Virtual Rapport

Jonathan Gratch[1], Anna Okhmatovskaia[1], Francois Lamothe[2], Stacy Marsella[1], Mathieu Morales[2], R. J. van der Werf[3], Louis-Philippe Morency[4]

[1] University of Southern California
[2]Ecole Spéciale Militaire de St-Cyr
[3]University of Twente
[4]Massachusetts institute of technology

Abstract. Effective face-to-face conversations are highly interactive. Participants respond to each other, engaging in nonconscious behavioral mimicry and backchanneling feedback. Such behaviors produce a subjective sense of rapport and are correlated with effective communication, greater liking and trust, and greater influence between participants. Creating rapport requires a tight sense-act loop that has been traditionally lacking in embodied conversational agents. Here we describe a system, based on psycholinguistic theory, designed to create a sense of rapport between a human speaker and virtual human listener. We provide empirical evidence that it increases speaker fluency and engagement.

1. Introduction

Conversations vary widely in terms of their quality. Sometimes we, er, um... We seem tongue tied. We stutter, pause and repeat our words. Other times, we feel in sync with our conversational partner and words flow without effort. Disfluency is typically a sign of cognitive load and can arise from a number of sources including the complexity of the subject matter or emotions arising from the social setting. One apparent influence on interactional fluency is the nonverbal behavior produced by participants (Chartrand and Bargh 1999). Fluent interactions typically involve nonverbal behavioral synchrony between the interactants. People mirror each other's postures and interject feedback such as nods or interjections (uh-huh) at just the right moment. In such situations, participants report feelings of rapport, like each other better, and are more likely to be persuaded by each other's assertions. Such findings have encouraged the development of embodied conversational agents that can reproduce such social influences.

When it comes to conversational gestures, most virtual human research has focused on half of the interactional equation. Systems emphasize the importance of nonverbal behavior in speech *production*. Only a few systems can interject meaningful nonverbal feedback *during* another's speech and when feedback exists at all, it typically occurs at utterance boundaries (eg.,Tosa 1993). Only a small number of systems have attempted to provide within-utterance listening feedback, and these methods usually rely on simple acoustic cues. For example, REA will execute a head nod or paraverbal (e.g. say "mm-hum") if the user pauses in mid-utterance (Cassell, Bickmore et al.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Virtual Rapport	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**14**	

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: Listening Agent Mapping

1999). Although there is considerable research showing the benefit of such feedback on human to human interaction, few studies have investigated their impact on human to virtual human rapport (cf. Cassell and Thórisson 1999; Bailenson and Yee 2005).

At last year's IVA conference we presented a Listening Agent that would try to create a sense of rapport simply by tying listening feedback to shallow features of a speaker's voice and bodily movements (Maatman, Gratch et al. 2005). Such an approach is clearly simpler than attempts to tie such feedback to a deep model of coordinated activity (Nakano, Reinstein et al. 2003; Heylen 2005). Here we present evidence that it can also be effective in positively influencing the quantity and quality of human speech.

2. Rapport Agent

The RAPPORT AGENT described here is an evolution of the LISTENING AGENT presented at IVA05 (Maatman, Gratch et al. 2005). The LISTENING AGENT was a simple approach to produce within-utterance listening behaviors based on real-time analysis of a speaker's voice, head motion, and body posture. The system was inspired by psycholinguistic findings that feelings of rapport are correlated with simple contingent behaviors between speaker and listener, including behavioral mimicry (Chartrand and Bargh 1999) and backchannel continuers (Yngve 1970). The LISTENING AGENT used a head-mounted motion tracker and signal processing of the speech signal to drive the listening mapping displayed in Table 1.

The system was directed at passing our proposed "Duncan Test," inspired by the work of Sue Duncan on studying rapport (Welji and Duncan 2004). Following the standard setup adopted by Duncan and McNeill, we suggest having a human participant watch a short cartoon and then describe it to a listening agent. To pass the test, the interaction between speaker and the agent should exhibit the same correlations between nonverbal behaviors, self and other reports of rapport, and social outcomes such as liking, persuasion and conversational fluency.

Preliminary evaluations suggested the LISTENING AGENT was viable for such a task, but also revealed limitations that we addressed in the creation of the RAPPORT AGENT:

- Contextual constraints on listening behavior: As the LISTENING AGENT employed a direct (i.e., stateless) mapping between detected features and responses, it couldn't account for important contextual features. For example, it might detect a "Speaker Gaze-Left" event, but could not condition its response on the state of the speaker (e.g., the speaker is silent), the state of the LISTENING AGENT (e.g., the agent is looking away), or other arbitrary features (e.g., the speaker's gender).
- Temporal constraints on listening behavior: the LISTENING AGENT had no notion of time, which, when coupled with lack of state, limited its ability to control the temporal dynamics of the listening behavior. For example, there was no easy way to constrain the number of behaviors produced within some interval of time.

Motion Features		Vocal Features	
Gestures	nod, shake	Intensity	silent, normal, loud
Head roll	upright, lean left, lean right	Range	wide, narrow
Gaze	straight, up, down, left, right	Other	backchannel opportunity

Table 2: Rapport Agent detected speaker features

- Variability of behavioral responses: the LISTENING AGENT enforced a 1-1 mapping between detected events and agent responses. This led to considerable repetition in the elicited behaviors and conveyed the sense that one was speaking to a robot.
- Portability: the LISTENING AGENT was restricted to a specialized room with a ceiling-mounted motion tracking system and a large screen to display the agent's graphical body. This hampered our ability to perform user testing in this heavily-utilized space and limited our ability to share the system with other colleagues.
- Feature detection: preliminary testing revealed shortcomings in the feature detectors. For example, the detection of speaker nods and shakes worked well for recognizing enacted (typically exaggerated) behavior but proved less reliable in recognizing more naturally elicited behavior. Further, head motions produced during speech introduced audio artifacts that influenced the detection of audio features.

These concerns led to a redesign of many of the LISTENING AGENT components. The resulting RAPPORT AGENT has an open modular architecture that facilitates the incorporation of different feature detectors and animation systems, and has an easily authored mapping between features and behavior. The behavior mapping language incorporates contextual features, probabilistic responses, and some control over the temporal dynamics of behavior. To address the issue of portability, we moved to a vision-based tracker and changed the setting from a standing interaction with a life-sized character to a seated interaction with a life-sized image of a character's head displayed on a computer monitor. Finally, we updated the original feature-detection algorithms and broadened the repertoire of recognized features. Here we give a high-level overview of the new architecture. Details can be found at (Lamothe and Morales 2006; van der Werf 2006). Figure 1 illustrates the basic outlines of the RAPPORT AGENT architecture.

Feature Detection

To produce listening behaviors, the RAPPORT AGENT first collects and analyzes the speaker's upper-body movements and voice to detect the features listed in Table 2.

For detecting features from the participants' movements, we focus on the motion of the speakers head. Watson, developed by Louis-Phillipe Morency, is an image-based tracking library that uses stereo images to track the participants' head position and orientation (Morency, Sidner et al. 2005). Watson also incorporates learned motion classifiers that detect head nods and shakes from a vector of head velocities. Other features are derived from the position and orientation of participant's head (filtered to reduce the impact of noise). For example, from the head position, given the participant is seated in a fixed chair, we can infer the posture of the spine.

Acoustic features are derived from properties of the pitch and intensity of the speech signal (the RAPPORT AGENT ignores the semantic content of the speaker's speech), using a signal processing package, LAUN, developed by Mathieu Morales. Speaker pitch is approximated with the cepstrum of the speech signal (Oppenheim

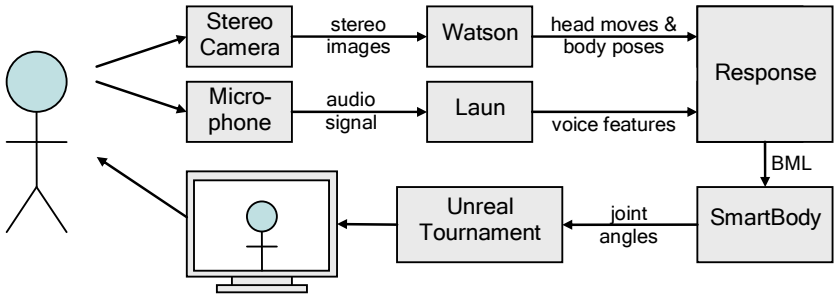

Figure 1: Rapport Agent architecture

and Schafer 2004) and processed every 20ms. Audio artifacts introduced by the motion of the Speaker's head are minimized by filtering low frequency noise. Speech intensity is derived from amplitude of the signal.

We split speech feature detections into two families: the "instant" are derived in real-time, and the "delayed" detections that can be analyzed at the end of the "sentences" featuring them. Instant features include silent/normal/loud speech (derived from signal intensity) and backchannel opportunity points (derived using the approach of Ward and Tsukahara 2000). In addition, we make a crude attempt to separate utterances based on silences and attempt to detect some features that hold across the utterance including pitch-range (positive affect is often associated with wider pitch-range).

Behavior Mapping

Recognized speaker features are mapped into listening behaviors through a set of authorable mapping rules. The language is based on five primitives.

- Each participant in the interaction is described by an *agent.* Agents consist of a set actions, states, animations and reactions. For the discussion that follows, we will assume two agents: the agent that represents the human speaker and the agent that represent the RAPPORT AGENT listener
- *Actions* represent discrete behavioral events that can be generated by an agent. These can consist of the detectable features of human behavior (Table 2) or arbitrary behavior outputs of the RAPPORT AGENT .
- *States* describe characteristics of an agent that can persist over time. Typically, states are asserted as consequences of actions (e.g., after detecting LeanLeft, the speaker is in the state of LeaningLeft). States can be constrained logically (e.g., the speaker cannot be simultaneously speaking and silent) and temporally (e.g., to ensure an agent stays in some state for some period of time).
- *Animations* are physical behaviors described in the Behavior Markup Language (BML) (Kopp, Krenn et al. 2006) that can be associated with agent actions. For example, a backchannel continuer might be associated with a nod animation.
- *Reactions* map from an action in one agent to an action in another agent. The mapping is conditional on the current state of one or more agents and can map, probabilistically, to one of a set of other actions.

Typically, reactions map actions of the speaker to (re)actions by the RAPPORT AGENT. For example, if LAUN detects a backchannel opportunity point in the speaker, this could cause the RAPPORT AGENT to react with a Nod with probability 0.6 or GazeUp with probability 0.2, assuming the RAPPORT AGENT is in the state of GazingForward. The framework, however, can support more general settings. For example, one could define mapping rules for multiparty settings (e.g., multiple speakers or multiple listening agents). Alternatively, one could transform the behavior of a human listener into some, perhaps altered animated behavior (c.f., Bailenson, Beall et al. 2004).

Animation
RAPPORT AGENT animation commands are passed to the SmartBody animation system (Kallmann and Marsella 2005). This is a virtual human animation system designed to seamlessly blend animations and procedural behaviors. These animations are rendered in the Unreal Tournament™ game engine and displayed to the Speaker.

3. Evaluation

While RAPPORT AGENT described above could be integrated into a wide variety of embodied conversational agent applications, there are a number of questions that need to be addressed first to ensure the suitability of such integration:

- Does the system correctly detect features of the speaker's behavior, such as head nods, shakes, pauses in speech, etc.?
- How well do behavior mapping rules approximate the behavior of human listeners?
- Is the agent's behavior judged to be natural when it is performed?
- Do listening behaviors of the agent have the predicted influence on the human speaker's behavior and perceptions?

Our preliminary analysis suggests that feature detection is reasonably accurate and we are currently collecting data on human face-to-face communication to address the second question. The study presented here focuses on the last two questions and attempts to replicate certain well-known findings in social psychology about the effects of listener's feedback in face-to-face communication. In this study we try to demonstrate that nonverbal behavior displayed by the RAPPORT AGENT contributes to its perceived believability, positively affects the speaker's motivation and speech fluency, and can induce subjective feelings of rapport in human participants.

Hypotheses
People are more willing to communicate when their conversational partners display interest and may be quite frustrated when such feedback is absent. Several studies have demonstrated increased speaker engagement when listeners provide feedback such as nods and mimicry, whether the interaction is between humans or between humans and synthetic agents. For example, Tatar (1997) has demonstrated that speakers talking about life experiences told shorter stories and reported that they were less engaged when the listener was distracted and, thus, provided less feedback. In a GrandChair project (Smith 2000), elderly people were found to tell longer stories to a virtual child agent that displayed active listening behavior.

Based on these findings we can hypothesize that human subjects would be more engaged in interaction with a responsive agent that displays positive listening behav-

iors, as opposed to no or inappropriate feedback. While this claim is very straightforward, it is less obvious how to measure the degree of speaker's engagement. Different studies have looked at self-reports, the amount of gesticulation, facial expression, posture, gazing behavior, speech production. In this study we focus on duration of interaction, arguing that engaged speakers would tend to speak more.

It is important to point out that the relation between the amount and type of listener's feedback on one hand and speech quantity on the other hand is complex. Contrary to Smith (2000) and Tatar (1997), some studies found that that speakers produced fewer utterances when provided with feedback, which was explained by arguing the feedback reduced ambiguity and promoted greater communicative efficiency (Krauss, Garlock et al. 1977; Cassell and Thórisson 1999). One may notice, however, that these two groups of studies have utilized rather different types of communicative tasks in their experiments. Both the speaker goals and the function of listener's feedback can vary considerably depending on the context of the task. In tasks where the primary goal is to convey information (e.g., Krauss, Garlock et al. 1977) or when faced with time pressure, one might expect feedback to promote efficient communication and thus reduced speech quantity. Listener's feedback under these conditions can indicate comprehension and allow the speaker to compact their speech and avoid repetitions. The story-telling tasks typically have different emphasis: the speaker is either explicitly or implicitly encouraged to speak more and provide more details. Listener's feedback in this case may serve as positive reinforcement and motivate the speaker to continue interaction.

The task used in current study (retelling a funny cartoon) most likely belongs to the second category. We could thus assume that longer interaction with the system reflects the subject's engagement and motivation, rather than inefficiency.

H1: People will interact with a responsive listener longer than an unresponsive one.

As the effects of listener's feedback on speech quantity can be quite complex, it is important to also look at speech quality. Studies show that in the absence of such feedback or when the feedback is incoherent, the speakers become disrupted, and their speech – less structured (Kraut, Lewis et al. 1982; Bavelas, Coates et al. 2000).

One possible explanation for this effect is that listening feedback provided in a timely fashion reduces cognitive load on the speaker. Sources of this load can vary depending on the task and social setting, but they all produce uncertainty that the speaker constantly needs to resolve (e.g. Did this person understand me? Does he/she agree with what I am saying? etc.). Following this explanation it can be expected that incoherent or inappropriate feedback can be even more disruptive than the absence of feedback.

In this study we focus on one particular aspect of speech quality – fluency. Improved speech fluency is a prominent characteristic of rapport interactions, and we expect to achieve similar positive effect of non-verbal feedback provided by the RAPPORT AGENT on the speaker's quality of speech in our study.

H2: People will speak more fluently when interacting with a responsive agent.

Thus far we have focused on objective characteristics of interactions that involve social rapport. However in addition, participants of such interactions typically experience subjective feelings of rapport, which are available via self-report. People point

out that they felt a connection with each other – that they "clicked". We hope to supplement our other findings by this self-reported sense of rapport.

H3: When interacting with a responsive agent people will indicate feelings of rapport in verbal self-reports.

Experimental Setup

In evaluating the system we adapt the "McNeill lab" paradigm (McNeill 1992) for studying gesture research. In this research, one participant, the Speaker, has previously observed some incident, and describes it to another participant, the Listener. Here, we replace the Listener with the RAPPORT AGENT system.

People can be socially influenced by a virtual character whether or not they believe it represents a real person (Nass and Reeves 1996) although there can be important differences depending on how the situation is framed. In this study, we use a cover story to make the subjects believe that they interact with a real human. The participants are told that the study evaluates an advanced telecommunication device, specifically a computer program that accurately captures all movements of one person and displays them on the screen (using an Avatar) to another person. According to the cover story, we were interested in comparing this new device to a more traditional telecommunication medium such as video camera, which is why one of the participants was sited in front of the monitor displaying a video image, while the other saw a life-size head of an avatar (see Figure 2).

The subjects were randomly assigned to one of two conditions labeled respectively "responsive" and "unresponsive". In a *responsive condition* the Avatar was controlled by the RAPPORT AGENT, as described earlier. The Avatar therefore displayed a range of nonverbal behaviors intended to provide positive feedback to the speaker and to create an impression of active listening.

Figure 2: Experimental setup

The Speaker and the Listener are separated by a screen. The Listener (left) can hear the Speaker (right) and see a video image of him/her. The Speaker instead sees an Avatar allegedly controlled by the Listener's behavior. In fact, the Avatar is controlled by the RAPPORT AGENT, although the Listener's behavior is recorded and will be used in future studies and to code discrepancies between the Listener and RAPPORT AGENT.

One stereo camera is installed in front of the Speaker (below the monitor) to track head movements. An identical camera is also placed in front of the Listener to make the subjects believe his/her behavior is being tracked. A video camera on the Speaker's side is directly connected to the Listener's monitor.

In an *unresponsive condition* the Avatar's behavior was controlled by a pre-recorded random script and was independent of the Speaker's or Listener's behavior. The script was built from the same set of animations as those used in responsive condition, excluding head nods and shakes. Thus, the Avatar's behavioral repertoire was limited to head turns and posture shifts.

Procedure

Each subject participated in an experiment twice: once in a role of a Speaker and once as a Listener. The order was selected randomly.

While the Listener waited outside of the room, the Speaker watched a short segment of Sylvester and Tweety cartoon, after which s/he was instructed to describe the segment to the Listener. The participants were told that they would be judged based on the Listener's story comprehension. The Speaker was encouraged to describe the story in as much detail as possible. In order to prevent the Listener from speaking back we have emphasized the distinct roles assigned to participants, but did not explicitly prohibit the Listener from talking. No time constraints were introduced.

After describing the cartoon (during which time the Speaker was sitting in front of the Avatar), the Speaker was asked to fill out a short questionnaire collecting the subject's feedback about his experience with the system. Then the participants switched their roles and the procedure was repeated. A different cartoon from the same series and of similar length was used for the second round.

At the end of the experiment, both participants were debriefed. The experimenter collected some informal qualitative feedback on their experience with the system, probed for suspicion and finally revealed the goals of the study and experimental manipulations.

Dependent Variables

The collected data can be grouped into 3 major categories:

1. *Duration of interaction.* To measure the duration of interaction, we record the total time it takes the subject to tell the story. To obtain a measure independent of individual differences in speech rate, we count the number of words in the subject's story. We also differentiate between total word count and the number of "meaningful" (lexical and functional) words. For the later, speech disfluencies, such as pause fillers and stutters are excluded.
2. *Speech fluency*. To assess the speaker's fluency we use two groups of measures: speech rate and the amount of speech disfluencies (Alibali, Heath et al. 2001). For speech rate we distinguish between overall speech rate (all words per second) and fluent speech rate (lexical and functional words per second). To measure the amount of disfluencies, we use disfluency rate (disfluencis per second) and disfluency frequency (a ratio of the number of disfluencies to total word count).
3. *Self-reported measures of rapport.* Included in this category are several items of the questionnaire (see Figure 3). The questionnaire includes both forced choice and free format open-ended questions. The later were used as a source of qualitative data.

The research hypotheses can be now operationalized in terms of dependent variables:

H1a: Total time to tell the story will be higher in responsive condition.

H1b: The recorded stories will be longer in responsive condition in terms of both total word count and the number of lexical and functional words
H2a: Overall and fluent speech rate will be higher in responsive condition
H2b: The disfluency rate and disfluency frequency will be higher in an unresponsive condition
H3a: The subjects in responsive condition will be more likely to report a sense of rapport on the questionnaire.

Subjects

The participants were 30 volunteers from among employees of USC's Institute for Creative Technologies. Two subjects were excluded from analysis due to an unforeseen interruption of experimental procedure. The final sample size was 28: 16 in a responsive and 12 in an unresponsive condition.

Results

Because of a relatively small sample size used for this study, we have refrained from making assumptions regarding data distribution, and used non-parametric statistics to evaluate the differences between two groups of subjects: Mann-Whitney U – for scale variables (length of interaction, speech fluency), and Chi-square – for nominal data (forced-choice questionnaire items). $p < .05$ was used as a criterion.

Table 3 summarizes the data on duration of interaction and speech fluency. Consistent with H1a and H1b, the subjects in responsive condition talked significantly longer both in terms of overall time and word count. An increase in word count was associated with the higher number of lexical and functional words, while the total number of filled pauses and other speech disfluencies remained the same.

Consistent with H2b, the disfluency rate was significantly higher in unresponsive condition. The same is true for the disfluency frequency. Contrary to H2a, the subjects in unresponsive condition tended to speak faster, not slower. This finding, however, is non-significant for both the overall speech rate and fluent speech rate.

Self-report data is presented in Figure 3. Several trends are worth mentioning:

- Subjects in the responsive condition were more likely to feel that they had a connection with their conversational partner, and to form an impression that the listener understood them. They also reported that they used the listener's feedback when they were telling the story.

var	Responsive[a]	Unresponsive[a]	Mann-Whitney U	Sig.[b]
total time	188.68	98.50	30.0	0.001*
N words	432	300	44.0	0.015*
N words - disfluencies	411	288	39.0	0.007*
Speech rate	2.55	2.77	57.5	0.074
Fluent speech rate	2.42	2.60	66.5	0.174
Disfluency rate	0.13	0.21	28.5	0.001*
Frequency of disfl.	0.05	0.08	48.0	0.026*

[a] – median used as a measure of central tendency
[b] – 2-tailed criterion
* - $p < .05$

Table 3: Duration of interaction and fluency of speech

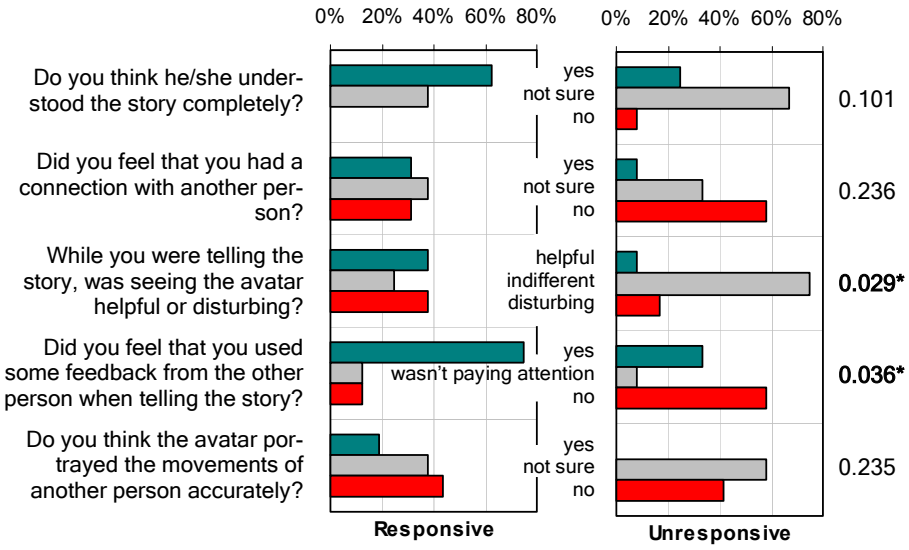

Figure 3: Summary of subjects' responses to selected questions

Chi-square statistics was used to compare frequency distributions between the groups. Significance values for each question are shown on the right ($p < .05$ marked with asterisks)

- Most subjects did not consider the avatar to be an accurate representation of a real listener; those few who did – all belonged to the responsive condition.
- Opinions on the helpfulness of the avatar were markedly different across the conditions. The subjects in responsive condition found the avatar to be either helpful or disturbing. In unresponsive condition 75% of the speakers had indifferent attitude.

Not all of the differences in self-reported measures reached statistical significance, and thus additional data may be needed to support these findings.

Discussion

The results obtained for the duration of interaction (word count and time) fully support our predictions, and are also consistent with some findings mentioned earlier (Smith 2000). The subjects spent more time talking to a responsive agent, and produced longer stories. What is important to note here is that there were significantly more "meaningful" words in these stories, suggesting that the increase in quantity of speech was not associated with a decreased quality.

We believe that this finding can be explained in terms of the subjects' willingness to interact with the listener (represented by an avatar in our experiment). The nonverbal behavior generated by the RAPPORT AGENT was intended to create an impression of an engaged and attentive listener and encourage the speaker. During the debriefing procedure after the experiment two subjects in the unresponsive condition (tested in different sessions) pointed out that they intentionally kept their stories short because the listener seemed to be uninterested. This observation brings to light an important

consideration in the design of embodied conversational agents: human observers tend to interpret not only the nonverbal clues displayed by the agent, but the absence of clues as well. The unresponsive agent in our experiment differed from a responsive one only by the absence of head nods, and randomized timing of posture and gaze shifts, so there weren't any specific behaviors that conveyed lack of interest or boredom. And yet, at least some subjects saw these signs in the agent's behavior. This suggests that one must carefully model the nonverbal behavior in embodied agents, since not only inappropriate behaviors, but sometimes just the lack of behaviors can produce undesirable effects in human observers depending on the context.

There is also evidence that human speakers were more engaged in conversation with a responsive agent, which is based on observations we made during the experiment. Several subjects in a responsive condition responded verbally to the feedback provided by the agent. In particular they could say "yes" and nod after the agent nodded. Or they could ask "Did you get it so far?", and then continue only after the agent nodded. This was not observed in an unresponsive condition. Since the experiment was built as a one-way communication, and such spontaneous interactions were actually discouraged by an instruction, they indicate a potential power of the system in producing social effects. These observations require further elaboration and formal experimental verification. Additional data on speaker's engagement may be obtained from analyzing gaze and gesturing behavior, which we plan to do in future studies.

We do not rule out additional explanations of these results. For instance, it is possible that speakers in the responsive condition remember the cartoon better and, thus, provide more details. This explanation does not exclude the one we presented before, and the next logical step for further research would be to find out what the weights of different factors underlying increased speech quantity are.

Our hypothesis regarding speech fluency was only partially supported: there was support for the amount of disfluencies (H2b), but not for speech rate (H2a). This suggests that speech rate may have a more complex relationship with conversational fluency than we believed. Indeed, speaking quickly does not necessarily mean speaking fluently. Particularly, in our study an increase in speech rate in the unresponsive condition was mainly due to more frequent inclusion of pause fillers, indicating that the subjects in this condition talked fast but with many disfluencies.

It is important to keep in mind that speech rate can be affected by a number of factors, in particular emotional. It is possible that the subjects in the unresponsive condition spoke faster because they felt uncomfortable and were trying to complete the task as quickly as possible. It was previously shown that synthetic agents can elicit anxiety in human users (Rickenberg and Reeves 2000) and, particularly, that unresponsive virtual audience produces greater anxiety in the speaker (Pertaub, Slater et al. 2001). Our results for the unresponsive condition are consistent with these findings.

As our experiment was not designed to control for social anxiety, presented explanations would need to be tested in further studies. In general, the problem of how social anxiety mediates the effects of listener's feedback on the speaker and how it interacts with rapport has not been yet investigated.

The results on self-reported feelings of rapport did not reach statistical significance, however the observed trends are consistent with our predictions. Increasing the sample size and using more fine-grained scales (compared to just "yes/no/unsure") may help obtain more conclusive results.

One particular finding derived from self-report data deserves attention: indifferent attitude towards the agent in unresponsive condition and either positive, or negative, and sometimes ambivalent – in responsive condition. The subjects seemed to ignore the agent when his behavior was unresponsive, but apparently could not do it when he was "actively listening". Several subjects in responsive condition admitted afterwards that they felt distracted by the agent and tried not to look at him to better concentrate on the story. This finding does not quite agree with the results on speech fluency and with our expectations for the responsive behavior to be helpful to the speaker. The question is: why such distraction occurs and what one can do to minimize it?

People appear to be more sensitive to some feedback – head nods and shakes – than to other components of listening behavior. As Chiu et. al. (1995) point out, it is hard for the speakers to ignore listener's feedback when it is relevant for the speech they are planning. Head nods and shakes typically appear to be of high relevance. When they are delivered at exactly the right moments, this improves interactional fluency. However if not perfectly timed, the head nods (and especially head shakes) are more disruptive than helpful. One obvious way to address this problem in RAPPORT AGENT is to make the head nod animation more subtle. At a deeper level, further work on improving feature detection and behavioral mapping rules is needed to ensure the agent's nonverbal feedback is in sync with the speaker's behavior.

We shall admit that overall self-reported measures of rapport and ratings of believability were lower than desired. Only about 30% of the subjects in responsive condition reported that they felt a connection with their conversational partner, and less than 20% considered the avatar to be an accurate representation of a real listener.

In order to find out what were the reasons for such results, we analyzed some qualitative data. In addition to answering formal yes/no questions, the subjects shared their comments on what difficulties they encountered when interacting with an agent, and what the reasons for his unnatural behavior were. The following factors seem to contribute to the subjects' overall impression of the agent and their experience of rapport:

- The Avatar did not display facial expressions, and many of the subjects felt that they were missing a significant part of the feedback the real listener was providing.
- The participants in responsive condition noticed imperfections in the animations: head nods seemed to be exaggerated and sometimes jerky, transitions between animations were not always smooth.
- Several subjects explicitly mentioned that some head nods were not properly timed.

These current limitations of the system will be addressed in our future work.

We have demonstrated that the RAPPORT AGENT exerts certain effects on the human speaker. However in order to further improve the system we need to know what it is about generated listening behavior that is responsible for these effects. Could the same results be achieved by manipulating the overall amount of movement displayed by the agent, or type of movement is important? Is it the mere occurrence of certain behaviors, or their timing that matters? How the results would change if the subjects believed they were talking to a computer and not to another human? We have already performed some additional analysis and are planning to gather more data to address these questions.

4. Conclusions

Presented in the current work is an Agent that aims at creating a sense of rapport in human speaker simply by tying nonverbal listening feedback to shallow features of a speaker's voice and bodily movements. This sense of rapport is believed to facilitate communication and to contribute to positive impression formation and trust between conversational partners.

We have conducted an empirical study, in which we attempted to replicate some of the known effects of rapport in human-to-virtual human communication. The results of this first round of system evaluation largely support our hypotheses. The RAPPORT AGENT was demonstrated to be effective in positively influencing the quality of their speech, their motivation and overall impression of communication. Noteworthy, the agent succeeded in achieving this effect without having a slightest idea of what the speakers were talking about.

The results also suggest how the system can be improved to further increase user satisfaction and subjectively perceived sense of rapport.

Acknowledgements

We would like to thank Susan Duncan, Jeremy Bailenson, Kris Thórisson, and Nigel Ward for very helpful feedback on this draft. Jillian Gerten provided crucial help in transcribing and analyzing subject dialogues. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Alibali, M. W., D. C. Heath, et al. (2001). "Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen." Journal of Memory and Language **44**: 169-188.

Bailenson, J., A. Beall, et al. (2004). "Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments." PRESENCE: Teleoperators and Virtual Environments **13**(4): 428-441.

Bailenson, J. N. and N. Yee (2005). "Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments." Psychological Science **16**: 814-819.

Bavelas, J. B., L. Coates, et al. (2000). "Listeners as Co-narrators." Jurnal of Personality and Social Psychology **79**(6): 941-952.

Cassell, J., T. Bickmore, et al. (1999). "Embodiment in Conversational Interfaces: Rea." *Conference on Human Factors in Computing Systems*, Pittsburgh, PA.

Cassell, J. and K. R. Thórisson (1999). "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents." International Journal of Applied Artificial Intelligence **13**(4-5): 519-538.

Chartrand, T. L. and J. A. Bargh (1999). "The Chameleon Effect: The Perception-Behavior Link and Social Interaction." Journal of Personality and Social Psychology **76**(6): 893-910.

Chiu, C., Y. Hong, et al. (1995). Gaze direction and fluency in conversational speech**:** Unpublished manuscript.

Heylen, D. (2005). "Challenges Ahead. Head Movements and other social acts in conversation." *AISB*, Hertfordshire, UK.

Kallmann, M. and S. Marsella (2005). "Hierarchical Motion Controllers for Real-Time Autonomous Virtual Humans." *5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, Springer.

Kopp, S., B. Krenn, et al. (2006). "Towards a common framework for multimodal generation in ECAs: The behavior markup language." *Intelligent Virtual Agents*, Marina del Rey, CA.

Krauss, R. M., C. M. Garlock, et al. (1977). "The Role of Audible and Visible Back-Channel Responses in Interpersonal Communication." Journal of Personality and Social Psychology **35**: 523-529.

Kraut, R. K., S. H. Lewis, et al. (1982). "Listener Responsiveness and the Coordination of Conversation." Journal of Personality and Social Psychology: 718-731.

Lamothe, F. and M. Morales (2006). Response Behavior. Marina del Rey, CA, University of Southern California**:** Technical Report ICT TR 01.2006.

Maatman, M., J. Gratch, et al. (2005). "Natural Behavior of a Listening Agent." *5th International Working Conference on Intelligent VirtualAgents*, Kos, Greece.

McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago, IL, The University of Chicago Press.

Morency, L.-P., C. Sidner, et al. (2005). "Contextual Recognition of Head Gestures." *7th International Conference on Multimodal Interactions*, Torento, Italy.

Nakano, Y., G. Reinstein, et al. (2003). "Towards a Model of Face-to-Face Grounding." *Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Nass, C. and B. Reeves (1996). The Media Equation, Cambridge University Press.

Oppenheim, A. V. and R. W. Schafer (2004). From Frequency to Quefrency: A History of the Cepstrum. IEEE Signal Processing Magazine. **September:** 95-106.

Pertaub, D.-P., M. Slater, et al. (2001). "An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience." Presence: Teleoperators and Virtual Environments **11**(1): 68-78.

Rickenberg, R. and B. Reeves (2000). "The effects of animated characters on anxiety, task performance, and evaluations of user interfaces,." *SIGCHI conference on Human factors in computing systems*, The Hague, The Netherlands.

Smith, J. (2000). GrandChair: Conversational Collection of Family Stories. Cambridge, MA, Media Lab, MIT.

Tatar, D. (1997). Social and personal consequences of a preoccupied listener. Department of Psychology. Stanford, CA, Stanford University**:** Unpublished doctoral dissertation.

Tosa, N. (1993). "Neurobaby." ACM SIGGRAPH: 212-213.

van der Werf, R. (2006). Creating Rapport with Virtual Humans. Marina del Rey, CA, University of Southern California**:** Technical Report ICT TR 02.2006.

Ward, N. and W. Tsukahara (2000). "Prosodic features which cue back-channel responses in English and Japanese." Journal of Pragmatics **23**: 1177-1207.

Welji, H. and S. Duncan (2004). "Characteristics of face-to-face interactions, with and without rapport: Friends vs. strangers." *Symposium on Cognitive Processing Effects of 'Social Resonance' in Interaction, 26th Annual Meeting of the Cognitive Science Society*.

Yngve, V. H. (1970). "On getting a word in edgewise." *Sixth regional Meeting of the Chicago Linguistic Society*.