

**ESTABLISHING RAPPORT WITH CONVERSATIONAL
AGENTS: COMPARING THE EFFECT OF ENVELOPE AND
EMOTIONAL FEEDBACK**

JOSHUA WONG WEI-ERN

(B. Soc. Sci (Hons.), NUS)

**A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF ARTS**

**DEPARTMENT OF COMMUNICATIONS AND NEW MEDIA
NATIONAL UNIVERSITY OF SINGAPORE**

2012

Acknowledgements

Thanks to my two supervisors, Timothy Marsh and Kevin McGee for the help and advice they have provided throughout the research process, and being willing to sort through my confusion as I tried to figure out what to do. Thanks to my lab mates at the ParTech lab – Alex Mitchell, Tim Merritt, Chris Ong, Chuah Teong Leong and Maryam Azh – for being there to bounce ideas off and the questions and critiques that have made me think through my research better. Thanks to my colleagues at CNM, for providing a warm and research-friendly environment to work in. Thanks to my family and my church cell members for constant encouragement to finish on time. And finally, thanks to God for His guidance, strength and motivation throughout the entire journey.

Table of Contents

Acknowledgements	i
Table of Contents	ii
Summary	iii
List of Figures	iv
List of Tables	v
1. Introduction	1
1.1. The History of Embodied Conversational Agents	1
1.2. Envelope versus emotional feedback in conversation.....	2
1.3. Affective Conversational Agents	3
1.4. Rapport.....	4
1.5. How to read this thesis	6
2. Related Work	7
2.1. Long-Term Rapport: Relational Agents.....	7
2.2. “Instant” Rapport-Building Agents.....	11
2.3. Summary of Related Work	14
3. Research Problem	15
4. Methodology	18
4.1. Participants.....	19
4.2. Materials	19
4.3. Session Protocol.....	23
4.4. Data Gathering	25
5. Results.....	26
6. Discussion.....	31
6.1. Conversational Grounding	31
6.2. Social Anxiety.....	34
6.3. Language Barriers	35
7. Limitations	37
8. Conclusion	39
Bibliography	41

Summary

Conversational agents have become increasingly sophisticated in interacting with humans since their early days as embodied interfaces. One way to improve the design of conversational agents is to study rapport – that feeling of being able to ‘click’ with someone – and how agents can be designed to build rapport with humans. So far, most research on rapport using agents has focused on *envelope feedback* – nonverbal behaviours that facilitate the process of communication without reference to the content of the conversation. This thesis will examine the effect of *emotional feedback* – nonverbal behaviours that indicate an emotional response to the content of a conversation – in an agent, and how this can affect rapport with humans in a storytelling scenario.

Thirty-six people took part in an experiment in which they re-told a sequence of events they had witnessed to an agent that was capable of producing both appropriate and inappropriate facial expressions. The rapport between the human and the agent was then measured through the duration of the story being told, the fluency of the speaker, and the self-reported feelings of rapport by the speaker. Results showed that inappropriate emotional feedback (in the form of facial expressions) caused the duration of the interaction to increase, which was the opposite of earlier studies on envelope feedback on rapport. Possible explanations for this effect could be attributed to emotional feedback’s greater impact on the conversational grounding process and on the speakers’ social anxieties, or by language barriers. This study shows that emotional feedback does have an impact on rapport, and in a way that is different from envelope feedback, and thus makes it an important factor to consider in the design of conversational agents.

List of Figures

Figure 1. Experimental Setup	20
Figure 2. Lab Layout	24
Figure 3. Speaker's evaluation of avatar's understanding of story	27
Figure 4. Speaker's evaluation of connection with other person through avatar	28
Figure 5. Speaker's evaluation of avatar's helpfulness when seen	28
Figure 6. Speaker's usage of feedback from avatar while telling story	29
Figure 7. Speaker's evaluation of accuracy of avatar's portrayal of Listener	29

List of Tables

Table 1. FACS Action Units used to create facial expressions.....	21
Table 2. Length of Interaction & Speech Fluency Results	26

1. Introduction

In the field of Human Computer Interaction, an emerging area of research is in the design of *conversational agents* – automated computer characters that can engage human users in conversation in a way that would be similar to another human talking with them. In designing these agents, various factors are being investigated. We begin by providing a quick overview of the history of conversational agents, and discuss the two elements of envelope versus emotional feedback in conversation. Then, we will examine embodied conversational agents that have an *affective* (emotion-related) component to them, which include agents that were built to display emotions, agents designed to persuade people, and agents designed to build long-term trust through their conversational interactions.

1.1. The History of Embodied Conversational Agents

Since the early days of conversational agents, there has been a growing trend in the complexity and variety of responses that these agents can display to humans interacting with them. Early conversational agents were mostly text-based, and became known as chatbots. Ever since the early anecdotal evidence of chatbots like ELIZA (Weizenbaum, 1966) being treated as if they were real people, and the pioneering work of Reeves and Nass in showing that people can treat computers as social actors (Reeves & Nass, 1996), research in the field of conversational agents has been aimed at finding ways to design agents that would respond like human beings.

This naturally led to the development of Embodied Conversational Agents (ECAs), which were conversational agents that had the same properties as humans in face-to-face conversation (Justine Cassell, et al., 1998), including:

- The ability to recognise and respond to verbal and non-verbal input
- The ability to generate verbal and non-verbal output
- The use of conversational functions such as turn-taking, feedback and repair mechanisms

- A performance model that allows negotiation of the conversational process, and contributions of new propositions to the discourse.

The beginning few Embodied Conversational Agents focused on two things: modelling real-time generation of verbal and non-verbal behaviours in agents (Justine Cassell, et al., 1994), and designing for task-oriented conversations, in which the agent was helping or assisting the user in accomplishing a task. These included agents like Gandalf (Justine Cassell & Thorisson, 1999), an agent which helped guide children through an interactive model of a solar system, and Olga (Beskow & McGlashan, 1997), an agent which provided information for consumers that wanted to purchase a microwave. For the most part, these agents were simply embodied animated interfaces. They focused on building what became known as the *conversational envelope* – the behaviours that surrounded and sustained the *process* of conversation, without much reference to the actual conversational content. To understand this further, a short introduction to feedback in conversation is needed.

1.2. Envelope versus emotional feedback in conversation

During a conversation, there are two possible forms of feedback that a listener can provide to the speaker. These have been variously defined as *general* versus *specific feedback* (Bavelas, Coates, & Johnson, 2000), or as *envelope* versus *emotional feedback* (Justine Cassell & Thorisson, 1999) for the specific case of nonverbal behaviours.

In the first case of *general* or *envelope feedback*, the response of the listener is unrelated to the actual content of the conversation, but rather serves as a comment or aid to the process of communication. Examples include head nods to indicate that the listener has heard and understood the message, body postural shifts that mimic the speaker's, and eye gaze behaviour that tracks the speaker's movements.

In the second case of the *specific* or *emotional feedback* (also called *semantic* or *content feedback*), the listener responds directly to the content of the conversation. This usually takes the form of an emotional display of some sort, such as a sharp intake of breath to indicate suspense or surprise, smiles (genuine or faked) that indicate interest and/or happiness, facial expressions, verbal exclamations, continuation of the narrative via back-channels, and others.

In the early days of Embodied Conversational Agents, most research concentrated on building envelope feedback behaviours – agents that would follow rules that allowed them to successfully take part in the *process* of communication, in a similar way to a human being. Later on, as design of conversational agents became more sophisticated, ECAs began to be designed to address the *content* of communications, and include some *affective* components – taking into account human emotions as part of the process of interacting with users.

1.3. Affective Conversational Agents

Following Rosalind Picard's seminal book on *Affective Computing* (Picard, 1997), emotions became recognised as an important part of the cognitive and decision-making processes among researchers working on agents, and there were several attempts to address emotional issues through the design or use of conversational agents. These ranged from designing agents like GRETA that were capable of displaying complex emotions through multimodal methods (Bevecqua, Mancini, Niewiadomski, & Pelachaud, 2007), to creating agents that were able to model emotion processing and arrive at different kinds of emotional states based on their appraisal of the situation (Bartneck, 2002).

Other researchers became interested not so much in the science of how to create characters that were able to model or display emotions, but rather what effects those displays of emotion by agents had on human beings interacting with these agents. These studies implemented certain agent behaviours and evaluated them with users in order to learn more about a particular socio-emotional phenomenon (such as persuasion, empathy or rapport) and designing agents to better suit those purposes.

For example, Krumhuber et al. (Krumhuber, Manstead, Cosker, Marshall, & Rosin, 2009) studied the effects of genuine versus faked smiles of synthetic characters on likability and trust in a simulated job interview setting. They found that human interviewers rated characters who had 'genuine' smiles more positively than those which had 'faked' smiles, and both better than those which had neutral expressions. Their results paralleled the pattern of findings found in studies on human smiles as well.

In the study of persuasive agents, Bailenson and Yee designed and evaluated an agent that mimicked user's head movements while reading a persuasive message to them (Bailenson & Yee, 2005). They

found that agents that mimicked users were rated as more effective than agents which did not, thus showing that the usage of envelope feedback (in the form of head movements and eye gaze) had a positive impact on the persuasiveness of an agent. Ochs and Prendinger (Ochs & Prendinger, 2010) also examined persuasion, but in the context of agents that followed different emotional strategies versus agents that acted impulsively on their own emotional states in trying to persuade another person in a negotiation. They found that the agent that intentionally followed an emotional strategy to be more successful than one which acted impulsively, and of the three emotional strategies that were used, the empathetic agent was the most persuasive.

In the study of empathy, Brave et. al (Brave, Nass, & Hutchinson, 2005b) used agents that played the card game blackjack against the player, and either made empathic (other-oriented) comments or self-oriented comments about the results of the game. They discovered that players thought of empathic agents as more caring, more likable, more trustworthy and more submissive than non-empathic agents. They also found that players felt more supported during interactions with an empathic agent, though not necessarily more positive. Similarly, Pereira et al. (Pereira, Leite, Mascarenhas, Martinho, & Paiva, 2010) conducted an experiment in which a robot cat (iCat) was commenting on a chess game between two human players. It would direct empathic comments towards one player, and neutral comments towards the other. The results showed marked differences in how the players rated the friendliness of the cat, with the player which the cat had empathized with rating the cat much higher than the player who had neutral comments.

1.4. Rapport

While research into socio-emotional aspects of conversational agents has proceeded in various directions, one approach that has been coming to light in the last decade has been to focus on the activity of conversation itself – how human beings converse naturally with each other, and how we can identify the traits or characteristics of a good conversational partner, regardless of its nature as human or machine. So what makes for a good conversation partner? Many people have had the pleasurable experience of engaging in a conversation with a partner that seemed to just ‘click’ with them – their thoughts seemed to be interesting and stimulating, they built off each others’ replies, and

there was a sense of mutual understanding and warmth in their interactions with each other during conversation. This feeling, which we have all experienced in some degree or another, is commonly called *rapport*.

But what exactly *is* rapport, and how can it be formally defined? Investigations into the nature of rapport have primarily been the work of psychologists Tickle-Degnen and Rosenthal (Tickle-Degnen & Rosenthal, 1990). They firstly stated that rapport exists only in the interactions between individuals, and not as a property or quality of the individuals themselves, although some people may be better than others at establishing rapport under certain conditions. Tickle-Degnen and Rosenthal then go on to describe three essential components of rapport, and its' behavioural correlates. The three components that they name as essential to rapport are: mutual attentiveness, positivity, and coordination.

When participants in an interaction are feeling a high degree of rapport, they become *mutually attentive* to each other: "Their focus is directed toward the other, is other-involved. They experience the feeling as one of intense mutual interest in what the other is saying or doing." (Tickle-Degnen & Rosenthal, 1990) Secondly, there is an element of *positivity* in high rapport – a sense of "mutual friendliness and caring". Lastly, the third element of rapport is *coordination* between the participants – the sense that each person is responding 'in sync' with the other's actions, as in the example of a well-conducted orchestra. As Park and Burgess describe it, "Rapport implies the existence of a mutual responsiveness, such that every member of the group reacts immediately, spontaneously, and sympathetically to the sentiments and attitudes of every other member." (p. 893) (Park & Burgess, 1924)

While all three components are necessary for rapport, Tickle-Degnen go on to describe the relative weightage of each component during the interaction. They suggest that, when it comes to determining the level of rapport at different stages of development in the relationship/interaction, participants weigh the three components differently. In the early stages of an interaction or a relationship, participants pay more attention to positivity (warmth and friendliness) than the level of coordination

between the parties, when determining the level of rapport. During later interactions, as participants have become more familiar with each other, the level of coordination becomes a stronger factor for indicating the level of rapport than positivity. Mutual attention remains constant and high throughout, as it is difficult to imagine rapport existing when one of the participants is not paying attention. The increase in coordination and decrease in positivity is further verified through a study done recently by Cassell, Gill and Tepper which measured behavioural interactions between friends and strangers at a direction-giving task on three separate occasions. (Justine Cassell, Gill, & Tepper, 2007)

In summary, we understand the nature of rapport as a property of an interaction between two or more parties where there is a high degree of mutual attention, positivity and coordination. Being able to build rapport with another person is one of the elements of a good conversational partner, and hence, it would be important for the design of conversational agents. Thus, the question we now turn to is: What has been done so far in the design and evaluation of conversational agents that try to build rapport with humans?

1.5. How to read this thesis

The rest of this document is structured as follows: Having clarified the nature of rapport and provided an overview of the development of conversational agents, we now address what has been done in the field of conversational agents that are designed for rapport in Chapter 2. This is followed by the statement of the research problem and the specific contribution that this thesis proposes to make in Chapter 3. Chapter 4 will outline the methodology through which the study will be accomplished, and Chapter 5 will report the results of the study. In Chapter 6, the study results will be discussed, to explain how emotional feedback relates to rapport. Chapter 7 will describe the limitations and possible critiques of the project, and finally in Chapter 8 we will discuss the implications of this study and future work required.

2. Related Work

This section provides an overview of the latest work in the design of conversational agents for the purposes of building rapport. It is important to note that work on rapport in conversational agents can be roughly divided into two categories: studies that look at multiple interactions over a long period of time, and studies that examine the rapport that is established during one session. The former category is known as the study of *relational agents* – agents that seek to build long-term relationships with users. The latter category was characterised by Cassell, Gill and Tepper (Justine Cassell, et al., 2007) as ‘instant rapport’.

This chapter is divided into two parts: in the first section, we will look at what has been done in the field of relational agents, and other studies of agents that attempt to build rapport over multiple interactions. In the second section, we examine studies of agents that have been designed for ‘instant rapport’ in a single interaction – primarily the Rapport Agent and how it has been used. Finally, at the end a quick summary is provided to wrap the two halves together, and lead towards the Research Problem.

2.1. Long-Term Rapport: Relational Agents

Relational agents are defined as “computational artifacts designed to build long-term, socio-emotional relationships with their users” (Bickmore & Picard, 2005). Primarily led by Timothy Bickmore and the Relational Agents group at Northeastern University, various studies have looked at how agents can be designed for long-term relationships by establishing rapport and trust with humans.

Bickmore & Gruber (Bickmore & Gruber, 2010) provide an overview of the work that has been done on building effective relational agents for therapy and summarise several key elements that go into designing such agents. They state that “strategies for forming a strong therapeutic alliance are generally those that enable the user to respond to expected social behaviours”. This includes various forms of verbal relationship-building behaviours, and non-verbal behaviours that project liking for the other person and engagement in the interaction, as described in the experiment with Laura later. Other

studies also found that humans are less likely to be annoyed by agents reminding them to take medicine if the agent followed human social conventions (Bickmore, Mauer, Crespo, & Brown, 2008; Liu & Picard, 2005), and animation in agents contributed significantly to the establishment of a working alliance (Bickmore & Mauer, 2006).

In addition to the factors mentioned above, they also state that “empathy is a key element in forming strong helping relationships”, citing studies which showed that empathy alone, when used by a computer agent, promoted liking of the agent, reduced frustration and encouraged continued interaction with the agent (Brave, Nass, & Hutchinson, 2005a; Klein, Moon, & Picard, 2002).

Bickmore & Gruber also point out that “maintaining relationships with agents over months and years requires additional strategies”, such as variability in agent responses (Bickmore & Schulman, 2009), self-disclosure by the agent (Bickmore, Schulman, & Yin, 2009), and referencing knowledge of prior interactions.

An early example of these is REA, an agent designed to engage users in small talk in order to build trust for the purposes of selling them real estate (Bickmore & Cassell, 2001). REA was a multimodal agent that was able to respond to the user’s speech, gaze shifts, gestures and non-speech audio, through its’ own combination of speech with intonation, hand gestures and facial displays (J. Cassell, et al., 1999). It was an attempt to build an agent that could merge both envelope and emotional feedback (also called *interactional* and *propositional feedback*). REA had a dialogue planner which was able to plan out what to say next and the conversational strategies it would follow, while handling turn-taking, dialogue repair, and appropriate nonverbal cues for both listening and speaking. The planning of the conversational strategy was based on three factors: the depth of self-disclosure achieved between participants, the amount of information known about each other through the conversation, and the solidarity or ‘like-mindedness’ of their dispositions. If the system detected a lack of familiarity with the user, it would engage in small talk first and pursue conversational strategies to increase the depth and breadth of knowledge it has about the user before settling down to the task of selling them a house.

In one study (Bickmore & Cassell, 2001), two versions of REA were evaluated – one which included small talk before getting down to the business of selling a house, and the other which was purely task-oriented and started task talk immediately. What they discovered was that while the inclusion of small talk had little effect on user trust in introverts, it had a large impact on the level of trust in extroverts – such that the level of trust in the agent was much less if REA had not engaged in small talk. Likewise, people who initiated conversational interactions preferred small talk versus people who passively let REA take charge of the conversation. In other words, people who had personalities that reached out more to other people were the ones who needed and benefited most from having an agent that could use relationship-building conversational strategies.

The studies by the Relational Agents group have primarily centred on building agents for clinical settings and healthcare counselling. An example of this is the MIT FitTrack system, featured an agent Laura which was designed as a personal exercise advisor to help users increase their levels of physical activity, through projecting care and concern and building a relationship of trust with them (Bickmore & Picard, 2004, 2005). Eighty-four participants interacted with the system for about 10 minutes every day for a month. Laura was equipped to show a wide variety of responses, which included engaging in small talk (social dialogue), talking about the relationship itself (meta-relational dialogue), different forms of address, politeness strategies, empathy exchanges, humour, greetings and acknowledgements of time apart (continuity behaviours), and exhibiting nonverbal feedback. The non-verbal feedback included body postural shifts, gestures, proximity changes, four different types of facial expressions, gaze shifts, and head nods/shakes. She also had a history function, to remember user preferences expressed in conversation over the course of their 1-month interactions. To make all these possible, the user's conversation with Laura was restricted to selecting from a set of text responses, to which Laura's replies could then be scripted and controlled.

In order to study the effectiveness of Laura, two conditions were used: a Relational condition, in which the agent pursued relationship-building strategies, and a Non-Relational condition, in which the agent did not attempt to build a relationship. The study measured the working alliance established between the agent and users after 7 days into the program, and nearing the end. They also measured

the relationship quality through a behavioural measure – whether the humans chose to say goodbye to Laura in a businesslike or ‘sentimental’ way. Lastly, they measured the efficacy of the program in supporting exercise, and how often participants chose to interact with it.

The study showed that the Relational agent that pursued relationship-building strategies was able to establish higher levels of friendship and trust than the Non-Relational agent, and both were higher than the control condition. Although the agents did help to cause a behavioural change, a test performed 15 days after the end of the treatment showed that lasting change was not achieved, though this is also common with human behavioural therapists, and thus indicative that Laura managed to simulate results that were similar to a human exercise coach.

Another study involved a home-based relational agent designed to help remind schizophrenic patients to take their medicine, and encourage physical activity as well (Bickmore & Pfeifer, 2008). The agent was designed for daily interaction for a period of one month, as a standalone system. It tracked patients taking their medication through self-report, and if the patient had not taken it, it would deliberately ask the patient to take the medication now while it waited. It also reminded them about getting refills and instructs them about behavioural self-maintenance, such as getting multi-compartmented pillboxes and other ways they could help themselves.

While the experimental sample size was small (20 participants), the results were quite promising. The self-reported satisfaction ratings averaged out to a high 4.0 on a 5-point scale, and user adherence to medication was 89%. System logs from the agents showed that on average, users talked to the agent 65.8% of the days available. Overall, these results indicate that the agent was effective and satisfying for users to interact with.

A third study featured the creation of a Virtual Discharge Nurse agent, to counsel hospital patients on their self-care regimen before they are sent home from hospital (Bickmore, Pfeifer, & Jack, 2009). It was specifically designed to be able to explain written hospital discharge instructions to patients who had low health literacy. Discharge conversations from existing nurses were recorded and provided the basis for a model conversation for the agent to follow. The agent spent on average half an hour with

each patient, reviewing the discharge instructions, testing them for comprehension, and flagging unresolved issues for a human nurse to follow up on.

Results of two rounds of pilot studies showed that patients found the system easy to use (averaging a 6.8 rating on a 7-point scale), and reported high levels of satisfaction (6.7 on a 7-point scale). 74% of patients reported that they preferred receiving the instructions from the agent than a doctor or nurse, and expressed appreciation for the time and attention provided by the virtual nurse. They also felt that it was an authoritative source of information, and the ones that had a nurse with displayed Relational behaviours also reported feeling that the agent cared for them significantly more than those who got a nurse in a Non-Relational condition. These results generally indicate the usefulness of relational agents who are able to spend additional time and care with the patient to answer their questions in a low-pressure setting, even if the nurses are automated.

Finally, in a related field, a social robot, Autom, was built to promote diet-tracking among overweight users (Kidd, 2007). It was programmed to display appropriate greetings and limited social chat. For evaluation, 45 obese participants interacted with either the robot, a touch screen computer, or used a paper journal to record their daily eating behaviour. Results showed that participants rated the robot significantly higher in the formation of a working alliance compared to the touch screen or a paper diary. Also, participants who interacted with Autom continued to record their eating behaviour much longer than the other two conditions. Although there was no significant difference in weight loss between the conditions, the study had value in showing that embodied agents were preferable to non-personified systems.

2.2. “Instant” Rapport-Building Agents

We now move on to agents which are constrained to smaller, more focused interactions within a single conversation. Work on conversational agents and instant rapport has primarily been led by Jonathan Gratch and his colleagues at USC, using the Rapport Agent (Gratch, et al., 2006; Gratch, Wang, Gerten, Fast, & Duffy, 2007). The Rapport Agent is an Embodied Conversational Agent

(ECA) which is able to produce real-time envelope feedback, with an emphasis on *contingent feedback* (feedback that is tightly-coupled to what the speaker is doing at that moment). In the case of Rapport Agent, it would detect head shifts, eye gaze, gestures, vocal intensity, range and backchannel opportunities while listening to the user. (Gratch, et al., 2006) It would then respond with a combination of eye gaze, head nods/shakes, and postural shifts. Using the Rapport Agent, Gratch and colleagues studied the effects on envelope feedback on rapport, measured through a set of behavioural observations as well as self-report questionnaires.

There were two studies done to evaluate the effectiveness of the Rapport Agent. The first study (Gratch, et al., 2006) aimed at validating the use of the Rapport Agent as a tool for studying the effect of envelope feedback on rapport, using a scenario where the users would tell stories to an agent that would provide contingent envelope feedback. Thirty subjects were evaluated, and measurement of rapport was done through analysis of the story length, speaker fluency, and five questionnaire items that included both self-reported feelings of rapport and evaluations of the agent's effectiveness.

The results of this study showed that agents which displayed responsive (contingent) envelope feedback ended up with speakers that told longer stories, both in terms of word count as well as time taken, compared to agents that were unresponsive. Speaker fluency was partially affected, with the number of disfluencies significantly increasing in the unresponsive condition compared to the responsive condition, but speech rate had a more complex relationship with envelope feedback which could not be disambiguated through the results of this study. The self-report questionnaire items did not show significance in the measures that linked to rapport (evaluation of the agent's understanding of the speaker and feeling a connection with the agent), but there was evidence that people paid more attention to feedback from the agent in the responsive condition than the unresponsive condition.

A second more complex study was performed to evaluate the Rapport Agent and how contingency of envelope feedback affects rapport (Gratch, et al., 2007). There were four conditions: a human Speaker talking face-to-face to a human Listener (Face to Face condition), a human Speaker talking to an avatar that mimicked the human Listener's head and body movements (Mediated condition), a human

Speaker talking to an avatar that displayed contingent envelope feedback based on the Speaker's own verbal and nonverbal cues (Responsive condition), and a human talking to an avatar that displayed envelope feedback that was recorded in response to a different Speaker's verbal and nonverbal cues (Non-Contingent condition). They constructed a 10-item rapport scale to measure emotional and cognitive rapport, and analyzed the stories that were told to get the behavioural cognates of rapport. They also constructed scales to measure Speaker's evaluations of the agent's likeability and trustworthiness, the Speaker's evaluation of their own storytelling performance, and the Speaker's shyness and personality, which were reported in other studies.

Altogether 131 people took part in this experiment. The results reported in this study on rapport and contingency showed that the Responsive agent was rated as good as a human Face-to-face when it came to building rapport, but the Mediated avatar was rated as the lowest among the four conditions with regards to rapport, and engendering more pause-fillers in conversation. The results also showed that contingency was a major factor in affecting rapport: more words were spoken in the Non-contingent condition, with more pausefillers and disfluencies than in the Responsive (contingent condition). Thus, the study showed that when designing for virtual rapport with agents, it was important for envelope feedback to be tied closely to what the human user is doing at the moment.

The results of the Rapport Agent experiment above have also been reported in various other papers, with different foci. In one paper on social anxiety (Sin-Hwa Kang, Jonathan Gratch, Nina Wang, & James H. Watt, 2008), they conducted multiple regression analysis relating Shyness (social anxiety) to evaluations of Self-Performance, Embarrassment, Trustworthiness of the listener, Likability of the listener, and Rapport while controlling for the effects of both Public and Private Self-Consciousness. They showed that Shyness significantly reduces the evaluation of Self-Performance and increases Embarrassment when users told stories to the Non-contingent agent, although it had no statistically-significant impact on either measure in the other three conditions. Shyness also decreased their sense of Rapport when interacting with the Non-contingent agent, though not significant in other conditions. Lastly, Shyness reduced the evaluation of Trustworthiness of the listener in the two human Face-to-Face condition, but had no significant influence elsewhere, or on Likability.

Another paper on gaze behavior (Wang & Gratch, 2010) involved 144 participants evaluating the Rapport Agent under three conditions: Responsive (similar to the condition above), Staring (continuous fixed gaze on the speaker while displaying idle-time animations), and Ignoring (gazing randomly around the room with occasional eye-flicks to the speaker while displaying idle animations). Results showed that people felt significantly more rapport with the Responsive agents that responded to their actions through head movements and posture mimicking than either of the two other agents, which had roughly equal scores. Users also reported being more distracted by the agent's feedback in the Staring and Ignoring conditions than in the Responsive one. The main contribution of this paper showed that continuous gaze behavior through staring was about as equally detrimental and distracting to rapport as an agent that ignored the Speaker.

2.3. Summary of Related Work

In summary, embodied conversational agents have risen in complexity and sophistication, from the early days as embodied interface elements to actual intelligent agents capable of pursuing conversational goals while interacting with users. While many agents are still task-oriented helpers, increasingly more agents are being designed for social and emotional pursuits, including empathy, persuasion and rapport. Of those agents designed for rapport, some have looked at long-term relationship building, using a variety of multimodal methods to create agents that can inspire trust and engage in a variety of relational strategies, while others have focused on short-term interactions, and built agents that can develop rapport quickly through the use of contingent envelope feedback behaviours. The next section will outline the specific contributions to knowledge made by this thesis towards the field of designing agents to build rapport.

3. Research Problem

While there have been some studies done on agents that use both envelope and emotional feedback (most notably REA and Laura, in the Relational Agents section above), those studies take into account user reactions to the agent as a whole. There has been little research on the specific effects of envelope feedback versus emotional feedback on rapport in agents. Most of the studies that examine rapport in conversational agents have focused primarily on envelope feedback or conversational strategies. There is as yet little information about how emotional feedback specifically plays a role in establishing and maintaining rapport in a conversation with an agent. This study therefore seeks to answer the question: **How is rapport affected when an agent displays emotional feedback?**

In this section, I quickly outline the justification for a study that specifically looks at the effect of emotional feedback on rapport. I then go into greater detail about the basis for the protocol I would be following in the study, which is adapted from Gratch et. al's first study with the Rapport Agent.

While Cassell and Thorisson (Justine Cassell & Thorisson, 1999) did do a study that compared the effects of envelope versus emotional feedback in an embodied conversational agent and found envelope feedback to be more effective, there were a few limitations to their study which can now be challenged: First, the range of emotional expressions of the agent in the study was limited to just two responses – puzzled or happy. This was also in part due to the nature of the scenario set forth in the experimental design – the users were supposed to interact with the agent “as naturally as possible” to discover more about the solar system, a task-oriented scenario. At that time, they argued that embodied conversational agents were most commonly-used in task-oriented settings (such as checking emails), and thus their work was based upon the most expected settings and tasks envisioned for ECAs at that time.

However, over the last decade there has been a paradigmatic shift in the field of conversational agents from task-oriented fields to more social and relational settings. Thus, the importance of the emotional and social aspects of embodied agents and their responses have grown in prominence, especially in non task-oriented settings, such as friendship, small talk, storytelling or persuasion. Similarly, the

research literature on emotional expressivity and agents has grown to the point where it is now feasible to conduct more in-depth studies on the usage of emotional feedback in conversational agents.

Furthermore, according to Tickle-Degnen's observations of the relative weightage of the three components of rapport (Tickle-Degnen & Rosenthal, 1990), while mutual attention remains constantly high throughout, during early interactions, positivity is weighted more heavily than coordination in the determination of rapport. For agents that are building instant rapport, therefore, it is theoretically more effective to build an agent that can respond with positivity than coordination. While envelope feedback can display mutual attention, and contingent feedback in terms of timing allows for some element of coordination, it doesn't do so well in showing positivity, being limited to only head nods. Thus, emotional feedback that can show greater or varying amounts of positivity may be more useful than envelope feedback, especially during one-time interactions with agents.

In order to help validate the usefulness of this study, it is important to have a measuring stick against which the effects of emotional feedback can be seen. Therefore, this study was modelled on the initial study done by Gratch et. al in the creation of the RapportAgent, except that instead of looking at envelope feedback, *emotional feedback* will be monitored. By doing so, this thesis aims to be able to compare the effects of emotional feedback against the results of envelope feedback produced by the RapportAgent.

This study focuses on dyadic interactions in a storytelling or narrative situation, where one participant retells a story to a listener. This situation was first proposed by Duncan and Welji, in their study of rapport among friends and strangers (Duncan & Welji, 2004), and was adopted by Gratch and colleagues to test their RapportAgent's performance. In the experimental design, one participant would watch a cartoon and then describe it to another person or listening agent who hadn't seen it. In their study, Gratch et al. created an agent which would make head movements (nods, shakes) and change its' gaze direction in response to the user communicating with it. They created two agent conditions – a responsive condition, in which the agent would provide responses that were contingent

and appropriate to the speaker's story, and an unresponsive condition, in which the agent would display pre-recorded responses to someone else's story. They then measured the rapport generated through a mixture of behavioural analysis – length of story and speech fluency of the speaker – and self-reported indications of rapport by the speaker. They hypothesised that speakers would speak longer if they felt more rapport with the agent, and they would also speak more fluently, and indicate higher levels of rapport in the self-report questionnaire.

Gratch et. al found that subjects who were exposed to an agent that responded with the appropriate envelope feedback talked significantly longer and used more meaningful words in conversation than those who had an unresponsive agent. The ones with a responsive agent also talked more fluently than those with an unresponsive agent. However, the self-reported feelings of rapport were mixed, with participants feeling sure that the agent had understood them, but not very certain about establishing a connection with the other person. They also had mixed feelings about whether the avatar's movements were helpful or disturbing, and whether it accurately represented a human being. Upon further questioning, Gratch et. al discovered that one of the major barriers that caused a loss of believability was the lack of facial expressions, and that many subjects felt that they were missing a significant part of the feedback that a real listener would provide.

This study therefore attempted to follow a similar protocol as the first Gratch et. al study, but used emotional feedback (specifically in the form of facial expressions) instead of the envelope feedback that the Rapport Agent used. In order to provide a comparison, this study sought to answer three research questions that were similar to the questions asked in the RapportAgent study:

RQ1. How would a listening agent that displays appropriate emotional feedback affect the length of the conversation compared to an inappropriate agent?

RQ2. How would a listening agent that displays appropriate emotional feedback affect the fluency of the speaker compared to an inappropriate agent?

RQ3. How would a listening agent that displays appropriate emotional feedback affect the self-reported rapport felt by the speaker compared to an inappropriate agent?

4. Methodology

This chapter describes the research protocol that was used to answer the three questions posed at the end of the last chapter. The study protocol that was used is similar to the one used by Gratch et. al. in their evaluation of the RapportAgent, which in turn was inspired by the work of Duncan and Welji in their studies of face-to-face rapport among friends and strangers (Duncan & Welji, 2004). It focuses on a storytelling scenario, in which one partner in a conversational dyad (a Speaker) describes a sequence of events he had witnessed to the other partner in the conversation (a Listener). While the Speaker told the story, the Listener would respond with different kinds of verbal or nonverbal feedback, and then the Speaker's experience of rapport with the Listener would be evaluated.

Thirty-six participants took part in this study, split into pairs (or dyads). Each participant took turns to be a Speaker and a Listener within their dyad. Speakers would watch a short video of a funny cartoon, then re-tell the story of what they saw to an avatar. They were told that the avatar represented the Listener's facial expressions in real-time. In actuality, the avatar was being controlled by a wizard, simulating the behaviour of an intelligent, computerized agent. The participants were randomly assigned to either an *appropriate* or *inappropriate* condition, and the agent followed certain rules in the type and frequency of facial expressions shown to the Speaker as they told the story. After one participant had described the video to the avatar, the Speaker and Listener would switch roles, and the new Speaker would repeat the experiment with a different video. A post-experiment questionnaire was administered to measure self-reported feelings of rapport, and transcripts of the stories told were analyzed for behavioural indicators of rapport such as the length of the conversation and the fluency of the Speakers.

In the following sections, we will go into details about how the participants were chosen, the materials used in the experiment and how the agent was created, the session protocol/script followed by the experimenter during the experiment, and the specific types of data that were collected and how they measured rapport.

4.1. Participants

There were 36 people recruited in this study, 27 females and 9 males. They were all undergraduates from the National University of Singapore, receiving course credit for participating in the study.

When a pilot study was run earlier, it was found that pre-existing friendships (or in one case, a romantic relationship) had a positive influence on the level of rapport reported by the participants in the experiment. It also reduced the believability of the avatar, as participants were familiar with the ways in which their friends responded, which did not always correspond to the feedback they were seeing from the avatar. Therefore, in order to eliminate the impact of any pre-existing friendships on the rapport built during conversation, the 36 participants were screened beforehand and paired to people whom they said were strangers, or at most casual acquaintances, to them.

4.2. Materials

For this experiment, the Speaker watched a video prior to the storytelling session. This video was a short 3½ min segment of a funny Tom & Jerry cartoon. There were altogether two videos used in the experiment – one for each participant in the dyad – both of equal duration and belonging to the same series.

For the storytelling session, two workstations were set up in adjoining rooms, each with a computer monitor display and a webcam. (Fig. 1) The Speaker's workstation (seen below on the left) showed a picture of the avatar and had headphones attached with in-built microphone. The Listener's workstation (seen below on the right) showed the feed from the webcam at the Speaker's station and had a standalone microphone and audio speakers. On the Speaker's side, although the headphones provided the illusion that the Speaker would be able to hear the Listener, in actuality they were not connected to the Listener's microphone and thus the Speaker was insulated from any sounds or verbal feedback that the Listener could have made. This was done to reduce the possibility of the Speaker detecting incongruities between the laughs they heard from the Listener, and the facial expressions being shown by the avatar.



Figure 1. Experimental Setup

This experimental setup is slightly different from the one that Gratch et. al. used for the Rapport Agent – the Rapport Agent studies placed the participants in the same room, separated by a screen. This allowed them to hear each other talk and laugh, which would not have affected their understanding of the Rapport Agent’s behaviour, as the Rapport Agent was limited to eye gaze and gross head movements (nods, tilts and shakes) only, thus not showing any facial expressions. However, since a laugh can be shown on the face quite clearly, the Speakers would expect to see a laugh appearing on the avatar’s face the moment they heard it from the Listener. Thus, in order to maintain the illusion that the agent was accurately representing the Listener’s facial expressions, the Speaker had to be isolated from hearing the Listener’s laughing, and therefore they were placed into separate rooms, with the Speaker wearing headphones that helped muffle any noise. This helped to focus the Speakers’ attention on the avatar’s facial expressions, as that was the only form of feedback and communication they had with the other party.

The avatar used in this study was originally developed by the Smartbody project (Thiebaux, Marshall, Marsella, & Kallmann, 2008). It was then modified for this study to be able to display two long-lasting moods as ‘idle’ animations, and several instantaneous emotional expressions, following the Facial Action Coding System for facial muscles developed by Ekman et. al. (Ekman, 2007; Ekman & Friesen, 1978) and exaggerated slightly for effect. Each expression of emotion was based upon the research done by Ekman and colleagues on facial expressions and the muscles they trigger. The emotions shown by the avatar trigger the same facial muscles, but had their intensity increased in

order to be more visible. The tables below show the facial expressions used as well as the Action Units (AU) that indicate which facial muscles were activated in the avatar. Each AU number corresponds to a particular facial muscle, and the letters indicate the intensity of the effect, from A (trace) to E (maximum):











Mood: Smiling	Mood: Frowning	Expression: Surprise	Expression: Smile	Expression: Sadness
				
AU: 6B, 12B	AU: 4E, 7D, 15E, 39E	AU: 1E, 2E, 5E, 26C, 27A	AU: 6E, 12C	AU: 1E, 4E, 10C, 15E, 17E
Expression: Laugh	Expression: Puzzlement	Expression: Anger	Expression: Fear	Expression: Disgust
				
AU: 6E, 10E, 12C, 23C, 26B. (Also, head tilted back)	AU: 1E (left only), 2E (left only), 4E, 7D, 23E	AU: 4E, 7C, 9E, 10E, 15E, 23C, 24E, 39E, 45A	AU: 1E, 2E, 4E, 5E, 7D, 15E, 20C, 26C	AU: 4E, 7E, 9B, 15E

Table 1. FACS Action Units used to create facial expressions

Participants were randomly assigned to one of two agent conditions: *appropriate* or *inappropriate*. In the *appropriate* condition, the agent attempted to show *positivity* and *coordination*, two of the correlates of rapport identified by Tickle & Degnen. To show *positivity*, the idle expression on the agent's face was set to a slight smile (Mood: Smiling). To show *coordination*, the agent displayed emotional reactions in sync with the Speaker's expressions while telling the story. Since the story being told was a funny cartoon, the expected facial expressions were those of good humour and possibly surprise, and the agent attempted to mirror the expressions displayed by the Speaker. (i.e. if the Speaker laughed, the agent responded with a laugh). In addition to this, the agent also detected natural backchannel opportunities – such as the ends of phrases or sentences – and responded with either a smile or a surprised expression, as appropriate to the story.

In the *inappropriate* condition, the agent attempted to show *negativity* and *un-coordination*. Firstly, to show *negativity*, the idle expression of the agent was a slight frown, instead of a smile expected of a listener to funny story. Then, in order to show uncoordinated responses, the agent either did not display a reaction when expected, or displayed emotional feedback that was incongruous with the story being told. When the Speaker displayed a facial expression, or when the agent detected a backchannel opportunity, the agent had a 50% chance of not displaying any reaction at all, and another 50% chance of displaying an incongruous reaction. The expressions used for incongruous reactions were sadness and puzzlement.

Due to the difficulty of achieving these responses algorithmically in real time, the agent's responses were controlled through a Wizard-of-Oz setup, where there was a wizard in another room observing the Speaker, and then pressing various buttons to generate pre-programmed emotional expressions in the avatar. Results from a study by Jonsdottir et. al. (Jonsdottir, Gratch, Fast, & Thorisson, 2007) indicate that the appropriate length of a response cycle between human participants for real-time backchannel feedback is between 560-2000 msec, approximately 1 second on average. The wizard was trained to respond as quickly as possible to achieve a similar rapidity of response. To reduce the experimental variability caused by a human, the wizard strictly followed the behavioural rules guiding agent behaviour – positive or negative idle expressions for the different conditions, mirroring of

Speaker's expressions for the *appropriate* condition, and showing either congruous, incongruous, or no expressions during backchannel opportunities.

Lastly, a note needs to be said about envelope feedback in this experiment. In this study, it is assumed that a smile and a look of puzzlement were facial equivalents to the nods and headshakes used in the Rapport Agent envelope feedback experiments, indicating acknowledgement or not. As such, in both conditions there was a high level of envelope feedback that was held constant, so that it could be treated as a neutral factor. Both the appropriate and inappropriate agent responded in a timely manner to opportunities for envelope feedback, and thus the differences lie not in the presence or absence of envelope feedback, but in the different types of emotions shown when there *is* feedback.

4.3. Session Protocol

This study was designed to make the participants think they were interacting with a human. Dyadic pairs of participants were introduced to each other at the beginning of the session. They were then told a cover story that they were helping to evaluate an advanced telecommunications system, specifically a computer program that accurately captures all the movements and facial expressions of one person and displays them on screen (using an avatar) to another person. According to the cover story, the study was interested in comparing this system against a more traditional telecommunications medium, such as video conferencing, which is why one of the participants would be seated in front of a monitor displaying a live feed from videoconferencing software, while the other saw a life-size head of an avatar.

After the briefing, participants were shown the two different workstations displaying the videoconferencing software and the avatar, and also watched a short video introducing the range of emotional expressions that the avatar could make to familiarise them with the avatar, prior to engaging in the experiment proper.

The participants were then split up into a designated Speaker and Listener. Each participant went through the experiment twice, once as a Speaker, once as a Listener. The order was chosen randomly. While the Listeners waited at the workstation showing the videoconferencing system, the Speakers

were taken to another room and asked to watch the first Tom & Jerry video. They were told that they would later be describing the story to the Listener, and that the system would be evaluated by the Listener's story comprehension. This was to encourage them to speak more, in order to tell a full and complete story, and provide sufficient material for analysis.

After watching the video, the Speaker was taken to an empty room with a single workstation and monitor which displayed the avatar. They were told that the avatar accurately represents the facial expressions of the Listener seated in the other room. The Speaker was then instructed to put on the headphones with attached microphone, and informed that they would be recorded via webcam and microphone. The Listener was seated in an adjoining room in front of the computer monitor that showed the feed from the Speaker's webcam, and was able to hear the Speaker through the system. The wizard was also sharing the same room as the Listener, hidden from the Listener's sight by a screen (Fig. 2). The wizard could thus see and hear what the Listener saw and heard through the video feed from the webcam, and could give instructions to the avatar, unbeknownst to the Listener. The Listener was told that the wizard would be video recording and taking notes of the story as it was being told. This accounted for any actions of the wizard overheard by the Listener, and none of the participants suspected deception due to the wizard's activities.

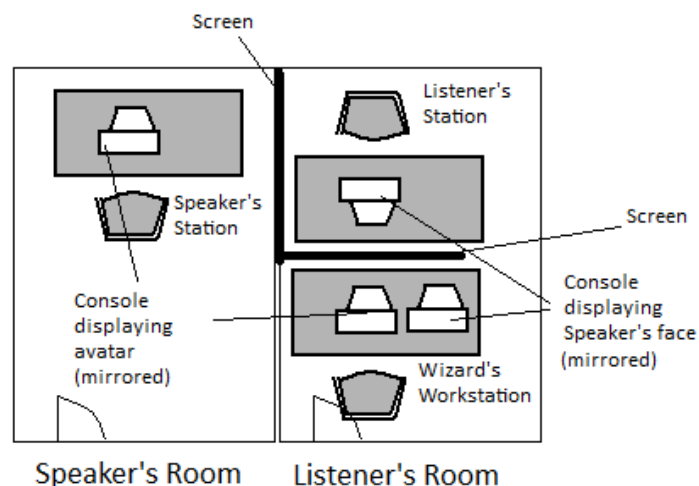


Figure 2. Lab Layout

After finishing the story, the Speaker was then instructed to fill out a post-experiment questionnaire, following which the Speaker and Listener changed roles and the procedure repeated with a different video clip of similar length from the same series. At the end of both stories, the participants were debriefed together and told about the deception and the real purpose of the experiment.

4.4. Data Gathering

At the beginning of the experimental session, basic demographic data was collected from the participants – age, gender and ethnicity. Ethnicity was particularly important as we wanted to be aware of any possible cross-cultural misunderstandings of facial expressions.

For behavioural measures, and to answer research questions 1 & 2, the Speaker was recorded via the webcam and microphone. Data collected can be grouped into three categories:

1. *Duration of interaction*: This includes total time to tell the story, the number of words in the story, and the number of “meaningful” (lexical and functional) words in the story.
2. *Speech fluency*: Two groups of measures are used – speech rate and amount of disfluencies (incomplete words and pausefillers like “um”). For speech rate, both overall speech rate as well as fluent speech rate (lexical and functional words per second) was measured. Likewise, for disfluencies, both disfluency rate (number of disfluencies per second) and disfluency ratio (number of disfluencies against overall word count) were measured.
3. *Self-report measures*: A post-experiment questionnaire was filled out by the Speaker, measuring their self-reported feelings for rapport. It consisted of 5 forced-choice items:
 - Do you think he/she understood the story completely?
 - Did you feel you had a connection with the other person?
 - While you were telling the story, was seeing the avatar helpful or disturbing?
 - Did you feel you used some feedback from the other person while telling the story?
 - Do you think the avatar portrayed the movements of another person accurately?

5. Results

Results showed a marked difference in the length of story and the self-reported feelings of rapport. The primary interesting result was that participants exposed to an *inappropriate* condition spoke much longer than those who were faced with *appropriate* facial expressions, which was surprisingly opposite of the results from studies on envelope feedback. Other results showed that the differences in speech fluency were nearly negligible, and of the five measures for rapport questionnaire, more participants reported positive indications of rapport for the appropriate condition than for the inappropriate condition.

One of the experimental results was contaminated due to the Speaker being able to hear the Listener laughing loudly from another room while being shown a frowning face by the avatar, and so had to be discarded. This left a sample size of 35 participants, 18 in the *appropriate* condition and 17 in the *inappropriate* condition.

Similar to the Gratch et. al study, because of the small sample size, non-parametric statistics were used to analyse the results: Mann-Whitney U for scale variables (length of interaction, speech fluency), and Chi-square for nominal variables (forced-choice questionnaire items). Table 2 shows a summary of these results:

Variable	Appropriate ^a	Inappropriate ^a	Mann-Whitney U	Significance ^b
Total time	67	108	103.0	0.099 *
Word count	176	253	121.5	0.298
Fluent word count	162	228	122.5	0.314
Speech rate	2.60	2.61	139.5	0.656
Fluent speech rate	2.32	2.36	138.0	0.620
Disfluency rate	0.20	0.22	143.0	0.741
Disfluency ratio	0.08	0.10	132.5	0.496

^a – median used as a measure of central tendency

^b – 2-tailed criterion

* - $p < 0.05$

Table 2. Length of Interaction & Speech Fluency Results

Only one of the results for the behavioural measures achieved statistical significance – the total time taken to tell the story. For time taken, the mean ranks were 15.22 for the appropriate condition, and 20.94 for the inappropriate condition. The word counts, while not significant, also showed greater numbers in the *inappropriate* condition than the *appropriate* condition.

For the five forced-item measures of rapport in the questionnaire, the graphs below show the results of the people who selected each option under different conditions, expressed as a percentage of the total number of people in that condition. The p-values for Chi-square tests are also shown on the graph. The ones which achieved significance of $p < .05$ are bolded. While only the first two of the five results achieved significance, some overall trends are worth reporting.

More people felt they were understood when faced with appropriate facial expressions than inappropriate ones:

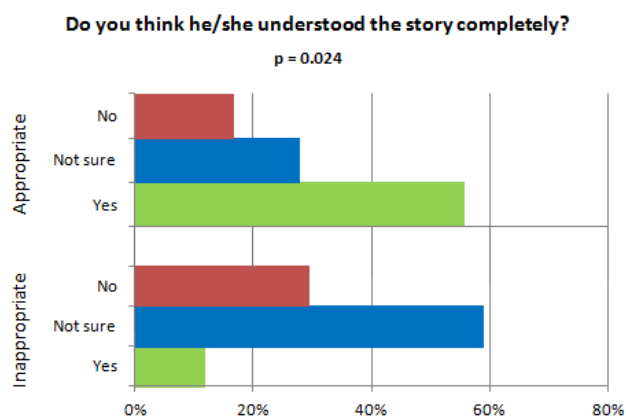


Figure 3. Speaker’s evaluation of avatar’s understanding of story

No one felt like they had a connection with the other person in the inappropriate condition, compared to about one-third in the appropriate condition who did feel some connection:

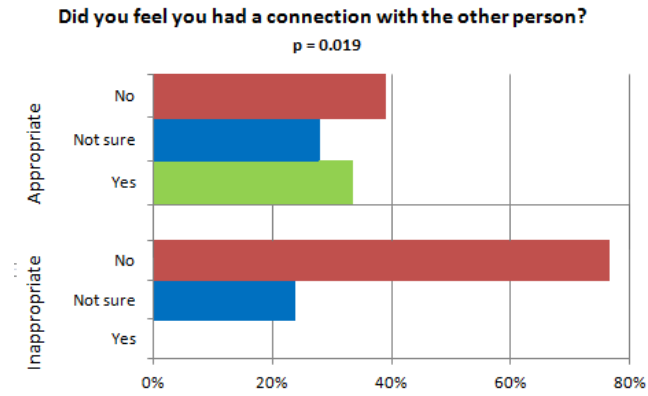


Figure 4. Speaker's evaluation of connection with other person through avatar

More people thought seeing the avatar was helpful in the appropriate condition than in the inappropriate condition:

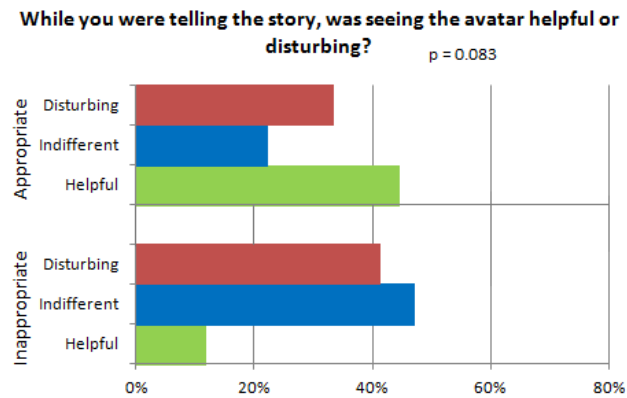


Figure 5. Speaker's evaluation of avatar's helpfulness when seen

More people felt that they used some feedback from the avatar in the appropriate condition than in the inappropriate condition:

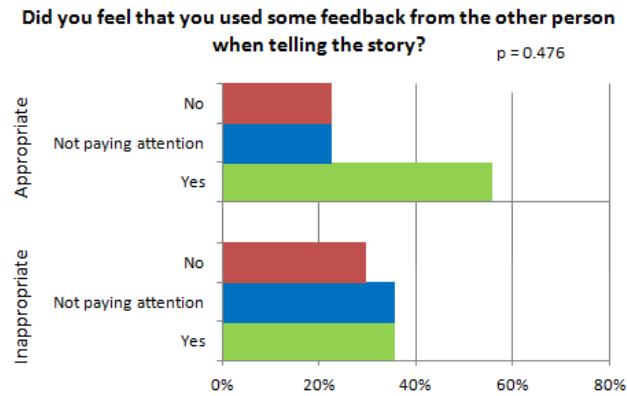


Figure 6. Speaker’s usage of feedback from avatar while telling story

And lastly, most were unsure about the accuracy of the avatar’s portrayal of the other person:

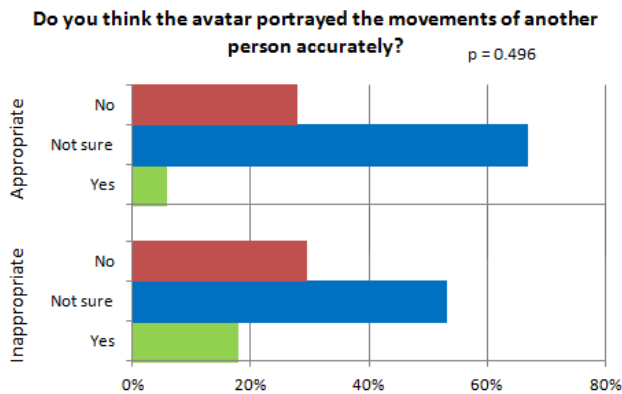


Figure 7. Speaker’s evaluation of accuracy of avatar’s portrayal of Listener

Overall, for the self-reported questionnaire, the first two results were statistically significant with $p < 0.05$, and the third result for a $p < 0.10$. Thus, many more participants felt that the other person (via the avatar) had understood them in the appropriate condition than in the inappropriate condition, and most people did not feel a connection with the avatar in the inappropriate condition, while results were mixed in the appropriate condition. Thirdly, more people rated the seeing the avatar as “Helpful” in the appropriate condition than in the inappropriate condition, although this result was not as strongly supported statistically as the first two. The final two factors – the usage of feedback from

avatar mentioned above and the accuracy of the avatar's portrayal of the other person – did not achieve statistical significance, but the overall trends warrant further study on the effects of facial expressions on attention, and work done to improve the believability of the avatars' reactions.

6. Discussion

The results above raise several points for discussion. For the different behavioural factors measured (Table 2), only the measurement of time taken to tell the story reached statistical significance.

However, looking at the overall trend for story length both in absolute time and word count, this indicates that the people who are faced with inappropriate facial expressions are taking *more* time and words to tell the story than people who are faced with appropriate facial expressions, which is different from the studies on envelope feedback. Reasons for this could include conversational grounding problems, social anxiety, or language barriers.

This chapter is split into three sections, each dealing with one of the factors above. In the first section, we go into the definition of conversational grounding, and the implications it has for the results of this experiment. In the second section, an alternative explanation is proposed in the form of social anxiety, which will also be described in detail. Lastly, in the third section a note is made of possible language barriers that may have influenced the results in comparison with those from Gratch's study of envelope feedback.

6.1. Conversational Grounding

One way to explain these results is to look at the theories of *conversational grounding* (Clark & Brennan, 1991). In order for collective actions (like conversation) to take place, the participants must assume they have a vast collection of mutual beliefs and information that are held in common. This is known as *common ground*. It is difficult to proceed in a conversation if one person understands the word "mouse" to mean a rodent while the other views it as a computer peripheral device. In order for the conversation to make sense, both participants need to coordinate the content of the conversation. *Grounding* is the process by which participants in a conversation establish and update common ground, and it happens constantly throughout the conversation.

In grounding, participants try to make sure that what has been said is what has been understood – that is to say, to make it part of their common ground. Grounding occurs in two phases – a presentation

phase, and an acceptance phase. During the presentation phase, the speaker produces an utterance for the listener to consider. In the acceptance phase, the listener provides evidence that he or she believes that they have understood the utterance, with the assumption that once the speaker receives that evidence, the speaker will also believe that the listener has understood. While that is the simple form, in practice Clark and Brennan note that it can get very complicated, with incomplete utterances, interruptions and self-repairs going on constantly.

Grounding is most evident in the acceptance phase, when listeners try to give evidence that they believe they understand the utterance. Speakers can look for *negative evidence* – evidence that they have been misunderstood or misheard. If they find negative evidence, the speaker engages in conversational repair, otherwise by default, they assume they have been understood and proceed on. However, ultimately, Speakers would prefer *positive evidence* – evidence that they have been understood. At its most basic, this takes the form of continued attention (through eye gaze, or other ways of indicating one is paying attention), but can also include acknowledgements (“uh huh”, “mm”, head nods, smiles), and initiation of the relevant next turn (e.g. answering the question posed, replying with a relevant response, etc.).

As noted earlier, many of the Speakers who took part seemed hesitant in their speech, speaking in rapid bursts and then pausing for confirmation. This can be explained by the grounding process. The feedback given by the agent in the *appropriate* versus *inappropriate* states is very different when viewed through the process of grounding. In the *appropriate* condition, the agent would smile, laugh or look (pleasantly) surprised when backchannel opportunities came up. In all cases, this was a form of *positive feedback*. The contingency of the responses provided adequate evidence of continued attention, a smile could be taken as an acknowledgement of the speaker’s utterance, and the laugh or look of surprise could be both acknowledgement as well as initiation of the next conversational turn – equivalent to saying “That’s funny”, or “Really? Wow, I didn’t expect that.” Furthermore, in the *appropriate* condition, there was no *negative* evidence to show that the agent did not understand the Speaker’s utterances.

However, this was not the case in the *inappropriate* condition. In the inappropriate condition, the agent would only respond half the time when presented with backchannel opportunities. This could be taken by the Speaker as a sign of inattention, and thus a lack of positive evidence that the agent was sharing the same common ground. Furthermore, during the times that the agent *did* respond, it was with either a sad expression or a puzzled one. Both were negative evidences of misunderstanding in different ways. A puzzled expression was a clear signal that the agent did not understand what the Speaker was saying. A sad expression indicated that the agent believed sadness was the appropriate response to make to the Speaker's utterance, which again could be taken as a misunderstanding of the common ground.

In their study of coordination and rapport, Cassell, Gill and Tepper (Justine Cassell, et al., 2007) note that among strangers, there was a much greater frequency of grounding exchanges used than among friends. The ones speaking would gaze at the receivers more frequently, to signal the need for feedback, and would carry on after they have received some sort of acknowledgement. Since this study was also conducted between strangers, it would be expected that there would be frequent pauses for grounding exchanges. However, due to the difference in the type of feedback received in the appropriate versus inappropriate conditions, it is likely that Speakers would wait longer in the inappropriate condition for any sort of positive feedback, and would also stop more often to ground the conversation. That could be another explanation why the length of the conversation in the inappropriate condition was longer than in the appropriate condition.

It is important to note that this result was caused by a difference in the methodology between Gratch's study and mine. In the study of the Rapport Agent, for the inappropriate/unresponsive agent condition, Gratch et. al. played a video showing the agent responding with postural shifts and head turns, based on the agent's responses to someone else's story, but *excluding head nods and shakes*. Therefore, the unresponsive Rapport Agent lacked positive feedback in grounded communications, but did not provide *negative feedback*. Whereas in this study, due to the nature of emotional feedback in addressing the content of the communication, negative feedback was present, and hence this would have increased the perception of misunderstanding between Speaker and agent. Thus, there was

greater hesitancy and many more efforts at conversational repair with agent showing inappropriate emotional feedback than the agent with appropriate emotional feedback.

6.2. Social Anxiety

Another possible explanation for this result is the social anxiety. Social anxiety has been defined as a condition in which “some people, especially those who are shy or easily embarrassed, feel anxious in almost any situation in which they might be evaluated.” (Myers, 1999) The American Psychiatric Association (DSM-IV, 1994) also define it similarly: “Social anxiety disorders or social phobias are characterized by intense and personal fear of social performance situations in which embarrassment may occur, typically fear of public speaking and/or situations where interactions with others may occur.”

As the Speakers were clearly told that they would be recorded while they re-told the story of a video they only saw once to a stranger, it is quite possible that the experimental setup may have been ripe for social phobias to be triggered. While not explicitly recorded, several participants did indicate that they were feeling anxious about the storytelling task – one or two even going so far as to request repeated viewing of the video before they were willing to go ahead and describe it to another person. Observations of the participants as they told stories indicated that participants in the inappropriate condition tended to speak jerkily – with rushed sentences followed by longer pauses – compared to those in the appropriate condition. This could have been an indication of the level of nervousness and embarrassment they felt.

It was shown in previous research that virtual agents could also cause anxiety in human users (Rickenberg & Reeves, 2000), and that an audience with hostile/bored facial expressions caused greater levels of anxiety than one showing appreciative expressions (Pertaub & Slater, 2001). More specifically, Kang and colleagues (Sin-Hwa Kang, Jonathan Gratch, Ning Wang, & James H. Watt, 2008) demonstrated that high levels of shyness/social anxiety felt by Speakers led to higher evaluations of embarrassment and lowered self-performance as well as lowered indications of rapport with the Rapport Agent in an unresponsive condition. Given that the emotional impact of negative

facial expressions in this study is likely to be even greater than the emotional impact of non-contingent head and posture shifts in the Rapport Agent, it is possible that the embarrassment felt by participants in this study would have been higher.

This could also have led to a shift in conversational mode by the Speakers, from relationship-oriented conversation to task-oriented conversation. Casual comments made by participants after the experiment indicated that they had stopped paying attention when the avatar started showing them negative faces, and concentrated on just telling the story instead. This may also be a contributing factor to the length of story seen – the negative facial expressions may have changed the mindset of the participants from *trying to relate* to the virtual character to focusing on *performing the task* of storytelling instead. This could be a quality particular to emotional feedback, rather than envelope feedback, since displaying emotions implies a value-judgment being made about the person's contribution to the conversation. Many people may not wish to constantly face the negative emotions they saw on the avatar's face, so focused on communicating the memory of what they saw instead. This was also borne out by the results in Fig. 6, which showed greater numbers of people who didn't pay attention to the avatar's feedback in the inappropriate condition than in the appropriate condition, which *wasn't* seen in the envelope feedback experiments.

All these and more imply that the social anxiety of the individual participants may have been a factor that would have impacted story length and rapport. However, as this experiment did not control for social anxiety, further research would need to be done to disentangle this as a possible cause for the story length result.

6.3. Language Barriers

A third explanation for the difference in story length is that there could have been a language barrier in operation here, which may have had opposite effects in the two conditions. The study was done in a university in Singapore, a multiracial, multilingual society which also serves as a transit hub for many international students. While ethnicity was investigated as a factor, there seemed to be no discernible pattern in terms of perception or interpretation of facial expressions. But what did have an impact was

the participants' level of familiarity with English. A few were clearly translating from their native tongue to English, indicated by occasional usage of foreign words, followed by English equivalents. The lowered familiarity with English would tend to cut short unnecessary words in the story (thus explaining the relative brevity of the appropriate condition), but if they detect that the Listener is not understanding them due to the facial expressions, they would instead spend more effort and words to try and make themselves understood, which could account for much longer times in the inappropriate condition.

This also had implications for speech fluency. It was difficult to distinguish between disfluencies caused by the emotional feedback and disfluencies caused by the lack of familiarity with the English language. The results for fluency (both speech rate and disfluency measures) were nearly equal and did not reach statistical significance. While not all the participants suffered from a lack of familiarity with English, there could have been enough in the small sample size of this study to make an impact on the results.

7. Limitations

While this study does contribute to our understanding of the difference between emotional and envelope feedback in rapport, there are a number of limitations to this study which must be addressed.

Firstly, it should be noted that the experiment design ended up being more about the measurement of rapport between humans mediated by a virtual character rather than the rapport between human and agents directly. This is due to the fact that participants were led to believe that the avatar represented the Listener's facial expressions, rather than an intelligent system designed to interact with them. This may have led to source orientation problems (Sundar & Nass, 2000), where it was difficult to determine whether the source of the low rapport scores in the inappropriate condition was due to participants believing they had poor rapport with the other person behind the agent, or because they felt the agent did not accurately represent the other person. As the results in Fig. 6 showed that most people (in both conditions) were unsure whether the agent accurately represented the other person or not, it casts doubt on the accuracy of the self-reported measures of rapport found here. This could also be the reason why there were generally less statistically-significant results.

However, as this experimental design was based on the first study of Gratch et. al's Rapport Agent, the comparison with the results of that study can be taken to hold this factor as constant. Thus, though both of our studies showed the effects of rapport between humans mediated through an avatar, the difference between envelope and emotional feedback still applies. Inappropriate envelope feedback caused the story length to decrease, while inappropriate emotional feedback caused the story length to *increase*.

Secondly, it can be argued that the measurement of rapport in the self-reported questionnaire doesn't correspond to any known scale that has been properly verified. Although the questions were the same as Gratch's study, it is a weakness of that design as well. This is acknowledged as a weakness, which is why the findings of the self-report questionnaire have not been discussed in detail, nor any conclusions drawn from it. Instead, the results focused on the behavioural measures. At best, it is legitimate to say that more people *appear* to find appropriate facial expressions helpful in conveying

understanding and facilitating a connection with the other person than inappropriate expressions, but more robust measurements are necessary to prove the point.

Thirdly, the recognition of the facial expressions by the Speaker can be called into question. It is possible that the expressions shown by the avatar may not be associated with the right emotions in the Speaker's mind, or the intensity of the facial expressions may have had an effect on the Speaker's perception. It is true that the study design did not explicitly test the participants on their recognition of the various facial expressions shown. However, the facial expressions used by the avatar were largely based on the Universal Facial Expressions that Ekman and colleagues had discovered across all cultures, so presumably are generalizable to any human. Furthermore, in pilot tests done prior to the study, the testers were able to discriminate between the different facial expressions. However, the effect of facial expression intensity, while important, would lie outside the scope of this study as this study is concerned more with broad strokes of positive or negative expressions rather than finer details on varying levels of intensity. Future work would need to be done to examine the effects of emotional intensity.

Lastly, there was the possibility that the study design would have been better if the experiment was intra-subject – that is, if participants went through both conditions of appropriate and inappropriate emotional feedback. While that is a valid approach, it would have introduced further complications in terms of order effects, and would also have diverged further from the methodology that Gratch et. al used in the study of the Rapport Agent and envelope feedback. As one of the goals of this study was to deliberately compare the results of envelope versus emotional feedback, it was decided that the same methodology should be adhered to as closely as possible (allowing for differences due to the nature of the feedback itself, of course). Thus, the participants, as in the Rapport Agent studies, only experienced one condition. However, a follow-up study may be done to confirm the results of this study through intra-subject experiments.

8. Conclusion

To sum up, this study was aimed at finding out how emotional feedback, in the form of facial expressions, would have an effect on rapport. To do this, an agent was simulated that would show both appropriate and inappropriate emotional feedback to a human user telling a story to it, and measured the rapport felt by the human user through measuring the *duration* of the story being told, the *fluency* of the speaker as the story was being told, and the *feelings of rapport* reported by the speaker afterwards. Although the results for fluency were not statistically-significant and the feelings of rapport were mixed, the most significant finding was that inappropriate facial expressions caused the *duration* of the story told by the human to increase, which was the opposite of studies done with agents displaying envelope feedback. This thus indicates that rapport *is* affected by emotional feedback, and in a way that is different from envelope feedback.

This thesis then argued that possible causes for the difference in behaviour could arise from the differences between emotional and envelope feedback – specifically, that emotional feedback (through facial expressions) has a greater impact on conversational grounding than envelope feedback does, and that it can engender greater feelings of social anxiety – because emotional feedback conveys a value judgement on the content of the person’s contribution to the conversation, and not just the process of communication. However, do note that the results may be partially explained by language barriers, and thus further research is needed into these factors to confirm the initial findings documented here.

Furthermore, while this study is necessarily limited to facial expressions in a retelling of a funny cartoon, it does not encompass the whole of emotional feedback. An earlier investigation (unpublished) indicated that facial expressions were actually considered the least emotionally-expressive out of the three options of facial expressions, vocal tone, and word choices. This was partially due to the subtlety and instantaneity of facial expressions, which were often harder to catch and interpret than more obvious emotional shadings of words and tone. Therefore, more research

would be needed to investigate the full range of modalities for emotional feedback – particularly speech prosody (tone of voice) and linguistic choices for verbal feedback.

Also, the range of emotions studied so far has been limited to a small subset of pleasant, humorous interactions involving mostly monologue storytelling. There is still a lot of work remaining to design agents capable of handling scenarios which are structurally different – dialogues, group conversations, public speaking or formal counselling situations, to name a few. There is also need to investigate emotional ranges that may not necessarily as pleasant, such as anger or grief, but which may prove more beneficial in the long run as conversational agents can be developed to build rapport with humans in socio-emotional situations where users may be uncomfortable talking to other humans.

In conclusion, the field of conversational agent design is still a vast and complex field as many researchers attempt to implement ways of making agents that can substitute for humans in conversational interactions. This study provides some insight into the nature of the relationship between emotional feedback and the establishment of rapport in human-agent interactions, but there is much more work to be done before we will be able to reach the goal of an agent who can be as satisfying a conversational partner as a human being would be.

Bibliography

1. Bailenson, J. N., & Yee, N. (2005). Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Psychological Science, 16*, 814-819.
2. Bartneck, C. (2002). *Integrating the OCC Model of Emotions in Embodied Characters*. Paper presented at the Workshop on Virtual Conversational Characters: Applications, Methods and Research Challenges.
3. Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as Co-Narrators. *Journal of Personality and Social Psychology, 79*(6), 941-952.
4. Beskow, J., & McGlashan, S. (1997). *Olga - A Conversational Agent with Gestures*. Paper presented at the IJCAI Workshop on Animated Interface Agents.
5. Bevecqua, E., Mancini, M., Niewiadomski, R., & Pelachaud, C. (2007). *An expressive ECA showing complex emotions*. Paper presented at the AISB'07 Convention workshop: "Language, Speech and Gesture for Expressive Characters".
6. Bickmore, T., & Cassell, J. (2001). *Relational Agents: A Model and Implementation of Building User Trust*. Paper presented at the CHI.
7. Bickmore, T., & Gruber, A. (2010). Relational Agents in Clinical Psychiatry. *Harvard Review of Psychiatry, special issue on Psychiatry and Cyberspace, 18*(2), 119-130.
8. Bickmore, T., & Mauer, D. (2006). *Modalities for Building Relationships with Handheld Computer Agents*. Paper presented at the Conference on Human Factors in Computing Systems (CHI), Montreal.
9. Bickmore, T., Mauer, D., Crespo, F., & Brown, T. (2008). *Negotiating Task Interruptions with Virtual Agents for Health Behavior Change*. Paper presented at the Autonomous Agents and Multi-Agent Systems (AAMAS).
10. Bickmore, T., & Pfeifer, L. (2008). *Relational Agents For Antipsychotic Medication Adherence*. Paper presented at the CHI'08 Workshop on Technology in Mental Health, Florence, Italy.

11. Bickmore, T., Pfeifer, L., & Jack, B. W. (2009). *Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents*. Paper presented at the Human Factors in Computing Systems (CHI), Boston, MA.
12. Bickmore, T., & Picard, R. (2004). *Toward Caring Machines*. Paper presented at the CHI 2004.
13. Bickmore, T., & Picard, R. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interactions*, 12(2), 293-327.
14. Bickmore, T., & Schulman, D. (2009). *A Virtual Laboratory for Studying Long-Term Relationships between Humans and Virtual Agents*. Paper presented at the Autonomous Agents and Multi-Agent Systems (AAMAS), Budapest, Hungary.
15. Bickmore, T., Schulman, D., & Yin, L. (2009). *Engagement vs. Deceit: Virtual Humans with Human Autobiographies*. Paper presented at the Intelligent Virtual Agents.
16. Brave, S., Nass, C., & Hutchinson, K. (2005a). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62, 161-178.
17. Brave, S., Nass, C., & Hutchinson, K. (2005b). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2), 161-178.
18. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsen, H., et al. (1998, Oct 12-15). *An Architecture for Embodied Conversational Characters*. Paper presented at the First Workshop on Embodied Conversational Characters, Tahoe City, California.
19. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsen, H., et al. (1999). *Embodiment in Conversational Interfaces: REA*. Paper presented at the CHI.
20. Cassell, J., Gill, A. J., & Tepper, P. A. (2007). *Coordination in Conversation and Rapport*. Paper presented at the Workshop on Embodied Natural Language, Prague.
21. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., et al. (1994). *Animated conversation: rule-based generation of facial expression, gesture & spoken*

- intonation for multiple conversational agents*. Paper presented at the Proceedings of the 21st annual conference on Computer graphics and interactive techniques.
22. Cassell, J., & Thorisson, K. R. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, 13, 519-538.
 23. Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. B. R. J. M. Levine & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington D.C.: American Psychological Association.
 24. DSM-IV (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.
 25. Duncan, S., & Welji, H. (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 Meeting of the Cognitive Science Society*.
 26. Ekman, P. (2007). The directed facial action task. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment*. Oxford; New York: Oxford University Press.
 27. Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System*. Palo Alto, California: Consulting Psychologists Press, Inc.
 28. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Werf, R. J. v. d., et al. (2006). *Virtual Rapport*. Paper presented at the 6th International Conference on Intelligent Virtual Agents.
 29. Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). *Creating Rapport with Virtual Agents*. Paper presented at the 7th International Conference on Intelligent Virtual Agents.
 30. Jonsdottir, G. R., Gratch, J., Fast, E., & Thorisson, K. R. (2007). Fluid Semantic Back-Channel Feedback in Dialogue: Challenges and Progress *Intelligent Virtual Agents*. Paris, France: Springer.

31. Kang, S.-H., Gratch, J., Wang, N., & Watt, J. H. (2008). *Does the Contingency of Agents' Nonverbal Feedback Affect Users' Social Anxiety*. Paper presented at the 7th International Joint Conference on Autonomous Agents and Intelligent Systems.
32. Kang, S.-H., Gratch, J., Wang, N., & Watt, J. H. (2008). *Does the contingency of agents' nonverbal feedback affect users' social anxiety?* Paper presented at the Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1.
33. Kidd, C. D. (2007). *Engagement in Long-Term Human-Robot Interaction*. Cambridge, MA: MIT Press.
34. Klein, J., Moon, Y., & Picard, R. (2002). This Computer Responds to User Frustration: Theory, Design, Results and Implications. *Interacting with Computers, 14*, 119-140.
35. Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., & Rosin, P. L. (2009). Effects of Dynamic Attributes of Smiles in Human and Synthetic Faces : A Simulated Job Interview Setting. *Journal of Nonverbal Communication, 33*, 1-15.
36. Liu, K. K., & Picard, R. (2005). *Embedded Empathy in Continuous, Interactive Health Assessment*. Paper presented at the CHI Workshop on HCI Challenges in Health Assessment, Portland, Oregon.
37. Myers, D. (1999). *Social Psychology*: McGraw-Hill College.
38. Ochs, M., & Prendinger, H. (2010). *A Virtual Character's Emotional Persuasiveness*. Paper presented at the KEER2010, Kansai Engineering and Emotion Research.
39. Park, R. E., & Burgess, E. W. (1924). Introduction to the Science of Sociology
40. Pereira, A., Leite, I., Mascarenhas, S., Martinho, C., & Paiva, A. (2010). *Using Empathy to Improve Human-Robot Relationships*. Paper presented at the 3rd International Conference on Human-Robot Personal Relationships, Leiden, Netherlands.
41. Pertaub, D.-P., & Slater, M. (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.
42. Picard, R. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

43. Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge University Press.
44. Rickenberg, R., & Reeves, B. (2000). *The effects of animated characters on anxiety, task performance, and evaluations of user interfaces*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems.
45. Sundar, S., & Nass, C. (2000). Source Orientation in Human-Computer Interaction. *Communication Research*, 27(6), 683-703.
46. Thiebaux, M., Marshall, A. N., Marsella, S., & Kallmann, M. (2008). *SmartBody: Behavior Realization for Embodied Conversational Agents*. Paper presented at the Autonomous Agents and Multi-Agent Systems (AAMAS), Estoril, Portugal.
47. Tickle-Degnen, L., & Rosenthal, R. (1990). The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, 1(4), 285-293.
48. Wang, N., & Gratch, J. (2010). *Don't just stare at me!* Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems.
49. Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communications between man and machine. *Communications of the ACM*, 9(1), 36-45.