# A Conversational System for Multi-Session Child-Robot Interaction with Several Games

Ivana Kruijff-Korbayová[1], Heriberto Cuayáhuitl[1], Bernd Kiefer[1], Stefania Racioppa[1], Piero Cosi[2], Giulio Paci[2], Giacomo Sommavilla[2], Fabio Tesser[2] Hichem Sahli[3], Georgios Athanasopoulos[3], Weiyi Wang[3], Valentin Enescu[3], Werner Verhelst[3] Lola Cañamero[4], Aryel Beck[4], Antoine Hiolle[4] Raquel Ros Espinoza[5], Yiannis Demiris[5]

[1] Language Technology Lab, DFKI, Saarbrücken, Germany
ivana.kruijff@dfki.de
[2] Istituto di Scienze e Tecnologie della Cognizione, ISTC, C.N.R., Italy
[3] Interdisciplinary Institute for Broadband Technology - IBBT,
Vrije Universiteit Brussel, Dept. of Electronics and Informatics, Belgium
[4] Adaptive Systems Research Group, School of Computer Science,
University of Hertfordshire, United Kingdom
[5] Department of Electrical and Electronic Engineering, Imperial College London, UK

## 1 Introduction

Children are keen users of new technologies and new technologies can provide interesting opportunities to enrich children's experience, e.g., for educational and therapeutic purposes.As children are not small adults, it is necessary to research their specific needs and develop systems that address them. The ALIZ-E project[6] develops cognitive robots for adaptive social interaction with young users over several sessions in real-world settings. We demonstrate a conversational system developed in ALIZ-E using the Nao robot[7]. It engages a user in the following activities (Fig.1):

- quiz: the child and the robot ask each other series of multiple-choice quiz questions from various domains, the robot provides evaluation feedback;
- imitation: either the child or the robot presents a sequence of simple arm poses that the other tries to memorize and imitate;
- dance: the robot explores various dance moves with the child and then teaches the child a dance sequence according to its abilities

These activities were chosen with regard to the target application domain of the system, namely long-term interaction with children hospitalized due to metabolic disorders, in particular diabetes. Quiz is a knowledge-exchange ativity meant to support learning of health-related concepts. Due to its prediminantly verbal character and constrained interaction structure it is a good testbed for

[6] The EU-FP7 project ALIZ-E (ICT-248116), http://aliz-e.org/
[7] http://www.aldebaran-robotics.com/en

**Fig. 1.** *Left to right*: Nao in the measurement setup in a sound lab at VUB and playing Quiz, Imitation and Dance during experiments in the San Raffaele hospital in Milan.

speech-processing technologies. Dance is an activity that promotes physical exercise. At the same time it provides a challenging domain for motion modeling and processing. Finally, Imitation on the one hand involves memory-exercise, and on the other hand provides a gentle introduction to physical movement for those users who are too shy to join the dance activity. It also involves an interesting mixture of verbal and non-verbal interaction, but more structured than Dance.

Besides activity-specific conversation, the interactions involve also a social component (greetings, introductions). During an activity, the robot provides performance feedback to the user. The social aspect here requires careful handling of the evaluation process so as not to discourage the user with negative feedback. As the system is designed to have multiple encounters with a user, the robot's behavior differs in various aspects from the first session (meeting for the first time) to the subsequent sessions ("knowing" the user and their performance).

## 2 System Description

Fig. 2 depicts the system components (more details below). We use the Urbi middleware [3] to implement an event-based approach to integration [10].

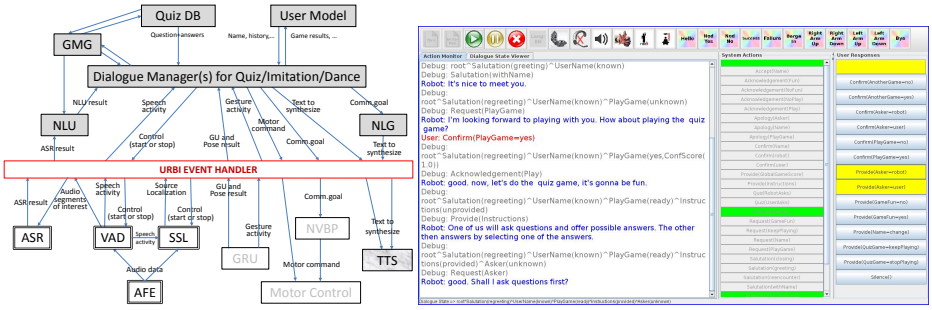


**Fig. 2.** *Left:* The components of the integrated system. Filled boxes: components implemented in Java, double-line boxes: C/C++, and plain boxes: UrbiScript. The TTS component is either the Acapela TTS on the Nao or the Mary TTS implemented in Java. *Right:* The Quiz Game Wizard GUI.

*Speech Signal Detection and Capture.* The Audio Front-End (AFE) component captures the speech signal from the microphones, makes preliminary preprocessing such as sample rate conversion, and sends the audio buffers to the Voice Activity Detection (VAD) component. VAD allows the robot to detect that dynamically varying sound sources for further analysis are active, using a robust energy based algorithm [8]. For the Sound Source Localization (SSL) component we implemented a Generalized Cross-Correlation based method with a set of pre-measured Time Delays On Arrival followed by parabolic interpolation [2].

*Spoken Input Processing* The Automatic Speech Recognition (ASR) component uses the Open-Source Large Vocabulary CSR Engine Julius[8] for which we trained an Italian child acoustic model. For demonstration purposes we also use an off-the-shelf ASR component for English.

Further processing in the Natural Language Understanding (NLU) component proceeds along two paths: For the recognition of quiz questions, answer options and answers we use fuzzy matching of recignized content words against the Quiz DB entries. For the recognition of other dialogue acts we use either partial parsing or keyword spotting as a fallback.

*Visual Input Processing* For the Gesture recognition and Understanding component (GRU) we have been experimenting with various methods. For example, one method to trace hands in the Imitation Game uses skin detection [9] enhanced with motion history [5]. An alternative method uses face detection and tracking algorithm [1] to define the vertical areas where the hands might move, and either motion detection or various optical flow algorithms.

*Dialogue Management* Speech and gesture interpretations go to the Dialogue Manager (DM) that bears primary responsibility for controlling the robot's conversational behaviour and the game progress. It keeps track of the interaction state, integrates the interpretations of the user's input/actions w.r.t. this state, and selects the next action of the system as a transition to another state, making progress towards a goal. How exactly this is done depends on the game. For example, selecting the next suitable question in Quiz is done by a separate Game Move Generator (GMG) component that also accesses the Quiz DB.

Inspired by the Zone of Proximal Development theory proposed by Vygotsky, the system takes the user's performance into account. For example, in the dance activity, the key point is to propose dance moves with gradually increasing or decreasing complexity based on the user's performance. On the one hand, the selected move should be within the user's capabilities to avoid discouraging the child, and on the other hand, challenging enough to maintain the child engaged in the task and willing to continue. The dance move selection mechanism thus considers: the level of complexity of the dance moves, a hierarchical representation of dance moves and the user's current potential capability to perform the different moves. The quiz question selection algorithm similarly takes into account the difficulty level of a question, whether it was asked already, and whether

[8] `http://julius.sourceforge.jp/`

the user knew the answer. In both quiz and imitation, the robot makes mistakes on purpose, to maintain a performance level approximating that of the user.

User-specific information (e.g., name, age) and interaction history (e.g., games played, achieved performance) are kept in the User Model (UM), which also receives updates from the DM. Game-specific information is also stored.

To experiment with the system without relying on fully automatic processing, we developed a Wizard-of-Oz interface (Fig. 2). Given a user input, the wizard can select the corresponding user dialogue act'. The DM then selects the next action. In the automatic mode, the DM passes a dialogue act to the NLG and NVBP components. In a non-automatic mode, the action selected by the DM is highlighted in the interface for the wizard to aprove or override. It is possible to switch between automatic and wizarded DM at any time during a session.

*Spoken Output Production* Spoken output is produced by the Natural Language Generation (NLG) and Text-To-Speech Synthesis (TTS) components. The system action selected by the DM specifies the type of dialogue act and the values of information state variables important for verbalization selection. Verbalization is determined by an utterance planner using a set of graph rewriting rules. The output is either a string passed directly to the TTS, or a logical form that serves as input to a grammar-based lexical realization component using OpenCCG[9].

Since repetitive verbalization of system output could be annoying and thus negatively influence engagement, we implemented a large range of verbal output variation. Selection among variants is either random or controlled by selectional criteria, taking into account the content to be conveyed and the dialogue context.

To foster a sense of familiarity between the robot and the user in interactions over multiple sessions, the robot explicitly acknowledges and refers to common ground with a given user, thus making it explicit that it is familiar with them.

For speech synthesis the commercial Acapela TTS system[10] is available by default on the Nao. However, we also integrated the open source Mary TTS platform[11], for which we developed a new Italian voice. Mary TTS supports state of the art HMM-synthesis technology, and enables us to experiment with the manipulation of para-verbal parameters (e.g. pitch shape, speech rate, voice intensity, pause durations) for the purpose of expressive speech synthesis, and the voice quality and timbre modifications algorithms [11] useful to convert an adult TTS voice into a child like voice.

To further enhance the contextual appropriateness of the output speech we experiment with modifications of the spoken output prosody using the support for controling the prosody of TTS voices with symbolic markup of speech rate, pitch and contour. So far, we implemented prosodic prominence modification (stress) on words that realize the focus of a sentence and emotional prosody modification according to the emotional state of the robot (sad and happy).

[9] `http://openccg.sourceforge.net/`
[10] `http://www.acapela-group.com/index.html`
[11] `http://mary.dfki.de/`

*Nonverbal Behavior Production* The Non-verbal Behavior Planning (NVBP) and Motor Control (MC) components produce arm gestures and head&body poses. Besides the game-specific moves and poses in the imitation and dance games, static key poses are produced to display emotions, namely anger, sadness, fear, happiness, excitement and pride [4].

## 3 Demonstrated Features

The present demonstration will focus in particular on spoken language input and output processing. This includes robust natural language understanding, dialogue management based on hierarchical reinforcement learning [6] using flexible hierarchical dialogue control [7], varied verbalization production, familiarity across multiple sessions and contextually controlled speech synthesis.

## References


1. OpenCV library website ((accessed 1052012)), `http://opencv.willowgarage.com`
2. Athanasopoulos, G., Brouckxon, H., Verhelst, W.: Sound source localization for real-world humanoid robots. In: Proceedings of 11th International Conference on Signal Processing (SIP 2012). pp. 131 – 136. WSEAS (Mar 2012)
3. Baillie, J.: URBI: Towards a Universal Robotic Low-Level Programming Language. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3219–3224. IEEE (2005)
4. Beck, A., Cañamero, L., Bard, K.: Towards an affect space for robots to display emotional body language. In: Proceedings of the 19th IEEE international symposium on robot and human interactive communication. pp. 464–469. Ro-Man 2010, IEEE (2010)
5. Bradski, G., Davis, J.: Motion segmentation and pose recognition with motion history gradients. Machine Vision and Applications 13, 174–184 (2002)
6. Cuayáhuitl, H.: Learning dialogue agents with bayesian relational state representations. In: Proceedings of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems (IJCAI-KRPDS), Barcelona, Spain. pp. 9–15 (2011)
7. Cuayáhuitl, H., Kruijff-Korbayová, I.: An interactive humanoid robot exhibiting flexible sub-dialogues. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Montreal, Canada (2012)
8. Dekens, T., Verhelst, W.: On the noise robustness of voice activity detection algorithms. In: Proceedings of 12th Annual Conference of the International Speech Communication Assosiation (INTERSPEECH). pp. 2649 – 2652. ISCA (Sep 2011)
9. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. Int. J. Comput. Vision 46, 81–96 (2002)
10. Kruijff-Korbayová, I., Athanasopoulos, G., Beck, A., Cosi, P., Cuayáhuitl, H., Dekens, T., Enescu, V., Hiolle, A., Kiefer, B., Sahli, H., Schröder, M., Sommavilla, G., Tesser, F., Verhelst, W.: An event-based conversational system for the nao robot. In: IWSDS 2011. Granada, Spain (Sep 2011)
11. Tesser, F., Zovato, E., Nicolao, M., Cosi, P.: Two Vocoder Techniques for Neutral to Emotional Timbre Conversion. In: Yoshinori Sagisaka, Tokuda, K. (eds.) 7th Speech Synthesis Workshop (SSW). pp. 130–135. ISCA, Kyoto, Japan (2010)