Toward Social Cognition in Robotics: Extracting and Internalizing Meaning from Perception

Jesús Martínez-Gómez, Rebeca Marfil, Luis V. Calderita, Juan Pedro Bandera, Luis J. Manso, Antonio Bandera, Adrián Romero-Garcés and Pablo Bustos

***Abstract*—One of the long-term objectives of artificial cognition is that robots will increasingly be capable of interacting with their human counterparts in open-ended tasks that can change over time. To achieve this end, the robot should be able to acquire and internalize new knowledge from human-robot interaction, *online*. This implies that the robot should attend and perceive the available cues, both verbal and nonverbal, that contain information about the inner qualities of the human counterparts. Social cognition focuses on the perceiver's ability to build cognitive representations of actors (emotions, intentions, ...) and their contexts. These representations should provide meaning to the sensed inputs and mediate the behavioural responses of the robot within this social scenario. This paper describes how the abilities for building such as cognitive representations are currently endowing in the cognitive software architecture RoboCog. It also presents a first set of complete experiments, involving different user profiles. These experiments show the promising possibilities of the proposal, and reveal the main future improvements to be addressed.**

***Index Terms*—Perception, verbal interaction, human-oriented perception, inner model.**

I. INTRODUCTION

Jesús Martínez is with both the Universities of Málaga and of Castilla-La Mancha. E-mail: Jesus.Martinez@uclm.es

Rebeca Marfil, Juan Pedro Bandera, Adrián Romero-Garcés are with University of Málaga. E-mail: {rebeca, jpbandera, ajbandera, adrigtl}@uma.es

Luis J. Manso, Luis V. Calderita, and Pablo Bustos are with University of Extremadura. E-mail: {lmanso, pbustos}@unex.es

COLLABORATION is an essential feature of social robotics [20]. Briefly, when two or more people agree on a common goal and a joint intention to reach that goal, they have to coordinate their actions to engage in joint actions, planning their courses of actions according to the actions of the other partners. The same holds for teams where the partners are people and robots [2], resulting on a collection of technical questions difficult to answer.

Human-robot collaboration requires the robot to coordinate its behavior to the behaviors of the humans at different levels, e.g., the semantic level, the level of the content and behavior selection in the interaction, and low-level aspects such as the temporal dynamics of the interaction [16]. This forces the robot to internalize information about the motions, actions and intentions of the rest of partners, and about the state of the environment. Furthermore, it is interesting for the robot to acquire knowledge about the profile (abilities or preferences) of their partners. Thus, the proper action will be chosen out of a collection of possible actions based on that knowledge. However, people are a dynamic factor in the environment with movements and manipulations complex to predict and understand. Therefore, robots need to form realistic expectations about human behaviour, but also to properly react to unexpected events [13]. Unfortunately, this level of flexibility is not provided by classical planning strategies.

The continuous coordination with humans emanates from the understanding of new actions in real-time, and it should provoke changes on the robot course of action at several layers of abstraction: from the sensorimotor level to the high-level planning one. Here, the goal is to integrate the responses emanated from all levels of abstraction. Furthermore, they must be included into a control architecture that provides the necessary functionalities for performing collaborative tasks: deep representations, domain knowledge and perception and action behaviours [3]. As M. Ali pointed out, *human-robot collaboration (HRC) does not only require good robotic platforms, but also needs supporting system software components* [1].

This paper describes how an initial corpus of functionalities for human-robot interaction has been incorporated within the cognitive architecture RoboCog [5]. According to their nature, functionalities are provided by two networks of components, one of them related with perception (people detection and tracking, and face classification), and the other one with verbal communication (speech recognition and synthesis). Briefly, these perceivers will capture the signals delivered by social targets to build meaningful percepts. With additional percepts, these representations will define the robotics inner experience of the world. Perception is then essentially the interface between outer and inner worlds [4].

The rest of the paper is organised as follows: Section II provides a brief description of the RoboCog architecture and its internal blocks. Section III describes how the functionalities for human-robot interaction are linked to the decision-making and executive modules of the architecture. Although the inclusion of both sets of functionalities is not restricted to a specific application, the present work is mainly focused on the collaboration into the framework of the ADAPTA project, id. number ITC-20111030. Hence, Section IV provides obtained results within this project. Finally, conclusion and future work are drawn at Section V.

II. The cognitive software architecture RoboCog

When making decisions that directly involve human users, the traditional 3-tier planning and plan execution scheme [8], which separates symbolic high-level planning from geometric plan execution, is not the best strategy. The generation of symbolic plans is relatively slow, thus the approach has to rely on a (almost) static world. Such an assumption is not only unrealistic, but also produces behavior that does not react to changes, which feels unnatural to humans. Motivated by human decision-making, the cognitive architecture RoboCog [5], depicted in Fig. 1, follows the guideline pointed out by Hayes-Roth and Hayes-Roth [9], which showed in protocol studies that when humans make plans, they consider different levels of abstraction in parallel and mentally simulate the execution of the task.

Within RoboCog, action execution, simulation, and perception are intimately tied together, sharing a common motor representation. This inner representation of the outer world is the central module of the architecture for action control, and it is organised in a hierarchical way. Thus, it provides different synchronised interfaces at levels of abstraction that range from the fine-grained aspects to symbolic high level. This central representation helps the robot to be aware of itself, but also to monitor its own capabilities and limitations. In Fig. 1, the representation provides two levels of abstraction that can be accessed through different interfaces. These interfaces deliver models of the outer world at a given abstraction level. Together with this central representation, the main elements of RoboCog will be the existence of a hierarchy of task-oriented modules, connected to the internal representation through these interfaces. The task-oriented modules (the so-called compoNets, as they will be composed by a set of software components) will

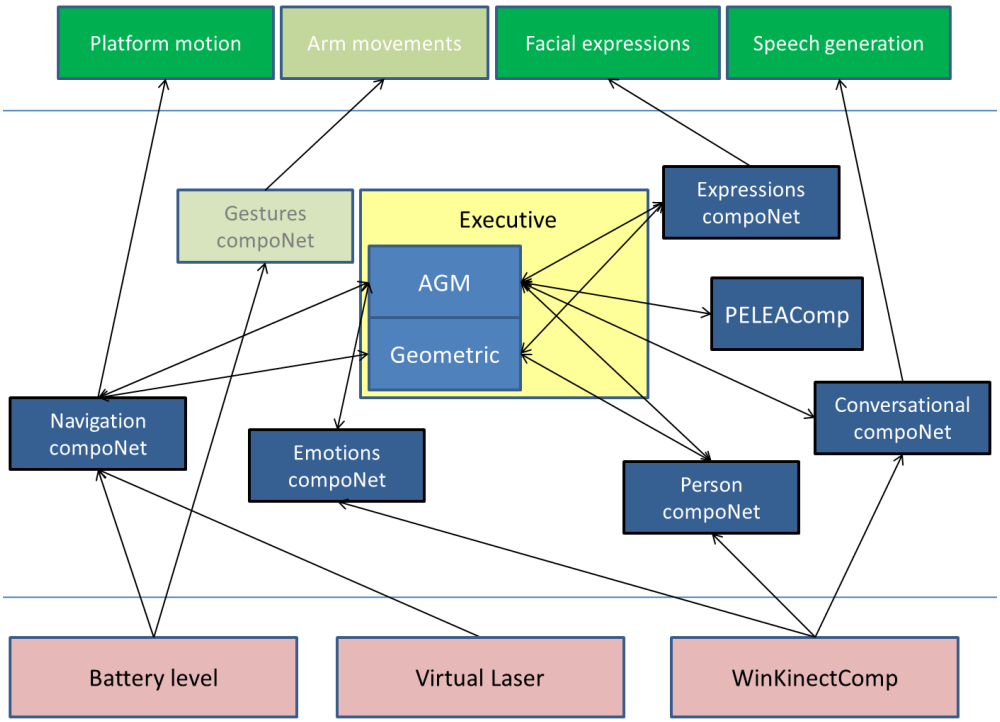

Fig. 1. Overview of the RoboCog architecture. Green upper blocks are low-level action components, pink low blocks are low-level perception components and blue blocks are networks of componets unfolding a specific functionality. The yellow block is the Executive center that sequentiates the activity and routes traffic among the blocks. Blurred blocks show modules not integrated in the specific instance of the architecture described in this paper.

be connected to the outer world through the Hardware Abstraction Layer. In the figure, this layer is divided up into action (motion or speech synthesis) and perception modules such as the virtual Laser or Battery level. The connection of the compoNets with the central representation is in charge of specific components, the agents.

The symbolic level encodes the world using a graph model (AGM, Active Grammar-based Model), whose evolution can be validated using an a priori set grammar [14]. Together with the Decision-making compoNet, PELEAComp, the aim is to provide an inherent trade-off between preconceived plans and reactive behavior. The concept of teleo-reactive plans also includes a learning mechanism, which learns preconditions and effects of actions (teleo-reactive operators). Here, the learning of plans will rather be targeted towards building a plan library, similar as in early AI planning systems. A prerequisite of this approach is the possibility to transform existing plans [15], which is provided in RoboCog by the integration of the PELEA framework [17] for planning (PELEAComp).

On the other hand, the geometric level encodes the world as a graph (the scene graph) where each relevant item (the robot and the people, but also objects and the environment) is a node linked to a kinematic tree. The whole representation can be animated as a virtual environment using the appropriate engine.

The responsibility of synchronising the

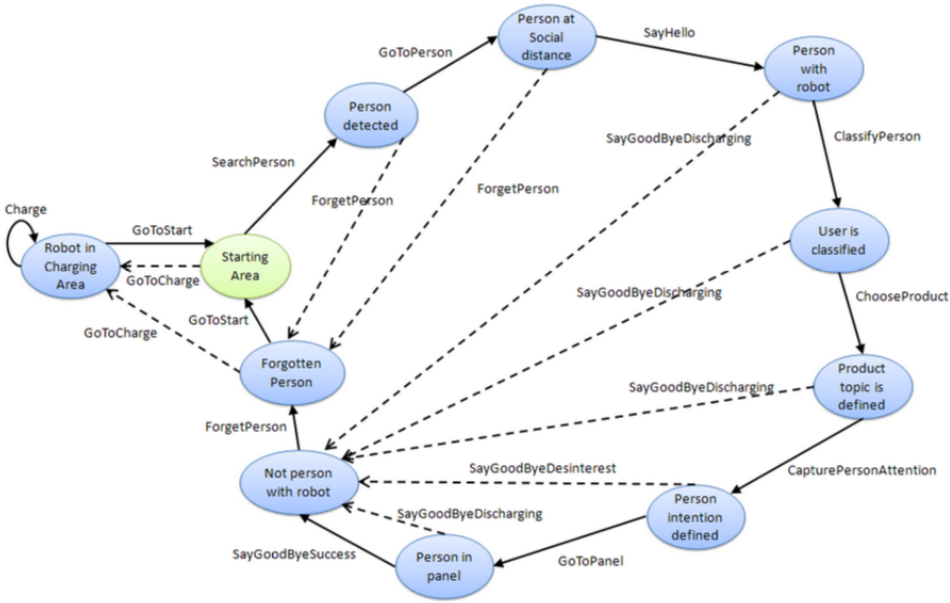

Fig. 2. The working scenario within the ADAPTA project

information stored at the symbolic and geometric levels relies on the agents. However, the Executive filters the information coming into the AGM graph, to maintain an internal representation of the outer world that is coherent with the domain theory. So far, there is no analogous mechanism to control the updates of the geometric representation, although some kind of filtering is locally done by the CompoNets. For example, could a perceived table be placed half buried into the room floor, in the internal geometric representation? Surely, some alerts must be triggered if an agent posts this information but, who should file the complaint? The question of whether a more centralized updating control is needed and of how it should operate is still under discussion.

III. Human-Robot Interaction within RoboCog: the ADAPTA scenario

In addition to plan transformation at a higher level, multi-modal interaction between humans and robots often needs quick, last minute adaptations due to unpredictable environment changes or human behavior. During dialogs, such adaptations are necessary when the human interrupts the conversation, unrequired inputs are provided in the form of backchanneling agreement or disagreement, or she shows other listener responses [7]. Other aspects of bodily behavior that are difficult to plan ahead are for example *behavior matching* and *synchrony* [19]. In such situations, transformational techniques based on constraint solvers where plans can be modified *on the fly* should be used [18], [22]. The proposed system addresses this continuous interaction behavior at a situational level, under the symbolic level provided by the AGM graph. Basically, this level is composed by symbolic attributes (e.g. emotional state, intention, ...). These attributes, that are provided with online computed predicates, are annotated on the nodes of the kinematic trees at the scene graph. The situational level provides the information needed to maintain a synchronised and fluent interaction with the human counterpart.

The proposed framework has been evaluated within the ADAPTA project. Before analysing the concrete features that the functionalities for HRI should fulfill, a brief presentation of the application developed for this project is provided.

A. The ADAPTA scenario

The goal of the ADAPTA project is the development of a social robot, named *Gualzru*, deployed in large Shopping Centers as a vendor. Fig. 2 shows the state diagram of this scenario. As depicted, Gualzru is firstly waiting on the Starting area (green state in Fig. 2) in the middle of an uncluttered corridor in the Shopping Center. The objective of Gualzru is to offer all people moving through the Shopping Center products and services. In fact, its aim is to drive potential consumers to an advertising panel that displays these products and services. As there are products for everybody, it could choose any person in the corridor. When Gualzru chooses a target, it moves towards that person. This displacement is very short (2-3 meters maximum) and should allow Gualzru waiting for the person in a static pose, facing her, but avoiding going very close to her (1,5-2 meters minimum). Thus, Gualzru could say "hello" to the person without scaring her.

If the person engages with it in this first contact, Gualzru will classify her into a group (using Gender and Age parameters) and will choose a Product Topic to offer, as Fig. 2 depicts. Product Topics provide Gualzru a general theme to speak with the person and to invite her accompanying it to the Panel. During this short conversation, Gualzru will be always ready to say goodbye to the user if she shows the intention of leaving the conversation or if the presented Product Topic is not interesting for her. It is important to consider that a person trying to continue a conversation after having shown no interest in it can be selected again as a target, but only after Gualzru has returned to the Starting area. On the other hand, Gualzru must also check its batteries level to say goodbye and move to the Charging area (close to the Starting area) if required.

If the person agrees on going with Gualzru to the Panel area, Gualzru moves to this area. There, it says the person goodbye and returns to the Starting area. After waiting for some time (i.e. the time the person requires to interact with the panel), Gualzru starts the process again to capture a new target. As before, if the batteries level is low, Gualzru moves to the Charging area to refill them (GoToCharge transition in Fig. 2).

B. The WinKinectComp component

The Gualzru robot uses a Kinect from Microsoft to extract data from the scene. This component implements a set of interfaces that are essential for the Person and Conversational compoNets, that are described below:

- MSKBodyEvent
  - This interface provides the list of bodies (with or without skeleton) in the scene.
- MSKFaceEvent
  - It generates the position and a list of features for the closest face detected by the Kinect.
- MSKASR
  - It provides the result of speech transcription to the Conversational compoNet.

The WinKinectComp component runs on a separate computer and the communication with the RoboComp components is carried out by using Ice [10]. Ice provides both a native client-server communication system. It also provides a publish-subscribe event distribution service named IceStorm[1], that decouples the connection among components. The WinKinectComp plays the publisher role while RoboComp components subscribe to their publications.

[1] http://www.zeroc.com/icestorm/

The WinKinectComp component has been developed using the Microsoft Kinect SDK. This SDK includes the management of asynchronous events related to people and audio detection. Each time a person or an audio source is detected, new events are triggered and data about these events are provided. Concretely, Kinect provides the 3D skeleton for the body and the coordinates for several facial features as the eyes, nose or mouth. Kinect generates also the most reliable speech transcription when using internal grammars. The IceStorm publication is then performed by processing these asynchronous events. Such processing includes feature selection and Ice data encapsulation.

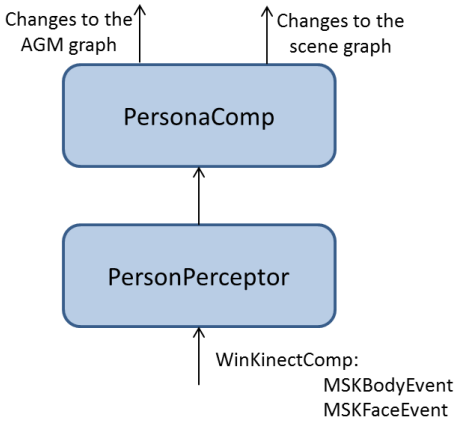

Fig. 3. Internal organisation of the Person compoNet

C. The Person compoNet

Within the Adapta scenario, the Person compoNet is in charge of detecting and tracking the human target during the execution of the whole process, until the person agrees or refuses to accompany the robot to the panel. Tracking is essential to avoid the robot keeping on speaking, even if the person has gone away. Furthermore, the Person compoNet will be responsible for classifying the person according to her gender and age. In the scenario at Fig. 2, the Person compoNet deals with the SearchPerson, GoToPerson, SayHello, ClassifyPerson, ChooseProduct and CapturePersonAttention processes.

Fig. 3 shows the internal structure of the compoNet. The PersonPerceptor component is in charge of acquiring the data from the WinKinectComp component and transforming them into a stream of feature vectors (person arrays) for PersonaComp. PersonaComp is in charge of choosing the human target, tracking her face and classifying her according to gender and age. Furthermore, the PersonaComp component implements the functionalities of the agent within this compoNet, being also in charge of modifying the scene graph and of proposing changes to the AGM graph.

1) PersonPerceptor: The PersonPerceptor component subscribes to two interfaces provided by the WinKinectComp component: the MSKBodyEvent and the MSKFaceEvent. This component works like a virtual sensor, using these data. Thus, the WinKinectComp is able to provide position and features of human bodies and faces, and the PersonPerceptor takes all these data to generate a 'person array' that is organised as follows: (i) A person whose skeleton and face has been detected (the Kinect sensor only detects the face of one person in the scene); (ii) People whose skeletons have been detected; (iii) People for whom no skeleton has been detected by the Kinect sensor.

The organisation of the person array fits the requirements of PersonaComp, and provides a direct mapping between the position on the array and the degree of interest that the component assigns to each person as a potential target. Each item in the array stores information about the distance to the robot. If a face has been detected for a person, the component evaluates if that face looks the camera and, if so, a normalised representation of the face is also stored.

2) PersonaComp: This component is in charge of (i) detecting and tracking the target; (ii) classifying the target according to gender and age; (iii) modifying the scene

graph within the geometric inner model; and (iv) proposing changes to the AGM graph. This last task implies that this component acts like an agent, detecting the next action to accomplish, according to the AGM graph, and proposing new changes to this graph when the action ends (with a positive or negative result). Hence, this component is continuously checking the graph to determine the next action to accomplish. When it detects one of these actions

→ SearchPerson
→ GoToPerson
→ SayHello
→ ClassifyPerson
→ ChooseProduct
→ CapturePersonAttention

then it activates the tracking function. This function will provide as output a proposal to change the graph. The possible proposals are

→ Person is detected
→ Person is classified
→ Person is lost

Furthermore, there is an action where this component does not track the person but can propose a change on the AGM graph

→ forgetPerson: this action occurs when the robot says ”goodbye” to the person. This action implies an initialisation of the graph state (i.e. a soft reset on the execution of the application).

Changes to the AGM are complemented with modifications on the scene graph. Fig. 4 shows how the geometric level includes new people (cylinders) when they are detected. For the AGM, however, there is only one target (the first detected person, correctly classified -middle row- as a young man).

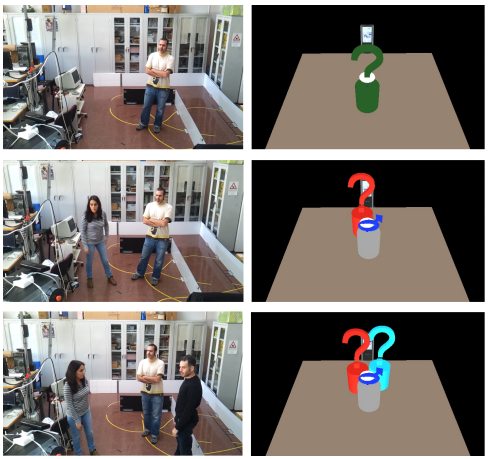

Fig. 4. Internalizing in the inner model from the Person perceiver

D. The Conversational compoNet

The distribution of the Conversational compoNet is based on the RoboCog architecture (Fig. 1). This CompoNet is connected to three main elements: a) the Executive, b) the WinKinectComp, and c) the Speech Generation. These three elements play specific roles in a conversation: the WinKinectComp senses audio data from the environment, while the Speech Generation is in charge of producing phrases. The Executive controls the overall conversational execution, stopping or starting the complete process if required.

Therefore, the Conversational processes the information provided by the WinKinectComp, generating phrases by means of the Speech Generation. All this process is totally driven by high-level goals, and each conversation can derive in different robot phrases. In the Adapta scenario, the robot speech depends exclusively on the product topic the robot is publicising, which depends on the age and gender of the user.

In this work, the speech recognition process is composed by two separate steps: transcription generation and comprehension (see Fig. 5). The first step processes the audio source and obtains the most reliable text transcription. This is completely carried out in the WinKinectComp using the Microsoft Kinect Speech SDK [2].

Transcription generation is performed by

[2]http://www.microsoft.com/en-us/download/details.aspx?id=14373

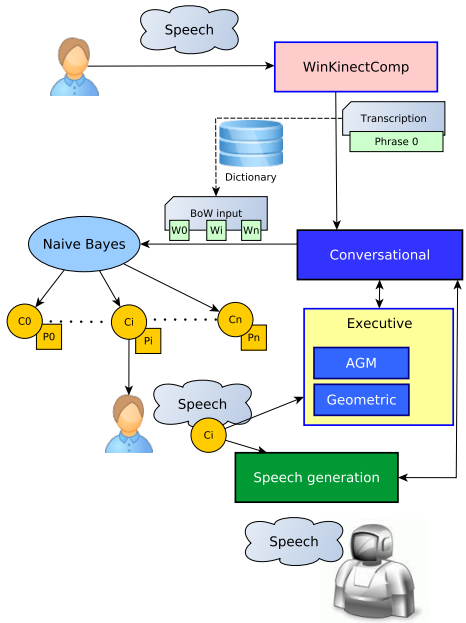

Fig. 5. Speech Recognition Procedure

using two internal key elements: the acoustic model and the language model. The acoustic model represents the probability of obtaining an input utterance x given a sequence of words w (transcription). It is directly provided by the Microsoft Speech SDK (for several languages such as Spanish or English). The language model scores the transcription w using the joint probability of the sequence of words. The probability for each word w_i depends on the list of previous words $w_{i-1}...w_{i-n}$.

The language model is generated by following the n-gram model [6], where n defines the number of words considered in the joint probability. The current proposal uses a 3-grams model generated from the COLA corpus [11]. Such corpus was selected because it includes informal ways of speaking. The 3-grams language model was then compiled using the Microsoft Speech SDK tools, obtaining a Kinect compatible grammar. The WinKinectComp uses this grammar to generate new transcriptions for each received audio that can be modelled with the grammar. That allows discarding environment noise.

Regarding the comprehension step, it is fully developed in the Conversational compoNet. Its uses as input the transcription generated with the WinKinectComp and assigns it a semantic label. This semantic information is then integrated in the system using the Executive. Thus, actions related to these semantic data can be performed. This step allows the robot understanding what the user expects from it after her speech. Some transcriptions can correspond to nonsense phrases (e.g. the user claims about the weather), but most of them are expected to affect the internal robot behaviour.

Within the Adapta scenario, several phrases with special meaning have been previously defined. The robot is able to answer questions about four topics related to the task: location of the Panel, requested service time, price of the service and extended information requirement. In addition to these topics, it is also necessary to detect when the user accepts or rejects the robot invitation. Therefore, comprehension is managed in this work as a classification problem with 6 different classes (4 questions and 2 user decisions).

The classification is performed using a Bag of Words (BoW) procedure [21] in conjunction with a Bayesian classifier. The dictionary for the BoW procedure is obtained with a variable selection process that removes useless words (as articles or connectors). Training and validation classifier sequences are generated using 750 sample phrases performed by more than 25 people. From these initial phrases, a grammar model is built for each one of the six classes. This model includes random variations. Finally, 1800 different phrases are generated (300 by class, using the obtained models) for training, and 600 additional ones for validation.

When a user generates a new speech. The WinKinectComp firstly generates the most reliable transcription and sends it to the Conversational. If the conversation is not

active (it depends on the current scenario), such transcription is discarded. Otherwise, the transcription is transformed using the BoW representation. If the obtained set of words does not include a minimum number of key words (included in the dictionary), the input phrase is directly labeled as nonsense. If not, the input set of words is processed using the Naive Bayes classifier and the set of output probabilities is studied. The input phrase is only classified as C_i when $P(C_i|w)$ clearly outperforms the rest of probabilities $P(C_j|w)_{i \neq j}$.

As a result of the classification, two different scenarios are identified: a) user question, and b) user decision. In both situations the compoNet requires the speech generation module to answer with an appropriate phrase. However, user decisions involve changes in the internal cognitive representation: the user will be labeled as interested or not interested in the Panel. Therefore, the Conversational would propose changes to the AGM graph through the Executive.

IV. Experimental results

Experimentation on this paper was carried out on a single day. People testing the system range from members of the project staff, who were used to the robot, to totally unrelated users. Tests were not conducted using the final version of Gualzru as the robot was not finished. A different platform was used instead. This platform is a modification of the Nomadics Nomad200 Tech (Fig. 6). This fact is important because the way of moving and, above all, the external appearance of the robot will be completely different for the completed Gualzru. The appearance is one of the main factors that influence the way the robot is perceived. Therefore, it influences the results. In any case, the robot moves, speaks, suggests to visit the panel and, according to provided outputs, says goodbye or accompanies the person to the panel.

The experiment involve a complete interaction with the robot for each person. Interaction ends when the person refuses to accompany the robot, or when the robot leaves the person by the Panel and says goodbye. Throughout the experiment, the technical performance of the robot is evaluated. When each test finishes, the person fills a questionnaire about it. The goals of the experiment are: (i) check the validity of the proposed system; (ii) get useful feedback in order to improve the final design and behavior of Gualzru robot.

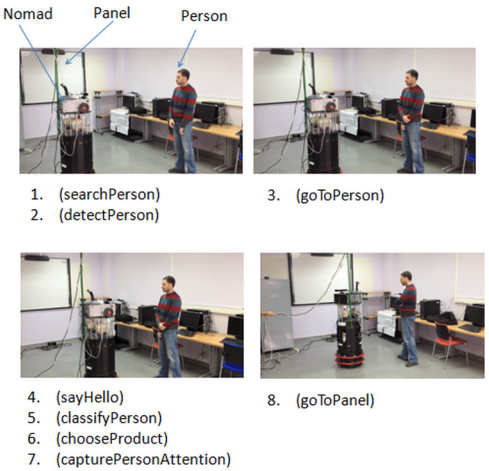

Fig. 6. Nomad200 driving the application with an user

The results obtained for a test group of 12 people are detailed below. Six of these people are familiar with the ADAPTA project, while the other six people are completely unrelated to the project. While the number of experiments is low to offer a complete evaluation of the system, it is high enough to offer a first impression, highlight advantages and drawbacks and suggest further improvements.

A. Evaluation of the HRI: Questionnaire

The experiment is evaluated using a questionnaire which responds to a model similar to that employed by Joosse et al. [12] to generate the database BEHAVE-II. Its main difference is that it has been created not from the point of view of the person observing the behavior of the user against the presence

of the robot, but from the point of view of the same user that interacts with the robot. In this sense, we can consider that collects influences of questionnaires of the Almere original model or the man-machine interaction. In particular, the questionnaire includes a collection of questions arranged in four blocks (navigation, conversation, interaction and general sensations) and two additional questions.

1) Navigation
   - Do you feel safe when the robot approaches you?
   - Does the robot invade your personal space?
   - Do you think robot movements are natural?
   - Have you stepped away from the robot during the interaction, because you feared you could collide?
2) Conversation
   - Have you understood what the robot told you?
   - Do you think the robot understood you?
   - Could you maintain a coherent conversation?
   - Do you think the robot has a pleasant voice?
3) Interaction
   - Did the robot get blocked during the interaction?
   - Do you think your interaction with the robot was natural?
   - Was the conversation fluent?
   - Did the robot seem to be controlled by a person?
4) Sensations
   - Did you enjoy the experiment?
   - Do you think the experiment was not interesting?
   - Would you like to repeat the experiment?
   - Would you recommend other people to interact with the robot?
5) Additional issues:
   - What task do you think the robot has to perform?
   - Do you think the robot finished its task? If not, what part of the task was not performed, in your opinion?

Questions were answered with a numeric value between 0 (no, not at all) and 5 (yes, of course).

B. Obtained results

Fig. 7 shows collected statistics about the responses to the questions in paragraphs 1-4 of the questionnaire.

With respect to the two additional questions for the user, the first question, 'What task do you think the robot has to perform?', was answered correctly by all users (guide to customers in shopping centers, advertising slogan, ...). About the second question, 'Do you think the robot finished its task? If not, what part of the task was not performed, in your opinion?', the robot successfully completed the task in 10 of the 12 cases. In the remaining two, the robot was not able to understand the person or was totally locked in one of the first actions of the working scenario (classification of people). In any case, the robot finished correctly the greeting and presentation, so the user came to know what the mission of the robot was.

V. DISCUSSION AND FUTURE WORK

Despite of the insufficient number of tests, there are some conclusions that can be extracted from the performed experiments.

- Navigation.- People perceived robot movements as safe but unnatural. The robot did not invade people personal space.
- Conversation.- The voice of the robot was understandable, but unpleasant. Regarding the robot comprehension, people interpreted that Nomad200 did not understand them.
- Interaction.- Human-robot interaction was not perceived as fluent and presented several blocking episodes. Some of these episodes were due to classification problems, but most of them were caused by the conversational module. On the other hand, users described Nomad200 as an autonomous robot (without external human control), which is a very positive point.
- General feeling.- Most of the users think the experiment was interesting,

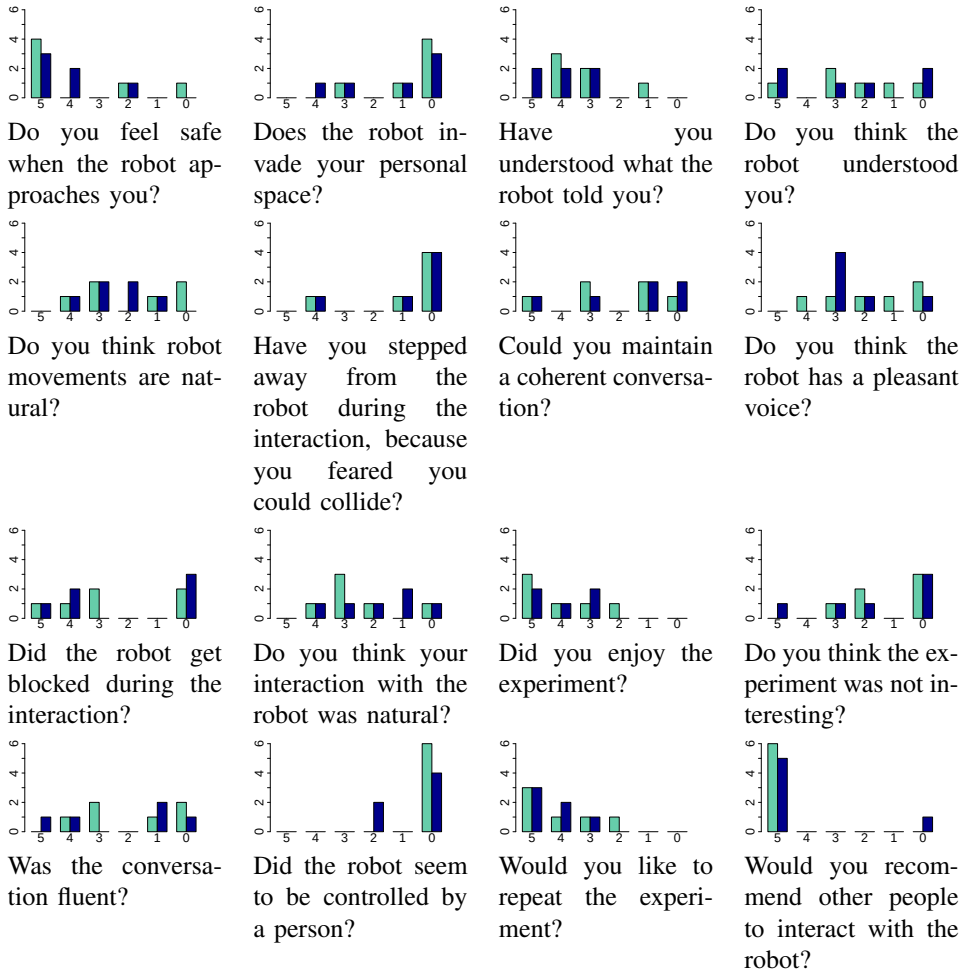

Fig. 7. Responses to the questionnaire for people familiar with the ADAPTA project (blue) and not (green).

and they would like to repeat the process. Moreover, most of them would recommend the process.

These results show that there are two keypoints that affect human-robot interaction: robot's appearance and robot's responses. Nomad200 moves and speaks in an unnatural way. That makes (along with the lack of facial expression) the user to feel uncomfortable during some parts of the process, and specially, during blocking situations. These undesired circumstances happened when Nomard200 responses did not correspond to the expected ones, due to comprehension or classification failures. In order to improve the interaction, Nomard200 should play a more active role during the process, detecting what the user expects from it in each moment and acting when necessary.

On the short-term, the architecture will be extended to include facial emotions and gesture recognition modules. The hardware for showing simple facial expressions should also be built and integrated on the real Gualzru robot. On the mid-term, future work will focus on modifying the structures that currently manage the inner representation to merge geometric and symbolic knowledge

inside the same representation. Both representations should be wrapped up into an adequate software package at RoboComp. This will allow the Executive to generate copies of the representation. These copies will be used by Emulation compoNets to launch simulations of alternative branches of the plan. These Emulators will access the representation using the same interfaces that are used to access the action and perception modules, while maintaining the capability to compute and update the logical symbols and predicates over the geometric structures.

ACKNOWLEDGMENT

This paper has been partially supported by the Spanish Ministerio de Economía y Competitividad TIN2012-TIN2012-38079-C03 and FEDER funds, and by the Interconecta Programme 2011 project ITC-20111030 ADAPTA.

REFERENCES

[1] M. Ali, Contribution to decisional human-robot interaction: towards collaborative robot companions, PhD Thesis, Institut National de Sciences Appliques de Toulouse, France, 2012

[2] A. Bauer, D. Wollherr and M. Buss, Human-robot collaboration: a survey, Int. Journal of Humanoid Robotics, 2007

[3] M. Beetz, D. Jain, L. Mösenlechner, M. Tenorth, Towards performing everyday manipulation activities, Robotics and Autonomous Systems, 2010

[4] G.V. Bodenhausen, and K. Hugenberg, Attention, Perception, and Social Cognition. Philadelphia, PA: Psychology Press, 2008

[5] P. Bustos, J. Martínez-Gómez, I. García-Varea, L. Rodríguez-Ruiz, P. Bachiller, L.V. Calderita, L.J. Manso, A. Sánchez, A. Bandera and J.P. Bandera, Multimodal interaction with Loki. Proc. of the XIV Workshop on Physical Agents, 2013.

[6] F. Jelinek, Statistical methods for speech recognition, MIT press, 1997

[7] D.J. Fujimoto, Listener responses in interaction: a case for abandoning the term backchannel, Journal of Osaka Jogakuin 2YColl 37, 35-54, 2007

[8] E. Gat, On three-layer architectures. Artificial Intelligence And Mobile Robots, 1997

[9] B. Hayes-Roth, and F. Hayes-Roth, A Cognitive model of planning, Cognitive Science 3 (4), 275-310, 1979

[10] M. Henning and M. Spruiell. Distributed programming with ice. ZeroC Inc. Revision, vol. 3. 2003

[11] K. Hofland, A. M. Jørgensen, and E. Drange, and A. Stenström, COLA: A Spanish spoken corpus of youth language, 2005

[12] M. Joosse, A. Sardar, M. Lohse and V. Evers, BEHAVE-II: The revised set of measures to assess users' attitudinal and behavioral responses to a social robot, Int. J. Social Robotics 5(3), 379-388, 2013

[13] A. Kirsch, T. Kruse, and L. Mösenlechner, An integrated planning and learning framework for human-robot interaction, 4th Workshop on Planning and Plan Execution for Real-World Systems (held in conjunction with ICAPS 09), 2009

[14] L.J. Manso, Perception as Stochastic Sampling on Dynamic Graph Spaces, PhD Thesis. Cáceres Polytechnic School, University of Extremadura, 2013

[15] A. Müller, A. Kirsch, M. Beetz, Transformational planning for everyday activity, In Proceedings of the 17th International Conference on Automated Planning and Scheduling (ICAPS'07), 2007

[16] A. Nijholt, D. Reidsma, H. van Welbergen, H. J. A. op den Akker, and Z.M. Ruttkay. Mutually coordinated anticipatory multimodal interaction. In Nonverbal Features of Human-Human and Human-Machine Interaction, volume 5042 of LNCS, pages 70-89, Berlin, 2008. Springer Verlag.

[17] E. Quintero, V. Alcázar, D. Borrajo, J. Fernández-Olivares, F. Fernández, A. García Olaya, C. Guzman, E. Onaindia, D. Prior, Autonomous Mobile Robot Control and Learning with the PELEA Architecture, Proc. Automated Action Planning for Autonomous Mobile Robots (PARM 2011), 2011

[18] D. Reidsma, H. Welbergen, and J. Zwiers, Multimodal Plan Representation for adaptable BML scheduling. Intelligent Virtual Agents, 11th International Conference, LNCS vol. 6895, pp 296-308, 2011

[19] R.C. Schmidt, and M. Richardson, Dynamics of interpersonal coordination, In A. Fuchs and V. Jirsa (eds) Coordination: Neural, Behavioral and Social Dynamics, vol. 17 of Understanding Complex Systems, pp 281-308 Springer Berlin Heidelberg, 2008

[20] M. Tomasello, M.Carpenter, J. Call, T. Behne, T. and H. Moll, Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences 28 (5), 675-691, 2005

[21] H. M. Wallach, Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning, 977-984, 2006

[22] J. Zwiers, H. Welbergen, D. Reidsma, Continuous interaction within the SAIBA framework, In 11th Int. Conf. on Intelligent Virtual Agents, LNCS vol 6895, 324-330, 2011