Long-Term Human-Robot Interaction with Young Users

Paul Baxter
University of Plymouth

Tony Belpaeme*
University of Plymouth

Lola Cañamero
University of Hertfordshire

Piero Cosi
National Research Council – Padova

Yiannis Demiris
Imperial College London

Valentin Enescu
Vrije Universiteit Brussel

ABSTRACT

Artificial companion agents have the potential to combine novel means for effective health communication with young patients support and entertainment. However, the theory and practice of long-term child-robot interaction is currently an under-developed area of research. This paper introduces an approach that integrates multiple functional aspects necessary to implement temporally extended human-robot interaction in the setting of a paediatric ward. We present our methodology for the implementation of a companion robot which will be used to support young patients in hospital as they learn to manage a lifelong metabolic disorder (diabetes). The robot will interact with patients over an extended period of time. The necessary functional aspects are identified and introduced, and a review of the technical challenges involved is presented.

1. INTRODUCTION

Robotic systems are playing an increasingly important role in healthcare, the use of robots in surgery is well established but an increasingly important aspect of medical robotics concerns artificial companion agents. Companion robots have a variety of potential healthcare applications including the provision of health education, supporting communication between patients and heathcare professionals and providing entertainment for patients in hospital. In order for robots to perform successfully as companion agents it is desirable that their interactions with human users be as naturalistic as possible. Humans are capable of maintaining social relationships over long periods of time and of adapting their behaviour to take account of previous encounters with other individuals. To date, one sigificant constraint on the efficacy of companion robots has been their inability to sustain non-continuous, temporally extended social interactions with users i.e. to engage in social exchanges extending beyond the scale of minutes and to adapt their interactive behaviour on the basis of previous encounters with a user.

The EU ALIZ-E project aims to contribute to the development of integrated cognitive systems capable of naturally interacting with young people in real-world situations, with a specific goal of supporting children engaged in a residential diabetes-management course. Fundamental to making human-robot interaction natural and integrated into the fabric of our lives, is that the robot can establish itself cognitively in the long term. Only if interaction provides a sense of continuity over longer periods of time, can it provide the resonance necessary for a constructive relationship between human and robot. It is commonly acknowledged that learning, adaptation, emotion, multi-modal dyadic and group interactions will be necessary to achieve this goal, but the field has not yet been presented with conclusive design paradigms, algorithms and test results showing how a robot can enter and successfully maintain an interaction spread beyond the current single episode interaction frame and stretching over several days.

To this end, several threads that research the different requirements in the support of long-term companion robot interactions are being persued. The aim of this paper is to illustrate the necessity of each of these different research threads, and to describe what each entails. Four distinct threads may be identified: (1) adaptive memory for long-term interaction; (2) adaptive user and task modeling; (3) adaptive non-linguistic behaviour; and (4) adaptive linguistic interaction. It is proposed that it is only through the combination of each of these that the eventual goal of long-term human-robot interaction may be achieved.

To demonstrate and evaluate the scientific approach, the ALIZ-E project develops and instantiates these methods in a succession of integrated systems that are tested through experiments with potential users 'in the wild', in the context of a medical healthcare setting. In the context of this project, long-term interaction means a non-continuous succession of interaction episodes over a period of up to 5 days. For the first round of data collection experiments and integrated system development tests several use cases have been designed that involve engaging the user(s) in simple games involving various combinations of verbal and non-verbal behavior. These case studies, and the ALIZ-E project as a whole, make use of the Aldebaran Nao small humanoid robot[1] as the test platform (controlled using Urbi[2]), as it provides a readily accessible and friendly interface for the target young users. The initial results of these experiments (described below) highlight the need for a multi-modal and integrated approach to the problem of naturalistic and extended interactions.

*Corresponding author: tony.belpaeme@plymouth.ac.uk.

[1] www.aldebaran-robotics.com

[2] Open Source, available from www.gostai.com/products/urbi

Figure 1: Children interacting with a WoZ operated Nao robot.

2. AN INITIAL CASE STUDY

Our initial target group are 8-12 year old children with metabolic disorders (i.e., diabetes and obesity). The initial use cases centre on engaging the child with four different games, involving both verbal and physical interactions. During game play, the robot will adapt its communication and interaction style to the child's behaviour. In particular the adaptation strategy will be used for two main classes of parameter: 1) communication, for example the robot will be able to adapt verbal parameters (e.g., repeating child's preferred expression, or adopting the same expression modality, for instance using more visual images), para-verbal parameters (e.g., speech rate, voice intensity, pause duration), and non-verbal parameters (e.g., posture, simulated breath rate, gestures); and 2) interaction parameters, e.g. adapting games (e.g. playing with or without verbal guidance, with or without physical action examples) and changing the type of game played (e.g., if the child is bored or if the game isn't challenging enough).

The first step in the evaluation of this strategy has been to conduct probe trials involving kindergarten children, using a Wizard of Oz approach[3]. The data obtained thus far appears promising indicating that the robot has the capacity to engage children from a wide age range and that it is perceived as an non-threatening source of entertainment (i.e. as a peer or toy). These observations lend support to the notion that such a robotic platform could be effective in communicating healthcare information to child users in entertaining and memorable ways. These observations also lend support to the notion that the capacity for adaptive and multi-modal behaviour is required for the agent to engage in, and maintain, long-term interactions.

3. MEMORY FOR LONG-TERM HRI

In human-robot interaction, it is the robot's capacity for adaptivity that enables interaction to be maintained after the initial novelty wears off. This is also apparent from the use case study presented above, where the children's attention waned over time unless the robot displayed some novel behaviour. The necessity for user driven adaptation of behaviour to support long-term, multi-episode interaction thus requires some capacity for memory in order to maintain pertinent information about the user and the context. A general functional definition of a memory system that supports this may be proposed: it is an agent-centred system that may store information, from various sources including interactions, for the purpose of adapting future behaviour. Whilst this definition does not specify any particular implementational mechanisms, its commitment to the purpose of future behavioural adaptation means that memory cannot be considered a pure passive storage structure. This contention underlies the approach taken to the memory system in the ALIZ-E project; one which is not generally reflected in current interpretations or implementations of memory.

Current computational architectures for human-robot interaction are typically reactive systems with no explicit temporal dimension to the production of behaviour. Implementation of memory systems for artificial agents have generally been restricted to general cognitive architectures, where social interaction is not the main purpose of the agent. These implementations implicitly have memory as a separable and passive storage element that is essentially dissociable from the cognitive architecture. For example, in the SOAR architecture, episodic memory is implemented as a database that periodically stores the entire contents of a central processing space, for recall at a later time if a similar condition is encountered. Similarly, the short-term memory system of the ISAC upper-torso humanoid robot is a symbolic database, where entries have a time-limited membership of the database, and are supplemented with a spatial component. Where memory systems have been implemented for social (and usually virtual) agents, they often have similar implementational characteristics to those described in the examples above. Exceptions do exist: [9], for example, stores an sensorimotor interaction history that informs action selection.

In contrast to existing approaches to memory implementations in cognitive architectures, the ALIZ-E memory system may be characterised as a set of active processes, and not as a passive storage device [13]. Three main principles underlie this assertion, and the approach taken. First is that memory is taken to be a distributed system, rather than a centralised store. Secondly, memory is proposed to underlie cognition and is an ongoing mediator of behaviour, rather than a system distinct from cognition. A fundamental mechanism proposed to underlie this functionality is that of cross-modal priming, where multiple modalities may be coordinated through the manipulation of synchronous distributed activation levels. Finally, the extended symbol grounding approach is used, where the learning agent aligns its internal representation to that of a group of other agents. Taken together, these principles, and the mechanisms derived from them, provide a framework for the development of the ALIZ-E memory system.

4. USER AND TASK MODELING

The robot has to adapt its non-linguistic and linguistic behavior to the user, the task and their interaction history. This functionality requires the support of an integrative adaptive user and task model.This section deals with how we select and test aspects of robot behaviour. Interviews with diabetes health care providers show that patient empowerment is very important. Most research on patient empowerment aims at increasing patient compliance with healthcare regimes. Children between 8-10 years of age are generally highly compliant with the requirements of their medical regime. However, it is known that the manner in which children and their carers deal with diabetes at this age influences the compliance at puberty, so the 8-10 age range is an effective point at which to introduce and test extra support via a robotic companion.

It is important that the child user is supported in learning how

[3]The results of this trial are reported in a companion paper submitted to the HRI2011 Robots and Children workshop.

to manage their health independently. Interviews diabetes care providers indicate that parents tend to take a controlling role in management of the child's condition even when the child is capable of performing that role herself. This can result in a reduction in compliance as the child reaches puberty and begins to rebel against parental control. Children who are empowered to take control of their condition at a younger age are less prone to neglect the management of their health during puberty and after. User modeling is most often used for rule based applications (e.g. intelligent tutoring). In such contexts it is clear what knowledge they have and what they lack. Such models can be augmented with affective user modeling i.e. to improve self-efficacy [8]. In ALIZ-E one goal is to engage the child in longer interactions and this is facilitated by the creation of an affective bond between robot and child [2]. User and task models interact when, for example, a usually introverted child is in a very happy mood and at ease and thus behaving in a more extrovert fashion than usual. The interaction with the robot should change to reflect the child's behaviour on that occasion and this may necessitate the selection of a new task or adaptation of an existing task.

5. NON-LINGUISTIC BEHAVIOR

Adaptive non-linguistic behavior refers to the robot's ability to analyze and adaptively generate non-verbal interaction elements such as emotions, body postures and actions. For example, the robot might involve the child in a game to prevent anxiety in a hospital environment, monitor his/her mood, and switch to another game when the child gets bored. In another example, in teaching the child a dance in order to promote physical activity, the robot should be able to assess the child's progress in learning sequence of moves, congratulate her/him for a positive outcome, or repeat the sequence while trying to keep the child engaged. In the following, we discuss various aspects of behavior recognition and production.

Behavior understanding implies the identification of various activities/gestures/postures of the child (such as sitting or walking), as well as some commands the child might give to the robot in a verbal manner (such as "stop"). Spatiotemporal (ST) visual descriptors based on gradients represent an effective way to capture the appearance and motion information without computing the optical flow (an ill-posed and computationally-intensive problem). So far, we have experimented with encouraging results the recognition of a reduced number of behaviors with a noise-filtered, ST descriptor. However, much work remains to be done in coping with view-independent recognition, background clutter, and intra-class variability caused by individual speed and appearance.

The production of expressive elements of behaviour has received considerable attention in the area of human-robot interaction literature [10]. We use the term "production of expressive behavior" to refer not only to the display of emotional expressions (with a particular focus on bodily expression and non-verbal sounds), but also to the internal processes that lead to the selection of the appropriate emotion-related behavior in the context of social interaction. Considering the chosen robotic platform, bodily expressions offer more potential and flexibility than facial expressions to provide the child with an adequate expressive feedback. Recent work using the Nao robot has shown a high level of recognition in adults rating the emotional body postures of the robot [1]. The characteristics of the body postures are directly related to an *Affect Space*, which controls the intensity of the emotion expressed (as a relative position from the neutral expression for instance).

Keeping in mind the initial goals of the project – establishing a bond and gaining the child's trust – the expressive responses of the robot have to be graded and appropriate to the current context (the game being played), and the emotional state of the child. The real-time emotion recognition system will play an important role towards that end. However, elements and mechanisms underlying the production of more subtle signals relevant to social-emotional interaction are investigated in order to provide the robot with a more continuous appraisal of the engagement of the child. As proposed in [3], the use of rhythm and other time-based interaction variables can inform the robot about the user's level of engagement and affect during the interaction. An important challenge in this regard is the production of timely and appropriate behaviour to elicit positive emotions and engagement from the child. It is hypothesized that as a result of this resonance, bonding between the dyad will grow with the extension of interactions.

6. LINGUISTIC INTERACTION

The cognitive system needs to cope with a wide range of linguistic inputs (along with non-verbal inputs) and respond to them in a way that the human finds natural and appropriate. This involves challenges at several levels: to begin with, what is "natural" and "appropriate" in human-robot interaction is not yet fully understood. Non-understandings and misunderstandings are frequent, and often difficult to overcome; trying to prevent them by very careful verification strategies results in unnatural and tedious interactions. This leads to the second challenge: how to process unconstrained spoken input in a robust fashion, while at the same time achieving deep enough comprehension. And, how to improve comprehension by making use of the robot's awareness (and knowledge) of the current physical situation, the user, the interaction context and the long-term experience. The final set of challenges pertains to the verbal output produced by the robot. The kind of dialogue contributions the robot makes, and their wording, provide a window into the robot's cognitive capabilities. It is known that the dialogue and output planning strategies used by a system have an effect on interaction success and user satisfaction, and that adaptivity to the user has a positive effect [12]. It is also known that the users' linguistic behaviour is influenced by that of the system they interact with [4].

Automatic speech recognition for young users is a central challenge. Both intra- and inter-speaker variability is much greater for children than for adults, and, in turn, much greater for young children than for older children, approaching adult values around the age of 13 [5]. In addition, there are currently very few child speech corpora available. We plan to expand these speech databases through the recording of new sentences, using the Nao robot's microphones. In order to support believable real-time interaction between robot and child we use a speech recognition system which implements Hidden Markov Models and n-gram statistical language models. In addition, speech adaptation methods will be applied for speaker-specific acoustic modelling. Extended and multiple interaction episodes can be used to improve the accuracy of subsequent speech recognition through the application of iterative, unsupervised speaker adaptation to the data collected.

We adopt a model in which incremental parsing, and dialogue interpretation are closely coupled [7]. The level at which verbal input is interpreted is selected on a spectrum between deep and shallow processing in a resource-sensitive and adaptive way. Although deep processing ideally yields the highest quality of interpretation, it may sometimes not be desirable, possible or even necessary in a given situation. Deep processing high demands on

system resources (modelled by the associated costs) and may be precluded by low quality of the received input. Deep processing may not be necessary when the verbally communicated content contributes nothing or very little to the topic of the interaction. We use a controller to guide processing. The controller determines how information flows between these processes to optimally discriminate between alternative hypotheses. It has access to information sources in- and outside dialogue to prime construction and discrimination of alternative hypotheses. These sources include nonlinguistic signs, user state, and memory context. The controller learns policies to balance between the costs of using processing resources and the expected gain for the interaction. During processing, it can reactively produce back channel signals or clarification requests, to signal (non-) understanding [6].

Synthesized speech triggers social identification processes, for this reason we would like to use the voice of a child for Nao. We would also like to express emotions vocally. However, the creation of a Text To Speech child corpus with accurate pronunciation is a substantial challenge and for this reason we are evaluating the possibility of using an adult female speaker and adapting her voice to sound like that of a child. Emotional speech synthesis must take into account the manipulation of para-verbal parameters like speech rate, voice intensity, pause durations, etc. The most suitable solution is Hidden Markov model (HMM) synthesis which allows stronger parameter control than Unit Selection synthesis. Finally, in order to obtain the emotive speech, HMM trajectory estimation techniques must be coupled with digital signal algorithms and speech models capable of implementing voice quality and timbre modifications [11] as well as general pitch shifting and time stretching algorithms. While it is true that HMM-based speech synthesis allows for more flexible voice control, data-driven speech synthesis allows for more natural sounding voice qualities. While synthesis of unrestricted natural-sounding expressive speech is still largely an unsolved problem, phrase concatenation could prove a viable solution for systems with small to moderate dialog models. Therefore, we intend to also consider phrase concatenation synthesis as well as hybrid methods in which parameter values from an HMM synthesizer are used in specifying the target units for Unit Selection synthesis.

7. CONCLUDING REMARKS

That companion agents, specifically robots, have the capacity to play a significant role in healthcare communication is clear. Furthermore, that role will require the ability to support long-term interactions over multiple distinct episodes. This paper has presented methods for the implementation of a set of functional capacities necessary in support of this, and proposed that the integration of these functions is essential for naturalistic interactions. The ALIZ-E project seeks to achieve this goal through the incremental development of an integrated system: the results of initial case studies have provided support for this methodology, both in terms of the definition of requirements, and the testing of the resultant system.

Acknowledgments

This work is supported by the EU Integrated Project ALIZ-E (FP7-ICT-248116).

8. ADDITIONAL AUTHORS

Antoine Hiolle (University of Hertfordshire), Ivana Kruijff-Korbayová (Deutsches Forschungszentrum für Künstliche Intelligenz), Rosemarijn Looije (Netherlands Organization for Applied Scientific Research), Marco Nalin (Fondazione Centro San Raffaele del Monte Tabor), Mark A. Neerincx (Netherlands Organization for Applied Scientific Research), Hichem Sahli (Vrije Universiteit Brussel), Giacomo Sommavilla (National Research Council – Padova), Fabio Tesser (National Research Council – Padova), Rachel Wood (University of Plymouth).

9. REFERENCES

[1] A. Beck, A. Hiolle, A. Mazel, and L. Cañamero. Interpretation of Emotional Body Language Displayed by Robots. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments*, 2010.

[2] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer, 2007a.

[3] A. Hiolle, L. Cañamero, P. Andry, A. Blanchard, and P. Gaussier. Using the interaction rhythm as a natural reinforcement signal for social robots: A matter of belief. In *Proceedings of the International Conference on Social Robotics*, 2010.

[4] I. Kruijff-Korbayová and O. Kukina. The effect of dialogue system output style variation on users' evaluation judgements and input style. In *Proceedings of SigDial'08*, Columbus, Ohio, 2008.

[5] S. Lee, A. Potamianos, and S. Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Amer., Vol. 105, March 1999.*, pages 1455–1468, 1999.

[6] L. Li, M. Littman, and T. Walsh. Knows what it knows: A framework for self-aware learning. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML-08)*, Helsinki, Finland, 2008.

[7] P. Lison and G. Kruijff. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008)*, Patras, Greece, 2008.

[8] S. Mcquiggan, B. Mott, and J. Lester. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, 18(1):81–123, 2008.

[9] N. A. Mirza, C. L. Nehaniv, K. Dautenhahn, and R. te Boekhorst. Grounded sensorimotor interaction histories in an information theoretic metric space for robot ontogeny. *Adaptive Behavior*, 15(2):167–187, 2007.

[10] C. Pelachaud and L. Cañamero. Achieving human-like qualities in virtual and physical humanoids. *Special Issue of the International Journal on Humanoid Robotics*, 3(3):1371–1389, 2006.

[11] F. Tesser, E. Zovato, M. Nicolao, and P. Cosi. Two Vocoder Techniques for Neutral to Emotional Timbre Conversion. In Y. Sagisaka and K. Tokuda, editors, *7th Speech Synthesis Workshop (SSW)*, pages 130–135, Kyoto, Japan, 2010. ISCA.

[12] A. Winterboer and J. Moore. Evaluating information presentation strategies for spoken recommendations. In *Proceedings of the ACM Conference on Recommender Systems*, 2007.

[13] R. Wood, P. Baxter, and T. Belpaeme. A developmental perspective on memory-centred cognition for social interaction. In *Proceedings of the Tenth International Conference on Epigenetic Robotics*, 2010.