2011 RO-MAN
20th IEEE International Symposium on Robot and
Human Interactive Communication,
July 31 - August 3, 2011, Atlanta, GA, USA

Simon plays Simon says: The timing of turn-taking in an imitation game

Crystal Chao, Jinhan Lee, Momotaz Begum, and Andrea L. Thomaz

Abstract— **Turn-taking is fundamental to the way humans engage in information exchange, but robots currently lack the turn-taking skills required for natural communication. In order to bring effective turn-taking to robots, we must first understand the underlying processes in the context of what is possible to implement. We describe a data collection experiment with an interaction format inspired by "Simon says," a turn-taking imitation game that engages the channels of gaze, speech, and motion. We analyze data from 23 human subjects interacting with a humanoid social robot and propose the principle of *minimum necessary information (MNI)* as a factor in determining the timing of the human response. We also describe the other observed phenomena of channel exclusion, efficiency, and adaptation. We discuss the implications of these principles and propose some ways to incorporate our findings into a computational model of turn-taking.**

I. Introduction and Related Work

Turn-taking is the fundamental way that humans organize interactions with each other. The idea that turn-taking is a deeply rooted human behavior is supported by research in developmental psychology. Studies of mother-infant dyads have defined interaction patterns in dyadic phases [1] and described rhythmic cyclicity and mutual regulation [2].

Extensive treatment of turn-taking can be found in the linguistics literature as well. Some work focuses on the structure of syntax and semantics in language usage [3], and other work additionally analyzes the contribution of other signals used by humans such as paralinguistic cues, gaze shift, and gesticulation [4], [5].

We believe that socially embedded machines that employ the same turn-taking principles will be more intuitive for humans to interact with. Implementations of turn-taking components come from many different approaches. Turn-taking is a highly multimodal process, and prior work gives much in-depth analysis of specific channels, such as gaze usage to designate speaker or listener roles [6] or speech strategies in spoken dialog systems [7]. Closely related is the problem of contingency or engagement detection, which requires implementing robot perception for awareness of the human's cue usage [8], [9], [10]. Turn-taking has also been demonstrated in situated agents [11], including management of multi-party conversation [12].

Eventually, some unifying efforts will be required to integrate the piecewise work into an architecture for physically embodied robots. Although it may be convenient for a roboticist to appeal to the argument that good turn-taking emerges automatically from reactive behavior [13], the reality is that turn-taking interactions with humans also involve constant overlapping, misunderstanding, and recovery [14]. An architecture designed for turn-taking should be able to handle these situations, both perceptually and behaviorally.

Fig. 1. A participant plays "Simon says" with the Simon robot.

In this paper, we present an experiment designed to inform a model of turn-taking behavior that can be implemented on a social robot. This model includes timing parameters that are measured relative to a variety of interaction signals. We note that such signals cannot be defined without developing a theory about the information conveyed by certain communication acts between the robot and the human. We develop such a characterization that permits us to effectively estimate the relevant timing parameters.

II. Approach

To provide context for our experiment, we describe our approach to achieving our longer-term research goal of natural, human-like turn-taking in robots. We intend to build a computational model of turn-taking interchanges that defines the relevant states of both the human and the robot, as well as how belief about those states should drive the behavior of the robot. Included in the model is a specification of the behavioral signals that indicate the state of the human partner. To leverage this model, we need to determine what signals are most indicative of important events in turn-taking, how feasible they are to perceive, and how they vary or stay constant across different types of domains.

A. Computational Model of Turn-Taking

The foundation of our computational approach to turn-taking is uncertainty modeling. Uncertainty in estimating when one or one's partner should speak is already an issue in human-human turn-taking; even with excellent perceptual capabilities, humans are still prone to unintended interruptions, simultaneous starts, and awkward silences [14]. On a robot, the problem of uncertainty is heightened even further by noisy and limited sensory data. To deal with uncertainty over

This work was supported by ONR Young Investigator Award #N000140810842.

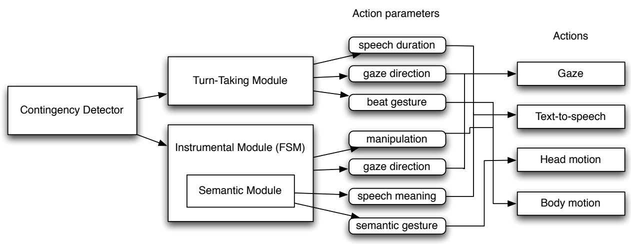


Fig. 2. A proposed architecture for turn-taking. Parameters are specified by a context-free Turn-Taking Module and context-dependent Instrumental Module. Parameters from both modules are used to instantiate robot actions.

time, we intend to use a probabilistic model that comprises a timing model and an observation model:

1) Timing Model: The timing model describes the fundamental timing or structure of when people naturally take turns. The robot can use the timing model to determine if a person is being contingent or engaged, as well as decide when it might want to take a turn in order to avoid a collision. When perception is ambiguous, timing provides a feed-forward signal for the robot to keep on interacting.

2) Observation Model: The observation model describes the robot perception required to determine when people are about to seize or pass the floor, or when they are acting engaged. The observations form a feedback signal that keeps the overall model updated and allows the robot to understand what is currently transpiring in the interaction.

Similarly to [10] and [12], our approach in this paper is to analyze interaction data in order to find general assumptions that can be used to construct such a model. We thus conduct an experiment in which we collect a diverse selection of turn-taking episodes, both good and bad, through a combination of teleoperation and randomly generated timing variations. We then hand-code this data to learn about human-robot turn-taking behavior that can later be executed autonomously.

B. Architecture for Turn-Taking

Figure 2 shows our current concept of an architecture for turn-taking. The 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, but we hope to extract as much domain-independent turn-taking behavior as possible in order to create a transferable module. In the future, we intend to analyze channel usage across multiple domains, such as teaching-learning interactions or collaborations involving object manipulations. In this paper, we focus on turn-taking in the specific domain of a "Simon says" game and present some analyses that will lead us closer to this goal.

III. Experiment

We ran a teleoperated data collection experiment in which our robot plays "Simon says" with a human partner. The game is attractive as an initial domain of investigation for its multimodality, interactive symmetry, and relative simplicity, being isolated from such complexities as object-based joint attention. We collected data from a total of 27 human subjects. For 4 subjects there was a problem that caused data loss with at least one logging component, so our analysis includes data from 23 subjects. We collected approximately 4 minutes of data from each participant.

A. Platform

The robot used is an upper-torso humanoid robot, Simon. It has two series-elastic 7-DOF arms with 4-DOF hands, and a socially expressive head and neck. The sensors recorded were one of Simon's eye cameras, an external camera mounted on a tripod, a structured light depth sensor ("Kinect") mounted on a tripod, and a microphone worn around the participant's neck. The computers used for logging data were synchronized to the same time server.

B. "Simon Says" Domain Description

The domain is an imitation game based on the traditional children's game "Simon says." Figure 1 shows the face-to-face setup. The game has a leading and a following role; the leader is referred to as "Simon." We divide the interaction into a game phase and a negotiation phase.

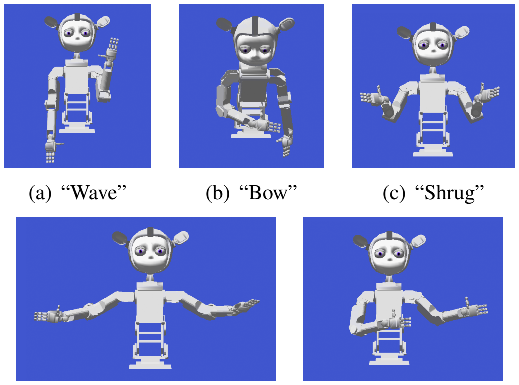
In the game phase, the leader can say, "Simon says, [perform an action]." The available actions are depicted in Figure 3. The follower should then imitate that action. The leader can also say, "[Perform an action]," after which the follower should do nothing, or else he loses the game. The leader concludes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

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 exchange roles at any time.

C. Robot Behavior

All of the robot's behavior is organized into states in a finite state machine (FSM). The 15 states available to the teleoperator are described in Table I. Each state in the FSM controls the robot's three channels of communication:

- *Body animation* – the actions of the game as shown in Figure 3. The speed of the animation was selected uniformly at random from a safe range.
- *Speech content* – an utterance randomly selected from the group of valid sentences for the state. Each state had 1-3 sentences as options.
- *Gaze direction* – gazing at the person's face using a visual servoing mechanism with the eye camera, or gazing away from the person.



(a) "Wave" (b) "Bow" (c) "Shrug"

(d) "Fly like a bird" (e) "Play air guitar"

Fig. 3. Actions in the "Simon says" game.

TABLE I

FSM STATES AVAILABLE TO TELEOPERATOR.

State	Description
Hello	Start the interaction ("Hello, let's play Simon says").
Bye	End the interaction ("Thanks, that was fun").
Request	Request to play Simon ("Can I play Simon now?").
Accept	Accept request ("That's fine with me").
Deny	Deny request ("No, not yet").
Simon says	Select an action command starting with "Simon says."
Do this	Select an action command.
Win	Conclude the set by winning ("Ha ha, I win").
Lose	Admit to losing ("Oh no, I guess you win").
Can't do	Say "I can't do that."
Bow	Perform "bow" action as a follower.
Bird	Perform "bird" action as a follower.
Guitar	Perform "air guitar" action as a follower.
Shrug	Perform "shrug" action as a follower.
Wave	Perform "wave" action as a follower.

Random delays were sometimes inserted before each channel, to increase variation in the robot's executed behavior.

One of the authors teleoperated the robot using a keyboard interface to select specific FSM states. The teleoperator additionally had the option of interrupting the current state, for a total of 16 keys. All of the keybinds were on one side of the keyboard to reduce the contribution of the keypress interface to the timing of the interaction.

D. Protocol

Participants were provided an explanation of the game and the available actions. They were not told that the robot was being teleoperated. The participants were told to adhere to a set of keywords when speaking to the robot. They were then given about a minute of practice with the robot to familiarize themselves with the interaction and memorize the five actions. During this time they were allowed to ask clarifying questions to the experimenters. After the practice session, data collection commenced, and they were told to avoid interacting with the experimenters.

After the data collection was complete, subjects completed a survey about their experiences. The questions were similar to those in [11]. We will be using these survey responses as a baseline for evaluating future implementations of autonomous turn-taking controllers.

IV. Results and Analysis

Because our goal here is to understand human timing in turn-taking, our analysis focuses on human responses to the robot's different signals. We ask the questions: Which signal is the most reliable predictor of human timing? What is the timing model and distribution? This informs how a robot should shape its expectations about the timing of human responses, as well as emulate these parameters in order to produce human-like behavior. In this section, we present an analysis of experiment results about several components that contribute to the manifested timing of turn-taking.

A. Data Coding

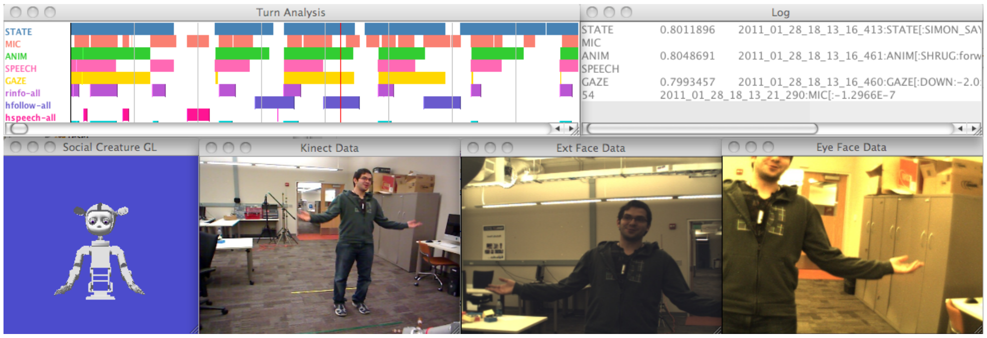
Figure 4 shows our interface for visualizing and coding the data. The data from the depth sensor, two cameras, and microphone can be played back in a synchronized fashion alongside an OpenGL visualization of the robot's joint angles and a live update of the text log of the robot behavior. The coders can scrub through the data and visually assess how their coded events align with other events in the data.

The specific data we examine is the human *response delay*, which is the time between a referent event and the start of the coded human response. We separate the data collected from this experiment into game phase data and negotiation phase data, which show two different types of turn-taking interactions. All events that needed to be coded (i.e. were not part of the logged robot behavior) were coded independently by two of the authors, and for each event that was agreed upon, the coded time was averaged. The events were:

1) Game phase response: a human event. 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 secondarily accompanied by 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% for game phase events, and the average difference in coded time was 123 milliseconds.

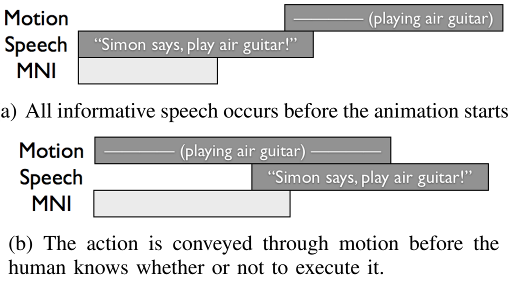
2) Negotiation phase response: a human event. In the negotiation phase, the exchanges are shorter, and the robot uses speech but not any body animations to communicate. Most robot utterances are also too short for the robot to have time to gaze away and back to the human, so the robot primarily gazes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

3) Minimum necessary information (MNI): a robot signal. This describes an interval during which the robot conveys the minimum amount of information needed for the human to respond in a semantically appropriate way. More explanation and reasoning for this signal is provided next in Section IV-B. Figures 5 and 6 show examples of MNI video coding. In the game phase, the human needs to know whether or not to respond as well the motion with which to respond,



```
STATE     0.8011896     2011_01_28_18_13_16_413:STATE[:SIMON_SAY
MIC
ANIM      0.8048691     2011_01_28_18_13_16_461:ANIM[:SHRUG:forw
SPEECH
GAZE      0.7993457     2011_01_28_18_13_16_460:GAZE[:DOWN:-2.0:
54        2011_01_28_18_13_21_290:MIC[:-1.2966E-7
```

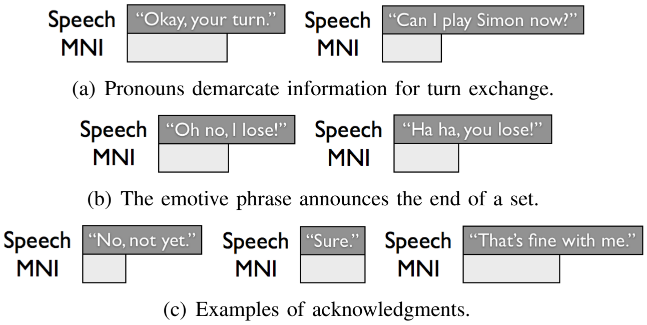
Fig. 4. Interface for visualizing and video-coding the collected data.



(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.

Fig. 5. Examples of coding robot MNI in the game phase.



(a) Pronouns demarcate information for turn exchange.

(b) The emotive phrase announces the end of a set.

(c) Examples of acknowledgments.

Fig. 6. Examples of coding robot MNI in the negotiation phase.

so the information end is the earliest point at which both of these are conveyed. In the negotiation phase, the information is usually marked by a pronoun. The coder agreement was 99.8% for robot MNI events, and the average difference in coded time was 202 milliseconds.

B. Minimum Necessary Information (MNI)

In order to characterize a predictive human response delay distribution, one needs to determine a reliable referent event. For example, some channel-based referent events are: 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 7 for both interaction phases. It becomes immediately apparent that not all of these signals are useful predictors. Specifically, a good referent event should yield distributions that have these properties:

1) *Nonnegativity* – If the response delay is negative, then this referent event cannot be the cause of the response.
2) *Low variance* – The distribution should have low variability to allow for more accurate prediction.
3) *Generality* – The distribution should be consistent across different types of interactions.

Responses to the motion event and the gaze event both violate nonnegativity (Figure 7). Gaze has been demonstrated to be an excellent indicator in multiparty conversation domains [6], [12], but it is less predictive in this particular dyadic interaction; we suspect that it might show greater impact in a dyadic object manipulation task. The best channel-based referent event is speech, but 41% of human responses still occur before the robot finishes speech in the game phase.

We thus argue for a concept called *minimum necessary information (MNI)* — the minimum amount of information needed to be conveyed by the robot for the human to respond in a semantically appropriate way (that is, discounting barge-ins or simultaneous starts). The best referent event to use is the end of the MNI signal. The response delay distributions to MNI endings are shown superimposed with the other distributions in Figure 7 and also fit to curves in Figure 8. MNI satisfies nonnegativity for both interaction phases and is relatively general. The means in Figure 8 are also within half a second from that of the distribution in [9]. We think this could be attributed to the higher processing requirement for the multimodal information content of this game.

C. Channel Exclusion

We also hypothesize that human turn-taking follows conventions for managing exclusions per channel. We observed that although subjects did not wait for the robot to finish speaking before they moved, they usually waited for the robot to finish speaking before they spoke. This accounted for the differences in the distributions of response delays to speech shown in Figure 7. For responses to speech, the negotiation phase distributions were shifted in the positive direction as compared to the game phase distributions.

Additionally, we observed that people tended to avoid simultaneous speaking after a simultaneous start. There were 23 instances of simultaneous speech in the data set, spread across 10 subjects. Of these, 7 (30%) constituted backchannel feedback. The remaining 16 instances were simultaneous starts. Of the simultaneous starts, 3 resulted in the teleoperator interrupting the robot speech, 8 resulted in the human

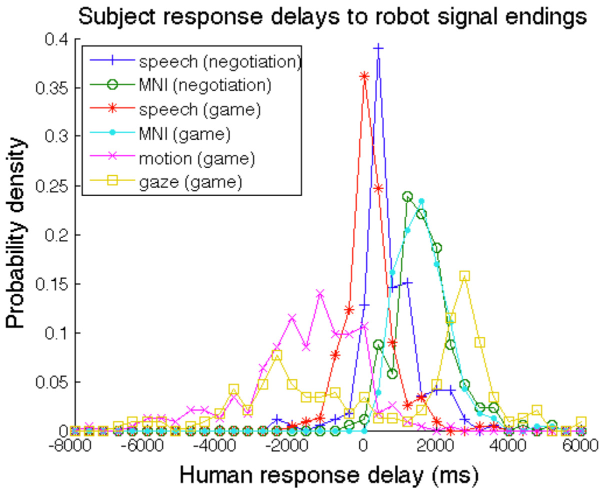


Fig. 7. Histograms of human response delays with respect to all potential robot referent signals. Negative delays indicate that subjects responded before the robot completed its turn-taking action within that channel.

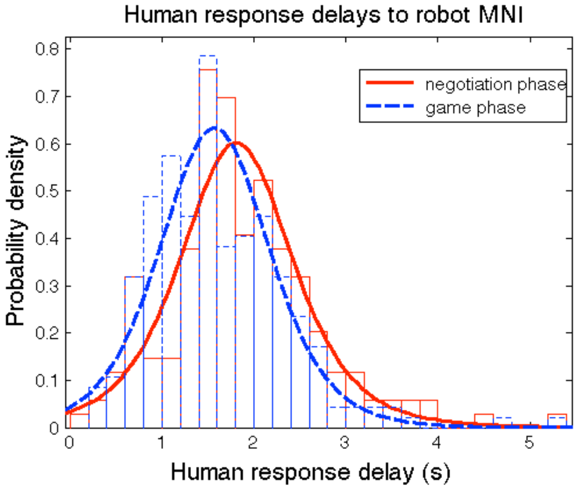


Fig. 8. The delays of human responses with respect to robot MNI endings in the negotiation and game phases. The curves represent maximum likelihood fits to Student's t probability density functions.

interrupting his own speech, and 3 resulted in a decrease in the human's speech volume. Although this is sparse data, this tendency to back off from simultaneous starts shows an adherence to channel exclusion.

This channel exclusion also has an effect on the response delay distributions to MNI. Compared to the game phase distribution, the negotiation phase distribution is slightly delayed due to this lock. However, the MNI is still relatively robust overall because the robot's speech contained a balance of shorter and longer utterances.

This domain had only one channel with a "lock," which was speech. One could envision a domain where there were exclusions in the motion channel. Both parties could need to move in the same space or need to use the same tool. These factors could lead to delayed responses. In addition, more or fewer exclusions in any channel could arise due to differences in cultural communication or personality.

D. *Efficiency vs. Adaptation*

Turn-taking is a dynamic process, and timing can evolve as the interaction progresses. If we believe that MNI endings are stable referent events, we can use response delays to them to investigate how human responses change over time.

One phenomenon we observed in the data was the notion of increasing efficiency or fluency, as described extensively in [15]. We can characterize a response's efficiency as the inverse of the response delay after the MNI end — the lower the response delay, the higher the efficiency. For some subjects, their time to react decreased with practice, as less information was needed from the robot to react, and the response delays showed a downward trend. An example is shown in Figure 9(a). Nine subjects (39%) exhibited this trend in their data.

Although this interaction was too short to see a significant difference, we think that a robot can expect this change in any domain involving repetitive behavior that leads to improvement in performance. Leiser observed in [16] that repeated information exchanges between humans cause abbreviations in language due to decreasing information requirements, which suggests that responses would approach MNI endings with repetition. A well-practiced human-robot dyad may operate at a periodicity close to the MNI point, with plenty of overlapping in any channel that did not present an exclusion.

We hypothesize that there is also another phenomenon of adaptation, where one party can adapt to and gradually approach the other party's timing. We observed that certain subjects started to imitate the robot's mannerisms of speech and motion and actually slowed down their timing to be more similar to the robot's. An example is shown in Figure 9(b). Seven subjects (30%) showed this trend. With a robot behavior control system that was sensitive to turn-taking timing, this could occur in both directions, with both parties converging on a timing between their prior distributions.

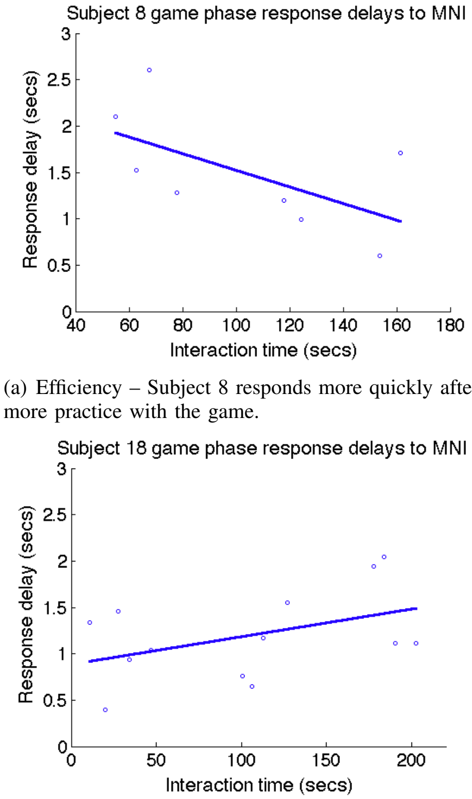
V. Discussion and Future Work

As previously described in Section II, our current computational model considers turn-taking in terms of a timing model and an observation model. The data from this experiment informs the directions of our future work to implement these models for enabling effective turn-taking on a robot.

With respect to the timing model, we observed in this experiment that timing can vary greatly across subjects. However, it may be useful to start with an informative prior such as that in Figure 8 and update the model as the interaction progresses, factoring in such effects as adaptation and increased efficiency with practice.

We intend to train an observation model from the sensor data based on video-coded labelings of turn-taking events. Some of the indicators are very subtle, such as eye gaze shifts, aspirations, and mouth openings, which have been described previously [4], [5]. These are plainly observable to the video coders, but it remains an open question whether the robot can feasibly recognize these autonomously.

Both the timing and observation models depend on the concept of MNI. Determining the MNI may not be easy for



(a) Efficiency – Subject 8 responds more quickly after more practice with the game.

(b) Adaptation – Subject 18 responds more slowly over time, adapting to the robot's behavior.

Fig. 9. Changes in interaction timing.

less structured domains. For a "Simon says" game with only a handful of actions and utterances, it can be pre-coded easily for autonomous behavior, but this coding may be too arduous for other tasks. It is possible that it can be learned on a per-action basis if the timing model is already known. We plan to verify the timing with experiments in other domains.

We do think that the principle of MNI is a useful way of understanding the process, even in cases when it is not a convenient signal to come by. It makes some recommendations for robot behavior. As an example, humans use "uh" and "um" followed by a pause as a strategy for seizing and holding turns when they are uncertain about what they are going to say [17]; the "uh" demarcates their acceptance of the turn and engagement in the interaction while denying information to their partner in order to hold the turn. The robot could similarly manage its own information transmission as a strategy for regulating the interaction. Conversely, the robot should relinquish its turn earlier if the human has clearly conveyed understanding, rather than always insisting on completing the current action in its FSM state. A turn-taking architecture would need to include support for smooth action interruptions to handle this dynamic turn-taking process.

We believe that the cues in the observation model serve as indirect mechanisms for the communication of intention and understanding, while information exchange is what truly drives all turn-taking. Each party manages its own information transmission, fully observable to itself, but must also interpret the other party's information reception, which is only partially observable to itself.

VI. Conclusion

We conduct a data collection experiment in which we collect and code data from 23 human subjects playing "Simon says" with the Simon robot. Our data suggest that minimum necessary information (MNI) is a robust indicator for determining the human response delay to the robot across multiple phases in the interaction. The data also show exclusions in the speech channel and point to ways of analyzing efficiency and adaptation. We intend to use these results to implement a computational model of turn-taking that will lead to an effective and generic controller for autonomous turn-taking in human-robot interaction.

References

[1] E. Tronick, H. Als, and L. Adamson, "Structure of early face-to-face communicative interactions," in *Before Speech: The Beginning of Interpersonal Communication*, M. Bullowa, Ed. Cambridge: Cambridge University Press, 1979, pp. 349–374.

[2] C. Trevarthen, "Communication and cooperation in early infancy: A description of primary intersubjectivity," in *Before Speech: The Beginning of Interpersonal Communication*, M. Bullowa, Ed. Cambridge University Press, 1979, pp. 389–450.

[3] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[4] S. Duncan, "On the structure of speaker-auditor interaction during speaking turns," *Language in Society*, vol. 3, no. 2, pp. 161–180, 1974.

[5] B. Orestrom, *Turn-taking in English conversation*. CWK Gleerup, 1983.

[6] B. Mutlu, T. Shiwa, T. K. H. Ishiguro, and N. Hagita, "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)*, 2009.

[7] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *Proceedings of the Human Language Technologies (HLT)*, 2009.

[8] J. Lee, J. F. Kiser, A. F. Bobick, and A. L. Thomaz, "Vision-based contingency detection," in *Proceedings of the 2011 ACM/IEEE Conference on Human-Robot Interaction (HRI)*, 2011.

[9] J. R. Movellan, "An infomax controller for real time detection of social contingency," in *Proceedings of the 4th International Conference on Development and Learning (ICDL)*, 2005, pp. 19–24.

[10] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proceedings of the 2010 ACM/IEEE Conference on Human-Robot Interaction (HRI)*, 2010.

[11] J. Cassell and K. R. Thorisson, "The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents," *Applied Artificial Intelligence*, vol. 13, pp. 519–538, 1999.

[12] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *Proceedings of the 12th International Conference on Multimodal Interfaces (ICMI)*, 2010.

[13] H. Kose-Bagci, K. Dautenhan, and C. L. Nehaniv, "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*, 2008, pp. 346–353.

[14] E. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, no. 1, pp. 1–63, 2000.

[15] G. Hoffman and C. Breazeal, "Effects of anticipatory action on human-robot teamwork: Efficiency, fluency, and perception of team," in *Proceedings of the 2007 ACM/IEEE Conference on Human-Robot Interaction (HRI)*, 2007.

[16] R. G. Leiser, "Improving natural language and speech interfaces by the use of metalinguistic phenomena," *Applied Ergonomics*, vol. 20, no. 3, pp. 168–173, 1989.

[17] H. H. Clark and J. E. F. Tree, "Using *uh* and *uh* in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.