Turn Taking Based on Information Flow for Fluent Human-Robot Interaction

Andrea L. Thomaz, Crystal Chao

■ *Turn taking is a fundamental part of human communication. Our goal is to devise a turn-taking framework for human-robot interaction that, like the human skill, represents something fundamental about interaction, generic to context or domain. We propose a model of turn taking, and conduct an experiment with human subjects to inform this model. Our findings from this study suggest that information flow is an integral part of human floor-passing behavior. Following this, we implement autonomous floor relinquishing on a robot and discuss our insights into the nature of a general turn-taking model for human-robot interaction.*

Turn taking, or use of reciprocal interactions of engagement, is the foundation of human communication. From the first months of life, infants learn to influence the timing, onset, and duration of turns in face-to-face interactions with caregivers through the use of cues such as eye gaze and vocalizations, and they express significant anxiety when deviations from the expected turn-taking pattern take place (Kaye 1977). Linguistics research with adults suggests turn taking is a dynamic and fluid process, whereby interactive partners alternately engage in the various phases of seizing, holding, and yielding the floor through turns and backchannels (Duncan 1974, Orestrom 1983). In order for socially situated, embodied machines to interact with humans properly, it seems logical that they should follow the same deeply rooted turn-taking principles that govern human social behavior.

Current interactions with robots are often rigid, ambiguous, and confusing. Humans issuing commands to robots often need to repeat themselves or wait for extended periods of time. Typically robots are unresponsive, opaque machines, and thus frustrating to deal with. The research goal in human-robot interaction (HRI) of unstructured dialogic interaction would allow communication with robots that is as natural as communication with other humans. While the field of HRI tends to focus on speech, gesture, and the *content* of interaction, our work additionally aims to understand how robots can get the underlying timing of social interaction right.

Our overall research agenda is to devise a turn-taking framework for HRI that, like the human skill, represents something fundamental about interaction, generic to context or domain. In this paper we first present our model of floor passing for a human-robot dyad and an experiment with human subjects to derive some of the parameters of this model. A primary conclusion of our data analysis is that human turn-taking behavior is dictated by information flow, and much of the robot's awkward

 ISSN 0738-4602

turn-taking behavior resulted from poor floor relinquishing. We present our follow-up implementation of autonomous floor relinquishing and its resulting behavior. Finally, we discuss our insights on turn taking for HRI.

Related Work

Turn taking is a common framework of interaction for social robots or robots designed to communicate with people. A seminal example is the Kismet robot that engages in childlike social interactions with a human (Breazeal 2002). Examples of human-robot turn taking are also seen in learning by demonstration/imitation interactions (Billard 2002, Nicolescu and Mataríc 2003, Breazeal et al. 2004), as well as recent works in human-robot music interactions that have either an implicit or explicit turn-taking component (Weinberg and Driscoll 2007; Michalowski, Sabanovic, and Kozima 2007; Kose-Bagci, Dautenhan, and Nehaniv 2008).

Often, examples of turn taking in social robotics take advantage of people's propensity to engage in turn taking, but there are some fundamental aspects of the ability still missing from the robot's perspective. The robot often takes its turn without recognizing that a person did or did not respond before continuing its behavior. The exchange does not influence the progression of the interaction (for example, engaging more with people or aspects of the environment that respond contingently). In general, turn taking is often specific to the domain behavior designed by the programmer, rather than an underlying part of the robot's behavior.

There are several examples of implementations of various turn-taking components. Prior works have done in-depth analysis of specific communication channels, such as gaze usage to designate speaker or listener roles (Mutlu et al. 2009) or speech strategies in spoken dialogue systems (Raux and Eskenazi 2009). Closely related is the problem of contingency or engagement detection, which requires implementing robot perception for awareness of the human's cue usage (Movellan 2005, Rich et al. 2010, Lee et al. 2011). Turn taking has also been demonstrated in situated agents (Cassell and Thorisson 1999), including management of multiparty conversation (Bohus and Horvitz 2010). Eventually, it will be necessary to integrate the piecewise work into an architecture for physically embodied robots.

Approach

Human-human communication is fraught with errors and uncertainty. Even with excellent perceptual capabilities, people still fall victim to unintended interruptions, speaking over each other, and awkward silences (Schegloff 2000). When moving to human-robot communication, the problems are intensified with noisy and limited sensor data. Thus the foundation of our computational approach to turn taking is the notion that the state of the world is only partially observable for the robot, and that there is much uncertainty in the problem of estimating which partner in the dyad has the floor to speak or act.

We describe turn dynamics as the first-order Markov process shown in figure 1. At each time step *t,* both the robot (R_t) and the human (H_t) can be in one of four floor states: seizing, passing, holding, listening. The structure of this model is inferred from prior knowledge about human turn taking (Duncan 1974, Orestrom 1983). Passing and seizing are the two transitory states where the floor is transitioning from one person to the other, while holding and listening are the two floor states of the dyad during a turn. Theoretically, R_t and H_t should always be in a seizing / passing or holding / listening configuration. But in reality many of the other "error" configurations will also have a nonzero probability. For example, at a pause in the dialogue it is common to see both parties try to seize the floor, and then one decides to relinquish to the other. Or the listening party may try to seize the floor before the holding party makes any passing cues, commonly called a barge-in. The research challenge is to learn the parameters of this model from data and involves two primary research questions: the timing model and the observation model.

The timing model represents how and when the human and the robot transition from state to state, that is, the human transition function $P(H_t \mid H_{t-1}, R_{t-1})$, and the robot transition function $P(R_t \mid R_{t-1}, H_{t-1})$.

The robot states are fully observable to the robot, but the robot has to infer the human's hidden floor state through sensory observations. The observation model, $P(O_t \mid H_t)$ models how the sensor data reflects the human floor state H_t.

Similar to Rich et al. (2010) and Bohus and Horvitz (2010), our approach is to analyze interaction data in order to find general assumptions and learn the parameters for this model. By tracking the status of the interaction, the robot can make better decisions about when to take turns in a separate action module (described later). To learn the model parameters, we conducted an experiment to collect a diverse set of turn-taking episodes, purposely including both good and bad examples, through a combination of teleoperated robot behavior and randomly generated timing variations. We then manually coded this data to learn about human-robot turn-taking behavior to inform our implementation of an autonomous

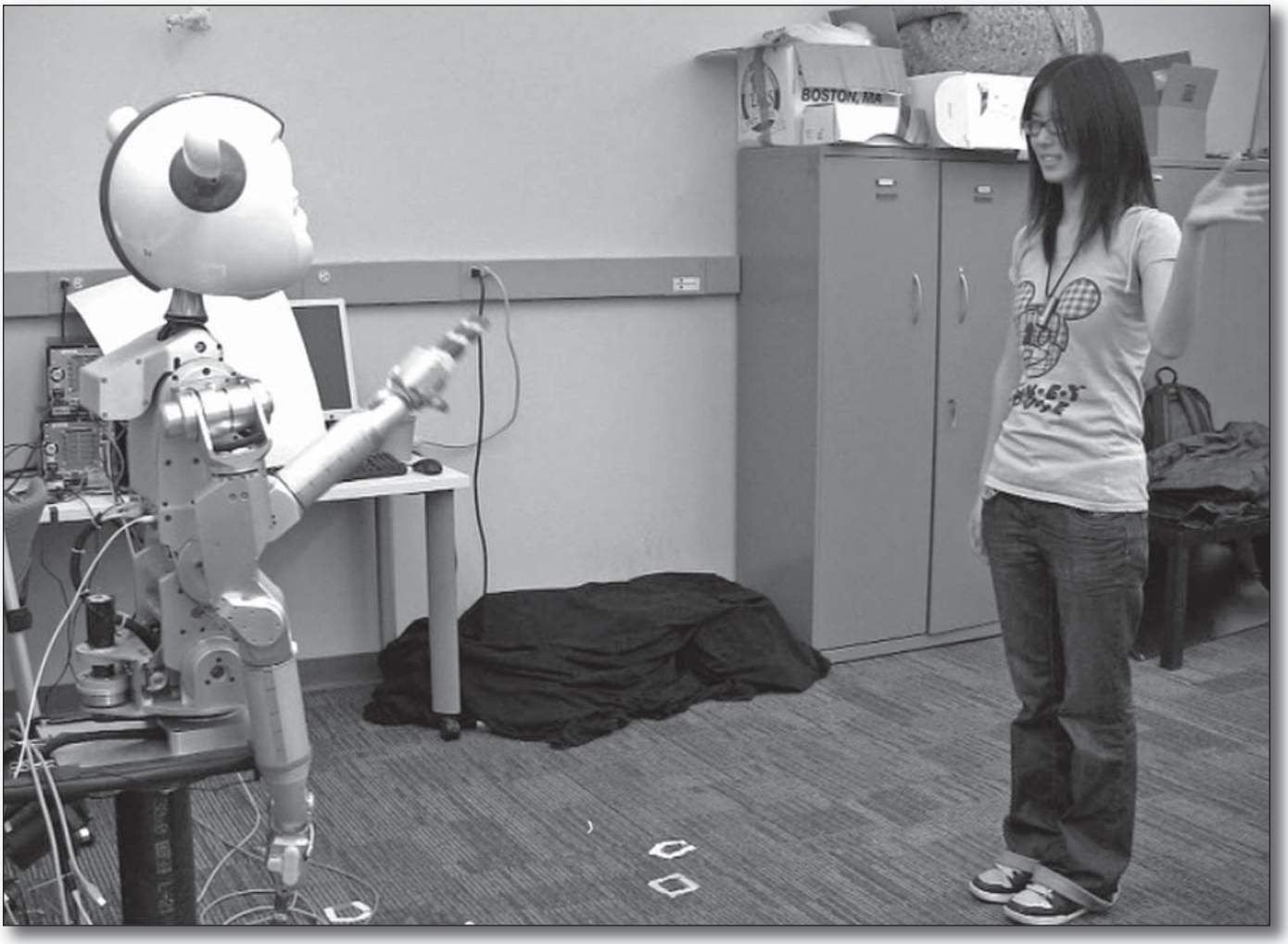

Figure 1. A Participant with the Simon Robot.

robot controller. In the remainder of this article we describe the data-collection experiment, present some results of our data analysis, and describe how this informs aspects of these model parameters and was used in a subsequent implementation to autonomously control the robot in the same scenario as our experiment.

Robot Platform

The robot used in this research is a Meka Robotics upper-torso humanoid robot we call Simon, seen in figure 2. It has two series-elastic 7-DOF arms with 4-DOF hands, and a socially expressive head and neck. The sensors used in this work are one of Simon's eye cameras, an external camera mounted on a tripod, a structured-light depth sensor (MS Kinect) mounted on a tripod, and a microphone worn around the human partner's neck.

Experiment

We ran a teleoperated data collection experiment (also known as Wizard of Oz) in which our robot played an imitation game based on the traditional children's game "Simon Says" with a human partner. We collected data from a total of 27 people.

For 4 participants there was a problem that caused data loss with at least one logging component, so our analysis includes data from 23 participants. We collected approximately 4 minutes of data from each participant.[1]

This domain has several desirable qualities for our investigation. It is multimodal, there is interactive symmetry between the human and robot, it is relatively simple, and it is isolated from such complexities as object-based joint attention. Figure 2 shows the face-to-face setup. The game has a leader and a follower role; the leader is called

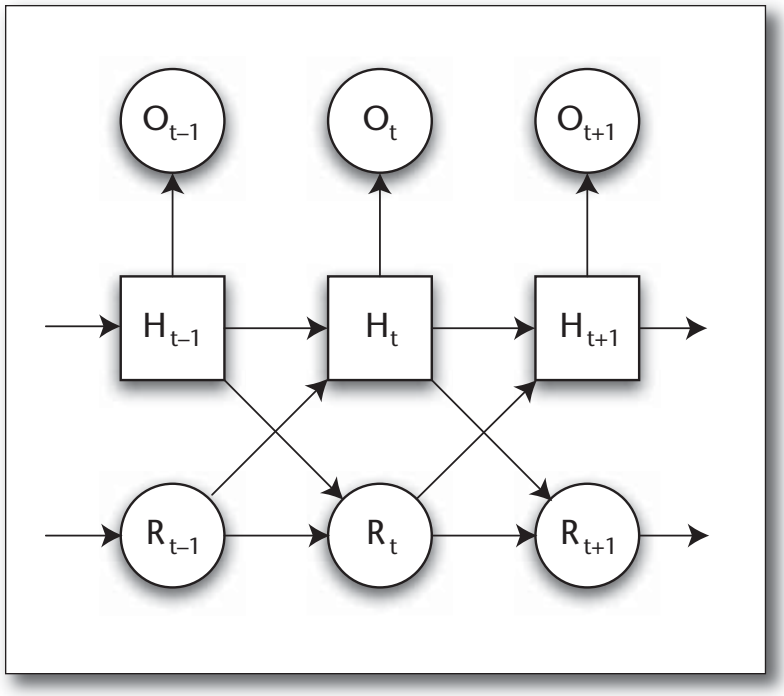

Figure 2. Our Model for Human-Robot Turn Dynamics Is a First-Order Markov Process.

At each time step both the robot (R_t) and the human (H_t) can be in one of four floor states: seizing, passing, holding, listening.

Simon. In our experiment, the human and the robot play both roles, dividing the interaction into a series of game and negotiation phases.

Game Phase. In the game phase, the leader can say, "Simon says, [perform an action]." The available actions are depicted in figure 3. The follower should then imitate that action. The leader can also say, "[Perform an action]," after which the follower should do nothing, or else he loses the game. The leader concludes the set after observing an incorrect response by declaring, "You lose!" or "I win!"

Negotiation Phase. In the negotiation phase, the follower can ask, "Can I play Simon?" or say, "I want to play Simon." The leader can then transfer the leadership role or reject the request. The leader also has the option of asking the follower, "Do you want to play Simon?" or saying to him, "You can play Simon now." The leader and follower can do this negotiation to exchange roles at any time during the game.

The robot's behavior in the game was implemented as a 15-state finite state machine (FSM) (Chao et al. 2011). For example, the *Hello State* starts the interaction and says something like "Hello, let's play Simon says," the *Simon Says State* selects an action command starting with "Simon says," and the *Bow State* performs the "bow" action as a follower. Each state includes a combination of actions from each of three channels: body motion (one of the actions shown in figure 3), eye gaze direction (either at the person or away), and speech (an utterance randomly selected from a group of 1–3 valid sentences for each state).

During the experiment one of the authors teleoperated the robot using a keyboard interface to signal which state the robot should be in. Our goal with this experiment was to collect a varied data set of both good and bad turn-taking episodes. We hypothesized that the exact timing and coordination between the three action channels would play a role in good/bad floor passing behavior on the part of the robot. It is hard for a teleoperator to generate random behavior. Thus, in order to collect wide distribution of examples, we inserted random delays before and after actions of the teleoperator, to vary the ways that these three channels lined up.

We logged data from a Kinect, an external camera, a camera in one of the robot's eyes, and a microphone as our observations of the human. We also logged the FSM states of the robot.

Data Analysis

Our analysis for this article focuses on human responses to the robot's different signals. The specific data we examine is the human *response delay*, which is the time between a referent event in the robot's behavior and the start of the human's response. This requires that we manually code the human response point for all the data. We separate the data collected from this experiment into game phase data and negotiation phase data, in order to analyze these two different types of turn-taking interactions. All events that needed to be coded (that is, were not part of the logged behavior) were coded independently by two people, and for each event that was agreed upon, the coded time was averaged. The coded events were the game phase response, negotiation phase response, and minimum necessary information.

Game Phase Response

In the game phase data, the robot plays the leader and communicates using a mixture of speech, motion, and gaze. The human plays the follower and responds primarily with a motion, which is sometimes also accompanied with a speech backchannel. For a more controlled data set, the game phase data includes only correct human responses to the robot's "Simon says" turns. The coder agreement was 100 percent for game phase response events, and the average difference in coded time was 123 milliseconds.

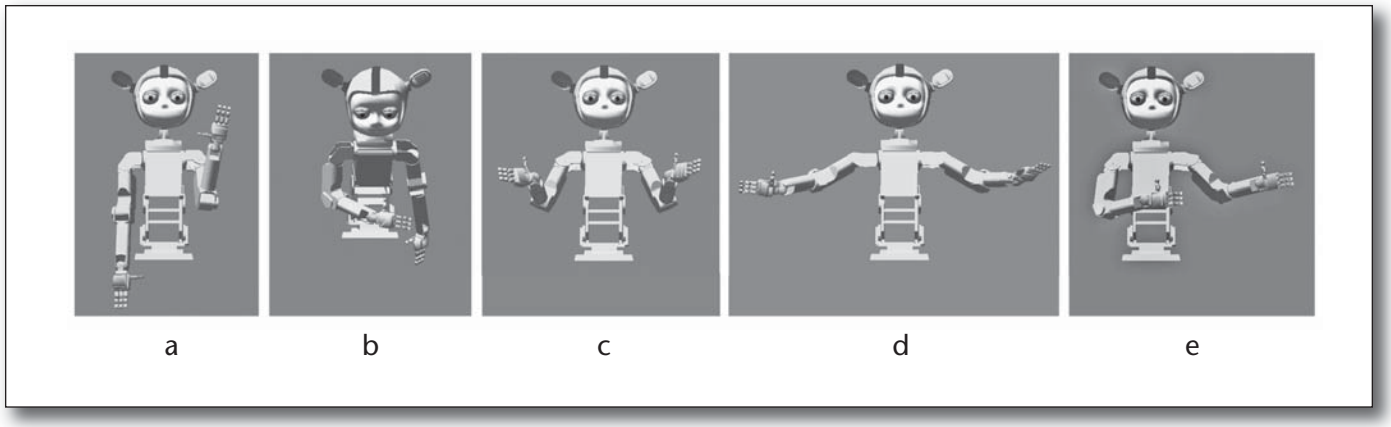

Figure 3. Actions in the "Simon Says" Game.

a. Wave. b. Bow. c. Shrug. d. Fly like a bird. e. Play air guitar.

Negotiation Phase Response

In the negotiation phase, the exchanges are shorter, and the robot uses speech but not any body animations to communicate. Most robot utterances are also too short for the robot to have time to gaze away and back to the human, so the robot primarily gazes at the human. The human's response is primarily in the speech channel. The coder agreement was 94.2 percent for negotiation phase response events, and the average difference in coded time was 368 milliseconds.

Minimum Necessary Information (MNI)

This is a robot event describing an interval during which the robot conveys the minimum amount of information needed for the human to respond in a semantically appropriate way. Figure 4 shows examples of MNI coding. In the game phase, the human needs to know whether or not to respond as well as the motion with which to respond, so the MNI is the earliest point at which both of these are conveyed. For example, if the robot says "Simon says, play air guitar," the person only needs to hear "Simon says, play" to respond appropriately. But if the robot first starts playing air guitar, then the person only needs to know whether or not to do it, which is communicated in the first syllable of the speech event. In the negotiation phase, the information is usually marked by a pronoun. The listener only needs to know whether the speaker wants to be the leader or not. The coder agreement was 99.8 percent for robot MNI events, and the average difference in coded time was 202 milliseconds.

Quantitative Results

We are interested in the timing of human turn delays, which need to be with respect to something — a referent event. As potential referent events we consider the MNI point, as well as channel-based referent events: the end of robot motion, the end of robot speech, or the moment when the robot gazes at the human after looking away.

Histograms of response delays with respect to these referent events are shown in figure 5, which separately depicts the delays for the game and negotiation phases. We see that not all of these referent event signals are good predictors of human response time. A good referent event should yield distributions that have three properties: nonnegativity, low variance, and generality. *Nonnegativity:* If the response delay is negative, then this referent event could not have caused the response. *Low Variance:* The distribution should have low variability to allow for more accurate prediction. *Generality:* The distribution should be consistent across different types of interactions (that is, we would like to see the same response delay distribution across both the game and negotiation phases).

In figure 5a we see that responses to the motion event and the gaze event both violate nonnegativity. Gaze has been demonstrated to be an excellent indicator in multiparty conversation domains (Mutlu et al. 2009, Bohus and Horvitz 2010), but it is less predictive in this particular dyadic interaction. We suspect that gaze will be a stronger predictor in a dyadic interaction where object manipulation or focusing attention on the workspace is a more prominent part of the task. The best channel-based referent event is speech, but 41 percent of human responses still occur before the robot finishes speech in the game phase. People tended to wait until the end of speech in the negotiation phase since their responses are speech-based, and there is a "lock" on the speech channel. But in the game phase they can respond in the motion channel before speech is finished.

The best referent event is the end of the MNI sig-

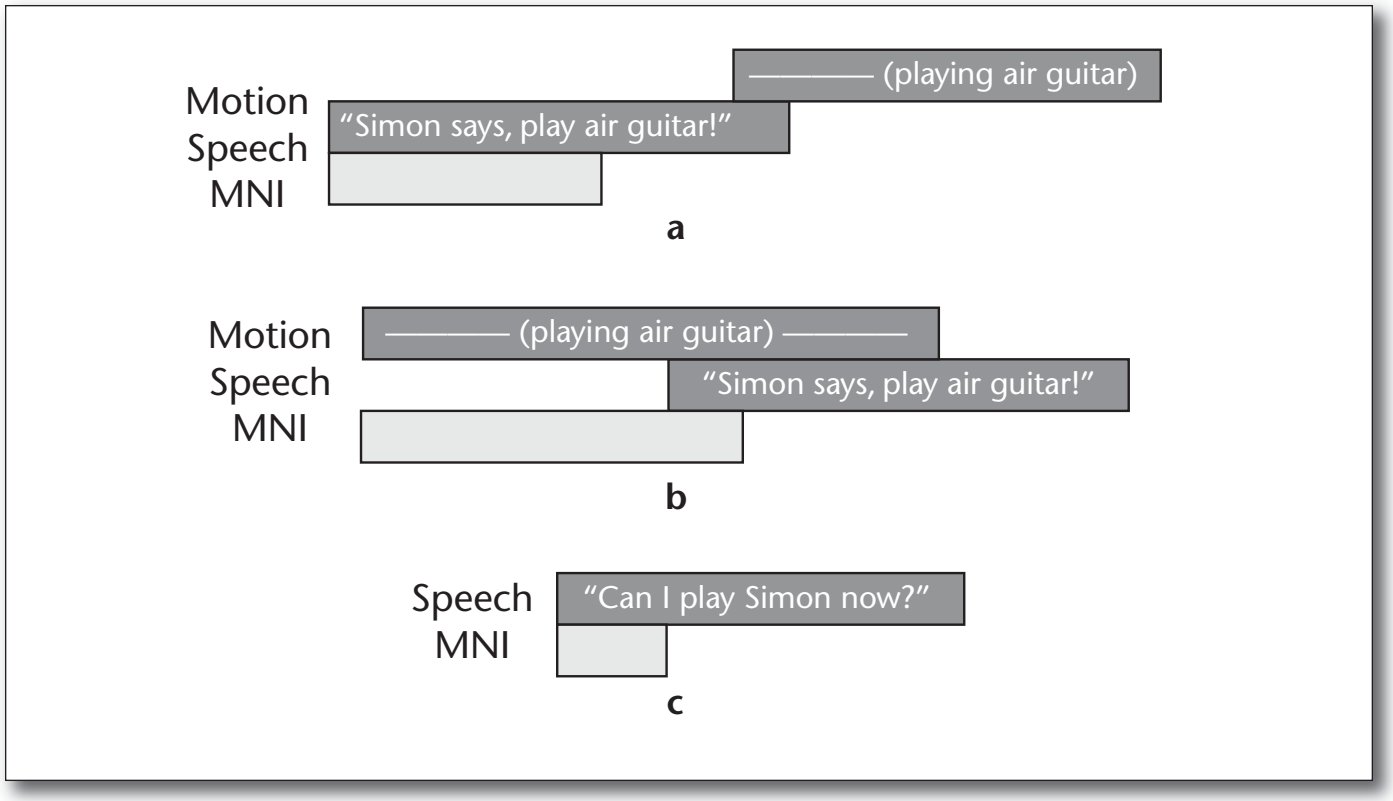

Figure 4. Examples of Coding Robot MNI in the Game Phase.

a. All informative speech occurs before the animation starts. b. The action is conveyed through motion before the human knows whether or not to execute it. c. MNI in the negotiation phase was often the pronoun.

nal. The response delay distributions to MNI endings are shown with the other distributions in figure 5 and also fit to curves in figure 5b. MNI satisfies nonnegativity for both interaction phases and is relatively general. The means in figure 5b are also within half a second from that of the distribution in (Movellan 2005). We think this could be attributed to the higher processing requirement for the multimodal information content of this game. Thus, we conclude that the timing of floor passing in a dyadic interaction like our game scenario is best predicted by information flow. We believe this will prove to be general across many interaction domains, but this generalization of results is left to future work.

Qualitative Examples

In addition to the quantitative results of response delay timing, we have several qualitative observations that illustrate how the MNI point and information flow dictate floor passing in a dyadic interaction. Given that people respond to the robot using the MNI point, we would expect awkward floor passing to result when the robot does not make use of this signal in its own floor passing behavior. This happened several times in our experiment in both the game phase and the negotiation phase.

In the game phase, this typically happened when the robot was the leader and continued to perform its gesture (such as playing air guitar) for too long after the human partner had already interpreted the gesture and completed the appropriate response turn. In many cases, subjects had to wait for the robot to finish once they were done with their response gestures. This also happened in cases where the human lost the game. We see examples where they notice the gesture the robot is doing, start doing the same gesture in response, then visibly/audibly notice they were instead supposed to remain still. But the robot still takes the time to finish its gesture before declaring the person lost. These examples illustrate the inefficiencies that the robot introduces when it does not have an appropriate floor-relinquishing behavior.

In the negotiation phase, the robot's awkward floor-relinquishing behavior results in dominance over the human partner rather than just inefficiency. As mentioned previously, in this phase the turns are primarily speech-based. Thus, simultaneous speech is a common occurrence. For example, after a pause both the robot and the human might

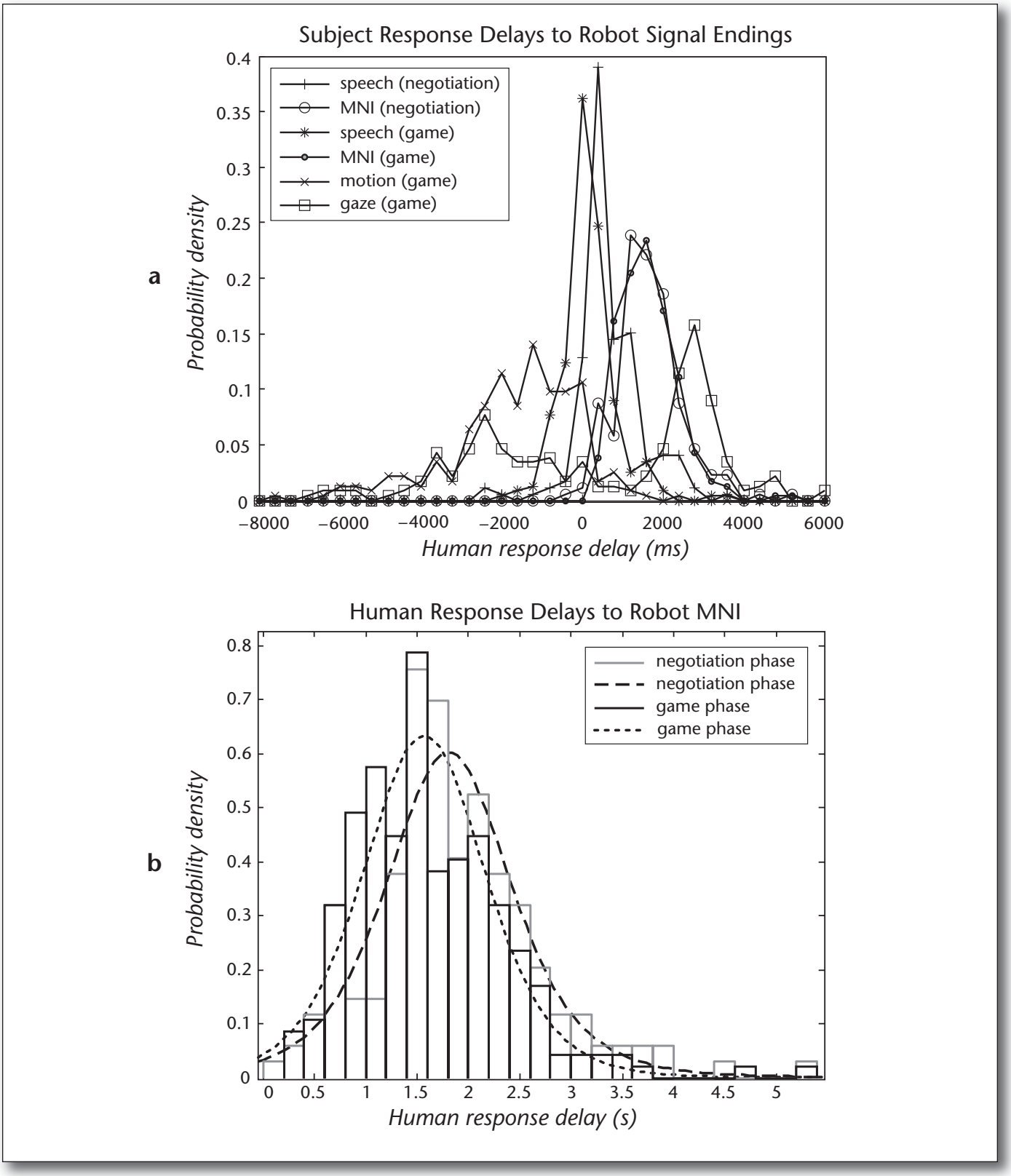

Figure 5. Human Response Delays.

(a) Histograms of human response delays with respect to all potential robot referent signals. Negative delays indicate that subjects responded before the robot completed its turn-taking action within that channel. (b) The delays of human responses with respect to robot MNI endings in the negotiation and game phases. The curves represent maximum likelihood fits to Student's *t* probability density functions.

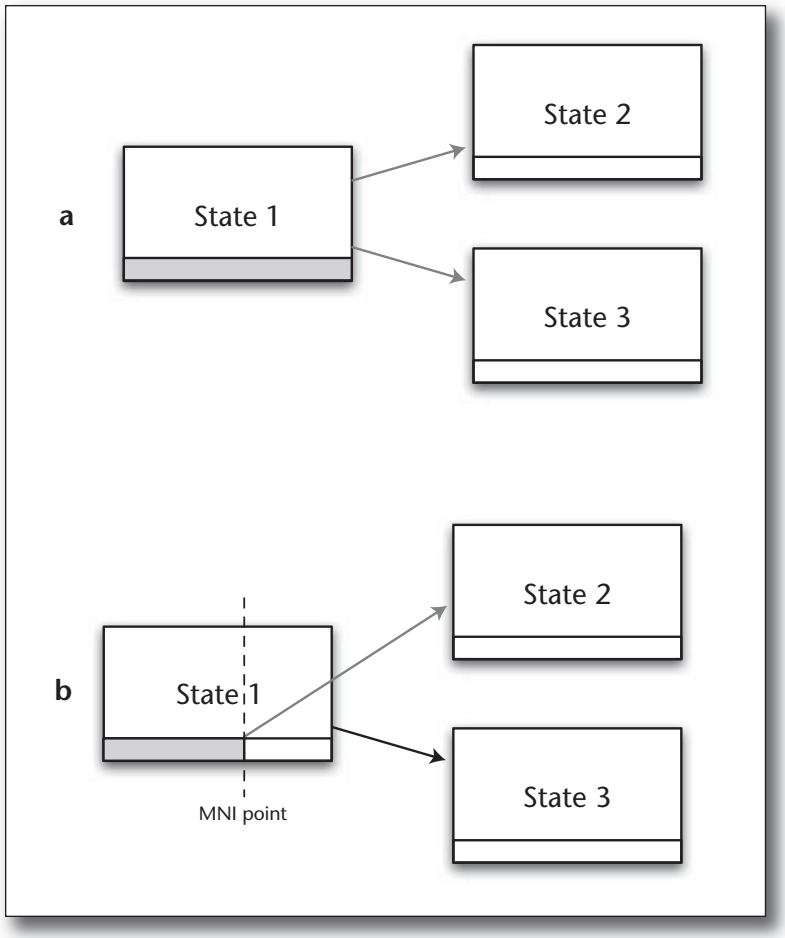

Figure 6. A Slice of an FSM.

The bars at the bottom of each state indicate the progress toward the completion of that state's actions. (a) on the left represents a typical FSM, where a state finishes executing and then evaluates to decide on a transition. (b) on the right represents our interruptible FSM implementation, to achieve floor relinquishing. A state transition has the option of evaluating prior to the completion of the previous state, and based on this can interrupt the current action to proceed to the next state without completing the first.

start asking "Can I play Simon now?" This state of both parties trying to seize the floor is typical in human communication and results in one of the parties relinquishing to the other (such as "Go ahead" or "Sorry, you were saying?"). However, in our experiment the robot's autonomous behavior was to keep going to the end of its turn, resulting in the human always being the party to back off. The few good examples of robot floor relinquishing were a result of the teleoperator acting fast enough to interrupt the robot's behavior and let the human have the floor.

Autonomous Floor Relinquishing

As illustrated with the above examples, the ability for the robot to interrupt its current action and appropriately relinquish the floor is important for achieving fluency in the interaction. In this section we describe a modification to the robot's FSM implementation that enables relinquishing. We demonstrate the results with both a gesture-based and speech-based example.

Our implementation of autonomous floor relinquishing is achieved by allowing transitions to interrupt states in the FSM. Figure 6 depicts the difference between the former state machine and the new implementation. Previously, a state had to run to completion before the transitions out of that state started being evaluated. We achieve floor relinquishing by allowing some state transitions to start being evaluated, and optionally interrupt the state, prior to the state's completion. If the transition does not fire early, however, the behavior is the same as a normal state machine. The interaction designer specifies whether or not any given state can be interrupted or any given transition can fire early.

The interruption of a state signals that the current state should terminate as soon as possible so that the next state can activate. This means that any active text-to-speech process is destroyed, and the current robot joint positions are quickly interpolated to the starting positions of the next state. The timing of an interruption should potentially also affect the robot's behavior. A robot may want to treat an interruption that occurs after the MNI point of the state as a normal part of information exchange and proceed with the domain-specific activity, whereas an interruption that occurs prior to the MNI point could indicate a simultaneous start from which the robot should back off or attempt to recover.

This FSM modification allows us to set up a "Simon Says" FSM very similar to the one we used in the experiment described earlier, but with the new feature that the robot can now have the expectation that after the MNI point in a particular speech or gesture being executed, there is the possibility that a transition to the next state will happen prior to action completion.

Figure 7 compares the timing of the FSMs with and without interrupts side by side to show how interrupts at the MNI point increase interaction efficiency. When the robot plays the leader in the "Simon Says" domain, the robot autonomously recognizes the human's game actions using the Kinect sensor and transitions as soon an action is recognizable (that is, before the human finishes her action by putting down her arms). Compared to the original behavior, where the robot's actions run to completion, the overall interaction becomes more efficient and timing becomes dictated more by the human responses with less human wait time.

Another useful application of interruptions is

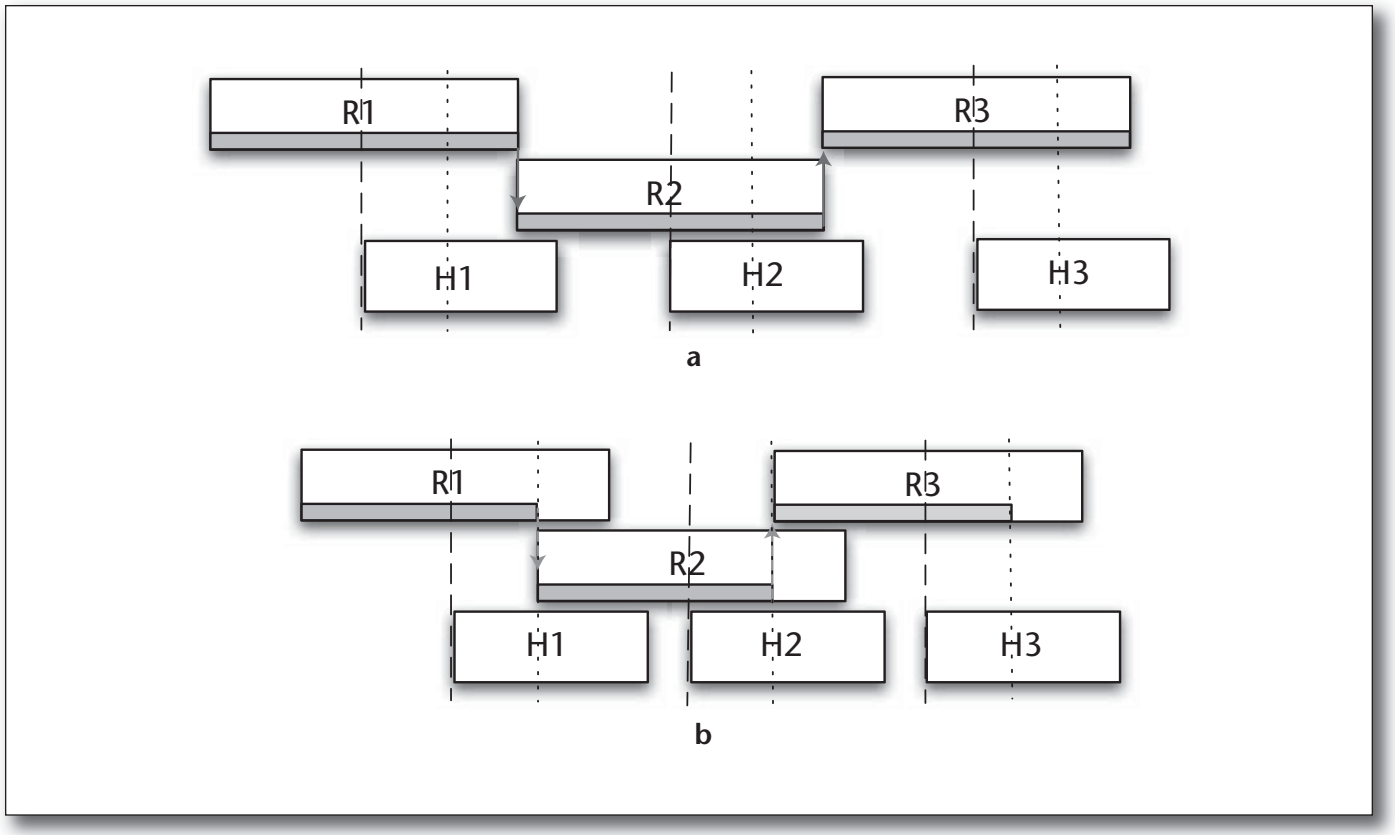

Figure 7. Turn Taking With and Without State Interruptions.

R indicates a robot turn, and H indicates a human turn. The dashed lines show robot turn MNI points and dotted lines show human turn MNI points. (a) shows a state machine without interruptions; even when the human MNI point passes, the robot continues to complete the state's actions. (b) shows how transitions that interrupt the current state can make the interaction more efficient.

resolving simultaneous starts. The robot senses the general volume level on the microphone in order to do incremental recognition of the human's speech act (that is, prior to recognizing the full sentence from the grammar, the robot can recognize from the mic level that the human is in the process of speaking). If the human asks the robot, "Can I be Simon now?" at the same time that the robot says something, the autonomous controller interrupts the robot's current speech to gain a clearer signal of the human's speech for speech recognition. Compared to the original behavior, the robot is now able to allow the human to barge in rather than always requiring that the human back off during simultaneous speech.[2]

Insights on Turn Taking for HRI

Our goal is to create a transferable domain-independent turn-taking module. Our initial work has taught us that a primary component of this module is the definition of how it should interact with elements of the specific context or domain. Thus, while we believe there are generic elements to an HRI turn-taking behavior, they are tightly coupled to the domain or context in which they are instantiated. Figure 8 shows our current concept of an architecture for turn taking.

This architecture focuses on the specific channels of gaze, speech, and motion, which are independently well studied in HRI. Actions in these channels are parametrized, such that specific parameters can be decided by either the domain-specific Instrumental Module or the generic Turn-Taking Module in order to generate the final behavior. The separation between the Instrumental Module and Turn-Taking Module highlights the principle dichotomy between domain-specific robot capabilities and context-free interaction behavior. In reality, the boundary between the two is not so pronounced. The turn-taking model needs to give floor state estimation, which drives the domain-specific FSM, but that FSM also needs to tell the turn-taking model about the flow of information in each context, which will usually be highly semantic-based information. Then collectively they contribute parameters for robot actions.

Going forward we are considering turn taking as

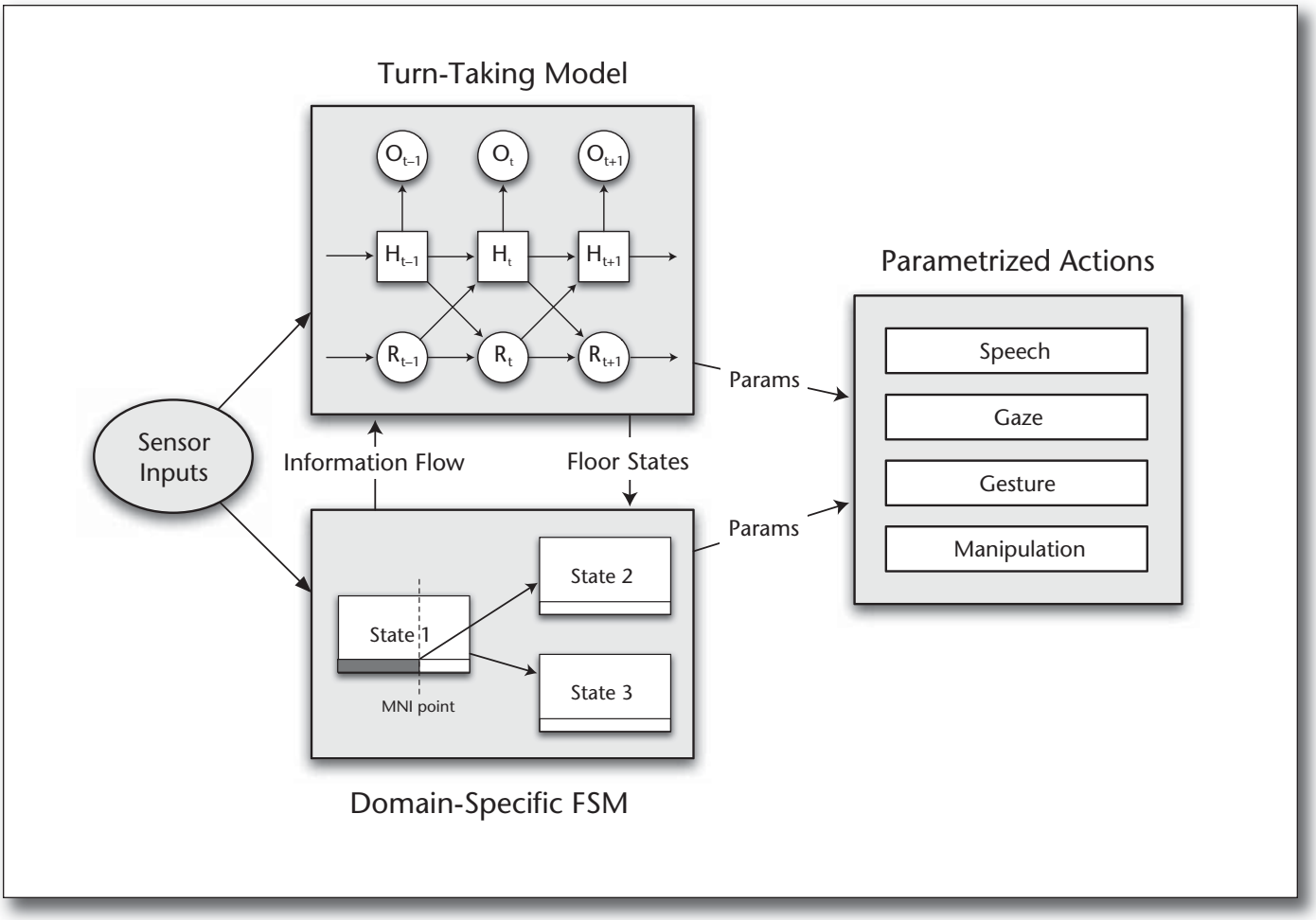

Figure 8. A Framework for Turn Taking In HRI.

The turn-taking model tracks floor state estimation, which drives the domain-specific FSM. The FSM provides the turn-taking model feedback about the flow of information in the domain. They collectively contribute parameters for robot actions.

a template that needs to be instantiated in specific domains. In the future, we intend to turn our attention to generality of this template, analyzing channel usage across other domains such as teaching-learning interactions or collaborations involving object manipulations. Perhaps the most important aspect of our model of turn taking for HRI is its multimodal nature. As was evident in our "Simon Says" example, turn taking is often more than just speech acts and dialogue. Gestures and nonverbal behavior are an essential part of passing the floor back and forth between two interacting partners. We expect that the role that speech versus gesture versus eye gaze plays in turn taking will vary considerably across the range of interaction domains in HRI. Some domains may be more like the negotiation phase of the "Simon Says" game, dominated by speech. In such a domain, overlapping of turns is less desirable because simultaneous speech interferes with the processing of speech content (in both humans and robots). Other domains may be more like the game phase (for example, a dyad collaborating on an object-based assembly task, or other colocated physical tasks). In these cases, gesture and nonverbal actions play a large role, and some turns may not even have a speech component at all. This can lead to more overlapping of turns and requires our floor state models for HRI to consider the timing of these nonverbal channels when tracking the state of the interaction.

Thus far, we have focused on the specific "Simon Says" domain and presented analyses that lead us closer to this goal of a general approach to turn taking for HRI. In particular, the domain drew important insights about information flow and the timing of floor state transitions. The minimum necessary information point is the earliest point at which the robot can assume that the human may want to take the floor and end the robot's turn. In

addition to this (which we have not modeled here), in a more complicated dialogue or interaction, the robot may need to have expectations about other kinds of barge-ins, such as clarifications or general back-channel comments. Unlike our "Simon Says" domain, with its relatively simple and short turns, these will be more relevant in dialogue-intensive interactions, or domains where the length of each partner's turn is longer.

Conclusions

Our goal is to devise a turn-taking framework for HRI that, like the human skill, represents something fundamental about interaction, generic to context or domain. The "Simon Says" experiment generated data about human-robot turn-taking dyads, and allows us to draw the conclusion that *minimum necessary information* is a significant factor in the human partner's response delay. Moreover, the robot's poor floor relinquishing behavior was often characterized by a lack of attention to the MNI point. Our autonomous floor relinquishing implementation centers on this and results in a more fluent "Simon Says" interaction. Finally, we have presented a framework for turn taking in HRI, highlighting its multimodal nature and the tightly coupled interplay between generic interaction principles and specific domain semantics.

Acknowledgments

This work was supported by ONR Young Investigator Award #N000140810842. The authors wish to thank Jinhan Lee and Momotaz Begum for their contributions to the "Simon Says" data collection.

Notes

1. See Chao et al. (2011) for additional details and analysis on the "Simon Says" experiment that are not presented here.

2. Video demonstrations of both of these examples can be found at www.cc.gatech.edu/social-machines/video/interrupt-gesture.mov and www.cc.gatech.edu/social-machines/video/interrupt-speech.mov.

References

Billard, A. 2002. Imitation: A Means to Enhance Learning of a Synthetic Proto-Language in an Autonomous Robot. In *Imitation in Animals and Artifacts*, ed. K. Dautenhahn and C. L. Nehaniv. Cambridge, MA: The MIT Press.

Bohus, D., and Horvitz, E. 2010. Facilitating Multiparty Dialog with Gaze, Gesture, and Speech. In *Proceedings of the 12th International Conference on Multimodal Interfaces*. New York: Association for Computing Machinery.

Breazeal, C.; Brooks, A.; Gray, J.; Hoffman, G.; Lieberman, J.; Lee, H.; Thomaz, A. L.; and Mulanda, D. 2004. Tutelage and Collaboration for Humanoid Robots. *International Journal of Humanoid Robotics* 1(2): 315–348.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge, MA: The MIT Press.

Cassell, J., and Thorisson, K. R. 1999. The Power of a Nod and a Glance: Envelope Versus Emotional Feedback In Animated Conversational Agents. *Applied Artificial Intelligence* 13(4–5): 519–538.

Chao, C.; Lee, J.; Begum, M.; and Thomaz, A. L. 2011. Simon Plays Simon Says: The Timing of Turn-Taking in an Imitation Game. In *Proceedings of the Twentieth IEEE International Symposium on Robot and Human Interactive Communication*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Duncan, S. 1974. On the Structure of Speaker-Auditor Interaction During Speaking Turns. *Language in Society* 3(2): 161–180.

Kaye, K. 1977. Infants Effects upon Their Mothers Teaching Strategies. In *The Social Context of Learning and Development*, ed. J. Glidewell. New York: Gardner Press.

Kose-Bagci, H.; Dautenhan, K.; and Nehaniv, C. L. 2008. Emergent Dynamics of Turn-Taking Interaction in Drumming Games with a Humanoid Robot. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, 346–353. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Lee, J.; Kiser, J. F.; Bobick, A. F.; and Thomaz, A. L. 2011. Vision-Based Contingency Detection. In *Proceedings of the 6th International Conference on Human Robot Interaction*. New York: Association for Computing Machinery.

Michalowski, M.; Sabanovic, S.; and Kozima, H. 2007. A Dancing Robot for Rhythmic Social Interaction. In *Proceedings of the Second ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. New York: Association for Computing Machinery.

Movellan, J. R. 2005. Infomax Control as a Model of Real Time Behavior. MPLab Tech Report 1. Machine Perception Laboratory, University of California, San Diego, La Jolla, CA.

Mutlu, B.; Shiwa, T.; Ishiguro, T. K. H.; and Hagita, N. 2009. Footing in Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. New York: Association for Computing Machinery.

Nicolescu, M. N., and Mataríc, M. J. 2003. Natural Methods for Robot Task Learning: Instructive Demonstrations, Generalization and Practice. Paper presented at the Second International Joint Conference on Autonomous Agents and Multiagent Systems, July 14–18, Melbourne, Victoria, Australia.

Orestrom, B. 1983. *Turn-Taking In English Conversation*. Lund, Sweden: CWK Gleerup.

Raux, A., and Eskenazi, M. 2009. A Finite-State Turn-Taking Model for Spoken Dialog Systems. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Rich, C.; Ponsler, B.; Holroyd, A.; and Sidner, C. L. 2010. Recognizing Engagement in Human-Robot Interaction. In *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction*. New York: Association for Computing Machinery.

Schegloff, E. 2000. Overlapping Talk and the Organization of Turn-Taking dor Conversation. *Language in Society* 29(1): 1–63.

Weinberg, G., and Driscoll, S. 2007. The Interactive Robotic Percussionist: New Developments in Form, Mechanics, Perception, and Interaction Design. In *Proceedings of the Second ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 97–104. New York: Association for Computing Machinery.

Andrea L. Thomaz is an assistant professor of interactive computing at the Georgia Institute of Technology. She directs the Socially Intelligent Machines lab, which is affiliated with the Robotics and Intelligent Machines Center and with the Graphics Visualization and Usability Center. She earned Sc.M. (2002) and Ph.D. (2006) degrees from the Massachusetts Institute of Technology. She has published in the areas of artificial intelligence, robotics, human-robot interaction, and human-computer interaction.

Crystal Chao received her B.S. in computer science from the Massachusetts Institute of Technology in 2008. She is currently a Ph.D. student in robotics at the Georgia Institute of Technology. Her research interests include interactive learning, human-robot dialogue, and turn taking in multimodal interactions.