GALA 2007 submission document

Title: The Rapport Agent
Track: Student
Category: Application

Authors: Ning Wang
5th year PhD Student
ICT, University of Southern California, USA

Contact data: **name:** Ning Wang
e-mail: nwang@ict.usc.edu
url: *http://www.isi.edu/~ning*
phone: +1 (310) 574-5700
postal address: *Institute for Creative Technologies*
13274 Fiji Way, Marina del Rey, CA USA

URL: *http://www.isi.edu/~ning*

Movie file submitted: *ningwang.divx*

Reference teacher: **Jonnathan Gratch**
affiliation: ICT, University of Southern California
e-mail: gratch@ict.usc.edu

Process of development: The current Rapport Agent was developed between August 2006 and May 2007 based on a previous implementation of listening agent. The author developed improved responsive and mimicking behavior of the agent and conducted evaluation studies of the agent.

Resources used:

1. Watson for tracking head position and orientation
   http://groups.csail.mit.edu/vision/vip/watson/
2. LAUN for detecting acoustic features from speech signals
3. BML for defining and planning animation
   http://wiki.mindmakers.org/projects:BML:main
4. SmartBody for producing animation defined by BML
   http://www.isi.edu/~amarshal/projects/SmartBody.html
5. Unreal Tournament™ game engine for rendering and displaying animation
6. Visual Studio for programming

Resources required: The Rapport Agent was developed on PCs with Windows XP Prof OS. In the evaluation study of the Rapport agent, two Videre Design Small Vision System stereo cameras were used to capture the subjects' movements. Headset with microphone was used to record subjects' speech. Three Panasonic PV-GS180 camcorders were used to videotape the subjects. Four desktop computers were used in the evaluation study: two DELL Precision 670 computers, one DELL Precision 690 and one DELL Precision 530 computer. 19-inch Dell monitors were used as display for three of computers. The Rapport agent was displayed on a 30-inch Apple display.

GALA 2007 submission document

The Rapport Agent

1. The application and context of the work

The Rapport Agent was designed to establish a sense of rapport with a human participant in "face-to-face monologs" where a human participant tells a story to a silent but attentive listener. In such settings, human listeners can indicate rapport through a variety of nonverbal signals (e.g., nodding, postural mirroring, etc.) The Rapport Agent attempts to replicate these behaviors through a real-time analysis of the speaker's voice, head motion, and body posture, providing rapid nonverbal feedback.

The Rapport Agent was part of the Stability and Support Operations – Simulation and Training (SASO-ST) project. SASO-ST is a large, interactive, immersive, virtual training environment prototype. In this environment, a trainee uses natural language to interact with a life-sized Virtual Human agent and perform a negotiation task. SASO-ST has developed a negotiation scenario where a trainee converses with an NGO doctor (Virtual Human agent) to convince him to move his clinic out of an area that will be in conflict. In this negotiation scenario, the trainee is a captain in the US Army who requires training in negotiation strategies. The SASO-ST system integrates a wide range of USC and ICT research efforts in areas such as natural language processing, emotion modeling, automated reasoning, speech recognition and computer animation to create virtual humans with believable behaviors. The Rapport Agent project was part of the effort to create believable listening behavior for the virtual agents.

The Rapport agent was evaluated in a series of studies. The results show that the Rapport Agent was as effective as human listeners in creating rapport.

2. Novelty

Emotional bonds don't arise from a simple exchange of facial displays, but often emerge through the dynamic give and take of face-to-face interactions. Rapport has been argued to lead to communicative efficiency, better learning outcomes, improved acceptance of medical advice and successful negotiations. "Embodied conversational agents" have attempted to generate nonverbal cues together with speech, but only a few have addressed the technical challenges of establishing the tight reciprocal feedback associated with rapport. For example, Neurobaby analyzes speech intonation and uses the extracted features to trigger emotional displays (Tosa, 1993). More recently, Breazeal's Kismet system extracts emotional qualities in the user's speech (Breazeal & Aryananda, 2002). A few systems can interject meaningful nonverbal feedback during another's speech. However these methods usually rely on simple acoustic cues. For example, REA will execute a head nod or paraverbal (e.g., "mm-hum") if the user pauses in mid-utterance (Cassell et al., 1999). Some work has attempted to extract extra-linguistic features of a speakers' behavior, but not for the purpose of informing listening behaviors. For example, Brand's voice puppetry work attempts to learn a mapping between acoustic features and facial configurations inciting a virtual puppet to react to the speaker's voice (Brand, 1999). Although there is considerable research showing the benefit of such feedback on human to human interaction, there has been almost no research on their impact on human to virtual human rapport.

The Rapport Agent explores the phenomenon of rapport. It was designed to establish a sense of rapport with a human participant by replicating a variety of speaker's nonverbal signals (e.g., nodding, postural mirroring, etc.) through a real-time analysis of the speaker's voice, head motion, and body posture. The Rapport Agent shows that a virtual character can be as effective as human listeners in creating rapport.

3. The architecture

The Rapport Agent uses a vision based tracking system and signal processing of the speech signal to detect features of the human speaker and uses a set of reactive rules to drive the listening mapping displayed in Table 1.

To produce listening behaviors, the Rapport Agent first collects and analyzes the speaker's upper-body movements and voice. (See the architecture of the system in Figure 1.) For detecting features from the speaker's movements, we focus on the speaker's head movements. Watson uses stereo video to track the participants' head position and orientation and incorporates learned motion classifiers that detect head nods and shakes from a

vector of head velocities. Other features are derived from the tracking data. For example, from the head position, given the participant is seated in a fixed chair, we can infer the posture of the spine. Thus, we detect head gestures (nods, shakes, rolls), posture shifts (lean left or right) and gaze direction.

Lowering of pitch → head nod
Raised loudness → head nod
Speech disfluency → posture/gaze shift
Speaker shifts posture → mimic
Speaker gazes away → mimic
Speaker nods or shakes → mimic

Table 1 Rapport Agent behavior mapping

Acoustic features are derived from properties of the pitch and intensity of the speech signal, using a signal processing package, LAUN, developed by Mathieu Morales. LAUN detects speech intensity (silent, normal, loud), range (wide, narrow), and backchannel opportunity points.

Recognized speaker features are mapped into listening animations through a set of authorable mapping language. This language supports several advanced features. Authors can specify contextual constraints on listening behavior, for example, triggering different behaviors depending on the state of the speaker (e.g., the speaker is silent), the state of the agent (e.g., the agent is looking away), or other arbitrary features (e.g., the speaker's gender). One can also specify temporal constraints on listening behavior: For example, one can constrain the number of behaviors produced within some interval of time. Finally, the author can specify variability in behavioral responses through a probability distribution of different animated responses.

These animation commands are passed to the SmartBody animation system using a standardized API. SmartBody is designed to seamlessly blend animations and procedural behaviors, particularly conversational behavior. These animations are rendered in the Unreal Tournament™ game engine and displayed to the Speaker.

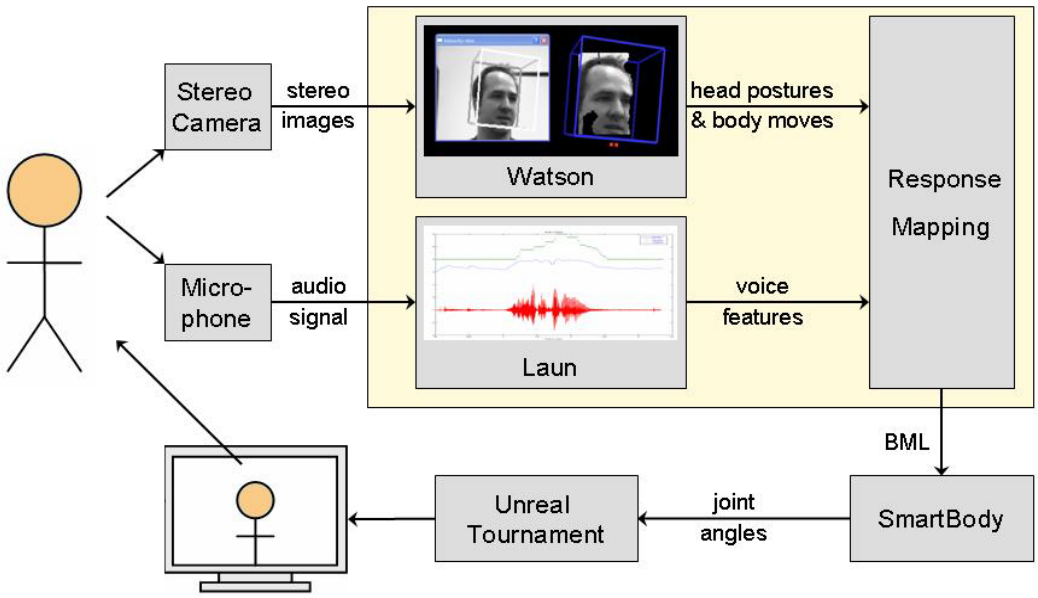

Figure 1 Rapport system architecture

4. Performance

The Rapport Agent was evaluated in a series of studies. Overall, our studies provide further evidence that the non-verbal behavior of virtual characters influence the behavior of the humans that interact with them. Our study results indicate that the Rapport Agent was as effective as human listeners in creating rapport. Further more, our study presents the first experimental evidence suggesting that contingency of agent feedback matters when it comes to creating virtual rapport.

One limitation of the Rapport Agent is its reliance on "mindless feedback" (i.e., it does not actually understand any of the meaning of the speaker's narrative). While this feedback can be quite powerful, it is insufficient for most potential applications of virtual humans. Such rapid, automatic feedback could be integrated with more meaningful responses derived from an analysis of the user's speech and facial expressions. But there are several technical obstacles must be overcome. For example, to provide within utterance feedback we see in rapportful interactions, systems would have to rapidly detect partial agreement, understanding or ambiguity at the word or phrase level. We are unaware of any such work.

References

1. Brand, M. (1999). Voice puppetry. Paper presented at the ACM SIGGRAPH.
2. Breazeal, C., & Aryananda, L. (2002). Recognition of Affective Communicative Intent in Robot-Directed Speech. Autonomous Robots, 12, 83-104.
3. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., et al. (1999). Embodiment in Conversational Interfaces: Rea. Paper presented at the Conference on Human Fac-tors in Computing Systems, Pittsburgh, PA.
4. Tosa, N. (1993). Neurobaby. ACM SIGGRAPH, 212-213.