DIARC: A Testbed for Natural Human-Robot Interaction

Paul Schermerhorn and **James Kramer** and **Christopher Middendorff** and **Matthias Scheutz**
Artificial Intelligence and Robotics Laboratory
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556, USA
Email: {pscherm1,jkramer3,cmidden1,mscheutz}@cse.nd.edu

Introduction

Autonomous human-like robots that interact in natural language with people in real-time pose many design challenges, from the functional organization of the robotic architecture, to the computational infrastructure possibly employing middle-ware for distributed computing, to the hardware operating many specialized devices for sensory and effector processing in addition to embedded controllers and standard computational boards. The task is to achieve a functional integration of very diverse modules that operate at different temporal scales using different representations on parallel hardware in a reliable and fault-tolerant manner that allows for *natural, believable human-robot interaction* (HRI). To achieve reliable, natural interaction with humans, several challenging requirements must be met, two of which are (R1) appropriate interaction capabilities, including *natural language capacity* (speech recognition and speech production), *dialog structure* (knowledge about dialogs, teleological discourse, etc.), *affect recognition and expression* (both for speech as well as facial expressions), and mechanisms for *non-verbal communication* (via gestures, head movements, gaze, etc.); and (R2) mechanisms for ensuring robust interactions, including *recovery from various communication failures* (acoustic, syntactic, semantic misunderstandings, dialog failures, etc.) as well as *software and hardware failure recovery* (crashes of components, internal timing problems, faulty hardware, etc.).

We are developing DIARC, a *distributed integrated affect*, *reflection*, *cognition* architecture for robots that interact naturally with humans (Scheutz *et al.* 2005; 2006). DIARC is a complete architecture that can be employed for HRI experiments without any modifications–robot behaviors can be expressed simply by virtue of *scripts* that contain general knowledge about conversations and action sequences. DIARC provides several features that are critical for the study of natural human interaction that are not easily found in other robotic systems. Some of these features are described below, and will be featured in the 2006 AAAI Robot Competition and Exhibition. Specifically, the robot will participate in the following categories of the Human-Robot Interaction competition: emotion recognition and appropriate emotion expression, natural language understanding and action execution, perceptual learning, and the integration challenge.

DIARC Implementation

DIARC is implemented in ADE, a Java-based infrastructure for the design, development, and execution of agent architectures (Scheutz 2006). To address requirement R2, ADE provides robust, reliable, fault-tolerant middle-ware services for the distribution of complex robotic architectures over multiple computers and their parallel operation, including monitoring, error detection, and recovery services that tie into the high-level action interpreter directly. A novel feature of ADE is a *reflective reasoning* component that is incorporated directly into the infrastructure, allowing it to maintain "facts" about the entire system, including both active and uninstantiated components, known hosts, relationships among components, etc. Furthermore, these "facts" can include rules (i.e., a "policy") that define reasoned, automatic reactions to changing conditions. For instance, if the speech recognition component fails, thereby rendering the agent unable to understand commands, the current policy may determine that it should be recovered on a different host. Moreover, the infrastructure is closely tied into the architecture, so if recovery is not possible locally, the speech recognition component can initiate actions in other components to respond appropriately to the failure (e.g., instantiating a new recovery goal, or switching the communication modality to a working component).

Emotion recognition and expression Because emotion is an important component in human-robot interaction (as stated in requirement R1), the system includes components to detect emotion in humans, and to express appropriate emotional responses. Emotion detection is bimodal, with components that analyze a speaker's voice and facial expression. Because the stress in a speaker's voice is marked by an increase in the mean of its fundamental frequency (pitch) and the intensity (volume) (Johnstone & Scherer 2000), tracking these features can provide information about the speaker's stress level. If the pitch and volume of an individual word is higher than their cumulative averages, it is marked as "stressed." The advantage of the employed system is that it requires no training corpus nor underlying training algorithm (although a short initial training phase can provide a useful speaker-dependent baseline).

Visual emotion detection identifies facial features such as eyebrow and mouth configuration (e.g., shape, size, and position). The positions of the features on the face are tracked to determine emotions; for instance, eyebrow raising signals "surprise", while the rate of position change provides a measure of the intensity. To compensate for varying lighting conditions, detection parameters are adapted via an online, real-time best parameter search using a swarm-based system (Middendorff & Scheutz 2006).

Emotion expression (as in requirement R1) applies an emotion filter, based on (Burkhardt & Sendlmeier 2000), to Festival speech output that alters various parameters tied to internal affective state. In particular, various degrees of emotional intensity are defined for "sad", "angry", "frightened", and "happy". To give the robot a "frightened" voice, pitch and speech rate are increased and "jitter" is added to give the voice a quivering sound.

Natural language understanding and action execution The robot interacts with humans by using spoken natural language in the course of executing conversation and action scripts. The *natural language processing subsystem* integrates and extends various existing components such as SONIC for speech recognition, the link parser for parsing, Verbnet and Framenet for semantic mapping, an enhanced and modified version of "thought treasure" for natural language understanding and sentence production, and a modified version of the Festival system for speech synthesis.

The *action control and goal management subsystem* is based on a novel *affective action interpreter*, which interprets scripts for natural language understanding and action control. Scripts in DIARC can include conditionals, allowing interaction to proceed in different ways depending on the current situation. In addition to modulating emotion expression as described above, the robot's affective states influence action selection via a *prioritized goal stack*. The robot's affective states change in response to factors such as successful or failed completions of tasks. As affect changes, the robot's assessments of individual goals' utilities will also change, automatically modifying their priorities and thus their order in the goal stack. This allows the robot to tailor its action selections to current conditions without requiring an expensive high-level evaluation process to examine, for example, the recent history of successes and failures.

Perceptual learning The robot also has the ability to learn to recognize objects visually, and subsequently respond appropriately to queries and commands regarding learned objects. Object recognition was implemented using *scale-invariant feature detection* (SIFT) points for the extraction of distinctive features from images (Lowe 2004). Once objects have been learned, the system is able to answer simple queries (e.g., "Is object X in the scene?") as well as more complex queries concerning spatial relationships between objects (e.g., "Is object X to the left of object Y?").

Integration challenge The capabilities described above each address some aspect of the individual categories of the human-robot interaction competition. However, the human-robot interaction integration challenge requires several of these individual categories to work together as a coherent system for human-robot interaction. To meet this challenge, a collaborative exploration task was designed that requires each of the above capabilities. For this exploration task, the pair forms a team that must traverse an extraplanetary environment while performing two parallel search tasks. The primary search is for a site with a sufficiently high signal strength to transmit information to an orbiting station. The human team member directs the primary search, telling the robot where to move and when to take readings of signal strength. For the secondary search, the human shows the robot an example of an object type and the robot is required to scan the environment for other objects of that type while responding to orders from the human related to the primary search ("move left," "check signal strength"). If the robot detects a match for the secondary search, it informs the human. The robot monitors the human's detected emotional state, responding appropriately to changes. For example, if the human becomes too stressed, the robot may offer suggestions or even assume greater degrees of autonomy to ensure successful task completion.

Conclusion

By demonstrating the the system's individual abilities in the emotion recognition and expression, natural language understanding and action execution, and perceptual learning categories of the HRI competition as well as its capacity to combine these abilities to perform the complex exploration task to meet the integration challenge, we have shown that DIARC and ADE form a viable platform in which to investigate architectures for human-robot interaction. Abilities such as those presented here (in addition to many others not addressed in this abstract) will ultimately be necessary to achieve human-like robotic systems that behave naturally, and thus are acceptable to humans.

References

Burkhardt, F., and Sendlmeier, W. 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*.

Johnstone, T., and Scherer, K. 2000. Vocal communication of emotion. In Lewis, M., and Haviland, J., eds., *Handbook of Emotion, 2nd ed.* Guilford. 220–235.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Middendorff, C., and Scheutz, M. 2006. Real-time evolving swarms for rapid pattern detection and tracking. In *Proceedings of ALife*.

Scheutz, M.; Schermerhorn, P.; Middendorff, C.; Kramer, J.; Anderson, D.; and Dingler, A. 2005. Toward affective cognitive robots for human-robot interaction. In *AAAI 2005 Robot Workshop*.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Middendorff, C. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. ACM Conference on Human-Robot Interaction (HRI2006).

Scheutz, M. 2006. ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence* 20(4-5).