The Virtual Human *Max* - Modeling Embodied Conversation

Stefan Kopp[1,2] , Christian Becker[1], and Ipke Wachsmuth[1,2]

[1]Artificial Intelligence Group, University of Bielefeld,
P.O. Box 100131, 33501 Bielefeld, Germany
{skopp,cbecker,ipke}@techfak.uni-bielefeld.de

[2]Center for Interdiscplinary Research (ZiF),
Bielefeld, Germany

Introduction

The Artificial Intelligence group at Bielefeld University has been developing the virtual human *Max* to study how the natural conversational behavior of humans can be modeled and made available for A.I. systems. This research activity embarks on the goal of building holistic, embodied agents that can engage with humans in face-to-face conversation and demonstrate many of the same communicative behaviors as displayed by humans. This activity has been pursued in a row of projects during which Max's conversational capabilities have been (and are being) steadily extended, allowing the employment of the agent in increasingly challenging scenarios. Originally started out as a platform for simulating speech and gesture generation [4], we brought Max to an application as a virtual receptionist that welcomes people in the hallway of our lab [3]. In the SFB 360 Situated Artificial Communicators, Max was utilized to study aspects of situated communication and collaborative, mixed-initiative dialogue in a VR construction task [6]. Since January 2004, Max has been applied in the *Heinz Nixdorf MuseumsForum* (HNF), a public computer museum in Paderborn (Germany), making the step from a lab-inhabiting research prototype to a system being confronted daily with real humans in a real-world setting [5].

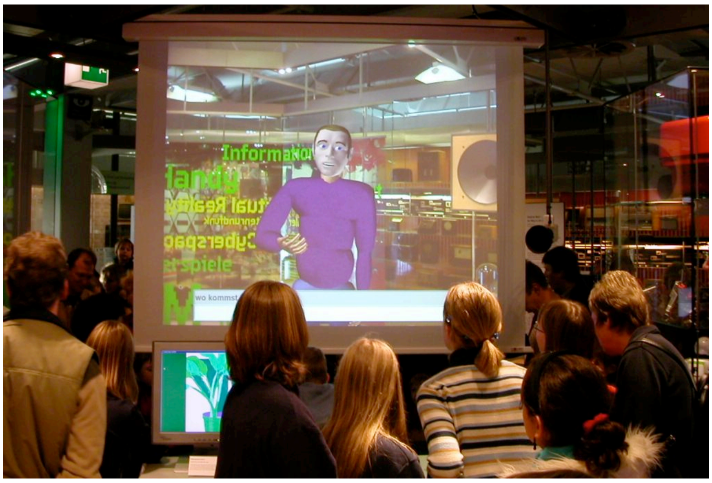

Fig. 1. Max interacting with visitors in the Heinz Nixdorf MuseumsForum.

Description of demonstration

Our demonstration will show Max as employed in the HNF. In this setting (see Figure 1), Max is visualized in human-like size on a static screen, standing face-to-face to his human interlocutors. The agent is equipped with camera-based visual perception and can notice people that are passing by. Max can start/end a dialogue and react to various input events (e.g., when the user starts or finishes typing). If there is no ongoing conversation, newly perceived persons are greeted and encouraged to start an interaction. Max tries to engage visitors in conversations in which he provides them with information about various topics of interest (in the HNF, about the museum or the exhibition). Visitors can give natural language input to the system using a keyboard, whereas Max will respond with a synthetic German voice and appropriate nonverbal behaviors like manual gestures, facial expressions, gaze, or locomotion.

The system performs mixed-initiative dialogue and is capable of initiating, holding, resuming and releasing topics and dialogue goals. Instead of being only reactive to user input, Max is thus able to keep up the conversation himself and to conduct a coherent dialog. In doing so, Max strives to create the impression of an enjoyable, cooperative interaction partner, being entertaining and fun to talk with. To this end, Max is capable of small talk and tailors his explanations to contextual factors like the visitor's interests and responds to questions, interruptions, or topic shifts. In discourse, the system draws upon knowledge about former episodes to answer questions like “How many people were here today?” or to derive user-related statements like “There were already five people here with your name”. Max has (and displays) an emotional state that is influenced by the presence of interlocutors and the current dialogue. For instance, insults by his dialogue partner lead to negative impulses that accumulate in Max’s emotion system, which can eventually result in Max leaving the scene in order to de-escalate the rude visitor behavior. Other features of the system include a guessing animal game, where Max asks questions to find out an animal that a visitor has in mind, or the internet lookup of up-to-date information (e.g. weather report).

Techniques being demonstrated

The embodied agent Max is a large-scale system that combines a multitude of A.I. methods and models. In our demonstration a variety of such techniques and methods can thus be seen at work in an integrated, comprehensive system. This includes approaches pertaining to autonomous agent architectures, multimodal behavior interpretation and production, knowledge representation, dialogue management, or cognitive and emotional modeling.

Cognitive Architecture

Max is based on an architecture that realizes and tightly integrates all faculties of perception, action, and cognition required to engage in embodied conversations. While at large employing the classical perceive-reason-act triad, it is conceived such that all processes are running concurrently. Perception and action are directly connected through a reactive component, affording reflexes and immediate responses to situation events or input by a dialogue partner. A keyboard is used as input device to constraint linguistic input as little as possible. Camera-based perception and real-time capable, image processing techniques are employed to find faces in front of the keyboard as well as a greater view at the exhibition area and to track them over time. All speech and visual inputs are sent to a perception module that utilizes sensory buffers, ultrashort term memories, to compensate for recognition drop-outs and to integrate both kinds of data. Reactive processing is realized by a behavior generation component, which is in charge of realizing the behaviors that are requested by other components. This includes feedback-driven reactive behaviors like gaze tracking the current interlocutor and secondary behaviors like eye blink and breathing. Additionally, to realize multimodal utterances, the Max system encompasses synthesis of prosodic speech and animation of emotional facial expressions, lip-sync speech, and coverbal gestures, as well as scheduling and executing all verbal and nonverbal behaviors in synchrony.

BDI-based deliberation

Deliberative processing of all events takes place in a central component. It determines when and how the agent acts, either driven by internal goals and intentions or in response to incoming events which, in turn, may originate either externally (user input, persons that have newly entered or left the agent's visual field) or internally (changing emotions, assertion of a new goal etc.). It maintains a dynamic spatial memory that contains all objects and persons in the agent's environmental context. All deliberative processes are carried out by a BDI interpreter [2], which continually pursues multiple, possibly nested plans (*intentions*) to achieve goals (*desires*) in the context of up-to-date knowledge about the world (*beliefs*). Most of the plans implement condition-action rules that test either user input or the content of a dynamic knowledge base; their actions can alter the dynamic knowledge structures, raise internal goals and thus invoke corresponding plans, or trigger the generation of an utterance (stating words, semantic-pragmatic aspects, and markup of the central part).

Multimodal dialogue

The deliberative component, running completely in the BDI framework, interprets an incoming event, decides how to react dependant on current context, and produces an appropriate response. It thereby combines pattern matching techniques to model robust small talk about large domains, with plan-based approaches to conduct longer, coherent dialogues and to act proactively, e.g. to take over the initiative, instead of being purely responsive as classical chatterbots are. The deliberative component draws upon long-term knowledge about former dialogue episodes with visitors or general capabilities of dialogue management, interpreting language input and generating behaviors for a certain communicative function. In addition, it maintains a dynamic knowledge base that includes a discourse model, a user model, as well as a self model that comprises the agent's world knowledge as well as current goals and intentions.

A set of skeleton plans realizes the agent's general, domain-independent dialogue skills like negotiating initiative or structuring a presentation. These plans are adjoined by a larger number of smaller plans implementing condition-action rules that define both, the broad conversation knowledge (e.g., dialogue goals that can be pursued, interpretations of input, small talk answers) as well as the deep knowledge about possible presentation contents. In its current state, Max is equipped with roughly 900 skeleton plans and 1.200 rule plans of conversational and presentational knowledge. At run-time, the BDI interpreter scores all plans depending on their utility and applicability in context. The most adequate plan is then selected for execution.

Emotions

Max is equipped with an emotion system that continuously runs a dynamic simulation to model the agent's emotional state. The emotional state is available anytime both in continuous terms of valence and arousal as well as a categorized emotion, e.g. happy, sad or angry (see [1]). The continuous values modulate subtle aspects of the agent's behaviors, namely, the pitch, speech rate, and band width of his voice and the rates of breathing and eye blink. The weighted emotion category is mapped to Max's facial expression and is sent to the agent's deliberative processes, thus making him cognitively "aware" of his own emotional state and subjecting it to his further deliberations. The emotion system, in turn, receives input from both the perception (e.g., seeing a person triggers a positive stimulus) and the deliberative component. For example, obscene or politically incorrect wordings in the user input lead to negative impulses on Max's emotional system.

Multimodal behavior generation

Max creates his multimodal communicative behaviors on-the-fly in order to fulfill a desired communicative function and to express his current emotional state. Drawing from a repository, nonverbal behaviors are added to support the given communicative function. Behavior planning further allocates bodily resources, taking account of the current movement and body context, and adapts deictic gestures to the current situational context. Combining means of speech synthesis and model-based computer animation, all planned behaviors are synthesized, scheduled, and executed from the scratch and automatically [4].

Acknowledgement

This work was partially supported by the *Heinz Nixdorf MuseumsForum* and the Deutsche Forschungsgemeinschaft in the Collaborative Research Center SFB 360.

References

1. C. Becker, S. Kopp, I. Wachsmuth: Simulating the Emotion Dynamics of a Multimodal Conversational Agent. Affective Dialogue Systems (2004)
2. M.J. Huber : JAM : A BDI-Theoretic Mobile Agent Architecture. Proc. Autonomous Agents'99, Seattle (1999)
3. B. Jung, S. Kopp: FlurMax: An Interactive Virtual Agent for Entertaining Visitors in a Hallway. In T. Rist et al. (eds.): Intelligent Virtual Agents, 23-26, Springer-Verlag (2003)
4. S. Kopp, I. Wachsmuth: Synthesizing Multimodal Utterances for Conversational Agents. Computer Animation and Virtual Worlds 15(1): 39-52 (2004)
5. S. Kopp, L. Gesellensetter, N. Krämer, I. Wachsmuth: A conversational agent as museum guide -- design and evaluation of a real-world application. Panayiotopoulos et al. (eds.): Intelligent Virtual Agents, LNAI 3661, 329-343, Berlin: Springer-Verlag (2005)
6. N. Lessmann, S. Kopp, I. Wachsmuth: Situated Interaction with a Virtual Human - Perception, Action, and Cognition, in Rickheit, Gert, and Wachsmuth, Ipke (eds.): Situated Communication, 287-323, Mouton de Gruyter (2006)