

Multimodal Human-Robot Interaction Framework for a Personal Robot

Javi F. Gorostiza, Ramón Barber, Alaa M. Khamis, María Malfaz
Rakel Pacheco, Rafael Rivas, Ana Corrales, Elena Delgado and Miguel A. Salichs

Abstract— This paper presents a framework for multimodal human-robot interaction. The proposed framework is being implemented in a personal robot called Maggie, developed at RoboticsLab of the University Carlos III of Madrid for social interaction research. The control architecture of this personal robot is a hybrid control architecture called AD (Automatic-Deliberative) that incorporates an Emotion Control System (ECS) Maggie's main goal is to interact in a natural way and establish a peer-to-peer relationship with humans. To achieve this goal, a set of human-robot interaction skills are developed based on the proposed framework. The human-robot interaction skills imply tactile, visual, remote voice and sound modes. The multi-modal fusion and synchronization are also presented in this paper.

I. INTRODUCTION

In recent years, human-robot social interaction has attracted considerable attention by the academic and the research communities. A social robot [1] has attitudes or behaviors that take the interests, intentions or needs of the humans into account. This robot must be able to interact with humans by following the social rules attached to its role. Bartneck and Forlizzi define a social robot as an autonomous or semiautonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact [2], page 2. The multimodality is considered as a main and an unquestionable feature of human-robot social interaction. Multimodality means providing the user with more than a single mode of interaction. Multimodal interfaces allow users to move seamlessly between different modes of interaction, from visual to voice to touch, according to changes in context or user preference. These interfaces have the advantage of increased usability and accessibility. Usability determines the overall utility of the system. It also determines the extent to which an interface supports its users in completing their tasks efficiently, effectively, and satisfactorily. In multimodal interfaces, the weaknesses of one modality can be offset by the strengths of another. For example, a person can order his/her personal assistant robot whose multimodal interface using gesture-based interface in a noisy environment where verbal communication can not be worked efficiently. Accessibility determines how easy it is for people to interact with the robot. The multimodal interfaces provide increased accessibility. For example, visually impaired users can rely on the voice modality while hearing-impaired users can use the visual modality. Multimodality implies the problem of integration and synchronization of different communication modalities both in perception and expression. Many robotic platforms have been built with different design considera-

tions, control architectures and capabilities to study human-robot social interaction. Kismet [3] and Leonardo [4] developed at MIT have an emotional reactive control architecture that integrates the visual and audio modes. While Kismet is able to react to human voice and movements simulating an infant behavior, Leonardo can detect gaze and non-verbal signal like turn taking in collaborative tasks. HERMES, an experimental robot of anthropomorphic size and shape developed at Bunderswehr University of Munich. This robot has hybrid control architecture (deliberative and reactive) and integrates remote mode (via internet), dialogs handling Natural Language Processing (NLP) and combination of vision and touch during the tasks of giving and taking objects [5]. Robonaut is a joint DARPANASA project designed to create a humanoid robot equivalent to humans during space walks activity. This robot is equipped with human-like hands and television camera eyes and has the option of rolling around Earth. This robot has been designed to assist astronauts in extra-vehicular activities. To do so, the robot is capable to handle natural language dialogs. Its control Architecture allows enhanced skills like perspective taking. Two models of human perspective taking are used in this architecture: jACT-R/S based on human representation models and Polyscheme based on human reasoning process [6]. Robovie is a humanoid robot that can communicate with humans and is designed to participate in human society as a partner [7]. This robot works by means of a behavior-based architecture. It responds to tactile events with predefined simple behaviors. Sparky is a social robot that uses both facial expressions and movements to interact with humans [8]. Rubi is another anthropomorphic robot with a head and arms designed for research on real-time social interaction between robots and humans [9]. Robota is a sophisticated educational toy robot designed to build human-robot social interactions with children with motor and cognitive disabilities [10]. In the Lino project, a robot head with a nice, cute appearance and emotional feedback can be configured in such a way that the human user enjoys the interaction and will more easily accept possible misunderstandings [11]. Other social robot designs rely on computer graphics and animation techniques. Vikia [12], Valerie Roboceptionist [13], Grace (Graduate robot Attending a ConferencE) or George [14] are some examples for computer graphic-based social robots. They all handle natural language dialogs and merge image animation with speech. All these projects pretend to develop robots that function more naturally and can be considered as partners for the human not just as mere tools. These robots need to interact with human (and perhaps with each other) through

similar ways by which humans interact with each other. The paper describes a framework for multimodal human-robot interaction. The remainder of the paper is structured as follows: in section 2, a brief description of R2H project is provided. Section 3 describes the interaction modes in the personal robot Maggie. The control architecture is presented in section 4 followed by examples for implemented skill in section 5. The integration between the implemented skills is described in section 6. Finally conclusions and future work are summarized in section 6.

II. PEER-TO-PEER ROBOT-HUMAN INTERACTION: R2H PROJECT

Traditionally, the human robot interaction systems are based on a master-slave idea. According to this idea, the human operators role is to supervise and give commands to the robot, while the robots role is to accomplish those tasks and eventually, to give the necessary information to the operator. The robot essentially acts as a tool used by the operator. In these systems, the interaction with the human appears to be a limiting factor that reduces the robot's autonomy. The goal of R2H (Robot-To-Human) project is to develop social robots with a high degree of autonomy. The robot's behaviors will be based on their own impulses and motivations. These motivations are the mechanisms of the robot to keep certain internal variables, related to its necessities, near to an ideal level. Inside the control architecture of the robot, human-robot interaction will be organized using the same principles of the interaction of the robot with the rest of the world. In the traditional control architecture, the human role is above the robot role. The new philosophy is to keep the human at the same level as any other environment object. This new approach to the human-robot interaction might be quite interesting for some kind of new robots, such as robots interacting peer to peer with humans, entertainment robots, teaching robots and even, therapeutic robots.

III. INTERACTION MODES IN THE PERSONAL ROBOT *Maggie*

This section explains how Peer-To-Peer Multimodal Interaction is implemented in Maggie, the Personal Robot Developed by the RoboticsLab in the University Carlos III of Madrid. We specify the relationships between the different Interaction Modes and Maggie's Hardware and Software Architectures.

To explain the different interaction modes, we take into account the information flux, that is, if the information goes from or to the user. This robot has been completely presented in [15] like a robotic platform for Human-Robot Social Interaction. We can see in Fig.1 Maggie has an artistic design of a 1.35 meters tall girl-like doll. It incorporates different interaction modalities such as verbal communication, emotion expression through head/arm/eyelids movement and audiovisual expression.

A. Human To Robot Information Flux

1) *Visual Mode: Proxemic and Kinesic perception:* The Visual Mode consists of every visible expression. This mode

is divided in kinesic: body gestures, and proxemic: body placing in the communication system.

The role of the proxemic and kinesic expression in the human-human interaction has been studied in several works [16] , [17]. It has been established the importance of body movements in the communication act because it contains a lot of information that flows very quickly. In [16] Bird-whistell argues that the 65% of the information in a human-human interaction is non-verbal. Visual gestures shows human thoughts, mood state, replaies, complements, accents and adjust verbal information.

Visual mode perception in Maggie is accomplish by the base sensors and a Webcam (Fig. 1) The base is equipped with 12 infrared optical sensors, 12 ultrasound sensors and 12 bumpers. Above the base, a laser range finder (Sick LMS 200) has been added. These sensors allows Maggie to identify environment entities like doors, walls, obstacles, humans, etc.

In the mouth, Maggie has incorporated a Webcam for visual detection and identification of environment objects. Near abdomen, a semi-opaque black sphere is added. This sphere is equipped with a color camera for future people tracking.

2) *Voice Mode: Automatic Speech Recognition (ASR):* This mode is in charge of verbal Human-Robot Communication. In her chaste Maggie incorporates a tablet-PC where Bluetooth enabled wireless microphone is connected. Maggie counts with an Automatic Speech recognition (ASR) module that aims at converting spoken words into text. There are many commercial automatic speech recognizers [18]. Dragon Naturally Speaking (DNS) Client SDK v.3.0 developed by Scansoft has been used as speech recognizer. The DNS engine is a speaker-dependent engine, which means that each user has to train the engine before use it. When ASR server is activated it listens continuously to the audio port and tries to interpret the received utterance into one sentence text. Once an utterance is detected, a socket sends the recognized text with an accuracy value to a specific net port. The embedded PC in the robot's base uses this port to take the recognized text and its accuracy value as the entries of the Dialog Manager System or Dialog Module.

3) *Tactile Mode: Skin Sensors and Tactile Screen:* Maggie is able to detect tactile events in two different ways: by means of her Tactile System distributed over her enclosure and by means of the tactile screen in her tablet-PC.

The Tactile System consists of several invisible capacitive sensors installed in the robot casing. There is one sensor on each shoulder, one on the top of the head, two on the chest, two close to the abdomen, two on the upper part of torsos back and three on each arm. Each sensor has an extensive active zone (5 cm²) and is activated by human touching (hand, cheek, arm, etc.) close to the active zone. Through the tactile screen, Maggie is able to detect drawing gestures, and mouse events.

B. Robot To Human Information Flux

1) *Visual Mode: Proxemic and Kinesic expression:* In the proxemic side, Maggie is able to express herself changing

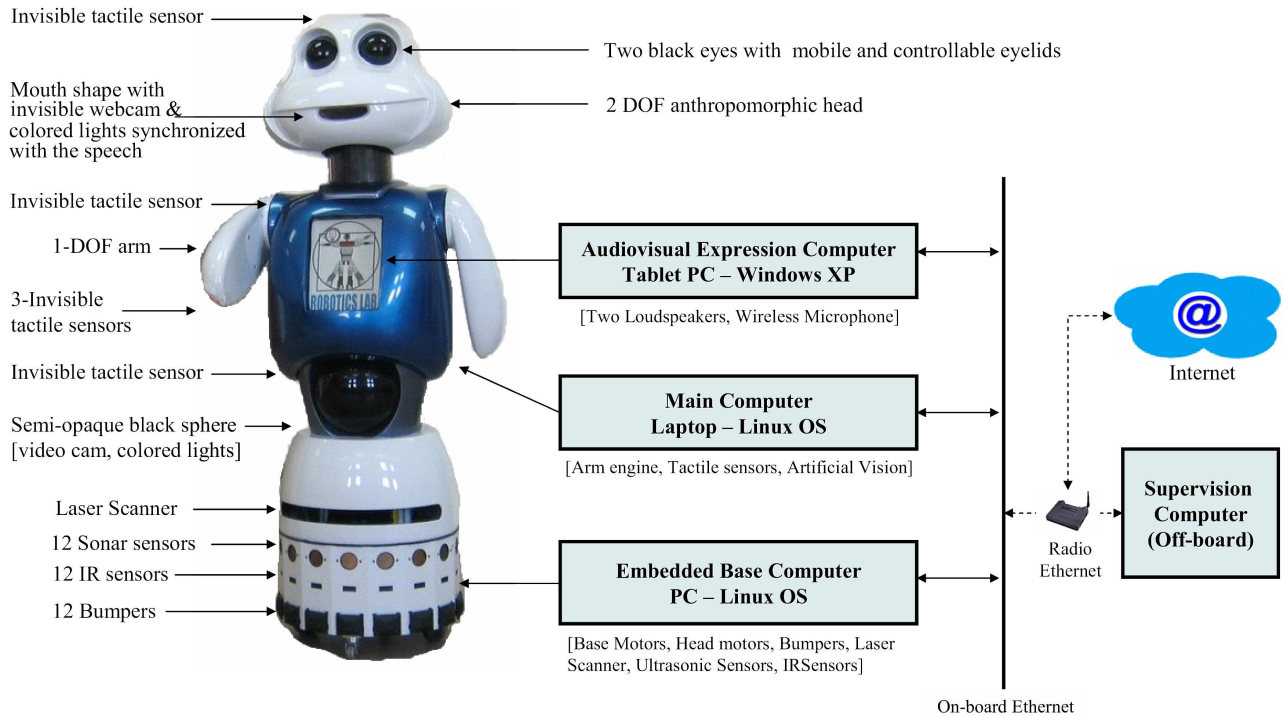


Fig. 1. Maggie Hardware Architecture

her position relative to the user. As her base is motorized by two differentially actuated wheels and a caster wheel on both sides, Maggie has a high degree of mobility. In the kinesic side, Maggie counts with a two DOF head: left/right and up/down; two black eyes with two mobile and controllable eyelids; and two 1-DOF arms without end effectors are built in the central part of the platform. These DOF allow Maggie to express herself through different kinesic gestures.

2) *Voice Mode: Text-to-Speech (TTS)*: Speech synthesis is realized by the tablet-PC when two speakers are connected. The Text-to-Speech Module is implemented in form of a server. This server is continuously reading the text sentence data in a specific port and converts it into audible speech. When the Dialog Module decides to synthesize a specific text sentence in speech, it sends a socket message to a specific port at which the TTS server is running. The TTS server encompasses VTxtAuto (VoiceText 1.0 Type Library) to generate speech from text. VTxtAuto is developed by Microsoft and distributed as part of the Speech API Software Development Kit (SDK). The VTxtAuto VTxtAuto Object allows the user to control certain parameters of voice synthesis such as the speech speed, pitch or volume.

3) *Audiovisual Mode: Images and Sound expression*: In her mouth, Maggie is equipped with several LEDs that are speech-synchronized. The semi-opaque sphere integrates colored lights for visual expression. The tablet-PC also provides audiovisual expressions through the screen photographs, video or/and animation images while

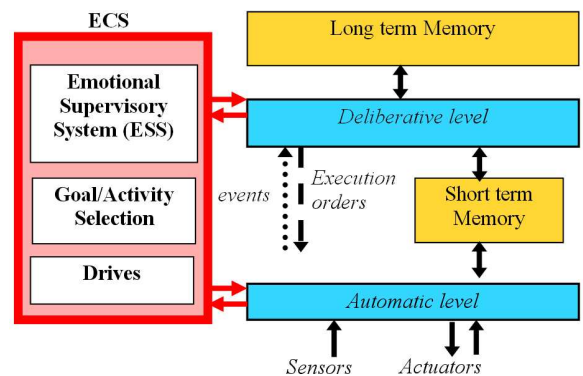


Fig. 2. Automatic-Deliberated Architecture

playing sounds and music. The software used to accomplish this is Pure Data and Graphic Environment for Multimedia (pd-gem), which is an open source audiovisual software tool.

IV. AD ARCHITECTURE SKILLS AND MULTIMODALITY

Peer-to-Peer Human-Robot Multimodal Interaction implies the presence of empathy between both, the user and the personal robot. To achieve this empathy we have chosen a hybrid control architecture called AD (Automatic-Deliberative) that is based on a psychological human model [19]. As shown in Fig. 2, the AD architecture is a two level architecture based on skills. A skill represents the robot's ability to perform a particular task. They are all built-in robot action and perception capacities. In the deliberative level there are skills capable of carrying out high level

tasks: planning, word model management, etc while at the automatic level there are reactive and sensory skills.

An emotional control system (ECS) [20] has been added to the AD architecture. Inside the ECS, there are three different modules: Drives, Activity Selection and Emotional Supervisory System (ESS). The Drives module is the one that controls the basic drives of the robot. The Activity Selection module on the other hand, determines goals and action tendencies of the robot. Finally, the ESS module generates the emotional state of the robot.

As mentioned previously, the AD Architecture is based on robotic skills distributed in two levels. In the automatic level, a skill can acquire sensory information of one or more sensors and take actions on one or more actuators. The skill structure is modular. A complex skill can be generated by combining different simple skills. For multimodal human-robot interaction, we have developed simple skills that provide different interaction modes described above, and then other complex skills can be easily constructed by integrating these simple skills. The complex skills can also be integrated in a more complex skill and so on.

V. SKILLS EXAMPLES

In this section some of simple and complex skills examples are presented.

A. Greeting Skill

This complex skill incorporates two simple skills: Tactile Skill and Arm Skill. The first one takes a skin tactile sensor event and shares an associated software event so any skill in the whole architecture is able to use this event and gives it meaning. The Arm Skill waits for an event associated to a shoulder capacitive sensor contact with the user hand and rise the arm. Then it waits for a timeout of hand tactile event for handshaking. When the user touch the hand, the robot handshakes and then the Arm Skill lowers the arm and finish.

B. Face Recognition and User Identification Skill

The user identification allows Maggie to memorize a set of well-known human individuals as identities in her world model. Face detection skill has been implemented using opencv-0.92 framework for image processing. The system is first trained using a neural network and several images samples. After the training phase, the system is able to identify a face using its data base by a rate of 90% and in an average time of about 2 secs. The user face has to be between 40cm and 3m and looking directly to Maggie's webcam in a well illuminated environment.

C. Dialogue Skill

The annotating dialog corpuses have been studied by many researchers [21]. In the draft of DAMSL (Dialog Act Markup in Several Layers) [22], utterance are tagged according to four main categories: communication status, information level, forward looking function and backward looking function. These categories describe the functions of

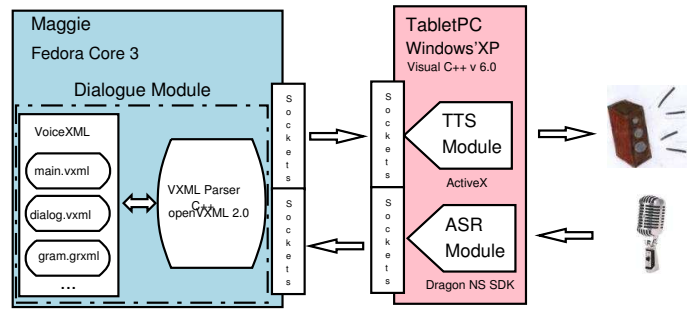


Fig. 3. Software Architecture for the Voice Platform

utterance mainly in terms of the speaker's intentions and the speaker's and hearer's obligations. But this a-priori made utterance structure implies a reactive speech response and we are looking for an adaptive dialog.

We use VoiceXML as a language of dialog representation. Using a Markup Language like dialog representation we are changing the involved functions of dialog flow to data representation. Then it is possible to make this data at runtime, that is, while the dialog is executing, and there we can incorporate the adaptation and learning features needed in a natural interaction.

A grammar markup language is used to specify the words and word syntax structures that the Cognitive Dialog Manager has to wait in the user utterance. The grammars are specify in specials files and can be activated or deactivated in runtime. Also the creation of these files is also possible in runtime. This allows Maggie to incorporate new items and new syntax structures while the dialog is executing.

As shown in Fig. 3, In Maggie's speech interface, the Dialog System closes the loop between recognized speech and synthesized voice. The dialog is implemented as a set of forms and menus. In both forms and menus there are a set of fields that has to be completed during the dialog interaction, but in the case of menus the prompts incorporates a fixed set of dialog topics.

D. Audiovisual Interaction Skill

This Skill merges screen tactile events with audiovisual actions. It presents to the user a GUI implemented with Qt designer tool. The GUI application displays a background image that is divided in two zones: center and edges (Fig. 4) The center zone mouse events are interpreted by the robot as pain tactile events, the edge zone mouse events are interpreted as pleasant gesture. The user can click or drop in the displayed image and te skill interprets the mouse event as a tactile gesture. A magnitude of the tactile gesture effect is also calculated depending on the velocity and position of the mouse event. In Table 1 it is showed a relation between each of the mouse events and the tactile effect.

E. Non-verbal visual expresion Skill

In the expression side, Maggie expresses herself to the human using non-verbal visual signals: gestures through head/arm/eyelids movements. So far, the gesture expression



Fig. 4. Tactile Skill: Image display in TabletPC

TABLE I
MOUSE EVENTS AND THEIR TACTILE EFFECT

Mouse Event	Zone	Tactile Effect
click	center	push
click	edge	tickle
small drag	center	cares
big drag	edge	scratch

skills can be designed through low-level C++ programs. A Java-based body gesture design interface is currently being developed for precise movement design and synchronization with other modes, such as voice mode. In this graphic interface, the user is going to be able to design the movement of each Degree Of Freedom (DOF) with custom precision. Notice that each movement designed is represented as a data matrix, not as a function. The transformation of closed functions in open data allows the overall system to incorporate adaptation and interoperability in a easier way.

F. Dancing Skill

Maggie has demonstrated its ability for closely cooperative dancing with humans. Maggie is able to change its movements as respond for events detected by tactile sensors as results of partner touching. A video demo is available at [23].

VI. MULTIMODAL INFORMATION INTEGRATION

To close the interaction loop where the robot response to the human requests and makes her own requests to the human, a world model is necessary. The information detected by the sensory skills has to be in a format and representation that allows to all the system components (other skills) use it. At the same time, motor skills has to be in a format and representation that other skills could call, activate or deactivate in execution time.

To achieve this integration, first of all, we are formulating every robot skills as data instead of functions. It can be accomplished using script languages, for example, standard markup languages. The advantages of using script programming languages are that they are very intuitive for the

developer and that the script can be autogenerated by other scripts, so adaptation and learning can be solved this way. Many researchers are using markup languages for improving human-machine interfaces. In [24] there are several markup language definitions that are used to integrate different animation characters DOF improving human-machine interfaces.

The W3C (World Wide Web Consortium) [25] provides specification and standards for many new markup languages designed for facilitating the development of multimodal interfaces.

The most important markup languages that are liable to be used in a Human-Robot interaction application can be enumerated as follows:

- **InkML** (Ink Markup Language): It represents the data of any notational language application like handwriting, gestures, sketches or music.
- **CCXML** (Call Control eXtensible Markup Language): It is intended to support telephony call control in HMI.
- **SRGS** (Speech Recognition Grammar Specification): It allows to detect special patterns of words or utterances.
- **InkXML** (Ink Markup Language): This language serves as the data format for representing ink entered with an electronic pen.
- **VXML** (Voice eXtensible Markup Language): This specification allows a dialog representation based in forms and menus.
- **NLSML** (Natural Language Semantic Markup Language): It is designed for the Speech Interface Framework. It is intended for use by systems that provide semantic interpretations for a variety of inputs, including but not necessarily limited to speech and natural language text input.
- **SSML** (Speech Synthesis Markup Languages): It allows an exhaustive text audio synthesis control.
- **EMMA** (Extensible MultiModal Annotation markup language): It represents raw input signal as speech, pen or keystroke input, gesture, etc. It integrates the different signal interpretations of the different media interpreters, speech or gesture recognition, semantic interpreters, etc.
- **SMIL** (Synchronized Multimedia Integration Language): It integrates different streaming media type to each other: images with text and video or audio.
- **SISR** (Semantic Interpretation for Speech Recognition): It describes the meaning of a Natural Language utterance. By "meaning", it is understood the rules of information computation to return to an application. It adds semantic tags to SRGS grammars.
- **SCXML** (State Char eXtensible Markup Language): It provides a state machine notation for control abstraction.

Fig. 5 shows the relationship between these languages. For example, NLSML contains the VoiceXML language, which contains SRGS language at the same time. While, SRGS represents low abstraction level information, NLSML represents higher one. Low abstraction level information

Markup Languages are related to the different robot modes directly.

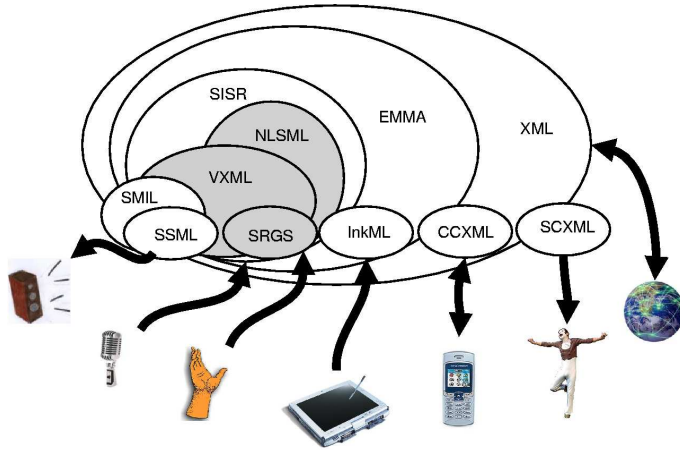


Fig. 5. Markup Language general scheme for Human-Robot Interaction

The semantic interpretation of the input is here considered as the performance of subsequent processing of the raw text. This semantic interpretation can be represented following SISR, described above. The output of the semantic interpretation processor may be represented using NLSML. These representations will allow resolving deictic and anaphoric reference.

VII. CONCLUSIONS AND FUTURE WORKS

Maggie can be considered as partner for the human not just as mere tool. This platform incorporates interaction multimodalities and has an attractive physical appearance social robot with which the human always enjoys interacting. This paper presented a framework for multimodal human-robot interaction, which is being implemented in the developed platform.

Of interest, for future work, is the development of new interactive scenarios based on the proposed framework. One of this can be a game scenario for Maggie training and choreography programming, where the user indicates a sequence to the personal robot and the robot responds and takes her own initiatives.

VIII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the funds provided by the Spanish Ministry of Education and Science (MEC) through the projects named "Personal Robotic Assistant" (PRA) and "Peer to Peer Robot-Human Interaction" (R2H), of MEC (Ministry of Science and Education)

REFERENCES

- [1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics & Autonomous Systems, Special issue on Socially Interactive Robots*, 42 (3-4), no. 42, pp. 143–166, 2003.
- [2] C. Bartneck and J. Forlizzi, "A design-centered framework for social human-robot interaction," in *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, Kurashiki, Okayama, Japan, 2004.
- [3] C. Breazeal and B. Scasselatti, "Infant-like social interactions between a robot and a human caretaker," MIT, Artificial Intelligence Laboratory, Tech. Rep., 1998.
- [4] C. Breazeal, A. Brooks, D. Chilongo, J. Gray, A. Hoffman, C. K. H. Lee, J. Lieberman, and A. Lockered, "Working collaboratively with humanoid robots," *ACM Computers in Entertainment*, vol. 2, no. 3, July 2004.
- [5] R. Bischoff and V. Graefe, "Hermes, a versatile personal robotic assistant," *IEEE*, vol. 92, no. 11, pp. 1759–1779, November 2004.
- [6] N. Cassimatis, J. Trafton, M. Bugajska, and A. Schultz, "Integrating cognition, perception and action through mental simulation in robots," Naval Research Laboratory, Technical Report, 2004.
- [7] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Development and evaluation of interactive humanoid robots," *IEEE*, vol. 92, no. 11, pp. 1839–1850, November 2004.
- [8] M. Scheeff, "Experience with sparky: A social robot," in *Proceedings of the Workshop on Interactive Robot Entertainment*, 2000.
- [9] J. C. B. Fortenberry and J. Movellan, "Rubi: A robotic platform for real-time social interaction," in *Third International Conference on Development and Learning (ICDL'04)*, 2004.
- [10] A. Billard, K. Dautenhahn, and G. Hayes, "Robota: Clever toy and educational tool," *Robotics & Autonomous Systems*, no. 42, pp. 259–269, 2003.
- [11] A. Breemen, M. Nuttin, and K. Crucq, "A user-interface robot for ambient intelligent environments," in *Proceedings of the 1st International Workshop on Advances in Service Robotics, ASER03*, Bardolino, Italy, 2003.
- [12] I. N. A. Bruce and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *IEEE International Conference on Robotics and Automation (ICRA'02)*. IEEE, 2002.
- [13] R. Gockley et al., "Designing robots for long-term social interaction," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2005.
- [14] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. Schultz, and J. Wang, "Grace and george: Social robots," in *AAAI Proceedings of AAAI'04. Mobile Robot Competition Workshop*, no. Technical Report WS-04-11, August 2004, pp. 15–20.
- [15] M. A. Salichs et al., "Maggie: A robotic platform for human-robot social interaction," in *IEEE International Conference on Robotics, Automation and Mechatronics (RAM 2006)*. IEEE, 2006.
- [16] F. Davis, *Inside Intuition-What we know about Non-Verbal Communication*. McGraw-Hill Book Co., 1971.
- [17] M. Kanpp, *Non-verbal communication*. Harcourt College Pub, 1996.
- [18] I. Even-Zohar. (2005, December) A general survey of speech recognition programs. [Online]. Available: <http://www.tau.ac.il/itamarez/sr/survey.htm>
- [19] R. Barber and M. A. Salichs, "A new human based architecture for intelligent autonomous robots," in *The Fourth IFAC Symposium on Intelligent Autonomous Vehicles*, 2001, pp. 85–89.
- [20] M. Malfaz and M. A. Salichs, "A new architecture for autonomous robots based on emotions." Lisbon, Portugal: Fifth IFAC Symposium on Intelligent Autonomous Vehicles, July 2004.
- [21] E. Atwell, "Machine learning from corpus resources for speech and handwriting recognition," *Thomas J. and Short M (editors), Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, 1996.
- [22] J. Allen and M. Core. (1997) Draft of damsl: Dialog act markup in several layers. [Online]. Available: <http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/>
- [23] (2005) Roboticslab: Maggie dancing video demo. [Online]. Available: <http://roboticslab.uc3m.es/roboticslab/gallery.php>
- [24] H. P. Mitsuru Ishizuka, *Life Like Characters: Tools, affective Functions, and Applications*. H. Prendinger; M. Ishizuka, 2004.
- [25] World wide web consortium (w3c). [Online]. Available: <http://www.w3c.com/>