the remote robot. In response to each distance reported, the tutor must give the robot prosodic feedback. These feedback utterances form the corpus of training examples to the prosody classifier. Because the robot's performance and the threshold are known before the tutor produces each training example, the examples are easily, automatically labeled.

#### 5.1.2.7 Reinforcement learning of waving behavior parameters

We demonstrate prosody as a feedback mechanism for the problem of refining Nico's social waving behavior. We define a waving behavior to be an oscillatory motion at Nico's elbow joint, around a fixed raised arm and hand position. A waving behavior can be parameterized by the amplitude (measured in joint angle degrees) and frequency of oscillation.

| | | |
|---|---|---|
| SMALL, FAST<br>(30°, 3.1 Hz) | MEDIUM, FAST<br>(40°, 3.1 Hz) | BIG, FAST<br>(70°, 3.1 Hz) |
| SMALL, MEDIUM<br>(30°, 1.9 Hz) | MEDIUM, MEDIUM<br>(40°, 1.9 Hz) | BIG, MEDIUM<br>(70°, 1.9 Hz) |
| SMALL, SLOW<br>(30°, 1.3 Hz) | MEDIUM, SLOW<br>(40°, 1.3 Hz) | BIG, SLOW<br>(70°, 1.3 Hz) |

Figure 5.3 A space of nine distinct waving behaviors. Each box represents one waving behavior state. In our experimental state space, waving behavior states are ordered from left to right with increasing amplitude, and from bottom to top with increasing frequency.

In this system, Nico is presented with a space of nine waving behaviors, combining three amplitudes ranging from small to large, and three frequencies ranging from slow to fast. The space of waving behaviors is organized as shown in Figure 5.3. Each box in the figure represents a single waving behavior. During each trial-feedback cycle, Nico can transition to a new waving behavior if it shares an edge with its most recent waving behavior's box, or Nico can choose to repeat the same waving behavior. This space can be thought of as a finite state machine with 9 states, and the learning task as the problem of identifying the goal state.

Before beginning an interactive tutorial with Nico, the tutor chooses a goal for what kind of waving behavior she would like Nico to perform. Nico initiates the tutorial by arbitrarily selecting a waving behavior and performing it.

Nico uses q-learning to discover the tutor's desired wave state. Therefore, Nico maintains an internal estimate of the utilities or q-values, $Q(s, a)$, for the transitions between

each waving behavior. Here, $s$ is a waving behavior state, and $a$ is an action or transition from s to Nico's next waving behavior state $s'$ (Russell & Norvig, 2003, pp. 763–784).

Following each transition and its demonstration of its newly selected waving behavior, Nico's q-learner receives prosodic feedback, $R(s,a) \in \{0,1\}$, which it treats as a q-learning reward signal and uses to update the q-value for its most recent transition between waving behaviors:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(R(s,a) + \gamma \max_{a'} Q(s',a')) \qquad (5.6)$$

where s' is the next waving behavior, which transition a leads to, from previous waving behavior $s$, and $a'$ is a transition leading from waving behavior $s'$ (Russell & Norvig, 2003, pp. 763–784).

The q-learning parameters $\alpha$ and $\gamma$ influence the sensitivity of q-values to changes in the q-values of successor states and transitions, and the number of predecessor states and transitions, whose values will be affected by updates to q-values, respectively (Russell & Norvig, 2003, pp. 763–784).

As for Nico's choice for its $l$th action, we have selected the following action policy: from its current waving behavior, with some probability $(1 - p[l])$ Nico selects the transition with the highest q-value. However, with probability $p[l]$, Nico instead uniformly randomly selects a transition.

Random exploration of the state space is important for two reasons. First, in the case where Nico misclassifies the prosodic affect, it can update its q-values to incorrectly prefer an undesirable transition. In such a case, random exploration can give Nico a new

opportunity to correct its error or return to an optimal path to goal. Secondly, if the state space of waving behavior should contain a local maximum, and if Nico finds itself performing the locally optimal waving behavior, a random transition away from the local maximum can give Nico an opportunity to seek the global optimum. In order to allow Nico to finally converge on some waving behavior, we allow the probability $p[l]$ of random exploration to decrease by a factor that grows geometrically with the number of trial-feedback cycles:

$$p[l] = p_* \xi^l \tag{5.7}$$

where $l$ is the number of cycles.

### 5.1.3 Validation experiment

We tested the recognition and learning systems in a closed, interactive feedback loop with a single human tutor (myself). Voice-activation detection was trained on 15s of background noise and 8s of continuous speech. Prosody classification was trained on 19 approving and disapproving feedback utterances, captured and labeled using the remote robot safety scenario.

For our q-learning system, we used the parameters $\alpha = .5; \gamma = .8; p_* = .7; \xi = .95$. We arbitrarily choose to let Nico start with a small, slow waving behavior, and we assign the tutor to prefer a big, fast waving motion.

#### 5.1.3.1 Voice-activation detector performance

The maximum-likelihood voice-activation detector exhibited only 3.4% error, including both false positives and negatives, when tested over the VAD training corpus itself. We did not

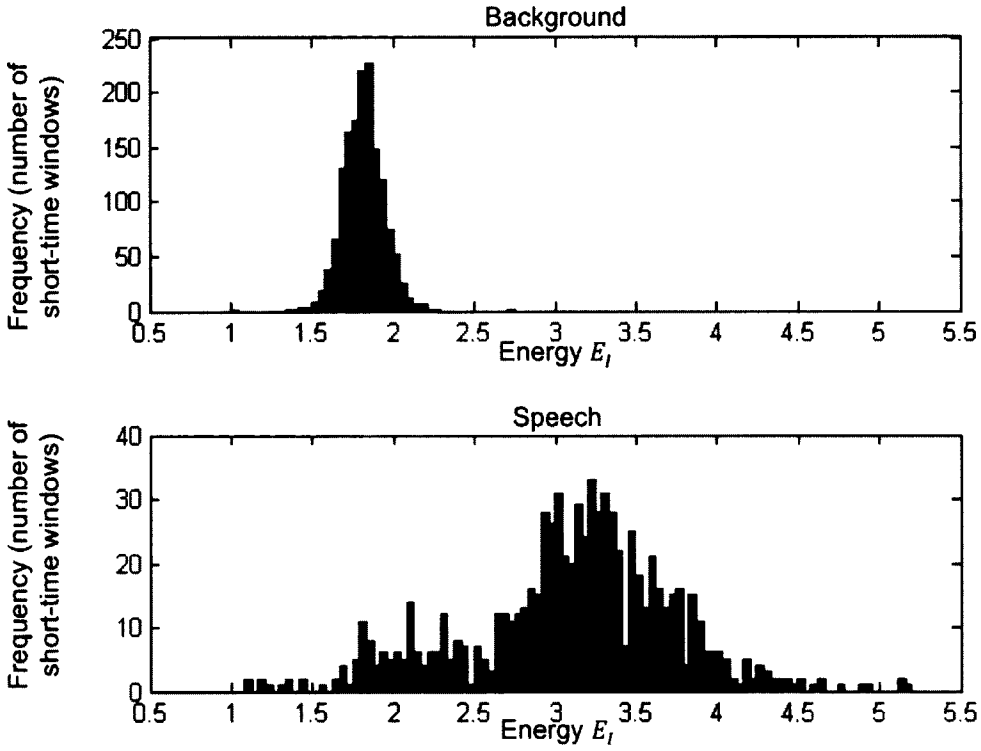



Figure 5.4 Training background (top) and speech (bottom) histograms over energy measurements, one of three voice-activation detection (VAD) features. The VAD derives Gaussians probability distributions from these sample distributions, and performs maximum-likelihood detection on novel short-time audio windows.

measure VAD error for any live interaction data, as we do not have a means to automatically acquire true voice activation labels in during tutorial. Figure 5.4 shows distributions of background noise and speech training samples for energy, the decision feature exhibiting the best separation between the noise and speech distributions. The background noise distribution is shown in the top plot, and the speech distribution is shown in the bottom plot.

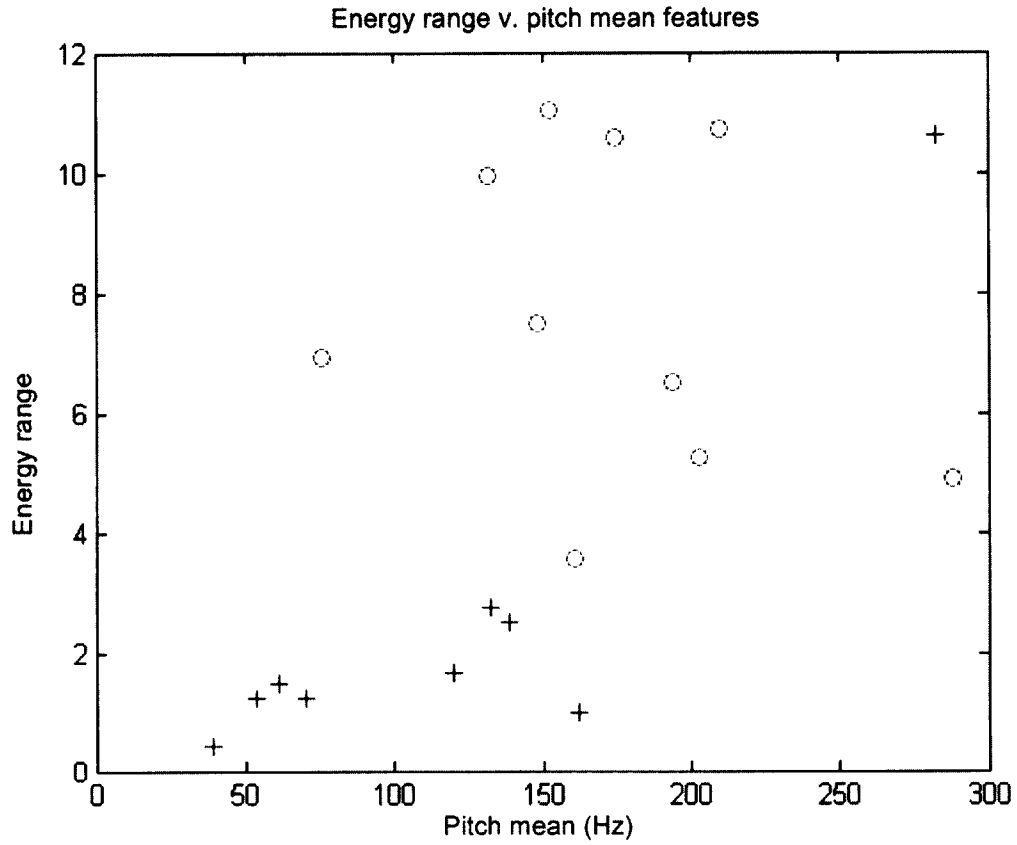



Figure 5.5 Prosody classifier training data distributed over *f0*-mean and energy-range features. Utterances featuring approving prosody are marked by "+"s and utterances featuring disapproving prosody are marked by "o"s. For these two features, the training data shows clear separation.

### 5.1.3.2 Prosody classification

The prosody classifier was trained on a corpus of 19 utterances, including 10 of approving and 9 of disapproving affect. Leave-one-out cross-validation over training data results in false positive and miss rates both of 5.3%.

Figure 5.5 shows the training utterances' distributions over *f0*-mean and energy-range features, two of the 15 training features consisting of simple statistics derived from acoustic measurements of *f0*, signal energy, and energy-weighted-*f0*. Over these two features, the training data shows clear separation.

Prosody classifier performance was measured over the particular tutoring interaction sequence presented in Figure 5.6. Presuming that the tutor's prosody was consistent with the desirability of each tried transition, the true prosodic values were estimated from Nico's sequence of transitions. For every transition (including random, as opposed to step-wise) which brought Nico closer to the desired waving behavior, the true prosody was estimated to be approving, and for transitions (including random, as opposed to step-wise) that brought Nico farther from the goal behavior, the true prosody was estimated to be not-approving.

A comparison of these estimated true prosody values with the actual output of the prosody classifier showed that the prosody classifier made 0 false positive errors and missed two, or 8.3%, of all approving utterances, falsely classifying them as not-approving.

#### 5.1.3.3 Learning the tutor's goal behavior

Figure 5.6 shows Nico's approach to the desired waving behavior, over the course of an interactive tutoring sequence. The plot shows the cumulative distance, measured in transitions, between Nico's current behavior and the goal behavior. The cumulative error curve has steep slope for trials during which Nico's waving behavior is very different from the goal behavior. On the other hand, the cumulative error curve is horizontal for those trials during which Nico is performing the goal behavior.

Figure 5.6 indicates that Nico performs the goal behavior five times in a row, from the 12th-17th trials, and then transitions away from the goal behavior, exploring behaviors which are far from the goal, until finally returning back to the goal behavior.

If Nico found the goal behavior, why did it later switch to another behavior? The answer is that Nico's action policy calls for it to select the transition with optimal q-value most of

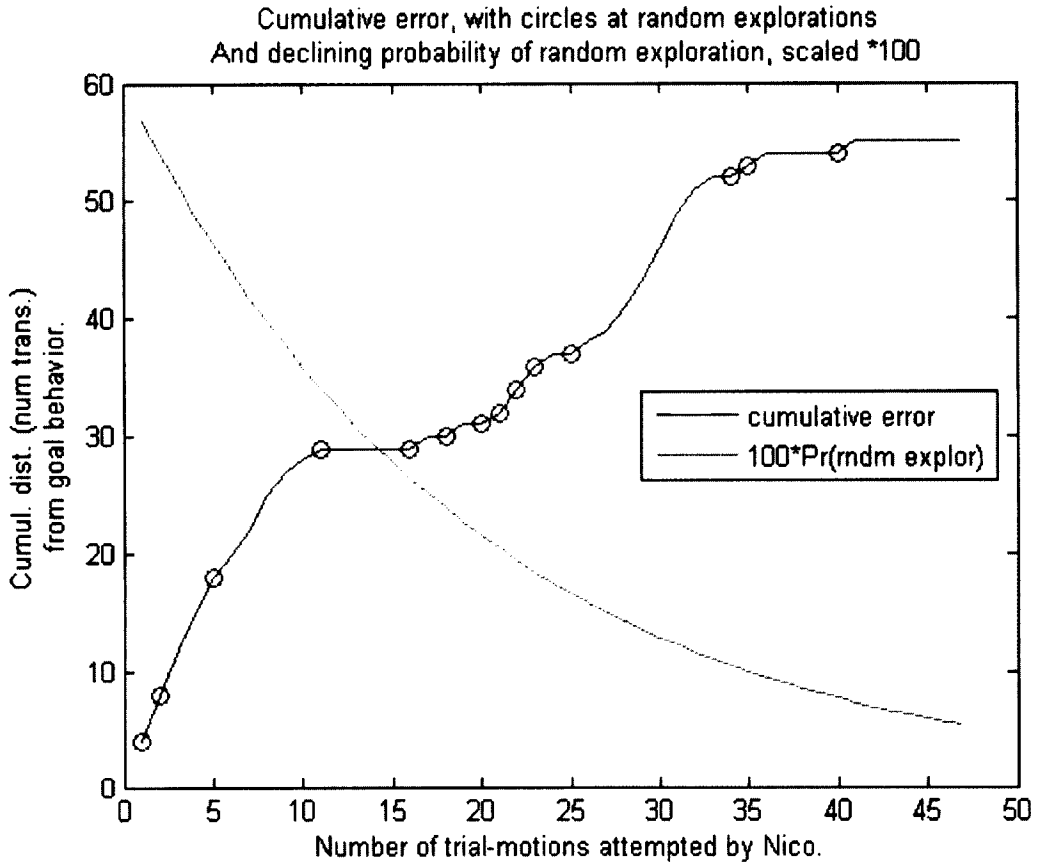



Figure 5.6 Convergence of waving behavior q-learner onto desired waving behavior. The blue line shows cumulative error versus number of trials in the tutorial sequence. Zero slope in the cumulative error curve indicates transition to the goal behavior, producing no additional error. Circles demarcate the trials during which Nico chose its next waving behavior uniformly at random. The red line indicates the declining probability (scaled by a factor of 100) of such random exploration, scaled by a factor of 100.

the time, but with some probability to select, uniformly at random, any transition, regardless of q-value. In Figure 5.6, circles mark the trials during which Nico randomly selected its next behavior. Note that from the beginning through the middle of the sequence, when Nico selects a transition at random, this results in the accumulation of error, as Nico explores far from the goal behavior. However, near the end of the sequence, Nico recovers to the goal behavior rapidly, as indicated by the horizontal cumulative error curve.

In general it is safe to expect that as the number of trials increases, Nico will find and stay fixated on the goal behavior. This is because the probability of randomly exploring away from the goal behavior decreases geometrically with time, and because as time passes, Nico enriches its model for the space of behaviors. For example, Nico's path indicates that during trials 25-33, Nico explored previously unvisited behavior states, causing it to learn q-values over these novel states.

This knowledge allowed Nico to recover immediately to the goal state, following random explorations away from it in trials 35 and 40. After 47 trials within this sequence, the probability of randomly choosing an action was only 6.3%.

### 5.1.4 Discussion

This system automatically recognizes positive and negative affect in a teaching interaction with a robotic learner. In Chapters 2 and 4, we described human-robot interactions over groups of adults with typical development, children with typical development, and children with ASD, in which participants directed affectively expressive prosody to robotic learners. Although we have not yet automated recognition of those participants' affective prosody, the system described in this chapter lays out a possible solution for that automation. Speech prosody is often observed to be oddly monotonic or inappropriately sing-song in individuals with ASD (Nadig & Shaw, 2012; Tager-Flusberg & Caronna, 2007), and has been cited as functionally detrimental to social interaction, and an important target for intervention (Paul, Shriberg, et al., 2005). Automation of affective expressions in prosody would allow users to practice speech prosody with an automatic judge outside the clinic, reducing labor in interventions.

#### 5.1.4.1 Prosody as feedback to drive machine learning

In addition, this system allows a person to use real-time speech prosody to drive a physically embodied machine learner. We have shown that it is possible to recognize affect in prosody in a real-time, interactive loop, with a high level of accuracy that makes possible its usage in a real-time learning system.

#### 5.1.4.2 Extension to other individuals

Initial explorations using other participants to tutor Nico have confirmed the importance of affective response from the robot, as previously demonstrated by Breazeal and Aryananda (2002). Even mothers of infants and highly expressive caretakers of pets, who are accustomed to speaking Motherese with their children or animals, indicate reluctance to express exaggerated affect to , in the absence of affective feedback from the robot.

#### 5.1.4.3 Extension to other affective states

Previous work in prosody classification has successfully classified over other affects, besides approval and disapproval. Thomaz, Hoffman, and Breazeal (2006) showed that humans often prefer to give guidance, which may be viewed as attentive affect, as well as positive or negative reinforcement, which may be viewed as approving or disapproving affect. It's possible that other prosodic affects may enrich a tutoring interaction with a robot, by providing feedback other than positive and negative reinforcement. Following the work described in this chapter, we studied the affective prosody that untrained adults directed toward a robotic learner and found that people provide affectively expressive prosody before the learner completes a trial action, and that the intensity and quantity of the affective expressions vary depending on the learner's performance (Chapter 2).

### 5.1.5 Implications for socially assistive robots

This system is a closed looped, using automatically recognized affective prosody to drive a learning algorithm. Our demonstration over a toy learning task and a user with expert knowledge has obvious methodological limitations. Nonetheless, the study proves the concept that we can close the loop on automated response to recognized social behavior.

As our understanding of the behaviors children with ASD use while interacting with robots we are simultaneously enriching the foundations for automating recognition of these behaviors. ˇ is a target for active research in intervention. Our proof-of-concept demonstration of a closed-loop, automatic recognition and learning system presented here endorses future work in the investigation and development of affective prosodic targets in robot-based intervention.

## 5.2 System 2: Recognition of mutual belief cues in infant-directed prosody

Deficiencies in theory of mind are thought to underlie many of the social difficulties experienced by individuals with ASD. Pierrehumbert and Hirschberg (1990) described prosodic expressions used by speakers to modify the information they share with listeners, that is, to manage mutual belief. These cues may be an important target for intervention, or a useful diagnostic indicator, in autism. Here we describe a novel system that classifies mutual belief management from infant-directed prosody. This investigation demonstrates the feasibility of automatic recognition of mutual belief management communications in prosody, and also suggests these as a behavior which deserves exploration among children with ASD, both within interactions with robots and generally.

We examined whether evidence for prosodic signals about shared belief can be automatically identified within the acoustic signal of infant-directed speech (E. S. Kim et al., 2008). We examined infant-directed speech because it is known to have exaggerated acoustic features and because studies have shown that some robots can elicit speech that is similar to Motherese, making Motherese a good match for human-robot interaction applications. Two transcripts of infant-directed speech for infants aged 1;4 and 1;6 were labeled with distinct speaker intents to modify shared beliefs, based on Pierrehumbert and Hirschberg's theory of prosodic expression of mutual belief cues. Acoustic predictions were made from intent labels first within a simple single-tone model that reflected only whether the speaker intended to add a word's information to the discourse (high tone, H*) or not (low tone, L*). We evaluated whether we could acoustically identify such cues in infant-directed prosody. We also predicted pitch within a more complicated five-category model that added intents to suggest a word as one of several possible alternatives (L*+H), a contrasting alternative (L+H*), or something about which the listener should make an inference (H*+L). The acoustic signal was then manually segmented and automatically classified based solely on whether the pitches at the beginning, end, and peak intensity points of stressed syllables in salient words were closer to the utterance's pitch minimum or maximum on a log scale. Evidence supporting our intent-based pitch predictions was found for L*, H*, and L*+H accents, but not for L+H* or H*+L. No evidence was found to support the hypothesis that infant-directed speech simplifies two-tone into single-tone pitch accents. Our system and demonstrations establish the feasibility of automatically recognizing mutual belief cues in infant-directed prosody, and also suggest future exploration in the production of such prosodic cues among individuals ASD.

### 5.2.1 Introduction

Prosody, or the melody of an utterance, can contain information about what the speaker thinks the listener knows (or does not yet know) about an utterance. For example, when introducing herself for the first time, a speaker might say "Hello, I'm Eli Kim" in a high pitch, indicating that she is communicating new information to the listener. When giving a talk before an audience that already knows her, however, she might instead begin with a desultory "Well, as you know, I'm Eli Kim" with low instead of high pitches on the name to indicating that she is reiterating information the audience already knows. Our system uses acoustic information to automatically detect prosodic signals of new or old information in infant-directed prosody.

The literature on adult-directed prosody has produced a rich classification scheme to associate acoustic cues in speech with specific intents to modify the listener's and speaker's shared beliefs (Pierrehumbert & Hirschberg, 1990). Children with autism frequently fail to produce prosodic signals that adequately differentiate between old and new information (McCaleb & Prizant, 1985), another deficient communication skill which marks prosody, and thus the speaking individual, as atypical or odd (Mesibov, 1992; Shriberg et al., 2001; Van Bourgondien & Woods, 1992).

As mentioned previously, infant-directed speech has exaggerated prosodic features. Investigations of infant-directed speech have focused on the use of these exaggerated features to emphasize turn-taking signals (Snow, 1977), speech stream segmentation (Fernald & Simon, 1984; Thiessen, Hill, & Saffran, 2005), and signals to attract and maintain an infant's attention and communication of affect. There has also been investigation of infant-directed prosodic signals of new versus old information (Fernald & Mazzie, 1991). However,

there has not been any investigation of whether infant-directed prosody contains the same breadth of shared belief modifying signals as does adult-directed prosody. Does infant-directed speech contain the same signals about mutual belief, or do parents simplify infant-directed prosody by reducing their selection of prosodic signals? The experiment described in this paper examines the prosodic patterns of infant-directed speech taken from the CHILDES database (MacWhinney, 2000) in order to determine whether infants as young as 16 months receive the full variety of pitch accents that signal, in adult-directed speech, speaker intent to modify shared knowledge.

Below, we provide background about prosody, including Pierrehumbert and Hirschberg's (1990) theory of the acoustic correlates of mutual belief cues, which our system semi-automatically recognizes. Section 5.2.2 describes our acoustic classification method. Section 5.2.3 will describe our experiment in which audio data from the CHILDES corpus was analyzed to determine whether the infant-directed speech matched the predictions implied by Pierrehumbert and Hirschberg's system. Section 5.2.4 will include our analysis of the data, and our conclusions in Section **Error! Reference source not found.** will discuss implications on socially assistive robots for autism interventions.

#### 5.2.1.1 Recognition of infant- and robot-directed prosody

In the robotics and cognitive science literature, previous computational research on infant- or infant-like, learner-directed speech has largely focused on communication of mood or affective intent. Systems have been built to recognize or describe the prosody of speaker approval (with sustained pitch peak intensity) and prohibition (with low, staccato tones) (Breazeal & Aryananda, 2002; E. S. Kim & Scassellati, 2007; Robinson-Mosher & Scassellati, 2004; Roy & Pentland, 1996; Slaney & McRoberts, 2003), bids to attract attention (with

rising pitch contours) (Breazeal & Aryananda, 2002; Ferrier, 1985; Robinson-Mosher & Scassellati, 2004; Slaney & McRoberts, 2003), and soothing intent (with falling pitch contours) (Breazeal & Aryananda, 2002; Papousek, Papousek, & Bornstein, 1985; Robinson-Mosher & Scassellati, 2004). Infant-directed prosody has also been investigated for cues to turn-taking (Snow, 1977), speech stream segmentation (Fernald & Simon, 1984; Thiessen et al., 2005), and new versus old information (Fernald & Mazzie, 1991).

There has been limited investigation into shared belief cues in infant-directed prosody. Adult-directed prosody is thought to convey information about what is mutually believed between speaker and listener (see Section 5.2.1.1). Whether such signals exist in infant-directed prosody has not been previously studied in the framework we discuss below, but there might be good reason to think that adults might modify their prosody to make it less complex. It is known that parents tend to speak to their children in exaggerated prosody, known as "Motherese" (Fernald, 1985). Infant-directed prosodic exaggeration has inspired some builders of robotic prosody classifiers to attempt to elicit Motherese-like speech with infant-like robots (Breazeal & Aryananda, 2002; E. S. Kim & Scassellati, 2007).

Whereas investigations of Motherese have suggested that infant-directed pitch contours are characteristic of specific affective intents, an alternative view may be that these contours are determined by the informational content of the speech. Pierrehumbert and Hirschberg have argued that the tune of adult-directed prosody cannot be explained either in terms of the speaker's speech acts or emotion alone, since the mapping from tune to speech act or emotion is at best one-to-many (Pierrehumbert & Hirschberg, 1990). Instead, to describe adult-directed prosody they proposed the system described below, in which prosody signals each word's relation to the speaker's intended modification of shared beliefs. To our

knowledge, before we commenced our own study, there had been no investigation of the extent to which Pierrehumbert and Hirschberg's (1990) model for adult-directed mutual belief cues also holds for infant-directed speech, though some similar observations about novelty affecting pitch have been made in the infant-directed literature (Fernald & Mazzie, 1991).

#### 5.2.1.2 Prosody, shared beliefs, and discourse structure

The following exposition of the shared belief information in prosody is based closely on that of Pierrehumbert and Hirschberg (1990), which has been empirically supported to some extent (Krahmer & Swerts, 2001). The labeling scheme summarized here is the basis for the popular ToBI representation of prosody (Beckman, Hirschberg, & Shattuck-Hufnagel, 2005; Silverman et al., 1992).

In English a speaker produces a pitch accent for at least one word in each utterance, marking it as salient. A high or low pitch on the stressed word conveys whether the speaker intends for the listener to add the word's information to their mutual beliefs (Pierrehumbert & Hirschberg, 1990). Accented words are perceived by listeners to be prominent, or stressed, with relation to other words. In English, every word has at least one stressed syllable; however, accented words receive an additional stress over other words. Stress of one word over others is conveyed though a combination of greater loudness, longer duration, and hyperarticulation of that word's stressed syllable. There are two simple pitch accents, H and L, and three two-tone pitch accents, which combine H and L pitches.

The H* pitch accent is used to convey the speaker's intent for the listener to add the accented information to their shared beliefs. Perceptually, an H*-accented word will feature a relatively high pitch at the perceptually prominent syllable in the prominent word. The '*'

diacritic indicates temporal alignment with the stressed syllable. This accent is commonly used when introducing new information, and frequently appears in declarative statements. For instance,

| Alice | likes | Bob |
|---|---|---|
| H* | | H* L L% |

Here, the speaker *S* intends for the listener *R* to add the fact of Alice's liking and the fact that Bob is liked to *R*'s beliefs. This utterance thus would be appropriate if, for example, neither person had been mentioned in the conversation previously. (The L L% at the end refers to the pitch of the phrase and whole utterance, respectively; we include these markings for completeness but will not discuss them.)

H* can be used to add connoted rather than denoted information to *R*'s beliefs. For instance, in this example, *S* tells *R* what *R* has done (and thus presumably already knows). Here *S* uses an H* accent to convey that *R* should add knowledge of *S*'s awareness to *R*'s beliefs.

| You | ate | my | cookie | on | purpose |
|---|---|---|---|---|---|
| H* | H* | | H* | | H* L L% |

The L* simple pitch accent is perceptually indicated by a prominent word that is close to the baseline pitch for the speaker. It indicates the speaker's intent for the listener not to add the accented item to his beliefs. This accent is commonly used when *S* is uncertain, such as in yes or no questions:

| Did | our | paper | get | rejected |
|---|---|---|---|---|
| | | L* | | L* H H% |

L* can indicate *S*'s belief that the expression is incorrect:

| I | guess | our | paper | just | isn't | good | enough |
|---|---|---|---|---|---|---|---|
| | L* | | L* | | L* | L* | L* L L% |

or when uttering information believed already known by *R*:

| I'd | like | coffee | and | I think I'll have a muffin |
|---|---|---|---|---|
| L* | L* | L* H H% | | |

In all these cases, *S* does not intend for *R* to add the L* accented information to their shared beliefs, since the L* accented items are uncertain, false, or previously added.

In two-tone, as in single-tone, pitch accents the '*' indicates temporal alignment with the stressed syllable. The L*+H pitch accent perceptually is perceived as a low frequency on a

stressed syllable, followed immediately by a rising pitch contour to a higher pitch. L*+H pitch accents indicate uncertainty in an implied comparison of scale. For example,

A: This talk is terrible.
B: The paper was good
L*+H L H%

The L*+H accent on good indicates *B*'s uncertainty as to the relevance of the paper's quality to the quality of the talk.

Likewise, L+H* pitch accents also signal an intended comparison of scale, but are instead conveyed with certainty, expecting the listener to add the accented item to *S* and *R*'s shared beliefs. For example,

A: This paper is awfully informal.

B: It's even chatty for a conference paper
L+H* L L%

The H*+L accent signals that the listener should infer support for the accented items from previously existing beliefs. Like the H* accent, H*+L signals that *R* should add the accented item to their beliefs, but also should make an inference based on the new information and existing beliefs, such as an implied course of action:

Your dinner's getting cold
H*+L L* H*+L H L%

Pierrehumbert and Hirschberg suggested that H+L* possessed a similar meaning to H*+L, but was used to convey information already known to the speaker (Pierrehumbert & Hirschberg, 1990). However, (Pierrehumbert & Hirschberg, 1990) also noted that "there is some difficulty in separating the meaning of H+L* from that of H*+L, because in many cases the phonological analysis is unclear" (p. 300). Moreover, in modern labeling conventions, the H+L* notation has been superseded with H+!H*, to note that this contour usually remains higher than other low tones (Beckman & Hirschberg, 1994). For these reasons, this tone was not predicted for any utterances in our experiment, though we did check for acoustic evidence of it.

### 5.2.2 Shared belief cue recognition algorithm

Usually pitch accents are manually classified by trained specialists using acoustic recordings and graphical representations of the pitch contour over time (Beckman et al., 2005). Manual classification tends not to produce high amounts of agreement among experts (Syrdal & McGory, 2000).

Attempts to automate pitch accent labeling from acoustic features have tended to focus on locating, rather than classifying, pitch accents (Ananthakrishnan & Narayanan, 2008; Hasegawa-Johnson et al., 2005), have classified only a very limited subset of pitch accents for adult-directed speech (L. X. Sun & Sun, 2002), or have classified pitch accents for languages beside English (B. Kim & Lee, 2006). Pitch accents have been statistically clustered with high agreement (78%) with listeners' judgments, suggesting acoustic regularities distinguishing pitch accent categories (Levow, 2006).

We have designed a partially automated method that allows American English pitch accents to be acoustically, quantitatively classified. This allows our hypothesis-testing to be free of bias introduced by our knowledge of the semantic content of the speech, and also is a step toward the fully automated classification of pitch accents. (The reader may find it useful to refer to Figure 5.7).

To begin, intensity and *f0* contours over time are estimated using Praat phonetic software (Boersma & Weenink, 2010). Utterances in the CHILDES transcripts are linked to their temporal positions in the recordings. CHILDES' CLAN software links text to Praat. Next, utterance-minimum and -maximum *f0* are extracted, using Praat, giving a baseline pitch and pitch range for the speaker at the time of utterance. Measuring pitch range locally within each utterance, instead of over the speaker's entire history, allows the range to adapt to the speaker's current affect and immediate auditory conditions, though it has the disadvantage of sometimes producing an falsely small range for utterances having no H* pitches. During this stage utterances within which *f0* estimation software clearly fails, are manually discarded: failures include sudden pitch drops below 75 Hz, sudden jumps to

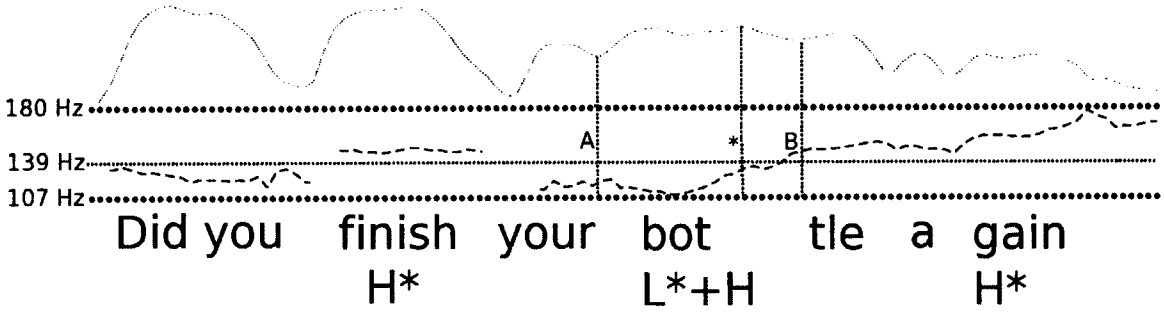



Figure 5.7 A sample utterance from MacWhinney's CHILDES corpus, with intensity (top) and *f0* (middle, dashed) extracted using Praat phonetic analysis software. The minimum and maximum *f0* of the utterance establish the baseline and range, and their average on a log scale gives the dividing line between L* and H* pitches. For two-tone classifications, the *f0* at the beginning of the stressed syllable (A) gives the first tone, and the end of the syllable (B) gives the second; the stress is placed based on the syllable's point of maximum intensity (*). Though the statement is phrased as a question (suggesting L*), in fact the speaker is essentially telling the infant that he is aware that the infant is done (H*) but is unsure whether the whole bottle is gone (L*+H).

overtones (doubling or halving errors), or misclassification of unvoiced noise as pitch. (Section 5.2.4 describes how frequently this occurred in our experiment.)

The next step is the segmentation of the stressed syllables in the selected words. Segmentation was done manually by listening to the audio and using cues from the intensity curves (e.g., "stops" such as "p" and "k" literally stop the air momentarily, and thus are clearly marked by low intensity). The stressed syllable of a selected word is the relevant part of the audio signal for pitch accent classification.

To classify a pitch accent as a single tone, *f0* at the point of maximum audio intensity within the syllable is compared to the baseline minimum and maximum *f0* over the whole utterance. After taking the logarithm of all three fundamental frequencies–minimum, maximum, and *f0* at time of maximum intensity–whether the maximum-intensity-*f0* is closer to the baseline minimum or maximum determines whether it is L* or H*. This comparison is done on a log scale, a method we introduce here for pitch accent classification because just-

noticeable differences for pitch are logarithmic in frequency in the 50-5000 Hz range (Clark, 2003), which covers the range of human speech, and because H* pitch accents are thought to actually be medium to high pitches within the speaker's range (Beckman & Hirschberg, 1994), which intuitively fits well with a log-scale model.

To classify a pitch accent as two-tone, our method examines *f0* at the beginning and end of the syllable as well. These points, which were manually identified for syllable identification, are subjected to the same logarithmic transformation, and classified as L or H based on whether they are above or below the log-transformed midpoint of the speaker's range. If the two endpoint classifications are the same, the pitch remains classified as a simple L* or H*. If they are different, then the pitch is classified as a two-tone accent, L+H or H+L. In the L+H case, the location of the accent mark is determined by the classification of the maximum intensity point. If the pitch at the time of maximum intensity is closer to the log-transformed pitch baseline, it is L*+H; otherwise, it is L+H*. The maximum intensity classification is similarly used to distinguish between H*+L and H+L*.

Both the acoustically simple one-tone method and the two-tone method were used and compared to our theoretical predictions in the experiments to be described below.

### 5.2.3 Experiment

Two transcripts of infant-directed speech from the CHILDES database (MacWhinney, 2000) were examined: one of a father speaking to his 16-month-old son (MacWhinney, 2000) and another of a mother speaking to her 18 month-old-daughter (Ratner, 1987). 165 words from these transcripts were chosen as targets for comparison of the theoretical predictions of the Pierrehumbert and Hirschberg model (Pierrehumbert & Hirschberg, 1990) to the observed acoustics. A word was chosen as a prediction target if it was central to the meaning of its

sentence, and if the transcript context made one pitch accent category seem more likely than the others. Single- and two-tone predictions were made for each selected word, given the conversational situation. In the forced two-choice prediction, pitch accent was predicted depending on whether the text suggested that the parent wished to introduce informational content with the word or not. In the five-category prediction, the experimenters made their predictions based on whether the word was being tentatively suggested as one of several specific alternatives (L*+H), being specifically suggested in contrast to another alternative (L + H*), was a reminder of something that the child should already know (H+L), or was otherwise introducing new information (H*) or not (L*). Predictions were made based on the textual transcripts alone, without having heard the audio recordings.

We note that our predictions assumed neither an accurate representation of the infant's belief state on the part of the speaker, nor expectations on the part of the speaker of adult-like belief state for the infant listener. Rather, we assumed that speakers tailor their representations of the listener's belief states to the individual listeners and conversations. Our predictions reflect only indications from local context in the transcripts (of up to a few preceding and following sentences) about the speaker's intents to modify what they apparently conceived to be mutual beliefs.

We also distinguish between introduction of a new word (for example, the naming of a novel object) and new information, a broader act, which can include, for example, newly achieved certainty in interpreting an infant's proto-linguistic requests for a bottle. Our H* predictions are of the broader sort.

Following transcript-based predictions, the utterances containing the selected words were then analyzed using the acoustic method introduced in Section 5.2.2, and the quantitative results compared to the theoretical predictions.

### 5.2.4 Results

Of the 165 utterances, 30 utterances were discarded because of auditory noise or incorrect segmentation within the corpus, leaving 135 data points for each of the single-tone and two-tone classification schemes.

The single-tone predictions of L* and H* coincided with the results of our single-tone acoustic analysis method (see Section 5.2.2) for 87 of the words, or 64% of the time; this was significantly more often than chance ($\chi^2$ (1, 135) = 8.61; $p < 0.005$). Though we had entertained the hypothesis that the difference might be attributed to whether the word was contained within a question or not, there was no evidence to support this idea ($\chi^2$ (1, 135) = 1.46, $p = 0.228$).

56 of the 135 two-tone predictions were correct, an occurrence highly unlikely to be due to chance because of the five categories ($\chi^2$ (20, 135) = 57; $p < 0.001$). Broken down into category-by-category comparisons, we found that the H*, L*, and L*+H predictions each produced significantly more correct responses than could be attributed to chance ($p < 0.005$; $p < 0.001$; $p < 0.005$, respectively), while the L+H* and H*+L predictions provided no such evidence of accuracy ($p = 0.656$; $p = 0.561$).

However, there was no evidence to support the hypothesis that parents tended to simplify their pitch accents toward their children, as there was no evidence that single tones

were more likely to be observed in the place of two-tone accents than vice versa ($\chi^2$ (1, 135) = 0.459; $p$ = 0.498).

Qualitatively, H* and H*+L were common pitch accents for introducing or reinforcing labels for objects (annotations are those provided by our acoustic method):

CHILD: What's that?
FATHER: Tape recorder over there
H* H*+L

L* was most common in cases when the parent was offering an interpretation of what the child was communicating:

FATHER: You like the soldiers?
L* L*

However, L* also occurred where we had predicted H* in cases where it seemed from the text that the parent was pointing out new information, but the parent was actually going through a ritual such as reading a familiar book:

MOTHER: And that's a rabbit with no face.
L* L*

L*+H was often used in its adult meaning of an alternative that the speaker was unwilling to support, but in cases where one might have expected L+H* to indicate correction, the speaker did not appear to follow through:

CHILD: dog? ...
FATHER: is that a doggy Honey ? ...
L*+H
FATHER: or is that [//] he's a kitty ?
L*

H*+L was very occasionally observed in the role of asking the child to make an inference, but this was not consistent:

MOTHER [pointing to mirror]: Who's in there? ...
H*+L

MOTHER: That's Amelia!
H*

As these examples illustrate, the instances in which the predictions failed to match the observations were often explainable by the ambiguity of the text, and not a failure of the theory or acoustic method.

### 5.2.5 Discussion

These results demonstrate that at least some of Pierrehumbert and Hirschberg's acoustic signals about shared belief, and our acoustic method for identifying pitch accents, hold for American English infant-directed prosody at ages 16-18 months. Our data shows strong evidence for H*, L*, and L*+H accents' usage for conveying the same information about mutual belief proposed in the adult-directed case, at least for the two speakers whose prosody we investigated thoroughly. These differences in pitch are not determined by a word's embedding in a question, but mark whether or not the speaker wishes to introduce new information with the word, or (in the case of L*+H) whether the speaker offers the word as one of several possible alternatives.

Table 5-1 Incidence of predictions and observations for Pierrehumbert and Hirscberg's six categories of pitch accent.

| Pred \ Obs | H | L | L*+H | L+H* | H*+L | H+L* | Total |
|---|---|---|---|---|---|---|---|
| H | 15 | 8 | 0 | 2 | 3 | 0 | 28 |
| L | 10 | 30 | 4 | 4 | 4 | 2 | 54 |
| L*+H | 8 | 6 | 7 | 2 | 0 | 2 | 25 |
| L+H* | 3 | 1 | 3 | 2 | 6 | 0 | 15 |
| H*+L | 5 | 3 | 1 | 0 | 2 | 2 | 13 |
| H+L* | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 41 | 48 | 15 | 10 | 15 | 6 | 135 |

Our findings show that even when speaking to infant listeners, with immature cognitive and linguistic capabilities, speakers signal their intent to modify listener's beliefs, in ways similar to those suspected to be used for adult listeners. In other words, infants are receiving cues about what is shared information even at an age when they are unlikely to have a concept of distinct states of knowledge between distinct individuals, which is demonstrated considerably later (Gopnik, 2001). It is therefore possible that children use pitch accent signals in learning to reason about shared and private information. Understanding the role of prosody in this process of reasoning about shared knowledge may be critical to understanding how theory of mind develops, and also to understanding why and how autistic children tend to demonstrate an impaired ability to reason about minds. A better understanding of how typical children integrate and react to infant-directed prosody may help early diagnosis of autism, which is known to include abnormal prosody as one of its symptoms (Shriberg et al., 2001). It is possible that among the prosodic difficulties found in

ASD are these markers of shared belief. And if so, these should also become a target for intervention, and the system for automatic recognition described here may facilitate robot-, or more generally, technology-based interventions.

The lack of evidence for H*+L and H+L* supports a recent tendency to view these particular pitch accents as less well empirically supported than others in Pierrehumbert and Hirschberg's scheme (Beckman & Elam, 1997), but it is somewhat unclear why L+H* poorly matched our predictions. There are several possible explanations. These accents may be particularly difficult to accurately predict from transcripts, since the difference between L*, L*+H, and L+H* might depend on how strongly the parent prefers an alternative. It is also possible that our acoustic method does not accurately describe L+H* accents. It is also possible that parents intentionally avoid this contour because of its negative connotation as a correction. This is a good question for future study.

What is clear is that American English infant-directed prosody contains some of the interesting signals about shared information theorized to exist in adult-directed prosody, and that a relatively simple method–comparing the log of the maximum intensity pitch to the speaker's maximum and minimum pitches–can extract them.

It may therefore be useful for creators of robotic systems to bear pitch accents in mind as an additional source of speech information. Though we have not yet measured agreement between our acoustic method with trained listeners' pitch accent judgments, our method offers quantitative, acoustic information about speaker intent. Automating the manual parts of our method, namely stressed syllable segmentation and discarding noise, are areas for future work.

## 5.3 Conclusions

In this chapter we have described a novel system that automatically recognizes affective, robot-directed prosody for use in a closed-loop machine learning system, and another novel system by which can recognize shared belief cues from infant-directed prosody. Production of affective prosody is known to be atypical among many individuals with ASD, to alienating effect. Evidence of theory of mind and prosodic deficits among individuals with ASD suggests the investigation of production of prosodic shared belief cues. The systems and demonstrations we have described in this chapter deepen our understanding of the production of these prosodic behaviors, and also indicate the feasibility of automatic recognition of these prosodic expressions. Automatic recognition, as well as the automatic responses we have demonstrated in System 1, can improve the feasibility of robot-based interventions targeting these behaviors.

# Chapter 6

# Interdisciplinary methodologies

While there is a rich history of studies involving robots and individuals with autism spectrum disorders (ASD), few of these studies have made substantial impact in the clinical research community. In this chapter we present an examination of collaborative challenges and strategies between clinicians and roboticists (E. S. Kim et al., 2012). We first examine how differences in approach, study design, evaluation, and publication practices have hindered uptake of these research results. Based on years of collaboration, we suggest a set of design principles that satisfy the needs (both academic and cultural) of both the robotics and clinical autism research communities. We developed these principles in the course of designing and executing the studies we described in Chapter 3 and Chapter 4.

## 6.1 A cultural divide

Any two distinct and mature research fields are likely to have substantially different methodologies and research cultures. In this section, we describe some of the critical differences between the ways the HRI and clinical communities typically plan, carry out, and report experimental studies. For simplicity, we will refer to the *robotics community* to indicate the fields, groups, and venues within which most of the extant findings on robotics and

autism have previously been published. This group is primarily represented by roboticists with backgrounds in computer science or engineering. The *clinical community* will refer to the fields and groups who conduct research on the diagnosis and treatment of autism, and the venues within which they share their findings. This group is represented by clinical practitioners, developmental psychologists, and social and behavioral therapists.

There is a growing body of research in effective inter- and trans-disciplinary scientific research, which has revealed, for instance, that collaborations are influenced by whether participants have faculty appointments in related academic departments and geographic proximity of collaborative partners (Stokols, Harvey, Gress, Fuqua, & Phillips, 2005; Stokols, 2006). In this chapter we describe collaborative challenges and strategies that are specific to clinical autism research and robotics.

We emphasize from the outset that our purpose is not to cast doubt over the methods and practices of either community. Rather, it is our position that to exclusively adopt the methods of one community or another would hinder progress towards the ultimate goal of partnership between these communities: using robots to aid the diagnosis and therapy of individuals with ASD. To conduct research and development according to only one community's standards would render results inaccessible to the majority of the other. Instead, we propose collaborative solutions—ways to negotiate logistical compromises and to design to each community's standards—that address some of the most pressing concerns of each group while making the results at least partially accessible to both. The following discussion organizes interdisciplinary differences into three critical areas: research approach, study design, and publication and dissemination.

### 6.1.1 Research approach

Within this collaborative space, the ultimate aim of both roboticists and clinicians is to determine the parameters within which, and the mechanisms by which, robots can improve interventions for, or assessment of, individuals with autism. Despite this shared ultimate goal, research approaches and motivations differ. While we sometimes loathe admitting it, research in robotics is often driven by the capabilities of our robots rather than the needs of target users. Funding awards, and their sponsored research endeavors, tend to focus on technological innovation, and the demonstration of feasibility of use. Each time a robot acquires a new capability, a search for applications that can take advantage of that new capability follows. The motivation for this approach is sensible: technological innovation can rapidly open new application areas and make fundamental changes to the kind of services that can be provided. Clinical research, on the other hand, is primarily driven by the specific needs of the target population. Funding and research efforts are directed toward questions that are most likely to reap substantial benefits for individuals with ASD. This fundamental and initiating distinction cascades into critical differences in the ways in which the two research communities approach collaborative works—as well as the ways funding agencies evaluate results. Clinicians have been hesitant to explore robotics technology in part because a clear case for the utility of robots in this area has not been made. What needs of a child with ASD does the robot fulfill, what support does it provide to the family, or what diagnostic value does offer to a clinician? To date there are no rigorous, controlled, sample-based demonstrations of a robot's improving symptoms, family support, or characterizations of individuals with ASD. Unfortunately, it is often not possible to answer these questions in advance of technology development. On the other hand, because little is known both about

how to design human-robot interactions for individuals with autism and about how these individuals will respond and benefit from interactions with a robot, technology cannot be developed strictly in advance of deployment to address specific needs of individuals with ASD. At best, contemporary research collaborations strive to utilize a design process that considers input from diverse stakeholders (including clinicians, families, and other users), and iteratively advances technology to meet needs that are, in turn, iteratively specified by the user community.

Differences in fundamental approach between robotics and clinical research communities lead, in turn, to differences in desired outcomes from studies. At present time, roboticists in clinical collaborations tend to seek proofs of concept, that is, demonstrations of a robot's successful engagement in interactions that are pleasant or socially appropriate, or that resemble an assessment, therapeutic or educational scenario. While engaging interactions are fundamental to effective interventions or assessments, a proof of concept alone will likely be insufficient to motivate clinical use. In clinical studies, research is validated only when a clearly specified benefit to the end user has been rigorously demonstrated. But demonstrations of engaging interactions with a robot do not necessarily show any specific clinical or functional benefit for the end user with ASD. From a pessimistic view a critic might claim that all existing robot-autism studies to date show only the ability for children with ASD to adapt to interactions with a robot, and that effectively training children to engage with robots will have no benefits to their ability to interact with other children or adults.

The study of HRI applications for autism is nascent. Given limited knowledge of the beneficial combinations of robotic form, type of interaction, and characteristics of affected

individuals, at present time research efforts necessarily tend to focus on proofs of concept. In addition, efforts to define a clear transition model between human-robot engagements and human-human engagements, plans for moving from dyadic child-robot interactions to triadic child-robot-adult interactions, or other structural mechanisms offer the possibility of moving collaborations toward demonstrated clinical utility.

Collaborative studies can provide data to support investigation of questions uniquely asked by each individual community, as well as questions shared by both. Roboticists seek to improve technologies, in order to better investigate the uses of HRI in autism, and clinicians investigate behavioral or biological markers which may distinguish individuals with ASD from those with typical development, as well as cognitive mechanisms which may be activated during interaction with a robot (Diehl et al., 2012). All these are important questions to answer en route toward demonstrations of the clinical utility of robots. As this interdisciplinary field gains knowledge and data, both technologically- and clinically-focused investigations can be iteratively advanced and refined. Studies can simultaneously acquire clinical interaction data necessary for shaping robotics development while investigating the parameters facilitating clinical utility. For instance, roboticists are interested in fully automating robotic perception of, and response to, human actions. However, better understanding of (and data from) heterogeneous behaviors among individuals with autism is needed, in order to inform and train such designs. In the mean time, robots often operate under secret, manual control which affords the (false) appearance of autonomous robotic behavior (the Wizard of Oz paradigm; Riek, 2012; Scassellati et al., 2012). While using Wizard-of-Oz-style control, clinicians can make detailed observations of children's responses to robots, roboticists can acquire data which can inform next-generation autonomous

perception and behavior selection technologies, and both sides of a collaboration can investigate proofs of concept.

### 6.1.2 Study design

A second set of differences exists regarding the typical methodologies employed in each discipline. Evaluations of robotics technology often focus on proof-of-concept, that is, on demonstrations of a system's effect for one or more people (often, $n < 5$). These small numbers are typically constrained by the investigation's focus on demonstrating the viability of the design and implementation, and by the tremendous engineering effort required to construct a robust, reliably functional device. Focus on the technology may initially cause roboticists to overlook the significant resources required to test with specialized populations, the difficulties associated with accessing the target population, and methodological rigor in user testing. To date, clinical validity and applicability have been difficult to gauge in studies of robotic applications for autism. This is due to insufficient provision of standardized characterizations of participants; or to insufficient control allowing comparison between a robot's effects on individuals with and without autism, or comparison between effects of interaction with a robot and that with an alternative device or person (Diehl et al., 2012). The gold standard for proving efficacy of a medical or behavioral treatment is consistency in findings from multiple, independently conducted, randomized, double-blind clinical trials, each of which requires experimenters blind to knowledge of individual participants' assignments to comparative groups, and participants blind to the parameters of the experiment. Practically, however, double blinding can be an extremely difficult standard to meet in autism research, because the differences between participants with ASD and controls is often apparent, and the nature and intention of a given task or intervention can

be obvious to participants and to the experimenter. For this reason, clinical research in autism frequently uses alternative designs in order to evaluate the efficacy of an intervention (or specificity and sensitivity of an assessment). However it is necessary to approach these designs with appropriate levels of clinical rigor. As discussed, for example, by Reichow, Volkmar, and Cicchetti (2008) the clinical autism research community has defined rubrics for evaluating the validity of evidence from experimental interventions, for the purpose of practical dissemination and application. Such standards include using adequately powerful sample sizes for group designs, using appropriate control conditions in both group and single subject designs, and generally obtaining standardized characterizations of participants which can be compared to other research (see Reichow et al., 2008). With respect to study design standards, robotics researchers face a long tradition and deeply ingrained methodology and must adapt to the practices of the clinical community. Clinical standards are also not negotiable within the space of collaboration with roboticists because clinical research standards impact legal, economic, educational, and medical decisions regarding the provision of care to affected individuals (Reichow & Volkmar, 2011). HRI studies for autism, with larger, statistically valid, comparisons have recently begun to emerge (e.g., Feil-Seifer & Matarić, 2011) and the reporting mechanisms for single subject, or case study, designs (which must meet specific design considerations to achieve traction within the clinical community; see Kazdin, 2011; Reichow et al., 2008) have also begun to gain acceptance within the robotics community.

In moving to studies that adhere to clinical standards, more standardized mechanisms for participant recruitment, for reporting population statistics, and for the analysis of data with respect to control groups will become necessary. Many current robot-autism studies

recruit participants in an ad hoc fashion, as obtaining access to populations for many groups is non-trivial, even within collaborations with clinicians. In addition, clear inclusionary criteria and recruitment procedures are essential to ensure a representative sampling, which is the basis of any statistical conclusion.

Along similar lines, clear characterization, as mentioned above, is fundamental for comparison among disparate research findings. Such comparisons, in turn, make possible definition and refinement of the parameters allowing effective application of HRI in autism treatment or assessment and the investigation into the cognitive mechanisms which such applications might engage (Diehl et al., 2012; Reichow et al., 2008). Participants in existing studies often have been described using a flat diagnostic label (or even just as "autistic"). As the expression of symptoms within ASD are extremely heterogeneous and the level of impairment ranges from very mild to very severe, these simple labels are typically not sufficient for providing a clear picture of the diverse abilities and selective deficits faced by these individuals (Diehl et al., 2012). In clinical autism research rigorous characterizations of socio-cognitive abilities is performed for all study participants, using externally validated protocols (Reichow et al., 2008). For example, assessment tools include the autism diagnostic interview–revised (ADI-R; Lord, Rutter, & Couteur, 1994), the childhood autism rating scale (CARS; Schopler, Reichler, & Renner, 1986), and the autism diagnostic observation schedule (ADOS; Lord et al., 2000b). These standardized tools allow for comparison of populations across research studies. These assessments can be lengthy and expensive, as each requires administration by a trained clinician, and each must have been performed close in time to the experimental study, as developmental changes in children with ASD can be substantial over short periods of time. Finally, most proof-of-concept studies from the robotics

community focus exclusively on children with ASD and do not provide a comparative sample of typically developing children or children without ASD having other impairments which are frequently comorbid or symptomatic of ASD, such as intellectual disabilities or specific language delays. A common objection to existing studies is that many of the effects seen when children with ASD interact with robots (especially increased attention, and high motivation) would be seen in any child when they are given a new robot toy to play with. The use of control groups as described above is standard practice in the clinical community, but has only begun to have more widespread, and increasingly standard, usage in robotics. In these aspects, robotics groups will most likely need to adopt the more standardized reporting mechanisms of the clinical community. However, some flexibility from the clinical community must be offered, as very few research groups have the resources to span the range of assessment, engineering design, and large-scale testing required to study a large statistical sample. For those robotics groups lacking access to highly experienced clinicians who have been specifically and rigorously trained in administering ADOS or ADI-R, CARS may present a slightly more accessible alternative, administrable by physicians, special educators, school pathologists, and speech pathologists who may have little experience with individuals with autism. Another, even more accessible but clinically comparable alternative is the Social Communication Questionnaire (SCQ; Rutter et al., 2003), which can be completed by parents or primary caregivers, and which is frequently used in clinical studies to affirm control participants' negative diagnoses.

Also frequently important in clinical research are measures of other kinds of cognitive development, frequently measured with IQ tests such as the Differential Abilities Scale (Elliott, 2007), Wechsler Intelligence Scale for Children (Wechsler, 2003), or with the Mullen

Scales of Early Learning (Mullen, 1995) where individuals may be too young for other tests. We advocate for the reporting of standardized IQ assessments, which we expect may be more readily available, given their utility in a broader range of disabilities and our assumption that professionals trained in their administration may be relatively accessible, particularly through schools.

Clearly there is a tradeoff to be made between resources devoted to characterization and comparability and specificity of characterization, and it is for each particular group of collaborators to negotiate this tradeoff.

### 6.1.3 Publication and dissemination

A final set of cultural differences concerns the timing and location of publication and dissemination of research results. Both the clinical community and the robotics community have their own established publication standards and venues, and the differences between these standards has implications for reporting results, for expectations of young researchers regarding tenure and promotion, and the evaluation of students. High-quality results in robotics typically appear as shorter length papers (6 to 12 pages) in annual conferences, many of which are peer-reviewed, highly competitive venues and result in archival publications. A robotics student might be expected to publish 1-2 such conference papers each year, and a lengthier journal article that covers multiple conference publications appearing every few years. In contrast, the clinical community typically publishes their primary results as longer manuscripts (10 to 30 pages) in monthly or quarterly peer-reviewed, and similarly highly competitive journals. A student in the clinical community might be expected to publish one such paper every few years and to support that publication with the presentation of unarchived posters and talks, at conferences and meetings. These differences

are perhaps the most difficult to overcome as they involve the expectations of the entire research communities who evaluate the work of these scientists, not just the researchers directly involved in the collaboration. An approach used in other interdisciplinary fields is to allow each collaborator to publish directly in their own preferred high-quality venue. This can be difficult in the case of robot-autism research, as publication in an archived computer science conference proceedings can at times block publication in a high quality clinical journal, which expects all of the data reported to be first-run material that has not appeared in other archived publications. It is our experience that publication challenges can be negotiated only by clear communication between the research collaborators about their expectations and needs regarding publication and clear communication of the difficulties involved in these interdisciplinary research issues to reviewers of student performance, tenure and promotion committees, and project reviewers.

### 6.1.4 Suggested bridges for collaboration

Methods in each community are valid within each, and funding and other resources reflect—indeed determine—the expectations each community must satisfy in their research. We suggest ways to negotiate the cultural differences we've outlined above, to foster collaborations which can further efforts toward demonstrated utility of robotic applications to intervention and assessment of ASD.

Ultimately, to be successfully accepted as a diagnostic or intervention tool, a robot's utility must be demonstrated with statistical significance over a large sample. This standard is generally required in the medical community to establish the evidence basis of any diagnostic tool or treatment's efficacy. Obviously there are personally affective, cultural, and legal implications to establishing any treatment as evidence-based. In the case of communication -

interventions, few treatments have met this rigorous standard, and typically only over narrowly targeted behaviors (Prelock et al., 2011). Along the way to this gold standard, there are other effective ways to establish validity within a clinical community. The key here is control. Interventions with broader behavioral targets frequently employ single case experimental designs (for example, changing criterion, reversal, multiple baseline, or alternating-treatment designs; see Kazdin, 2011) to establish non-statistical control over the many other changes developing children with autism may experience at the same time during which they receive treatment. Roboticists facing limited access to clinical resources may wish to consider single subject designs with rigorous control, such as a reversal (ABA) design, in which each participant's behavior is observed (A) before introduction of treatment (e.g., interaction with a robot), (B) just after or while treatment is being applied, and then (A) again, well after treatment has been withdrawn.

With respect to characterization and participant selection, researchers in both fields often face logistical (and funding) limitations on the assessments they can provide, as well as the participants they can recruit. As our understanding of the parameters allowing viable interactions between individuals with autism and robots improves, and as questions of utility become thus more possible to answer, we expect funding to explore specific subpopulations will become increasingly available. In the meantime, often given limited funding, both roboticists and clinicians must collaborate with other ongoing clinical studies having funding which can support expensive assessments. Thus, access to experimental participants is limited to collaboration with existing assessments. Here we suggest a compromise to both communities: that they recognize the intent of most current studies, in the application of robotics to autism, is to establish proof of concept, and that they allow incremental

evolution in the specification of viable parameters; that is, that they forgive such proof of concept studies when their experimental samples are broader or slightly different from what in principle may be the ideal population for the application in question. To make such proofs of concept viable and useful, current research should seek as detailed a characterization as possible, to help further both communities' understanding of the technological and clinical parameters that allow individuals with autism to successfully interact with robots. Generally, we suggest that researchers from both communities recruit the largest number of participants that their resources allow, from the subpopulation whom they anticipate will demonstrate the greatest utility of the robotic application. Where *n* is small, we suggest that researchers design according to well-established single-case methodologies (Kazdin, 2011).

Publication may be the most challenging arena in which to negotiate collaboration. Typically, funding agencies supporting each party will expect first-author publication. How can collaborators split results into two publications without compromising ethics by withholding results from the first publication? There is no perfect solution to this problem. Rather, it is our experience that pre-nuptial agreements can be made (and often require adjustment, depending on results of primary and exploratory analyses), and will often be determined based on funding allocation and who is putting in the most effort and resources. Part of this negotiation can be to identify which research questions are better suited to which community, and then to design experiments and plan analyses according to the planned order of publication. Mechanistic or explanatory analyses tend to require much greater effort, which may be better supported by staff in larger clinical groups. Thus, proof of concept questions, which may require less effort to answer, may be better targets for

robotics publications, especially because roboticists may be less interested in some of the finer analyses. Of course there is a lot of overlap, so negotiation is needed.

As technologies and proofs of concept evolve, collaborations between roboticists and clinicians may find greater opportunities to answer questions about the utility of robots in intervention and assessment. We expect that both communities will find it increasingly useful, at this point, to publish such findings in clinical venues, while technical innovations will likely make a greater impact within robotics venues.

## 6.2 Our collaborative strategy

The studies involving children with ASD that we have presented in this dissertation (Chapter 3 and Chapter 4) illustrate our approach to collaboration, which we hope will help other roboticists and clinical researchers to understand and navigate the cultural differences between their respective fields. Here we present our examination of specific points that we highlighted in our description of differences in practice, using Chapter 3 as a case study of a collaborative strategy, maximizing the advantages of both fields while eliminating the greatest barriers from collaboration and communication.

In terms of our research approach, we chose to focus on proof of concept, that school-aged children with high functioning and ASD would engage and enjoy a verbal task with an inexpensive, commercially produced robot under seamless interactive control. Although we are interested in automating the robot's perception and behaviors, we chose to focus on the proof of concept by using Wizard-of-Oz-style control, and to use our investigation of proof of concept to gather data that may support future technological research into automation. We also furthered our clinical agenda by collecting copious speech data and interaction data,

which we will continue to analyze, to understand in greater detail the ways that participants interacted with the robot and with people afterward.

Our agreement was to publish results from our study (Chapter 3) first as a proof-of-concept manuscript in a robotics-oriented publication venue (E. S. Kim et al., 2012). This was acceptable, and necessary, for several reasons. First, as one of the few larger-*N* studies of robot-child interaction in autism research, the study in Chapter 3 highlights the applicability, acceptability, and potential of social robots to effect meaningful change in children with ASD. Publishing sooner rather than later enables other researchers to see the advantages of these larger designs and the advantages of detailed clinical characterization in informing our understanding of what works and for whom. Second, it is important that roboticists, who will be on the front line of implementing the technically challenging but critical elements of HRI studies of autism, be given ample information regarding the details and hurdles that will help them design similar studies. Early dissemination of the study protocols and provision of usable (though not ideal) metrics of evaluating change will help these researchers adapt their own platforms and speed up development and evaluation time. Third, and perhaps one of the key issues informing our decision to publish these results in a robotics venue first, is that we estimated that the design, creation, implementation, verification, and evaluation of more detailed measures of interview dynamics, prosody, and semantic content could take approximately 3 months of time at our available level of funding. Factoring in additional statistical analyses and rigorous accounting of individual participant characterization variables, we estimate that the next iteration of these study results could take 5 months.

This decision did not come about haphazardly, but instead reflects our lengthy discussions and a priori agreements well in advance of the start of the study. Of course, research is not a static process, and, when dealing with such a new field such as HRI studies of autism, it is difficult to predict exactly what methods, techniques, protocols, and measures will bear fruit. Here we were guided by clinical insights that informed our study design in advance, and a long-term collaboration built around understanding each party's expectations. We expected that it would be necessary to publish preliminary analyses and proof-of-concept before a final, more detailed examination could fully explore the space of our results. The clinical members of our research team, in turn, expect (and it is our expectation, will receive) our full support in the second iteration of analyses.

Such agreements come also with consequences. First, because we froze the current state of analyses to publish the study in Chapter 3, the measures that we employ are necessarily coarse, and to an extent, incomplete. This study could benefit, for instance, by detailed ratings of affect and engagement during the interviews. Consistent with validity standards in the field, this would also require a second rater to confirm the accuracy and reproducibility of the more qualitative assessments. This study could likewise benefit from a careful transcription of verbal exchanges during interviews, complete with timings of utterances. We could then distill from these data sets measures relating to the frequency of verbal production by the children, the semantic content of their speech, and the dynamics of the conversation between the child and the interviewer. Finally, difficulties in obtaining reliable operationalized protocols for evaluating prosodic quality in interviews for children with ASD suggest that standard approaches need to be adopted to capture more subtle prosodic differences between study participants with ASD and the control group.

The second consequence of our decision to publish these results in a robotics venue first is that it may make it more difficult to later publish results regarding the second iteration of analyses in a more clinically themed journal. It was the opinion of the clinical members of our team that though this concern was valid, the more detailed and clinically-oriented second round of analyses and interpretations should make the second manuscript quite distinct from the first. In other words, it was a risk that everyone was willing to take.

### 6.2.1 Understanding Differences in Approach

At the intersection of robotics and autism research, differences in approach result in a number of potential pitfalls. Researchers in engineering fields typically focus on the development of methodologies, approaches, and processes. By contrast, researchers from clinical fields focus on specific issues relating to clinical populations. While the robotics community tends to focus on novel platforms for delivering treatment, the clinical research community focuses primarily on the treatment itself. A researcher in the robotics community gains greatly from expanding the vision of the possible, and so a successful proof of concept is in many ways a sufficient enterprise in and of itself. Yet, applications tied only to proof-of-concept studies, even though they may provide great benefits to a clinical population, may be left languishing in the land of "potential ideas" for years without a direct translation of those ideas into clinical applicability. This is quite a dangerous position, because without feedback from researchers focusing on clinical utility, the robotics community may drive novel technologies in unproductive directions while neglecting application areas that may have demonstrable clinical impact. Similarly, approaches that focus exclusively on tried-and-true engineering tools and platforms may be left languishing in the equally perilous land of "outdated technology" when more modern and capable technologies provide possible

solutions that could not have been considered with more mature technologies. Without attending to the rapidly changing landscape of technical advancement, clinicians face the difficult prospect of struggling to adapt technologies that have already been replaced with more convenient, efficient, or capable solutions.

The study outlined in Chapter 3 illustrates a way in which healthy collaborations between robotics labs and clinical enterprises can be formed. Beyond the typical skills that are necessary for any collaboration to succeed (e.g. mutual respect, open dialogue, rapid feedback), it was necessary for both our groups to understand our respective differences at a much deeper level.

### 6.2.2 Understanding differences in study design

As mentioned above, the focus on technical novelty and innovation in the robotics community differs from the focus on clinical utility in the clinical research community. This also has implications for the methods that are the standard for each field. With respect to design, we chose to prioritize proof of concept over technological development. For instance, we feel that speech recognition innovations will be required in order to replace Wizard of Oz with automation, but we have decided to justify such an investment first with a demonstration of a highly socially responsive robot, whether automated or manually operated.

#### 6.2.2.1 Sample sizes

In the robotics community, a proof-of-concept paper may include 1-6 participants with developmental disabilities. This is sufficient to illustrate the technical advances of the robotics platform, show feasibility, and provide a glimpse at the potential of the advances.

However, studies that aim to demonstrate clinical utility involving just a few participants are often regarded by clinicians and developmental researchers as being questionable and insufficiently powered to identify reasonable trends, even if effect sizes are large and results are statistically significant. A recent survey by Diehl et al. (2012) indicated in an extensive review of robotics work in autism that only six studies have involved more than six participants with ASD, and in this context discusses the need for larger and more rigorous studies to better define the role robotics can play in autism research.

In Chapter 3's study, we collected data from nearly 20 participants with ASD and 10 TD controls. This represents the largest group of participants with ASD in a robotics study to date. We should note that while a large sample size is advantageous for identifying robust positive findings, it is even more valuable in the context of interpreting negative findings. In Chapter 3's study, we found that TD children did not increase in their post-robot interview time as compared to their pre-robot interview time, whereas participants with ASD did. We went so far as to mention that we may be less enthusiastic about the negative result identified in the TD group, given the small sample size. However, it is important to note that this is still far larger than 90% of control groups employed in robotics papers reviewed by Diehl et al. (2012). In this fashion, our perspective on sufficient group sizes was heavily influenced by our clinical team members' expertise, a perspective that helps us to strive for higher standards in robotics-autism research.

#### 6.2.2.2 Clear characterization

One of the most pressing challenges presented to researchers studying ASD is, as identified by the Interagency Autism Coordinating Committee (2011), the heterogeneity present in the disorder. While the definition of autism spectrum disorders can be neatly summarized by a

single reference to the DSM-IV (American Psychiatric Association, 2000)[7], the complexity and heterogeneity of the autism spectrum (e.g., see Happé, Ronald, & Plomin, 2006) is easily overlooked by researchers with limited autism experience. Characteristics of individuals with ASD range from extremely high intelligence and relatively subtle communicative or social difficulties, to no language ability, comorbid and debilitating intellectual disabilities, and almost non-existing social function. Even within a relatively "high-functioning" group of individuals with ASD, behavioral and cognitive characteristics can range widely: verbal communication can be difficult to elicit or flow unceasingly, visual-spatial competency can be average or remarkably superior, adaptive functioning can be well preserved or severely impaired. Understanding the nature of these individual characteristics can often be a nuanced and subtle process, requiring high levels of clinical insight and care to decipher (e.g., see Karmiloff-Smith, 2006). In other words, knowing that the target population has ASD is necessary but not sufficient to understand all of the complexities of an experimental interaction with robots. Ideally, we would know in advance which subpopulations to target, and the expected behaviors of the targeted subpopulations on selected outcome measures. However, given the nascent state of our interdisciplinary field, such knowledge, is often unavailable at the time of experimental design. For this reason, larger-*N* proofs of concept and exploratory investigations are critical for the understanding of heterogeneity in behavioral responses among individuals with ASD, and thus essential to the advancement of robotics research in autism.

---

[7] Again, please note that all participants with ASD presented in this dissertation, including Chapter 3's study, were diagnosed according to DSM-IV. DSM-5's definition of autism spectrum disorders differs from that in DSM-IV.

In Chapter 3's study we collaborated with leading experts in speech and language pathology in ASD, coordinating with a team of expert clinicians and researchers in ASD. Their added insight was extremely valuable, and greatly enhanced the interpretability of this study. For instance, the clear clinical guidelines they provided indicated that the individuals comprising the ASD group indeed were all affected by ASD. We were also able to establish that, despite the several negative findings involving between-group differences of affect, engagement, and pre- and post-robot interview times, these results did not hold up for all individuals with ASD. Instead, we found that the higher-functioning participants with ASD responded more enthusiastically to the study, possibly suggesting that the particular paradigm employed in Chapter 3's study might be most engaging for individuals with PDD-NOS or Asperger syndrome, who typically exhibit less severe autism symptoms than children with autism (Walker et al., 2004).

#### 6.2.2.3 Rigorous metrics and statistical considerations

Data from HRI studies typically employ a structure that lends itself to standard statistical analyses; participant groups are of equal size, drawn from the same population and tested under equal experimental condition. The standard statistical analyses conducted on these studies (typically, t-tests and ANOVAs) are subject to assumptions based around this standard format. Even within the clinical literature, standardized approaches rely on statistical methods that provide value only when these assumptions hold. As studies at the interface of robotics and clinical research must often depart from these traditional formats, whether due to the heterogeneity and availability of the target population or the experimental and adaptive nature of the technology, analysis and interpretation of even large volumes of data must be done carefully and with respect to these underlying assumptions.

The study in Chapter 3, while somewhat elementary in its statistical needs, benefited from a careful examination of assumptions inherent in the selected statistical tests. In addition, the choice to use pre- and post-interview times as surrogates for self-motivated verbal elaborations, in the absence of more refined measures on changes in behavior, was aided by perspectives provided by multiple investigators. As this study matures in its analysis, the benefits and interpretability of the results will be greatly aided by collective emphasis on rigorous statistical modeling and the selection of the most appropriate outcome measures for analysis. Similarly, the lessons learned from this study, in partnership with clinical experts, will help pave the way for the design of future studies aimed at isolating specific properties of robots that are most important to effecting change in children with ASD.

### 6.2.3 Understanding perspectives on publication and dissemination

At the fundamental level, researchers from the robotics community and clinical researchers have a lot in common. They share the same high levels of inquisitiveness and curiosity, the same desire for rigorous truth, and the same goal of leveraging science to improve our understanding of the world and the lives of others. Yet, despite this, the language and perspectives of robotics researchers and clinical scientists can be very disconnected and a concerted effort to educate our collaborators in both fields must be made regarding publication venues.

First, clinical researchers may not understand the scope and magnitude of a robotics conference paper. To gain that perspective, they sometimes have to be informed that high-impact conferences may have similarly, or even more, competitive submission processes than prestigious journals. Furthermore, it is often not clear to clinical collaborators the great importance that conference publications have for career advancement among junior

roboticists, engineers, and computer scientists. For this reason, clinical partners to robotics laboratories may question whether it is worthwhile to devote time and resources toward the development of a well-written conference paper; Patterson, Snyder, and Ullman (1999) provide a succinct discussion of the impact of conference publications on the evaluation of computer scientists.

Second, whereas publishing an abstract in a psychology or other social science conference typically will not hinder publication of a corresponding journal article, in submitting a full-length, archived computer science conference paper that summarizes all clinical results may preclude publication in a peer-reviewed journal. The reason is that many, especially high-profile, journals have extensive requirements for innovation and novelty of work presented; that is, journals tend to actively prohibit the reporting of results which have been detailed in print elsewhere, whether prior to, during, or immediately after submission of the journal manuscript.

There are several options joint robotics-clinical collaborations can choose when deciding where and when to publish. First, they can forego conferences altogether, in favor of waiting to submit results to an appropriate journal. This has the advantage of maximizing the chances that study will be able to be accepted to journals, but runs the risk in fast-paced technology research areas of closing opportunities to be the first group in the field to publish concomitant technological advancements, while waiting for journal publication, which typically take longer than conference papers to submit, review, and publish. In addition, a publication in a journal with a clinical focus may not contribute to evaluations of a robotics researcher, when competing for grants and positions, under evaluation by other computer scientists and engineers; these evaluators may prefer high-quality conference publications in

technological fields. A second option is to publish the study in a conference first. This, of course, may raise problems concerning the novelty of following journal submissions, which will likely impact clinical scientists most. A third collaborative solution, such as the approach that we took (E. S. Kim et al., 2012), is to publish work-in-progress that can document the sophistication and innovation of the technical aspects of the study, over a preliminary population or analyses in progress; later, a following journal submission can represent results from a larger sample or more extensive analysis, either of which is likely to be considered a significant advance over—and thus a finding distinct from—the initial conference publication. In the case of the study described in Chapter 3, we chose to present data most relevant to the robotics community (i.e., findings about gross engagement with the robot and about possible indicators for the most appropriate target population) within this robotics venue, while reserving additional analyses of specific behavioral impact for a later publication in a clinical venue. Note, we do not advocate hiding, or "trimming" of data to achieve this collaborative negotiation; such an approach could present ethical challenges, since scientists are expected to report results as fully as possible. Rather, as we did, we suggest targeting research questions to robotics and clinical publication venues during experimental planning, and, where necessary and possible, the freezing of analyses while publications are pending.

In all cases, collaborators should establish a clear dialogue early, and should negotiate publication plans in advance, to best avoid conflict and to maximize the mutual benefits of the joint project. Roboticists' careers, and their relationships with funders and other evaluators, could be injured by surprise decisions, at the conclusion of extensive technological development and data collection, that results cannot be published for as long as a year. Likewise, clinicians who have heavily invested time and resources into a study

would face problems with career development and evaluation, if faced with a surprise rejection from a journal due to previous technical conference publication.

### 6.2.4 Establishing common ground by minimizing risk

Collaboration can succeed only if involved parties communicate effectively; this, in turn, requires that each understand the others' motivations, needs, and resources. A common ground, though perhaps not as noble as one would like, is mutual self-interest: the roboticist very much wants to see his or her platform used; the clinical researcher very much wants to provide new avenues for effecting positive changes in the population that is his or her expertise. It is important to consider that such a pairing poses significant risks to both sides of the collaboration: by pairing with clinicians and developmental experts, the roboticist takes a chance that his or her proof of concept may ultimately advance to a demonstration of non-effectiveness; the clinical expert, by wagering on a new technology, risks spending valuable clinical resources (especially personnel time and access to participants from a small and specialized population) on the exploration of nascent technology, instead of on investigation of better understood, and thus less risky, paradigms.

It is useful to understand the risks each community faces from a financial perspective. Robotics work is design- and development-heavy: much of the costs associated with creating a new robotics platform involve design work, machining, programming, and countless hours of trouble-shooting. Clinical work, especially experimental trials, are delivery-heavy: much of the costs associated with running a successful clinical research enterprise involve careful study design, an extended period of experimental delivery, and rigorous statistical analysis and interpretation. Development time, in the robotics community, is measured in months, and experiment delivery time is measured in weeks. In the clinical research community, these

time frames tend to be reversed. This means that in robotics work, time is largely spent in the process of rapid prototyping, deployment, and re-development. On the other hand, clinical partners will spend most of their time conducting the same trial over and over again.

Our collaboration in Chapter 3's study began with considerations to minimize risk to our clinical partners. First, this entailed ensuring that the robot platform was free from glitches, crashes, and other issues that might interfere with the delivery of the experimental protocol. Our debugging and testing phases were far more extensive than would have been usual for a non-collaborative proof-of-concept study. Second, interfaces between the robot and the experimenter controlling the robot were robustly designed; guaranteeing that rapid response to the behaviors of children could be accommodated. Roboticists without extensive clinical experience may overlook the potentially terrific expense required to conduct a rigorous experiment with special populations. Recruitment can be difficult, especially for less prevalent disorders. Access to a specific age-range or a subgroup of individuals with specific characteristics in addition to the disorder itself (e.g. *higher functioning 10- to 12-year-old children with ASD*), which is useful in controlling the experiment from a statistical vantage, can make recruitment even more difficult. Furthermore, clinical characterization requires both tremendous coordination of staff and considerable personnel costs. In other words, even besides study and platform design costs, expenses per participant can be quite high (upwards of several hundred dollars per participant). These costs, in addition to the importance of consistency, make mistakes in this work prohibitively expensive. Third, while roboticists often benefit from demonstrating innovation or proof-of-concept using expensive, one-of-a-kind prototypes, the potential cost for damage to or destruction of these prototypes makes involvement in a clinical environment potentially prohibitively expensive. Our efforts

leveraged a commercially available robot platform that could be easily replaced with minimal cost when damaged during a clinical visit. While this risk analysis is particular to our two research groups, a similar analysis of risk can be of great benefit in advancing an initial interdisciplinary conversation to a long-term and viable collaborative effort.

## 6.3 Conclusions

In addition to the technological and experimental efforts described in this dissertation, we have found that robot-autism research requires thoughtful, systematic project management. In this chapter we have described observations and suggested resolutions, developed in collaboration with my advisor Brian Scassellati and our clinical partners Rhea Paul and Frederick Shic, to differences between roboticists' and clinical autism researchers' investigatory goals, approaches, and expectations. We share a hope that these novel analyses and guidelines will help accelerate future investigations of socially assistive robots to toward clinical efficacy.

# Chapter 7

# Discussion

We have been looking at HRI applications for autism intervention along four overarching research questions. We have presented five novel studies, which contribute new knowledge and techniques in all four of these domains. In addition, we have described the complex challenges of interdisciplinary efforts shared between clinical and robotic researchers, and offered general solutions as well as two specific examples of successful collaborations.

**Question 1: How engaged and motivated are participants?**

A child's motivation to engage in an intervention can influence the intervention's efficacy (R. L. Koegel, O'Dell, et al., 1987). Particularly for the behavioral therapies targeted by social robotics applications, participants learn by practicing the targeted behaviors. The more motivated a child is to perform a behavior, the more opportunities therapists have to help him or her improve. Given the high prevalence of interest in machines and devices found among children with ASD, the use of robots as a highly motivating object is the fundamental motivation of HRI research autism interventions. In this dissertation we have presented a study (in Chapter 3) that shows, over large groups, that school-aged children

with ASD ($n$ =18) and children in a control group ($n$ = 11) are motivated to socially interact directly with robots, which indicates the potential to use robots as developmental scaffolding, that is as practice social partners en route to improved social interaction with humans. Chapter 4 replicated this finding (though only for children with ASD). Both chapters also provide evidence that children with ASD also interact more with a confederate in a therapist-like role while or immediately after, respectively, interacting with a robot. These findings, in turn, indicate the potential of a social robot as an embedded reinforcer. An embedded reinforcer can operate either by reinforcing children's social interacting with other humans or with the robot itself, by being implicitly interesting and by interacting with them socially.

Evidence presented in Chapter 4 also indicated that children with ASD increased time spent speaking with an examiner, following the robot interaction, more so than did the control group, suggesting that interaction with the robot has slightly different effects on TD and ASD children's affect or social motivation immediately following interaction.

Because children with ASD often face extraordinary challenges in daily living, because research with individuals having ASD requires additional ethical consideration and assessment, because children's participation in research requires parents' attendance, because the size of the population having ASD is limited, and finally because it is more it is before examining them, we first sought to establish typical adults' motivation to interact with robots. Chapter 2 presented new findings, over a group ($N$ = 27) of typically developing adults, that they spontaneously use speech, and express intense affect, to interact with a robot that they cannot touch. These findings indicated typical adults' motivation to interact

socially with a robot, and motivated us to pursue similar investigations with children with ASD and TD.

Chapters 2, 3, and 4, were all novel in their demonstration of these phenomena over large groups. Chapter 2 provided the first group evidence, to our knowledge, that adults will spontaneously use speech to interact with a robot. Chapter 3 presented the first large, controlled, group study to demonstrate a social robot's ability to facilitate social interaction with another person. This is also the first study to show this effect for older and higher-functioning children with ASD, whereas previous demonstrations have been presented in small-number case studies of younger children with lower functioning (Feil-Seifer & Matarić, 2009; Kozima et al., 2009). Chapter 4 presented the first large, controlled, group study to demonstrate a social robot's ability to elicit robot-directed speech and social engagement over a large group of children with ASD.

Chapter 3 also contrasted behavior with a robot against behavior with a computer game. Participants' far-reduced speech toward both the therapist-like confederate and the device in the computer game, in comparison to the robot condition, affirms our theory that despite their interest in the video game, the interaction, designed without sociality, could not sustain social interaction with the device or mediate social interaction with another person. This design is limited in the sense that it cannot speak fully to the question of robots as embedded reinforcers: a better test would pit a computer game against a robot in interaction, in parallel semi-structured interaction, as was enforced between the adult and robot conditions. A more parallel robot-computer game comparison would allow us to discern whether and to what extent a robot can afford an advantage in eliciting or mediating social interaction from children with ASD. That is a worthwhile question to answer because

computer games may be easier and less expensive to maintain and modify than robots and their forms and behaviors. However, we have reason to believe that for some social behaviors, such tests will reveal that robots do pose a clinically important advantage over computer games. Our group has shown that a group of typical adults is less willing to afford authority and personal space expected for an agent when interacting with a live, responsive video-feed of a robot on a computer monitor than with a robot whose body is physically present, that is, collocated (Bainbridge, Hart, Kim, & Scassellati, 2011).

### Question 2: How do participants behave?

Both this and the question of motivation examine participants' behaviors. Question 2 describes what specifically participants do (toward the robot, other people, the environment, and generally, during therapy). Answers to this question can inform possible target behaviors in therapies, both in interventions which will use robot-directed behaviors as stepping stones toward improved human-directed behaviors, as well as on future interventions which will use the robot to mediate interaction with other people.

In Chapter 2, we observed that a large group of naïve typical adults spontaneously used speech and affective prosody to interact with a robot. This was the first group study to document untrained people's spontaneous use of either speech or affective prosody toward a robot. Whereas Breazeal and Aryananada's investigation of affective prosody in robot-directed speech was collected by explicitly directing two adult females to perform prosodic expressions of five types of affect to interact with a robot, in our study in Chapter 2, 27 naïve, untrained adults spontaneously directed affectively expressive prosody toward a robot. Also, the audio recordings in Chapter 2 provide a sample of spontaneously produced,

affectively expressive prosody, from which we may be able to train automatic systems to recognize and classify affect.

Chapter 3 presented results that for a group of school-aged children with high functioning ASD, a social robot can mediate greater verbalization than a social (but less preferred) interaction partner, an adult human. We have shown that a robot elicits greater verbalization than a preferred but asocial interaction partner, a computer game. More importantly, a social robot increases social interaction with another person, more so than an adult or a computer game. These findings suggest that robots, with appropriate clinical guidance, may make useful supplements to communication and social skills interventions by facilitating social interaction with an adult, and by eventually being developed into uniquely embedded social reinforcers. Children with high functioning ASD will speak to a robot without special training.

Chapter 4 contributed evidence that children with TD and with ASD will direct affectively expressive speech to a social robot. The study in Chapter 4 also contributes data which can be used in the future for explorations of automatic affective prosody recognition in children with ASD. We also found that children with ASD extend their conversational engagement with a clinician following interaction with the robot, more than do their typically developing peers, and choose to freely play with the robot longer than their typically developing peers. Chapter 4 also introduced face-to-face orientation as a novel measure of social engagement, which we observed far less in participants with ASD than in their peers with TD.

In Chapter 5, we described two systems for analyzing speech prosody. The first, an online robotic learner uses real-time, automatic perception of affective prosody as an input

for online learning, the first system to learn from prosody. The system was demonstrated to refine a single-arm, social waving behavior, taught by a single user, distilling the human teacher's speech into a binary valence signal, which it treated as a reward in reinforcement learning. Despite its limitations, this system demonstrated that is possible to recognize affective valence from speech prosody in real time, with minimal training data from the speaker. This indicates further exploration of automatic, real-time recognition of prosodic expressions.

In the second study introduced in Chapter 5, we describe a novel algorithm for semi-automatically recognizing a speaker's prosodically expressed mutual belief signals. A speaker uses these to communicate whether the speaker intends the modulated utterance to contribute new information; to contrast with previously shared information; or to be considered sarcastically or doubtfully. The system's modest success in identifying these shared belief cues in infant-directed speech suggests the possibility of doing the same in robot-directed speech.

Both studies described in Chapter 5 indicate that spoken behaviors may be viable targets for robot-based interventions. In addition, these studies provide data that may provide the basis for automating recognition of speech prosodic expressions.

**Question 3: What design elements support interaction?**

Whereas the previous two questions address human participants' behaviors in HRI, this and the next question address elements important in robot and interaction design.

Chapter 2's study revealed contradictions to assumptions made by classic machine learning algorithms. When teaching, people provide other inputs well before feedback, which is typically modeled as following the completion of a learning trial. People's spoken input to

a robotic learner decreases from one trial to the next, if a learner consistently performs optimally. On the other hand, if a poorly performing learner suddenly improves, people increase the amount of input they provide to the learner. In other words, human teachers' spoken inputs to a learner should not be modeled as independent from one trial to the next; human teachers' spoken input depends on the learner's performance history.

As we discussed under Question 1, the difference between ASD participants' responses to the robot and video game conditions in Chapter 3 tells us that we must design sociality into any interaction with an engaging technology. And that we should expect that even with sociality designed into interaction with an engaging computer game, there will be some social behaviors that physical embodiment can elicit better than can two-dimensional graphics (Bainbridge et al., 2011).

In Chapter 5's first study, we found that differences and improvements in our learner's waving behaviors were hard for even an expert to discern. When designing a robot for human interaction, we must make the robot's behaviors legible and possibly check for people's understanding of them.

**Question 4: How should a robot adapt to maintain a long-term relationship?**

Chapter 2 suggested that human expectations of a learner's behaviors depend on past performance. Ignoring or violating these expectations is likely to limit a robot's ability to maintain a long-term relationship with a human partner. Also, if we do not develop our understanding of such typical social expectations, then an intervention focused on interaction with a robot risks missing the opportunity to reinforce adaptive social behaviors, or worse, risks reinforcing maladaptive behaviors. Our findings, thus indicate a need to

further explore and describe typical human expectations of a social partner's behaviors over the course of a longer relationship.

Chapter 3 and Chapter 4 described experiments in which children with ASD (and TD in Chapter 4) briefly interacted with a robot during a single visit. They do not provide much information about how robots must adapt to maintain long-term relationships. However, the engineering and personnel in delivering a single, brief robot interaction suggest the importance of striving to develop greater reliability in the operation of robots and longer battery life (for example, as the Pleo robot's batteries aged, we could not reliably expect more than 30 minutes of battery life), as well as the importance of automating robot behaviors to reduce the training and time required to manually operate the robot. Both of these engineering endeavors will offer savings in labor that scale at least linearly over the course of a treatment requiring many repeated interactions, and allow researchers to devote more resources toward adapting the robot's behaviors to support a long-term relationship between participants and the robot. Unfortunately, of course, both engineering goals extend along a distant horizon. In the meantime, our experience with single, brief robot interactions also indicate that careful protocol design can go a long way to smooth over minor technical failures. For instance, because we hid the television remote that controlled the Pleo robot, the signal occasionally failed to reach the robot's infrared receiver, introducing a delay in the robot's response to a participant's action or utterance. In such cases, the confederate could ask the participant to repeat the action or utterance, or could provide an interpretation of the robot's delay to ameliorate disruption of the illusion of the robot's autonomy (e.g., "Oops, Pleo must not have heard you!"). As we await each next update in the reliability and automation of robotic technology, we can design interaction protocols that allow the

interactions and robot behaviors to adapt and maintain a long-term series of therapeutic interactions.

In the first study in Chapter 5, we learned that it is possible to learn from input prosodically expressed affective feedback. This proof of concept motivated us to design the second study in Chapter 5, to better understand how people use prosody to teach a robotic learner. As described in greater detail, in Section 1.3, machine learning from human input may help us automate adaptive robotic behaviors to support long-term human-robot relationships

## 7.1 Design and methodological contributions

In Chapter 6 we described methodological, funding, and publishing differences between the autism clinical research community and technological researchers, and suggested ways collaborations can work through them. Although all parties developing HRI for autism share a common, ultimate goal of assisting individuals and families affected by autism, there are also fundamental differences, which lead to differences in approach: roboticists build their careers on innovating new technologies, while clinicians and clinical researchers build theirs on studying questions that are most likely to reap substantial benefits for individuals with ASD. This results in vastly different allocation of resources and different publication timescales and priorities (rapid high-quality conference publications in robotics me, and journal publication which may take a year longer than in robotics to prepare and go to print). By understanding these differences, collaborators can negotiate each side's contribution of resources and a publication schedule that serves both parties.

## 7.2 Conclusions

This dissertation contributes behavioral observations, robotic and algorithmic designs and systems, theoretical frameworks, and collaborative strategies that develop the clinical utility of socially assistive robots for autism interventions. In three well-controlled, large group studies, we have described a social robot's elicitation and mediation of social behavior, engagement in structured tasks, and enjoyment, among adults and children with typical development and among children with ASD. These are the first studies of large groups to demonstrate these behaviors, whereas previous demonstrations of such behaviors were limited to small numbers ($N \leq 6$). Our findings establish the acceptability and usability of social interactions with robots for these populations, and propel the feasibility of clinically useful robots forward by a large, if still early, step. Engagement and enjoyment are indications of motivation to participate. Motivation, in turn, is considered fundamental to social skills or communication interventions. Therefore, our findings suggest that robots reinforce communication and social engagement with other people, as well as with the robots themselves, all within structured interactions. Most importantly, we have shown, in comparisons against another person and another novel and engaging technology, that a social robot can facilitate greater communication with an interventionist. This suggests that social robots may be uniquely reinforcing of social interaction with other people.

We have described four broad research questions whose investigations have and will continue to underlie the development of clinically effective socially assistive robots. The behavioral observations and robotic and algorithmic systems we have contributed further our knowledge along all four of these questions. We have also described a novel theoretical framework of embedded reinforcement through which to understand and explore the

therapeutic utility of robots that elicit robot-directed communication and social behaviors, and that mediate or facilitate human-directed behaviors.

This dissertation describes two systems that automatically recognize prosodic communication and one that utilizes the output of its prosody classifier as input for online, real-time learning from a human tutor. These systems provide proofs of concept that prosodic communications can be automatically classified. The learning system provides proof of concept that in real time, a robot can use automatic perception to drive behavior selection and response. Although these systems and demonstrations are limited and have not been extended to children with ASD, they suggest the feasibility and future exploration of such extensions and future exploration.

Finally, we present systematic suggestions, developed with my advisor Brian Scassellati and our collaborators at the Yale Child Study Center, Rhea Paul and Fred Shic, that address challenges in interdisciplinary collaboration, which we argue have historically slowed or even stymied clinical uptake of assistive technologies. We have developed collaborative strategies over the course of our investigations of human-robot interaction among children with autism, which have progressively increased clinical interest in our research. We argue that consideration of the parameters of collaboration, which we have identified as particularly challenging, will help drive socially assistive robotics further toward clinical efficacy.

# Bibliography

American Psychiatric Association. (1994). Diagnostic and Statistical Manual of Mental Disorders: Diagnostic Criteria From DSM-IV. American Psychiatric Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Arlington, VA: Author.

Ananthakrishnan, S., & Narayanan, S. S. (2008). Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *Audio, Speech, and Language Processing, IEEE Transactions on, 16*(1), 216 –228. doi:10.1109/TASL.2007.907570

Andrews, S., Warner, J., & Stewart, R. (1986). EMG biofeedback and relaxation in the treatment of hyperfunctional dysphonia. *International Journal of Language & Communication Disorders, 21*(3), 353–369. doi:10.3109/13682828609019847

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents. *International Journal of Social Robotics, 3*(1), 41–52. doi:10.1007/s12369-010-0082-7

Bartelmus, C., & Scheibler, K. (2008, June). LIRC - Linux Infrared Remote Control. Retrieved June 1, 2008, from http://www.lirc.org/html/

Beckman, M. E., & Elam, G. A. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation, 3*. Retrieved from http://128.59.11.212/~julia/courses/old/cs4706/hw/tobi/labelling_guide_v3.pdf

Beckman, M. E., & Hirschberg, J. (1994). The ToBI annotation conventions.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). Retrieved from http://ling.ohio-state.edu/~tobi/JunBook/BeckHirschShattuckToBI.pdf

Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M. P., & Tomlinson, B. (2002). Integrated learning for interactive synthetic characters. *ACM Trans. Graph.*, *21*(3), 417–426. doi:10.1145/566654.566597

Boersma, P., & Weenink, D. (2010). {P}raat: doing phonetics by computer. Retrieved from http://www.praat.org

Breazeal, C., & Aryananda, L. (2002). Recognition of Affective Communicative Intent in Robot-Directed Speech. *Autonomous Robots*, *12*(1), 83–104. doi:10.1023/A:1013215010749

Broekens, J. (2007). Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning. In T. S. Huang, A. Nijholt, M. Pantic, & A. Pentland (Eds.), *Artifical Intelligence for Human Computing* (pp. 113–132). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-72348-6_6

Campbell, D. J., Shic, F., Macari, S., Chang, J. T., & Chawarska, K. (2012). Subtyping Toddlers with ASD Based on Their Scanning Patterns in Response to Dyadic Bids for Attention. Presented at the International Meeting for Autism Research (IMFAR), Toronto, ON, CA.

Carter, A. S., Davis, N. O., Klin, A., & Volkmar, F. R. (2005). Social development in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive*

*developmental disorders* (3rd ed., Vols. 1-2, Vol. 1, pp. 312–334). Hoboken, NJ: John Wiley and Sons.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669. doi:10.1017/S0305000903005701

Clark, G. (2003). Cochlear Implants: Fundamentals and Applications. Springer.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. doi:10.1037/h0026256

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces* (pp. 193–200). New York, NY, USA: ACM. doi:10.1145/169891.169968

Dautenhahn, K., & Billard, A. (2002). Games children with autism can play with Robota, a humanoid robotic doll. *Universal access and assistive technology*, 179–190.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, *12*(1), 1–35. doi:10.1075/pc.12.1.03dau

Dediu, H. (2011, 08). Nearly 75% of iPhones are in use outside the US | asymco. Retrieved from http://www.asymco.com/2011/01/08/nearly-75-of-iphones-are-in-use-outside-the-us/

Diehl, J. J., & Paul, R. (2011). Acoustic and Perceptual Measurements of Prosody Production on the Profiling Elements of Prosodic Systems in Children by Children with Autism Spectrum Disorders. *Applied Psycholinguistics*, *FirstView*, 1–27. doi:10.1017/S0142716411000646

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, *6*(1), 249–262. doi:10.1016/j.rasd.2011.05.006

DogsBody & Ratchet Software. (2008). MySkit - Performance Editor for PLEO. Retrieved April 27, 2013, from http://www.dogsbodynet.com/myskit/index.html

Doniec, M. W., Sun, G., & Scassellati, B. (2006). A demonstration of the efficiency of developmental learning. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on* (pp. 5226–5232). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1716827

Duquette, A., Michaud, F., & Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, *24*(2), 147–157. doi:10.1007/s10514-007-9056-5

Elliott, C. D. (2007). *Differential Ability Scales-II (DAS-II)*. San Antonio, TX: Pearson. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8338-820

Feil-Seifer, D., & Matarić, M. J. (2009). Toward Socially Assistive Robotics for Augmenting Interventions for Children with Autism Spectrum Disorders. In O. Khatib, V. Kumar, & G. J. Pappas (Eds.), *Experimental Robotics* (Vol. 54, pp. 201–210). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://www.springerlink.com/content/l2k004r536p73nl6/

Feil-Seifer, D., & Matarić, M. J. (2011). Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features. In

*Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 323–330). Lausanne, Switzerland: ACM. doi:10.1145/1957656.1957785

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. doi:10.1016/S0163-6383(85)80005-9

Fernald, A. (1989). Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message? *Child Development*, *60*(6), 1497–1510. doi:10.2307/1130938

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. doi:10.1037/0012-1649.27.2.209

Fernald, A., & Simon, T. (1984). Expanded Intonation Contours in Mothers' Speech to Newborns. *Developmental Psychology*, *20*(1), 104–13.

Ferrier, L. J. (1985). Intonation in discourse: Talk between 12-month-olds and their mothers. In *Children's language, Vol. 5* (pp. 35–60). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Fleishman. (2000, December 14). Furby Hacker Tinkers, Then He Simplifies. *The New York Times*. Retrieved from http://www.nytimes.com/2000/12/14/technology/furby-hacker-tinkers-then-he-simplifies.html

Garnica, O. K. (1977). Some prosodic and paralinguistic features of speech to young children. *Talking to children: Language input and acquisition*, 63–88.

Gentil, M., Aucouturier, J.-L., Delong, V., & Sambuis, E. (1994). EMG biofeedback in the treatment of dysarthria. *Folia Phoniatrica et Logopaedica*, *46*(4), 188–192. doi:10.1159/000266312

Goldstein, H. (2002). Communication Intervention for Children with Autism: A Review of Treatment Efficacy. *Journal of Autism and Developmental Disorders, 32*(5), 373–396. doi:10.1023/A:1020589821992

Gopnik, A. (2001). Theory of mind. In (F. C. Keil & R. A. Wilson, Eds.)*The MIT encyclopedia of the cognitive sciences.* Cambridge, MA, USA: MIT Press.

Grossman, R. B., Bemis, R. H., Plesa Skwerer, D., & Tager-Flusberg, H. (2010). Lexical and Affective Prosody in Children With High-Functioning Autism. *J Speech Lang Hear Res, 53*(3), 778–793. doi:10.1044/1092-4388(2009/08-0127)

Hagedorn, J., Hailpern, J., & Karahalios, K. G. (2008). VCode and VData: Illustrating a new framework for supporting the video annotation workflow. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 317–321). Napoli, Italy: ACM. doi:10.1145/1385569.1385622

Handleman, J. S. (1979). Generalization by Autistic-Type Children of Verbal Responses Across Settings. *Journal of Applied Behavior Analysis, 12*(2), 273–282. doi:10.1901/jaba.1979.12-273

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience, 9*(10), 1218–1220. doi:10.1038/nn1770

Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.-S., Cohen, A., ... Chavarria, S. (2005). Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. *Speech Communication, 46*(3–4), 418–439. doi:10.1016/j.specom.2005.01.009

Howlin, P., & Rutter, M. (1989). Mothers' speech to autistic children: A preliminary causal analysis. *Journal of Child Psychology and Psychiatry, 30*(6), 819–843.

IguanaWorks. (2008, June). Iguanaworks. Retrieved June 1, 2008, from http://iguanaworks.net/

Innvo Labs. (2012). PLEOworld. Retrieved February 22, 2012, from http://www.pleoworld.com/pleo_rb/eng/index.php

Interagency Autism Coordinating Committee. (2011). 2011 IACC strategic plan for autism spectrum disorder research.

Isbell, C., Shelton, C. R., Kearns, M., Singh, S., & Stone, P. (2001). A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents* (pp. 377–384). New York, NY, USA: ACM. doi:10.1145/375735.376334

Joseph, R. M., & Tager-Flusberg, H. (1997). An investigation of attention and affect in children with autism and Down Syndrome. *Journal of Autism and Developmental Disorders, 27*(4), 385–396. doi:10.1023/A:1025853321118

Kahney, L. (1999, July 6). Furby: It's Not Just a Toy. *WIRED*. Retrieved from http://www.wired.com/science/discoveries/news/1999/07/20572

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: a field trial. *Human-Computer Interaction, 19*(1), 61–84. doi:10.1207/s15327051hci1901&2_4

Kanner, L. (1943). Autistic disturbances of affective content. *Nervous Child, 2*, 217–250.

Kaplan, F., Oudeyer, P.-Y., Kubinyi, E., & Miklósi, A. (2002). Robotic clicker training. *Robotics and Autonomous Systems, 38*(3–4), 197–206. doi:10.1016/S0921-8890(02)00168-9

Karmiloff-Smith, A. (2006). Atypical epigenesis. *Developmental Science, 10*(1), 84–88. doi:10.1111/j.1467-7687.2007.00568.x

Kasari, C., Sigman, M., Mundy, P., & Yirmiya, N. (1990). Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children. *Journal of Autism and Developmental Disorders, 20*(1), 87–100. doi:10.1007/BF02206859

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kim, B., & Lee, G. G. (2006). C-TOBI-Based Pitch Accent Prediction Using Maximum-Entropy Model. In M. Gavrilova, O. Gervasi, V. Kumar, C. J. K. Tan, D. Taniar, A. Laganá, ... H. Choo (Eds.), *Computational Science and Its Applications - ICCSA 2006* (pp. 21–30). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11751595_3

Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., & Scassellati, B. (2013). Social Robots as Embedded Reinforcers of Social Behavior in Children with Autism. *Journal of Autism and Developmental Disorders, 43*(5), 1038–1049. doi:10.1007/s10803-012-1645-2

Kim, E. S., Gold, K., & Scassellati, B. (2008). What prosody tells infants to believe. In *7th IEEE International Conference on Development and Learning, 2008. ICDL 2008* (pp. 274 –279). Presented at the 7th IEEE International Conference on Development and Learning, 2008. ICDL 2008. doi:10.1109/DEVLRN.2008.4640842

Kim, E. S., Leyzberg, D., Tsui, K. M., & Scassellati, B. (2009). How people talk when teaching a robot. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 23–30). New York, NY, USA: ACM. doi:10.1145/1514095.1514102

Kim, E. S., Paul, R., Shic, F., & Scassellati, B. (2012). Bridging the Research Gap: Making HRI Useful to Individuals with Autism. *Journal of Human-Robot Interaction*, *1*(1). doi:10.5898/jhri.v1i1.25

Kim, E. S., & Scassellati, B. (2007). Learning to refine behavior using prosodic feedback. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on* (pp. 205 – 210). doi:10.1109/DEVLRN.2007.4354072

Klin, A., Lang, J., Cicchetti, D. V., & Volkmar, F. R. (2000). Brief Report: Interrater Reliability of Clinical Diagnosis and DSM-IV Criteria for Autistic Disorder: Results of the DSM-IV Autism Field Trial. *Journal of Autism and Developmental Disorders*, *30*(2), 163–167. doi:10.1023/A:1005415823867

Koegel, L. K., Koegel, R. L., Harrower, J. K., & Carter, C. M. (1999). Pivotal Response Intervention I: Overview of Approach. *Research and Practice for Persons with Severe Disabilities*, *24*(3), 174–185. doi:10.2511/rpsd.24.3.174

Koegel, R. L., Dyer, K., & Bell, L. K. (1987). The influence of child-preferred activities on autistic children's social behavior. *Journal of Applied Behavior Analysis*, *20*(3), 243–252. doi:10.1901/jaba.1987.20-243

Koegel, R. L., Koegel, L. K., & McNerney, E. K. (2001). Pivotal Areas in Intervention for Autism. *Journal of Clinical Child & Adolescent Psychology*, *30*(1), 19–32. doi:10.1207/S15374424JCCP3001_4

Koegel, R. L., O'Dell, M. C., & Koegel, L. K. (1987). A natural language teaching paradigm for nonverbal autistic children. *Journal of Autism and Developmental Disorders*, *17*(2), 187–200. doi:10.1007/BF01495055

Koegel, R. L., Vernon, T. W., & Koegel, L. K. (2009). Improving Social Initiations in Young Children with Autism Using Reinforcers with Embedded Social Interactions. *Journal of Autism and Developmental Disorders, 39*(9), 1240–1251. doi:10.1007/s10803-009-0732-5

Kozima, H., Michalowski, M. P., & Nakagawa, C. (2009). Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics, 1*(1), 3–18. doi:10.1007/s12369-008-0009-8

Kozima, H., Nakagawa, C., & Yasuda, Y. (2005). Interactive robots for communication-care: a case-study in autism therapy. In *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005* (pp. 341– 346). Presented at the IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005, IEEE. doi:10.1109/ROMAN.2005.1513802

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*(4), 391–405. doi:10.1016/S0167-6393(00)00058-3

Levow, G.-A. (2006). Unsupervised and semi-supervised learning of tone and pitch accent. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 224–231). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1220835.1220864

Liscombe, J., Venditti, J., & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of EUROSPEECH* (pp. 725–728). Retrieved from http://www1.cs.columbia.edu/~jaxin/www/papers/liscombe_etal_2003.pdf

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000a). The Autism Diagnostic Observation Schedule—Generic: A Standard

Measure of Social and Communication Deficits Associated with the Spectrum of Autism, 205–223.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... Rutter, M. (2000b). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223. doi:10.1023/A:1005592401947

Lord, C., Rutter, M., & Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659–685. doi:10.1007/BF02172145

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Transcription, format and programs* (3rd ed., Vol. 1). Mahwah, NJ: Lawrence Erlbaum. Retrieved from http://books.google.com/books/about/The_CHILDES_Project_Transcription_format.html?id=3skQRRMRvnAC

Maione, L., & Mirenda, P. (2006). Effects of Video Modeling and Video Feedback on Peer-Directed Social Language Skills of a Child With Autism. *Journal of Positive Behavior Interventions, 8*(2), 106–118. doi:10.1177/10983007060080020201

McCaleb, P., & Prizant, B. M. (1985). Encoding of new versus old information by autistic children. *Journal of Speech and Hearing Disorders, 50*(3), 230.

Menn, L., & Boyce, S. (1982). Fundamental Frequency and Discourse Structure. *Language and Speech, 25*(4), 341–383. doi:10.1177/002383098202500403

Mesibov, G. B. (1992). Treatment issues with high-functioning adolescents and adults with autism. In E. Schopler & G. B. Mesibov (Eds.), *High-functioning individuals with autism* (pp. 143–156). New York: Springer. Retrieved from http://books.google.com/books?hl=en&lr=&id=wugfR4WnBQ8C&oi=fnd&pg=PA143&dq=mesibov+1992+High-functioning+individuals+with+autism+schopler+Treatment+Issues+with+High-Functioning+Adolescents+and+Adults+with+Autism&ots=S1NKFAUZRL&sig=4MIsZM3WCgm9gWF8GTLtsjqyZLQ#v=onepage&q=mesibov%201992%20High-functioning%20individuals%20with%20autism%20schopler%20Treatment%20Issues%20with%20High-Functioning%20Adolescents%20and%20Adults%20with%20Autism&f=false

Mixdorf, H. (2002). Speech technology, tobi and making sense of prosody. Aix-en-Provence. Retrieved from http://sprosig.isle.illinois.edu/sp2002/papers.htm

Morgan, J. L. (1986). *From Simple Input to Complex Grammar.* Cambridge, MA: The MIT Press.

Mullen, E. M. (1995). *Mullen Scales of Early Learning* (AGS.). San Antonio, TX: Pearson.

Mundy, P., Sigman, M. D., & Dawson, G. (1989). Specifying the nature of the social impairment in autism. In *Autism: New perspectives on nature, diagnosis, and treatment* (pp. 3–21). Retrieved from http://books.google.com/books?id=eEWjVdwA8tEC&pg=PR15&lpg=PR15&dq=Autism:+New+perspectives+on+nature,+diagnosis,+and+treatment&source=bl&ots=IA1WNv0d28&sig=ZCrUumyoyew_LyQ3h0mwsRjNej0&hl=en&sa=X&ei=7vj9T6_eK8Gw6wG-

1vDQBg&ved=0CEYQ6AEwAA#v=onepage&q=Autism%3A%20New%20perspectives%20on%20nature%2C%20diagnosis%2C%20and%20treatment&f=false

Mundy, P., Sigman, M., Ungerer, J., & Sherman, T. (1986). Defining the Social Deficits of Autism: The Contribution of Non-verbal Communication Measures. *Journal of Child Psychology and Psychiatry*, *27*(5), 657–669. doi:10.1111/j.1469-7610.1986.tb00190.x

Nadig, A., & Shaw, H. (2012). Acoustic and Perceptual Measurement of Expressive Prosody in High-Functioning Autism: Increased Pitch Range and What it Means to Listeners. *Journal of Autism and Developmental Disorders*, *42*(4), 499–511. doi:10.1007/s10803-011-1264-3

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). New York, NY, USA: ACM. doi:10.1145/191666.191703

Noll, A. M. (1967). Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, *41*(2), 293–309. doi:10.1121/1.1910339

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1967). *The Measurement of Meaning*. University of Illinois Press. Retrieved from http://www.press.uillinois.edu/books/catalog/32mtm7sx9780252745393.html

Papousek, M., Papousek, H., & Bornstein, M. H. (1985). The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech. In *Social perception in infants* (pp. 269–297).

Patterson, D., Snyder, L., & Ullman, J. (1999). Best practices memo: Evaluating computer scientists and engineers for promotion and tenure. *Computing Research News*.

Paul, R. (2005). Assessing communication in autism spectrum disorders. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., Vols. 1-2, Vol. 2, pp. 799–816). Hoboken, NJ: John Wiley and Sons.

Paul, R. (2008). Interventions to improve communication in autism. *Child and Adolescent Psychiatric Clinics of North America, 17*(4), 835–856.

Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 35*(2), 205–220. doi:10.1007/s10803-004-1999-1

Paul, R., Orlovski, S., Marcinko, H., & Volkmar, F. (2009). Conversational behaviors in youth with high-functioning ASD and Asperger syndrome. *Journal of Autism and Developmental Disorders, 39*(1), 115–125. doi:10.1007/s10803-008-0607-1

Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., & Volkmar, F. (2005). Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 35*(6), 861–869. doi:10.1007/s10803-005-0031-8

Pierrehumbert, J., & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.), *Intentions in Communications* (pp. 271–313). MIT Press.

Prelock, P. A., Paul, R., & Allen, E. M. (2011). Evidence-based treatments in communication for children with autism spectrum disorders. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 93–169). New York, NY: Springer US. Retrieved from http://www.springerlink.com/content/g520308543808731/abstract/

Quatieri, T. F. (2002). *Discrete-time speech signal processing: principles and practice.* Upper Saddle River, NJ: Prentice Hall.

Ratner, N. B. (1986). Durational cues which mark clause boundaries in mother-child speech. In *Journal of Phonetics.* Presented at the 1985 American Speech-Language Hearing Association Annual Convention (1985, Washington, DC)., Elsevier Science. Retrieved from http://psycnet.apa.org/index.cfm?fa=search.displayrecord&uid=1988-01123-001

Ratner, N. B. (1987). The phonology of parent–child speech. In K. E. Nelson & A. E. van Kleeck (Eds.), *Children's language* (Vol. 6, pp. 159–174). Retrieved from http://www.psypress.com/books/details/9780898597608/

Reichow, B., Volkmar, F., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38*(7), 1311–1319. doi:10.1007/s10803-007-0517-7

Reichow, B., & Volkmar, F. R. (2011). Evidence-based practices in autism: Where we started. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 3–24). New York, NY: Springer US. Retrieved from http://www.springerlink.com/content/w34707658g646lg5/abstract/

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction, 1*(1).

Robins, B., Dautenhahn, K., & Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies, 7*(3), 509–542.

Robins, B., Dautenhahn, K., te Boekhorst, R., & Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help

encourage social interaction skills? *Universal Access in the Information Society*, *4*(2), 105–120. doi:10.1007/s10209-005-0116-3

Robinson-Mosher, A. L., & Scassellati, B. (2004). Prosody recognition in male infant-directed speech. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings* (Vol. 3, pp. 2209–2214 vol.3). Presented at the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. doi:10.1109/IROS.2004.1389737

Roy, D., & Pentland, A. (1996). Automatic spoken affect classification and analysis. In , *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996* (pp. 363–367). Presented at the , Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996. doi:10.1109/AFGR.1996.557292

Russell, S., & Norvig, P. (2003). *Artificial Intelligence - A Modern Approach* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc. Retrieved from http://catalogue.pearsoned.co.uk/preface/080536031X.pdf

Rutter, M., Bailey, A., & Lord, C. (2003). *Social Communication Questionnaire: SCQ (W-381)*. Los Angeles, CA: Western Psychological Services. Retrieved from http://portal.wpspublish.com/portal/page?_pageid=53,70432&_dad=portal&_schema=PORTAL

Sasson, N. J., Elison, J. T., Turner-Brown, L. M., Dichter, G. S., & Bodfish, J. W. (2011). Brief Report: Circumscribed Attention in Young Children with Autism. *Journal of Autism and Developmental Disorders*, *41*(2), 242–247. doi:10.1007/s10803-010-1038-3

Scassellati, B. (1996). Mechanisms of shared attention for a humanoid robot.

Scassellati, B. (2005). Quantitative metrics of social response for autism diagnosis. In *IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005* (pp. 585 – 590). doi:10.1109/ROMAN.2005.1513843

Scassellati, B., Admoni, H., & Matarić, M. J. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering, 14*.

Schopler, E., Reichler, R. J., & Renner, B. R. (1986). The Childhood Autism Rating Scale (CARS): For diagnostic screening and classification of autism. New York, NY: Irvington.

Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., & Volkmar, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome, 1097.

Sigman, M., & Mundy, P. (1989). Social attachments in autistic children. *Journal of the American Academy of Child & Adolescent Psychiatry, 28*(1), 74–81.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., … Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the 1992 international conference on spoken language processing* (Vol. 2, pp. 867–870).

Slaney, M., & McRoberts, G. (2003). BabyEars: A recognition system for affective vocalizations. *Speech Communication, 39*(3–4), 367–384. doi:10.1016/S0167-6393(02)00049-3

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language, 4*(01), 1–22. doi:10.1017/S0305000900000453

South, M., Ozonoff, S., & McMahon, W. M. (2005). Repetitive Behavior Profiles in Asperger Syndrome and High-Functioning Autism. *Journal of Autism and Developmental Disorders, 35*(2), 145–158. doi:10.1007/s10803-004-1992-8

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales (Vineland-II)* (2nd ed.). San Antonio, TX: Pearson. Retrieved from http://psychcorp.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=Vineland-II

Stanton, C. M., Kahn Jr., P. H., Severson, R. L., Ruckert, J. H., & Gill, B. T. (2008). Robotic animals might aid in the social development of children with autism (p. 271). ACM Press. doi:10.1145/1349822.1349858

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 101–108). San Diego, CA: ACM. doi:10.1145/1514095.1514115

Stern, A., Frank, A., & Resner, B. (1998). Virtual petz (video session): a hybrid approach to creating autonomous, lifelike dogz and catz. In *Proceedings of the second international conference on Autonomous agents* (pp. 334–335). New York, NY, USA: ACM. doi:10.1145/280765.280852

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: infant age and context related changes. *Journal of Child Language*, *10*(01), 1–15. doi:10.1017/S0305000900005092

Sun, G., & Scassellati, B. (2005). A fast and efficient model for learning to reach. *International Journal of Humanoid Robotics*, *02*(04), 391–413. doi:10.1142/S0219843605000569

Sun, L. X., & Sun, X. (2002). Pitch Accent Prediction Using Ensemble Machine. In *in Proceedings of ICSLP-2002* (pp. 16–20).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (1st ed.). Cambridge, MA, USA: MIT Press. Retrieved from http://mitpress.mit.edu/books/reinforcement-learning

Syrdal, A., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 3, pp. 235–238). Beijing China.

Tager-Flusberg, H., & Caronna, E. (2007). Language Disorders: Autism and Other Pervasive Developmental Disorders. *Pediatric Clinics of North America, 54*(3), 469–481. doi:10.1016/j.pcl.2007.02.011

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and Communication in Autism. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., Vols. 1-2, Vol. 1, pp. 335 – 364). Hoboken, NJ: John Wiley and Sons.

Tapus, A., Matarić, M. J., & Scassellati, B. (2007). Socially assistive robotics [Grand challenges of robotics]. *Robotics Automation Magazine, IEEE, 14*(1), 35 –42. doi:10.1109/MRA.2007.339605

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*(1), 53–71.

Thomaz, A. L., & Breazeal, C. (2006a). Transparency and socially guided machine learning. In *Fifth International Conference on Development and Learning (ICDL)*. Bloomington, IN.

Thomaz, A. L., & Breazeal, C. (2006b). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In

*Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, p. 1000). Retrieved from http://www.aaai.org/Papers/AAAI/2006/AAAI06-157.pdf

Thomaz, A. L., Hoffman, G., & Breazeal, C. (2006). Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006* (pp. 352–357). Presented at the The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006. doi:10.1109/ROMAN.2006.314459

Turner-Brown, L. M., Lam, K. S. L., Holtzclaw, T. N., Dichter, G. S., & Bodfish, J. W. (2011). Phenomenology and measurement of circumscribed interests in autism spectrum disorders. *Autism*, *15*(4), 437–456. doi:10.1177/1362361310386507

UGOBE Life Forms. (2008). PLEOworld. Retrieved June 1, 2008, from http://www.pleoworld.com/pleo_rb/eng/index.php

Van Bourgondien, M. E., & Woods, A. V. (1992). Vocational possibilities for high-functioning adults with autism. *High functioning individuals with autism*, 227–242.

Volkmar, F. R. (1998). Categorical Approaches to the Diagnosis of Autism An Overview of DSM-IV and ICD-10. *Autism*, *2*(1), 45–59. doi:10.1177/1362361398021005

Volkmar, F. R., & Klin, A. (2005). Issues in the classification of autism and related conditions. In F. R. Volkmar, R. Paul, A. Klin, & D. J. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd ed., Vols. 1-2, Vol. 1, pp. 335–364). Hoboken, NJ: John Wiley and Sons.

Volkmar, F. R., Lord, C., Bailey, A., Schultz, R. T., & Klin, A. (2004). Autism and pervasive developmental disorders. *Journal of Child Psychology and Psychiatry, 45*(1), 135–170. doi:10.1046/j.0021-9630.2003.00317.x

Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes. Harvard University Press.

Walker, D. R., Thompson, A., Zwaigenbaum, L., Goldberg, J., Bryson, S. E., Mahoney, W. J., … Szatmari, P. (2004). Specifying PDD-NOS: A Comparison of PDD-NOS, Asperger Syndrome, and Autism. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(2), 172–180. doi:10.1097/00004583-200402000-00012

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI).* San Antonio, TX: Pearson Assessment. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8981-502

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (WISC-IV)* (4th ed.). San Antonio, TX: Pearson Assessment. Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8979-044

Werry, I., & Dautenhahn, K. (1999). Applying mobile robot technology to the rehabilitation of autistic children. In *Proceedings of the 7th Symposium on Intelligent Robotic Systems (SIRS99).* Retrieved from https://uhra.herts.ac.uk/dspace/handle/2299/1946

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Test (GORT-4)* (4th ed.). San Antonio, TX: Pearson Assessment. Retrieved from

http://psychcorp.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8116-577&Mode=summary

Wood, D., & Middleton, D. (1975). A Study of Assisted Problem-Solving. *British Journal of Psychology*, *66*(2), 181–191. doi:10.1111/j.2044-8295.1975.tb01454.x

Yirmiya, N., Kasari, C., Sigman, M., & Mundy, P. (1989). Facial expressions of affect in autistic, mentally retarded and normal children. *Journal of Child Psychology and Psychiatry*, *30*(5), 725–735. doi:10.1111/j.1469-7610.1989.tb00785.x