Generating Speech And Gesture for Robotic Communication

Why do robots need to communicate?

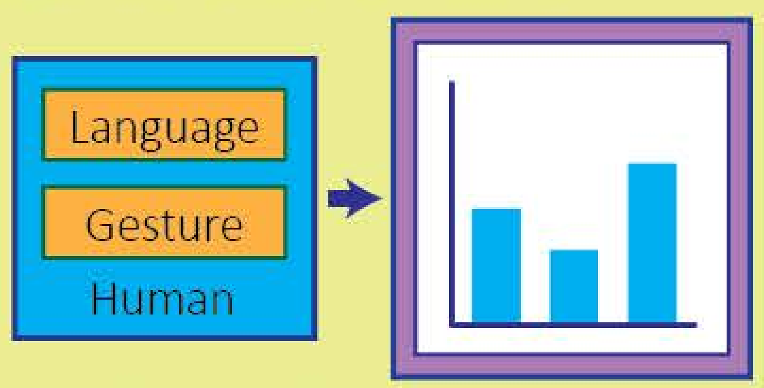
- to allow them to interact naturally with us
- to enable them to relay their current state
- to enable them to ask for help when needed

Communication is a two-way process:

- Listening: understanding speech and gesture
- Speaking: producing speech and gesture

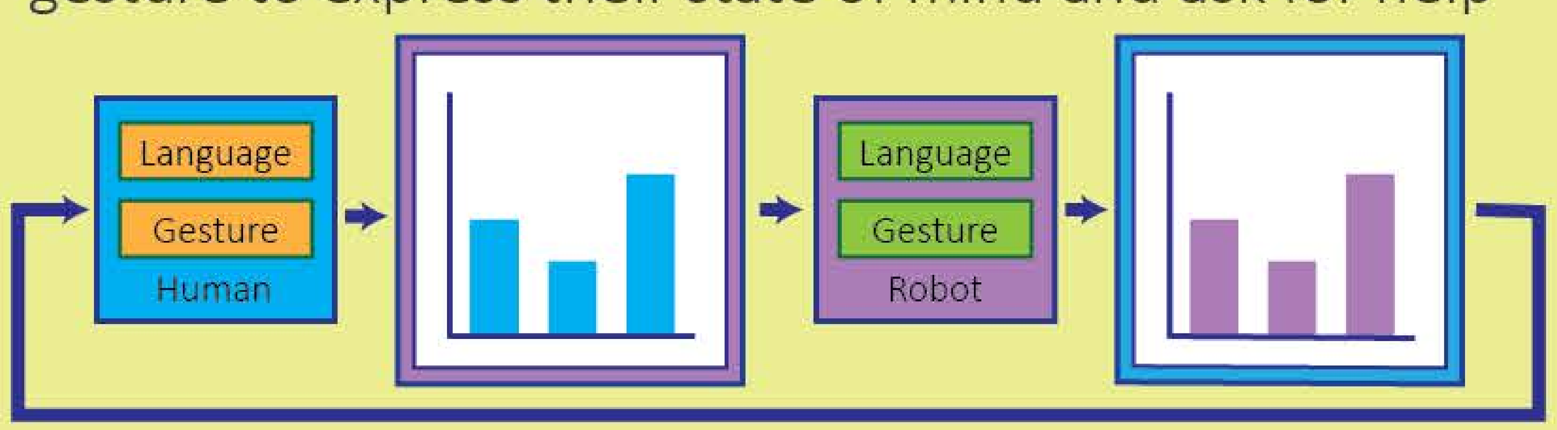
Existing work allows robots to listen.

Speech and gesture are interpreted to estimate the human participant's state of mind



In our project, the robot can speak.

We formulate a model to allow robots to use speech and gesture to express their state of mind and ask for help

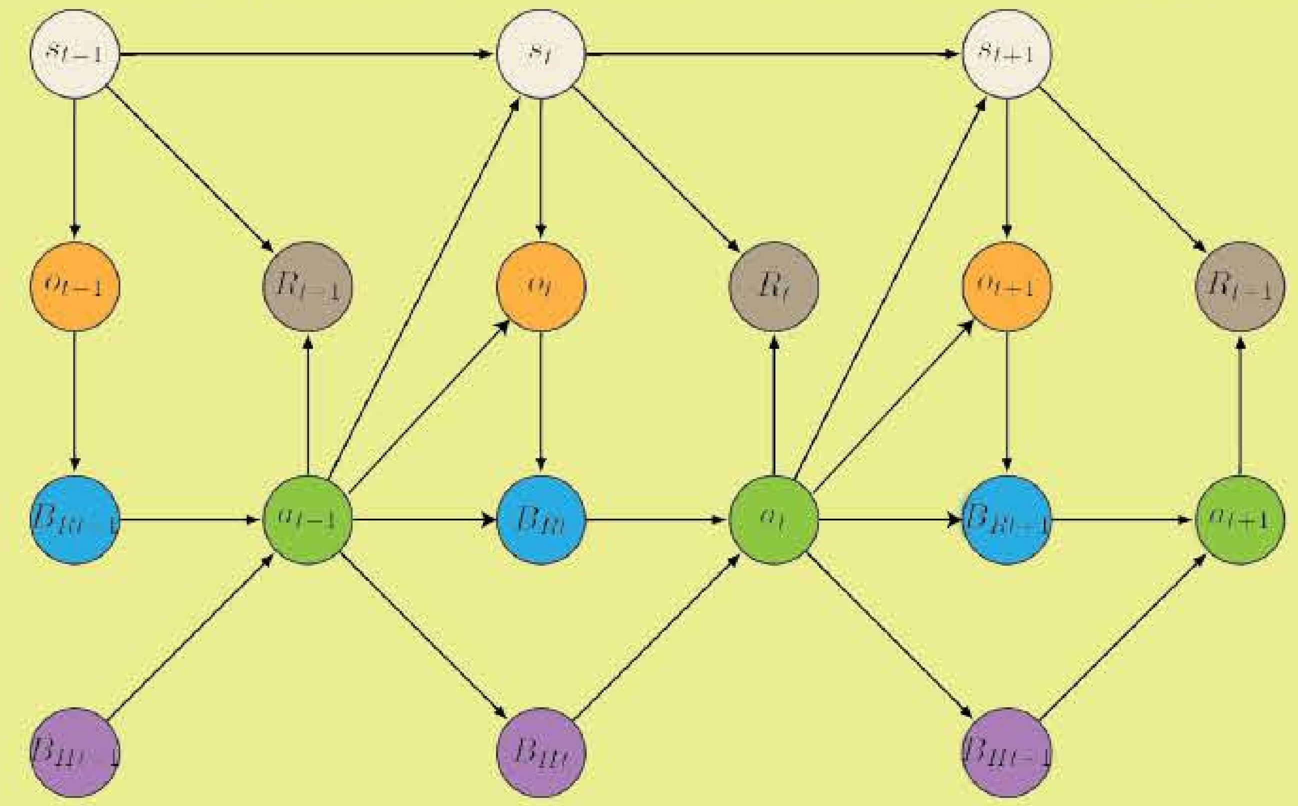


The task:

- Objects are arrayed on a table within the robot's reach
- The human requests an object from the robot using speech and gesture
- The robot must determine which object the human is referring to and hand it to the human
- *If the robot is unable to determine which object the human is requesting, it can use speech and gesture to communicate this uncertainty or ask for help*

Communication Model: MOMDP

(Partially Observable MDP with Mixed Observations)



States (s) are composed of a *hidden* component

- which object the human wants (ω)

and a *visible* component

- which objects are on the table (Θ)
- (distribution over) which object the human believes the robot will hand them (h)

Observations (o) inform our belief about which object the human wants (ω). They are composed of *language* (L)

$$p(L|s) = p(L|\omega) = \frac{\#\ Ls\ describing\ \omega}{\#\ words\ describing\ \omega}$$

and *gesture* (G)

$$p(G|s) = p(G|\omega) \approx N(\omega, v)$$

where N(ω,v) is a 2d normal distribution centered at ω's location with variance v

Actions taken affect the state, including the (distribution over) the object the human believes we will hand them (h)

The robot can take the following actions:

- Hand over an object
- Wait
- Speak (Ask a question)
- Gesture (Point at something)

We calculate how these actions affect the human's state (the distribution over h) by assuming the human interprets them the same way the robot interprets the human's actions

Asking a question and gesturing serve a dual purpose:

- information gathering: they elicit a human response to help inform the robot
- they communicate to the human the robot's current state, allowing for richer communication

Optimizing Performance for Real Time

The robot must respond quickly for to communicate effectly. However, Partially Observable MDPs (POMDPs) are notoriously difficult to solve optimally.

To solve a POMDP, determine the optimal action for every state. Most methods use a form of tree search.

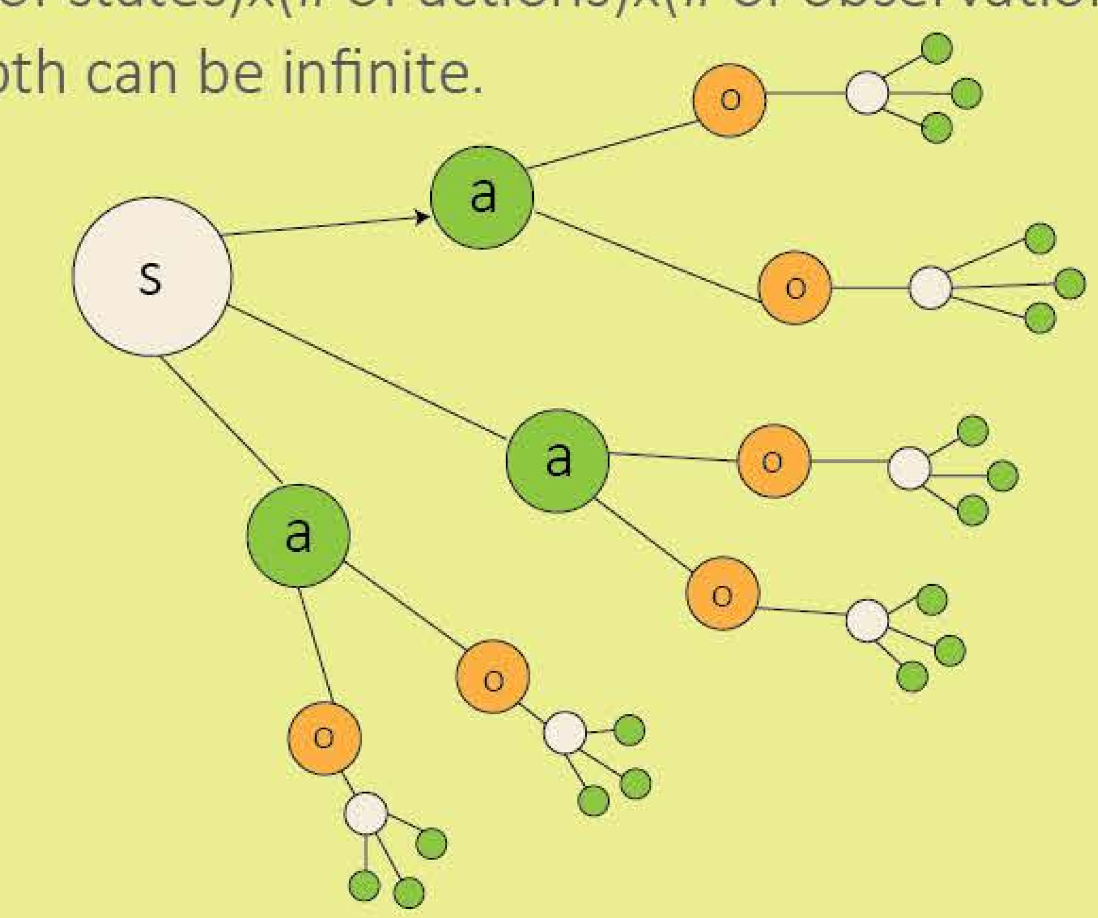
A POMDP state is a distribution over hidden states

----there are infinite number of states

The branching factor is

(# of states)x(# of actions)x(# of observations)

The depth can be infinite.



We reduce the number of calculations by pruning the tree.

- Explore only a finite horizon, only expand to depth d
- Assume some attributes of the state (e.g., distribution over h) are observable, reducing the state space
- Use Belief Sparse Sampling algorithm to only sample c observations instead of considering all of them
- reduce number of actions via Macro Actions
  - 1 Handoff action per Object =>
    - 1 Hand off most likely object action
  - 1 Question per L =>
    - 1 Ask the L with expect resulting smallest expected entropy

Training and Classification

- Precompute and record belief state and chosen action
- At runtime, use KNN to calculate the chosen action using the belief state as a feature vector
- Allows for extremely fast response times, as the tree does not need to be expanded