Can Virtual Humans Be More Engaging Than Real Ones?

Jonathan Gratch[1], Ning Wang[1], Anna Okhmatovskaia[2], Francois Lamothe[3], Mathieu Morales[3], R.J. van der Werf[4], and Louis-Philippe Morency[5]

[1] University of Southern California
[2] McGill University
[3] Ecole Spéciale Militaire de St-Cyr
[4] University of Twente
[5] Massachusetts Institute of Technology

Abstract. 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. This article explores the phenomenon of rapport, a feeling of connectedness that seems to arise from rapid and contingent positive feedback between partners and is often associated with socio-emotional processes. Rapport has been argued to lead to communicative efficiency, better learning outcomes, improved acceptance of medical advice and successful negotiations. We provide experimental evidence that a simple virtual character that provides positive listening feedback can induce stronger rapport-like effects than face-to-face communication between human partners. Specifically, this interaction can be more engaging to storytellers than speaking to a human audience, as measured by the length and content of their stories.

1 Introduction

Emotion is often studied as something that arises within an individual's head and "leaks" to the outside world through their nonverbal behavior (Ekman, 1992), possibly leading to better communication and coordination with sensitive observers (R. Frank, 1988; Keltner & Haidt, 1999). But emotion sometimes seems to emerge from the space between individuals. For example, feelings of antagonism may arise as facial expressions and postures are exchanged, as "one person's gradual leaning forward first leads to withdrawal until ground is held" (Parkinson, 2001). In such a fashion, feelings may arise, not from each individuals' understanding of how the interaction relates to their personal beliefs and desires (as argued by many theories and computational models of emotion, e.g., Ellsworth & Scherer, 2003; Gratch & Marsella, 2004; Neal Reilly, 1996), but via interpersonally distributed cognition where "neither person's separate cognitive processes can entirely explain the resulting co-regulated reaction" (Fogel, 1993 in Parkinson, 2001).

Rapport is one crucial characteristic of successful interactions that seems to emerge outside of conscious awareness through the subtle interplay of nonverbal emotional signals. Cappella (1990) states rapport to be "one of the central, if not *the* central, constructs necessary to understanding successful helping relationships and to explaining the development of personal relationships." Rapport is correlated with characteristic

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 286–297, 2007.

nonverbal behaviors in face-to-face interactions. Participants seem tightly enmeshed in something like a dance. They rapidly detect and respond to each other's movements. Tickel-Degnen and Rosenthal (1990) equate rapport with behaviors indicating positive emotions (e.g. head nods or smiles), mutual attentiveness (e.g. mutual gaze), and coordination (e.g. postural mimicry or synchronized movements). Rapport is argued to underlie social engagement (Tatar, 1997), success in negotiations (Drolet & Morris, 2000), improving worker compliance (Cogger, 1982), psychotherapeutic effectiveness (Tsui & Schultz, 1985), improved test performance in classrooms (Fuchs, 1987) and improved quality of child care (Burns, 1984).

Two lines of research suggest that virtual characters could establish rapport with humans, and thereby attain rapport's beneficial influence over communication, persuasion and learning. On the one hand, studies suggest that rapport can be experimentally induced or disrupted by altering the presence or character of contingent nonverbal feedback (eg., Bavelas, Coates, & Johnson, 2000; Drolet & Morris, 2000). On the other hand, research on the social impact of virtual characters suggests that people, in some sense, treat virtual characters as if they were real people, and exhibit many of the subtle social influences that arise in human-to-human interaction (Kramer, Tietz, & Bente, 2003; Nass & Reeves, 1996). "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). Whenever the speech recognizer detects a pause in the speech, the previous utterance is classified (within one or two seconds) as indicating approval, an attentional bid, or a prohibition. This recognition feature is combined with Kismet's current emotional state to determine facial expression and head posture. People who interact with Kismet often produce several utterances in succession, thus this approach is sufficient to provide a convincing illusion of real-time feedback. Only a few systems can interject meaningful nonverbal feedback *during* another's speech and 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 (cf. Bailenson & Yee, 2005; Cassell & Thórisson, 1999).

There is some reason to believe, however, that, at least in certain contexts, a virtual human could promote more rapport than might be found in normal human-to-human interactions. Rapport is something that typically develops over time as inhibitions break down and partners begin to form emotional bonds. Strangers rarely exhibit the characteristic positivity, mutual attention or nonverbal coordination seen amongst friends (Welji & Duncan, 2004). Virtual humans, in contrast, can be programmed to produce such behaviors from the very beginning of an interaction. Further, some

researchers have suggested that virtual humans may be inherently less threatening than other forms of social interaction due to their game like qualities and the inherent unreality of the virtual worlds they inhabit (Marsella, Johnson, & LaBore, 2003; Robins, Dautenhahn, Boekhorst, & A. Billard, 2005). Alternatively, people might find an immediately responsive agent disconcerting or insincere, working against the establishment of rapport. In this article, we assess the potential of the RAPPORT AGENT to create more engagement and speech fluency than might be found between typical strangers. In the study presented here, we test the hypothesis that a virtual human could be more engaging than a human listener.

The next section describes the technical capabilities of the RAPPORT AGENT. We then describe a study that assesses the engagement and speech fluency of story tellers speaking to a positive active listening agent, an unresponsive agent or an unfamiliar human listener. We conclude with a general discussion and future thoughts.

2 Rapport Agent

The RAPPORT AGENT (Figure 1) 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 (Gratch et al., 2006). 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. Creation of the system is inspired by findings that feelings of rapport are correlated with simple contingent behaviors between speaker and listener, including behavioral mimicry (Chartrand & Bargh, 1999) and backchanneling (e.g., nods, see Yngve, 1970). The RAPPORT AGENT uses a vision based tracking system and signal processing of the speech signal to detect features of the speaker and then uses a set of reactive rules to drive the listening mapping displayed in Table 1. The architecture of the system is displayed in Figure 2.

To produce listening behaviors, the RAPPORT AGENT first collects and analyzes the speaker's upper-body movements and voice. For detecting features from the participants' movements, we focus on the speaker's head movements. Watson (Morency, Sidner, Lee, & Darrell, 2005) 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.[1]

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. Speaker pitch is approximated with the cepstrum of the speech signal (Oppenheim & Schafer, 2004) and processed every 20ms. Audio artifacts introduced by the motion of the Speaker's head are minimized by filtering out low frequency noise.

[1] Note that some authors have argued that higher-level patterns of movement may play a more crucial role in the establishment of rapport and would be overlooked by this local approach (Grammer, Kruck, & Magnusson, 1998; Sakaguchi, Jonsson, & Hasegawa, 2005).

Speech intensity is derived from amplitude of the signal. LAUN detects speech intensity (silent, normal, loud), range (wide, narrow), questions and backchannel opportunity points (derived using the approach of Ward & Tsukahara, 2000).

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 (Kallmann & Marsella, 2005) using a standardized API (Kopp et al., 2006). 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.

Fig. 1. A child telling a story to the RAPPORT AGENT

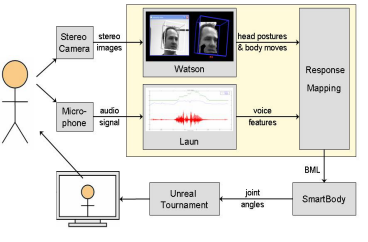

Fig. 2. Rapport Agent architecture

Table 1. Rapport Agent Mapping

Silence → gaze up/straight
Raised loudness → head nod
Backchannel → head nod
Ask question → head nod
Speaker shifts posture → mimic
Speaker gazes away → mimic
Speaker nods or shakes → mimic

3 Evaluation

Experimental Setup: In evaluating these hypotheses, we adapted the "McNeill lab" paradigm (McNeill, 1992) for studying gesture research: a speaker explains to a listener a previously watched film clip. As people can be socially influenced by a virtual

character whether or not they believe it represents a real person (Nass & Moon, 2000), we used a cover story to make the subjects believe that they were interacting with a real human. Participants were told that the purpose of the study was to evaluate 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. In line with the cover story, it was explained that 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 seated in front of the monitor displaying a video image, while the other saw a life-size head of an avatar (see Figure 3).

The subjects were assigned[2] to one of three conditions labeled respectively "face-to-face", "responsive" and "unresponsive." In all conditions, subjects sat across a table, separated by 8 feet (Figure 3). In a *face-to-face* condition, the listener and speaker could see each other (the screen and monitors in Figure 3 were removed). In the *responsive* and *unresponsive* conditions, the Speaker and the Listener were separated by a screen and didn't see each other directly. Rather, the Listener could hear the Speaker and see a video image of him/her. The Speaker could see an avatar on the monitor, sized to approximate the same field-of-view as the face-to-face condition. In the *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. In the *unresponsive* condition, the Avatar's behavior was controlled by a prerecorded 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.

Fig. 3. Experimental setup. The Listener (left) sees a video image of the Speaker (right). The Speaker sees an Avatar allegedly displaying the Listener's movements. Stereo cameras are installed in front of both participants (Listener data is ignored but stored for data collection/analysis).

Subjects. The participants were 48 adult volunteers from the University of Southern California's Institute for Creative Technologies. Two subjects were excluded from analysis due to an unforeseen interruption of experimental procedure. The final

[2] Subjects were not randomly assigned to the three conditions. Rather, the responsive and unresponsive conditions were part of an earlier study (Gratch et al., 2006). Here, we contrast face-to-face subjects with these earlier results. This methodological choice does limit the strength of our conclusions (see Section 5).

sample size was 46: 16 in a responsive and 12 in an unresponsive condition, 18 in face-to-face condition.

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

While the Listener waited outside of the room, the Speaker watched a short segment of a 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 talking back we 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 (while the Speaker was sitting in front of the Avatar), the Speaker was asked to complete 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: Engagement was indexed by total time it took the subject to tell the story and total number of words in the subject's story (independent of individual differences in speech rate).

To assess conversational fluency, we used two groups of measures: speech rate and the amount of speech disfluencies (Alibali, Heath, & Myers, 2001). Speech rate was indexed by the overall speech rate (all words per second) and the fluency speech rate (lexical and functional words per second). Amount of disfluencies was indexed by the disfluency rate (disfluencies per second) and the disfluency frequency (a ratio of the number of disfluency to total word count).

Subjective sense of rapport was measured through self-report using the forced-choice questionnaire items: "Did you feel you had a connection with the other person?" and "Did you think he/she [the listener] understood the story completely?". Additionally, the questionnaire included several open-ended questions, which were used as a source of qualitative data.

Thus the hypotheses were operationalized in terms of these measured variables, in the following ways:

H1a: Total time to tell the story will be longest in the *responsive* condition, followed by the *face-to-face* and then the *unresponsive* condition.

H1b: The recorded stories will be the longest in the *responsive* condition in terms of total word count followed by the *face-to-face* condition and then the *unresponsive* condition.

H2: The disfluency rate will be the highest in the *unresponsive* condition followed by the *face-to-face* condition and then the *responsive* condition.

H3: The subjects in *face-to-face* condition are most likely to report a sense of rapport on the questionnaire, followed by *responsive* condition and then *unresponsive* condition.

Results: The Tukey test was used to compare responses pairewise across the 3 conditions[3]. To satisfy the independence assumption required for the statistical analyses we conducted, analyses were conducted on speakers' data only. Table 2 summarizes the significant differences in duration and interaction fluency.

From Table 2, we can see that speakers in the *responsive* condition spoke longer and used more words than those in *unresponsive* and *face-to-face* condition. This result is consistent with H1a and H1b with the exception that the difference between the *face-to-face* and *unresponsive* conditions did not reach significance. However, this result demonstrates that speakers are more engaged when speaking to the responsive agent than to a real human listener.

In terms of disfluency, there are significant differences among three conditions in terms of speaker's disfluency rate, with speakers in the *unresponsive* condition having the highest disfluency rate. There are no significant differences between speakers in *face-to-face* and *responsive* condition. This is only partially consistent with H2: the speaker interacting with a real human listener didn't display more disfluency than those interacted with responsive agent, contrary to our predictions.

We further analyzed the different causes of disfluency. We counted the number of pause filler and incomplete words in the speech. There was no significant difference in number of pause fillers among the three conditions, but speakers in *unresponsive*

Table 2. Engagement and disfluency of speech

Variable	Responsive	Face-to-Face	Unresponsive
Duration (in seconds)	189.86_a	101.61_b	94.26_b
Number of Words Spoken	481.00_a	288.89_b	273.33_b
Number of Relevant Words[1]	456.13_a	277.44_b	251.33_b
Overall Speech Rate	2.62_a	2.91_a	2.94_a
Disfluency Rate[2]	0.1258_a	0.1130_a	0.2044_b
Number of Pause Fillers[3]	17.63_a	8.11_a	19.67_a
Pause Fillers Rate	$9.44E\text{-}2_a$	$8.57E\text{-}2_a$	$17.91E\text{-}2_b$
Number of Incomplete Words[4]	7.25_a	3.33_a	2.33_a
Incomplete Words Rate	$3.13E\text{-}2_a$	$2.73E\text{-}2_a$	$2.53\ E\text{-}2_a$
Number of Prolonged Words[5]	10.88_a	3.22_b	2.83_b
Prolonged Words Rate	$5.69\ E\text{-}2_a$	$3.68E\text{-}2_a$	$3.12E\text{-}2_a$

[1] Number of Relevant Words = n – pf - iw (n = Number of words spoken, pf = Number of Pause Fillers, iw = Number of Incomplete Words)
[2] Disfluency Rate = (pf + iw)/d (pf = Number of Pause Fillers, iw = Number of Incomplete Words, d = Duration)
[3] Example of Pause Fillers: "um" and "er".
[4] Example of Incomplete Words: "univers-".
[5] Example of Prolonged Words: "I li::ke it", where ":" signifies lengthened vow "i".
[Note] Means in the same row that do not share subscripts differ at $p < .05$ in the Tukey honestly significant difference comparison.

[3] Data from Responsive and Unresponsive conditions was analyzed and published before. This is a secondary analysis on the data set with additional data from face-to-face condition.

condition used significantly more pause fillers per second than the *face-to-face* and *responsive* conditions. No differences were found in either total number of incomplete words or frequency of incomplete words among three conditions. Also speakers in *responsive* condition used significantly more prolonged words than *face-to-face* and *unresponsive* condition, but there's no significant difference in prolonged words rate among these three conditions.

Contrary to H3, there are no significant differences emerged on self-report of rapport among speakers from different conditions. This may have been a byproduct of the low reliability of the self-report measure, e.g. each measure is a single item scale.

4 Discussion

The main hypothesis (H1) that a virtual human could be more engaging than a human listener was supported, suggesting that such technology can serve, both as a methodological tool for better understanding human-computer interaction, and as a means to establish rapport and its associated a range of socially desirable consequences, including improved computer-mediated learning and health interventions.

Contrary to our predictions (H2), however, face-to-face interactions did not differ significantly in disfluency with those with the RAPPORT AGENT, suggesting a more nuanced relationship between feedback, engagement, and fluency. The literature suggests that rapport serves as a mediating factor: contingent nonverbal feedback promotes rapport, which in turn, promotes beneficial social outcomes such as engagement and speech fluency. Rather, this result is more consistent with linguist theories that argue that nonverbal feedback such as head nods serve a variety of functions based on context. For example, Allwood and Cerrato (2003) argue that head nods serve a rapport-like function, conveying that the listener is paying attention. In contrast, nods can also convey semantically relevant information: that specific content was received or to show attitudinal reactions such as agreement or refusal. One could argue that the former function is more important for engagement, whereas the latter is more important for fluency (see also Bavelas et al., 2000; Cassell & Thórisson, 1999).

Indeed, anecdotal observations are consistent with the interpretation of Allwood and Cerrato (2003). Many head nods seem to be interpreted as an "I'm paying attention" signal that helps promote engagement. We further suspect that it is the frequency, rather than the timing of such gestures that are important for promoting this function. Indeed, the responsive agent generated far more head nods than human listeners. On the other hand, some agent gestures were clearly interpreted as conveying semantic meaning, and many speech disfluencies seemed to arise from the apparent inappropriateness of meaning conveyed by these gestures. Consider the following exchange taken from one of the speakers in the responsive condition: "... and the cat overhears this, and, so, Sylvester goes upstairs dressed as a bellhop [agent: shakes head] [speaker pauses] YES[emphatically], [speaker pause/smile] uh, so, uh, the, he knocks on the door." The agent detected a head shake as the speaker spoke "bellhop" (she actually made a slight side-to-side movement with her head at that moment) and attempted to mimic this gesture. This was interpreted as disagreement by the listener, causing the speaker to apparently lose her train of thought. We suspect such "inappropriate semantic feedback" is responsible for the higher disfluency rate in the responsive condition when compared to face-to-face interaction.

Though our findings on engagement are tantalizing, several methodological factors qualify the generality of our findings and must be considered before attempting to translate them into specific applications. Subjects were not assigned randomly across conditions – the face-to-face condition was run separately and after the other conditions. Thus, we cannot strictly rule out the impact of other incidental factors that might have systematically changed in this condition (e.g., weather, time of year). Additionally, subjects in the responsive and unresponsive conditions were led to believe they were speaking to a human controlled avatar, whereas the behaviors were, in fact, controlled by an intelligent agent. Several studies suggest that people will show similar social effects even if they are aware they are interacting with an agent, although the effects tend to be attenuated (Nass & Moon, 2000). Findings on self report did not reach significance, perhaps do to the lack of precision of our questionnaire (subsequent studies are using a licker scale), however several authors have noted that virtual characters often produces measurable behavioral effects even though subjects may not register awareness of these influences through self-report (Bailenson et al., 2005). Finally, the unresponsive condition varied both the contingency and the frequency of behaviors (e.g., the unresponsive agent did not nod). As the absence of behavior also communicates information, we cannot say definitively if it is the presence or contingency that promotes engagement. A larger study is currently underway that should address these methodological concerns and tease apart the factors that contributed to the observed effect.

This study focused on engagement and speech fluency; however rapport is implicated in a number of social effects including enhanced feelings of trust, greater persuasiveness and greater cooperation during negotiations. It should be straightforward to assess the impact of agent behavior on these other factors. For example, Frank, et. al (1993), showed that short face-to-face interactions enhance subsequent cooperation on simple social games (e.g., Prisoner's Dilemma). An obvious extension to the current study is to include a subsequent negotiation game as an indirect measure of trust/cooperation.

A key 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 but there are several technical obstacles must be overcome. Most speech recognition systems operate in batch mode and only extract meaning several hundred milliseconds after an utterance is complete. We have begun to experiment with continuous speech recognition but this considerably lowers the word accuracy rate. Even if text can be rapidly recognized, techniques for extracting useful meaning typically requires complete sentences. To provide this type of 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. Thus, systems that attempt to create the rapid contingent feedback associated with rapport (e.g., technological advances in incremental speech understanding), must find a way to integrate delayed semantic feedback with the more rapid generic feedback that can be provided by systems like the RAPPORT AGENT.

As this research advances, it will be important to develop a more mechanistic understanding of the processes that underlie rapport and social engagement. Although it is descriptive to argue that rapport somehow emerges in "the space between individuals," this observation is not that helpful when trying to construct a computational model of the process. Clearly, rapportful experiences are emotional experiences, but the specific emotion judgments underlying such experiences elude us. Scherer points one way forward (Scherer, 1993), arguing that subtle nonverbal cues are appraised (consciously or nonconsciously) with respect to each participants own needs, goals and values. For example, subtle cues of engagement may covey to the speaker that they are a good storyteller, and impacts their self-concept or self-esteem. We therefore regard an investigation of the appraisals that underlie feelings of rapport or engagement as an important subject for future research.

Although work on "virtual rapport" is in its early stages, these and related findings give us confidence that virtual characters can create some of the behavioral and cognitive correlates of successful helping relationships. We have explored factors related to effective multi-modal interfaces and assessed these properties in terms of explicit specific social consequences (i.e., engagement and fluency). Findings such as this can inform our understanding of the critical factors in designing effective computer-mediated human-human interaction under a variety of constraints, (e.g., video conferencing, collaboration across high vs. low bandwidth networks, etc.) by helping to identify crucial factors that impact social impressions and effective interaction. Given the wide-ranging benefits of establishing rapport, applications based on such techniques could have wide-ranging impact across a variety of social domains.

Acknowledgements. We thank Wendy Treynor for her invaluable help in data analysis. 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

1. Alibali, M.W., Heath, D.C., Myers, H.J.: 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 (2001)
2. Allwood, J., Cerrato, L.: A study of gestural feedback expressions. In: Paper presented at the First Nordic Symposium on Multimodal Communication, Copenhagen (2003)
3. Bailenson, J.N., Swinth, K.R., Hoyt, C.L., Persky, S., Dimov, A., Blascovich, J.: The independent and interactive effects of embodied agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in Immersive Virtual Environments. PRESENCE: Teleoperators and Virtual Environments 14, 379–393 (2005)
4. Bailenson, J.N., Yee, N.: Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science 16, 814–819 (2005)
5. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as Co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
6. Brand, M.: Voice puppetry. In: Paper presented at the ACM SIGGRAPH (1999)

7. Breazeal, C., Aryananda, L.: of Affective Communicative Intent in Robot-Directed Speech. Autonomous Robots 12, 83–104 (2002)
8. Burns, M.: Rapport and relationships: The basis of child care. Journal of Child Care. 2, 47–57 (1984)
9. Capella, J.N.: On defining conversational coordination and rapport. Psychological Inquiry 1(4), 303–305 (1990)
10. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., et al.: Embodiment in Conversational Interfaces: Rea. In: Paper presented at the Conference on Human Factors in Computing Systems, Pittsburgh, PA (1999)
11. Cassell, J., Thórisson, K.R.: 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 (1999)
12. Chartrand, T.L., Bargh, J.A.: The Chameleon Effect: The Perception-Behavior Link and Social Interaction. Journal of Personality and Social Psychology 76(6), 893–910 (1999)
13. Cogger, J.W.: Are you a skilled interviewer? Personnel Journal 61, 840–843 (1982)
14. Drolet, A.L., Morris, M.W.: Rapport in conflict resolution: accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Experimental Social Psychology 36, 26–50 (2000)
15. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–2000 (1992)
16. Ellsworth, P.C., Scherer, K.R.: Appraisal processes in emotion. In: Davidson, R.J., Goldsmith, H.H., Scherer, K.R. (eds.) Handbook of the affective sciences, pp. 572–595. Oxford University Press, New York (2003)
17. Fogel, A.: Developing through relationships: Origins of communication, self and culture. Harvester Wheatsheaf, New York (1993)
18. Frank, R.: Passions with reason: the strategic role of the emotions. W. W. Norton, New York (1988)
19. Frank, R.H., Gilovich, T., Regan, D.T.: The evolution of one-shot cooperation: an experiment. Ethology and Sociobiology 14, 247–256 (1993)
20. Fuchs, D.: Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education 7, 90–104 (1987)
21. Grammer, K., Kruck, K.B., Magnusson, M.S.: The courtship dance: Patterns of nonverbal synchronization in opposit-sex encounters. Journal of Nonverbal Behavior 22, 3–29 (1998)
22. Gratch, J., Marsella, S.: A domain independent framework for modeling emotion. Journal of Cognitive Systems Research 5(4), 269–306 (2004)
23. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., et al.: Virtual Rapport. In: Paper presented at the 6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA (2006)
24. Kallmann, M., Marsella, S.: Hierarchical Motion Controllers for Real-Time Autonomous Virtual Humans. In: Paper presented at the 5th International Working Conference on Intelligent Virtual Agents, Kos, Greece (2005)
25. Keltner, D., Haidt, J.: Social Functions of Emotions at Four Levels of Anysis. Cognition and Emotion 13(5), 505–521 (1999)
26. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., et al.: Towards a common framework for multimodal generation in ECAs: The behavior markup language. In: Paper presented at the Intelligent Virtual Agents, Marina del Rey, CA (2006)
27. Kramer, N.C., Tietz, B., Bente, G.: Effects of embodied interface agents and their gestural activity. In: Paper presented at the Intelligent Virtual Agents, Kloster Irsee, Germany (2003)

28. Marsella, S., Johnson, W.L., LaBore, C.: Interactive pedagogical drama for health interventions. In: Paper presented at the Conference on Artificial Intelligence in Education, Sydney, Australia (2003)
29. McNeill, D.: Hand and mind: What gestures reveal about thought. The University of Chicago Press, Chicago, IL (1992)
30. Morency, L.-P., Sidner, C., Lee, C., Darrell, T.: Contextual Recognition of Head Gestures. In: Paper presented at the 7th International Conference on Multimodal Interactions, Torento, Italy (2005)
31. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues 56(1), 81–103 (2000)
32. Nass, C., Reeves, B.: The Media Equation. Cambridge University Press, Cambridge (1996)
33. Neal Reilly, W.S.: Believable Social and Emotional Agents (Ph.D Thesis No. CMU-CS-96-138). Carnegie Mellon University, Pittsburgh, PA (1996)
34. Oppenheim, A.V., Schafer, R.W.: From Frequency to Quefrency: A History of the Cepstrum. IEEE Signal Processing Magazine, pp. 95-106 (September 2004)
35. Parkinson, B.: Putting appraisal in context. In: Scherer, K., Schorr, A., Johnstone, T. (eds.) Appraisal processes in emotion: Theory, methods, research, pp. 173–186. Oxford University Press, London (2001)
36. Robins, B., Dautenhahn, K., Boekhorst, R.t., Billard, A.: Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills? Special issue, Design for a more inclusive world of Universal Access in the Information Society, 4(2) (2005)
37. Sakaguchi, K., Jonsson, G.K., Hasegawa, T.: Initial interpersonal attraction between mixed-sex dyad and movement synchrony. In: Anolli, L., Duncan Jr, S., Magnusson, M.S., Riva, G. (eds.): The hidden structure of interaction: from neurons to culture patterns. Amsterdam (2005)
38. Scherer, K.: Comment: interpersonal expectations, social influence, and emotion transfer. In: Blanck, P.D. (ed.) Interpersonal Expectations: theory, research, and applications, pp. 316–333. Combridge University Press, Paris (1993)
39. Tatar, D.: Social and personal consequences of a preoccupied listener. Stanford University, Stanford, CA (1997)
40. Tickle-Degnen, L., Rosenthal, R.: The Nature of Rapport and its Nonverbal Correlates. Psychological Inquiry 1(4), 285–293 (1990)
41. Tosa, N.: Neurobaby. ACM SIGGRAPH, 212–213 (1993)
42. Tsui, P., Schultz, G.L.: Failure of Rapport: Why psychotheraputic engagement fails in the treatment of Asian clients. American Journal of Orthopsychiatry 55, 561–569 (1985)
43. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 23, 1177–1207 (2000)
44. Welji, H., Duncan, S.: Characteristics of face-to-face interactions, with and without rapport: Friends vs. strangers. In: Paper presented at the Symposium on Cognitive Processing Effects of 'Social Resonance' in Interaction, 26th Annual Meeting of the Cognitive Science Society (2004)
45. Yngve, V.H.: On getting a word in edgewise. In: Paper presented at the Sixth regional Meeting of the Chicago Linguistic Society (1970)