Grounded Situation Models for Robots: Where words and percepts meet

Nikolaos Mavridis
Cognitive Machines Group, MIT Media Laboratory
20 Ames Street, Cambridge, MA 02139-4307
nmav@mit.edu

Deb Roy
Cognitive Machines Group, MIT Media Laboratory
20 Ames Street, Cambridge, MA 02139-4307
dkroy@media.mit.edu

***Abstract*— Our long-term objective is to develop robots that engage in natural language-mediated cooperative tasks with humans. To support this goal, we are developing an amodal representation and associated processes which is called a *grounded situation model* (GSM). We are also developing a modular architecture in which the GSM resides in a centrally located module, around which there are language, perception, and action-related modules. The GSM acts as a sensor-updated "structured blackboard", that serves as a workspace with contents similar to a "theatrical stage" in the robot's "mind", which might be filled in with present, past or imagined situations. Two main desiderata drive the design of the GSM: first, "parsing" situations into ontological types and relations that reflect human language semantics, and second, allowing bidirectional translation between sensory-derived data/expectations and linguistic descriptions. We present an implemented system that allows of a range of conversational and assistive behavior by a manipulator robot. The robot updates beliefs (held in the GSM) about its physical environment, the human user, and itself, based on a mixture of linguistic, visual and proprioceptive evidence. It can answer basic questions about the present or past and also perform actions through verbal interaction. Most importantly, a novel contribution of our approach is the robot's ability for seamless integration of both language- and sensor-derived information about the situation: For example, the system can acquire parts of situations either by seeing them or by "imagining" them through descriptions given by the user: "There is a red ball at the left". These situations can later be used to create mental imagery and sensory expectations, thus enabling the aforementioned bidirectionality.**

I. Robots, Language and Modularity

As robots grow in ability and complexity, natural language is likely to assume an increasingly central role in human-robot interaction. Our current work is part of a larger effort to develop conversational interfaces for interactive robots ([3], [6], [11], [8]). Robots that understand and use natural language may find application in entertainment, assistive, and educational domains. Such interactive robots are prime examples of systems where integration of numerous technologies in complex ways is required, and thus well designed modularity is necessary. One of the main challenges that one faces when designing such a system, is interfacing perceptual/motor with speech modules: existing natural language processing (NLP) systems cannot simply "plug and play".

One historical reason behind this incompatibility is that the development of NLP and robotics have proceeded with relatively little interaction. NLP deals with the *discrete*, symbolic world of words and sentences whereas robotics the *continuous and stochastic*: one must confront the noisy, uncertain nature of physically embodied systems with sensory-motor grounded interaction. Current computational models of semantics used in NLP are variants of "dictionary definitions", essentially structured networks of word-like symbols. It is impossible to directly apply these approaches in any principled way to endow robots with linguistic skills since the underlying theories of semantics in symbol-based NLP provide no appropriate "hooks" for action and perception [9].

We posit that an important step towards bridging the language-robot divide is to develop new knowledge representations that facilitate cross-modal interoperability. Motivated by these concerns, we have developed a *grounded situation model* (GSM) which lies at the center of our proposed modular architecture. The GSM acts as a sensor-updated "structured blackboard", that serves as a workspace with contents similar to a "theatrical stage" in the agent's "mind", which might be filled in with present, past or imagined situations. It provides connection points to both speech-processing (through discrete verbal categories) as well as to perceptual/motor subsystems (through continuous / stochastic descriptions).

When building conversational systems, one must select a subset of human language to be produced/comprehended. Although we may eventually want robots that can converse about a range of abstract topics, a natural starting point is to develop means for robots to talk about their immediate physical and social context. This parallels the development of semantics from concrete to abstract domains in children [10].

II. Challenges in Cross-Modal Representation

A central problem in connecting language and perception is the potential for *mismatched levels of specificity*. For example, the descriptive phrase "there is a cup on the table" and the visual observation of the cup will lead to consistent knowledge but with very different levels of specificity. The linguistic description does not provide information about the size, orientation, and color of the cup, and provides only bounds on its location (it's *somewhere* on the table). Visual perception in contrast will provide far more detail. How is a robot to translate between such varying sources of knowledge? More generally, sensory, motor, and linguistic sources differ in levels of specificity/ambiguity, yet must be aligned sensory verification, cross-modal belief propagation, and action.

In our approach, a grounded situation model serves as a mediating amodal representation that connects sensory-derived percepts with linguistic structures as well as action parameters. The GSM is amodal in the sense of being neither a viewer-dependent 2D image, nor an absolute 3D spatial model (it even includes invisibles such as the beliefs of others), nor an unstructured list of language-like propositions describing beliefs. It is a representation accumulating information coming in from multiple modalities (vision, touch, proprioception, language), which has a structure analogous to the situation the robot is embedded in or is imagining.

The overall GSM design was driven by **two desiderata**: first, "parsing" situations into ontological types and relations that reflect human language semantics, and second, allowing bidirectional translation between sensory-derived data/expectations and linguistic descriptions. The GSM design was then further specified by a set of **behavioral goals** for a manipulator robot under development in our lab [8] which we anticipate will serve as a basis for developing more sophisticated linguistic abilities in future work. The robot's world consists of a table top on which various objects are placed and manipulated. We set the following behavioral goals:

Answering questions about the present physical context: such as, "What color is the one on the left?" (a question about objects on the table).

Quantifying and expressing confidence in beliefs: Thus, when asked about the location of an object that it hasn't seen for a while, it might answer, "Probably at the left, but maybe not", expressing uncertainty since the object might have been moved while the robot was not looking.

Respond to spoken requests: such as "Look to the left" or "Hand me the small red one" with situationally appropriate motor actions.

Imagining situations described through language: so that the robot can understand commands such as "Imagine an object at the left", or descriptions such as "There is a small object at the right". Such speech acts must be translated into representations that may later be related to sensory input.

Remembering and resolving temporal referents: so that the robot can keep track of salient past events and talk about them. This would enable the robot to answer questions such as "What color was the object that was at the center of the table when the red one appeared?".

The GSM and associated cross-modal belief update and language processing algorithms we have developed enable each of these situationally-grounded linguistic behaviors.

III. Relation to Previous Work

The notion of a situation or mental model has been proposed by cognitive psychologists ([12], [5]) in this spirit, but most such work focuses only on the connection between mental models and language. For example, Johnson-Laird provides an elaborate overall account, but mainly focuses on language understanding and inference making. Most behavioral experiments reviewed in [12] probe the structure of human mental models, and assess the relevance of their prime "dimensions" (space, time, protagonist etc.). However, again most of these experiments involve only language (story understanding). In contrast, in our work, the processes providing sensory-motor grounding of situation models are also specified.

Below, a short review of existing robots with conversational abilities is given. The approaches taken towards connecting language with perception and action will be briefly examined, as well as their behavioral repertoires.

In [3], the authors propose a natural-model semantics which they apply to the interpretation of robot commands, in two robotic aids for the disabled. As the above robots are not equipped with perceptual systems, a model of the environment consisting of 3D object positions and properties is entered manually into a knowledge base. Total confidence and complete knowledge is assumed. In [6], a Bayesian network interconnects visual to verbal information about objects. The system can interpret gestures, and includes visual attention mechanisms, but can only handle action requests. In [11], an occupancy map built by range sensor data plays part of the role of a GSM. Objects are individuated, and spatial relations are exploited in answering questions and interpreting action requests. The robot Leonardo [1] uses a cognitive architecture built on top of the *c5* codebase, an extension of *c4*[2]. A centrally located "Belief system" module interconnects speech, vision and action. Hierarchical structures called "percept trees" classify sensory inputs to "snapshots" which are fed to the belief system, which decides whether to create or update beliefs about objects and their properties. The system models not only robot beliefs but also human beliefs, through representations having the same structure (which our system accomplishes by using embedded GSM's). Also, the system models attentional/referent focus, which our system does not.

However, our system has **three novel abilities** compared to all of the above mentioned systems. These were already explicated in the behavioral specification given in the previous section, under the headings: "Quantifying and expressing confidence in beliefs", "Imagining situations described through language" and "Remembering and resolving temporal referents". Through the second of these, objects instantiated through language can be referred to, acted upon, and can also be visualised in mental imagery (thus enabling bidirectionality between language and vision, the second design desideratum). The GSM has enabled our system to attain these abilities.

IV. Embodiment

The robot is an arm with 7 degrees of freedom, equipped with force feedback actuators, a gripper with force-sensitive touch sensors integrated into each finger tip, joint angle encoders, and dual cameras mounted around the gripper.

A layer of low-level software consists of numerous modules that run on a set of networked computers running Linux. Front end visual processing is carried out by the following modules: camera capture, color-based segmentation, face detection , and 2D region detection and tracking. Currently only one of the robot's cameras are used for visual perception. The output of the visual subsystem is a stream of detected faces and regions

at 20 frames per second. Low-level motor control is based on PID controllers. At a higher-level, motor primitives such as "pick up" have been coded as parameterized action schemas. In addition to looking at, grasping, and moving objects, the robot can also weight objects by lifting.

The robot's environment consists of a table populated by objects, and a human interacting with the robot and the objects, who is standing near the edge of the table. The robot's purpose is to serve as a "conversational helping hand".

V. GSM Representations

As mentioned before, the overall GSM design was driven by two desiderata, and then was further specified by the behavioral goals. In order to fulfill the **first desideratum**, the GSM should reflect the natural hierarchy of agents, bodies, body parts, and properties that is implied in human languages.

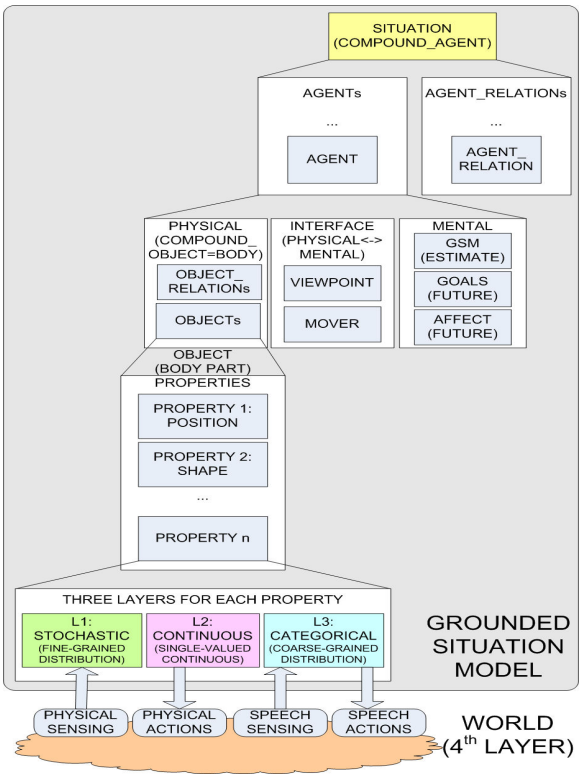

Fig. 1. Hierarchical structure of the GSM

Thus, at the highest level, the situation model consists of agents and relations among them. Any inanimate object is modeled as being potentially agentive. The GSM contains all the information the robot has acquired about itself and its environment, in the form of one agent structure for the self, another agent structure for the human, and one more for each inanimate object. Every agent structure breaks down to a three-part representation, consisting of the body (physical realm), the mind (mental realm), and the interface (between physical and mental). The body consists of simple objects (body parts) and spatial object relations. The mental realm is represented by a recursively *embedded GSM* associated with each body, enabling the robot to maintain a "theory of mind" about other entities. The interface consists of the specifications of contact points between mental and the physical, i.e., "sensory entry" and "action output" parameters. At present, the only element of the mental realm that is fully functional is the ability for the robot to model another agent's visual point of view. In this paper, we will focus on only the physical realm of the GSM since the motivating behaviors defined earlier deal with physically-grounded semantics. In future work in which we are planning to develop socially-grounded linguistic behaviors, the mental realm of the GSM will become crucial.

Objects in the physical realm bottom out in clusters of properties. For example, the representation of a ball will bottom out in a set of properties that model the look, feel, and location of the ball. In order to fulfill the **second desideratum**, i.e. allowing bidirectional translation between sensory-derived data/expectations and linguistic descriptions, each property is encoded by a set of *three layers* of linked representations:

Layer 1 (L1) maintains **stochastic** representations of properties, suited for sensory measurements. Let us assume that we have acquired multiple noisy measurements of the position property of a particular object by computing the centroid of a tracked visual region over time. We would like to encode our knowledge of the position in a summary form, which should give little weight to non-persistent outliers, which should not cause any significant loss of measurement resolution, and which should still retain an ability to remember the spread of sensed values and our confidence in them. We should also be able to lower our confidence when measurements become outdated, and furthermore actively drive the acquisition of more current sensory information whenever required. To satisfy the above requirements, it would be reasonable to represent position property as a stochastic variable, through a probability distribution (e,g., a continuous parametric form, or as we have implemented it, a discretized histogram).

Layer 2 (L2) maintains **continuous single-valued** encodings of properties, suited for use as action control parameters. Consider a scenario where we want to execute an action which requires the position of the object as a parameter. For example, we might want to use the object as a target for a lift motor routine. The stochastic distribution must be sampled in order to guide action. In our current implementation, the continuous layer may be generated by selecting the maximum density point from L1. A second motivation for maintaining L2 is to support simulation based reasoning. To simulate interaction of objects over time, a single value for properties such as size, orientation, and position leads to computationally tractable physical simulation, whereas stochastic representations would be far more complex and time-consuming to manipulate.

Layer 3 (L3) maintains **discrete, categorical** encodings of properties, suited for interfaces with natural language. Consider the scenario of asking the robot where an object is. To respond, a verbal spatial category must be produced and communicated by the robot (e.g., "at the left"). We need to be able to provide a single discrete value corresponding to the verbal category chosen, or better yet, provide a probability distribution over multiple spatial categories. This is what the

categorical layer, L3, accomplishes. It represents a property as a distribution over a number of verbal categories (while in L1 we had a fine-grained distribution over sensory-derived measurements). For example, we might have "left", "right", "center" in the case of position, or "red", "blue" in the case of color etc. We have suggested that the categorical layer is motivated by the need for production of verbal descriptions. It is equally motivated by the converse need, translating from verbal descriptions to property representations: the robot might be told that "there is an object at the center". If there is total confidence in the linguistic source, the robot can represent the information as a discrete distribution over the categories, with $P(location = center) = 1$ and all other probabilities zero.

To summarize, the GSM represents a situation as a hierarchy of objects in the physical realm linked, optionally, to a mental realm. The realms bottom out in a linked three-layered representation comprising stochastic (L1), continuous (L2), and categorical (L3) levels. A particular configuration of the GSM represents a *moment* in time – a snapshot of the state of the situation. An *event* is a structure providing landmarks on the sequence of moments. It consists of an event type ID, start/end time indices, and a list of participants (agents or bodyparts).

VI. GSM Algorithms

The GSM is used for two basic purposes, belief maintenance and action control. The constituents of the GSM hierarchy, and each object's layered property representation is created and maintained using update procedures described in this section. Conceptually, we treat the robot's external physical world as a fourth property layer ("L0") that interacts with L1 via sensory-motor processes. In this conceptualization, perception is seen as a bottom up process caused by the physical world and propagating through layers of representation and hierarchical structure. Action, on the other hand, is seen as top down, starting with encoded new desired states in the hierarchical GSM which are "pushed down" to effect change in the physical environment to effect desired change.

A. Situation Model Updating

We will use the updating of an object's position property as an illustrative example (Figure 2). We will adopt the notation Ci/Rj for the columns/rows of this figure. Pseudo-code is available online at http://www.media.mit.edu/∼nmav.

Sensory information updates of the stochastic layer: Given no information (sensory or linguistic) about the position of an object, we are faced with a choice: what should be the initial probability distribution on positions? In our robot's particular case, objects are assumed to be on the table - thus the object's location must be bounded in space defined by the surface of the table. As a first approximation the a priori probabilities of unseen object positions are spread uniformly across the table

Now let us suppose that an estimate of the position of an object is generated by the visual system. How should the probability distribution of the stochastic layer be updated? We have chosen to calculate the new distribution as the weighted sum of the old distribution with a rectangular envelope centered at the new measurement. In the limiting case, this envelope consists of only one bin, namely the bin which contains the new measurement. The weight and envelope can be adjusted to fit the noise and temporal characteristics of the measurement.

As a general rule, we assume that over time, knowledge becomes less reliable without information refreshing. For example, let us suppose that sensory information is not currently available about an object's position because the robot is not looking at it. Over time, the robot's confidence in knowing the position of the object should decrease (someone might move it while the robot is not moving, etc.). To model this confidence decay in L1, we use a diffusion process. The new value of each element of the position distribution in L1 is given by the weighted sum of its old value with that of its neighbors within a pre-specified neighborhood. The expected rates of change dictate the settings of the weights. Diffusion parameters are set separately for each property modality. Color and shape beliefs are diffused much more slowly since they are far less likely to shift over time (but color, will, for example, shift in perception as lighting conditions change).

For example, in C1 an object has been visible for some period of time and is still visible. In R2C1, the resulting distribution has become very sharp after the object was stable and visible for some time - in fact it consists of a single bin (under the cross-hair). The robot knows where the object is with certainty. In contrast, in R2C2 and R2C3, the robot's head has looked away, and the object has not been visible for some time (C2), and even more time (C3). The diffusion process has taken over and spread out the distribution.

Speech-derived information updating the categorical layer: The categorical layer consists of a distribution over a set of verbal positional categories ("right", "center" etc.) . If the robot receives information that the property value "left" was given through speech for the object under consideration, then the robot sets P("left") = 1 while the probability of other categories is set to zero. If such information is absent, it has two choices. Either the system can assume an empirical prior over the verbal categories, or it can use a non-informative uniform prior, and again we have chosen to implement the latter. In C4, the position is specified by the verbal information "...at the center". Thus, in R4C4 we have P("center")=1 while P(other category)=0. In contrast, when no spatial information is given through speech we get a uniform pdf (see R4C5).

The stochastic layer (L1) feeds the categorical layer (L3): Whenever information enters the GSM (either via L1 or L3) or when a change occurs due to diffusion, the three layers must be updated in order to ensure cross-layer consistency. If the change has occurred at the stochastic layer, then update information feeds the categorical and vice-versa. The continuous layer is always fed via the stochastic. The stochastic layer contains more specific information than the categorical, and thus the forward feeding process is many-to-one and straightforward. Each property has an associated classifier. The classifier maps continuous sensory-derived values to categories. The classifier could in principle be implemented by any algorithm, such as SVM's, neural networks, etc. For simplicity

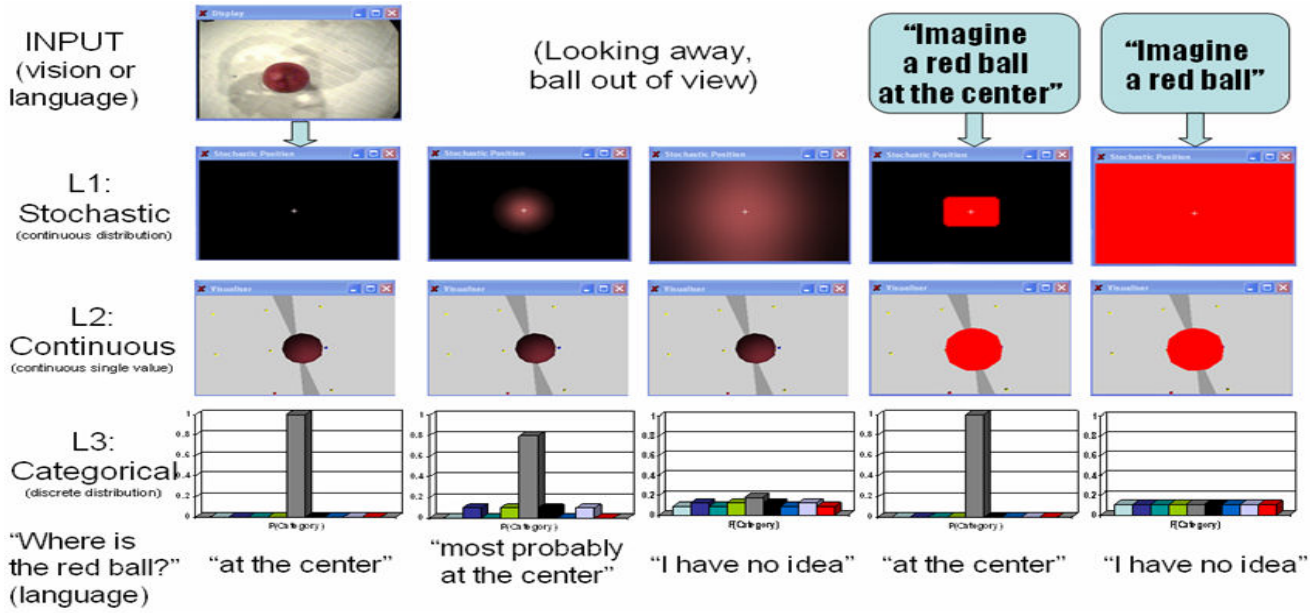

Fig. 2. GSM layer contents: objects instantiated through vision, persistent objs, objs instantiated on the basis of speech.

we have implemented nearest neighbor classification around predetermined centers (for more refined spatial models, see [7]. Initially, all verbal categories are assigned zero probability. Then, each bin of the stochastic layer is considered. The probability of the verbal category associated with the center of the bin (according to the classifier) is increased by the amount of probability that corresponds to the bin of the stochastic layer that is under consideration. As a result, we obtain probabilities of verbal categories as the sum of the probabilities of their corresponding bins in the stochastic layer. The narrowly-spread stochastic distribution in C2R2 has created the narrowly-spread categorical in R4, and the wide-spread of C3R2 leads to the one in R4.

The categorical layer feeds the stochastic layer: If we try to invert the previous transformation, a one-to-many mapping results. In order to achieve uniqueness, we enforced the constraint that the stochastic layer bins that correspond to the same verbal category should be equiprobable. Originally, the stochastic layer elements are all assigned zero probability. Each category is considered in turn. The elements that correspond to the category under consideration are marked, and the probability of the category under consideration is spread equally among them. In C4, when R4 is fed to R2, the area corresponding to the bins whose centers would be classified as belonging to the "center" spatial category is filled with equal probability. In C5, each category corresponds to a rectangle such as the one shown in the C4R2 for "center", thus the whole of C5R2 is equiprobable.

Translation from the categorical layer to descriptive speech: Consider the case of R4C1. Unquestionably, as P("center") approaches unity, the robot can describe its position as "at the center". But things are less clear in C2 and C3. There, according to a decision tree created with preset thresholds on the probabilities of the three most highly probable categories and the entropy of the distribution, numerous different resulting verbalizations occur. For example, if P(most likely category) > 0.7 and < 0.9, then we get "most probably at the <spatial category>" (C2). As a further example, when the distribution is almost equiprobable as quantified by its entropy, then we get "I have no idea" (C3). The decision thresholds that were currently arbitrarily set, but could be empirically learned.

The stochastic layer feeds the continuous layer: Here, we are seeking a single representative value for the distribution of the stochastic layer. Here we have chosen the statistical mean (and not mode), as no bimodals distributions arise. In our example, all the distributions shown in R2 share the same mean, i.e. the center of the table. Thus, if the robot were to look at an object, in both cases the same target fixation point would selected to guide the motor system.

B. Temporal model construction

Moments are created in a straightforward manner. The current GSM state is copied and time-indexed. In the current implementation, moments are stored forever. For round-the-clock operation, some form of memory filter / consolidation must be added, but this has not been explored yet.

Events are created and continuously updated based on the current and previous moments, through *event classifiers*. Events might be instantaneous or optionally encode duration. For example, when velocity (approximated by positional differences) rises above a preset threshold, it triggers the creation of the instantaneous "start moving" event. In contrast, an event having duration is first created, and then its end time is continuously updated as long as the event holds (e.g., the "is moving" event has duration equal to the period that an object is observed in motion). The event classifiers are again very simple in this first prototype. However, they are plug-in replaceable by more complicated trainable classifiers, utilizing hidden markov models, stochastic context free grammars etc.

C. Spoken Language Processing

We use the Sphinx 4 continuous speech recognizer to convert incoming speech into text transcripts. Keyword-based semantic frames are used to parse speech transcripts.

After passing through the recognizer, utterances are then classified in terms of their speech act type: questions ("Where is...", "What color is...", etc.), action requests ("Touch...", "Look at...", etc.), information about the situation ("There is..."), viewpoint-dependent actions ("Touch the one on my left", etc.). Tense information (present/past) is also extracted.

Object reference resolution: Reference to an object can be resolved to any part of the three main agents of the situation model: me (robot), you (human partner) and others (objects on the table). It might be resolved to one, many, or no such parts. It might be referred to either through "part names" (my head, your arm) or through "definite descriptions" (the small red one, the large ones at the top). The simple objects (body parts) of the robot and the user are usually referred to by part names, while the objects on the table (others), are referred through attributive descriptions. Consider the question "Where was the blue object when your head started moving?". In this case, both part names ("your head") as well as attributive descriptions ("blue object") are used, one for each object referent. The robot might either ask a disambiguating question (supplemented with deictic pointing by the robot) until it narrows down to a single referent, or it might carry out the requested action in succession on all the referents fitting the description. The course of action taken depends on the action requested, on whether it can accept groups of objects as arguments, and also on whether plural or singular was used. For example, assume that three objects are on the table - a small red sphere, a large red sphere, and a blue sphere. If the human requests "Touch the red one!", the robot will answer "do you mean this one or that one?" while pointing to the two red spheres in succession. Then, the human can narrow down by saying "Touch the small red one". Else, if the human had requested "Touch the red ones!" then the robot would touch both red spheres in succession. These behaviors are selected via a decision tree which is driven by the number of matching referents, the plural or singular number, and the possibility or not of carrying out the specified action with multiple referents.

Temporal reference resolution: In the case of questions or actions involving the past, temporal references must also be resolved. Their existence is detected through the keyword "when". After "when", an event description involving object referents should follow. Consider the meaning of the phrase "when your head started moving". This amounts to going back in time until a matching event is found, and resolving to the time of this event. The referrable event classes can be found in the appendix. The participants are either the objects, the user, or the robot itself. In the case multiple candidate events are found, only the most recent is reported. If the requested action is not a question, then one further condition should hold: the referred object should still exist, so that it can be acted upon.

VII. Modular Implementation Architecture

The software implementation of the GSM and its associated algorithms is organized around a set of modules (Figure 3):

Situation Model: the module holding the current state of the GSM. This module broadcasts its contents to other modules over network connections, and processes requests for object creation/deletion/updates from the modality-specific modules in order to maintain the GSM object hierarchy.

Visor, Proprioceptor, Imaginer (modality-specific modules): Each of these modules propose changes to the current GSM state, which is broadcast from the Situation Model. Visor listens to the visual stream, while Proprioceptor connects to the robot's position and force encoders. Imaginer processes linguistic descriptions about real or imaginary situations. Via the imaginer, the situation model can now be fed not only through the senses but also through linguistic descriptions, and be later updated by either.

Inquirer: Provides the capability of answering simple questions about the present, such as "What color are the objects at the left?", and also of acting on objects described through the present: "Touch the blue one". Carries out object referent resolution, and requests appropriate actions.

Rememberer: Through this module, the past becomes accessible. It uses the "event" lists in order to resolve temporal referents such as "when the red one appeared" etc. Then, and after also having resolved the object referents at the right times, it feeds the appropriate "moments".

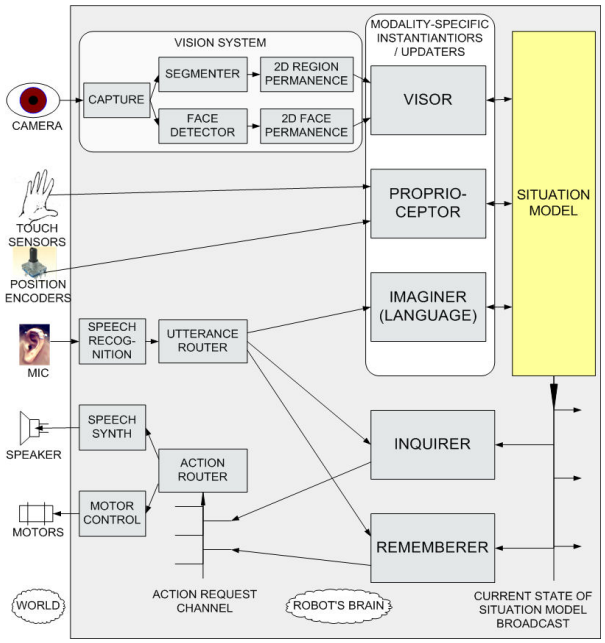

Fig. 3. Modular implementation architecture.

The primary data structure that is exchanged among modules is the present state of the situation. Changes to this are proposed by the various sensory-specific modules (visor, imaginer etc.), which then drive both language and motor actions (through the inquirer and the rememberer). Moments and events are only held in the rememberer.

VIII. Current Performance

The implementation of the GSM and its associated algorithms may be evaluated at various levels. Although none of the conversational robotics papers that we refer to include quantitative evaluations, we could attempt quantifying the performance of our system's components (speech recognition, vision etc.). However, the main focus of this paper has been the design of the representations, the algorithms and the architecture to operationalize the GSM concept for a robot. To evaluate this more holistic goal, we believe a functional (behavioral) evaluation of the complete system is more appropriate.

One approach to such behavioral evaluation is to use standard language comprehension tests administered to children. For example, the Token Test [4] is commonly used to assess language skills of young children who exhibit language acquisition difficulties. To administer the test, the evaluator arranges a set of physical tokens on a table and asks the subject to perform various manipulation tasks ("When I touch the green square, you take the white one", etc.). The Token Test is an ideal evaluation for our system since it evaluates basic language-to-world mapping skills and does not rely on social or cultural knowledge.

The Token Test is divided into five parts ordered in increasing difficulty. Using the GSM based system we have described, our robot is now able to pass the first two parts. For example, it responds appropriately to requests such as: "Touch the large red circle!". As a whole, the robot might make some errors due to failures of various subsystems. Speech recognition errors or visual processing errors are two most common causes since the dialog structures are quite simple. But the main point we would like to emphasize is that the GSM and related algorithms provides our robot with the *capacity* for passing two of five parts of the test. Below, we suggest next steps for tackling the remaining parts.

However, our implementation based on the GSM can achieve more than simply respond to Token Test style requests. A human communication partner can also ask questions about what it sees, knows, and remembers about its table top world. Furthermore, the human can describe parts of the environment that the robot can't see, causing our robot's imagination module to instantiate categorical beliefs which can be verified and enriched by consequent perception.

A. Detailed Example of Performance

This example is part of the accompanying video. A user informs the robot that "there is a blue object at the left" (which is fed to the imaginer). Thus, the categorical layers are filled with the values corresponding to the verbal categories given, i.e. "left" for position, "blue" for color, and all categories equiprobable for size. Thus, if the robot is asked "What color is the one at the left?" it will answer "blue", even though it has never seen the object yet, and doesn't know exactly what shade of blue it is. However, when the robot is asked "How big is the one at the left?" it promptly responds "I have no idea" given that this information was not linguistically transmitted and that all size categories are a priori equally likely for blue objects at the left. Later, when it will see the object, it would answer "It is small", as it has captured adequate sensory information.

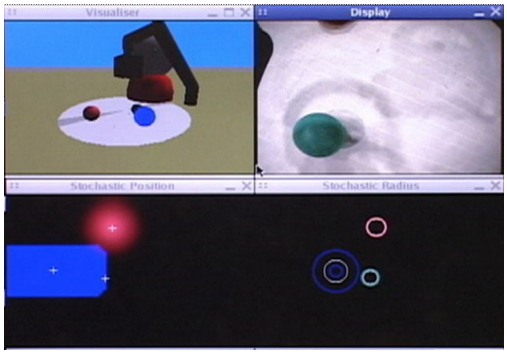

Fig. 4. GSM contents after the robot is told that, "There is a blue object at the left".

In Figure 4, the robot has already seen a red and a green object, as can be seen in the GSM. Furthermore, the user has informed the robot that "there is a blue object at the left", and the robot has created a representation for it. Notice that the left area of the table is not currently visible due to the field of view of the robot. Thus, the blue object that was described through language has not been seen yet. At the "stochastic position" window, the blue rectangular area corresponds to the position values classified as belonging to the spatial category "left". Notice how this differs from the single-point distribution for the green object (which is currently visible by the robot's camera as seen in the "display" window). Also, notice how it differs from the cloud-like distribution for the red object, that hasn't been seen for a while (and thus its distribution has diffused, as it might have moved in the mean time). At the "stochastic radius" window, the area between the inner and outer blue circles correspond to the possible radii the blue object might take. Notice how the radius of the green is already determined by previous observations, and thus the inner and outer green circles coincide (similarly for the red). Thus, when the robot is asked "How big is the blue one?" it responds with "I have no idea", while if it is asked "How big is the green one?" it gives a specific answer, i.e. "small" in this case.

In Figure 5, the robot has now moved its head, and the blue object that it had previously imagined (after "there is a blue object at the left") has now been seen. Compare to Fig. 4: At the "stochastic position" window the blue rectangular area in Fig. 4 has shrunk to a single point (the point under the leftmost cross in Fig. 5). Thus, the robot doesn't only know that the blue object is somewhere at the left, but is much more certain about exactly where it is. At the "stochastic radius" window, the outer and inner blue circles that existed in Figure 4 have shrunk and expanded in order to coincide with each other, and their current radius happens to be within the "small" category. Thus, when the robot is now asked "what size is the blue one?" it will respond with "small" (and not "I have no idea" as it would before seeing the blue object and after hearing "there is a blue object at the left").

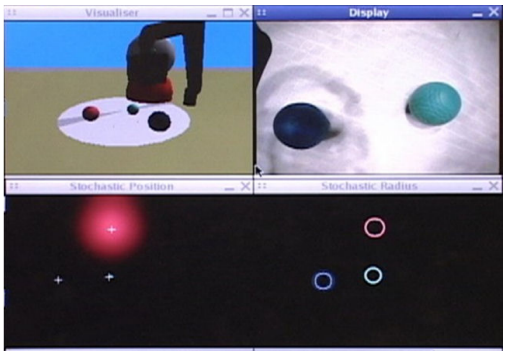

Fig. 5. GSM contents after the robot moves its head and sees the blue obj.

IX. Future Directions

Our current work is focused on three main objectives. First, we aim towards enabling the system to handle richer representations of object shapes, acquired through multiple views, integrated in an active-vision framework. Richer shape capabilities will enable the incorporation of richer spatial relations, such as containment and support, which in turn figure prominently in natural language semantics.

Second, we are enhancing the representation of other agents (i.e. the human user), in order to include not only their viewpoint towards the world, but also a complete embedded GSM ascribed to the other agent. Using embedded GSMs, it will for example be possible for the robot to encode differences in beliefs it holds from those it believes its human partner holds. Language planning can then take into account agent-dependent GSM contents to choose appropriate words.

Our third objective is behavioral: to develop the GSM to a stage that enables the robot to pass all five sections of the standard Token Test for Children [4]. We think that the required extensions can be designed atop of the current GSM-based architecture in a principled way.

X. Conclusion

We believe that the primary obstacle towards effective human-robot communication in natural language lies in the traditional separation of language from sensing and acting. Our main thesis is that special amodal knowledge structures representing situations are needed as bridges. We have presented the design and implementation of such a structure, namely a grounded situation model, that serves as a bridge for an interactive conversational robot, and which resides in a centrally located module in the implemented modular architecture. The overall design of the GSM was driven by two desiderata, and then the design of the specific GSM was further refined through a set of behavioral goals that were explicated. In the resulting system, all behavioral goals have been achieved, and a summary of all implemented behaviors is given in the appendix. The robot is currently able to pass the first two parts of the Token Test, a standard test used to assess early situated language skills. The robot is also able to answer questions about the present and past, act on objects and locations, and integrate verbal with sensory information about the world. As explicated before, the robot has three novel abilities compared to other existing conversational robots, and the GSM together with the implemented modular architecture have been instrumental in attaining them. We believe that the suggested GSM design, with its hierarchical object structure, three-layered property representations, and recursively embedded GSM's, provides an important step towards endowing robots with physically and socially grounded language skills, and ultimately towards truly cooperative conversational robots.

Appendix

A. Current Behavioral Repertoire

The system responds to the following utterance classes:

1. Questions (present/past):
<question> is/are the <obj des>
<question> was/were the <obj des> when <event des>
2. Action requests (referent description in present/past):
<action> the <obj des>
<action> the <obj des> when <event des>
3. Imaginer request: Imagine/There is a <obj des>
4. Location-centered "look at": Look at <location>
5. Viewpoint: <action> the one on my/your left/right
6. Basic mode switching:
Wake up, sleep, relax, look at me/the table
Types:
<question> belongs to {where, how big, what color}
<action> belongs to {touch, pick up, hand me, look at}
<obj des> contains <size> <color> <object> <locus>
<event des> contains <actor> <event type>
<size> belongs to {small, medium, large}
<color> belongs to {red, green, blue}
<locus> belongs to {center, left, right, top, bottom}
<actor> is either an <obj des> or <agent part>
<agent part> belongs to {my, your} × {head, arm}
<event type> belongs to {appeared, disappeared, started moving, stopped moving, came in view, came out of view}

References

[1] C. Breazeal et al. Humanoid Robots as Cooperative Partners for People. In *IJHR*, 2004.
[2] R. Burke et al. Creature Smarts. In *Proceedings Game Developers Conference*, 2001.
[3] C. Crangle and P. Suppes. *Language and Learning for Robots.* CSLI Publications, Stanford, CA, 1994.
[4] F. DiSimoni. *The Token Test for Children.* DLM Teaching Resources, USA, 1978.
[5] P. N. Johnson-Laird. *Mental Models.* Cambridge University Press, Cambridge, MA, 1983.
[6] P. McGuire et al. Multi-modal human-machine communication for instructing robot grasping tasks. In *Proc. IROS*, 2:1082-1088, 2002.
[7] T. Regier and L. Carlson. Grounding spatial language in perception. *J. Experim. Psych.*, 130(2):273-298, 2001.
[8] D. Roy, K.Hsiao and N. Mavridis. Mental Imagery for a Conversational Robot. *IEEE SMC B*, 34(3):1374-1383, 2004.
[9] D. Roy. Semiotic Schemas: A Framework for Grounding Language in Action and Perception. *Artificial Intelligence*, 167(1-2):170-205, 2005.
[10] C. Snow. Mothers' speech to children learning language. *Child Development*, 43:549–565, 1972.
[11] D. Sofge et al. An Agent-driven Human-centric Interface for Autonomous Mobile Robots. In *Proceedings SCI.*, 2003
[12] R. A. Zwaan and G. A. Radvansky. Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, 123(2):162–185, 1998