A Model of Continuous Intention Grounding for HRI

Julian Hough and David Schlangen
Dialogue Systems Group // CITEC
Faculty of Linguistics and Literature
Bielefeld University
firstname.lastname@uni-bielefeld.de

ABSTRACT

We present a formal model of intention grounding for robots in HRI which models the user and robot's own current intentions simultaneously. Central to the model are interactive state machines, and strength-of-evidence functions, which are evaluations of the degree to which an agent's current intention is being communicated publicly. The model has so far been used in a simple pick-and-place robot with a speech interface, but we argue this can be generalized to other robots such as semi-autonomous vehicles, providing potential for automatic evaluation and active learning.

1. INTRODUCTION

In HRI, the intentions of humans and robots are fundamentally different in terms of how they are arrived at, and how they are internally represented. However, if appropriate assumptions are made when designing a robotic system, it is possible to achieve sufficient overlap (or homomorphism, in an informal sense) between the different agents' intentions in order to carry out a task together. The inherent differences in the internal representations and processing of humans and robots [10] makes arriving at these overlapping intentions an even greater challenge. However, one can imbue robots with communicative *grounding* mechanisms in the sense of [2, 1], which are the abilities to build and align internal representations towards shared information or "common ground".

In robots with speech interfaces, there has been much work on user intention recognition, forming part of the subfield of Spoken Language Understanding (SLU). Conversely, on the side of the robot, grounding the robot's intention can be achieved through robots having high legibility in their actions to make them 'intent-expressive' to users [4, 3]. What is required in real-world HRI is to model both the user and robot's intentions simultaneously, and the public evidence for them within a single system– SLU runs online during the robotic action, where the robotic action immediately and seamlessly updates the context in which the words being spoken are interpreted. Achieving a closed-loop interaction which ensures user intentions are dynamically recognized accurately, and also quickly, is a challenge we address here.

2. A MODEL OF CONTINUOUS INTENTION GROUNDING FOR HRI

We build on the proposal of [9] for a communicative grounding model for HRI, inspired by recent attempts to incrementalize grounding strategies in dialogue models [6, 5, 8], which can be purposed for simple robots with speech interfaces if certain modifications are made. The first modification is that the robot's actions have the same status as dialogue acts. The second is that commitment to goals can be real-valued rather than absolute, and this commitment can be evaluated by *strength-of-evidence functions* which monitor the degree to which each agent is showing commitment to their goal at a given time.

Statecharts with strength-of-evidence functions.

We follow work using Harel statecharts [7] for dialogue control in robotic systems by [11, 12]. Fig. 1 defines the grounding state machine for a simple robot which interprets a user's speech to carry out actions. Here we characterize the user and robot as having *parallel* states, represented either side of the dotted line.

Fig. 1 shows the states and "triggering conditions" that must be satisfied to allow state transitions (written on the arcs between state boxes, where specific guards are in square brackets). The main motivation of the model is to explore the criteria by which the robot judges both their own and their interaction partner's goals to have become *publicly manifest* (though not necessarily grounded) in real time, and therefore when they are showing commitment to them. To determine which grounding state each agent is in, we use evaluation functions Ev for each agent's state within the triggering conditions– these are *strength-of-evidence* valuation functions that return a real number value indicating the degree to which the agent has displayed their goal publicly, according to the robot's best knowledge. Goals are hidden and estimated in the case of the user state and observed in the case of the robot, yet both have to be evaluated for the degree to which they are manifest to allow appropriate interpretation of the user's speech.

$UserGoal$ is estimated as the most likely user intention from a set of possible goals $Intentions$, given the current utterance u, the robot's state $Robot$, user's state $User$ and the current task's state $Task$, as in (1), with $Ev(UserGoal)$ characterized as its probability in (2). $Intentions$ is the set of intentions specified on a degree of abstraction deemed relevant by the system designer– for example a possible intention could be $TAKE(X)$ for a robot capable of taking object X. While they are important, the lower-level intentions required to realize the higher-level ones need not be discussed for the overview of the model here.

$$UserGoal := \underset{i \in Intentions}{\arg\max}\ p(i \mid u, Robot, User, Task) \quad (1)$$

$$Ev(UserGoal) := \max_{i \in Intentions} p(i \mid u, Robot, User, Task) \quad (2)$$

While the estimated user's goal is continuously being up-

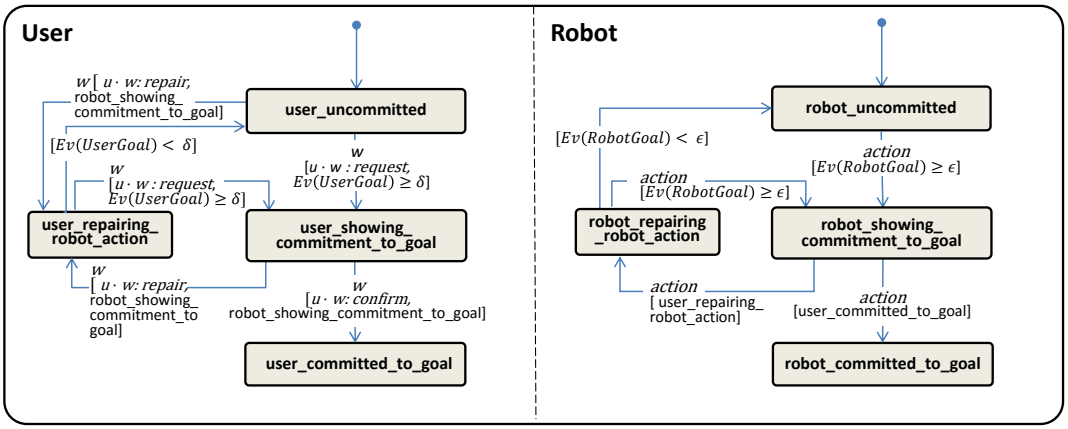

Figure 1: Interactive Statechart as modelled by the robot consisting of two parallel, concurrent states, one for each participant. The triggering events and conditions in the transition functions (the directed edges) can reference the other state.

dated through new evidence, this goal can only be judged to become sufficiently mutually manifest with the robot when a certain confidence criterion has been met– here we characterize this as a real-valued threshold δ. Using a real-valued threshold allows experimentation into increasing responsiveness of the robot by reducing it [8]. As Fig. 1 shows, once $Ev(UserGoal) \geq \delta$ then the state `user_showing_commitment_to_goal` can be entered. In a fully cooperative system one can assume the assignment $RobotGoal := UserGoal$ is then carried out upon entering the state (though we omit this from the core grounding model given cooperativity is not assumed).

Conversely, the Robot's view of its own grounding state uses the function $Ev(RobotGoal)$ and its own threshold ϵ. Unlike the user, the robot's goal is taken to be fully observed to itself as some $i \in Intentions$, however it must still estimate when i is made public by its actions, and we assume this is a probability function which considers the states of $Robot$, $User$ and $Task$ as the user's state machine does, but also depends on its most recent action $action$:

$$Ev(RobotGoal : i) := p(i \mid action, Robot, User, Task) \quad (3)$$

Given this is an approximation to the user's inference function from the robot's action to the most likely goal (rather than the predictability of the action given the goal), it can be seen as a measure of the legibility of the robot's action [3]. Once ϵ has been reached, or the action has become sufficiently legible, the robot may enter `robot_showing_commitment_to_goal`. Once in this state it is permissible for the user state to either commit to the mutually manifest goal and trigger grounding, else engage the robot in repair, entering `user_repairing_robot_action`. Then, as soon as is physically possible in the motor plan, the robot state will become `robot_repairing_robot_action`. The repairing state's internal processes are identical to the initial `user_uncommitted` one, except the first action upon entry is to prune $Intentions$ such that:

$$Intentions := \{i \mid p(RobotGoal \mid i) = 0\} \quad (4)$$

(4) removes all those intentions which would eventually lead to entry to the repaired $RobotGoal$ intention. The robot will remain in this repairing state until the user's state has exited `user_repairing_robot_action`, triggering the end of the user-initiated repair interaction. Note that it is only possible for the user state to repair the $RobotGoal$, rather than $UserGoal$– the user can repair the latter through self-repair, but that is currently not represented as its own state. Repair of the robot's current action is only possible through knowing it had shown commitment to a goal which caused it (i.e. been in the state `robot_showing_commitment_to_goal`), otherwise, as per normal principles of situated dialogue, it would not be able to interpret the utterance as a repair. The strength-of-evidence function $Ev(RobotGoal)$ and the threshold ϵ are therefore of tantamount importance, as they determine when confirmations and repairs can be interpreted as such, and consequently determine the interactive dynamics of the system.

Fluidity through incremental processing.

We achieve fluidity in this grounding process through incremental processing. The increment of the triggering events in the $User$ state is the latest word w in current utterance u (as opposed to the latest complete utterance). The principal Natural Language Understanding (NLU) decisions are therefore to classify incrementally which type of dialogue act u is, (e.g. u : *confirm*), whether w begins a new dialogue act or not, and estimate $UserGoal$ from the set $Intentions$, whatever they may be in the given application. The grounding statechart is then checked to see if a transition is possible from the user's current state as each word is processed, akin to incremental dialogue state tracking [13].

3. EXAMPLE APPLICATIONS

For a given robot, while the interactive statechart defines the transitions between grounding states, their triggering criteria require the definition of the variables of the estimated current $UserGoal$ from a set $Intentions$, its strength-of-evidence function $Ev(UserGoal)$, threshold δ, and the analog for the robot state, $Ev(RobotGoal)$ and threshold ϵ. The below subsections give two examples, however we emphasize these are two simple sets of choices, and the potential for optimizing these provides a useful avenue for research, and the potential for learning and selecting some of these elements through interaction is also possible.

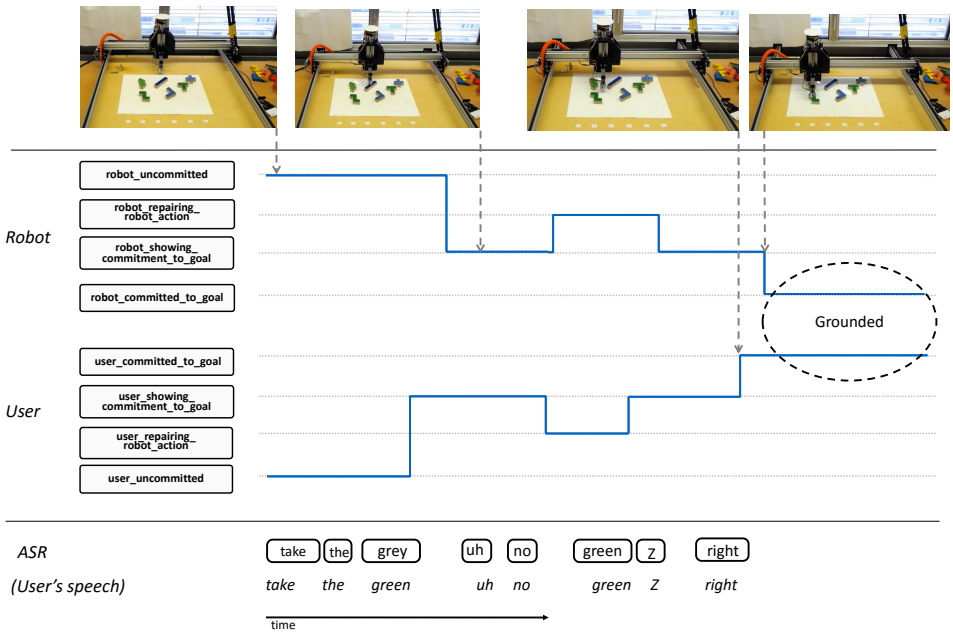

Figure 2: Concurrent User and Robot grounding states during an interaction where an initial mis-recognition of 'green' as 'grey' by the ASR, and confusion over colours in reference resolution where 'grey' gives higher probability to a blue object. The recognition of repair allows the participants to become grounded again.

3.1 Pick-and-place robot with voice interface

Our first example application can be seen in Fig. 2, a simple pick-and-place robot which performs natural language understanding incrementally as words are spoken. The robot's principal objective is to resolve references to the objects in the scene and place them at target locations.

We characterize $UserGoal$ as taking or placing the most likely object out of the referent set R according to the robot's reference resolution's output distribution, given the user's utterance u so far (e.g.(5)) and *Intentions* is simply the set of possible actions applicable to all the possible referents. When the action is $TAKE$, the Ev function is the probability of the most likely object being referred to as in (6). It is an empirical question to find a suitable δ which (6) should reach to allow entry into `user_showing_commitment_to_goal`– making this lower means the robot is quicker to react to the user's speech, though could be incorrect in estimating the user's intended referent, while higher values may result in better accuracy but with a noticeably less fluid interaction [8].

$$UserGoal := TAKE(\arg\max_{r \in R} p(r \mid u)) \tag{5}$$

$$Ev(UserGoal) := \max_{r \in R} p(r \mid u) \tag{6}$$

$UserGoal$ is obtained incrementally with a simple NLU method using the results from the robot's reference resolution and the *Robot* and *User*'s current grounding states. Firstly, sub-utterance dialogue act (DA) classification is performed, judging the utterance u to be in $\{request, confirm, repair\}$. Then the state machine is queried to see if transitioning away from the current state is possible and the $UserGoal$ is updated – see [9].

The robot's state machine module partially consists of the *Robot* grounding statechart in Fig. 1. When the *User* state is `user_showing_commitment_to_goal`, the *RobotGoal* is set to $UserGoal$, whereupon it plans to grab or place the estimated referent. The robot can estimate the time needed to pick up the referent r_i based on a vector v containing its arm's current position and velocity with a function $MinTimeToGrab(r_i, v)$. We use this to characterize a simple strength-of-evidence function for the robot's intention in (7). This approximates the predictability that its current action will lead to picking up its target r_i with a softmax function over the negative time estimations for grabbing each object still in play.

$$Ev(RobotGoal : TAKE(r_i)) := \frac{e^{-MinTimeToGrab(r_i,v)}}{\sum_{r_j \in R} e^{-MinTimeToGrab(r_j,v)}} \tag{7}$$

The threshold ϵ determines how quickly the robot state machine can enter `robot_showing_commitment_to_goal`, which is the point where the robot's action can be repaired or confirmed by the user. If ϵ is low, then it may optimistically interpret repair and confirmation acts as referring to its goal early in its movements, while if it is too high, repairs and confirmations may only be interpreted as such when the arm is very near the target object, making the interaction safer but more cumbersome [8].

Fig. 2 shows the state dynamics for the concurrent statechart during an interaction with repair. Notice how the *Robot* state mirrors, though slightly lags, the *User*, by virtue of the fact that it takes time to demonstrate commitment to a given goal with a sufficiently strong $Ev(RobotGoal)$, or legibility. The robot's ASR error leads to it showing commitment to picking up the wrong object. A user-initiated repair interaction begins. During the repair, *Intentions* is pruned such that all intentions implying the current *RobotGoal* are removed as in (4). The robot consequently changes its goal to match $UserGoal$ and re-enters `robot_showing_commitment_to_goal` once

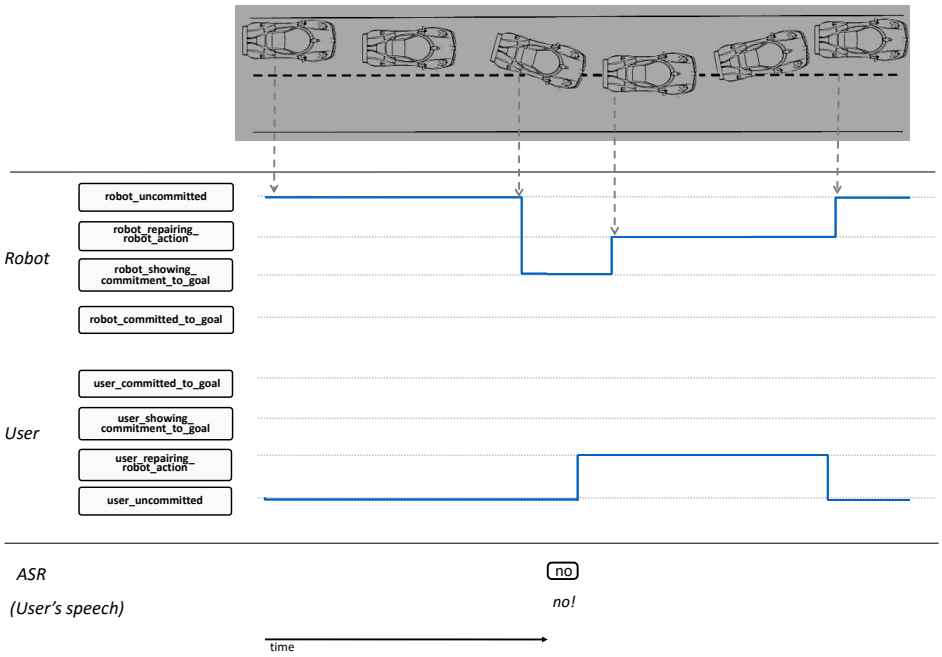

Figure 3: An example of the grounding state model in operation for a semi-autonomous vehicle which has taken the initiative to change lanes, which is repaired by the user.

its new movement has become legible. In this state, the user's confirmation "right" is interpreted as referring to the revised *RobotGoal*, triggering the entry to `robot_committed_to_goal`.

In experiments with this robot, [9] show that user ratings of the robot's perceived understanding correlate very strongly with a simple internal measure of understanding derivable from the grounding state machine as the number of state entries into the `robot_committed_to_goal` state over time. This understanding measure could be used for automatic evaluation and as a reward in reinforcement and active learning paradigms in future work.

3.2 Semi-autonomous car with voice interface

We now briefly outline a possible application to semi-autonomous cars. We envisage a situation where a car is left to make autonomous decisions, while the user can intervene when they wish to with voice commands.

We show an example where the car decides to change lanes automatically, upon which the user sees an oncoming traffic jam in the lane being moved to beyond the robot's field of view, and then repairs the lane change to go back to the original lane– see Fig. 3.

Focusing on the robot's intention to change to lane l_i, given a function which estimates the minimum time needed for the car to join a given lane l_j's trajectory from its current position and velocity in vector v, $MinTime(l_j, v)$, we posit the simple strength-of-evidence function in (8).

$$Ev(RobotGoal : ChangeLane(l_i)) := \frac{e^{-MinTime(l_i,v)}}{\sum_{l_j \in Lanes} e^{-MinTime(l_j,v)}} \tag{8}$$

We note here that as this is an approximation to the user's inference function from the snippet of the car's movement trajectory to the car's goal, more sophisticated approaches to legibility of movement shown in [4] with continuous functions could be used here.

In Fig. 3, once the *RobotGoal* to change lanes becomes legible, with (8) reaching ϵ, the user's repair act "no!" can be interpreted as referring to that goal. The robot then engages in repair until both agents can become uncommitted again. Repairing the *RobotGoal* again uses (4) to prune $Intentions$, which in the absence of another sufficiently strong $UserGoal$ causes reversion to the original uncommitted state rather than a correction of the action as in Fig. 2.

4. CONCLUSION

We have presented an abstract model for intention grounding in HRI which allows investigation into continuous intention recognition and expression for different robots with speech interfaces. The choice of the possible intentions, strength-of-evidence functions and thresholds will vary with the affordances of the robot and its primary tasks. However, provided the robot has the ability to monitor the ongoing progress of its actions, and provided those actions are interruptible, our model can be used as a framework for investigating the optimal strength-of-evidence functions for intentions in HRI.

Furthermore, we propose that in future work, rather than restricting the definition of all the elements of the model to be defined at design time, the strength-of-evidence functions and the thresholds could be learned during HRI using model selection and reinforcement learning methods.

Acknowledgments

We thank the two anonymous reviewers for their helpful comments. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG), and the DFG-funded DUEL project (grant SCHL 845/5-1).

5. REFERENCES

[1] H. H. Clark. *Using language*. Cambridge university press, 1996.

[2] H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991), 1991.

[3] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58. ACM, 2015.

[4] A. D. Dragan, K. C. Lee, and S. S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[5] A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, London, UK, 2015. ACL.

[6] J. Ginzburg, R. Fernández, and D. Schlangen. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9), 2014.

[7] D. Harel. Statecharts: A visual formalism for complex systems. *Science of computer programming*, 8(3), 1987.

[8] J. Hough and D. Schlangen. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 288–298, Los Angeles, September 2016. Association for Computational Linguistics.

[9] J. Hough and D. Schlangen. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*, 2017.

[10] G.-J. M. Kruijff. There is no common ground in human-robot interaction. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, 2012.

[11] J. Peltason and B. Wrede. Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL, 2010.

[12] G. Skantze and S. Al Moubayed. Iristk: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012.

[13] J. D. Williams. A belief tracking challenge task for spoken dialog systems. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*. ACL, 2012.