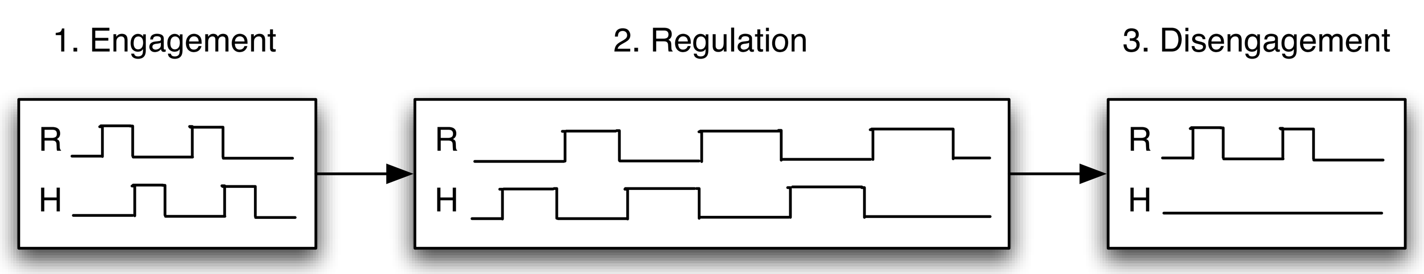
Turn-Taking for Human-Robot Interaction

Crystal Chao and Andrea L. Thomaz
Socially Intelligent Machines Lab, Georgia Institute of Technology

Turn-Taking

An interaction can be viewed as being in one of three stages: engagement, regulation, and disengagement.



A contingency detector can tell a robot whether to engage or disengage a person (stages 1 and 3). After engaging, the dyad performs turn-taking to regulate interaction timing (stage 2).

The turn-taking process is highly multimodal and uses many channels for communicating turn state, including gaze, prosody, and whole-body gesture.

Architecture

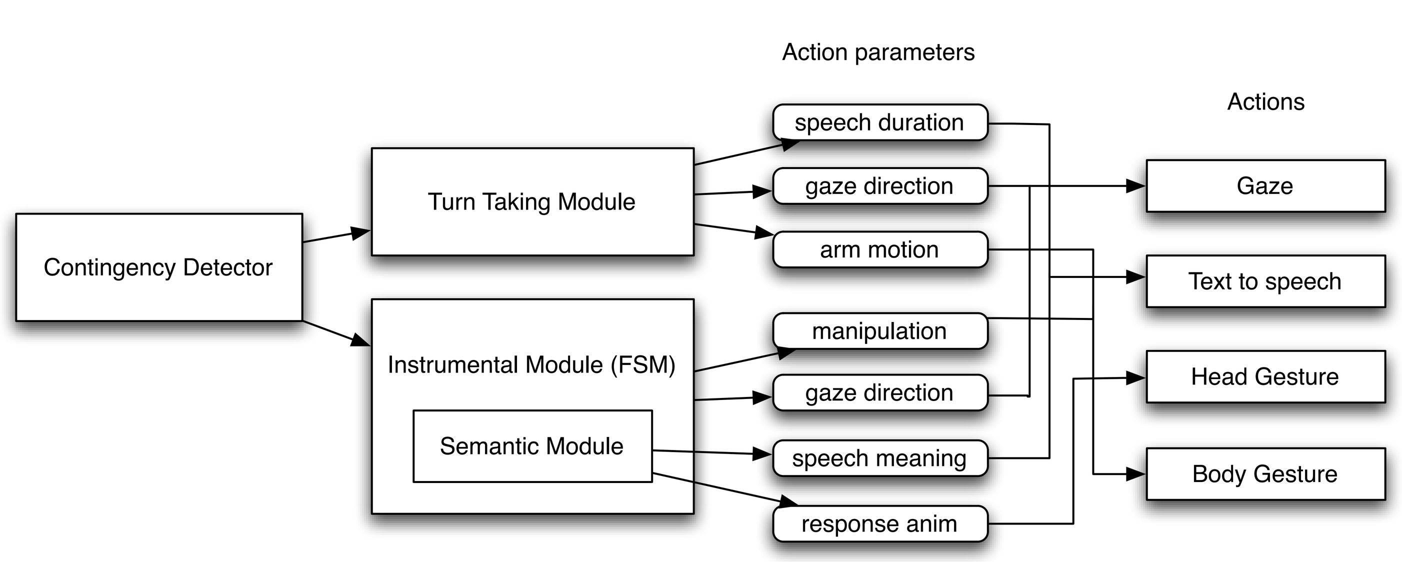
We hypothesize that there is some context-free component of turn-taking, which allows us to modularize the behavior and reuse it in a way that is generic to multiple types of interactions.

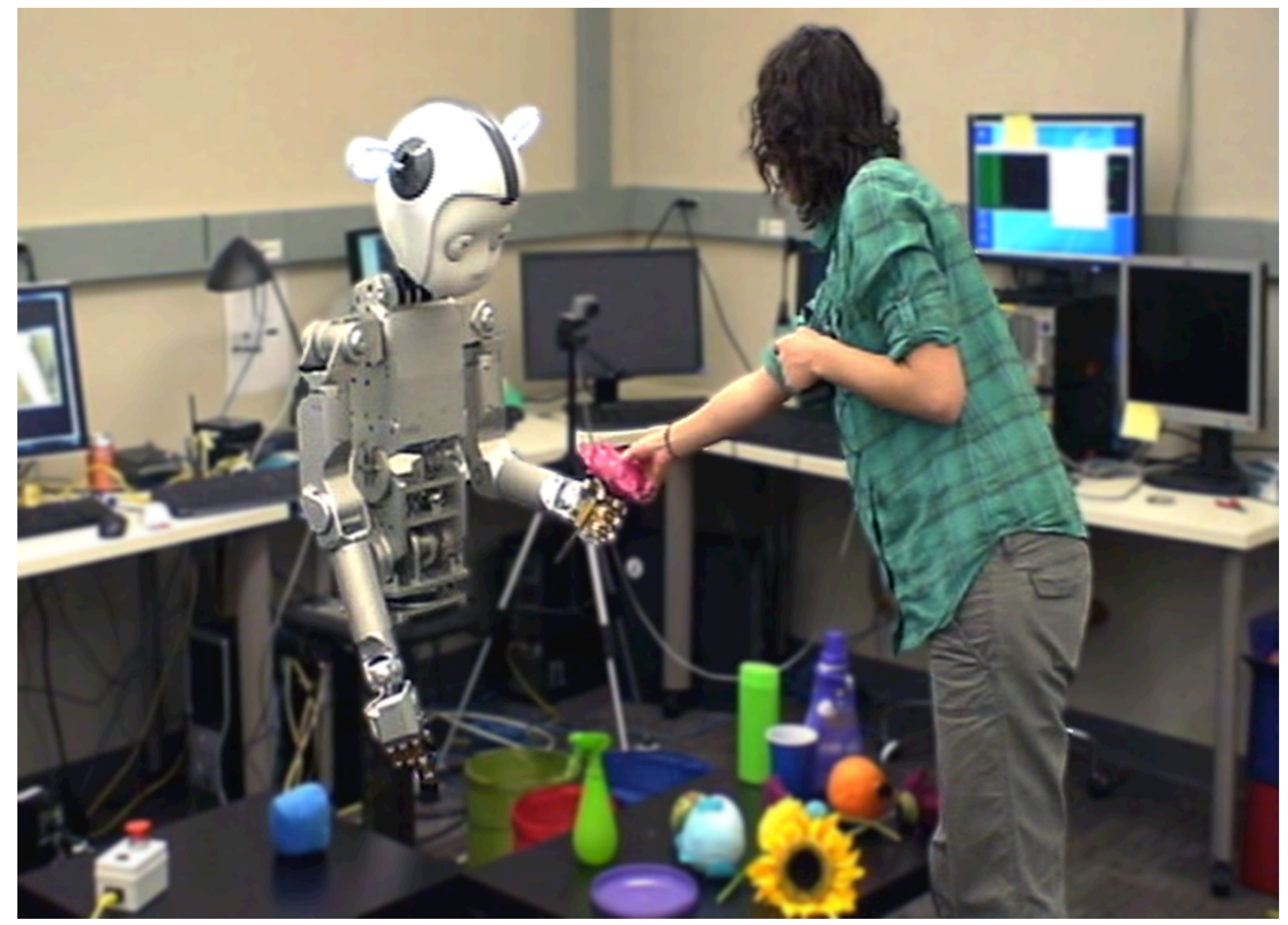
Components

Contingency Detector – triggers stage 2 modules
Turn-Taking Module – context-free turn regulation
Instrumental Module – task-based finite state machine
Semantic Module – library of ways to communicate a meaning

Action Selection

The context-based (instrumental) and context-free (turn-taking) modules generate action parameters, which combine to select actions that are executable by the robot.



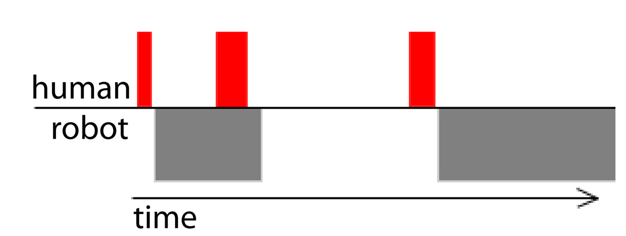


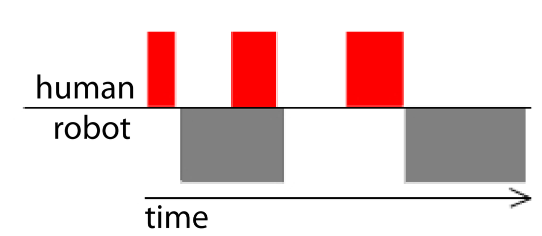
Goals

Interaction improvement

A model for turn-taking may allow robots to have smoother, more efficient interactions that are less frustrating for humans.

Some common error patterns in our interactions include the human waiting for too long, or the human repeating himself:





Ease of programming

For each new domain or task, the programmer should be able to implement only domain-specific behaviors without spending time coding turn-taking behaviors in the instrumental FSM.

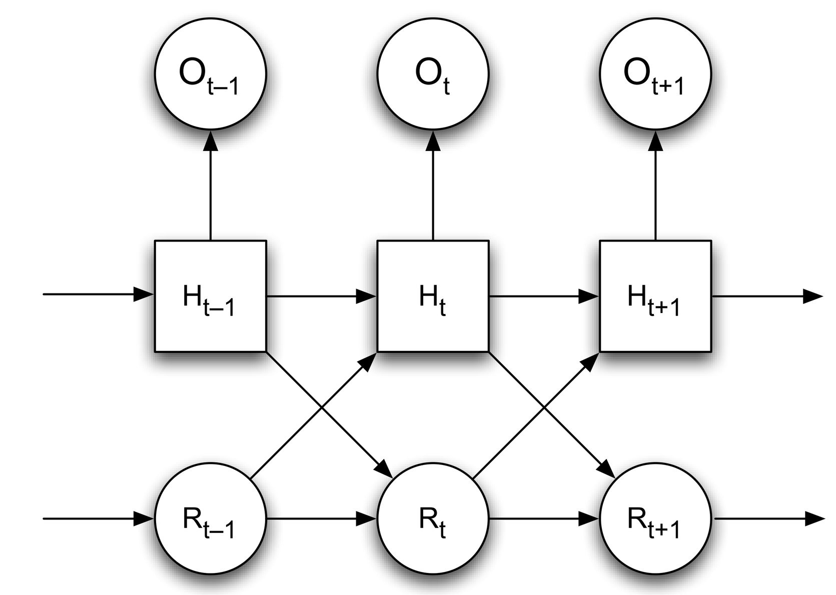
Domains

Turn-taking is often investigated in the free conversation domain. Other interaction domains may have significant effects on dynamics and observation distributions, such as for gaze and motion.

Some domains we are considering include learning from human demonstrations, imitation games, play-based interactions with children, and providing information to passersby.

Temporal Inference Model

We describe turn dynamics as a first-order Markov process.



At time *t*:

O_t = observation
H_t = human turn state
R_t = robot turn state

The hidden state H_t is filtered over time.

States: R_t, $H_t \in$ {Seizing, Passing, Holding, Listening}

Observations: Vector of binary features, e.g. O_t = <1,0,1,1,0>
Features: <mic level, gaze direction, speech recognized by grammar, hand motion, prosody derivative>

Conditional probability distributions

Three functions to train from video-coded experiment data, in which the robot is tele-operated to do good/bad turn-taking:

Observation function $P(O_t|H_t)$
how well the sensor data reflects the human turn state H_t

Human transition function $P(H_t|H_{t-1},R_{t-1})$
how the human takes turns

Robot transition function $P(R_t|R_{t-1},H_{t-1})$
how the robot should do turn taking. a good and a bad model

Extensions

Semi-Markov process: A Markov chain assumes that the times between events are geometrically distributed. Other distributions can be modeled using a semi-Markov process that depends on an additional node Q_t representing time spent in a state.

Switching model: The human transition function may be vastly different across different people; certain people may be more aggressive or more passive turn-takers. The robot can have different models and select the maximum likelihood model.

State space: Backchannels could be explicitly represented.