Expectation-driven Autonomous Learning and Interaction System

Bram Bolder, Holger Brandl, Martin Heracles, Herbert Janßen, Inna Mikhailova, Jens Schmüdderich, Christian Goerick

2008

Expectation-driven Autonomous Learning and Interaction System

Bram Bolder, Holger Brandl, Martin Heracles, Herbert Janssen, Inna Mikhailova,
Jens Schmüdderich, and Christian Goerick
Honda Research Institute Europe
Carl-Legien-Straße 30
D-63073 Offenbach am Main, Germany
bram.bolder@honda-ri.de

Abstract— **We introduce our latest autonomous learning and interaction system instance ALIS 2. It comprises different sensing modalities for visual (depth blobs, planar surfaces, motion) and auditory (speech, localization) signals and self-collision free behavior generation on the robot ASIMO. The system design emphasizes the split into a completely autonomous reactive layer and an expectation generation layer. Different feature channels can be classified and named with arbitrary speech labels in on-line learning sessions. The feasibility of the proposed approach is shown by interaction experiments.**

I. INTRODUCTION

In the recent years the research on humanoid robots made a considerable progress in the hardware, sensor-processing, and control related domains. The focus of the research now moves from enhancing isolated abilities of the robots towards the integration of different abilities into one complex system. Several questions have to be addressed on the system level, e.g. which system architecture supports incremental building, robustness and stability of the overall behavior, coupling of the processing and the integration of control on different time scales.

In [1] we presented a first step towards our long-term goal of incrementally creating an autonomously behaving system that learns and develops in interaction with a human user as well as based on internal needs and motivations. Here we develop this approach further and extend the system ALIS ("Autonomous Learning and Interaction System") by a mechanism of expectation generation, a learning speech classifier, and an extended processing and representation of visual sensor data. The expectation generation aims at an incremental step towards the integration of goal-driven behavior into a system that uses reactive controllers, while the learning of the speech classifier significantly increases the possibilities to interact with the system. The resulting system comprises audio saliency for gaze selection, a visual proto-object based fixation and short term memory of the current field of view, the classification of the proto-object's features and executed actions, the online learning of the speech labels, and an interaction-oriented control of the humanoid body including the generation of the gestures. The focus of this paper is not on the single elements but rather on the system design and the key properties of the architecture. For a detailed discussion of the the visual and speech processing of our system please see [2].

In section III we give a short review on the SYSTEMATICA framework, the design principle we built our systems on. Based on this framework, we describe the implementation of our current system ALIS 2 in section IV. Section V then illustrates and evaluates the system based on some experimental runs. We conclude by giving an outlook to our future research in section VI.

II. RELATED WORK

There exist a huge number of applications aiming at learning speech via robot-human interaction. The main difference of our work is that we are interested in a system architecture that allows the learning and the usage of mental concepts in general, whereas we see the speech learning as a particular case where an auditory utterance is a part of a mental concept. We pay special attention to the interplay between the parts of the system. One aspect of system integration is the question how the local classifications in different feature channels can cooperate and be bound into a mental concept. Another aspect of integration is on the control side: how a reactive control layer that uses no (or only simple) models and an anticipative control that uses mental concepts as world model can run in parallel and profit from each other. We discuss these two points below.

One major feature of our approach is the cooperation of different classifiers for different feature channels. These classifiers can operate asynchronously and have different output spaces. The system treats all classifiers as equal and can combine results from multiple classifiers dynamically. All the control is done via local decisions. We furthermore embed learning and the capability of detecting and resolving conflicts of classification results.

Existing work on mixture of experts [3] or ensemble learning architectures [4] use the cooperation of different classifiers, however the output spaces of all classifiers are identical. Classifiers whose input hardly contains enough information to classify the global outcome are still forced to do so. Furthermore the global classification always has to wait for the slowest classifier, even if it would not contribute at all to the result. Other systems exist that can handle independent classifiers [5], [6]. Here however exists a large asymmetry between the classifiers. On the one hand there is an unreliable and flexible classifier (in both [5] and [6] the speech/audio signal classifier), all the other classifiers are treated as reliable. This makes the system less flexible and expandable.

During the learning phase we use an additional attention cue that specifies which classifiers can provide teaching signals to the classifier that learns. Hence we do not force all classifiers to do the same job. This results in a synergy of classifiers without enforcement of equal classification performance.

All the visual information in our system is represented in form of proto-objects. These are also used for the target selection and fitness evaluation of some of the behaviors. Proto-objects are a concept originating from psychophysical modeling [7], [8], [9]. They can be thought of as coherent regions or groups of features in the field of view that are trackable and can be pointed or referred to without identification. Orabona et al. [10], [11] developed a system that uses proto-objects — in their case colored blobs — to let a robot learn the notion of an object consisting of possibly multiple proto-objects using statistical means. In our point of view, proto-objects simply describe entities in the outer world that can be interacted with. They are a representation of visual information that are sufficient to be used for the generation of behaviors.

Now we turn to the question of behavior control organization that integrates both reactive control and expectation-driven control. Some recent approaches in Reinforcement Learning use a reactive layer for the description of the state of the system-environment interaction as well as for execution of plans, e.g. [12], [13]. These approaches do not switch between reactive and anticipative modes; they are forced to always evaluate the future reward and always plan ahead. In contrast our system behaves in a reactive manner as long as it has no expectations, switches to expectation driven behavior if a feature channel generates expectations of associated features, and switches back to reactive mode ones the expectations are fulfilled.

The behavior-based approaches, e.g. [14], provide a possibility for the planning layer to manipulate the action selection in the reactive layer. However, the anticipation does not influence the perception that is separated into a 'symbol converter'. In our system the expectations participate also in the perception part. The bottom-up classification is processed in a different way depending on the expectations. Expected features are considered as reliable and are used as a teaching signal, whereas the unexpected features trigger the behavior that resolves the conflict between the expectation and the reality. This mechanism can later be used for disambiguation and hypothesis testing.

Common to the approaches discussed above is the fact that the extension of the reactive behavior aims directly at planning. However, from the evolutionary perspective, the anticipation may first be used simply to detect an inappropriate behavior. The model described in [15] goes in this direction. The focus is set on expectancy learning and the interplay between the expectancy system, the perception system, and the control that does not require extensive planning (e.g. conditioning, habituation, behavior suspension in case of the expectation mismatch). The planning is seen as a next incremental step. Our approach shows some parallels to this work. One of our original contributions is the active resolution of mismatch situations in a way that has not been proposed before.

Our conviction that one of important goals of current research is understanding the design principles of integration of subsystems into a large system is shared by [16]. The authors argue that the system architecture has to support the parallelism and conversion to shared representations instead of enforcement of the same data format in subsystems. As discussed above, our architecture was built according to the similar principles and fulfills these requirements. The main difference of our approach is that we deal more explicitly with the problems of behavior organization. To our knowledge, the implementations based on [16] are restricted to the parsing of commands to the robot that plans the action and monitors the goal achievement in a rather restricted scenario. Although the general CoSy framework, which is the base for the work in [16], aims at the integration of reactive and deliberative layers, the implementations give an impression that the system follows only a functional decomposition and there is no transitions between the behaviors generated on different layers. In contrast, our framework SYSTEMATICA supports both functional and behavioral decomposition that helps to design the system in the way that the complexity of representations match the complexity of the desired behavior. Further, instead of using 'blackboards' for information sharing SYSTEMATICA favors structured information communication. Finally, SYSTEMATICA supports incremental steps in system building. In our implementation we extend the reactive layer first by expectation-driven behavior before we will go further to the integration of planning abilities.

III. SYSTEMATICA

In this section we give a short review of the SYSTEMATICA framework, see [1] for a more detailed discussion. The framework describes the control architecture with the help of processing units. The unit n is characterized by the following elements:

- internal dynamics D_n;
- the used subspace $S_n(X)$ of the complete input space X spanned by extero- and proprioception;
- publicly accessible representations R_n the unit can create;
- issued top-down information $T_{n,l}$ to other units, $n > l$;
- accepted top-down information $T_{m,n}$, $m > n$;
- issued motor commands M_n with priority P_n.

The elements described above may be empty. For example a unit may use no perceptive input $S_n = \varnothing$ but only read the accessible representations of other units, or it can only create representations but emit no motor commands, $M_n = \varnothing$. The index n represents the index of the creation in the incremental system building. Therefore, units with a lower index n cannot observe the representations R_m of units with a higher index m. The unit can autonomously emit some externally observable behavior B_n by issuing the

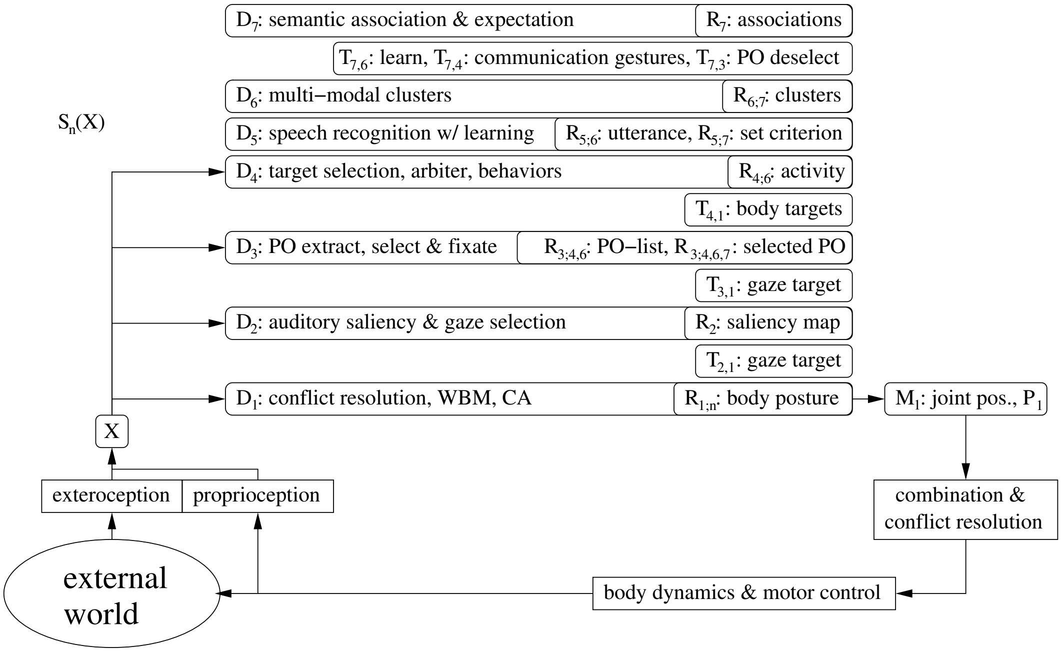
D_7: semantic association & expectation
R_7: associations
$T_{7,6}$: learn, $T_{7,4}$: communication gestures, $T_{7,3}$: PO deselect
D_6: multi–modal clusters
$R_{6;7}$: clusters
$S_n(X)$
D_5: speech recognition w/ learning
$R_{5;6}$: utterance, $R_{5;7}$: set criterion
D_4: target selection, arbiter, behaviors
$R_{4;6}$: activity
$T_{4,1}$: body targets
D_3: PO extract, select & fixate
$R_{3;4,6}$: PO–list, $R_{3;4,6,7}$: selected PO
$T_{3,1}$: gaze target
D_2: auditory saliency & gaze selection
R_2: saliency map
$T_{2,1}$: gaze target
D_1: conflict resolution, WBM, CA
$R_{1;n}$: body posture
M_1: joint pos., P_1
X
exteroception
proprioception
combination & conflict resolution
external world
body dynamics & motor control

Fig. 1. ALIS 2 overview using the SYSTEMATICA framework.

motor commands or by providing top-down modulation. The behavior B_n may have different semantics Z_j depending on the current situation or context C_i, i.e. the behaviors B_n represent skills or actions from the system's point of view rather than observer dependent quantities.

The SYSTEMATICA framework allows to characterize the system architecture with respect to the following issues: Find a system's decomposition or a procedure to decompose or construct units n consisting of $S_n(x), D_n, B_n, R_n, M_n, P_n, T_{m,n}$ such that

- an incremental and learning system can be built;
- the system is always able to act, even if the level of performance may vary, i.e. they should treat top-down information $T_{m,n}$ only on a voluntary basis;
- lower level units n provide representations and decompositions that
  - are suited to show a certain behavior at level n,
  - are suited to serve as auxiliary decompositions for higher levels $m > n$, i.e. make the situation treatable for others, provide an 'internal platform' so that higher levels can learn to treat the situation.

If these requirements are met, the system will have several benefits. It is robust against failure of units, i.e. if a unit n fails, the units 1 to $n-1$ should remain unaffected. Only the overall performance of the system would be reduced. Due to the loose coupling between the units, asynchrony or latencies should not break the overall performance. The hierarchical design and the sharing of representations allows the units to be of reduced complexity.

The figure 1 shows the SYSTEMATICA-based formalization of the actual instance of our system ALIS 2. The units one to four are to a large extent reused from the previous version of ALIS [1]. These units can work independently from units five to seven that are a result of the latest research. Hence, our system building is indeed incremental. At the same time, in difference to the subsumption architecture, the units are not independent, but build on the representations of underlying layers. The newly added unit six performs the multimodal feature classification and thus provides the representations suited for the expectation generation in unit seven. Unit seven in turn sends the top-down information to lower level units in form of learning signals and modulation of basic behaviors. The individual units are described in details in the next section.

IV. IMPLEMENTED SYSTEM UNITS

A. Conflict Resolution, Whole Body Motion, Collision Avoidance (Unit 1)

The unit with the dynamics D_1 is the whole body motion control of the robot, including a basic conflict resolution for different target commands and a self collision avoidance of the robot's body. It is identical to the unit D_1 in the previous ALIS instance. See [1] for a more detailed description. The representation R_1 used and provided is a copy of the overall posture of the robot. The top-down information $T_{n,1}$ provided to the unit has the form of targets for the right and left hand respectively, the head, and the walking. The implemented conflict resolution takes care that for each of the four targets individually at most one is selected for execution. Without top-down information, the robot is standing in a rest position with a predefined posture at a predefined position. If the top-down information is switched off, the robot walks back to this predefined home position.

B. Auditory Saliency and Gaze Selection (Unit 2)

The second unit with dynamics D_2 computes an one-dimensional (azimuth) auditory localization and stores it in a saliency map R_2. See [17] for details. Current research focuses on extracting the elevation as well [18]. It provides the gaze target $T_{2,1}$ based on the peak of a saliency map to direct the gaze into the direction of the sound origin. This allows interaction from any position around the robot in order for instance to guide visual attention to a region outside the current visual view.

C. Proto-Object Extraction, Selection, and Fixation (Unit 3)

Unit three extracts proto-objects from the current visual scene and performs a temporal and spatial stabilization of these using short term memories. The concept of proto-objects as we employ it for behavior generation is explained in more detail in [19]. Three different visual cues enter the system, each with their own short term memory. Depth proto-objects are based on contiguous regions of depth values in a restricted range we call the peripersonal space. This overlaps roughly with the manipulation range of both arms. The second kind of proto-objects are based on object proper motion, i.e. contiguous image regions with similar movement relative to the robot [20]. These proto-objects allow an interaction over a larger range. One can attract the robot for instance by waving. The third kind of proto-objects are based on textured or non-textured planar surfaces. Although the method can extract planar surfaces in arbitrary orientations [21], we restrict ourselves here to roughly horizontal surfaces. These proto-objects allow the robot to identify behaviorally relevant support surfaces like chairs, tables, and the floor. The proto-objects from the three sources are then merged, i.e. those that probably describe the same entity in the world are merged into one proto-object. The complete list of proto-objects is made available inside R_3 for all interested units.

A simple proto-object attention mechanism selects one of the currently available proto-objects. Its unique ID, the 'selected-ID' is also made available in R_3. The selection mechanism stays on the same proto-object as long as it is available or the top down influence $T_{7,3}$ deselects the current ID. As long as there are proto-objects that were not deselected, a next proto-object is then selected based on an arbitrary metric regarding status and distance to the robot. If no proto-object is available, the selected-ID is set to the value 'invalid'. The selection mechanism allows stable interaction with any single proto-object.

If the selected-ID is valid, a gaze command $T_{3,1}$ is generated to direct the gaze towards the proto-object location. This allows a simple visual tracking of the attended entity. $T_{3,1}$ has a higher priority than $T_{2,1}$.

D. Target Selection, Arbiter, Behaviors (Unit 4)

Unit four with D_4 governs the control of the robot's body except for the gaze direction. This is achieved by deriving targets from the proto-object representation R_3 and sending them as top-down information $T_{4,1}$ for the right and the left hand as well as for walking to unit one. Details of the internal dynamics D_3 can be found in [22]. Unit four consists of a group of internal behaviors, each sending a fitness value to the arbiter that signals if they are able to be executed successfully. The arbiter resolves conflicts and sends an activity value to each of the behaviors. The activity pattern is made available in R_4. The top down information $T_{7,4}$ acts as a bias on the activities. This allows certain behaviors to be executed with preference without interfering too much with the originally active behaviors. The overall behavior B_4 then consists of the complete set of internal behaviors and their interaction with the arbiter.

Some of the behaviors are used for direct interaction. Peripersonal and planar surface proto-objects will be pointed to with the most appropriate hand — depending on which side the proto-object is while regarding a hysteresis. The robot will also adapt the distance towards these proto-objects, i.e. walk towards it if it is too far away and walk backwards if it is too close.

E. The Units Introduced so far

The combination of the units one to four realizes the framework for autonomous interaction with the robot in a reactive way. Interaction can be initiated in different spatial regions around the robot and trigger different response behaviors. All interactions have in common, that the gaze is directed to the currently attended stimulus. Visual proto-objects will be preferred over auditory stimuli. If neither a visual nor an auditory stimulus is available, the robot will go back to or remain at a predefined home position.

F. Speech Recognition with Learning (Unit 5)

Unit five consists of a speech recognition system including an online learning of arbitrary utterances. Two kinds of utterances are distinguished and treated differently. Predefined utterances are mapped to predefined labels available in R_5. These are used to trigger the learning according to different criteria. All other utterances are communicated to the speech classifier in unit six via R_5.

G. Multi-Modal Clusters (Unit 6)

Using the representations R_3, R_4, and R_5, Unit six is able to classify different features. These features are compared with predefined or online trained clusters separately for the different feature channels. The classification results are then made available as population codes in R_6, i.e. for each cluster, a confidence value based on the similarity to the current input features is calculated and stored. The population code is very flexible, since it allows to express a best candidate, ambiguities, and the fact that nothing could be classified. The latter is used for instance to encode missing sensor values such as in the case of the speech recognition when nothing is spoken.

In our ALIS 2 implementation, five feature channels are extracted. For the proto-objects, several clusters are predefined that describe relative properties (left and right position), absolute properties (horizontal planar surfaces of different height such as a table, chair, or step), or the movement status (moving or still). The internal state of the behavior generation activity is classified with respect to the two behaviors of walking towards a target and returning to the home position.

The choice of these clusters is arbitrary but fixed, the learning of the clusters is subject to current research. For speech, no clusters are predefined. These can be trained online during a learning session by providing a few (three to five are sufficient) repetitions of the utterance and a feedback for the desired output activity. Note that one can train synonyms, i.e. different utterances for the same desired output activity.

At any time, R_6 will represent all classified information for the current scene. All the feature channel classifiers constantly update their classification results but can do so at different speeds. There is no need for synchronization.

H. Semantic Association and Expectation (Unit 7)

The unit with the dynamics D_7 is able to associate the results from the classifiers in R_6 with each other. An association matrix converts between the results of the different feature channels. The classification results can then generate expectations for each of the other channels. Which classifiers can generate expectations is arbitrary, in our implementation ALIS 2 we restricted this to only the auditory utterances, although different other combinations have been successfully tested. The expectations are locally compared to the current results of each feature channel. If the local classification matches the expectation, a certain gesture is requested via $T_{7,5}$ — here the nodding of the head.

In case of a mismatch, a process is triggered to try to resolve it. If then the mismatch is with respect to a proto-object feature, a communicative gesture — here the shaking of the head — is requested via $T_{7,5}$ and the current proto-object is deselected via $T_{7,3}$. This is repeated until the expectation is met. A time-out prevents the system to stick with an expectation that it can not resolve by itself anytime soon. If the expectation is with respect to the internal behavior activation state, then this behavior is simply requested via $T_{7,5}$. The mismatch case for the internal behavior state can thus be interpreted as providing a command to the robot. The behavior that the robot returns to its home position when no proto-object is available can be associated with an utterance. Providing this command is the only way for the user to let the robot disengage interaction by retreating.

For each of the four non-speech feature channels, a predefined utterance available in R_5 can trigger a learning session. This learning session simply raises the expectation for any result for the respective channel, i.e. all classification results are permitted, and a zero expectation for the others. The matching classification results are then transformed into the respectively desired utterance classifier output. This is then communicated via $T_{7,6}$ in order to provide a learning signal to the speech cluster learning. A time-out after the last provided utterance completes the learning session.

The correlation matrix used to convert the classification results between the different feature channels was chosen to be fixed. The only information that it contains however is that there is a correlation between auditory clusters and the others. This simply encodes the namability of features, i.e. the fact that the human interactor uses certain utterances for certain features.

I. Overview over the Complete Dynamics

The different dynamics in the units described above and their interplay result in an overall system with many features. A completely reactive layer (units one to four) allows simple interaction. The upper layer (units five to seven) allow to combine information extracted from the environment and from internal states in order to modulate the behaviors. Should any part of the upper layer fail, the system would still be usable.

The expectation generation allows the system to correlate and evaluate properties, to learn new features, to command the robot to activate certain behaviors, and to extend the system beyond reactivity. The latter allows the step from reactive to goal directed behaviors.

V. RESULTS

A typical example of a continuous interaction sequence is depicted and commented in figure 2.

Figure 3 shows the most important system states during this sequence. In the sequence labels for all modalities were trained, while the use of the labels for expectation generation was deliberately limited to a few trials to keep the graph readable. For the same reason we also did not include learning of synonyms in this experiment, although this is possible at any time.

The basic interaction scenario is easily described:

First, by default the system is 'reactive' to direct visual and auditory stimuli: if an object is presented within a short range of the robot (about 1m) or a plane within a larger range (about 2.5 m) ASIMO will directly react by gazing at, pointing at, and approaching the target. Movement like a walking person or hand-waving as well as any acoustic noise will change ASIMO's gaze but not trigger pointing or approaching.

Secondly a human can trigger a learning session for any modality by speaking a specific key phrase such as 'learn where this object is' or 'learn a name for your action'. ASIMO will react by showing his attention with a gesture (slight raising of his arms). The human can now utter the label that ASIMO shall associate to this state. We get fairly reliable recognition for three to five repetitions of the label during a learning session. If the human does not speak for a few seconds ASIMO will automatically terminate the learning session.

Third one can utter an already learned label. ASIMO will then compare his current sensor/action state to the specified expectation state. If it is a sensor state (vision input), ASIMO will either nod (match) or shake (mismatch) his head. In case of a mismatch ASIMO will also try to match alternative visual input (proto-objects) until either a match is achieved or a few seconds have expired. In case of an action label ASIMO will directly execute the specified action (e.g. walk back) while continuing to react to the environment (e.g. still gaze and point at objects).

The above elements can be used by the human in any order and without delays — the whole learning and evaluation process in online and real time.

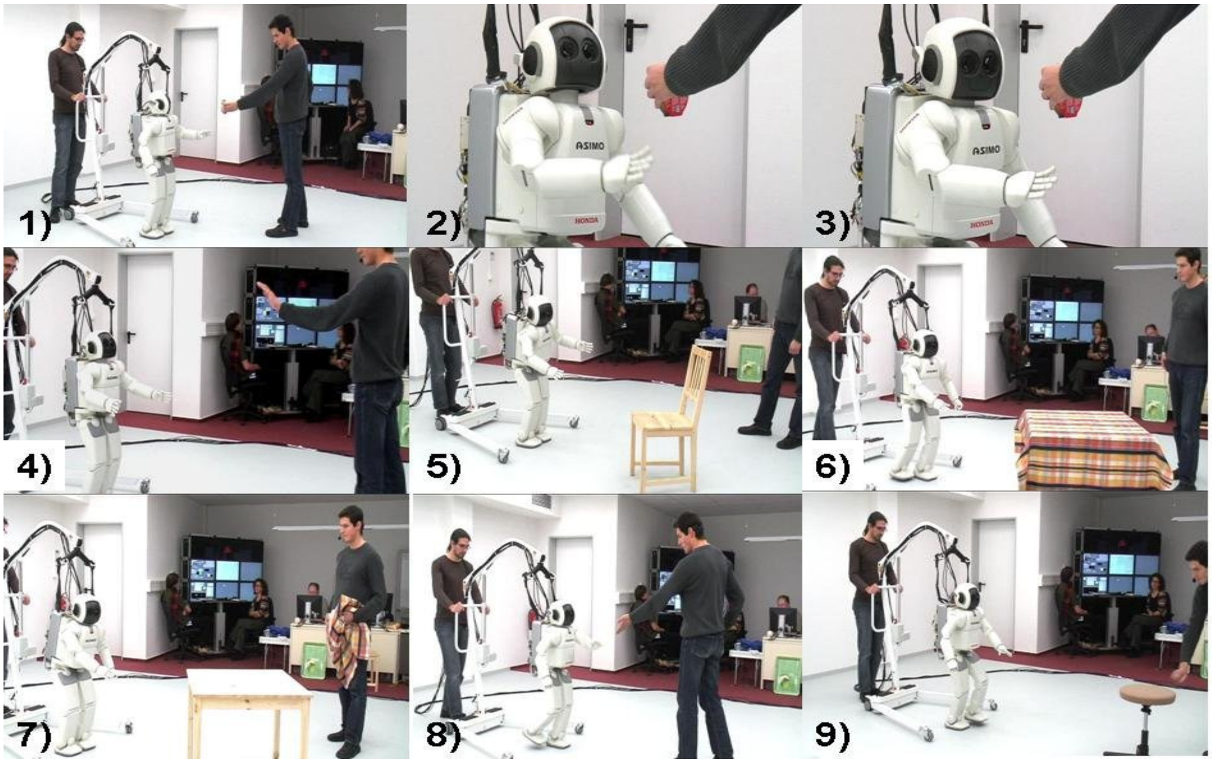

Fig. 2. Image series from the experiment on speech learning and evaluation. 1) Fixation and pointing while learning the speech label for left position. 2,3) Expectation evaluation: nodding as ‘Yes’ gesture to show the expectation match. 4) Interaction with a motion proto-object. Learning of ‘still’ and ‘moving’ labels. 5,6,7) Interaction with planar proto-objects. 5) Learning of the ‘chair’ label. 6) Fixation and adjustment of the interaction distance to the table and learning of the ‘table’ label. 7) Testing the independency from visual appearance. 8) Approaching the user and learning ‘forward’ label. 9) Evaluation of learned labels on the unknown object.

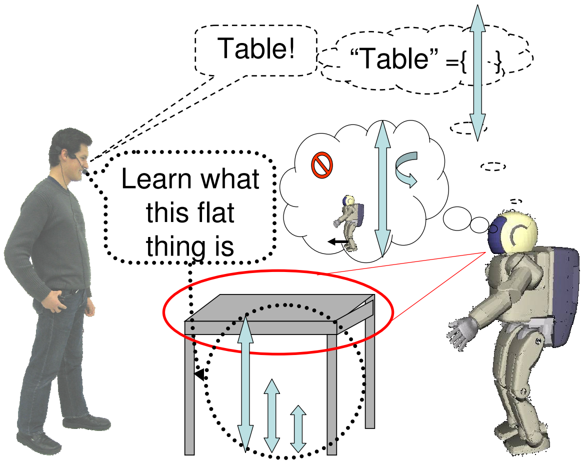

Fig. 4. Example of learning the association of the audio label ‘table’ with the current result for the planar surface classifier. The current situation (solid lines) is restricted by the learning criterion (dotted lines) to form the association to be learned (dashed lines). See text for a more detailed discussion.

The whole sequence consists of first learning ‘left’ and ‘right’, then evaluating them, then learning other modalities with a quick evaluation of ‘table’ in between and finally a longer evaluation sequence. This final evaluation sequence was done with a single previously not shown stool object to show that the modalities are independent of appearance and are applicable to any target object.

In the first evaluation at second 72 the user says ‘right’ while showing an object on the left. This creates a mismatch that triggers head shaking (label ‘N’ in behavior activation plot) and stops tracking. The request stays active until the robot finds the object on the right (second 80) and nods (label ‘Y’). After second 277 we evaluate some of the learned labels while ASIMO is tracking an unknown seen object. This shows the results are independent of the visual appearance.

Around second 304 the evaluation utterance ‘still’ is at first mis-recognized as ‘approach’ but this has no adversary effect and repeating the utterance led to a correct recognition.

Figure 4 illustrates the system behavior during learning. The table is the currently selected and fixated proto-object (solid ellipse). The current classification results (in solid lined cloud) are: the object is not moving, approach behavior towards the table is active, the object consists of a surface with a large height, and the object is e.g. on the left side. The user now tells ASIMO with the utterance ‘learn what this flat thing is’ that it should associate the following utterances with the results of the planar surface classifier (dotted lines). Upon uttering ‘table’ a few times, the system is able to associate this utterance with the planar surface classification result of something with a large height (dashed lines).

To show the flexibility and reliability of the system we depicted another short sequence in figure 5. Here the user first teaches ‘left’ but actually moves the object around

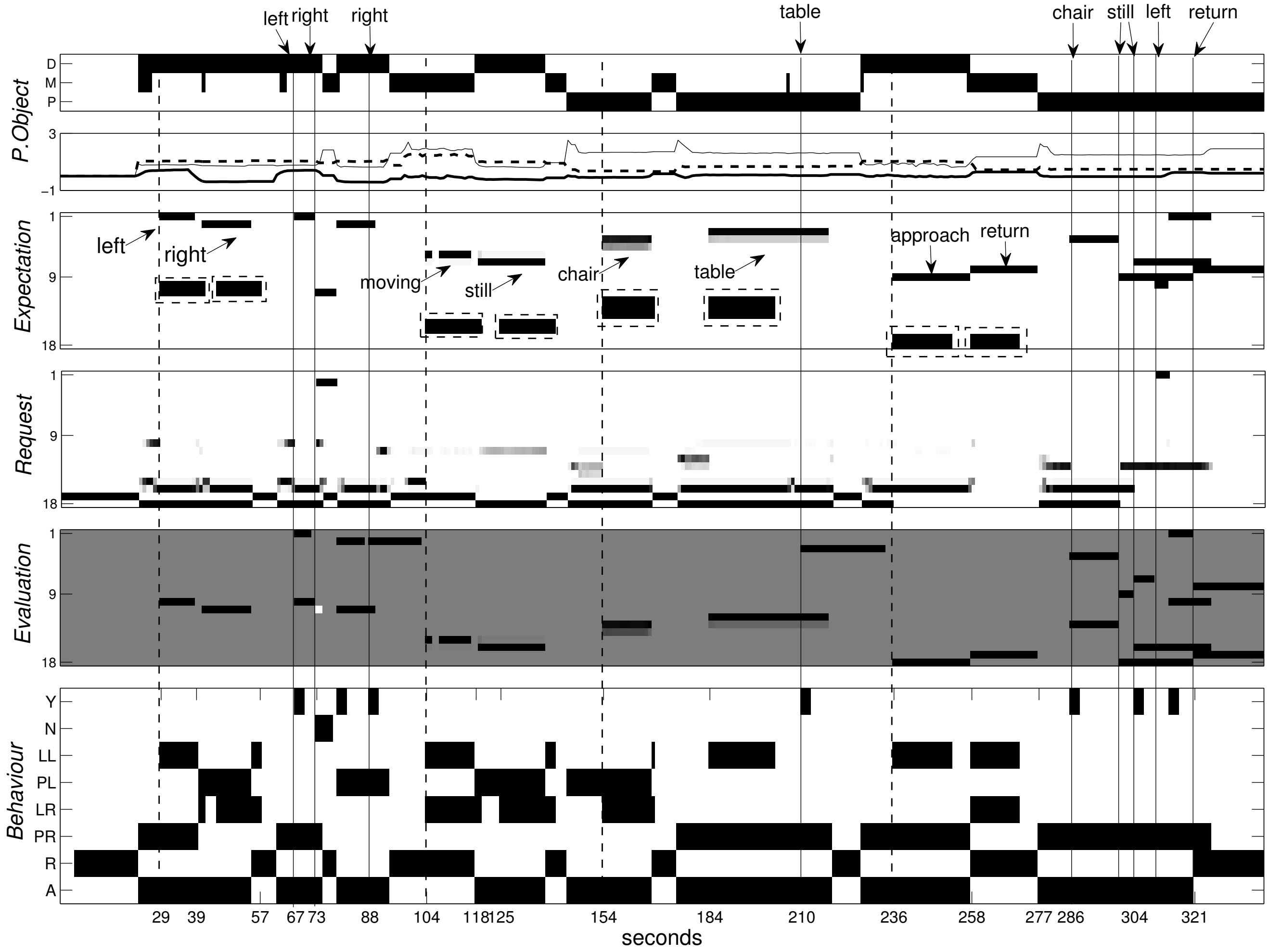

Fig. 3. One run of the experiment: speech learning and evaluation — for explanation also see text. Depicted is the system state over time. Above the graph the utterances for evaluation are shown. In the expectation plot the speech input for learning is shown, however in the experiment each utterance is repeated three to five times. The upper two plots display information about proto objects. The first plot depicts the source of the current proto object: 'D'- Depth, 'M'- Motion, 'P'- Plane. The second plot displays the object's position in cylindric coordinates relative to the robot's torso: the thick line shows the angle (rad), the dashed line the height, and the thin line the distance. The plot on the very bottom shows the state of the behavior activations: 'Y'-nod,'N'-shake', 'LL','LR'- learning gestures with left/right hand, 'PL','PR'-pointing with left/right hand, 'R'-return, 'A'-approach. The middle 3 plots show the state of the upper layer (units five to seven). The expectation plot shows which states are currently expected, the request shows which states have not been confirmed, and the evaluation plot shows both the expectation match (dark) and mismatch (bright). The first 9 values of expectation-, request-, and evaluation- vector correspond to the speech channel. The dashed boxes show the expectations in non-speech channels generated by the learning criteria. The corresponding expectation in the speech channel is used as a teaching signal.

during learning and even holds it to the right of the robot for some time as can be seen from the proto object position and also from the bars in the expectation plot. Since the association is between the 'label' class and the highest accumulated evidence this still gives stable results which is important for an interactive robot that may move around freely while interacting. Secondly 'right' is trained but this time by hand waving in some distance to the right of the robot, thus only generating movement proto objects as can be verified in the top plot of the graph. Finally both labels are evaluated and produce the correct response.

The total system was tested extensively during many interactive test runs and presentations to visitors of our lab and performs reliably. Please also see the accompanying video.

VI. Discussion and Outlook

We showed the implementation of a robustly interacting and learning system with both reactive and expectation driven behaviors. Our main focus lies on the design of the complete system and not on the single building blocks. The design principle SYSTEMATICA has proven to be an efficient and flexible way to consider system design and its implementation. By taking care of the coupling between the different system units, i.e. which representations to made

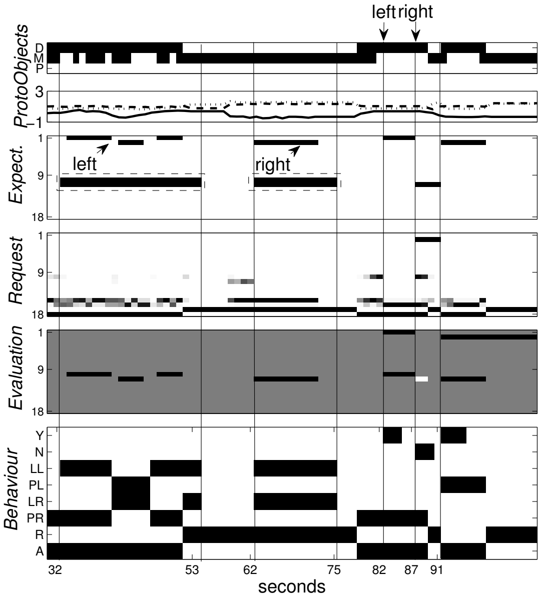

Fig. 5. Test of the system's stability. See text for explanation.

available and which top-down feedback should be used, a robust system could be built. ALIS 2 successfully shows this flexible and robust behavior during interaction. At any time, the system is in a usable state and follows the dynamics of the interaction.

The standardized data format used for the classification results, the uniform handling of the classifiers, and the usage of expectations as an interface for specifying goals, testing hypotheses, and teaching signals have proven to be an efficient way of designing the abstraction away from purely reactive systems. The simplicity of these methods allow a flexible extention in future systems.

The ALIS 2 system as described in this paper is in principle just a snapshot of our current research platform. It is used as a basis for integrating additional functionality or for replacing existing parts by better ones. The ideas behind the latter are twofold. One is to make the system more flexible and extensible, the other is to replace currently taken shortcuts by proper functionality. The idea behind this is not to redesign the system from scratch, but to take a working system and then replace or add parts that keep the system working but perform better. One example is our current research on replacing the predefined clusters for the non-speech cases by a general concept of learning those clusters. Preliminary results already look promising. Other current work includes using 2d sound localization, speech production instead of or additional to gestures, and incorporating tactile sensors. On the system level, we are starting to investigate ways of decoupling the functional units even further in order to enhance robustness even more.

ACKNOWLEDGMENT

We would like to thank our colleagues Mark Dunn, Achim Bendig, Michael Gienger, Benjamin Dittes, Marcus Stein, Antonello Ceravola, Martin Heckmann, Tobias Rodemann, and Frank Joublin for fruitful discussions and their practical support.

REFERENCES

[1] C. Goerick, B. Bolder, H. Janssen, M. Gienger, H. Sugiura, M. Dunn, I. Mikhailova, T. Rodemann, H. Wersing, and S. Kirstein, "Towards incremental hierarchical behavior generation for humanoids," in *IEEE-RAS International Conference on Humanoids*, 2007.

[2] J. Schmüdderich, H. Brandl, B. Bolder, M. Heracles, H. Janssen, I. Mikhailova, and C. Goerick, "Organizing multimodal perception for autonomous learning and interactive systems," in *IEEE-RAS International Conference on Humanoids*, 2008.

[3] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm, Tech. Rep. AIM-1440, 1993.

[4] R. Polikar, D. Parikh, and S. Mandayam, "Multiple classifier systems for multisensor data fusion," in *Sensors Applications Symposium. Proceedings of the 2006 IEEE*, 2006, pp. 180–184.

[5] D. H. Ballard and C. Yu, "A multimodal learning interface for word acquisition," in *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003.

[6] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2007, pp. 2415–2420.

[7] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Research*, vol. 40, pp. 1469–1487, 2000.

[8] A. Clark, "Feature-placing and proto-objects," *Philosophical Psychology*, no. 4, pp. 443–469, December 2004.

[9] Z. W. Pylyshyn, "Visual indexes, preconceptual objects, and situated vision," *Cognition*, no. 1, pp. 127–158, June 2001.

[10] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *CVPR*, 2005.

[11] L. Natale, F. Orabona, F. Berton, G. Metta, and G. Sandini, "From sensorimotor development to object perception," in *Humanoids*, 2005.

[12] S. Hart, S. Ou, J. Sweeney, and R. Grupen, "A framework for learning declarative structure," in *Robotics: Science and Systems - Workshop on Manipulation for Human Environments.* Philadelphia, Pennsylvania., August 2006.

[13] M. Toussaint and A. J. Storkey, "Probabilistic inference for solving discrete and continuous state markov decision processes," in *ICML*, 2006, pp. 945–952.

[14] P. Ulam and R. Arkin, "Biasing behavioral activation with intent," *to appear in Intelligent Service Robotics*, 2008.

[15] C. Balkenius, *Natural Intelligence in Artificial Creatures.* Lund University Cognitive Studies 37, 1995.

[16] N. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. Kruijff, M. Brenner, G. Berginc, and D. Skočaj, "Towards an integrated robot with multiple cognitive functions," in *AAAI.* AAAI Press, 2007, pp. 1548–1553.

[17] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin, and C. Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS).* IEEE, 2006.

[18] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2008), accepted.* IEEE, 2008.

[19] B. Bolder, M. Dunn, M. Gienger, H. Janssen, H. Sugiura, and C. Goerick, "Visually guided whole body interaction," in *IEEE International Conference on Robotics and Automation*, 2007.

[20] J. Schmüdderich, V. Willert, J. Eggert, S. Rebhan, C. Goerick, G. Sagerer, and E. Körner, "Detecting objects proper motion using optical flow, kinematics and depth information," *IEEE Trans. Man Cybern.*, vol. 38, no. 4, 2008.

[21] M. Heracles, B. Bolder, and C. Goerick, "Robust detection of arbitrary planar surfaces in real-time from unreliable 3D data," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009, submitted.

[22] T. Bergener, C. Bruckhoff, P. Dahm, H. Janssen, F. Joublin, R. Menzner, A. Steinhage, and W. von Seelen, "Complex behavior by means of dynamical systems for an anthropomorphic robot," *Neural Networks*, no. 7, pp. 1087–1099, 1999.