Controlling Social Dynamics with a Parametrized Model of Floor Regulation

Crystal Chao, Andrea L. Thomaz
Georgia Institute of Technology

Turn-taking is ubiquitous in human communication, yet turn-taking between humans and robots continues to be stilted and awkward for human users. The goal of our work is to build autonomous robot controllers for successfully engaging in human-like turn-taking interactions. Towards this end, we present CADENCE, a novel computational model and architecture that explicitly reasons about the four components of floor regulation: *seizing* the floor, *yielding* the floor, *holding* the floor, and *auditing* the owner of the floor. The model is parametrized to enable the robot to achieve a range of social dynamics for the human-robot dyad. In a between-groups experiment with 30 participants, our humanoid robot uses this turn-taking system at two contrasting parametrizations to engage users in autonomous object play interactions. Our results from the study show that: (1) manipulating these turn-taking parameters results in significantly different robot behavior; (2) people perceive the robot's behavioral differences and consequently attribute different personalities to the robot; and (3) changing the robot's personality results in different behavior from the human, manipulating the social dynamics of the dyad. We discuss the implications of this work for various contextual applications as well as the key limitations of the system to be addressed in future work.

Keywords: Turn-taking, engagement, situated dialogue, floor exchange, backchannel, multimodal systems, timed Petri net, architecture, human-robot interaction

1. Introduction

A substantial percentage of action taken by humans in their daily lives is social action—informing, promising, convincing, asking, teaching, and learning from others. Natural language and dialogue are constructs that exist to facilitate social action, and their sophistication in humans reflects a unique reliance upon social action to accomplish their goals (Clark, 1996). Turn-taking through situated dialogue is one of the most common protocols by which humans communicate and exchange information, which occurs naturally in humans from a very young age (Trevarthen, 1979; Tronick, Als, & Adamson, 1979). Yet turn-taking between humans and robots remains an awkward and confusing experience for human users.

A reason for this continued awkwardness is that turn-taking is often relegated to the status of emergent behavior in human-robot interaction (HRI) systems, rather than treated as an interaction process to be explicitly controlled. If a robot does not have the capacity to adapt to the human's

Journal of Human-Robot Interaction, Vol. 1, No. 1, 2012, Pages 78-95. DOI 10.5898/JHRI.1.1.Tanaka

style or take the initiative to repair breakdowns, then the onus is on the human to adapt to the robot's incidental turn-taking dynamics that occur as a side effect of the robot's other behaviors. The longer-term intent of this work is thus the continued development of explicit control for turn-taking that plays an integral part in the robot's social decision-making process.

Turn-taking typically refers to spoken dialogue and the fluent exchange of the speaking floor. We believe that the embodied nature of turn-taking in HRI necessitates a broader perspective, in which turn-taking is a phenomenon that arises in the presence of any bottlenecking resources that are exchanged by interaction participants. These include the speaking floor in addition to shared resources in other embodied modalities, such as control over shared physical space or objects. The robot needs to be aware of and manage these shared resources—exhibiting *reciprocal* behavior. Turn-taking in embodied interaction is also a *multimodal* process, both due to the diversity of bottleneck types in embodied interaction as well as the gaze, verbal, and gestural cues used to signal turn-taking intent. The lofty goal is for an autonomous robot cooperating with a human to achieve the same fluent and seamless turn-taking of shared resources that occurs between two humans.

Our research focuses on building autonomous controllers for human-robot turn-taking interactions. To limit the scope of the problem, we restrict our work to dyadic interactions between one human and one robot. Each interaction participant is described as being in one of four states: *seizing* the floor, *holding* the floor, *yielding* the floor to the partner, *auditing* the partner's turn. Our prior work focused substantially on yielding the floor. We previously discovered that robots always fully completing their temporally extended actions forced humans to yield to the robot (Chao, Lee, Begum, & Thomaz, 2011; Thomaz & Chao, 2011). This led us to implement and evaluate a turn-taking controller than could yield the floor in a collaborative manipulation domain, leading to a reversal of dynamics—humans exerted increased control over task balance, acting as the ultimate decision-makers as to whether to take more initiative or leave more of the work to the robot (Chao & Thomaz, 2012).

Our prior work argues strongly for robots able to *yield* control to humans, but still a central problem remains: the timing of when to *seize* the floor. A robot that only yields does not take initiative to recover from lapses, introduce new information or action, or provide backchannel feedback. The challenge, then, is how to reintroduce initiative into a robot's turn-taking without inappropriately detracting from the human's control of the interaction. From a usability perspective, the system should be held responsible for taking initiative to structure the interaction in order to recover from moments of ambiguity or confusion, as well as to make the interaction state self-evident.

In this article, we present CADENCE, our Control Architecture for the Dynamics of Embodied Natural Coordination and Engagement. CADENCE, as described in Section 3, is a new computational model for floor regulation that explicitly represents and reasons about all four components of the floor regulation problem. Prior work in human social psychology has shown how dominant or deferent conversational styles correlate with influence over a task (Mazur & Cataldo, 1989). We thus posit that one's implementation of these turn-taking behaviors for a robot significantly affects the overall social dynamics of the dyad. Turn-taking is characterized by a fundamental tension between who is initiating action or communication (seizing and holding), and who is being supportive and contingent upon the other (yielding and auditing). The decision of when to seize and how often to seize should differ depending on the relative status of the interaction partners. Nearly all social transactions between humans are defined by relative status, and communicating the appropriate status can be essential to the success of the interaction (Johnstone, 1987).

CADENCE features a parametrization of turn-taking that enables the robot to control its behavior for achieving a desired social dynamic. In Sections 4 and 5, we present an experiment within a highly open-ended domain of a robot and a human playing together with toys on a tabletop. Our results in Section 6 highlight some of the effects that exhibiting different personality or status behavior can have on humans. The experiment is designed to explore the range of interaction dynamics made possible by CADENCE, and thus features two somewhat extreme parametrizations rather than attempting to test any hypothetical "optimal" setting. Realistically, the most suitable parametrization will vary significantly for different social contexts (a discussion point we raise in Section 7.3). We emphasize that the experiment and results do not aim to prescribe or advocate any specific social dynamic for HRI, but simply to characterize interaction differences between contrasting turn-taking styles made possible by the system.

We start by presenting related work on floor exchange systems in Section 2. Next, we describe our implementation of parametrized floor regulation as a generic system in Section 3. This is followed by a description in Section 4 of how we instantiate this model for turn-taking interactions in the particular task domain of playing with objects on a tabletop. We use this domain to evaluate the system in an experiment with 30 human participants. Our results show that: (1) manipulating our floor regulation parameters results in different robot behavior; (2) people perceive this difference in behavior and attribute different personalities to the robot; and (3) changing the robot's personality results in different behavior from the human, manipulating social dynamics of the dyad. Section 7 then identifies system shortcomings and applications to be addressed in future work.

2. Related Systems

Turn-taking is complex to control for an artificial agent due to a number of challenges. Turn-taking interactions feature sensitive low-latency timing requirements. The process is highly multimodal and relies on the precise control of gaze, gesture, and speech cues. In addition, it requires multimodal perception of the human's cues that can be noisy or unreliable, or simply does not yet exist.

One of the earlier architectures for controlling multimodal behavior is BEAT (Cassell et al., 1999), which was used to control virtual conversational characters that were extended by real-world perception of the human. Such virtual characters, also referred to as embodied conversational agents (ECAs), can now perform more complex reasoning in turn-taking for situated multiparty conversation by using a cost structure to make turn decisions (Bohus & Horvitz, 2011). The notion of cost has also been used to control turn-taking for a spoken dialogue system based on a finite state machine (Raux & Eskenazi, 2012).

Research on turn-taking in social robots is motivated by similar goals to the research on ECAs or speech systems, but some of the challenges differ. The robot's physical embodiment leads to differences in behavioral timing and social impact on the user, and thus may require action control architectures of a different nature. The work of (Rich, Ponsler, Holroyd, & Sidner, 2010) and (Holroyd, Rich, Sidner, & Ponsler, 2011) has addressed some of these challenges by identifying and generating "connection events" in order for a robot to maintain engagement with a human interaction partner. Other systems have also been developed to control multimodal dialogue for social robots, such as the work of (Kanda, Ishiguro, Imai, & Ono, 2004) that controls dynamic switching of behaviors in the speech and gesture modalities, and the framework of (Nakano et al., 2011) that controls task-based dialogue using parallelized processes with interruption handling.

In a vast number of social robot systems, turn-taking is an emergent behavior that arises from other interacting processes as opposed to being explicitly controlled. A seminal example of such a robot is Kismet (Breazeal, 2003). Kismet's control architecture did not represent or control the conversational floor, and the combination of an emotion regulation system with reactive behavior produced some engaging and expressive interactions with people. However, human dialogue is vastly more complex than simply alternating reactive responses and features dynamic adjustments to simultaneous starts, overlaps, and silences (E. Schegloff, 2000). In modeling such complexities, we aim to progress the field beyond emergent turn-taking.

Other prior research lends much focused insight into subcomponents of floor regulation problem by studying individual modalities in a controlled fashion; for example, the work of (Mutlu, Shiwa, Ishiguro, & Hagita, 2009) analyzes the function of gaze behaviors in designating speakers and auditors. Drawing from prior work in the field, our goal is to combine these components into an integrated system, with the hope of iteratively increasing the communicative capabilities and overall social competence of a robot. In performing this integration, we also enable the study of interesting dynamics between subparts of the system, such as the effects of embodied robot action taken across different modalities.

3. CADENCE System Description

3.1 Infrastructure

CADENCE is designed to support interactions with our upper-torso humanoid robot Simon, shown in Figure 6. Simon has two 7-DOF arms with 4-DOF hands, which are used for gesturing in addition to picking, placing, and pointing at objects. The arms and hands have compliant series-elastic actuators for safe operation alongside a human. Simon also has a socially expressive head and neck for gaze and head gestures. The head includes two 2-DOF ears that also contain LED arrays. Two speakers near Simon's base are used for communicating through text-to-speech.

The robot hardware is controlled through the C6 (Creatures 6) software framework, a cognitive architecture for building intelligent interactive characters (an earlier version, C4, is described in (Blumberg et al., 2002)). Within C6, sensor data is received over the network in the sensory system. These are organized into a percept tree in the perception system, where percepts describe semantic features of the sensor data (e.g. object location, voice activity presence). Percepts are clustered and filtered over time to form persistent beliefs in the belief system. The action system then uses these beliefs for decision-making and action selection. Performing actions causes changes to the robot's joint positions in the motor system, which are then sent to the robot hardware over the network through the motor renderer. The full C6 sensory-motor loop runs at approximately 60 Hz.

The C6 action system layer is changed to suit the needs of a particular interaction. The work in this article specifically describes one action architecture implementation based on a timed Petri net (TPN), an extension of Petri nets with additional modeling of timing. This decision is based on the modeling power of TPNs for combinations of temporally extended actions. Embodied cooperative acts require synchronization over bottlenecks, cross-modality and cross-participant concurrency, condition sequencing, and timing management. TPNs are uniquely able to model these requirements in a way that overcomes limitations of scalability and generalizability faced by state-based representations such as Markov models and finite state machines. Where state-based representations must take crossproducts of conditions to achieve synchronization and concurrency simultaneously, Petri nets offer an elegant formalism for an interaction designer to model interacting variables in behavioral processes.

A basic Petri net, or Place/Transition (P/T) net, is a bipartite multigraph comprising two finite disjoint sets of nodes, places and transitions. A multiset of directed arcs connects the node types in an alternating fashion. Places are used to represent robot state and can contain a natural number of tokens; control is transferred through token movement throughout the graph. The mapping of places to tokens is called a marking $M : P \rightarrow \mathbb{N}$, and is the Petri net's implicit state representation. More formally, a Petri net is a 5-tuple $N = (P, T, I, O, M_0)$, where:

- P is a finite set of places,
- T is a finite set of transitions, where $P \cup T \neq \emptyset$ and $P \cap T = \emptyset$,
- $I : P \times T$ is the input function directing incoming arcs to transitions from places,

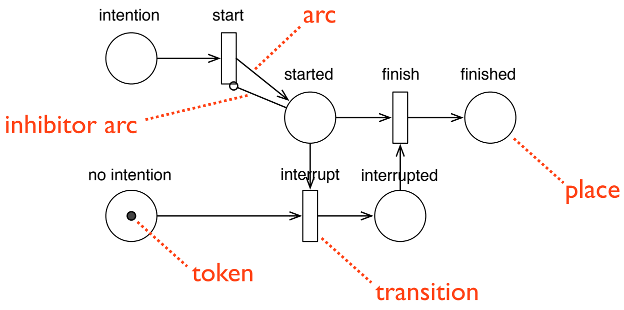

Figure 1. An interruptible action template from our previous work, with Petri net primitives labeled.

- $O : T \times P$ is the output function directing outgoing arcs from transitions to places, and
- M_0 is an initial marking.

A transition $t = \{\mathcal{G}(I), \mathcal{F}(M, I, O)\}$ is controlled by two functions: a guard function $\mathcal{G}(I) \rightarrow \{0, 1\}$ and a firing function $\mathcal{F}(M, I, O) \rightarrow M'$. The guard function enables the transition (allows it to fire) as a function of the inputs. The firing function runs until the guard function is no longer satisfied, which requires the transition t to alter the graph marking in a way that changes the guard function inputs. This results in the transition disabling. A transition can induce such marking changes by transferring ownership of tokens between places, by destroying unneeded tokens in input places, or by spawning new tokens. In our system, places and tokens are also typed; this variant is known as colored Petri nets. The control logic for guard functions depends only on the presence and absence of tokens in places, but the firing logic for a transition can unpack the typed data contained within tokens to perform an operation.

In the timed Petri net extension, the guard and firing functions are associated with timers for the enabling delay and the firing delay. Restrictions to the time distributions result in different classes of Petri nets, such as stochastic Petri nets (exponential) and time Petri nets (deterministic intervals) (Wang, 1998). A system clock $C(i, \tau) \rightarrow \tau'$ controls how time τ updates to new time τ' at each cycle i. In our TPN, we additionally track intervals of time $[\tau_\alpha, \tau_\beta)$ for which places contain tokens and which transitions are enabled, allowing operations on the historical activation times of these nodes. For the purposes of controlling robot behavior, we sometimes design transitions that are intended to fire continually for an extended duration (i.e. while engaged with a human or throughout an entire experiment); this fits within the semantics of TPNs but contrasts somewhat with traditional Petri net modeling, in which incoming tokens to a transition are intended to be consumed immediately.

Petri nets have a specific visualization scheme in which places are drawn as circles, transitions as rectangles, directed arcs as arrows, and tokens as small filled circles inside of places. Inhibitor arcs are drawn with a circular endpoint. We have labeled these graph primitives for the reader in Figure 1, which depicts an interruptible action template from our previous work (Chao & Thomaz, 2012). We refer the reader to that same article for more details and arguments for using this representation to control robot behavior. More details on standard Petri nets and their applications can also be found in a survey by (Murata, 1989).

3.2 Turn-taking process

Turn-taking is the dynamic process by which interaction participants exchange shared resources. The conversational floor, the focal point of turn-taking in linguistics, is a shared resource due to the cognitive difficulties of processing simultaneous speech (Baddeley & Della Sala, 1996). In addition, the floor owner has the opportunity to exhibit greater control over additional shared resources such as shared physical space and objects through accompanying gesticulation and manipulation actions. Floor exchange represents shifting control over the outcome of an interaction.

While turns themselves may be highly task-oriented, they are also accompanied by signals across multiple modalities that communicate one's desire to own the floor. People may exploit beat gestures or gaze aversion to suppress an interaction partner's attempts to seize the floor, or gaze back at a partner and alter one's prosody to signal an imminent intent to yield (Duncan, 1974; Orestrom, 1983). In this section, we describe the components of the turn-taking Petri net that are used to monitor the floor and regulate its ownership. Because the Petri net behavior model is a flat model comprising many connected subgraphs, we have annotated common nodes between Figures 2, 3, and 4 with a consistent shading scheme in order to highlight the connections between the subgraphs.

3.2.1 Floor state representation The cornerstones of turn-taking behavior are seizing the floor, holding the floor, yielding the floor, and auditing the current owner of the floor. CADENCE models seizing and yielding as transitions (t_{seize} and t_{yield}) that lead to the states of holding ($p_{holding}$) and auditing (represented by $p_{yielded}$, which activates t_{audit}). This execution flow is discussed more in Section 3.2.3. The balance between the time one spends exhibiting holding versus auditing behavior is critical to the social dynamics of a turn-taking interaction. The combinations of the user's and robot's attempts at holding and auditing additionally result in the meta-states of *conflict* (both taking a turn), a *lapse* (neither taking a turn), or one or the other owning the floor.

Figure 2 depicts the relationship between engagement, individual turn states, and dyadic floor states in the system. Each boxed area indicates a set of places between which a single token is shared; hence, these places are mutually exclusive. The dyadic floor state is only updated if the robot is currently engaged with an interaction partner. This floor state is determined as a function of the time that the robot and the user have spent in their respective current turn states. For example, simultaneous holding that exceeds a duration referred to as *conflict time* results in a dyadic state of conflict, and similarly for *lapse time* and lapses. The dyadic floor states then drive other decisions made in the robot's turn-taking, such as whether to interrupt itself or take more initiative. Parameters related to conflicts and lapses are summarized in Section 3.2.6.

3.2.2 User modeling The perceptual signals used to monitor the user's behavior include speech presence, gesturing, and whether the user is gazing at or away from the robot. The particular implementation of these signals can be expected to vary across domains. Details for one relatively domain-invariant implementation of these signals are given in Section 4.1.2. These low-level features are fundamental to attaining and communicating attention in the visual and auditory channels.

Figure 3 shows the process that models the user's turn-taking state. While the user is engaged, the firing function of t_{signal} accesses the belief system to classify perception of the user's speech and motion into the three states of $p_{suppressing}$, p_{idle}, and $p_{signaling}$. These three places share a mutually exclusive token assigned by t_{signal} at each clock cycle that indicates whether the user's actions appear to be suppressing robot turn attempts ($p_{suppressing}$), the user is not performing actions (p_{idle}), or that user action has been perceived at a lower strength or consistency than would be interpreted as suppression ($p_{signaling}$). For efficiency in this pattern of TPN substructure, a single token can be moved between the three mutually exclusive places (rather than spawning and destroying tokens at each state change). The place $p_{signaling}$ owns the token as a precursor to $p_{suppressing}$ owning it; this is used in the robot's control process to determine whether the robot should hesitate while taking a turn (see Section 3.2.4).

The places $p_{holding}$ and $p_{yielded}$ in the user model also share a token, a structure that is mirrored in the robot's control process (see Figures 2 and 4). While the place $p_{holding}$ has a token, the

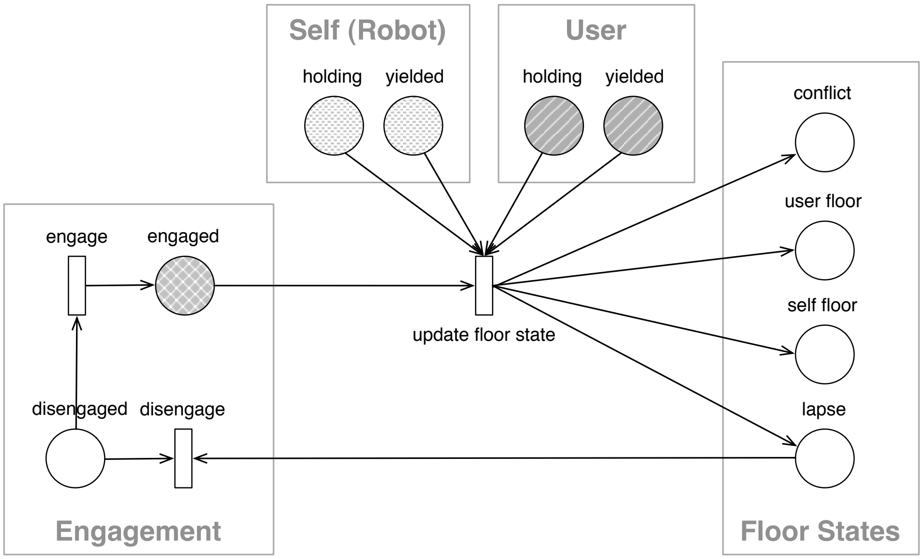

Figure 2. This diagram shows the relationship between engagement, turn states for the robot and the user, and dyadic floor states. The floor state update is based only on the time that $p_{holding}$ and $p_{yielded}$ from the *Robot* and *User* processes contain tokens. The full *User* model is shown in Figure 3, and the full *Robot* turn-taking control process is shown in Figure 4.

transition $t_{segment}$ determines whether the user is currently inside or outside of a speaking segment. This allows the modeling of states in which a user appears to be holding the floor through gesture or gaze cues but is not currently speaking (for example, after a statement of "Um...", which is typically followed by a pause and a longer spoken turn) (Clark & Tree, 2002).

3.2.3 Full turns versus backchannels CADENCE makes a distinction between turn-taking style and turn-taking content. The turn-taking control mechanisms for seizing, holding, yielding, and auditing as described in Sections 3.2.4 and 3.2.5 function to regulate the flow of turn content. The turn content itself falls into the categories of either full turns or backchannels in the system. Execution chains for both are are depicted in Figure 4. The interruption chain for full turns is addressed in more detail in Section 3.2.4.

In each of these execution chains in Figure 4, the tokens and places are of type `Turn`. This construct is defined to comprise a set of acts (of type `Act`), where an act is a modality-specific action that can be started and stopped. Each act is associated with a function that returns the act's start time, an offset defined relative to the beginning of the turn. Thus, while the robot is holding the floor by running the transition t_{hold}, the turn-taking process delegates each act to its modality-specific execution process, each of which is a connected TPN subgraph that handles act execution and interruption correctly for the resources that it controls (Chao & Thomaz, 2012). Running t_{yield} causes modality-specific execution to interrupt current acts and abandon future acts in the Turn. CADENCE act types currently include `SpeechAct`, `GazeAct`, `GestureAct`, and `ObjectAct` for manipulation (some examples are given later in Section 4.1.1). Since the floor regulation model is intended to be generic, it relies on a separate context model to provide context-appropriate turn content. Sometimes default turn-taking behavior is specified, which can be overridden by acts; for

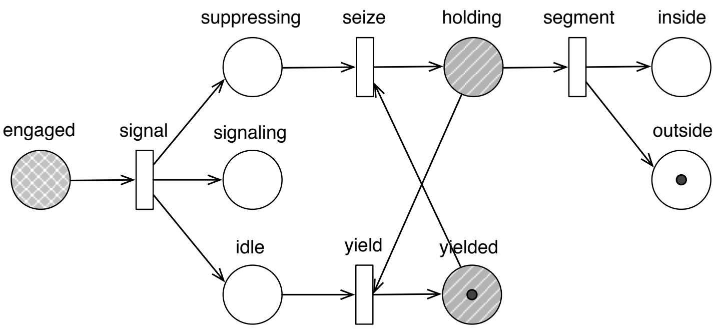

Figure 3. The user state model is based on perceptual signals for the user speaking, gesturing, and gazing away or at the robot. The places $p_{holding}$ and $p_{yielded}$ are used in conjunction with those of the robot to determine the dyadic floor state, as shown in Figure 2.

example, if no `GazeAct` is specified for a given turn, the default behavior of the turn-taking process is to gaze away from the partner's face while taking a turn and to gaze back at the partner's face when yielding the floor.

Backchannels are a special form of action that communicate a speaker's current desire for, or avoidance of, the floor. Originally, backchannels were considered to be behavior that supported and maintained engagement with the current floor owner (Yngve, 1970). However, subsequent analyses of backchannels have showed a diversity of purposes (Bangerter & Clark, 2003; Jurafsky, Shriberg, Fox, & Curl, 1998). There can also be ambiguity in the difference between backchannels and short-duration full turns. For example, the same spoken utterance of "uh huh" can denote a semantic affirmative or simply an acknowledgment that the other speaker said something, depending on the particular context and speakers.

CADENCE currently supports backchannels as either continuers or incipient speakership markers, which communicate contrasting intentions towards floor ownership. The *continuer* is used when auditing to communicate that the current floor owner should continue holding the floor. A commonly used continuer in English is "mhm." The *incipient speakership marker* is used to acknowledge that the floor owner has held the floor for some time and to communicate a desire or intention to seize the floor. An example occurring in English is the construct, "Yeah, but..." followed by a full turn.

In the system, the backchannel is a subclass of Turn because it comprises the same types of acts, such as head gestures and spoken utterances. However, backchannels are run through a separate control chain in the turn-taking process. This design decision was made so that time spent backchanneling would not count against floor time. Another reason is that these backchannels are not interruptible and do not communicate domain-specific information, which contrasts with the full-turn execution chain. After performing our system evaluation, we consider that some of these assumptions may need to be revisited; these points will be raised in Section 7.1.

3.2.4 Yielding and auditing In previous work, we investigated the role of a robot's self-interruptions as a mechanism for yielding the floor (Chao & Thomaz, 2012). Robot self-interruptions are possible in our system when the *interrupt self* parameter is set to true.

We extend our previous work on action interruptions by including hesitations as a precursor to the robot's interrupting its turn. This logic is shown in the interruption chain at the top of Figure 4. The motivation for this addition was the observation that completely aborting the current turn was too extreme of a reaction when responding to short signals from the human. For example, sensor noise or fidgeting from the human could cause an extended manipulation action to abort. To reduce this level of commitment, when t_{hold} is active (indicating that the robot is taking a turn), the robot also decides whether or not to hesitate. This decision is based on $p_{signaling}$ from the user process owning a token, as described in Section 3.2.2 and shown in Figure 3. Hesitating results in pausing the current turn, which pauses active acts and prevents later acts in the turn from starting. Within a small *hesitation resolution* deadline, the robot must then decide whether to interrupt or resume its turn. If the user state transitions to $p_{suppressing}$ before the deadline, the robot proceeds to interrupt itself; otherwise, it resumes the turn. In practice, the change in the robot's behavior resulting from

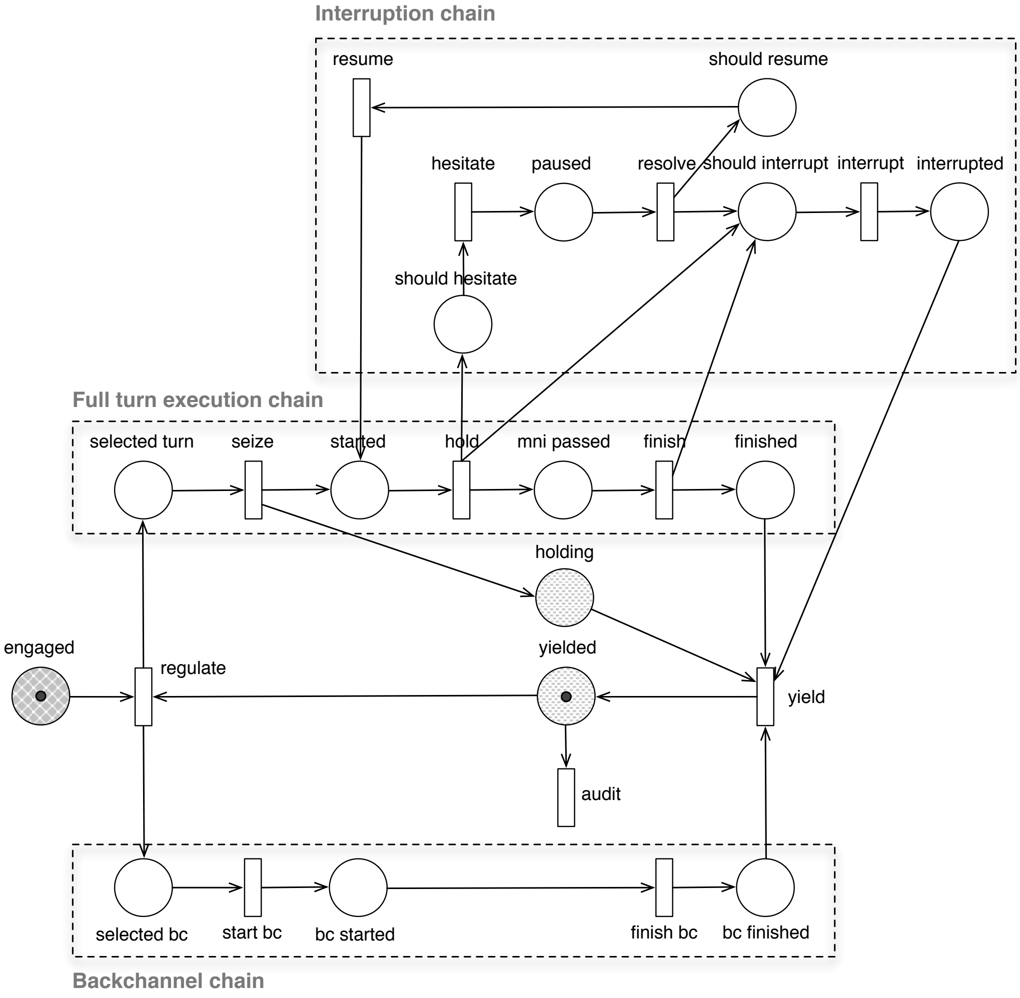

Figure 4. This diagram shows control chains for the robot's turn execution. The primary control chain is the full-turn execution chain, which is used for the playback of full turns. A full turn is moved through the interruption chain if the robot determines at some point while holding that it needs to yield the floor. The backchannel chain is an abbreviated alternative control flow for short, uninterruptible turns that do not convey domain information.

hesitation provides feedback to the user that some signaling was detected and allows the user to decide whether to back off or try seizing the floor.

The pausing and resumption behavior for acts varies depending on a particular act's modality. For gesture and manipulation, the robot pauses by maintaining its current pose. For gaze acts, the robot looks at the human's head when pausing and returns to the gaze act target when resuming. For speech acts, the robot stops speaking just as it would when fully interrupting itself. However, if a speech act is resumed, the robot starts again at the beginning of that particular speech act. This works well when turns comprise multiple speech acts in the form of phrases or turn construction units (Sacks, Schegloff, & Jefferson, 1974), which we implement for our contextual instantiation described in Section 4. In the linguistics literature, the retake of these utterances is sometimes referred to as "recycled turn beginnings" (E. A. Schegloff, 1987).

CADENCE also supports mechanisms for turn interruption from our prior work. When running t_{hold}, the robot may skip directly to interrupting its current turn if the dyad has been in a state of conflict for some amount of time, which we define as the *conflict tolerance* parameter. The robot may also interrupt its turn after the point of *minimum necessary information (MNI)* has passed and the user is ready to proceed, which indicates that the goal of the turn has been achieved (Chao et al., 2011; Thomaz & Chao, 2011). This shaves time off of the ends of turns in order to support increased fluency when the dyad is well-practiced.

After the robot has yielded the floor, whether through a mechanism of interruption or from completion of turns or backchannels, the robot runs the transition t_{audit} to control behavior that supports the user's holding of the floor. This involves gazing in directions appropriate for establishing joint attention within the context, such as the user's hands, as well as periodic glances at the floor holder's face. The approach taken here for auditing agrees with the model proposed by Holroyd et al., which was derived from video analysis of human-to-human turn-taking interactions (Holroyd et al., 2011).

3.2.5 Seizing the floor After having yielded the floor, the robot must decide when to take another turn. Its moment-to-moment options are to take a full turn, to backchannel, or to wait another cycle and delay the decision. In Figure 4, the transition $t_{regulate}$ is responsible for making this decision and placing the result in $p_{turn-selected}$ or $p_{selected-bc}$ if a turn or backchannel is selected. In making this decision, the robot tries to maintain a *floor factor* k_f, which describes the ratio between the robot's holding of the floor and the user's holding of the floor. That is, given all intervals $[r_\alpha, r_\beta)$ denoting $p_{holding}$ owning a token for the robot and $[u_\alpha, u_\beta)$ denoting the same for the user, we define a floor factor difference on a turn T with duration L_T:

$$\Delta(T) = \left| k_f - \frac{L_T + \sum_i r_{\beta_i} - r_{\alpha i}}{\sum_j u_{\beta_j} - u_{\alpha j}} \right| \tag{1}$$

Similarly for a backchannel BC, we define the difference as:

$$\Delta(BC) = \left| k_f - \frac{\sum_i r_{\beta_i} - r_{\alpha i}}{L_{BC} + \sum_j u_{\beta_j} - u_{\alpha j}} \right| \tag{2}$$

These values, in addition to $\Delta(W)$ using Equation 2 for the option of waiting to delay the decision until the next cycle, are used in the regulatory decision-making process. The relevance of a full turn, backchannel, or delay at any given moment additionally depends on multiple conditions with timing constraints. The details of this process are depicted in Figure 5. Paths throughout the tree in the figure are terminated with the following strategies for seizing or avoiding the floor, resulting in selection of backchannels or full turns:

- **Response** – a full turn that is taken in response to the user's previous turn. This occurs after a duration known as the *response delay* has passed since the user yielded the floor.
- **Deflection** – a backchannel continuer that is taken in lieu of a response, in order to avoid seizing the floor. Backchannels cannot occur more often than a period defined by the *backchannel spacing* parameter.
- **Support** – a backchannel continuer inserted between the user's speaking segments while the user continues to hold the floor in other modalities. This conveys support for the user's floor ownership.
- **Deep interruption** – a full turn that is started while the user is holding the floor. The parameter *interrupt user* must be set to true for this strategy to be used, and the user must have been holding the floor for at least a duration referred to as *interrupt patience*.
- **Interjection** – an incipient speakership marker followed by a full turn. Like deep interruption, *interrupt user* must be set to true, but interjection occurs between user speaking segments (such as the support strategy) rather than during them.
- **Lapse recovery** – a full turn that is taken after having been in a lapse for longer than a duration referred to as the robot's *lapse tolerance*. This allows the robot to take initiative to recover from a period of awkward extended inaction. Lapses longer than 3 to 4 seconds are associated with lower communicative competence in humans (Wiemann, 1977).

For all strategies based on full turns, the setting of the *require gaze to seize* parameter to true further restricts the robot's seizing of the floor to moments when the user is gazing at the robot. If none of these conditions is satisfied, the robot waits without selecting a full turn or a backchannel, and the decision is repeated again at the next clock cycle.

3.2.6 Turn-taking parameters As a summary, the following set of system parameters controls the dynamics of the turn-taking system, which results in different turn-taking styles. The majority of these are values or ranges of time specified in milliseconds. The purpose of each parameter is defined below with an explanation for its expected impact on interactions. In our experimental evaluation described in Section 5.1, we specify settings of these parameters for producing two contrasting robot behavior styles. The parameters that differ across the conditions in this experiment are demarcated below with an asterisk.

- *Floor factor** – the robot's desired ratio of itself holding the floor to the user holding the floor.
- *Response delay* – how much time to wait after the user yields the floor before seizing the floor. Rather than being necessary for computational resource reasons, this value is used to determine how much opportunity to allow the user to seize the floor again after yielding.
- *Interrupt user** – whether the robot can try to seize the floor when the user is holding the floor.
- *Interrupt self** – whether the robot hesitates or interrupts its current turn if the user tries to seize the floor.
- *Conflict time* – how much time both the robot and the user spend continuously holding the floor before the robot considers the dyad to be in a state of conflict.

- *Lapse time* – how much time both the robot and the user spend continuously auditing before the robot considers the dyad to be in a lapse.
- *Conflict tolerance** – how much time the robot tolerates a state of conflict before being forced to interrupt its current turn.
- *Lapse tolerance** – how much time the robot tolerates being in a lapse before being forced to seize the floor.
- *Interrupt patience* – the minimum amount of time the user has spent continuously holding the floor before a deep interrupt is allowable. This parameter is used only if interruptions of the user are permitted.
- *Hesitation resolution* – the deadline after the robot hesitates that the robot must decide whether to resume or interrupt its currently paused turn.
- *Act spacing** – a uniformly sampled range of time that separates acts within a turn. Higher values encourage interjections from the user.
- *Backchannel spacing** – a uniformly sampled range of time that separates consecutive backchannels.
- *Require gaze to seize** – whether the user must be gazing at the robot in order for the robot to seize the floor.

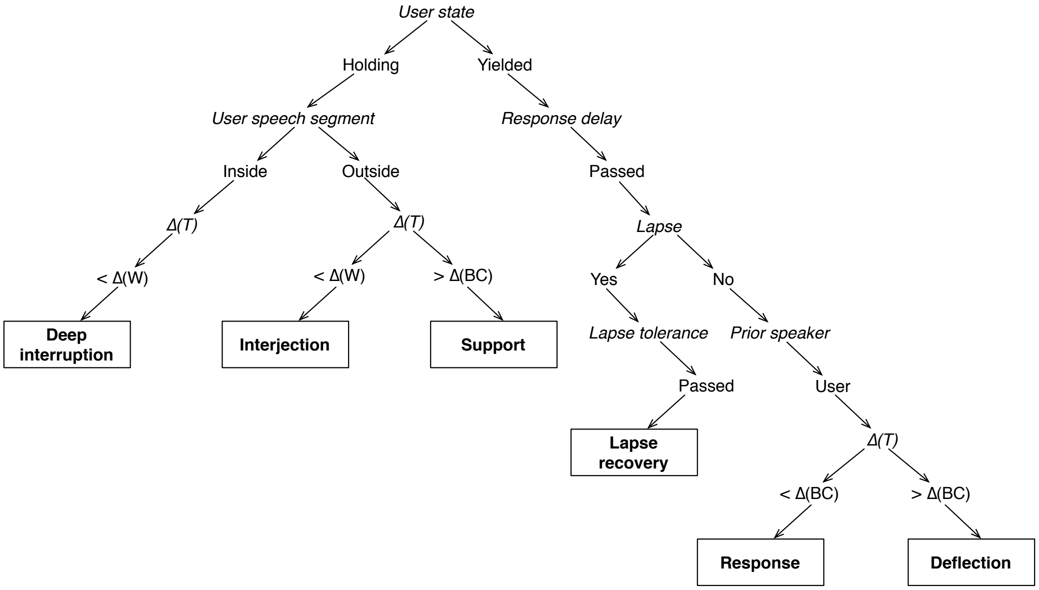

Figure 5. The decision-making process inside of the transition $t_{regulate}$ regulates floor ownership based on differences from a predetermined *floor factor* parameter that relates the robot's and the user's holding of the floor. Other conditions additionally constrain the selection of a full turn or backchannel, which leads to placing a token in $p_{turn-selected}$ versus $p_{selected-bc}$.

4. Contextual Instantiation

A central challenge in the design of an integrated turn-taking system is the role played by context. A system that regulates turn-taking inseparably from domain specifics has little utility, as it must be redesigned for each new domain. Thus, we strive in the design of our turn-taking controller to minimize the effort required for this transfer. In our system, the aim is for the robot's behavioral processes to be reusable across domains, and the context model is a TPN subgraph connected to these behavioral processes that gets replaced for each new domain (Chao & Thomaz, 2012).

Related to the notion of abstracting the skill of turn-taking from domain-specific knowledge is the sensitive balance between using bottom-up and top-down perception for driving robot behavior. As an example showing this contrast, a top-down perceptual process might be a grammar-based speech recognizer that relies on endpointing before returning results, while its bottom-up counterpart may be a filter for the presence of speech in the audio signal. Of course, the spectrum of semantic knowledge encapsulated by differing perceptual techniques is more fluid than these extremes.

In this paper, we deliberately design an experimental context to be as open-ended as possible to focus on turn-taking driven by bottom-up signals. This includes a decision to have the robot speak an artificial language to circumvent top-down speech recognition. Perhaps contrary to intuition, this design actually allows us explore a space of more complex interactions, as compared to domains bounded by task constraints that we have used in previous work. We find that this open-ended domain uncovers the innate sense of obligation to speak, act, or yield that is driven by a human's intuition for turn-taking without being complicated by issues in task and natural language understanding. Certainly a future goal of our work is the design of interfaces to support more context-appropriate turn-taking behavior, as well as the application to CADENCE to more practical domains.

In this section, we describe the implementation of contextual components for this domain, to be used in an experiment with users described in Section 5.

4.1 Setting

The interaction setting is intended to support a relatively open-ended multimodal dialogue about toys at a tabletop. Participants have access to a bin of objects containing toys such as blocks and small plush animals, which they use to play with the robot. The catch is that the robot and the human do not speak the same language, so the domain is free from task semantics or the need for natural language understanding of the user's turns.

4.1.1 Robot actions Robot turns in this context are constructed as random combinations of acts in the modalities of speech, gesture, manipulation, and gaze. Figure 6 has examples of the resulting behavior. Full turns contain the following types of acts:

- Acts in the speech modality are phrases in an artificial language. These phrases were pregenerated by sampling random strings of phonemes. The phrases vary from approximately 1–5 seconds in duration and are grouped by the prosodic endings of ellipsis, exclamation, interrogation, and statement. Each turn consists of 1–3 of these phrases, of which the last phrase is always one of either exclamation, interrogation, or statement, and its antecedents all have elliptical prosody. The *act spacing* parameter is used to set the timing for these phrases.

- Head gestures include a head nod (looks like "yes"), a side-to-side head shake (looks like "no"), and several for communicating uncertainty through head tilt and sideways eye motions.

- Arm gestures are animations previously retargeted from human motion capture that were selected based on their interpretability as attitude towards an object or event. The communicative intentions of the human performing the gestures were shrugging, "aww shucks," "phooey," and presentation.

- Object-directed arm actions include picking, placing, or pointing at objects on the table. A manipulation action is accompanied with gaze toward the object of reference unless another gaze act is specified for the turn.
- A gaze act toward one of the objects on the table may also be selected; thus, an arm or head gesture can be interpreted as being directed toward the object.

Backchannels were restricted to head nodding or shaking gestures and 1–3 phonemes sampled from a limited phoneme set. Incipient speakership markers were sampled from the space of English vowels. Continuers were sampled from the consonants /m/, /n/, /h/ and the vowel /∧/. These different phoneme sets were intended to show the backchannels' contrasting functions, reflecting the different backchannel distributions that occur in natural languages (Jurafsky et al., 1998; Clancy, Thompson, Suzuki, & Tao, 1996).

4.1.2 Perception A Microsoft Kinect was used for tracking the human's skeleton using the Kinect Software Development Kit. Specifically, the human head and both hand positions were used for the robot's auditing behavior. The head and shoulder positions relative to the participant's hips were also used to determine whether participant gaze was oriented toward or away from the robot's head.

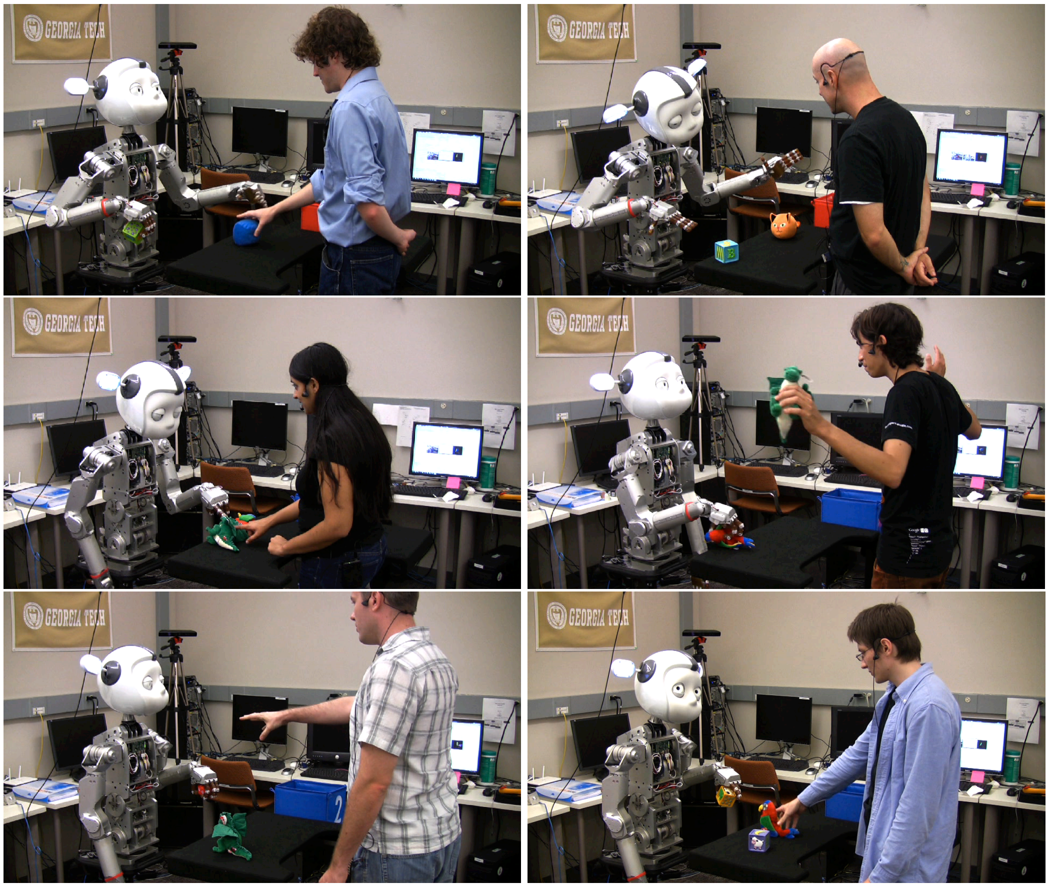

Figure 6. Examples of participants interacting with Simon in the context of tabletop object play.

The human was considered to be gesturing if either of the hands were in motion over the past 800 milliseconds, or if the hands were outstretched over the table. These cues were used in t_{signal} in Figure 3 of the user model.

The signal for the presence of user speech was detected through a Pure Data module (Puckette, 1996) for determining the pitch of an audio signal. This signal was used in t_{signal} and $t_{segment}$ in Figure 3. Participants wore a headset with a directional microphone to minimize the detection of the robot's speech. The audio signal was preprocessed with an amplitude filter that was tuned to ignore the robot's voice.

To detect tabletop objects for attention and manipulation, tabletop segmentation was performed using an overhead Asus Xtion. This object perception was domain-specific and thus occurred within the swappable context model process. The table was detected using a plane extraction technique and subtracted to yield 3D point clusters representing the objects (Trevor, 2012). Only clusters detected within the boundaries of the table plane and above the table were considered in the context. Additional tracking was performed to reason about occluded objects from the perspective of the Asus sensor using knowledge of both robot kinematics and human kinematics (from the Kinect skeleton). Clusters within a certain distance from robot or human link positions were not considered to be objects, but previously detected clusters exceeding a distance from agent links were considered to be occluded by agents and thus preserved.

5. Experiment

To evaluate the preliminary implementation of CADENCE as described in Section 3, we designed a between-groups user study in which our robot Simon used the system to control its autonomous behavior within a situated dialogue about objects. The primary purpose of this experiment is system evaluation. To validate that our parametrized model of floor regulation is effective in achieving different social dynamics, we compared the behavior exhibited by the robot across contrasting parameter settings, as well as the effects of these differences on user behavior and user perceptions. The experiment also enables us to analyze any turn-taking errors that occur in either of these parametrizations, for insights on future work.

5.1 Parameter groups

The user study contained two conditions designed to investigate situations in which a robot shows different levels of initiative or control:

- *Active condition* – The robot tries to act twice as often as the human and deliberately interrupts the human to maintain this ratio.
- *Passive condition* – The robot tries to act half as often as the human, hesitates and interrupts its own actions, and often backchannels to avoid seizing the floor.

Table 1 shows the specific parameter differences between the conditions. We recognize that these settings represent only two points in the large space of turn-taking styles possible.

5.2 Procedure

In total, there were 30 participants who interacted with Simon in this user study (15 per condition), of which 8 were female (4 per condition). The age range of participants was 17 to 45 years old, with a mean age of 23.5. Ten participants reported experience interacting with young children (5 per condition), such as teaching or babysitting. The participants were recruited from the campus community through mailing lists.

Table 1: Parameter settings that differed between the two experimental conditions.

Parameter	**Active condition value**	**Passive condition value**
Floor factor	2.0	0.5
Interrupt user	true	false
Interrupt self	false	true
Conflict tolerance	N/A	1000 ms
Lapse tolerance	500 ms	4000 ms
Act spacing	50–250 ms	500–1000 ms
Backchannel spacing	2000–4000 ms	4000–6000 ms
Require gaze to seize	false	true

Each participant was randomly assigned one of the conditions and interacted with Simon for two sessions of three minutes each within that behavioral condition. The only difference between the sessions was that the set of objects was changed, and participants were informed that this was the only difference. Participants were told that they should teach Simon about the objects as if he were a young child of about three years of age, but they would not understand what Simon was saying because he would be speaking a foreign language. They were encouraged to talk about properties of the objects, tell stories about interactions between them, or otherwise play with them in a way that was appropriate to a young child.

In addition, participants were primed in ways that constrained their behavior. They were told that Simon could see both their hands, their head, and objects that were on the table, but that if they wanted Simon to attempt to interact physically with any objects, those objects needed to be on the table and their hands could not be covering the objects. This instruction was to prevent users from attempting handoffs to the robot, which were not supported in this study. Finally, participants were instructed to continue engaging the robot and to avoid turning to the experimenter for the entirety of the interaction sessions, even if they were uncertain what they should do.

To trigger the interaction, participants were told to wave to the robot and say “Hello Simon” after the robot’s ear lights turned on. For all sessions, the robot started with an uninterruptible greeting turn, comprising a wave gesture with a spoken exclamation. A three-minute timer was started at the end of this turn. At the end of the three minutes, Simon completed his current turn and turned off his ear lights to signal the end of the session.

5.3 Measures

5.3.1 Post-study questionnaire After the two interaction sessions were completed, users were asked to fill out a survey with the following questions:

1. How did you find the pacing of the interaction? (slow, medium, fast)
2. Who led the interaction? (Simon, me, about equal)
3. Please rate the following statements about the interaction with Simon. (1 = strongly disagree, 7 = strongly agree)
    (a) Simon was responsive to my actions.
    (b) I had influence on Simon’s behavior.
    (c) Simon had influence on my behavior.

(d) Simon listened to me.

(e) Simon talked over or interrupted me.

(f) I had to spend time waiting for Simon.

(g) Simon had to spend time waiting for me.

(h) The interaction pace felt natural.

(i) There were silences where nothing happened.

(j) There were overlaps where we both tried to act.

(k) There were awkward moments in the interaction.

4. How would you classify... (1 = strongly introverted, 7 = strongly extroverted)

   (a) ... Simon's personality?

   (b) ... your own personality?

5. (Open-ended) List some adjectives describing Simon's personality.

6. (Open-ended) Please provide a critical review of Simon's social skills.

5.3.2 Logged data Important system events were also logged for each interaction session with timestamps to millisecond precision. These system logs included:

- Petri net events, including transition enables and disables, tokens changing places, tokens changing values, and tokens being spawned or destroyed;

- events for each act being started, paused, resumed, or stopped;

- and reason codes for each full turn or backchannel taken, according to the strategies specified in Section 3.2.3 and Figure 4.

The perceptual data for the human was logged at the framerate of the system, which averaged 30 Hz. This data included pitch and onsets detected from the microphone and all transforms for human skeletons detected from the Kinect. The robot's joint positions were also logged at this rate. A video was taken of each interaction session for future video coding analysis.

6. Results

In this section, we present some of the results of our evaluation based on participants' subjective responses, the robot's behavioral data, and participant speech data. We observed that the speech presence signal was extremely robust during the study, but the human skeleton data generally was not (due to arm occlusions and noise). Thus, we can reliably examine human spoken turns, but not human floor-holding that relies on gesture or gaze. The latter requires video coding of the data for accurate analysis, which we leave to future work.

6.1 Differences in robot behavior

Our first analysis is a manipulation check that examines whether or not the system's control of these floor regulation parameters actually resulted in different robot behavior across these two conditions. For each modality, Figure 7 compares the fraction of time spent in each behavioral state across all subjects' data for each condition. It can clearly be seen that the active robot spent more time attempting to hold the floor, resulting in taking more full turns; this then led to increased gesturing, speaking, and gazing away from the person's body. In contrast, the passive robot maintained a closer balance between holding the floor and auditing, resulting in more backchannels and gaze at the person's head or hands relative to the active condition.

The fraction of time spent in each of these states differed significantly across conditions for all

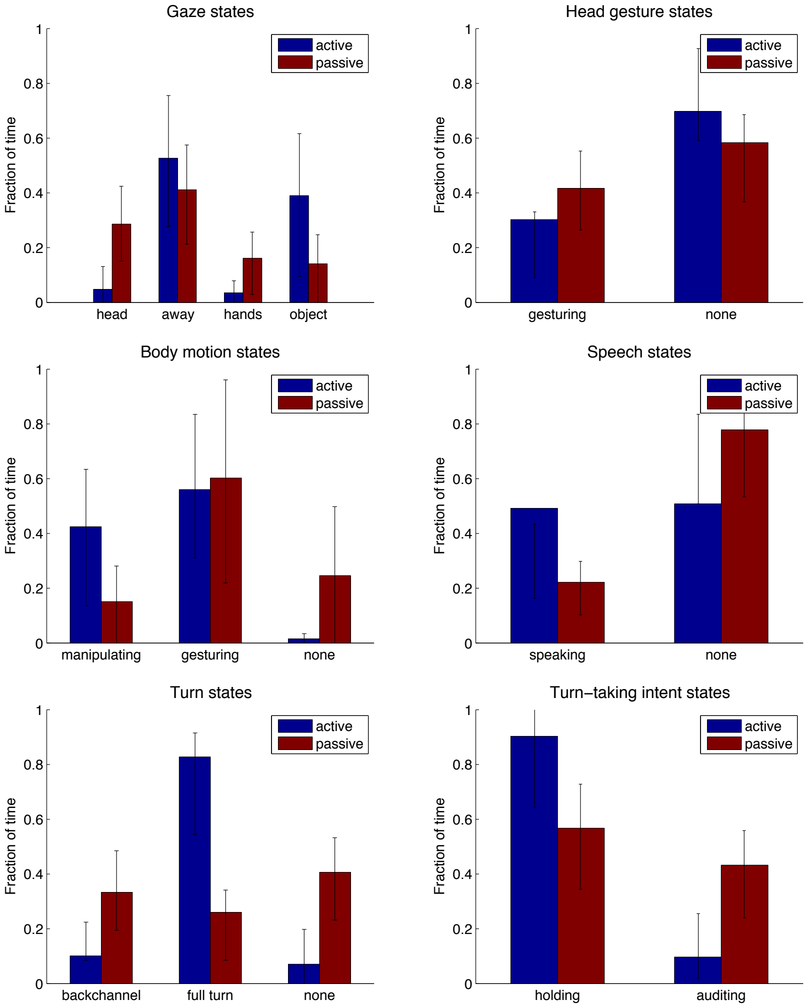

Figure 7. The time that the robot spent in each state, compared across conditions. Each chart shows data for a specific modality. Differences across conditions are significant to $p < .01$ for all modality states except for arm gesturing.

Table 2: Adjectives describing the robot's personality that were reported in only one condition.

Adjective	**Active**	**Adjective**	**Passive**
aloof, spacey, distant	3	shy	3
outgoing, extroverted	2	moody, temperamental, flighty	3
gregarious, loud	2	unresponsive, silent	2
bold, confident	2	responsive	1
enthusiastic	1	misunderstood	1
cautious	1	sweet	1
slow	1	naive	1
introverted	1	confused	1
		helpful	1

Table 3: Adjectives describing the robot's personality that were reported in both conditions.

Adjective	**Active**	**Passive**
curious, inquisitive	6	6
talkative, rambling	5	3
unattentive, absent-minded, ADD, distracted	3	1
childish, child-like	2	1
attentive, observant	1	2
stubborn, willful	1	2
playful	1	1
contemplative	1	1

modality states except for arm gesturing. We also confirmed that the robot did not behave significantly differently across the two interaction sessions for either condition. Because our subsequent analysis focuses on speech data, we additionally state here that the robot spoke significantly more in the active condition (M = 87.0 sec, SD = 20.1) than in the passive condition (M = 39.6 sec, SD = 9.4), $t(13)$ = 1.30, $p <$ .001.

6.2 Perception of behavioral differences

Subjective responses indicate that the difference in behavior was perceptible to participants. They were significantly more likely to agree with the statement, *"Simon talked over or interrupted me,"* in the active condition (M = 6.00, SD = 1.07) than in the passive condition (M = 4.80, SD = 1.21), $t(14)$ = 2.96, p = .01. We also found marginal significance for agreement with the statement, *"There were silences where nothing happened,"* where the passive condition reported a higher average value (M = 4.53, SD = 1.64) than the active condition (M = 3.40, SD = 1.76), $t(14)$ = 2.02, p = .06.

Moreover, this difference in robot behavior impacted their perception of the robot's personality. Participants in the active condition perceived Simon as significantly more extroverted (M = 4.93, SD = 1.53) than participants in the passive condition (M = 3.46, SD = 1.06), $t(14)$ = 4.01, p = .001. Subjective reports of participants' own personality introversion ratings did not differ significantly across the conditions. When the participant's self-rating was subtracted from Simon's rating, the average difference was M = 1.13, SD = 1.85 in the active condition and $M = -1.11$, SD = 1.67 in the passive condition, $t(14)$ = 3.86, p = .002.

Adjectives reported by participants to describe Simon's personality are shown in Tables 2 and 3. Synonyms are grouped in these listings. Table 2 lists all adjectives reported in only one condition, and Table 3 lists all adjectives that were reported in both conditions. In some cases, opposites were reported within the same condition (*extroverted* and *introverted* in the active condition, *responsive* and *unresponsive* in the passive condition, *attentive* and *inattentive* in both conditions), showing the breadth of subjective experience in the study. Although it is difficult to make strong claims about patterns across these open-ended responses, our impression is that the passive condition elicited more sympathetic and underdog-like descriptors, whereas the active condition resulted in more dominant adjectives.

Overall, these results confirm that the system is capable of manipulating the robot's initiative and status within an interaction, and that humans can perceive the effects of this manipulation.

6.3 Impact on human behavior

We have determined that the robot behaved differently across the conditions and that humans could perceive these differences, but we additionally want to analyze the extent to which the manipulation of floor regulation impacted the behavior of the human. As mentioned previously, we found that we were able to track the human's speaking turns reliably. We analyzed occurrences of robot and user speech across the conditions based on logged data. The starts and ends of user speech segments were determined from the speech presence signal based on a window gap size of 250 milliseconds. Two user logs (one per condition) were generated incorrectly and thus are omitted from this analysis.

In examining this data, we find that participants spoke significantly more in the passive condition ($M = 59.5$ secs, $SD = 26.2$) than in the active condition ($M = 40.1$ secs, $SD = 18.1$), $t(13) = 3.70$, $p = .003$. This is likely due to human aversion to overlapping speech, which led to inhibition of user speech in the active condition but created more opportunities for user speech in the passive condition. In fact, we found that the user spoke slightly but significantly more in the second session in the passive condition ($M = 64.2$ secs, $SD = 28.6$) when compared to the first session ($M = 54.7$ secs, $SD = 23.5$) , $t(13) = 3.72$, $p = .003$, but this was not true for the active condition. This could be explained by the users exhibiting more tentative and uncertain behavior in the first encounter with the robot but taking more control after having seen the robot's passive behavior. As can be expected from these results, the ratio of robot speaking to user speaking also differed significantly across conditions. This ratio[1] was $M = 3.28$, $SD = 3.80$ in the active condition and $M = 0.85$, $SD = 0.56$ in the passive condition, $t(13) = 4.01$, $p = .001$.

We also discovered that there was significantly more overlapping speech in the active condition ($M = 13.1$ secs, $SD = 7.6$) than in the passive condition ($M = 8.3$ secs, $SD = 4.9$), $t(13) = 3.97$, $p = .002$. We did not analyze whether these overlaps were robot interruptions, user interruptions, or simultaneous starts, since this requires more contextual knowledge and would need to be determined through video coding.

Given that the user spoke more to a passive robot than an active one, we hypothesized that there may have been regularities in the robot's modality-specific actions that encouraged the user to seize the floor. We compared robot behavioral states at the starts of user speech turns to the robot's overall modality-specific behavior state distributions for each condition and found that this hypothesis was not supported by the data. Figure 8 shows the high similarity between the distributions for the active condition (data for the passive condition is even more similar, and thus is not provided). An exception is the speech modality, where the absence of robot speech favors a user seize attempt. Hence, in this data set featuring open-ended human-robot turn-taking, a simple "nod and a glance" (Cassell & Thorisson, 1999) does not suffice for controlling or predicting when users will take turns.

[1] Note that these ratios differ from the floor factor parameter setting because this result only takes into account speaking turns, whereas the floor factor ratio also accounts for holding across other modalities.

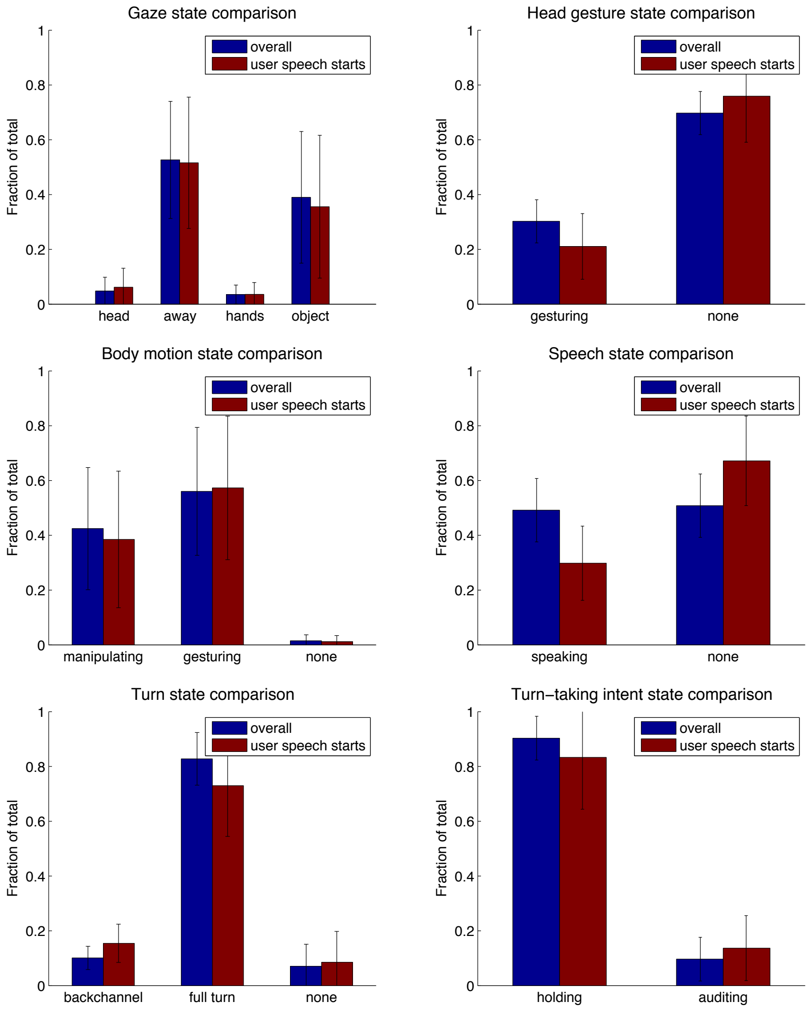

Figure 8. This data is taken from the active condition only. The figures compare distributions of robot behavior data at the times of user speech starts to the overall distribution for the active condition. Overall, the distributions appear highly similar, making it difficult to predict onsets of user speech. The most substantial difference can be seen in the speech state comparison.

7. Discussion and Future Work

Results indicate that CADENCE is able to manipulate social dynamics through variations in floor regulation behavior. Experimental manipulation of our model parameters resulted in significantly different robot behavior that caused participants in different conditions to attribute significantly different personality types to the robot. Importantly, this manipulation in robot behavior succeeded in eliciting different behavior from the human partner across the two groups, changing the social dynamics of the dyad. The human partner acted longer and more often when interacting with a passive robot than with an active one.

Here we make several additional observations about challenges for future work.

7.1 Improvements to backchanneling

In general, there were fewer significant differences in subjective ratings across conditions than we had expected. In both conditions, subjects thought the interaction was slow to medium in pacing, about equal in who was leading, and had overlaps in turns. Subjects also did not consider Simon more of a listener in the passive condition. In examining the data in conjunction with anecdotal observations of the study, it seems that spoken backchannels did not function as supportive auditing behavior as much as hypothesized. This may explain why several participants perceived Simon as *"talkative"* even in the passive condition (Table 3); users may have perceived Simon's backchanneling as taking floor time, albeit in short spurts. In light of this, it may have been more appropriate for backchannels to count against the robot's floor time after all, but perhaps with reduced weight.

Improving timing and comprehensibility of backchannels seems necessary for generating more successful auditing behavior. Some part of the lack of accessibility of backchannels in this study can be attributed to the artificial language. When using true linguistic backchannels in a task context, it will be important to consider the different functional types of backchannels and the information they convey (Bangerter & Clark, 2003). In addition, backchannel timing is a sensitive issue. Backchanneling to maintain engagement after a lapse could be inappropriate if the user is thinking silently; such a backchannel could be perceived as disruptive or annoying. On the other hand, it can be completely appropriate to backchannel with quick spacing in response to short, high-information utterances from the human. More modeling of these types of information exchanges may be needed for the robot to be able to time backchannels appropriately.

7.2 Modality-specific bottlenecks

One simplification we made in our current implementation was to treat the conversational floor as a singular resource to be negotiated by the two parties. This resulted in action across all modalities being combined to classify holding behavior. Realistically, interaction dynamics are also strongly defined by modality-specific bottlenecks. For example, overlapping speech is avoided, as well as close proximity that can lead to physical collision or uncomfortable social distance, but cross-modality simultaneity such as speech from one party and gesturing from another does not necessarily constitute a conflict. On the other hand, correlation of actions would still be expected across modalities due to the innate structure of information bottlenecking in the turn-taking interaction—for example, in the way that gaze behavior accompanies speech turns, and in the way that speech is synchronized with deictic gestures (Mondada, 2007).

A future challenge is to define the roles of these modality-specific bottlenecks more clearly when planning higher-level cross-modality turns. For example, seizing the floor to speak may inhibit the partner's tendency to speak but does not necessarily require that the partner stop acting in the workspace. Similarly, taking a turn using a particular object prohibits the partner from using that object but does not prohibit him from taking a speaking turn or from performing an action with another object in the workspace. Indeed, there were many instances in the study of a user speaking over Simon's manipulation actions (e.g. providing words of encouragement or semantic object labels) that were socially appropriate and non-conflicting. On the other hand, rigidly enforcing complete independence of individual modalities is similarly oversimplifying, as gaze and body motions do influence verbal expression and suppression; this explains why telephone conversation dynamics differ so significantly from face-to-face dynamics (ten Bosch, Oostdijk, & de Ruiter, 2004). Modeling appropriate constraints between the floor resource and modality-specific resources will be necessary in order to capture the fluid human-robot interaction that we desire from our system.

7.3 Contextual parameter setting

Our study demonstrated only two possible parameter settings of the system. We emphasize again that we do not universally advocate one of these conditions over the other, but simply state that our model allows the robot to exert some control over social dynamics and influence who takes more initiative in the interaction. The appropriate parameter setting should ultimately be dictated by social context. A point of interest for our future work is such a contextual setting of these parameters by recognizing detectable and generalizable characteristics of a social situation or an interaction partner.

As an example, a tour guide robot may benefit from "active" floor regulation parameters while leading a crowd, whereas a butler robot may need to rely on "passive" parameter settings when serving its owner. We believe that the system parameters are intuitive, and as such can be set to constants defined by a robot designer to suit a particular application or culture; or perhaps a personal service robot can be set to use custom parameters preferred by its owner. We are also interested in further exploring appropriate turn-taking behavior for a robot learning from a human teacher; a spectrum of passive to active behavior is possible for a robot active learner, not all of which is actually conducive to successful learning (Cakmak, Chao, & Thomaz, 2010).

In addition to static parameter settings based on a task domain, it may be useful to modulate the parameters dynamically throughout an interaction. This technique could be used to support status-elevating and status-lowering transactions (Johnstone, 1987). In addition, such modulation could allow better adaptation of the robot to the human. It is known that people gradually synchronize the timing of communicative behaviors to interaction partners over time (Burgoon, Stern, & Dillman, 1995), and we have also observed such convergence in our previous work (Chao et al., 2011); this capability could potentially improve a dyad's fluency. It may also be useful to consider adapting to the human's affective state, if it is perceivable through cues like vocal prosody. A stressed or angered human may desire different turn-taking dynamics than a relaxed one. Conversely, the control of the robot's turn-taking behavior could be modulated by an emotion model, serving a communicative purpose regarding the robot's internal state.

8. Conclusion

Embodied turn-taking between humans is characterized by seamless exchange of shared resources, including the conversational floor. Humans achieve fluency in turn-taking of such resources through multimodal and reciprocal behavior. We aim to enable robots to achieve similar fluency in human-robot dyadic interactions through a methodology of iteratively building autonomous robot behavior controllers and studying their resulting dynamics with human users, the results of which inform the next round of system implementation.

To this end, we have developed CADENCE, a Control Architecture for the Dynamics of Embodied Natural Coordination and Engagement, based on a timed Petri net representation. CADENCE controls a robot's multimodal turn-taking for dyadic face-to-face interactions with humans and includes a novel computational model that explicitly reasons about the four components of floor regulation: *seizing*, *yielding*, *holding*, and *auditing*. The turn-taking model-controller is intuitively parametrized to allow the robot to achieve a range of different social dynamics, which can be altered to target specific interaction scenarios.

We applied two contrasting parameter settings of CADENCE to a 30-participant experiment within an open-ended domain of tabletop object play with a human. Our system evaluation demonstrates that: (1) manipulating these floor regulation parameters results in significantly different robot behavior; (2) people are able to perceive this difference, as well as attribute different personality types to the robot; and (3) changing the robot's personality results in different behavior from the human, manipulating the social dynamics of the dyad. Our results confirm the utility of CADENCE for controlling turn-taking dynamics but also point to shortcomings specifically related to backchannel communication and the appropriate modeling of modality-specific bottlenecks. As we iterate on the system, we seek to address these issues in hopes of improving human-robot fluency and increasing the applicability of CADENCE to practical domains of human-robot cooperation.

Acknowledgements

This work was supported by the Office of Naval Research, grant number N00014-12-1-0484.

References

Baddeley, A., & Della Sala, S. (1996). Working memory and executive control. *Philosophical Transactions: Biological Sciences*, *351*(1346), 1397–1404, http://dx.doi.org/10.1098/rstb.1996.0123.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, *27*(2), 195–225, http://dx.doi.org/10.1016/S0364-0213(02)00118-0.

Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M., & Tomlinson, B. (2002). Integrated learning for interactive synthetic characters. In *Proceedings of the 29th annual conference on computer graphics and interactive techniques (SIGGRAPH)* (pp. 417–426, http://dx.doi.org/10.1145/566654.566597).

Bohus, D., & Horvitz, E. (2011). Decisions about turns in multiparty conversation: From perception to action. In *Proceedings of the 13th international conference on multimodal interfaces (ICMI)* (pp. 153–160, http://dx.doi.org/10.1145/2070481.2070507).

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1–2), 119–155, http://dx.doi.org/10.1016/S1071-5819(03)00018-1.

Burgoon, J., Stern, L., & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. New York: Cambridge University Press.

Cakmak, M., Chao, C., & Thomaz, A. L. (2010). Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 108–118, http://dx.doi.org/10.1109/TAMD.2010.2051030.

Cassell, J., Bickmore, T., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Requirements for an architecture for embodied conversational characters. In *Proceedings of computer animation and simulation* (pp. 109–120, http://dx.doi.org/10.1007/978-3-7091-6423-5_11).

Cassell, J., & Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, *13*, 519–538, http://dx.doi.org/10.1080/088395199117360.

Chao, C., Lee, J. H., Begum, M., & Thomaz, A. (2011). Simon plays Simon says: The timing of turn-taking in an imitation game. In *Proceedings of the IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 235–240, http://dx.doi.org/10.1109/ROMAN.2011.6005239).

Chao, C., & Thomaz, A. L. (2012). Timing in multimodal turn-taking interactions: Control and analysis using timed Petri nets. *Journal of Human-Robot Interaction*, *1*(1), 4–25, http://dx.doi.org/10.5898/JHRI.1.1.Chao.

Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese and Mandarin. *Journal of Pragmatics*, *26*, 355–387, http://dx.doi.org/10.1016/0378-2166(95)00036-4.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Tree, J. E. F. (2002). Using *uh* and *uh* in spontaneous speaking. *Cognition*, *84*, 73–111, http://dx.doi.org/10.1016/S0010-0277(02)00017-3.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, *3*(2), 161–180, http://dx.doi.org/10.1017/S0047404500004322.

Holroyd, A., Rich, C., Sidner, C., & Ponsler, B. (2011). Generating connection events for human-robot collaboration. In *Proceedings of the IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 241–246, http://dx.doi.org/10.1109/ROMAN.2011.6005245).

Johnstone, K. (1987). *Impro: Improvisation and the theatre*. New York: Routledge.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of the ACL/COLING-98 workshop on discourse relations and discourse markers* (pp. 114–120).

Kanda, T., Ishiguro, H., Imai, M., & Ono, T. (2004). Development and evaluation of interactive humanoid robots. In *Proceedings of the IEEE* (Vol. 92, pp. 1839–1850, http://dx.doi.org/10.1109/JPROC.2004.835359).

Mazur, A., & Cataldo, M. (1989). Dominance and deference in conversation. *Journal of Social and Biological Systems*, *12*(1), 87–99, http://dx.doi.org/10.1016/0140-1750(89)90023-7.

Mondada, L. (2007). Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. *Discourse Studies*, *9*(2), 194–225, http://dx.doi.org/10.1177/1461445607075346.

Murata, T. (1989). Petri nets: Properties, analysis and applications. In *Proceedings of the IEEE* (Vol. 77, pp. 541–580, http://dx.doi.org/10.1109/5.24143).

Mutlu, B., Shiwa, T., Ishiguro, T. K. H., & Hagita, N. (2009). Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 2009 ACM/IEEE conference on human-robot interaction (HRI)* (pp. 61–68, http://dx.doi.org/10.1145/1514095.1514109).

Nakano, M., Hasegawa, Y., Funakoshi, K., Takeuchi, J., Torii, T., Nakadai, K., et al. (2011). A multi-expert model for dialogue and behavior control of conversational robots and agents. *Knowledge-Based Systems*, *24*(2), 248–256, http://dx.doi.org/10.1016/j.knosys.2010.08.004.

Orestrom, B. (1983). *Turn-taking in English conversation*. Lund: CWK Gleerup.

Puckette, M. (1996). Pure Data: another integrated computer music environment. In *Proceedings of the international computer music conference* (pp. 37–41).

Raux, A., & Eskenazi, M. (2012). Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, *9*(1), 1–23, http://doi.acm.org/10.1145/2168748.2168749.

Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *Proceedings of the 2010 ACM/IEEE conference on human-robot interaction (HRI)* (pp. 375–382, http://dx.doi.org/10.1109/HRI.2010.5453163).

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696–735, http://dx.doi.org/10.1017/S0047404500001019.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, *29*(1), 1–63.

Schegloff, E. A. (1987). Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organization. In G. Button & J. R. E. Lee (Eds.), *Talk and social organisation* (pp. 70–85). Clevedon, England: Multilingual Matters.

ten Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In P. Sojka, I. Kopecek, & K. Pala (Eds.), *Text, speech and dialogue* (Vol. 3206, pp. 563–570, http://dx.doi.org/10.1007/978-3-540-30120-2_71).

Thomaz, A., & Chao, C. (2011). Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine Special Issue on Dialogue With Robots*, *32*(4), 53–63.

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal communication* (pp. 389–450). New York: Cambridge University Press.

Trevor, A. J. B. (2012). Fast segmentation of organized point cloud data. In *Proceedings of the international conference on robotics and automation (ICRA), Advanced 3D point cloud processing with Point Cloud Library (PCL).*

Tronick, E., Als, H., & Adamson, L. (1979). Structure of early face-to-face communicative interactions. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal communication* (pp. 349–374). New York: Cambridge University Press.

Wang, J. (1998). *Timed Petri nets: Theory and application*. Norwell, MA: Kluwer Academic Publishers.

Wiemann, J. M. (1977). Explication and test of a model of communicative competence. *Human Communication Research*, *3*(3), 195–213, http://dx.doi.org/10.1111/j.1468-2958.1977.tb00518.x.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567–577).

Authors' names and contact information: Crystal Chao, School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA. Email: cchao@gatech.edu. Andrea L. Thomaz, School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA. Email: athomaz@cc.gatech.edu.