Turn Taking for Human-Robot Interaction

Crystal Chao and **Andrea L. Thomaz**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia 30332, USA
cchao@gatech.edu, athomaz@cc.gatech.edu

Introduction

Applications in Human-Robot Interaction (HRI) in the not-so-distant future include robots that collaborate with factory workers or serve us as caregivers or waitstaff. When offering customized functionality in these dynamic environments, robots need to engage in real-time exchanges with humans. Robots thus need to be capable of participating in smooth turn-taking interactions.

The research goal in HRI of unstructured dialogic interaction would allow communication with robots that is as natural as communication with other humans. Turn-taking is the framework that provides structure for human communication. Consciously or subconsciously, humans are able to communicate their understanding and control of the turn structure to a conversation partner by using syntax, semantics, paralinguistic cues, eye gaze, and body language in a socially intelligent way. Our research aims to show that by implementing these turn-taking cues within a interaction architecture that is designed fundamentally for turn-taking, a robot becomes easier and more efficient for a human to interact with. This paper outlines our approach and initial pilot study into this line of research.

Approach

Turn-taking is the fundamental way that humans organize interactions with each other. Turn-taking routines, especially in mother-infant gaze systems, have been studied extensively in cognitive science (Trevarthen 1979). Deviations from the expected turn-taking process have been found to cause anxiety in infants, leading to the conclusion that turn-taking is natural and fundamental behavior. Thus it seems logical that socially situated, embodied machines, meant to interact with humans, should use the same deeply rooted turn-taking principles of human social behavior.

Extensive treatment of turn-taking can be found in the linguistics literature as well. Some work focuses on the structure of syntax and semantics, and other work additionally analyze the contribution of paralinguistic cues, gaze shift, and gesticulation (Orestrom 1983). Researchers state that turn-taking is a dynamic and fluid process, including the various complexity levels of floors, turns, and backchannels.

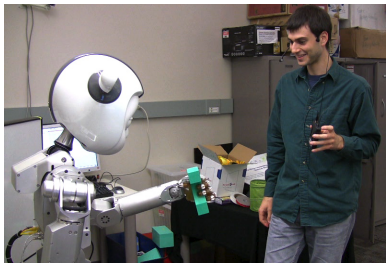


Figure 1: Simon engaging in a turn-taking interaction with a human subject.

In HRI, work on turn-taking needs to be approached from two directions. The first is awareness of the human's cue usage. This is a perception problem closely related to recognizing contingency and engagement, as in (Rich et al. 2010). The second is executing turn-taking cues in a socially intelligent manner. For example, (Mutlu et al. 2009) showed robots using gaze cues to control speaker-listener roles .

The piecewise work eventually needs to be integrated into a broader turn-taking architecture. Using a naive reactive architecture is insufficient in an HRI setting, as missed cues can result in awkward and inefficient timing. Using simple finite state machines leads to confusion when the machine state is not transparent to the human, resulting in repeated commands or pauses. An architecture specifically designed for turn-taking should be able to sit on top of the robot's existing functionality and handle turn timeouts, cue management, and timing of continuous seamless feedback to the human about internal machine state.

Pilot Study

This pilot study focuses on one aspect of turn-taking: effective robot to human turn passing.

Platform

The robotic platform for this research is "Simon," an 38-DOF upper-torso humanoid social robot. Simon's behavior is controlled using the C6 software system. Simon uses vision from one eye camera to identify objects. While an object is directly in front of the camera, Simon moves it and a

color histogram is sampled from the optical flow segmentation. For speech recognition, we use Microsoft Speech API under Windows 7 with a predefined grammar. Both recognized and rejected speech matches are sent to the robot, along with the durations of the speech input for tracing human turn start times.

Interaction Scenario

We use a teaching interaction, where the human teaches Simon to sort colored objects into three bins. The teacher stands opposite Simon for face-to-face interaction. He is responsible for handing objects from the table to Simon when prompted, stating which bins it goes in, and testing Simon.

Simon generalizes the target goal using unsupervised learning on the color histogram to determine the color category and Bayesian maximum-likelihood learning on the target bins to determine the goal location. More details on Simon's learning system can be found in (Chao, Cakmak, and Thomaz 2010).

Study Conditions

The pilot study used a repeated measures design with a turn-taking and a baseline condition. In the turn-taking condition, the robot looks away from the human teacher's face during its own turn. Only when the robot needs human input to proceed does it look back at the person. In the baseline condition, Simon looks at the human teacher in all the instances in which he would look away in the turn-taking condition.

Measures

The following events were logged: human speech, torque spikes for object handoffs, robot speech, robot gestures, robot gaze, and robot learning state transitions.

The start of a robot turn is the time that the first speech command is initiated for the turn, and the end of the turn is when the robot starts awaiting human input. For human speech turns, the end of a human turn is the time that a speech recognition packet reaches the robot controller. The start of the turn can be traced back from the end time based on the duration of the utterance. The torque spike on the left arm chain was also recorded as a human turn boundary.

Subjects were asked to complete a survey after the study. The questions were taken from a survey designed and used by (Cassell and Thorisson 1999) pertaining to perceived life-likeness of a virtual agent.

Preliminary Results

The pilot study included eight subjects. All subjects were males with ample robotics experience (which likely had an effect). There was a robot error and shutdown during one subject's run, so that data is not included. None of this data is statistically significant or should be used to form conclusions about turn-taking in HRI. However, qualitative examination of the turn data logged suggests quantitative analysis that may be possible with a more rigorous future experiment.

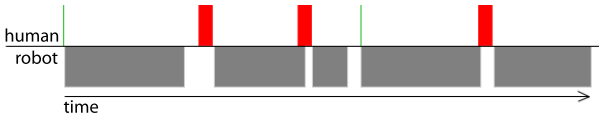
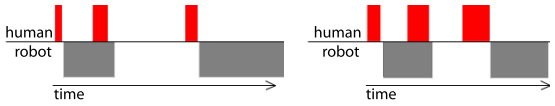


Figure 2: An example of good turn-taking.



(a) Long pause. (b) Broadening.

Figure 3: Examples of undesirable turn patterns.

Turn Patterns

One of the interim goals of this research is characterizing the quality of a turn-taking interaction. The recurring turn-taking problems in the pilot study across subjects suggest universal templates for turn-taking quality that can be automatically detected. The diagrams in Figures 2 and 3 show visualizations of such patterns.

The pattern in Figure 2 demonstrates an example of smooth turn-taking in this domain. There is minimal overlapping of turns as well as minimal time spent between turns, characteristics of an efficient interaction.

Figure 3 shows two undesirable effects of poor turn-taking. Figure 3(a) shows an overlapping turn followed by a long pause before another human turn attempt. This happens when the human starts and finishes issuing a command to the robot before the robot is ready for input. The human waits a while for the robot to respond to the command and when the robot does nothing, the human repeats himself.

Figure 3(b) has the same cause but the effect is that the human repeats himself more slowly each time, such that the red area becomes broader and broader. When the robot is unresponsive to human speech, the human tends to experiment with articulation and speed of his own speech. The effect of this is that the human thinks that speaking slower is more successful, when really speaking later is the key.

Activity

The average number of extra turn attempts was slightly higher for the baseline condition (10.1%) than the turn-taking condition (7.4%). For this particular task, this indicates how often the human had to repeat himself, with turn-taking requiring fewer repetitions.

The breakdown of activity across all the subjects for each condition is shown in Table 1. As can be seen, most of the time was spent on the robot's turn, since this particular task requires the robot to do much more than the human. One might expect the *Intersecting* and *Neither* categories to be minimized during good turn-taking compared to poor turn-taking. The *Neither* category should still be nonzero due the latency that occurs in natural turn-taking. The value of more or less *Intersecting* is debatable (see Challenges).

Table 1: Percentage of time spent under each type of activity

Activity Type	Baseline	Turn-taking
Intersecting	4.7%	4.1%
Robot Only	63.0%	66.2%
Human Only	10.5%	9.7%
Neither	21.9%	20.0%

Survey Response

Only two subjects claimed they could consciously detect a difference between the two conditions. One of the two said that he felt a difference between the conditions relating to eye contact, feeling less eye contact in the baseline condition than the turn-taking condition. Perhaps people feel more like the robot is turning to look at them if it looks away first.

The rest of the subjects rated the two conditions with the same value for all of the survey questions. This suggests the repeated measures format may not be appropriate unless perhaps more cues are implemented such that the turn-taking signal is more observable.

Summary of Challenges and Future Work

This pilot study highlighted many challenges of working in this domain. We outline these as a roadmap for future work.

One problem with turn analysis is that human speech alone does not accurately portray what the human is doing. A person might in fact be doing other things such as examining the objects on the table to decide which one to select, or unsuccessfully pushing objects in the robot's hand. Whether or not these should be construed as the human taking his turn is an open question. In general when gestures occur in parallel with speech, does speech or gesture finality indicate a turn end? Many subjects interrupted Simon's shrug animation, taking the end of his speech to be the end of his turn.

Less ambiguously, human gaze can typically signal turn-taking. For HRI studies this could be done with time-consuming video coding, but ultimately the robot needs to detect these events real-time in order to act upon them. This perception problem is an essential one to solve in order to create a turn-taking architecture.

When turns can be completely logged and analyzed for both parties, an important question becomes how to quantify how good the interaction is. What is an appropriate metric for good turn-taking? Some analysis from linguistics literature counts the number of interruptions and false starts as indicators of failed turn exchange; overlapping here is considered negatively because it corrupts perceptual data. However, overlapping may not always be undesirable. Highly fluent human-robot collaboration results in significant overlapping of activity and thus highly efficient task execution (Hoffman and Breazeal 2007). In this study, the entire interaction would actually be completed more efficiently if the start of human speech occurred before the end of the robot's turn, but the end of the human speech fired after the robot's turn was relinquished. Turn-taking analysis thus needs to take these factors into consideration.

A less dire problem was subjects not looking up at the robot and noticing the cues. It may be better for the subjects to memorize the speech recognition grammar. Usually the initial turns are accompanied with hesitation and looking towards the experimenter, and after several repetitions subjects did not need to refer to the grammar but would be drawn to look at it anyway. In a future experiment, the subjects should practice using the grammar before data collection begins.

The social intelligence of the subjects is also a factor. Humans who cannot act appropriately on human turn-taking cues may also have trouble with human-robot turn-taking. It may be informative to evaluate the subject's baseline behavior with another human prior to any experiment.

Conclusion

Turn-taking is a fundamental skill for human communication and is one that robots will need in order to achieve natural communication with humans. Robots embedded in HRI scenarios stand to benefit from an architecture designed specifically for turn-taking with humans. The work discussed in this paper takes a first step towards such an architecture by implementing turn passing and conducting some preliminary analysis on data from a pilot study. The analysis demonstrates several ways to investigate turn patterns, human and robot activity, and subjects' perceptions, as well as recommends several revisions of the pilot study for a future experiment.

References

Cassell, J., and 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.

Chao, C.; Cakmak, M.; and Thomaz, A. L. 2010. Transparent active learning for robots. In *Proceedings of the 2010 ACM Conference on Human-Robot Interaction (HRI)*.

Hoffman, G., and Breazeal, C. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, 1–8. New York, NY, USA: ACM.

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 2009 ACM Conference on Human-Robot Interaction (HRI)*.

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

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

Trevarthen, C. 1979. Communication and cooperation in early infancy: A description of primary intersubjectivity. In Bullowa, M., ed., *Before Speech: The Beginning of Interpersonal Communication*. Cambridge: Cambridge University Press. 389–450.