It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot

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

ABSTRACT

For effective HRI, robots must go beyond having good legibility of their intentions shown by their actions, but also ground the degree of uncertainty they have. We show how in simple robots which have spoken language understanding capacities, uncertainty can be communicated to users by principles of grounding in dialogue interaction even without natural language generation. We present a model which makes this possible for robots with limited communication channels beyond the execution of task actions themselves. We implement our model in a pick-and-place robot, and experiment with two strategies for grounding uncertainty. In an observer study, we show that participants observing interactions with the robot run by the two different strategies were able to infer the degree of understanding the robot had internally, and in the more uncertainty-expressive system, were also able to perceive the degree of internal uncertainty the robot had reliably.

Keywords

Communicative grounding; Uncertainty; Incrementality

1. INTRODUCTION

In human-human interaction, understanding is not an all-or-nothing affair. When following a request to do something, we can act tentatively, displaying uncertainty about our understanding. In HRI, much existing work is concerned with the legibility of robot actions, under the assumption that the robot's current goal is always certain to the robot, as is the robot's basis for that goal [6, 5]. In this paper we investigate how simple robots with spoken language understanding capabilities can best communicate their internal uncertainty to users and onlookers with their non-verbal, task actions. Particularly, we investigate how the degree of a robot's uncertainty can be communicated through the manner of execution of its actions during interaction.

We take uncertainty, or the converse property, confidence, to be a first class citizen of HRI. Humans and robots generate different internal representations and use qualitatively different types of processing when interacting with each other [14], giving rise to different types and degrees of uncertainty for each agent. In an interaction with a manipulator robot which responds to a user requesting manipulation of real-world objects through speech, the types of uncertainty are on multiple levels. These levels include: uncertainty over which words are being spoken as they are recognized from an automatic speech recognizer (ASR); the current real-world location and visual properties of the objects from a computer vision processor; which target objects or locations are being referred to (language grounding, or reference resolution), and, recognition of intentions as to what to do with the objects (dialogue act and intention recognition).

In this paper we explore how simple robots can communicate their internal uncertainty to human interaction partners. After providing background on grounding uncertainty in §2, we present a grounding model for HRI which draws on dialogue systems research in §3. The model is designed to allow a robot's internal uncertainty to be communicated to users as it interprets speech, and in an online, fluid manner. In §4 we describe a proof-of-concept implementation of the model in a simple robot whose only communicative channel is the execution of its task actions, and then in §5 describe a study which proves its efficacy in communicating its uncertainty to observers of the robot's behaviour.

2. GROUNDING UNCERTAINTY IN HRI

Communicative grounding in the sense of [4, 3] is the way in which interaction participants build and align their internal representations towards shared information or "common ground". In HRI, grounding is a particular challenge, given the fundamental differences in the internal representations and processing of humans and robots [14]. Given these differences, perceived behaviour of modern robots may not reflect their internal states appropriately, as system designers may feign competences to make interaction less cumbersome, whilst obscuring the robot's real level of understanding. For robots which understand speech, the trade-off between "keeping up appearances" by displaying apparently human-like dialogue behaviour, and on the other hand grounding the robot's uncertainty, can be weighted more towards the former objective. Consequently, users may not be aware how or to what degree a robot has misunderstood them, which could have negative consequences if robots are to engage in active learning through interaction [23, 22].

In HRI, grounding research has focused on communicative problems in language grounding, for example on how per-

HRI '17, March 06 - 09, 2017, Vienna, Austria

ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020214

spective taking and frame of reference differ between robots and humans [13, 17, 16]. Also, the improvement of grounding intentions through increasing the legibility of actions has received attention– [6, 5] show the importance of legible robot motion which is more 'intent-expressive' to users.

We argue in this paper that a robot not only needs to monitor when its internal goal is becoming legible, but the robot should also be able to ground the *degree of commitment to its goal* with the user when it is uncertain, so it may get the required aid, and do so in a fluid and non-cumbersome way.

In simple robots, uncertainty can be grounded in the manner of execution of the task actions themselves, and here, our motivation is similar to [20]'s position paper on how a robot's uncertainty can be indicated through hesitation before moving. Technically, we also aim to achieve the trade-off between 'safety' and speed of movement similar to [19]'s system, however, our notion of safety here, rather than being the avoidance of physical hazards, is the state of internal certainty the robot has about the user's current intention, according to their ongoing speech. We propose this can be achieved through incorporating a fluid grounding mechanism into the robot's architecture.

3. A MODEL FOR FLUIDLY GROUNDING UNCERTAINTY FOR SIMPLE ROBOTS

We propose a communicative grounding model to make a robot's internal uncertainty common ground with the user. We draw on computational models of grounding [27, 21] and recent attempts to incrementalize grounding strategies in dialogue models [8, 7, 10], 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 point in the interaction. From the definition of these functions and the grounding model, internal measures of *understanding* and *confidence* can be calculated on-line by the robot. We explain the elements in turn below.

Statecharts with strength-of-evidence functions.

For our grounding model, we follow work using Harel statecharts [9] for dialogue control in robotic systems by [22, 26]. 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. This allows the robot to estimate which grounding state the robot and human are in concurrently, without having to explicitly represent the Cartesian product of all possible states.

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 conditions or "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 in the triggering conditions on the state transitions– 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 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 in the set of possible goals $Intentions$, given the current utterance u, the robot's state $Robot$ and the current task's state $Task$, as in (1). $Intentions$ is the set of user 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.

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

While the estimated user's goal is continuously being updated 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 [10]. 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, however it must still estimate when $RobotGoal$ is made public by its action, and once ϵ has been reached, 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 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:

$$Intentions := \{i \mid p(RobotGoal \mid i) = 0\} \tag{2}$$

(2) 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

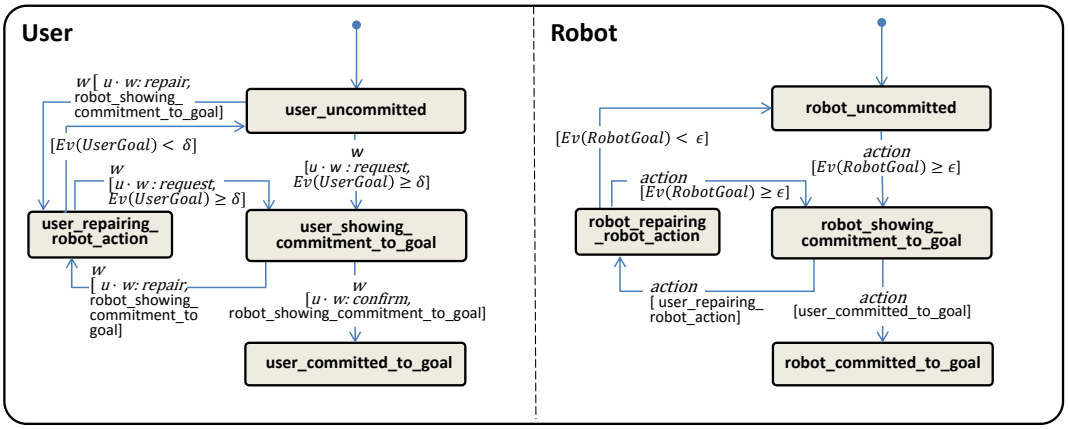

Figure 1: Interactive Statechart as modelled by the Robot. The statechart consists of two parallel, concurrent states, one for each participant. The triggering events and conditions in the transition functions (the directed edges) can reference the other state.

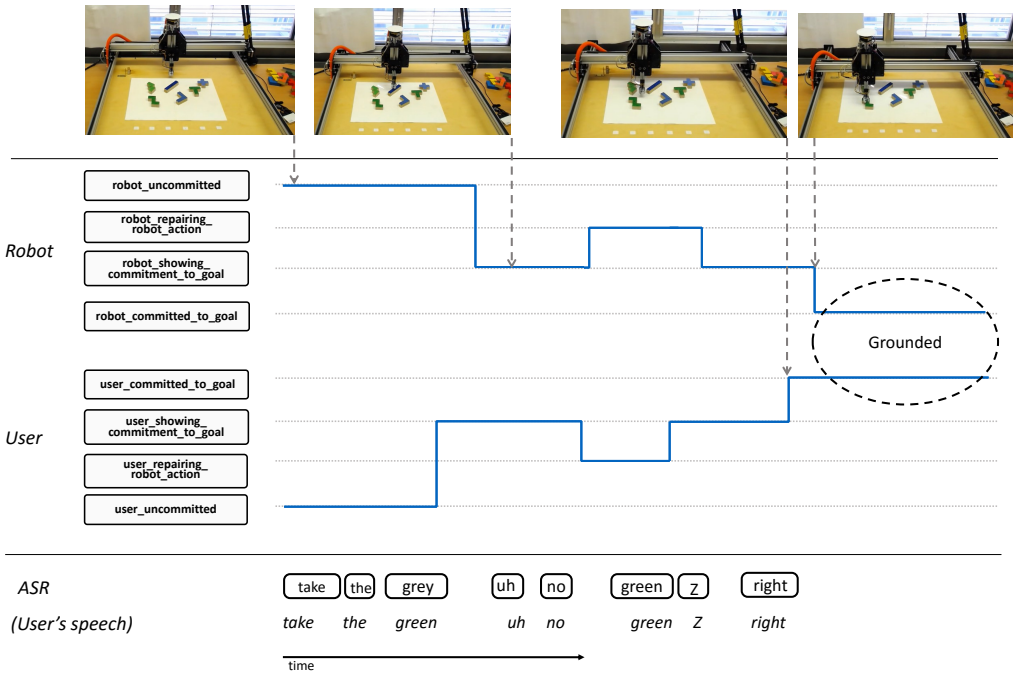

Figure 2: Concurrent User and Robot grounding states during an interaction where an initial mis-recognition of 'green' as 'grey' by the ASR, and confusion over colours in reference resolution where 'grey' gives higher probability to a blue object. The recognition of repair allows the participants to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

such, and consequently determine the interactive dynamics of 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 [28].

For an example of the grounding model in action see Fig. 2. This shows the state dynamics for the concurrent statechart during an interaction with repair.[1] The robot's

[1]Notice how the *Robot* state mirrors, though slightly lags,

ASR error leads to it showing commitment to picking up the wrong object, where upon user-initiated repair interaction begins. From its repair state, the robot changes its goal and re-enters `robot_showing_commitment_to_goal` once its new movement has become legible. In this state, the user's confirmation "right" is interpreted as referring to the current *RobotGoal*, triggering the entry to `robot_committed_to_goal`.

In addition to the word-by-word, left-to-right incremental language understanding ability [15] describe, our system can simultaneously act after each word is processed (like [2]'s model, which is not implemented in a robot), but the novelty is that the robotic action then dynamically updates the context used to understand the following words in the utterance. This is consistent with our approach that the robot's non-verbal actions are treated with the same status as dialogue/speech acts in dialogue and speech act models.

Internal Measures of Understanding and Confidence.

With this grounding model and dialogue act and intention recognition adequate for the robot to know it is in a given grounding state, it is possible to derive measures of understanding and confidence in its own actions, in line with our motivation of making robots monitor their own uncertainty.

Unless being explicitly informed, a robot will not have direct knowledge of its own level of understanding of the user's speech. However, using our grounding model it can estimate this simply as the efficiency in transitioning within its *Robot* state machine from the initial `robot_uncommitted` state to the `robot_committed_to_goal` state for a given goal. So, the level of understanding for a given time period $t_i..t_n$ is simply as in (3), where $S(t_i..t_n)$ represents the states transitioned to over that time.

$$U(t_i...t_n) = \frac{2\times \mid \{s_j \mid s_j \in S(t_i...t_n) \wedge s_j : robot\_committed\_to\_goal\} \mid}{\mid S(t_i...t_n) \mid} \tag{3}$$

This simple measure is based on the following assumption: if the level of understanding is perfect, only two grounding state entries are required to go from being uncommitted to `robot_committed_to_goal` (see Fig .1). In the example in Fig. 2, the understanding measure for this time period would be $\frac{2}{4} = 0.5$, as the total number of state transitions is 4, and there is only one `robot_committed_to_goal` state. Understanding is therefore a measure of grounding efficiency.

Confidence, which is the inverse of uncertainty, is characterized in (4). It is simply the sum of all the strength-of-evidence measures for the user's goal which have been recorded in the state history so far, normalized by the number of state transitions. Confidence in the hypothesized $UserGoal$ is the basis on which the robot's actions are made.

$$C(t_i...t_n) = \frac{\sum_j s_j.Ev(UserGoal) \in S(t_i...t_n)}{\mid S(t_i...t_n) \mid} \tag{4}$$

(4) is a summary of the evidence for the user's goal so far– simply the mean of the strength-of-evidence values that the actions were based on. If there is no uncertainty at all in the evidence about the user's goal when these actions were taken, then this measure would simply average at 1. Lower $Ev(UserGoal)$ values reduce this ongoing average confidence. We use averaging as opposed to a standard product for probability values in Markov chains as we assume confidence is additive, and the length and efficiency of the interaction is better captured by our simple understanding metric in (3).

Showing uncertainty in action execution.

When the robot repairs its action during the user's speech, this is a clear sign of unsuccessful grounding. Legible change of intent shows a lack of absolute commitment to the original goal. However, there are also ways in which simple robots can display the *degree* of uncertainty about their goal and communicate the value of $Ev(UserGoal)$ which their actions are based on. A robot can do this by waiting longer periods for confirmation before acting, as in [20], or by moving more slowly. We explore these two strategies in our experiment in §5 to communicate the robot's current internal confidence. More complex robots could display low levels of confidence through other means such as a confused facial expressions, or requesting clarification for robots with spoken language generation abilities [18].

3.1 Managing uncertainty at different levels with the Incremental Unit framework

To manage the processing in our robotic system which the grounding state machine is housed in, we use the Incremental Unit (IU) framework [25]. The input and output of each module are incremental units (IUs), which are packages of information with a pre-defined level of granularity– e.g. a *wordIU* can represent a single incremental ASR word hypothesis. IUs created in the output buffer of one module trigger downstream processing (and creation of new IUs) in other modules with access to that buffer.

IUs can be defined to be connected by directed edges, called *Grounded In* links, which in general take the semantics of "triggered by" from the source to the sink IU. Grounded In links are useful in cases where input IU hypotheses may be *revoked* (for instance, by changing ASR hypotheses), as reasoning can be triggered about how to revoke or repair actions that are Grounded In these input IUs in downstream modules– see e.g. [11].

To implement the grounding strategies above, we recast the standard Grounded In dependencies: while the output IUs are taken as Grounded In the input IUs which triggered them (from sensor modules to actuator modules), as per standard processing, in our system the reverse will also be true: consistent with the statecharts driving the behaviour, the interpretation of the user's speech is dependent on the robot's latest or currently ongoing robot action. Consequently interpretation IUs can be grounded in action IUs– see the reversed feedback arrow in Fig. 3.

To deal with concurrency issues that this closed-loop approach has, the IU modules coordinate their behaviours by sending event instances to each other, where events here are IU edit messages. The edit messages consist in *ADD*s where the IU is initially created, *COMMIT*s if there is certainty they will not change their payload, and, as mentioned above *REVOKE*s may be sent if the basis for a previously ADDed IU becomes unreliable. IUs also have different temporal statuses of being either *upcoming*, *ongoing* or *completed*, a temporal logic which allows the system to reason with the status of the actions being executed or planned by the robot.

the *User*, by virtue of the fact that it takes time to demonstrate commitment to a given goal with a sufficently strong $Ev(RobotGoal)$ (or legibility [6]).

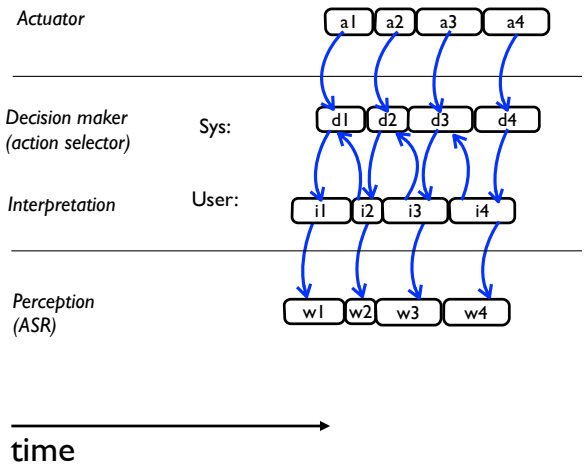

Figure 3: The addition of tight feedback over standard IU approaches helps achieve requirements of fluid interaction and situated repair interpretation. Grounded In links in blue.

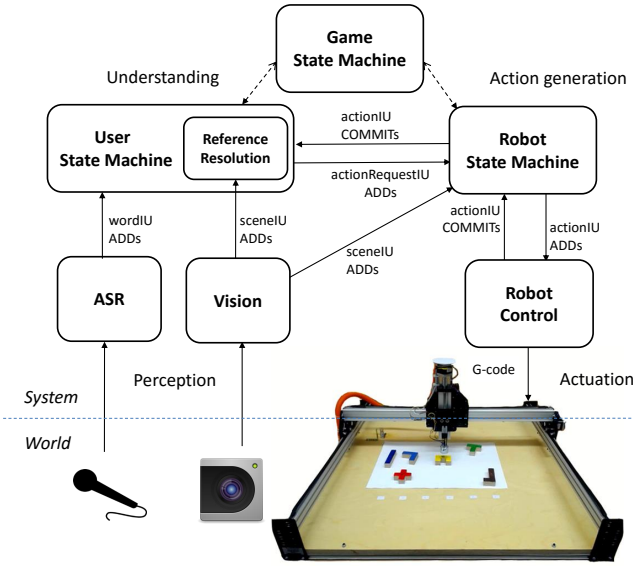

Figure 4: Robot architecture.

4. IMPLEMENTATION IN A SIMPLE PICK-AND-PLACE ROBOT

We implement the above grounding model and incremental processing in a real-world pick-and-place robot, *PentoRob* [10], the architecture of which can be seen in Fig. 4. The domain we use in this paper is grabbing and placing real-world magnetic Pentomino pieces at target locations, however the system is adaptable to novel objects and tasks.

For the robotic arm, we use the ShapeOko2,[2] a heavy-duty 3-axis CNC machine, which we modified with a rotatable electromagnet, whereby its movement and magnetic field are controlled via two Arduino boards. The sensors are a webcam and microphone.

4.1 System components

The robot was implemented in Java using the InproTK [1] dialogue systems toolkit.[3] The modules involved are described below, in terms of their input information or IUs, processing, and output IUs.

Incremental Speech Recognizer (ASR).

We use Google's web-based ASR API [24] in German mode, in line with the native language of our experimental participants. It achieves a Word Error Rate in our target domain of 20%. While it has slightly sub-optimal incremental performance, this did not incur great costs in terms of the grounding we focus on here.

Computer Vision (CV).

We utilize OpenCV in a Python module to track objects in the camera's view. This information is relayed to InproTK from Python via the Robotics Service Bus (RSB),[4] which outputs IDs and positions of objects it detects in the scene along with their low-level features (e.g., RGB/HSV values, x,y coordinates, number of edges, etc.), converting these into *sceneIU*s which the downstream reference resolution model consumes. The Robot State Machine also uses these for reasoning about positions of the objects it plans to grab.[5]

Reference resolution (WAC).

The reference resolution component consists of a Words As Classifiers (WAC) model [12]. Our robot's WAC model is trained on a corpus of Wizard-of-Oz Pentomino puzzle playing interactions. During application, as a referring expression is uttered and recognized, the classifier for each word in the expression is applied to all objects in the scene, which after normalisation, results in a probability distribution over objects. [12] report 65% accuracy on a 1-out-of-32 reference resolution task in this domain with the same features.

User State Machine.

We implement the principal NLU features within the User State Machine module, which runs the $User$ state of the interactive grounding statechart as in Fig. 1. While the statechart defines the transitions between states, their triggering criteria require the variables of the estimated current $UserGoal$ from a set *Intentions*, its strength-of-evidence function Ev and threshold δ to be defined. In our domain we characterize $UserGoal$ as simply taking or placing the most likely object in the referent set R being referred to according to WAC's output distribution given the utterance u so far (e.g. (5) and *Intentions* is simply the distribution over the possible actions of the Pentomino pieces still in play. When the action is $TAKE$, the Ev function is the probability value of the highest ranked object in WAC's distribution over its second highest rank as in (6). We experimented to find a suitable δ which (6) needs to reach as 0.05 [10].

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

[2]http://www.shapeoko.com/wiki/index.php/ShapeOko_2
[3]http://bitbucket.org/inpro/inprotk
[4]https://code.cor-lab.de/projects/rsb
[5]The objects' positions are calculated accurately from a single video stream using perspective projection.

$$Ev(UserGoal) = \text{Margin}(\arg\max_{r \in R} p(r \mid u)) \quad (6)$$

When the action of the goal is $PLACE$ and the system is waiting to recognize a target location as a numbered location based on the incoming $wordIU$s, we use a simpler $Ev(UserGoal)$: this remains 1.0 until 4 seconds after the last robotic action, then decays gradually to 0.05 after 10 seconds. We assigned these bounds after a short pilot study. Given the uncertainty lies in incoming wordIUs for the simple location references, a function based on the ASR confidences could be used, however for our purposes the time elapsed alone was a good enough indicator of uncertainty.

We obtain $UserGoal$ incrementally with a simple NLU method, which uses the results from the WAC reference resolution and the *Robot* and $User$'s current grounding state. Firstly, sub-utterance dialogue act (DA) classification is performed, judging the utterance to be in $\{request, confirm, repair\}$ after every word. The classifier is a simple segmenter which uses key word spotting for $confirm$ words and common *repair* initiating words, and also classifies a *repair* if the word changes the $UserGoal$, else outputting the default *request*. Given the DA classification, the state machine is queried to see if transitioning away from the current state is possible.

If a successful state change is achieved and $UserGoal$ has changed or been instantiated in the process, a new *ActionRequestIU* is made available in the module's right buffer, where the payload is a frame with the dialogue act type, the action type (`TAKE` or `PLACE`) and optional arguments `target_piece` and `target_location`.

Robot State Machine.

The robot state machine module partially consists of the *Robot* grounding statechart in Fig. 1, having access to triggering conditions involving the User's state through ActionRequestIUs in its input buffer. When the $User$ state is `user_showing_commitment_to_goal`, the *RobotGoal* is set to $UserGoal$, and through a simple planning function, ActionIUs are created to achieve it. It sends these as RSB messages to the robotic actuation module and once confirmed, also via RSB, that the action has begun, the ActionIU is *committed* and the robot's *action state*, orthogonal to its grounding state, is set to one of the following, with superstates in brackets:

```
{stationary_without_piece |
moving_without_piece |
moving_to_piece (taking) |
over_target_piece (taking) |
grabbing_piece (taking) |
stationary_with_piece(placing) |
moving_with_piece (placing) |
over_target_location (placing) |
 dropping_piece (placing)}
```

For estimation of its own grounding state, we define the robot's strength-of-evidence function as in (7):

$$Ev(RobotGoal) = \begin{cases} 1 & \text{if over\_target\_piece,} \\ 1 & \text{if over\_target\_location,} \\ 0.5 & \text{if taking} \wedge legible(RobotGoal), \\ 0.5 & \text{if placing} \wedge legible(RobotGoal), \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Here we take *legible* to be a very simple characteristic: that the elapsed time from the beginning of the current action is over 50% of the expected duration of the entire action. The simplistic function (7) embodies the assumption that there is absolute certainty that the robot's goal has been demonstrated when its arm is directly over the target pieces and locations, else if it is legibly moving to these positions, there is some evidence, else there is none. While sufficient for our purposes, more sophisticated approaches can be seen in [5]. In this paper, based on experimentation we set the threshold ϵ which (7) must reach to be 0.5 [10].

Robot actuation module.

The module controlling the robotic actuation of the ShapeOKO arm is a Python module with an Arduino board G-code interface to the arm. This sends RSB feedback messages to the robotic control module to the effect that actions have been successful or unsuccessfully started, and with their estimated finishing time.

5. EXPERIMENT: AN ONLOOKER STUDY ON INFERRING UNCERTAINTY

To test our model of grounding uncertainty in our implementation with users we carry out an onlooker experiment with two versions of our model in PentoRob. Both versions have the same model of internal uncertainty as described above, however they differ as to how they ground this uncertainty with the user in the following ways:

Uncertainty through repair only: only grounds uncertainty by allowing repairs to change its goal and therefore change its action. The maximum speed of movement is the same for every action, and it has a default waiting time over its goal pieces and locations of 1.5 seconds.

Uncertainty through movement: also allows repairs to change its goal but also exhibits its own level of confidence about its goal through its speed of movement and waiting time before acting. Its speed is proportional to its confidence about the piece or target location being referred to, according to equation (4), and its waiting time over target pieces and areas is inversely proportional to this confidence, with a maximal wait time λ seconds before taking the initiative with action a and a degradation down to 0 (no wait) according to (8) below. In this study we set λ to 3 seconds.

$$\text{WaitTime}(a) = \lambda \times (1 - Ev(UserGoal)) \quad (8)$$

The aim of the study was to evaluate our model for its ability to ground system internal measures of understanding and confidence, and whether altering simple parameters of the action execution can communicate the degree of uncertainty. Our concrete hypotheses were:

H1 Our grounding model allows observers to rate the robot's level of understanding reliably without the robot knowing whether it had successfully achieved the user's goal. Due to this measure being based on grounding efficiency in terms of number of state transitions (as in eq. (3)) there should be no difference between the two settings.

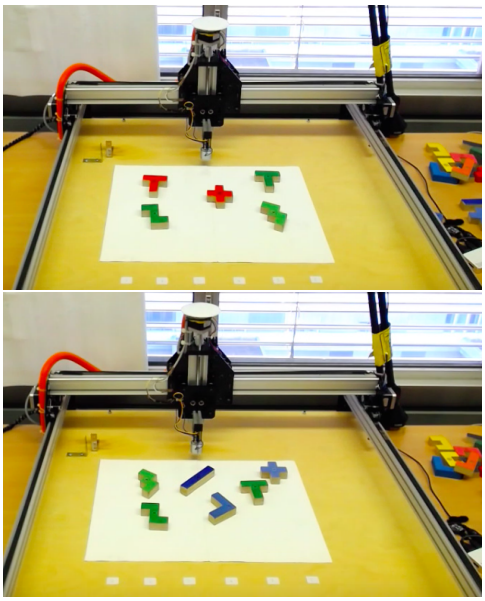

Figure 5: Setting 1 (top) and 2 (bottom).

H2 In our *uncertainty through movement* setting, users would be able to rate the robot's perceived level of confidence more closely to its internal measure of confidence (as in eq. (4)) than in the *uncertainty through repair only* setting. Given the former system maps its confidence onto parameters of its actions, this should be noticeable to users.

5.1 Method

We conducted a video observation experiment with 12 participants who were undergraduate aged students (4 male, 8 female), all of whom understood German (9 native). The participants rated videos of our robot interacting with an undergraduate aged female native German speaker. The instruction giver was not visible in the video as it was captured via a camera placed at her head level, so effected a 'point-of-view' shot where her voice could be heard – see Fig. 5. Each video showed the robot placing a number of pieces as instructed by the instruction giver. Direct interaction with the robot was not used to avoid the difficulty of giving ratings during the interaction in a way that was natural. We understood from the outset that this was not direct HRI, but believe it still has concrete implications for it.

Participants watched and rated four videos in total. The study was a within-subjects design in that participants watched videos in two halves, where they were told the first two videos showed the robot controlled by one system, and the final two showed the robot being controlled by a different system. Both pairs of videos had the same task and starting configurations for two different situations as shown in Fig 5:

1 The initial scene consists of 5 Pentomino pieces: 3 green, 2 red, where there are distractors in terms of shape, with two T-shaped pieces in each colour (introducing human-like uncertainty). The green Z and M pieces according to our computer vision module are confusable (introducing robot-specific uncertainty). All 5 pieces are to be placed in specified numbered boxes at the bottom of the screen.

2 The initial scene consists of 6 Pentomino pieces: 3 green, 3 blue. While there are no distractors in terms of shape, our computer vision module often confuses green and blue. The first 3 pieces, all green, are to be placed in the specified numbered boxes.

For all four interactions, the words recognized, dialogue acts classified, grounding state changes, robotic action requests and callbacks, understanding and confidence measures were logged by our system. These were used to test our hypotheses.

Participants watched both situations in the same order but with a randomly selected order of the system presentations. The experimenter, not co-present, paused the video after each piece was placed in its target location. During the pause, participants would rate the robot's behaviour during the clip they had just observed under two dimensions of Understanding and Confidence on a scale from 1 to 7, writing them down on a form. Before the experiment the meanings of the two ratings dimensions were explained as follows:

- Understanding- to what degree did you feel the robot understood what it had to do?
- Confidence- to what degree did you feel the robot had confidence in its decisions to act?

The purpose of the clip-by-clip rating, rather than using overall interaction-final impressions, is that we wanted to get close to the live monitoring of understanding and uncertainty required in performing joint tasks with robots, in line with our motivation in §2. We separate understanding and confidence ratings in order to see how well our model's internal measures of both dimensions correlate with the users' perceptions, emphasizing these as separate aspects, rather than conflating the ratings into one.

5.2 Results

One participant did not follow the instructions, so his results were discarded. For the remaining 11 participants, we performed Z-score normalization over their ratings, and then correlated these Z-scores with the internal understanding and confidence ratings for each of the 8 clips.

For understanding, the internal values of the system's understanding for each clip as given in (3) correlated very strongly with user ratings of perceived understanding for both systems. The *uncertainty through movement* setting correlating linearly (Pearson's R) very strongly at 0.921 ($n=88$ for this and all results below, $p<0.0001$), and in a non-parametric Spearman's rank correlating strongly at 0.832 ($p<0.0001$). Our *uncertainty through repair only* system's internal understanding for each clip also correlated strongly linearly and non-parametrically (Pearson's $R= 0.841$ ($p<0.0001$) and Spearman's rank $= 0.765$ ($p<0.0001$))– see Fig 6 (top). According to a Fisher r-to-z transformation, the two Spearman's rank correlations were not significantly different ($z=1.210$. $p=0.226$). This result gives evidence for H1 being the case in our simple domain.

For confidence, the inverse of uncertainty which we are investigating, there is a different pattern, in that while the *uncertainty through movement* setting's internal confidence correlates strongly with the observer judgements (Pearson's $R=0.726$, $p<0.0001$; Spearman's rank $= 0.704$, $p<0.0001$), the *uncertainty through repair only* setting only does so moderately (Pearson's $R=0.398$, $p<0.001$; Spearman's rank $=$

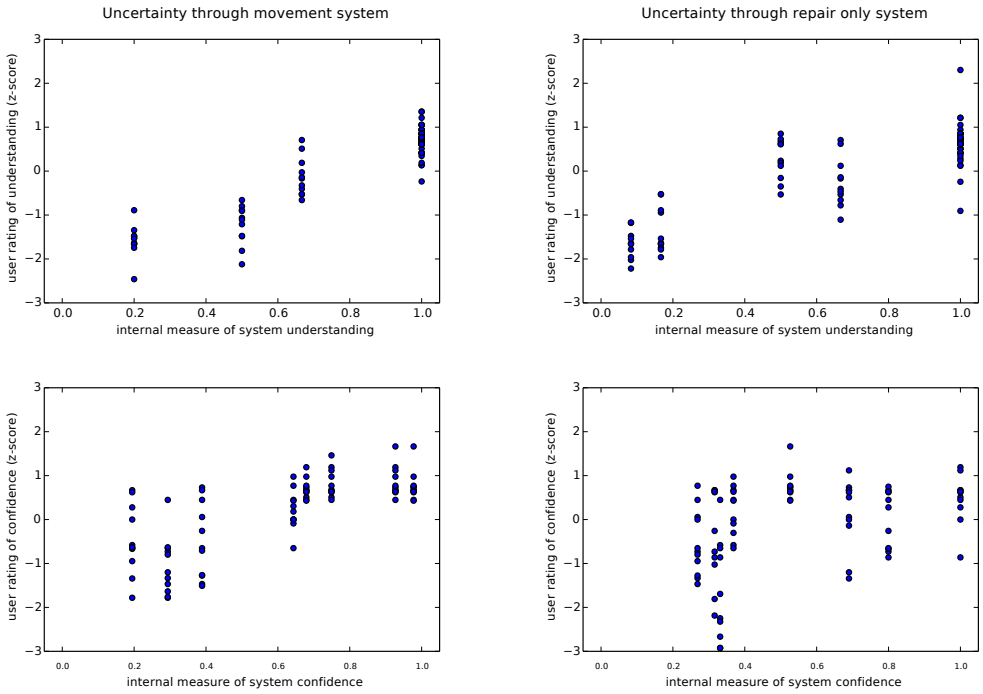

Figure 6: Top: plot of the internal understanding measures of the system (x-axis) vs. observed rated level of understanding by participants (y-axis). Bottom: internal confidence measure of the system (x) vs. observed rated confidence (y). Left-hand plots are uncertainty through movement system, right-hand repair only.

0.4018, $p<0.001$)– see Fig. 6 (bottom). A Fisher r-to-z transformation on the Spearman's rank coefficients shows the *uncertainty through movement*'s correlation was significantly stronger ($z=2.93$, $p<0.01$). This gives some evidence to supporting H2, that the uncertainty is more easily groundable through mapping it onto real-valued parameters in the robot's movement.

5.3 Discussion

We can tentatively conclude that uncertainty can be recognized in the *movement* system more reliably than the *repair only* system, given the stronger correlation between internal confidence values and the ratings of confidence. The correlation suggests people can not only perceive the presence of uncertainty, but its degree. Given the *repair only* system 'hides' its uncertainty in its action, this result is what we would expect, and serves as evidence that one can systematically ground levels of certainty through simple parameter changes in robotic movement alone.

There are several limitations to the study. Firstly, due to the fact video-recorded interactions were rated, one can only tentatively draw conclusions for actual HRI. It is difficult to design scenarios where users could rate the robot's levels of understanding and confidence online while they carry out a time-critical task with the robot, however other measures of uncertainty derivable from sensory information could be used. Various options such as gaze data or acoustic measures from the user's speech could be used in a further study. A future study could also be designed where uncertainty is contingent on some aspects of interactive success for the user, such as perceived rapport and trust with the robot.

Secondly, the study could be scaled to a bigger participant pool. Using crowd sourcing could potentially be useful for informing our grounding model with enough data. In spite of this limitation, from our proof-of-concept we can draw positive conclusions that on-lookers can perceive uncertainty with some reliability.

6. CONCLUSION

We have presented a grounding model for HRI and show how, when implemented in a simple robot, it can communicate a robot's levels of uncertainty, in terms of understanding and confidence, purely through the robot's task actions. While parameterization of robotic movement by confidence levels has been used in previous studies for giving extra time for sensing [19], here we do this for communicative purposes.

Our study has some pleasing potential consequences for HRI with simple robots in general. Any system with adjustable parameters in its actions could in principle adopt our grounding model. The choice of the 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 adapted to novel cases.

Acknowledgments

We thank the HRI reviewers for their helpful and detailed comments. We thank Casey Kennington, Oliver Eickmeyer and Livia Dia for contributions to software and hardware and Gerdis Anderson for help running the experiments. 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).

7. REFERENCES

[1] T. Baumann and D. Schlangen. The inprotk 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*. ACL, 2012.

[2] T. Brick and M. Scheutz. Incremental natural language processing for hri. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 263–270. ACM, 2007.

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

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] C. Kennington, S. Kousidis, T. Baumann, H. Buschmeier, S. Kopp, and D. Schlangen. Better driving and recall when in-car information presentation uses situationally-aware incremental speech output generation. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2014.

[12] C. Kennington and D. Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the Conference for the Association for Computational Linguistics (ACL). ACL, 2015.

[13] T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010.

[14] 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.

[15] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symposium on Language and Robots*, 2007.

[16] C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL, 2012.

[17] C. Liu, J. Walker, and J. Y. Chai. Ambiguities in spatial language understanding in situated human robot dialogue. In *AAAI Fall Symposium: Dialog with Robots*, 2010.

[18] M. Marge and A. I. Rudnicky. Towards overcoming miscommunication in situated dialogue by asking questions. In *AAAI Fall Symposium: Building Representations of Common Ground with Intelligent Agents*, 2011.

[19] J. Miura, Y. Negishi, and Y. Shirai. Adaptive robot speed control by considering map and motion uncertainty. *Robotics and Autonomous Systems*, 54(2):110–117, 2006.

[20] A. Moon, B. Panton, H. Van der Loos, and E. Croft. Using hesitation gestures for safe and ethical human-robot interaction. In *Proceedings of the ICRA*, pages 11–13, 2010.

[21] T. Paek and E. Horvitz. Grounding criterion: Toward a formal theory of grounding. Technical report, MSR Technical Report, 2000.

[22] 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.

[23] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1613–1622. ACM, 2010.

[24] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. Your Word is my Command: Google Search by Voice: A Case Study. In *Advances in Speech Recognition*. Springer, 2010.

[25] D. Schlangen and G. Skantze. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1), 2011.

[26] 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.

[27] D. R. Traum. A computational theory of grounding in natural language conversation. Technical report, DTIC Document, 1994.

[28] 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.