Parasocial Consensus Sampling: Modeling Human Nonverbal Behaviors from Multiple Perspectives

by

Lixing Huang

Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

Janurary 2013

Dedication

To my beloved parents Xiaoyang Huang and Caixia Lu...

Acknowledgements

About four and half years ago, the airplane I took landed in LAX. On my very first night in Los Angeles, I slept in a rented car, outside the old ICT building. I thought about a lot of things, a little worried... I had completely no idea how I could finish my PhD program and how long it would take. Now, I am about to finish my dissertation, trying to recall *the* "big thing" that magically helped me find the path. But all I can remember are a lot of generous help, great advice, and the hard working on many small things...

First of all, I want to thank my PhD advisor, Professor Jonathan Gratch. Without his support and mentoring, this dissertation would not have been possible. I really appreciate he gives me the freedom of choosing the research topic, yet maintains a high standard of requirement. I will never forget the time when he sat down beside me and taught me how to look for the facts hidden behind those plain numbers. My research interest may change in the future, but the skills he taught me will always be helpful. I will always remember the very first paper he, Professor Louis-Philippe Morency and me wrote together, which is now the foundation of my dissertation. That was really fun and surprising experience. I enjoy the years working with him, because he is not only a very good academic advisor but also a nice friend. Thank you, Jon.

I also want to thank the members of my qualifying and dissertation committees: Professor Louis-Philippe Morency, Professor Stacy Marsella, Professor Shri Narayanan, and Professor Gérard G. Medioni. They provided invaluable feedback on my research work and always gave me insightful comments in a very constructive way. Especially, I want to thank Professor Louis-Philippe Morency, who is always patient, very passionate,

and pursuing perfection. He spent a lot of time advising me in the first two years of my PhD program, generously sharing his experiences with me and supporting my research.

I owe my deepest gratitude to Lila Brooke, Edward Fast, Alesia Egan, Jamison Moore, and Jill Boberg. They helped create the awesome software system, high quality dataset, and comfortable working environment. It would be extremely hard to perform my research work without the support from all these talented colleagues.

Last but not least, I am fortune enough to have many great fellow students, especially the ICT PhD student group. My personal life would not have so much fun without the friendships I've made during my days in USC.

After I complete my PhD research work, I realize how little this piece of work contributes to the human knowledge and how much hard work it is required to perform original research. I admire and respect the researchers who are always curious about the unknown world and working hard to explore it bravely.

Table of Contents

List Of Tables

List Of Figures

Abstract

Virtual humans are embodied software agents designed to simulate the appearance and social behaviors of humans, typically with the goal of facilitating natural interactions between humans and computers. They play an important role in the advancement of today's immersive virtual worlds, including domains such as virtual training (Swartout et al., 2006), education (Rowe et al., 2010), and health care (Bickmore et al., 2010).

One of the key challenges in creating virtual humans is giving them human-like nonverbal behaviors. There has been extensive research on analyzing and modeling human nonverbal behaviors. Some of them rely on results from observing and manually analyzing human behaviors, while others approach the problem by exploring advanced machine learning techniques on large amounts of annotated human behavior data. However, little attention has been paid to the "data" these systems learn from.

In this thesis, we propose a new methodology called *Parasocial Consensus Sampling (PCS)* to approach the problem of modeling human nonverbal behaviors from the "data" perspective. It is based on previous research on *Paraosical Interaction* theory (Horton & Wohl, 1956). The basic idea of Parasocial Consensus Sampling is to have multiple independent participants experience the same social situation parasocially (i.e. act "as if" they were in a real dyadic interaction) in order to gain insight into the typicality of how individuals would behave within face-to-face interactions.

First, we validate this framework by applying it to model listener backchannel feedback and turn-taking behavior. The results demonstrate that (1) people are able to provide valid behavioral data in parasocial interaction, (2) PCS data generates better

virtual human behaviors and (3) can be used to learn better prediction models for virtual human. Second, we show that the PCS framework can help us tease apart the causalities of the variability of human behavior in face-to-face interactions. Such research work would be difficult to perform by traditional approaches. Moreover, PCS enables much larger scale and more efficient data collection method than traditional face-to-face interaction. Finally, we integrate the PCS-data driven models into a virtual human system, and compare it with a state-of-the-art virtual human application, the Rapport Agent (Gratch et al., 2007) in real interactions. Human subjects are asked to evaluate the performance of each agent regarding the correctness of the agents' behaviors, the rapport they feel during the interactions and the overall naturalness. The results suggest that the new agent predicts the timing of backchannel feedback and end-of-turn more precisely, performs more natural behaviors and thereby creates much stronger feeling of rapport between users and agents.

Chapter 1

Introduction

1.1 Motivation

Virtual humans are embodied software agents designed to simulate the appearance and social behaviors of humans, typically with the goal of facilitating natural interactions between humans and computers. They play an important role in the advancement of today's immersive virtual worlds, including domains such as training (Swartout et al., 2006), education (Rowe et al., 2010), and health care (Bickmore et al., 2010).

One of the key challenges in creating virtual humans is giving them human-like nonverbal behaviors. In face-to-face interactions, our actions often speak louder than our words (Mehrabian, 2007) (Burgoon et al., 2010). A speaker's nonverbal cues - like facial expression, gestures, and posture - can dictate the meaning of an utterance and add a richer dimension of understanding. But such cues are not simply modifiers to speech, they also have interactional functions. For example, even silent listeners communicate a wealth of information back to the speaker through their facial expressions, gestures and posture. This feedback can be *generic*, meaning it conveys a general stance toward the speaker (e.g., nodding signals "go on") or *specific*, which is tied to a deeper understanding of, and reaction to, the personal relevance of what the speaker is saying (Bavelas et al., 2000) (Bavelas & Gerwing, 2011). Either way, such feedback helps co-construct how an interaction unfolds. Indeed, the presence and synchrony of such nonverbal cues

in face-to-face interactions is an important predictor of how successful the interaction will be (Tickle-Degnen & Rosenthal, 1990).

If a virtual human is capable of recognizing (Morency et al., 2007), understanding (Ang et al., 2002) and exploiting (Cassell et al., 2000) conversational nonverbal behaviors appropriately, it can evoke a range of social behaviors in users which is typically only seen in face-to-face interactions with other humans (Bailenson & Yee, 2005) (de Melo et al., 2009) (Bickmore et al., 2010) (Gratch et al., 2006) (Gratch et al., 2007) (Wang & Gratch, 2010) (Von der Putten et al., 2009). Thus, considerable effort has been expended toward the goal of building better nonverbal behavioral models for virtual humans (Kipp et al., 2007) (Morency et al., 2008) (Lee & Marsella, 2009) (Lee & Marsella, 2006) (Cassell et al., 2001) (de Kok & Heylen, 2009).

Modeling nonverbal behavior is a hard problem. First, human behavior contains variability and nonverbal displays are affected by many internal factors, such as emotion, personality, physiology and interactional goals, and external factors, such as the presence of others and others' responses. Second, these factors interact both within and across participants - for example the emotions of one participant in a conversation can spill over and alter the behavior of other actors - making it difficult to isolate and model these sources of variability. Third, nonverbal behavior is subtle and dynamic, such that similar displays can convey very different meanings. For example, minor variations in the performance of a smile can alter perceptions of trust (Krumhuber et al., 2007). Unfortunately, this "micro structure" is rarely studied and poorly understood. Together, these characteristics make modeling nonverbal behavior challenging.

Methodologies for studying nonverbal behavior have evolved dramatically over the years. Early seminal, and still influential, research relied on observations of natural behavior "in the field" (Smith et al., 2010) (McDougall, 2003) (Darwin et al., 2002); whereas more recent work systematically recorded and annotated behaviors in laboratory settings designed to approximate real-world interactions (Kendon, 2004) (McNeill, 1992) (Duncan, 1972). This latter work often involves elaborate annotation schemes

(e.g., the facial action coding system (Ekman & Friesen, 1977)) that require considerable effort to train and execute. Such methods enable the recording of relatively large amounts of human behavior data and facilitate micro-analysis (e.g. frame-by-frame), via manual annotation, with relative ease. Such analyses usually yield descriptive rules, which are more helpful as general theoretical constraints but not as helpful to inform the moment-to-moment generation of virtual agent behavior.

Very recently, there has been an explosion of interest in using automated techniques such as signal processing and machine learning to both acquire and analyze large datasets of nonverbal behavior (Morency et al., 2008) (Lee & Marsella, 2009) (Kipp et al., 2007) (Jonsdottir et al., 2008) (de Kok & Heylen, 2009). The advantages of such approaches are multifold. For example, by dispensing with the need for manual analysis of data, they enable the examination of much larger datasets than would be possible by manual techniques. Also, machine learning techniques usually generate quantitative models which can be implemented in programming languages and directly deployed in virtual human systems. Moreover, one can easily build separate models for different contexts through individual alternations to a dataset. Unfortunately, the promise of such techniques remains unfulfilled except in the case of fairly simple behaviors (e.g., smiles and nods) and most contemporary "automated" techniques still require extensive manual annotation and analysis.

1.2 Open Challenges

The primary contribution of this thesis is methodological: to address current challenges in collecting and annotating large datasets of face-to-face nonverbal behavior. Although previous research mostly focused on the techniques for learning from data, there has been less attention to the data these systems learn from. Virtually all research on virtual human nonverbal behavior is based on passive observations of natural interactions (e.g., recording two participants talking), however there are several reasons why such data is problematic for learning general behavior models. First, there is considerable

variability in human behavior. Any given conversation may contain many behaviors that are idiosyncratic, atypical, or worse, negative examples of the concept we wish to learn. For example, if the goal is to learn to produce the feeling of rapport, it is important to realize that many face-to-face interactions fail in this regard. Ideally, observed behaviors would be separated into good and bad instances of the target behavior, but it is not always obvious how to make this separation and even manual human coders may disagree on what constitutes a good behavior. Second, face-to-face interactions are "co-constructed" in that moment-to-moment behaviors depend not only on characteristics of the individual, but on their contingent reactions to the behavior of the other party in the interaction. For example, even in a monolog, a speaker will often attend to the reactions of his listeners and adjust his behavior accordingly (Bavelas et al., 2000). This mutually-contingent nature of social interactions amplifies the underlying variability of human behavior but also makes it difficult to tease apart the eliciting conditions for any given behavior (i.e., is this person a non-engaging speaker, or is he reacting to a disengaged audience). These drawbacks are not insurmountable but they imply that we will have to collect large amounts of behavioral data to surmount them, which brings us to the third drawback of data-driven approaches: the traditional way of recording face-to-face interaction data is very expensive and time-consuming. It usually takes months to recruit pairs of participants, followed by an extensive period of manually-annotating the resulting recordings.

In this thesis, we propose a new methodology, inspired by the concept of the "wisdom of crowds" (Surowiecki, 2004), to address these challenges. The intuition behind the "wisdom of crowds" is that, under the right circumstances (e.g. individuals make independent decisions and there is a great diversity among them), the aggregation of opinions from many individuals, although with varied quality, approximates the objective truth. This concept has been applied to problems in computer vision (Sorokin & Forsyth, 2008) and natural language processing (Snow et al., 2008), where many non-experts are recruited to perform trivial tasks, such as image labeling. Although

individually they are not as good as experts (e.g. they provide noisy labels), the aggregated version removes the noise and works as well as the data provided by experts.

Before applying the same concept to human behavior data, it is important to note that the characteristics of nonverbal behavior mentioned above make it fundamentally different from other problems utilizing the "wisdom of crowds" approach, such as image labeling. First, human behavior contains more variability and there is unlikely to be a single "objective truth" as to what constitutes correct nonverbal behavior. At any moment, there are always different ways to behave. For example, by showing a funny picture to different people, some may laugh and some may not. Both reactions are probably reasonable. It is only possible to tell what is the most likely behavior under certain circumstance rather than what is the correct behavior. Second, human behavior is contextual and contingent. For example, many of the nonverbal behaviors seen in a conversation are contingent reactions to the behavior of the other party (e.g. smile when smiled at). Therefore, it is difficult to tease apart if a given behavior was driven by properties of the individual (e.g., the speaker is in a positive mood) or of the immediate social context (e.g., he is speaking to a receptive audience). Because of these fundamental differences, it is not clear how to apply the concept of the "wisdom of crowds" to human behavior.

One way to solve the problems is to somehow break down the contingency in face-to-face interactions. For example, to determine if a smile is an appropriate reaction to a speaker's utterance, we would ideally observe a crowd of listeners reacting to the same utterance (e.g., we could observe faces on the Washington Mall reacting to the President's inaugural speech). In such situations we could simply tally the distribution of smiles or frowns in the audience at each moment. Such an approach would be adequate if we merely wish to simply study an audience's passive reactions to a speaker, but our goal here is more general. We wish to inform the design of interactive systems where multiple participants jointly create the social interaction. In such cases, each conversation, even if on the same topic, will be unique and co-constructed by both

speakers. We will never be able to observe the identical utterance produced in the same context as it will, at some level, depend on the history of what has come before, and each data point will in some sense be idiosyncratic. But what if we could somehow do a "what if" analysis where we could insert a crowd of participants into each moment in the history of an unfolding conversation and observe their immediate reactions. If possible, much as with the audience, we could calculate the probability of each possible behavior, and further, identify characteristics of individuals that tend to elicit these behaviors (e.g., most people interrupt at a given point but not if they are introverted). The main contribution of this thesis is a methodological approach that enables the collection and analysis of this sort of "what if" data.

1.3 Approach and Objectives

In our work, we propose a general data collection framework, called *Parasocial Consensus Sampling (PCS)*, which successfully applies the concept of the "wisdom of crowds" to human behavior modeling problem. Parasocial Consensus Sampling is based on the theory of *parasocial interaction* introduced by Horton and Wohl (Horton & Wohl, 1956), in which they argued that people exhibit a natural tendency to interact naturally with pre-recorded videos of people as if they were interacting with the actual person face-to-face. A classic example would be nodding or yelling at a presidential candidate during a televised debate. Reeves and Nass's *Media Equation Theory* (Reeves & Nass, 1996) also suggests that social responses are automatically elicited by any cues that are related to human characteristics, such as the appearance of a human face on a screen. Therefore, by asking people to pretend to interact with media representation of social interactions (e.g. pre-recorded videos), it may be possible to elicit behaviors that are consistent with those elicited in natural interactions. The basic idea of Parasocial Consensus Sampling is to have multiple independent participants experience the same social situation parasocially (i.e. act "as if" they were in a real dyadic interaction) in order to

gain insight into the typicality (i.e. consensus view) of how individuals behave within face-to-face interactions.

The research addresses three questions:

1. *Can we rely on Parasocial Consensus Sampling framework to collect valid human behavior data, and use it to generate better virtual human behavior and learn better nonverbal behavior prediction models?*

2. *Can we take advantage of Parasocial Consensus Sampling framework to understand the causalities of the variability of human nonverbal behaviors in face-to-face interactions?*

3. *Can we apply the PCS data to advance the development of virtual human systems?*

The first question is addressed through applying the framework to model conversational listening behaviors. Specifically, we apply PCS to collect listener backchannel feedback and turn-taking behaviors. Participants are first informed of the interactional goal (e.g. show interest in the speaker while listening or taking appropriate conversational turns) and the target behaviors (e.g. backchannel feedback, such as nodding or brief sound like "uh-hum", or take a turn appropriately), and then they interact with pre-recorded videos parasocially. Their parasocial responses are used to drive the virtual human behavior directly and to learn prediction models. We then run both subjective and objective experiments to evaluate the performance of the virtual human and the prediction models.

PCS framework allows multiple independent participants to experience the same social situation (that is, one side of the interaction keeps consistent) parasocially; therefore, the variability of their nonverbal behaviors can be attributed to the individual differences among these participants. To investigate the second question, we focus on whether and how personality traits influence nonverbal behaviors, such as head nod, headshake, and smile. Specifically, we group participants based on their personality traits, and compare the consensus behavior of these groups.

The third question is the ultimate goal of our research work. Compared to traditional approaches, PCS helps collect better human behavior data in a more effective way. Ideally, these data should help build better nonverbal behavior models and thus improve the performance of the current virtual human systems. To answer the third question, we integrate the PCS-data driven models into a virtual human system, and compared it with a state-of-the-art virtual human application, Rapport Agent (Gratch et al., 2006), in real interactions. Human subjects are asked to evaluate the agents subjectively regarding the correctness of the agents' behaviors, the rapport they feel during the interaction and the overall naturalness. Following Bavelas' terminology (Bavelas et al., 2000) of generic versus specific feedback, the behavior models considered in this thesis will focus on generic feedback (meaning that the results do not require the virtual human fully comprehend the content of the human speech). Obviously, specific feedback is more challenging and demands advanced natural language techniques. We will revisit the question of specific feedback towards the end of the thesis.

1.4 Major Contributions

The dissertation's contributions are to:

- Propose a novel methodology, called Parasocial Consensus Sampling (PCS), for human nonverbal behavior data collection, which has already inspired some of the work in virtual human research area (de Kok & Heylen, 2011) (de Kok et al., 2010) (Ozkan & Morency, 2011);
- Demonstrate the validity of the PCS framework in modeling human nonverbal behaviors. Specifically, we apply the PCS framework to model *generic* listener feedback;
- Investigate the causalities of the variability of human nonverbal behaviors in face-to-face interactions by taking advantage of PCS. Such research work would be difficult to perform by traditional approaches;

- Use the methodology to dramatically improve the behavior of an important virtual human application, the Rapport Agent (a system that has been used by multiple international groups to study the social effects of virtual human on people).

1.5 Outline

The remainder of the dissertation is organized as the following:

1. Chapter 2 reviews related work.
2. Chapter 3 describes the *Parasocial Consensus Sampling* framework.
3. Chapter 4 and 5 validate the PCS framework by applying it to model backchannel feedback in face-to-face conversation and turn-taking behavior in the interview setting.
4. Chapter 6 and 7 explore and extend the PCS framework, where we investigate how personality traits influence listener backchannel feedback in face-to-face interaction and how the measurement channel (an important element in PCS which will be described in Chapter 3) affects the PCS data.
5. Chapter 8 describes the development of a new version of virtual human that integrates the backchannel model and turn-taking model built from the PCS data and compares the performance of the new virtual human with the old one, Rapport Agent.
6. Chapter 9 discusses conclusions and proposes future work.

Chapter 2

Related Work

2.1 Nonverbal Behavior in Face-to-Face Interaction

Nonverbal behavior plays a critical role in successful face-to-face interactions. Nonverbal signals encode propositional content and help to regulate the give and take of conversations. Indeed, in many contexts, more than half of the information exchanged in an interaction is encoded in the nonverbal channel (Burgoon & Hoobler, 1994) (Mehrabian, 1981). Thus, nonverbal fluency is an essential skill for any entity, human or virtual, that aspires to successfully understand or participate in face-to-face interactions.

Nonverbal behavior's propositional function is to convey the actual thoughts or internal states. When speaking, we make gestures, accompanying the speech, to compensate or enhance the verbal content. For example, we sometimes use hands to measure the size of a symbolic space while saying "it was this big"; we use head movements to signal "yes or no", to make superlative or intensified expression (e.g. "very very old") or to express the concepts of inclusivity such as "everyone" or "everything" (Heylen, 2005) (Kendon, 2002) (McClave, 2000). Nonverbal behavior is also considered as the overt reaction of our internal states. For example, there is a widely accepted assumption that emotion and facial expression are closely connected. Emotion states can lead to specific facial expressions (e.g. sadness is expressed by raising inner brow and/or pulling brow

together). Besides emotion, nonverbal behavior has also been argued to express personality to some extent for a long time, although with mixed results (Duncan & Fiske, 1985). (Gifford, 1994) applied the lens model to study the encoding and decoding of interpersonal dispositions in nonverbal behavior and demonstrated that the encoding of some interpersonal dispositions, such as dominant and submissive, in nonverbal behavior is moderate.

Nonverbal behavior also serves interactional functions such as providing rapportful responses and regulating conversation turns. When engaged in a fluent conversation, we feel "in sync", or rapport, with the conversational partner. The feeling of rapport is argued to underlie many desirable social effects (Drolet & Morris, 2000) (Burns, 1984) (Tatar, 1997) (Bernieri & Rosenthal, 1991). Tickle-Degnen and Rosenthal (Tickle-Degnen & Rosenthal, 1990) suggested a three-factor theory of rapport and theoretically linked nonverbal behaviors to the three factors, which are mutual attentiveness (e.g. mutual gaze), positivity (e.g. smile and head nod) and coordination (e.g. postural mirroring and interactional synchrony). Not only the forms, the timing of nonverbal behavior also shows coordination between conversational partners. For example, Brunner (Brunner, 1979) observed that listeners time their smiles to the speakers so that these smiles serve as backchannel responses (Yngve, 1970). (Bavelas et al., 2002) showed that the speaker usually seeks a response at some key points by looking at the listener, creating a short period of mutual gaze, in which the listener inserts brief responses such as "mhm" or head nod. Without such precisely timed responses, (Bavelas & Gerwing, 2011) found the speakers were unable to perform their tasks well.

In face-to-face interaction, the roles of speaker and listener are regulated seamlessly by a negotiation process of turn-taking. Much empirical evidence showed that nonverbal behaviors, such as prosody, eye-gaze and head gestures, play an important role in regulating turns. In the auditory channel, (Duncan, 1972) found that a rising or falling pitch at the end of a sentence serves as a turn-yielding signal; similarly, (Beattie, 1982) claimed that a falling intonation pattern at the end of a clause always indicates the

end of a turn. In the visual channel, gaze has been identified as a useful mechanism to coordinate turns in conversation. When one conversant finishes his current turn, he will often look towards the other; later, the established mutual gaze between them will be broken by the other conversant when she begins to talk. This is called "mutual gaze break" (Novick et al., 1996). It was found that looking-towards ("gaze-back") and looking-away ("gaze aversion") are correlated to end of turn and beginning of turn, respectively. This pattern occurs at approximately 42% of the turn changes in (Novick et al., 1996). Head movements sometimes accompany the change of eye gaze and acts as turn-yielding signals as well (Duncan, 1974). Speaker often make tiny head nods or shakes at the end of turn to elicit confirmation from the listener (Heylen, 2006), or to signal to the listener the turn-taking channel is open again.

When serving interactional functions, nonverbal feedback can be classified into two classes (Bavelas et al., 2000). One is generic feedback, which is not closely connected to what is being said. Such generic behaviors don't convey any specific meanings, and would be appropriate in different scenarios. The other one is specific feedback, which is tied to a deeper understanding of, and reaction to, the personal relevance of what is being said. Such specific behaviors usually depend not only on the understanding of the semantic meanings but also on our own role and participatory goals, which may change as the conversation unfolds.

2.2 Nonverbal Behavior of Virtual Humans

Given the important function nonverbal behaviors serve, it is not surprising that considerable efforts have been directed at endowing virtual human with such ability to give appropriate nonverbal behaviors. Similar to face-to-face interaction, virtual human's nonverbal behavior also serves propositional and interactional functions.

When serving propositional functions, nonverbal behavior either compensates or synchronizes with the agent's verbal content or reflects its internal state. For example, BEAT (Cassell et al., 2001) is a nonverbal behavior generation engine, which extracts

linguistic and contextual information, such as clause boundary and rheme/theme, from input text and suggests appropriate hand gestures, eye gaze and other nonverbal behaviors. NVBG (Lee & Marsella, 2006) is another nonverbal behavior generator which generates nonverbal behaviors such as head movements and hand gestures according to the communicative functions inferred from a surface text analysis. (Strauss & Kipp, 2008) presented an affective embodied agent for real-time commentary. It is able to produce coherent natural language and nonverbal behaviors to reflect its emotions, mood and personality.

When serving interactional functions, nonverbal behavior is usually used to regulate conversations or provide contingent feedback to the human partner. Previous work mostly focused on enabling virtual humans to provide generic feedback. For example, REA (Cassell, 2000) is such a human-like agent who can use nonverbal behaviors to regulate conversations. For example, she uses glance around to break away from the conversation; she raises hands into the gesture space as a signal for turn-taking. Gandalf agent (Thorisson, 2002) was able to decide when to take a turn according to its conversational partner's behaviors. It is a layered architecture with several update loops, each designed for a different sensor (e.g. speech detector, gaze detector, etc.) operating at a different speed. A multimodal turn-taking model combines those unimodal features based on heuristic rules to make turn-taking decisions. Rapport Agent (Gratch et al., 2007) was a listening agent who provides contingent nonverbal generic feedback (e.g. head nod or posture shift) in real-time based on the human speaker's behavior. Recently, (Wang et al., 2011) addressed the problem of modeling specific listener feedback and described the Listener Feedback Model for virtual agents which generates both generic and specific feedback conditioned on a variety of factors, including the speaker's behavior, the listener's role, the listener's desire to participate in the conversation and the unfolding comprehension of partial utterances.

It has been demonstrated that with appropriate nonverbal behaviors, virtual humans can better interact with human partners and thus produce more desirable social effects

than those without appropriate nonverbal behaviors. For example, (Bailenson & Yee, 2005) showed that by mimicking a participant's head movements at a 4-s delay, the virtual agent is perceived as more persuasive. (de Melo et al., 2009) demonstrated that humans are more cooperated with the agent who displays moral facial expressions in a social dilemma game. (Gratch et al., 2007) (Gratch et al., 2006) found that the virtual human with contingent positive feedback could induce the subjective feeling and many of the behavioral benefits of the psychological concept of rapport, which was not seen in cases when no such contingent feedback occurs.

2.3 Modeling Nonverbal Behavior for Virtual Humans

The problem of how to build nonverbal behavior models for virtual human has been investigated for a long time. Originally, researchers depend on the findings from the social psychology literature. For example, (Pelachaud, 1996) (Cassell et al., 1994) developed virtual agent systems using descriptive rules like "speaker looks away from the hearer at the beginning of a long turn and looks towards the hearer at the end of the turn" to coordinate the conversation. NVBG (Lee & Marsella, 2006) summarized various findings from the literature to implement the mappings between nonverbal behaviors and communication functions. (Gratch et al., 2007) also derived a set of rules, such as mimicking the human speaker's nod, from the literature to drive the Rapport Agent to provide contingent feedback. However, such descriptive rules from the social psychology literature are more helpful as general theoretical points than to directly drive a virtual agent's behavior as they typically describe general findings and do not precisely characterize the specific circumstance and timing information for when such behaviors should be employed. Furthermore, although very instructive for describing typical human face-to-face interactions, they are not necessarily derived from the same context in which the virtual agent will be applied.

Recently, researchers start to explore more advanced machine learning techniques to learn behavior models from large amounts of annotated human behavior data. For

Figure 2.1: Anvil User Interface

instances, (Nishimura et al., 2007) proposed a unimodal decision tree approach for producing backchannels based on prosodic features, the system analyzes speech in 100ms intervals and generates backchannels as well as other paralinguistic cues (e.g. turn taking) based on pitch and power contours. (Lee & Marsella, 2009) developed a HMM model using the AMI data set (Carletta, 2007) to predict the timing of the speaker's head nods. They predicted whether to nod or not based on a set of linguistic features extracted from the speaker's verbal content. Evaluation results showed that their HMM model performed better than the previous rule-based one. (Morency et al., 2008) presented a discriminative probabilistic framework to predict listener backchannel feedback based on the conversational partner's behavior. They developed an automatic feature selection strategy and trained a Latent Dynamic Conditional Random Field (LDCRF) using multimodal features (lexical words, prosodic features and eye gaze) to learn the dynamic structure of the interaction. (Jonsdottir et al., 2008) built a reinforcement learning model that learned the optimal pause duration by relying on pitch slope and pitch value, which could take turns with human-like speed and reliability. Although previous research mostly focused on the learning techniques, there has been less attention to the data these systems learn from.

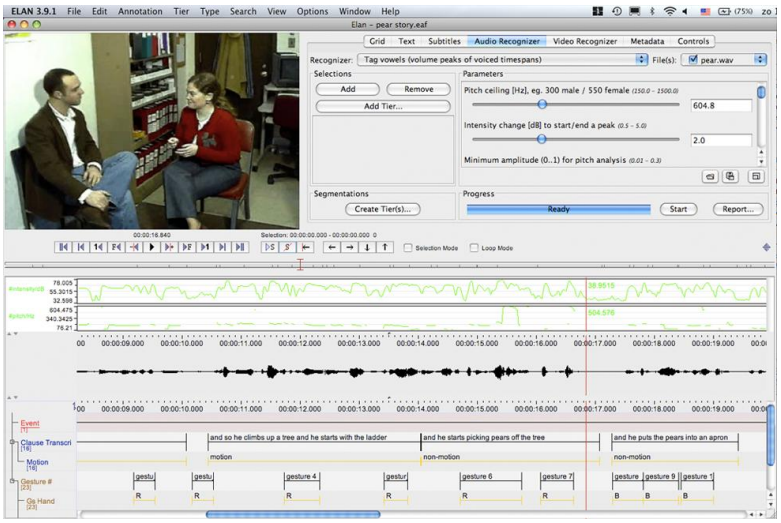

Figure 2.2: Elan User Interface

2.4 Human Behavior Dataset

The construction of annotated human behavior dataset is a very time-consuming and expensive process. Usually, the dataset consists of many video recordings of human interactions. Coders are hired to watch those videos frame by frame and label specific human behaviors manually by using software such as ANVIL (Figure 2.1) or ELAN (Figure 2.2). However, human nonverbal behavior is subtle and ambiguous, which makes the task even harder, and sometimes, introduces the problem of inter-coder reliability. (Afzal & Robinson, 2009) showed that the inter-coder reliability for facial expression is relatively low, and suggested that the low reliability, rather than being an error of measure, is in fact an acknowledged observation reported for natural human behavior data (Abrilian et al., 2005) (Cowie et al., 2005).

Besides hiring coders to label data manually, there are some efforts in applying machine learning techniques to automatically label human behavior. For example, (Littlewort et al., 2011) presented a software tool for automatic facial expression recognition, which can code the intensity of 19 different facial actions from the Facial Action Unit Coding system (FACS) and 6 different prototypical facial expressions. Although it

shows promising results, there are still challenges, such as relatively low accuracy and generality, in replacing manual labor with such techniques.

Recently, there is a new trend to explore the concept of the "wisdom of crowds" for data annotation (Sorokin & Forsyth, 2008) (Snow et al., 2008). The idea is to have multiple independent coders make judgments on the same object (e.g. an image), then the aggregation of those opinions, although with varied quality, approximates the objective truth. However, the dataset most of the work in this area focused on is fundamentally different from the human interactional data (e.g. face-to-face interaction) we are interested in. First, human behavior contains more variability. At any moment, there are always different ways to behave, such that it is only possible to tell what is the most likely behavior under certain circumstance rather than what is the correct behavior. Second, human behavior is contingent. In face-to-face interaction, the speaker's behavior not only affects but also is affected by the listener's behavior at the same time. Because of these differences, researchers need come up with special experiment designs in order to apply the concept of the "wisdom of crowds" on human interactional data. An example is the Parallel Listener Corpus created by (de Kok & Heylen, 2011). In their experiment, a speaker talked to one listener (i.e. the real listener) via a camera and they both could see each other on the screen; whereas, at the same time, two other listeners (i.e. concealed listeners), sitting separately, watched the speaker's behavior as well and were convinced that each of them was actually the only listener (as shown in Figure 2.3). In this way, they collected listener feedback from three independent listeners who "interacted" with the *same* speaker simultaneously. By aggregating the responses from the three listeners, they built a consensus view of how an average listener would respond and thus learnt a better backchannel prediction model (de Kok et al., 2010). Though partially solving the problem of collecting multiple perspectives on a single interaction (i.e. it obtains multiple views of a speaker but a single view of each listener), such a setup is quite elaborate and limited to contexts where one participant cannot speak. In

Figure 2.3: An example of the Parallel Listener Corpus. The top-left one is the speaker, the bottom-right one is the real listener, and the other two are the concealed listeners.

next chapter, I will show how Parasocial Consensus Sampling provides a more flexible and efficient approach for collecting human behavior data from multiple perspectives.

2.5 Conclusion

Nonverbal behaviors have important functions in face-to-face interaction as well as in human-agent interaction. Not surprisingly, lots of efforts have been devoted to model human nonverbal behaviors and apply such models to build virtual human systems. Although previous research mostly focused on the techniques for learning from data, there has been less attention to the data these systems learn from. In our work, we propose a new methodology for human behavior data collection, which takes advantage of the concept of the "wisdom of crowds" to facilitate the modeling of nonverbal behaviors.

Chapter 3

Parasocial Consensus Sampling Framework

3.1 Background

Parasocial Consensus Sampling is inspired by the research work on *parasocial interaction* (Horton & Wohl, 1956). They observed the interactions between "media users" and "media figures" (e.g. presenters, actors and celebrities) and found such interaction can yield a form of parasocial relationship, where the users respond as if they were in a social relationship with the media figures.

Subsequent to Horton and Wohl, this phenomenon has been explored by many researchers in media research and communication research. It has been observed that people exhibit a natural tendency to interact with media representation of people as if they were interacting with the actual person face-to-face. For example, (Masters & Sullivan, 1990) showed that people do react with corresponding emotions to televised emotional expressions of political leaders; (Sundar & Nass, 2000) observed that individuals apply politeness norms, often only seen in social interactions, to computers; (Levy, 1979) also described that people behave as if they were having a two-way conversation with a television news anchorperson while watching the person on TV. Cognitively, (Bargh, 1988) argued that humans are fundamentally using the same cognitive processes in both interpersonal and mediated communication. This was later confirmed in Perse

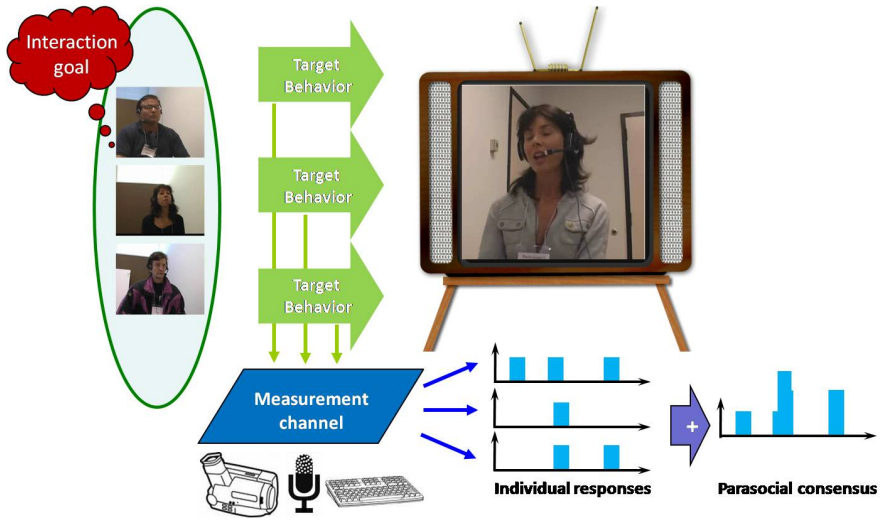

Figure 3.1: Parasocial Consensus Sampling framework. Given some interactional goal, the participants are guided to interact with media representation of people (e.g. video clips) parasocially. Their behaviors are measured via some measurement channel and later combined into a consensus view.

and Rubin's study (Perse & Rubin, 1989), where they tested two interpersonal cognitive frameworks to both social and parasocial relationships. The results highlighted the validity of applying interpersonal cognitive frameworks to the context of mediated communication. The similarity between responses in both social and parasocial interactions can also be viewed as an example of the Media Equation theory (Reeves & Nass, 1996). They argued that social responses can be automatically elicited by any cues related to human characteristics, such as the appearance of a human face. Recently, (Hartmann & Goldhoorn, 2010) analyzed the causes of a parasocial experience in TV viewers. They suggested the addressing styles and perceived attractiveness of the media figure and the viewer's perspective-taking ability are three factors influencing the intense of parasocial experiences and invented the Experience of Parasocial Interaction scale (ESPI) to measure the parasocial experience.

3.2 Parasocial Consensus Sampling

Parasocial Consensus Sampling (PCS) is a novel methodological approach to eliciting information about the typicality of human responses in social interactions. Unlike traditional virtual human design, where interaction behaviors are deduced by observing how individuals respond in a social situation, PCS allows multiple individuals to vicariously experience the same social situation to gain insight on the typicality (i.e., consensus view) of how individuals would behave within face-to-face interactions. By eliciting multiple perspectives, this approach can help tease apart what is idiosyncratic from what is essential and help reveal the strength of cues that elicit social responses.

The idea of parasocial *consensus* is to combine multiple parasocial responses to the same media clip in order to develop a composite view of how a typical individual would respond. For example, if a significant portion of individuals smile at a certain point in a videotaped speech, we might naturally conclude that smiling is a typical response to whatever is occurring in the media at that moment. More formally, a parasocial consensus is drawing agreement from the feedback of multiple independent participants when they experience the same mediated representation of an interaction. The parasocial consensus does not reflect the behavior of any one individual but can be seen more as a prototypical or summary trend over some population of individuals which, advantageously, allows us to derive both the strength and reliability of the response.

Although we can never know how everyone would respond to a given situation, *sampling* is a way to estimate the consensus by randomly selecting individuals from some population. Thus, parasocial consensus sampling is a way to estimate the consensus behavioral response in face-to-face interactions by recording the parasocial responses of multiple individuals to the same media (i.e., by replacing one partner in a pre-recorded interaction with multiple vicarious partners). By repeating this process over a corpus of face-to-face interaction data we can augment the traditional databases used in learning virtual human interactional behaviors with estimates of the strength and reliability of

such responses and, hopefully, learn more reliable and effective behavioral mappings to drive the behavior of virtual humans.

Formally, we define Parasocial Consensus Sampling as a data collection protocol involving five elements (as shown in Figure 3.1):

- *An interactional goal*: this is the intended goal of the virtual human interactional behaviors. For example, (Gratch et al., 2007) created an agent that conveys a sense of rapport and engagement. Participants in parasocial consensus sampling should be implicitly or explicitly encouraged to behave in a manner consistent with this goal (e.g., if the goal is to increase rapport, participants could be instructed to respond as though they are interested in the pre-recorded speaker).

- *A target behavioral response*: this is the particular response or set of responses that we wish our virtual human to generate. For example, if we are trying to create a virtual human that knows when to interrupt his conversational partner, participants should be encouraged to produce such behavior. Candidate behavioral responses include backchannel feedback, turn taking or evaluative facial expressions etc.

- *Media*: this is the set of stimuli that will be presented to participants in order to stimulate their parasocial responses. Ideally this would be a media clip derived from a natural face-to-face interaction where the participants can view the clip from a first-person perspective and have the feeling of being directly addressed. For example, if the original interaction was a face-to-face conversation across a table, the camera position should approximate as close as possible the perspective of one of the conversational partners.

- *A target population*: this is the population of individuals we wish our virtual human to approximate. This might consist of members selected from some particular group (e.g., women, speakers of African American vernacular, or patients with clinical depression). Participants should be recruited from this target population.

- *A measurement channel*: this is the mechanism by which we measure the parasocial response. The most natural way to measure the response would be to encourage participants to behave as if they were in a face-to-face interaction and record their normal responses. However, a powerful advantage of imaginary nature of parasocial interactions is that participants might be encouraged to elicit responses in a more easily measured fashion. For example, if we are interested in the consensus for when to smile in an interaction, we can ask participants to exaggerate the behavior or even press a button whenever they feel the behavior is appropriate. Candidate measurement channels include the visual channel (e.g. videotaping), audio channel (e.g. voice recording) or mechanical channel (e.g. keyboard response).

Given these components, PCS proceeds as follows: for each parasocial stimuli of interest, draw multiple participants from the target population, induce the interactional goal and allow them to experience the media stimuli while measuring the target behavioral response through the selected measurement channel.

3.3 Advantages of PCS

Compared with face-to-face interaction data, PCS has several advantages.

1. It captures the variability in human behavior. In face-to-face interaction, for each speaker, only one listener's feedback data is collected. As shown in the upper part of Figure 3.2, what the data provides us is binary values over time, that is, giving feedback or not. However, that is not what we really want. Human behavior is flexible so that it is not appropriate to restrict it to a yes or no question. Instead, the listener's feedback needs to be associated with probability representing how likely the feedback will be given over time. With multiple independent participants' feedback, this can be done by building a histogram, which we call parasocial consensus, over time. This shows how many participants agree to give feedback

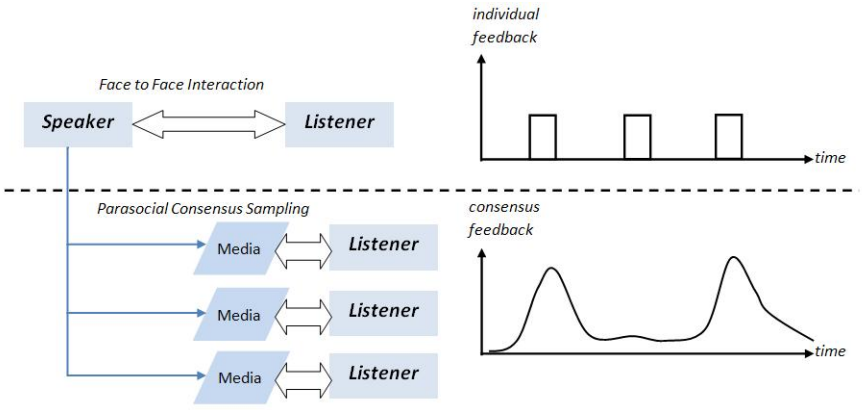

Figure 3.2: Comparison between Parasocial Consensus Sampling (PCS) and conventional Face-to-Face Interaction. Unlike face-to-face interaction, where interaction behaviors are deduced by observing how individuals respond in a social situation, parasocial consensus sampling allows multiple individuals to vicariously experience the same social situation to gain insight on the typical (i.e., consensus view) of how individuals behave within face-to-face interaction.

at time t. The more the number of participants agreeing to give feedback at time t, the higher probability the feedback has.

2. It helps tease apart the social causality of human behavior by holding some variable consistent. In face-to-face interaction, speaker and listener are paired. Their behaviors depend not only on their own specific characteristics, but also on their contingent reactions to the behavior of the other party in the interaction. This mutually-contingent nature of face-to-face interaction makes it difficult to tease apart the causality (i.e. is this listener an non-engaging people, or is he reacting to a disengaged speaker). While in PCS, speaker and listener are de-coupled. By having multiple listeners interact with the same speaker, we hold the speaker's behavior consistent so that it is possible to analyze how the listener's characteristics (e.g. personality) affect their behaviors. By having the same listener (or same set of listeners) interact with multiple speakers, we hold the listener's characteristics consistent so that it is possible to analyze how the speaker's characteristics affect

the listener's behaviors. We will explore this advantage in Chapter 6 and 7 to reveal how personality traits influence listener backchannel feedback.

3. It enables larger scale and more efficient data collection method. The advancement of multimedia technologies, such as YouTube service and its APIs, allow participants to experience media representation parasocially online. Without the limitation of bringing participants into the lab, it potentially enables much larger scale of data collection. Moreover, in parasocial interaction, participants interact with media representation instead of real people so that they can experience the same social situation independently and simultaneously. Therefore, we are able to collect behavior data in a parallel fashion. We will demonstrate this characteristic in Chapter 6 by crowdsourcing listener backchannel feedback to public via Amazon Mechanical Turk. This allows us to collect data from hundreds of participants in a fairly short time, which would be difficult to perform by traditional approaches.

3.4 PCS Tools

3.4.1 PCS Data Collection Tool

As mentioned before, PCS enables larger scale and more efficient data collection method. In order to fully take advantage of this characteristic, our data collection tool has to be a web application instead of the traditional desktop software. In the following studies, we either deploy the data collection tool on Amazon Web Service (aws.amazon.com) or in a local area network. It has two basic functionalities: presenting media representation and measuring target behavioral responses. The content of media representation is hosted on a web server (could be our own server or online service such as YouTube) and presented using a Flash Player. The player must be fully controlled by JavaScript, which can programmatically load content, decide when to start playing the content and when to stop and so on. This allows us to implement different logics in different scenarios. Our tool implements two methods to measure human behaviors: one is via keyboard

Figure 3.3: PCS Visualization Tool.

and the other one is via camera. When using the keyboard, which will be shown in Chapter 4, 5 and 6, participants are instructed to press a button whenever they feel like to perform the target behavior (e.g. head nod). The JavaScript code then records the timestamp when the button is pressed and sends it to a central server. When using the camera, which will be shown in Chapter 7, we develop a Microsoft Silverlight plug-in to take pictures of the participants at a rate of about 20 fps. The images can later be converted to videos for future analysis. We will compare the data collected via keyboard and camera in Chapter 7.

3.4.2 PCS Visualization Tool

The idea of parasocial consensus is to combine many parasocial responses (e.g. hundreds of participants respond to one video) to the same media clip to develop a composite view. It is difficult to reveal truth from such massive dataset, and thus a visualization tool is necessary to help us validate, analyze and understand both the composite view and the individual difference among participants. Therefore, we develop the PCS visualization tool (as shown in Figure 3.3). It consists of five components. By selecting a video from the video table (Component 4), the corresponding media clip (Component 1) and

participants (Component 5) will show up. Component 2 represents the composite view of all participants. The x-axis is the timeline, and the y-axis represents the number of participants who agree to perform the target behavior at that moment. By selecting the individual participant from the participant table (Component 5), the individual data will show up in Component 3 and we can compare the data from different participants easily. As the media clip plays, a timeline (the red vertical line) will move correspondingly in both Component 2 and 3 so that we can compare parasocial responses with the stimuli behavior (e.g. the speaker's behavior) in real time.

3.5 Conclusion

In this chapter, we introduce the general framework of Parasocial Consensus Sampling and compare the data collected under this framework with the traditional face-to-face interaction data. Although PCS can better capture the variability in human behavior, help tease apart the social causality and allow us to collect better data in a more efficient way, it has several drawbacks. For example, PCS breaks the contingency in face-to-face interaction by having participants interact with pre-recorded videos parasocially; participants in such parasocial interactions are motivated by artificial interactional goals, and so on. It is not clear yet whether the data collected under this framework could help us build better nonverbal behavior models, understand the variability of human behaviors and eventually improve the performance of virtual human systems. In next five chapters, we are going to validate, explore, extend, and apply the PCS framework from these aspects.

Chapter 4

Validating Parasocial Consensus Sampling: Modeling Listener Backchannel Feedback

4.1 Data Collection and Processing

Backchannel feedback, such as head nod or paraverbals like "uh-huh", is an important kind of nonverbal feedback within face-to-face interactions that signals a person's interest, attention and willingness to keep listening. We often see appropriate backchannel feedback when the listener connects with the speaker and is engrossed in the conversation; and such behavior will in turn influence the speaker's behavior positively. Backchannel feedback is very flexible: different people may have different responses when interacting with the exactly same speaker. If one listener doesn't provide backchannel feedback at this moment, it is not necessarily correct to conclude that that moment is not a good opportunity; and vice versa. A better solution would be to have multiple independent opinions on when to give backchannel feedback, and then use the consensus as a more reliable estimate. Also, from the consensus, we could infer the quality of the behavior cues from the speaker in terms of backchannel opportunity. This is exactly what Parasocial Consensus Sampling intends to be.

As discussed in Section 3.2, Parasocial Consensus Sampling is defined by five key elements: interactional goal, target behavioral response, media, target population and measurement channel. In this study, we customize it as follows:

- *Interactional Goal*: Creating rapport;
- *Target Behavioral Response*: Backchannel feedback;
- *Media*: Pre-recorded videos;
- *Target Population*: General public;
- *Measurement Channel*: Keyboard;

The video set (ICT Rapport data set 06-07) [1] used in our study was previously collected and used for studying how humans create rapport during face-to-face interactions. Each video records a human speaker retold a story to another human listener. While collecting the data, after each interaction, the speaker was asked to assess the level of rapport (i.e. rapport score) s/he felt when interacting with the listener.

We recruited 9 fluent English speakers (2 females, 7 males) from a local temporary employment agency to participate in the parasocial interactions with the human speaker videos. The average age of the participants is 45.2 years old, and the standard deviation is 12.6. The participants' behavior data was collected using the tool described in Section 3.4. They were instructed to pretend they were in a video teleconference with the speaker in the video and to establish rapport by conveying they were actively listening and interested in what was being said. To convey this interest, participants were asked to press the keyboard each time they felt like providing backchannel feedback such as head nod or paraverbals (e.g."uh-huh" or "OK"). In a ***one-day*** experiment, each of the 9 participants interacted with a total of 45 videos. This is much more efficient than the original approach to collect human behavior data from face-to-face interactions. To assess participants' subjective impressions on the task and their competence, they were asked three questions after watching each video:

- *Competence*: do you find the task easy or hard?

[1] Datasets are available for research purpose at rapport.ict.usc.edu

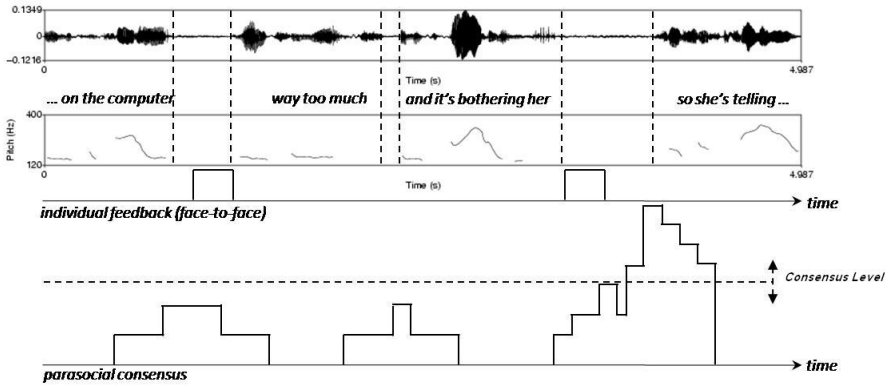

Figure 4.1: Example segment illustrating a parasocial consensus of listener backchannel varies over time. While individual feedback (from the original face-to-face interaction) only gives discrete prediction; our parasocial consensus shows the relative importance of each feedback. By applying a consensus level to our parasocial consensus, we get only important feedback.

- *Missed Opportunities*: do you think you missed good opportunities to provide feedback?

- *Timing*: do you think you gave feedback at points where you should not have?

Each question was answered using a 5-point Likert Scale. At the end of the experiment, participants were offered the opportunity to make general comments about the study.

The backchannel PCS dataset of each video consists of N (here N=9) sets of parasocial responses: $T_1, T_2, ..., T_N$, where N is the number of participants. For each parasocial interaction T_i, the PCS dataset contains the response timestamps T = t_1, t_2, ... indicating when the participant gave a response. These response timestamps are combined to create the parasocial consensus following a three-step approach:

(a) *Convert timestamps*: Each response timestamp can be viewed as a window of opportunity where backchannel feedback is likely. We create a one second time window (Ward & Tsukahara, 2000) centered around each timestamp. The timeline is then sampled at a constant frame rate of 10Hz.

(b) *Correct for individual differences (optional)*: Our current data collection requires participants to press a button when they expect a response and it is well known that individuals can differ significantly in their reaction time on such tasks. Therefore, the quality of consensus data can be improved if we first factor out these individual differences before combining response timestamps into a consensus. One way to estimate this delay (i.e. reaction time) is to compare the parasocial interaction with the face-to-face interaction. We follow (Ward & Tsukahara, 2000) approach to count how often PCS data matches the real listeners feedback and find the time offset that maximizes this score. This process was repeated independently on the PCS data of the nine participants. Their reaction time values varied from 600ms to 1200ms, with average of 970ms.

(c) *Build consensus view from multiple interactions*: a histogram is computed over time by looking at all the parasocial interactions. Whenever there is backchannel feedback occurring on a sample (sampled at 10Hz), the histogram of that sample is increased by 1. Thus, each sample is associated with a number indicating how many participants agree to give backchannel feedback at that point. Figure 4.1 shows an example of one parasocial consensus and compares it to the backchannel feedback from the real listener in the original face-to-face interaction. By looking at the real listener's feedback, it seems that pause is a good elicitor of listener feedback, but the relative strength of this feature is unclear. In contrast, the parasocial consensus clearly shows that the pauses differ in their propensity to elicit feedback. Looking more carefully at the example, we see the utterances before the first two pauses are statements, while the last one expresses an opinion, suggesting that pauses after opinions may be better predictors of listener feedback. Also, the speaker expressed emphasis on the third utterance. This result gives us a tool to better analyze and understand features that predict backchannel feedback.

Before applying the parasocial consensus to generate virtual human behavior or learn a prediction model, a consensus level is set to filter out the backchannel feedback whose probabilities are low. The probability is determined by the number of participants

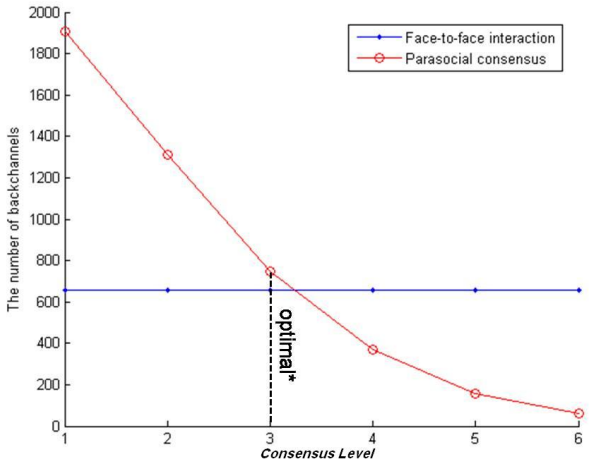

Figure 4.2: Selecting the consensus level. When the consensus level is set to 3, the number of backchannels from parasocial consensus data is closest to the number from face-to-face interaction data

agreeing to give that feedback. In the consensus data, feedback is associated with various probabilities. The higher the consensus level is, the fewer and more reliable backchannel feedback is selected. (Morency et al., 2008) explained consensus level as a way to make the virtual human have different expressiveness; the more frequent the feedback is, the more expressive the virtual human will be. We follow the concept here and select the consensus level so that the parasocial consensus data is as expressive as the original listener's behavior. By testing different values, as shown in Figure 4.2, the consensus level is set to 3 in all of the following experiments.

4.2 Experiments

In this section, we run a series of experiments to answer three questions: (1) Can people provide valid backchannel feedback in parasocial interaction? (2) Can we use the behavioral data collected in PCS to generate better virtual human behavior? (3) Can we learn a better prediction model from the PCS data for the virtual human?

	Competence	Missed Opportunities	Timing
Mean	4.30	1.30	1.10

Table 4.1: Results of the self-assessment questionnaire (in 5-point Likert scale)

4.2.1 Experiment 1

As mentioned in previous section, we ask participants three questions after watching each video to assess their subjective impressions on the task and their competence. From the participants' self-report, it is clear (shown in Table 4.1.) that they think the task is easy, and the number of missed opportunities and wrong feedback are small. In other words, they do feel like they can do this task quite well. Some comments indicated that after watching the first video and being accustomed to the special way to interact with the speaker in the video, it is easy to follow that routine later.

Besides the self-assessment, we also run an objective evaluation to measure the quality of the parasocial responses. In this experiment, participants were explicitly instructed to create a sense of rapport. One way to assess the data quality is to compare the consensus with the real listener's behaviors: if the behavior of an original listener closely approximates the consensus behavior, we would expect that the corresponding speaker rated high rapport score; if it differed significantly from the consensus, the speaker should rate low rapport score. We run the evaluation in four steps:

1. Separate videos into a low-rapport set and a high-rapport set: we sort the videos in ascending order based on the rapport score that the original speaker rated, and group the first half into the low-rapport set and the second half into the high-rapport set.

2. Predict backchannel: we set the consensus level to 3. The peaks in parasocial consensus whose values are larger than that are selected as the predicted backchannel feedback.

3. Compute correlation: the correlation is the percentage of the predicted backchannel feedback which can find matches in the original listener's behavior.

4. Compare the correlation between low-rapport set and high-rapport set: following the previous 3 steps, each video has a correlation measurement between parasocial consensus data and the original listener's behavior. ANOVA test is applied to find whether there is significant difference for the correlation between the two video sets. The mean value of the correlation for low-rapport set is 0.35, and the mean value for high-rapport set is 0.46, p-value is 0.03. This shows that parasocial consensus data correlates with the listener's behavior in the high-rapport dataset better than with those in the low-rapport set.

For question 1, we get an affirmative answer, that is, people are able to provide valid backchannel feedback in parasocial interaction.

4.2.2 Experiment 2

In this experiment, we drive the virtual human using parasocial consensus data and evaluate its performance regarding the quality of backchannel feedback.

Five speaker videos are randomly selected from the pre-recorded face-to-face interactions. For each speaker video, the virtual human, as a virtual listener, is driven by four sets of backchannel data respectively:

- *PCS*: the backchannel from parasocial consensus where the consensus level is set to 3;
- *F2F*: the original listener's backchannel feedback;
- *PCS all*: the backchannel from parasocial consensus where the consensus level is set to 0;
- *Random*: randomly generated backchannel feedback;

Figure 4.3: Videos for subjective evaluation

The four versions of virtual human are composed together with the corresponding speaker's video, illustrating a human interacting with a virtual listening agent (as shown in Figure 4.3). In a within-subjects design, 33 participants were recruited to evaluate the quality of each set of backchannel feedback. Each participant watched the four versions (presented in a random order) of one of the five videos. Before watching, the participants were told that "In each video, there is a speaker telling a story and a virtual human trying to give feedback to the speaker using head nods. The speaker will be the same in each video, the only difference is the virtual human's head nods. You will evaluate the timing of head nods by answering 4 questions after watching each video". The 4 questions we used in this evaluation are:

- *Rapport*: How much rapport do you feel between the agent and speaker while watching the video? (From 1(Not at all) to 7(Very much))
- *Believable*: Do you believe the agent was listening carefully to the speaker? (From 1(No, I don't believe) to 7(Yes, absolutely))
- *Wrong Head Nods*: How often do you think the agent head nod at inappropriate time? (From 1(Never inappropriate) to 7(Always inappropriate))
- *Missed Opportunities*: How often do you think the agent missed head nod opportunities? (From 1(Never miss) to 7(Always miss))

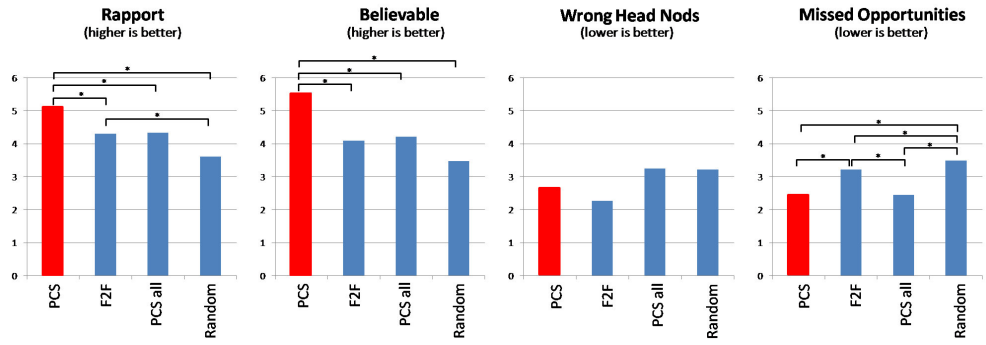

Figure 4.4: the subjective evaluation results for rapport, believable, wrong head nods, and missed opportunities of the four versions: PCS, F2F, PCS all, and Random. The star(*) means there is significant difference between the versions under the brackets.

General Linear Model repeated measure is used here to find whether there is significant difference among the four versions. The results are summarized in Figure 4.4.

Rapport: the mean of rapport level of the virtual human driven by PCS is 5.121, the mean of rapport level of the virtual human driven by F2F is 4.303, the mean of rapport level by PCS all is 4.333 and the mean of rapport level by random data is 3.606. The rapport level from PCS is significantly larger than the other three versions, and the rapport level from F2F is significantly larger than the random data.

Believable: the mean of believable level of the virtual human driven by PCS is 5.55, the mean of believable level by F2F is 4.09, the mean of believable level by PCS all is 4.21, and the mean of believable level by random data is 3.48. The believable level of PCS is significantly larger than the other three versions.

Wrong Head Nods: the mean of inappropriate head nods of the virtual human driven by PCS is 2.667, the mean of inappropriate head nods by F2F is 2.273, the mean of inappropriate head nods by PCS all is 3.242, and the mean of inappropriate head nods by random data is 3.212. There is no significant difference among the four versions, though.

Missed Opportunities: the mean of missed opportunities of the virtual human driven by PCS is 2.455, the mean of missed opportunities by F2F is 3.212, the mean of missed

opportunities by PCS all is 2.455, and the mean of missed opportunities by random data is 3.485. The missed opportunities of random data is significantly larger than the other three versions, the missed opportunities of F2F is significantly larger than that of PCS and PCS all.

From the rapport and believable questions, it is obvious that the virtual human driven by PCS creates the most rapport and people find it more believable than other versions. This demonstrates that the parasocial consensus sampling generates better listener backchannel data and thus achieves the interactional goal better than the conventional face-to-face interaction data does. Not surprisingly, random head nods produce the worst result, which matches previous work (Gratch et al., 2007) where the authors found that the contingency of agent feedback matters when it comes to creating virtual rapport. Interestingly, the virtual human driven by PCS all has similar performance as the F2F data. This confirms the importance of selecting a good consensus level.

When looking at the wrong head nods and missed opportunities questions, we find that all four approaches have approximately the same number of wrong head nods (false positive). The difference is in the missed opportunities (false negative) where both PCS and PCS all significantly outperform F2F and random data. This indicates that individuals cannot always catch all the good backchannel opportunities; whereas, by aggregating the feedback from multiple independent participants, we get a more complete picture. Also it is worth noticing that the number of missed opportunities in PCS is identical to that of PCS all, suggesting the consensus level did not filter important backchannel feedback.

In all, the result shows that PCS is able to generate better virtual human behavior.

4.2.3 Experiment 3

In this experiment, we show how to learn a predictive model from parasocial consensus data and evaluate the response of a virtual human driven by the predictive model.

For virtual human to taking advantage of the consensus data, we have to learn a predictive model that takes as input the speaker's nonverbal features and then predict when to give backchannel feedback. We choose Conditional Random Field (CRF) because of its advantages in modeling the sequential aspects of human behavior.

In the Rapport data set, we labeled the start and end time of the speaker's nonverbal behaviors, such as eye gaze, head nod and pause and so on. The CRF model learns the relationship between the occurrence of those nonverbal behaviors and backchannel feedback derived from the parasocial consensus data. Before learning, the nonverbal behaviors are first encoded using three encoding dictionaries (Morency et al., 2008). The idea of feature encoding is to explicitly describe the long-range dependency between the speaker's features and the listener's response. For example, the listener may decide to give backchannel feedback 1 second after the speaker looks back at him. The three encoding dictionaries are described as follows:

Binary Encoding: the value of the encoded feature is 1 between the start and the end time and 0 elsewhere. This encoding models the fact that the appearance of the nonverbal behavior is a predictor for giving backchannel feedback.

Step Encoding: this encoding generalizes binary encoding by adding two parameters: width and latency. Width represents the length of the encoded feature and the latency represents the delay between the start of the feature and its encoded version. It is useful if the influence of the feature on giving backchannel feedback is constant but with a certain delay and some duration.

Ramp Encoding: this encoding is similar to step encoding other than the encoded value increase within a period of time (width) linearly. It is useful if the influence of the feature on giving backchannel feedback is varying over time.

To apply encoding dictionaries, the only needed information is the starting time except the binary encoding. Binary encoding simply detects on and off of the features. In all cases, no future information is needed. The width and latency parameters for

step and ramp encoding are pre-defined. Therefore, all feature encoding can be done in real time.

Taken the encoded features, CRF model outputs a sequence of probabilities indicating the likelihood of giving backchannel feedback. We extract the backchannel predictions (i.e. give backchannel feedback now) from the sequence in two steps. First, local maximum (peaks) is selected as candidates; then, a threshold is set so that only the local maximum whose probability is larger than that is picked up as the final output. In our experiment, we select the threshold so that the number of feedback from CRF model is closest to that from the real listeners in the original face-to-face interaction.

As suggested by (Morency et al., 2008), four speaker features are used in the CRF model, they are pause using binary encoding, speaker looking at the listener using ramp encoding (width=2s delay=1s), lexical word "and" using step encoding (width=1s delay=0.5s) and speaker looking at the listener using binary encoding. All features were hand labeled by coders. Following the classic work (Lafferty et al., 2001), We train the CRF model and optimize the L_2 regularization term (10^k, k=-1..3) by applying 4-fold cross-validation on the training set. The performance of prediction model is measured by F_1 score, which is the harmonic mean of precision and recall. Precision is the probability that predicted backchannel feedback corresponds to actual listener's behavior; recall is the probability that backchannel feedback produced by an actual listener is predicted by the model.

We constructed a subjective evaluation experiment to assess whether the virtual human driven by our CRF model can achieve the interactional goal, creating rapport, better. From the Rapport data set, ten speaker videos were selected which were not used in the training set. As mentioned in Section 4.1, when the original face-to-face interactions were conducted, speakers were asked to rate rapport score after interacting with listeners. Of the ten videos, five were those with the lowest rapport scores and five were those with the highest rapport scores.

Our CRF model is compared with two models:

Rapport Agent: This is the baseline model. Rapport Agent (Gratch et al., 2007) used a rule-based model to predict when to give backchannel feedback. For example, if the speaker nods, the listener should nod back; if there are backchannel opportunities in the speaker's speech, the listener should nod back.

Real listener's behavior: This is the backchannel feedback of the human listeners from the original face-to-face interactions.

To evaluate different models, we composed videos illustrating a human speaker interacting with the virtual human (as shown in Figure 4.3). For each speaker video, the virtual human was driven by three models: PCS-CRF (CRF model trained on parasocial consensus data), Rapport Agent (rule-based model) and Natural (real listener's behavior). 17 participants were recruited to evaluate the performance of each version and each of them watched all three versions of the ten videos. Before watching, they were told you are going to evaluate different versions of a virtual agent in the context of interacting with a human speaker. In each video, there is a speaker telling a story and the virtual agent giving nonverbal feedback to the speaker by nodding. We need you to evaluate the timing of the agents head nods. Participants evaluated the virtual humans behavior by answering 7 questions:

Rapport Scale:

- *Close Connection*: Do you feel a close connection between the agent and the human speaker? (1(not at all) 7(yes, definitely close connection));
- *Engrossed*: Did the agent appear to be engrossed in listening to the story? (1(not engrossed at all) 7(very much engrossed));
- *Rapport*: Did there seem to be rapport between the agent and the speaker? (1(no rapport at all) 7(yes, theres rapport));
- *Listen Carefully*: Did the agent appear NOT to be listening carefully to the speaker? (1(No, he doesnt listen at all) 7(Yes, he is listening very carefully));

Perceived Accuracy:

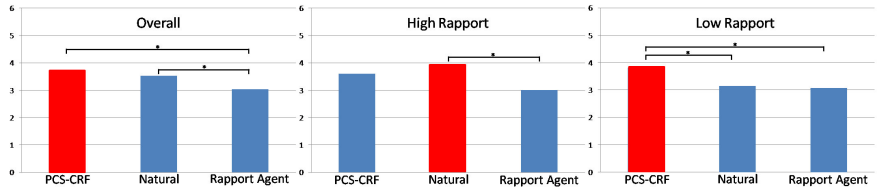

Figure 4.5: Rapport Scale. Overall, the virtual human driven by CRF is significantly better than Rapport Agent. For low-rapport videos, the virtual human driven by CRF is significantly better than the one driven by real listener's behavior.

- *Precision*: How often do you think the agent nodded his head at an inappropriate time? (1(always inappropriate) 7(always appropriate));
- *Recall*: How often do you think the agent missed head nod opportunities? (1(missed a lot) 7(never missed));

Naturalness:

- Do you think the virtual agent's behavior is natural? (1(not natural at all) -7(yes, absolutely natural));

ANOVA test was applied to find whether there is significant difference among the three versions. The four items related to rapport, showing good reliability (Cronbach's alpha = 0.98) among each other, are collapsed into one single scale. The results are summarized from Figure 4.5 to 4.8. In each figure, from left to right, they are mean values for all 10 videos (Overall), 5 high-rapport videos (High Rapport), and 5 low-rapport videos (Low Rapport) respectively. The star (*) means there is significant difference between the versions under the bracket.

Rapport Scale: Overall, the mean of the rapport scale for PCS-CRF is 3.71, the mean for Natural is 3.52, and it is 3.04 for Rapport Agent; both PCS-CRF and Natural are significantly better than Rapport Agent. For the high-rapport data set, the mean of the rapport scale for PCS-CRF is 3.60, it is 3.93 for Natural and 3.00 for Rapport Agent; the Natural is significantly better than Rapport agent, while there is no significant difference between PCS-CRF and Natural. For the low-rapport data set, the mean for

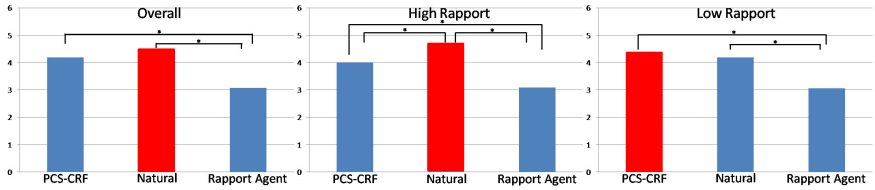

Figure 4.6: Precision. The virtual human driven by CRF provides backchannel feedback more precisely than the Rapport Agent.

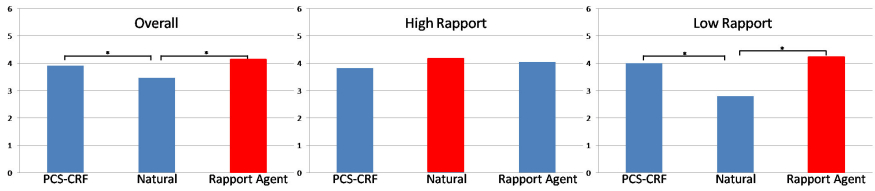

Figure 4.7: Recall. The virtual human driven by real listener's behavior misses more opportunities to provide backchannel feedback than the other two versions do.

PCS-CRF, Natural and Rapport Agent are 3.81, 3.14 and 3.07 respectively. PCS-CRF outperforms the other two significantly.

Precision: Overall, the mean of the precision for PCS-CRF is 4.19, the mean of the precision for Natural is 4.45, and it is 3.07 for Rapport Agent; both PCS-CRF and Natural are significantly better than Rapport Agent. For the high-rapport data set, the mean of the precision for PCS-CRF is 4.00, it is 4.72 for Natural and 3.09 for Rapport Agent; natural outperforms the other two significantly, and the PCS-CRF is significantly better than Rapport Agent. For the low-rapport data set, the mean of the

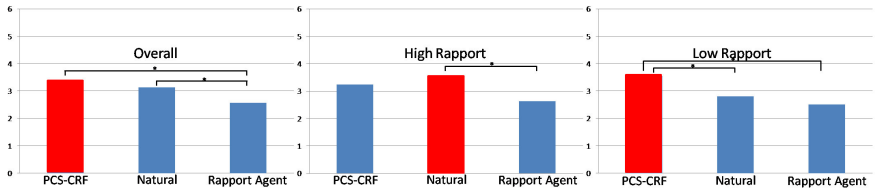

Figure 4.8: Natural. Overall, The virtual human driven by CRF is more natural than Rapport Agent. For low-rapport videos, the virtual human driven by CRF is more natural than the one driven by real listener's behavior.

precision are 4.37, 4.19 and 3.06 for PCS-CRF, Natural and Rapport Agent respectively; both PCS-CRF and Natural are significantly better than Rapport Agent.

Recall: Overall, the mean of the recall are 3.92, 3.46 and 4.15 for PCS-CRF, Natural and Rapport Agent respectively; Natural is significantly worse than the other two. For the high-rapport data set, the mean are 3.83, 4.19 and 4.05 for PCS-CRF, Natural and Rapport Agent respectively; there is no significant difference among them. For low-rapport data set, the mean are 4.00, 2.79, and 4.24 for PCS-CRF, Natural and Rapport Agent; Natural is significantly worse than the other two.

Naturalness: Overall, the mean of naturalness are 3.40, 3.13 and 2.57 for PCS-CRF, Natural and Rapport Agent respectively; Rapport Agent is significantly worse than the other two. For high-rapport data set, the mean of naturalness are 3.24, 3.48 and 2.64 for PCS-CRF, Natural and Rapport Agent respectively; Natural outperforms Rapport Agent significantly, and there is no significant difference between PCS-CRF and Natural. For low-rapport data set, the mean are 3.55, 2.81 and .252 for PCS-CRF, Natural and Rapport Agent respectively; PCS-CRF is significantly better than the other two.

Overall, the virtual human driven by the CRF model (PCS-CRF) is significantly better than the Rapport Agent. It demonstrates a better prediction model can be learned from parasocial consensus sampling data. If applied to virtual human systems, it has the potential to create better social effects than the Rapport Agent did. By looking at the virtual human driven by PCS-CRF and the one driven by real listeners behavior, we don't see significant difference overall. However there is significant difference between the two in the low-rapport videos, which shows PCS-CRF can do as well as real human listeners who succeed in creating rapport but better than those who fail to.

For the Precision question, PCS-CRF does significantly better than the Rapport Agent; while there is no difference between the two for the Recall question. The Rapport Agent gave responses whenever he saw the speaker nodded or the presence of backchannel opportunities. Such simple rules may lead to many unnecessary head nods

so that the recall is high while the precision is low. This explains the reason why PCS-CRF outperforms Rapport Agent.

By comparing the virtual human driven by CRF and the one driven by real listener's behavior, we don't see significant difference between them for the Precision question, which is expected, since real listeners are not likely to give wrong feedback in natural face-to-face interactions. However, there is significant difference between the two for the Recall question, and the difference mainly comes from the low-rapport videos. This explains why PCS-CRF does better than real listener's behavior in the low-rapport videos. Real listeners sometimes don't catch all appropriate backchannel opportunities within the interactions and thus fail to create rapport. On the other hand, PCS-CRF is learned from consensus data which is not likely to fail in this regard unless most of the parasocial interactions fail to give necessary feedback at the same time.

By comparing the Naturalness question with the Rapport-related question, we find the virtual human is perceived more natural when it creates more rapport. This confirms the previous finding that creating rapport does lead to positive social effects.

The results of Experiment 3 show that we can learn a better prediction model from PCS data for the virtual human.

4.3 Conclusion

This chapter describes the Parasocial Consensus Sampling framework and our first efforts in applying the PCS framework to model listener backchannel feedback. In a series of experiments, we show that (1) people are able to provide valid backchannel feedback in parasocial interaction, (2) PCS data generates better virtual human behavior and (3) can be used to learn a better prediction model for virtual human.

Chapter 5

Validating Parasocial Consensus Sampling: Modeling Turn-taking Behavior

5.1 Data Collection and Processing

Human conversation is a cooperative and fluent activity. The speaker monitors feedback from the listener and adjusts his behaviors accordingly while producing utterances; the listener provides moment-to-moment feedback to alter and help co-construct the subsequent speech while processing the utterances from the speaker. People rarely speak simultaneously. Rather, the roles of speaker and listener are regulated seamlessly by a negotiation process of turn-taking. To further validate the PCS framework, we apply it to model the turn-taking behavior.

As stated in Section 3.2, Parasocial Consensus Sampling consists of five key elements, we customize them as follows to collect human turn-taking behavior.

- *Interactional Goal*: In our work, we want the virtual human to learn a turn-taking strategy similar to that of a human and avoid interruptions and long mutual silences during the conversation;
- *Target Behavioral Response*: Take the conversational turn;
- *Media*: Pre-recorded videos of a series of interviews;
- *Target Population*: General public;

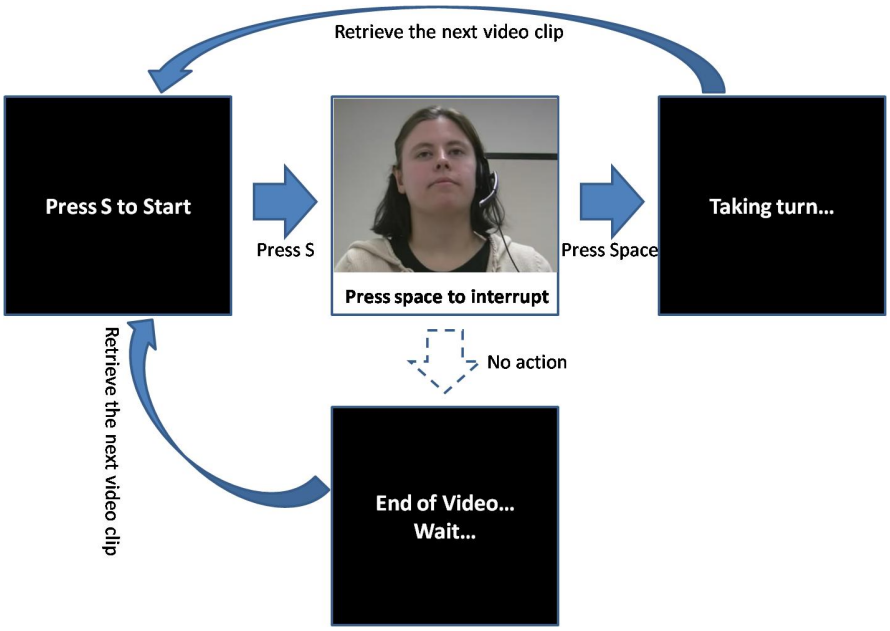

Figure 5.1: The procedure of the designed interactive experience. We provide some feedback to the participants whey they are taking turns.

- *Measurement Channel*: Keyboard;

The videos are from an interview data set collected in a previous unrelated study (Kang et al., 2009) on intimate self-disclosure. In the dataset, participants recruited from the general population were asked a series of 10 increasingly intimate questions, either by a human confederate via a two-way video conference or by a virtual human interviewer controlled by a human confederate through a wizard-of-oz interface. In the virtual human condition, the interviewer pressed a button to retrieve pre-recorded voice messages when he wanted to take turns and the virtual human displayed corresponding lip-sync movements at the same time. In both conditions, the interviewer listened silently until they felt it was appropriate to proceed to the next question. All interviewee's behaviors were videotaped. These videos are used in the following experiment.

In our parasocial interactions, in order to strengthen the parasocial effect, we customized the PCS data collection tool described in Section 3.4 to created an interactive interface that highlighted the feeling that coders were engaged in an actual interview,

although all coders knew they were interacting with pre-recorded media. This coding procedure is illustrated in Figure 5.1. Interviewee videos are displayed in a web browser. Coders press the "S" key to "ask" a question (upon pressing the key they hear an audio recording of a question asked in the original interview) and see the interviewees reactions as they respond to the question. At any point during the answer, the coder may press "Space" key to indicate when it is a proper point to interrupt. At this point, the interface automatically loads a new randomly selected question and waits till the coder presses "S" again to ask this next question. If the coder does not press Space key, the video will keep playing until reaching the end. At this point the video ends and the coder is prompted to push "S" to ask the next question.

We recruited 9 participants from a local temporary agency. First, they signed confidentiality forms protecting the video materials they were about to watch. Then, they were taught how to use the interactive tool and provided instruction regarding the task: "You are going to watch some videos clips where an interviewee is answering questions. While watching, please press space bar as soon as you think the interviewee has finished his answer to the current question". Every participant watched all videos in ***one-day*** experiment. None of them reported difficulty in finishing the task or following the routines of our interactive tool.

There are 46 videos in total in our data set, each contains 10 questions and was watched by all 9 participants. Therefore, there should be 4140 ($46 \times 10 \times 9$) turn-taking opportunities (i.e. if each participant took a turn at the end of the answer to each question). The turn-taking data we collected contains approximately 3300 records, which suggests that individual data contains variability. Not everyone agreed to take a turn at the end of each question. By combining all individual data together, we can develop a more reliable version of how a typical individual would take turns.

Similar to the approach in modeling backchannel feedback (Section 4.1), we compute a histogram of multiple participants' responses to build the consensus. We did this by

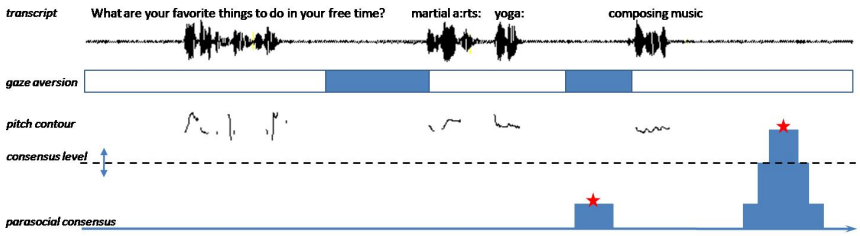

Figure 5.2: An example of a parasocial consensus. The consensus peaks (stars) are possible turn-taking points and their height is a measure of their quality (representing the agreement between PCS coders). The first peak, which co-occurs with the speaker's gaze aversion, was not seen as a good turn-taking point by most coders. By setting the consensus level, we can remove such outliners from the consensus view.

converting the video time line into samples at a sample rate of 30Hz. Each turn-taking point (the time when a participant pressed the button) is centered in a 1 second window, that is, a window of 30 samples. Whenever a turn took place on a sample, the histogram of that sample increases by 1. Therefore, each sample is associated with a number indicating the number of participants who agreed to take the turn here. Figure 5.2 shows an example of a parasocial consensus. The parasocial coders identified two possible turn-taking points (the one within the pause following "yoga:" and the other one within the pause following "music"), but they are associated with different probabilities. Clearly, the consensus view favors the second one probably due to the fact that the speaker is averting gaze during the first place. Overall, the histogram can be considered as a measurement of probability to take a turn. By setting a proper consensus level, we can separate the turn-taking places with low probabilities (i.e. bad instances) from the ones with high probabilities (i.e. good instances), which is better for further analysis and learning the prediction model. For the 46 videos, there are 460 turn-taking opportunities (ten questions in each video) in total. We choose the consensus level so that the number of turn-taking from the consensus data is closest to 460. By testing different values, the optimal consensus level is 3 with 479 turn-taking places.

	Turn-taking Pauses	Non-turn-taking Pauses
Looking-away	3% (11)	59% (598)
Looking-towards	27% (114)	12% (123)
Nods	17% (71)	8% (77)
Pitch slope Up	38% (160)	22% (227)
Pitch slope Down	35% (149)	24% (238)
Pitch slope Straight	27% (113)	54% (547)
Average Pitch Above	14% (60)	11% (108)
Average Pitch Below	38% (162)	32% (321)
Average Pitch At	48% (200)	57% (583)
Syntax Completion	98% (416)	64% (648)

Table 5.1: The percentage of turn-taking pauses and non-turn-taking pauses that co-occur with different features. The absolute number is shown in parentheses.

5.2 Analysis of the PCS Turn-taking Data

In preparation for analysis, interviewee's answers were transcribed and gaze and head nods were manually annotated. Prosodic features (pause and pitch contour) were labeled automatically by Aizula (Ward, 1997). We use similar intonation patterns as mentioned in (Jonsdottir et al., 2008), where the pitch contour in the most recent tail (last 300 msec) of speech right before a pause is used to find the slope and the average pitch value. The slope is categorized into 3 classes: Up, Straight, and Down; the average pitch value in the tail is categorized into Above, At, and Below by comparing with the average pitch value of the whole speech segment before that pause. The completion of syntactic structures is extracted from the transcripts. Since the basic units in the transcripts are always complete clauses, their boundaries can be considered as syntax completion points. We end up with 10 features in the analysis as shown in Table 5.1.

We classify pauses as a turn-taking pause if and only if a peak from the parasocial consensus occurs within the duration of the pause; otherwise we classify it as a non-turn-taking pause. There are 1434 pauses in our data, of which 422 are the turn-taking pauses. Of all turn-taking pauses, 95% are longer than 1.5s; while only 30% of non-turn-taking pauses are longer than that. Our analysis is based on two measurements:

1. the frequency of co-occurrence between each of the 10 features and turn-taking and non-turn-taking pauses;

2. the average pause duration before taking a turn when one of the 10 features is present.

The analysis is done on four aspects:

Gaze: As shown in Table 5.1, 59% of non-turn-taking pauses co-occur with looking-away, whereas the percentage is only 3% for turn-taking pauses. This suggests that looking-away is a very effective turn-holding signal. By looking away, the speaker closes the turn-taking channel temporarily and thus can focus on recalling information while holding the floor. For turn-taking pauses, 27% of them co-occur with looking-towards; while the percentage is only 12% for non-turn-taking pauses. Therefore, looking-towards is not randomly distributed among pauses, it provides discriminative information. Interestingly, at turn-taking pauses, we find different gaze patterns influence the pause duration people wait to take turns. When the interviewee is staring at the interviewer, the average pause duration is about 1.3s; when looking-towards happens, the average pause duration is about 1.7s; in a few cases when looking-away happens, the average pause duration is about 2.0s. This suggests people employ different turn-taking strategies by observing different gaze patterns. Looking-away makes people hesitate to take turns; people make turn-taking decision more quickly when the interviewer is staring at him or her. Although looking-towards is a predictive cue for turn-taking, the average pause duration in this condition is not the shortest. We find people tend to wait for a period of time after the interviewee looks back. This may be due to the needs to confirm there is no more information forthcoming after the interviewee's gaze aversion. The average waiting time is about 0.8s.

Prosody: As Table 5.1 shows, 73% of turn-taking pauses associated with rising or falling pitch contours; the percentage is only 46% for non-turn-taking pauses. This suggests that the slope of pitch contour is a predictive cue for turn-taking. This is in line with (Duncan, 1972) where they considered rising or falling intonation as one of

the turn-yielding signals. For the average pitch value, we do not observe significantly discriminative information. At turn-taking pauses, 38% co-occur with low pitch, 14% co-occur with high pitch, and 48% co-occur with average pitch; a similar pattern occurs with non-turn-taking pauses, 32% co-occur with low pitch, 11% co-occur with high pitch, and 57% co-occur with average pitch. Similarly to the impact of gaze patterns on pause duration, different slope patterns influence the time people use to take turns as well. We find the average pause duration is 1.37s when the slope is down, 1.45s when the slope is up, and 1.56s when it is flat. People tend to be more hesitated to take turns when the slope is flat.

Nod: We find that head nods happen at 17% of turn-taking pauses but only at 8% of non-turn-taking places. This suggests that head nod serves as a predictive cue for turn-taking. The average pause duration people wait to take turns is about 1.7s when head nod occurs, while it is shorter (about 1.4s) when there is no head nod. We observe that people usually wait until the head gesture finishes before taking a turn, which explains why the duration is longer when head nods appear. However, they immediately take the turn after the gesture finishes. The average gap between turn-taking and the end of a head nod is about 0.2s.

Syntax: Completed syntactic structures like "sentences, clauses, phrases, and even one-word constructions" (Sacks et al., 1974) are considered to be possible turn-taking opportunities. In our data, the percentage of turn-taking pauses that co-occur with syntax completion is high, but there are also 64% non-turn-taking pauses following completed syntactic structures. In interviews, a turn is usually consist of several complete clauses; thus, many syntax completion points are inside a turn. Although it is a necessary condition for turn-taking, it is not sufficient enough.

5.3 Build Multimodal End-of-Turn Prediction Model

We apply the same techniques (i.e. CRF and encoding dictionary (Section 4.2.3)) as building the backchannel prediction model to model turn-taking behavior. The initial

feature set contains looking-away, looking-towards, head nod and prosody. Each of them is encoded by 13 encoding dictionaries, which are binary encoding, step encoding with six different sets of parameters ([0.5 0], [1.0 0], [0.5 0.5], [1.0 0.5], [0.5 1.0], [1.0 1.0], each pair represents the width and the latency) and ramp encoding with the same six sets of parameters. Before training, we apply a feature selection algorithm to find the most informative features to avoid overfitting. The feature selection algorithm is based on regularization technique. Regularization is considered to smooth the model by introducing a penalty term into the optimization function. The penalty will increase if the complexity of the model increases. It is useful when the model has a high complexity (i.e. lot of features) and the size of the training data is limited (i.e. a small number of training samples). One of the most used regularization forms is L_1 regularization which is defined as:

$$R(\theta) = \lambda||\theta||_1 = \lambda\Sigma|\theta_i| \tag{5.1}$$

where θ is the model parameters (here they are the weights of features) and $\lambda > 0$. L_1 regularization has been used for feature selection in nonverbal behavior analysis before (Ozkan & Morency, 2010). The feature selection process starts with a high regularization parameter λ where all feature weights are zeros and then gradually reduces the value of λ until all weights become non-zeros. During the process, the features whose weights turn to non-zeros earlier are considered more important. For explanation purpose, Figure 5.3 shows the change of weights of a few features during the feature selection process. From left to right, the value of λ gradually decreases. When λ is 2000, only two features' weights are non-zeros, they are pause with binary encoding and looking-away with binary encoding, which suggests pause and looking-away are two of the most important features. The values of the two weights reflect how they influence turn-taking behavior, that is, the occurrence of pause is likely to induce turn-taking (with positive weight), while the occurrence of looking-away is not (with negative weight). Later, the weights of pitch slope, looking-towards, and head nod turn to non-zeros. The order of turning to

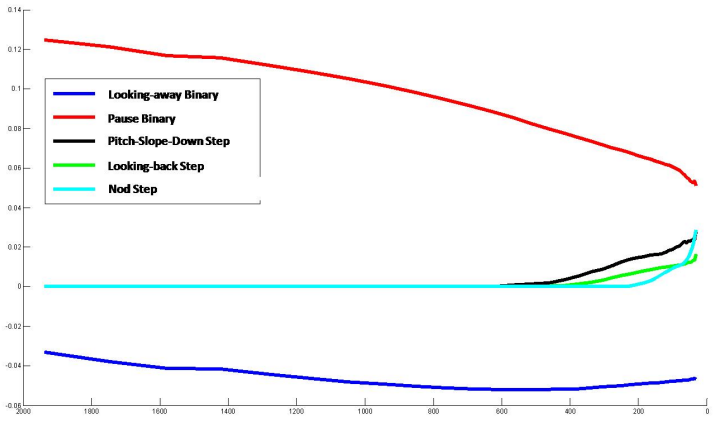

Figure 5.3: Visualization of the feature selection process.

non-zeros can be considered as a measurement of the relative importance. The number of features is selected automatically by comparing the performance of models on the validation set. In our experiment, 10 features are automatically selected as the final feature set. Interestingly, the selected features have a good match with the analysis in Section 5.2. For example, looking-away with binary encoding is selected, which matches the close correlation between looking-away and non-turn-taking pauses; the rising (Up) and falling (Down) pitch slopes are both selected, which matches the observation that these two patterns are correlated with turn-taking pauses.

In the training process, we randomly split all videos into 4 folders, 3 of them are used as developing set and the remaining one is used as testing set. The process repeats 4 times so that every video appears in the testing set once. Following (Lafferty et al., 2001), we train the CRF model and optimize L_2 regularization term (10^k, k = -1..3) by applying 10-fold cross validation on the developing set. The performance of the model is measured by F_1 score, which is the weighted harmonic mean of precision and recall. Precision measures the percentage of correct turn-taking predictions among all predictions; recall measures the percentage of turn-taking places in the testing set that

are correctly predicted. A prediction is considered correct if the predicted turn-taking occurs within the 1 second window around the ground truth.

5.4 Evaluation

We compare the performance of two models learned from PCS (PCS-Multimodal model and PCS-Pause model) with two baseline models (prosody model and syntax model) from previous literature.

PCS-Multimodal: This is the model we learned in the previous section;

PCS-Pause: The pause model is created by choosing an optimal length of pause duration. It classifies a pause to be a turn-taking pause if its duration is longer than the threshold. As analyzed in Section 5.2, the optimal pause threshold is selected to be 1.5s, which is the average time that the participants took to take turns during parasocial interactions.

Prosody Model: Prosody model is trained the same way as PCS-Multimodal model but with only prosodic features. These prosodic features were based on (Jonsdottir et al., 2008).

Syntax Model: Syntax model is based on the previous work of (Sacks et al., 1974), where syntax completion points, such as the end of "sentences, clauses, phrases, and one-word constructions", are suggested as possible turn-taking places. The syntax completion points are determined by the transcripts of the interviewee's speech. Since the basic units in the transcripts are always complete clauses, their boundaries can be considered as syntax completion points.

Predictions from each model are evaluated by two criteria. The first one is turn-taking pause prediction, where the results are compared with the confederate interviewer's turn-taking behavior in the original interactions. The predicted time is considered correct if happening during the same pause as when the confederate interviewer took a turn. One should note that the confederate interviewer's turn-taking behavior

	Precision	Recall	F_1
PCS-Multimodal Model	0.78	0.81	0.80
PCS-Pause Model	0.59	0.901	0.71
Prosody Model	0.58	0.77	0.67
Syntax Model	0.29	0.97	0.45

Table 5.2: Evaluations for Turn-taking pause prediction: F_1 score of PCS-Multimodal is significantly better than that of the other three models

	Precision	Recall	F_1
PCS-Multimodal Model	0.43	0.42	0.43
PCS-Pause Model	0.32	0.46	0.38
Prosody Model	0.33	0.42	0.37

Table 5.3: Evaluation for Turn-taking time prediction: F_1 of PCS-Multimodal is better than that of the other two models.

is only one of the possible ways to behave. That is why we also compare with the parasocial consensus which aggregates multiple coders.

The second criterion is turn-taking time prediction, where the results are compared with the turn-taking opportunities from the parasocial consensus. This is a more challenging criterion, since the predicted time is considered correct only if it is within the 1 second window around the turn-taking opportunities from parasocial consensus. Note that the syntax model cannot predict exact turn-taking time, since it is evaluated at the clause level. For this reason, we did not include it in this condition.

Turn-taking pause prediction: As Table 5.2 shows, F_1 score of the PCS-Multimodal model is better than that of other three models. Paired T-Test comparisons between PCS-Multimodal model and the other three models ($p = 0.05$ for PCS-Pause, $p < 0.01$ for the other two) suggest the difference is statistically significant. This indicates syntax or prosody cannot provide enough information by themselves to predict the turn-taking pauses. In our data, a turn is usually consist of several clauses, where the syntax completion points are not always followed by the transition of turns. As suggested in (Selting, 2000), turn-constructional units in some activities, like story-telling, where a turn is always consist of several of them, should be categorized into non-final and final

Figure 5.4: Turn-taking time with staring gaze: The interviewee is staring at the interviewer at the end of this answer, our model predicts the turn-taking time about 0.9s after the end of the answer and the participants in parasocial interaction took about 1.0s.

ones and the end of non-final turn-constructional units are not followed by transition of turns. Although (Jonsdottir et al., 2008) showed prosodic features are good at end-of-turn prediction, it is not the case in our data where the change of prosody may sometimes be caused by the completion of clauses inside a turn but not the end of turn. By leveraging the multimodal features, our PCS-Multimodal model performs the best.

Turn-taking time prediction: PCS-Multimodal model is significantly better than prosody model (p=0.02) but not better than PCS-Pause model in predicting turn-taking time. However, the PCS-Pause model cannot represent the flexibility of human behavior; that is, in real conversations, the turn-taking time is not fixed (e.g. 1.5s) but varies based on the other conversational partner's behavior. PCS-Multimodal model is able to generate human-like turn-taking strategy by taking into account the interviewee's nonverbal features. For example, in Figure 5.4, the interviewee is staring at the interviewer when she answers the question, the model waits for only 0.9 second to take the turn; in Figure 5.5, the interviewee looks towards the interviewer about 1 second after he finishes his answer, the model waits for about 2 seconds to take the turn. The variation of the predictions from PCS-Multimodal model matches the trend we find in Section 5.2.

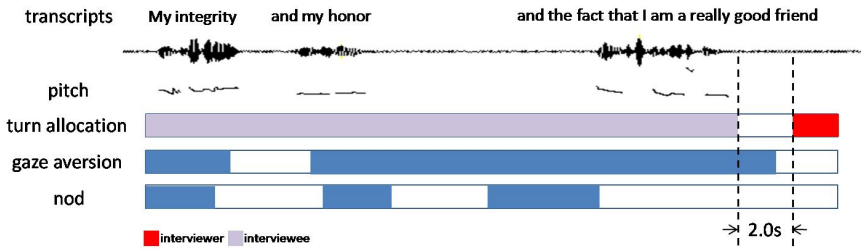

Figure 5.5: Turn-taking time with looking-towards: The interviewee looks towards the interviewer at the end of his answer, our model predicts the turn-taking time about 2.0s after the end of answer and the participants in parasocial interaction took about 1.9s.

5.5 Conclusion

This chapter describes our second study in applying the PCS framework to model turn-taking behavior. By analyzing the consensus view of how a typical individual would take turns, we found multimodal features provide predictive information to learn the end-of-turn prediction model. Our PCS-Multimodal model achieved a high accuracy in predicting the turn-taking pauses and had a human-like turn-taking strategy, which further validates this new methodology.

Chapter 6

Exploring Parasocial Consensus Sampling: Crowdsourcing Backchannel Feedback

In previous two chapters, we validated the PCS framework by applying it to model listener backchannel feedback and turn-taking behavior. The results imply that (1) people are able to provide valid behavior data in parasocial interactions, and (2) compared to face-to-face interaction data, PCS data generates better virtual human behaviors and (3) can be used to learn better prediction models for virtual human. The better performance of PCS data is probably due to the fact that it captures the variability of human behavior better than face-to-face interaction data does. As mentioned in Chapter 3, PCS has two other characteristics that have not yet been fully explored. First, PCS helps tease apart the social causality of human behavior by holding some variables consistent (e.g. the speaker's behavior keeps the same when different listeners interact with him parasocially). Second, PCS enables much more efficient data collection mechanism than the traditional approaches. These characteristics inspire the research work described in this and the next chapter, which would be extremely difficult to perform by traditional approaches.

6.1 The Problem

As mentioned before, human behavior contains significant variability, especially in face-to-face interactions. Such variability comes from both sides involved in the interactions, and it is difficult to reveal the causalities using face-to-face interaction data because of its mutually-contingent nature. Here we use listener backchannel feedback as an example to demonstrate how PCS data could help us tease apart the social causality and understand how each side influence the behavior respectively.

Chapter 4 has suggested several speaker features, such as pause and eye gaze, as listener backchannel feedback elicitors. However, this is not enough to explain the individual variability observed in listener backchannel feedback. We argue that the listener backchannel feedback (BC) is a function of both speaker's behavior and the listener's characteristics:

$$BC(listener) = f(speaker's behavior) + g(listener's characteristics) \tag{6.1}$$

The listener's characteristics may include a lot of attributes. Here we focus on how listener's personality traits influence the backchannel feedback.

6.2 Data Collection and Processing

Although we have demonstrated the efficiency of PCS framework in previous work, the size of the samples (i.e. population) is still small (e.g. in both backchannel and turn-taking studies, the number of participants is only 9). PCS allows us to collect data in a much larger scale. Here we crowdsource backchannel feedback to public using Amazon Mechanical Turk (AMT), which enables us to collect hundreds of responses to each video in a much faster and less expensive way.

Amazon Mechanical Turk is a web service created by Amazon, which provides a marketplace for those who (i.e. requester) want to post tasks to be completed and

Big Five Personality Traits	Extroversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Self-Consciousness	Self-directed, Other-directed
Parasocial Experience	Parasocial experience scale (Hartmann & Goldhoorn, 2010)
Other	Shyness, Self-monitoring, Gender

Table 6.1: The attributes of each coder we measured before they started interacting with the speakers parasocially.

specify prices for completing them. The idea is to utilize people's (i.e. coder) small trunk of time, typically from 5 minutes to 1 hour, to finish trivial tasks, such as image tagging. The price of each task is often on the order of a few cents. Therefore, it is possible to have many coders repeat the same task. Although the individual coder is usually not an expert for the task, one often can achieve expert-level results by relying on the wisdom of crowds (Surowiecki, 2004).

We used the data collection tool described in Section 3.4 and integrated it with AMT. Coders can find our tasks on the marketplace and follow a link to the web application. First, they finish a 90-item questionnaire (see Appendix A) that assesses several individual differences that we expect to influence backchannel behavior (listed in Table 6.1). They next watch an example video illustrating the process of interacting with a human speaker parasocially. Next, they watch 8 videos in random order, each about 2 to 3 minute long. Each video features a human speaker telling two stories. Coders press a button whenever they feel like to provide feedback, such as nodding or uttering "uh-hum". After interacting with each video, coders answer a 6-item questionnaire to measure their parasocial experiences (Hartmann & Goldhoorn, 2010). At the end of the task, they leave comments about the study. Coders need to finish the study within 90 minutes in order to get paid, and we pay 4 US dollars for their work. Following the procedure, we initially constructed a dataset of 350 coders providing backchannel data to 8 speaker videos. To better understand speaker variability, we subsequently coded additional 16 speaker videos (in two rounds) using 100 coders each. For the analysis that follows, we collapse these three collections into a single dataset of 24 speaker videos.

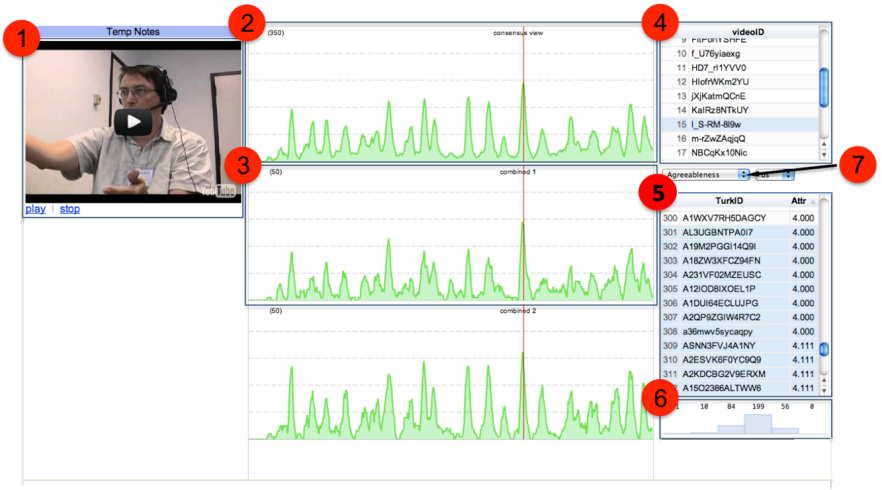

Figure 6.1: The interface of the extended PCS visualization tool. By selecting a video ID from the video table (Component 4), the corresponding speaker video (Component 1) and coders (Component 5) show up. Component 2 represents the consensus view of all coders for the speaker video. By selecting a group of coders, a histogram (Component 3) will be computed by aggregating the responses from those coders. We measure several personality attributes of each coder. By selecting an attribute (Component 7), the coder table (Component 5) will be populated with the corresponding values. And a histogram (Component 6) will be displayed, indicating the distribution of coders along the selected attribute. In this figure, two histograms below Component 2 represent the consensus of two sub-groups of coders: those that are least and most agreeable, respectively.

We extend the PCS visualization tool described in Section 3.4 to help us analyze how the coders' personalities influence their own behaviors. In this extended version (as shown in Figure 6.1), each coder is associated with a set of attributes (i.e. personality measurements). By selecting an attribute (Component 7), the value of the attribute of each coder will be displayed (Component 5) and a histogram (Component 6) will show up to illustrate the distribution of coders along the selected attribute. User can sort coders by the values of the selected attributes (Component 5). If select a group of coders, a histogram (Component 3) will be computed by aggregating the responses from those coders. With this tool, it is easy to compare the average behaviors from two different groups (e.g. the least and most agreeable coders).

6.3 Data Analysis

We analyze the causalities of listener backchannel feedback from both speaker side and listener side respectively.

6.3.1 Speaker's Features

Clearly, some of the variance in listener backchannel feedback is driven by features of the speaker's verbal and nonverbal behavior. We leave verbal analysis to future work but here analyze what speaker nonverbal features trigger backchannel feedback. Our analysis is based on the frequency of co-occurrence between speaker features and listener backchannel feedback. For each speaker feature, we have the starting time t_s and end time t_e that have been labeled by human annotators. For listener backchannel feedback, we record the time (t_b) when the coder pressed the button. We count it as a match if,

$$t_s <= t_b <= t_e + window \tag{6.2}$$

That is, if the backchannel feedback occurs within the speaker feature or right after it, the feature is considered as triggering the feedback. Inspired by the idea of encoding dictionary (Morency et al., 2008), we add the variable *"window"* to count the case where backchannel feedback is triggered by speaker features but with a certain delay. In our work, *window* is set to be 500ms.

For each coder, we count the co-occurrence between the backchannel feedback and each of the speaker features. If a speaker feature always co-occurs with backchannel feedback, it is considered as an important feature that the coder relies on. We measure the importance of a feature as follows:

1. P = *# of co-occurrence / # of occurrence of a feature*
2. R = *# of co-occurrence / # of backchannel feedback*

And then, the importance $\boldsymbol{I}$ is calculated as the harmonic mean of P and R. By ranking the speaker features on the importance measure $\boldsymbol{I}$, we find 99% coders depend on "pause" and "speaker eye gaze" to provide backchannel feedback, and 73% coders depend on "pause", "speaker eye gaze" and "speaker head nod" to provide feedback. This is in line with the analysis described in Chapter 4. The result suggests that listeners use almost the same subset of speaker features to decide when to give feedback, which cannot fully explain the individual variability in backchannel feedback.

6.3.2 Listen's Personality Traits

Here we examine how listener personality traits might impact backchannel production. Previous research has demonstrated the close relationship between backchannel feedback and personality. For example, (Chartrand & Bargh, 1999) found empathic individuals exhibit more mimicry behavior during interaction to a greater extent than not empathic individuals. (Borkenau & Liebler, 1992) showed that high extroversion is associated with greater level of gesturing, more frequent head nods, and a great speed of movement. Table 6.1 lists several individual traits of the coders that we currently investigate.

Except gender, every other attribute is measured using standard psychometric scales. In this way, each coder can be characterized by a set of values. For each attribute, coders are grouped as follows: we calculate the mean μ and the standard deviation σ from all coders, those whose values are less than $\mu - \sigma$ are grouped into the *low_group*, while those whose values are larger than $\mu + \sigma$ are grouped into the *high_group*. We compute the average number of feedback for each group as follows:

1. For each group, a histogram is computed by using the data from the coders in the corresponding group;

2. We sum up the histogram computed in the first step to get the total number of backchannel feedback. The average number of feedback is calculated by dividing the total number by the number of coders in the corresponding group.

Trait	Low	High	p-value
Extroversion	172.61	173.51	$p = 0.88$
Agreeableness	**166.59**	**190.72**	**p=0.004**
Conscientiousness	**178.76**	**207.72**	**p=0.0002**
Neuroticism	182.02	180.02	$p = 0.67$
Openness	**171.52**	**203.58**	**p<0.0001**
Self-consciousness	**173.71**	**210.17**	**p=0.0008**
Other-consciousness	171.04	205.76	$p = 0.012$
Shyness	178.26	175.51	$p = 0.66$
Self-monitor	**166.97**	**206.03**	**p<0.0001**
Parasocial Experience	**163.50**	**203.06**	**p<0.0001**
Gender(F/M)	182.71	181.73	$p = 0.69$

Table 6.2: Compare the average number of feedback between the *low_group* and *high_group* with respect to each attribute.

The average number of feedback is computed for every video. Finally, t-test is used to find whether there is significant difference between the *low_group* and *high_group*. The result is summarized in Table 6.2.

We highlight the rows where significant differences are found between the *low_group* and *high_group* regarding the corresponding attribute. Usually, when running t-test, we consider the difference is significant if p-value is less than 0.05. However, since we have a relatively large number of hypotheses in this test, the likelihood of witnessing a rare event also increases. According to Bonferroni Correction (Bonferroni, 1935), we should test each individual hypothesis at a statistical significance level of 1/n times (i.e. here n = 11) what it would be if only one hypothesis were tested. Therefore, we only consider the difference is significant when p-value is less than 0.0045.

In line with previous research, our consensus data shows similar impact of personality on backchannel feedback. For example, (Chartrand & Bargh, 1999) found empathic individuals exhibit more mimicry behavior during the interaction to a greater extent than not empathic individuals; accordingly, our data suggests that agreeableness, conscientiousness, and openness all have similar influence on the number of feedback. This result matches (Chartrand & Bargh, 1999) because (Del Barrio et al., 2004) showed

significant positive correlation between empathy and agreeableness, conscientiousness, and openness. Interestingly, the expected significant difference between extroverted and introverted participants does not show up. We will discuss it later in Chapter 7. Self-monitoring is a kind of psychology measurement of the ability of being aware of the social situation and monitoring themselves accordingly. People who can closely monitor themselves often behave in a manner that is highly responsive to social cues. Similarly, our result indicates that the people who are good at self-monitoring can provide significantly more feedback during the interaction. (Hartmann & Goldhoorn, 2010) suggested Parasocial Interaction Scale to measure parasocial experience when interacting with media parasocially, and they found stronger parasocial experience resulted in higher commitment to social norms and a greater enjoyment of the exposure situation. Similarly, we find that the participants who have better parasocial experiences give significantly more backchannel feedback than those who have not.

6.4 Conclusion

From the analysis on both speaker side and listener side, we show that coders depend on almost the same subset of speaker features to provide backchannel feedback, indicating there may be some consistent backchannel feedback rules; while the significant individual variability of backchannel feedback may be due to the individual differences, such as personality traits and parasocial experience, among those coders. By using the behavior data from different groups (e.g. the most and least conscientious group), it is possible to train different prediction models for virtual humans with different personality traits. Such analysis can only be done by using PCS framework because (1) it breaks down the mutual-contingency in face-to-face interaction and (2) it allows much faster and more efficient data collection method. We collected about 550 listeners' behavior data in just several weeks, and these data needs no manual annotation. Only with such amount of data can we observe the significant individual differences.

Chapter 7

Naturalistic Behavioral Measurement

7.1 The Problem

As mentioned before, the PCS framework consists of five elements: media, an interactional goal, target behavioral responses, target population and a measurement channel. It enables crowd-sourced responses from some selected population to the fixed media in accordance with some specified interactional goal. Previous chapters have demonstrated the generality of PCS by varying the interactional goal, target behavioral responses, and target population:

- Interactional Goal: create rapport (Chapter 4) or take the conversational turn (Chapter 5);
- Target Behavioral Responses: backchannel feedback (Chapter 4) or turn-taking (Chapter 5);
- Target Population: different personality groups (Chapter 6).

This chapter examines alternative procedures for measuring parasocial responses. In previous work, PCS coders simply pressed a button to signal behaviors. Here we examine a more naturalistic approach. Specifically, we ask participants to act as if they were in a real conversation (e.g., smile if they feel like smiling) and record coders' nonverbal responses. Indeed, the original research on parasocial interaction emphasized

that people can easily engage in this "as if" game and produce natural conversational behaviors (Masters & Sullivan, 1990) (Sundar & Nass, 2000) (Levy, 1979).

There are several differences between using keyboard and using natural behaviors:

1. By using camera as the measurement channel, it is possible to measure multiple behaviors at the same time by videotaping the participant's behavior. This is difficult for keyboard, since pressing different buttons for different behaviors is likely to place too much cognitive load on the participants.

2. Using keyboard demands an explicit conscious decision from a coder whereas actual nonverbal behaviors are typically implicit and automatic. Thus it is possible that keyboard only assesses their beliefs about the behavior rather than the actual behavior.

3. By using keyboard as the measurement channel, we only record the starting time of the behavior; while by using camera as the measurement channel, we can capture not just the presence but something about the characteristics of the behavior, including for example its duration and intensity.

In this work, we focus on listener backchannel feedback. As mentioned in point 1, we simultaneously measure listener's head nods, headshakes, and smiles as we change the measurement channel from keyboard to camera. To examine point 2, we will compare the backchannel feedback data collected via keyboard with the data collected via naturalistic responses to investigate whether or not there is significant difference between them and what may cause such difference. We leave point 3 to future work.

7.2 Data Collection and Annotation

7.2.1 Data Collection

The media used in this study is a subset of 8 randomly selected videos from the 24 videos used in Chapter 6. Each video features a speaker telling two stories. To select

Figure 7.1: An example of the parasocial interaction. The participant (right side) interacted with the speaker video (left side) parasocially, and her nonverbal behaviors were recorded by a camera. In this example, the speaker paused and tried to remember the details of the story he was supposed to tell. He had an embarrassed smile because it took him a relatively long time, and the participant smiled back, probably to reassure him. Although the participant was aware that the interaction was not real, she displayed such facial expressions seemingly automatically. We use the OKAO vision system from Omron Inc (Lao & Kawade, 2005) to detect smiles, which can infer the level of smiling (continuous value from 0 to 100)

PCS coders, we recruited 28 participants via www.craigslist.com from the general Los Angeles area. Before beginning the study, the participants were required to read the instructions and ask questions about anything they do not understand. They were informed beforehand that they would be videotaped and instructed to pretend to show interest and create a sense of rapport with the videotaped speaker by showing backchannel feedback such as head nod, head shake, and smile and so on. They first finished the 90-item personality inventory (see Appendix A) described in Section 6.1. Next, they watched the 8 speaker videos in sequence in a random order. Their nonverbal responses to the speakers were recorded by our PCS data collection tool (the measurement channel is a camera) described in Section 3.4. At the end of the study, the participants were debriefed and each was paid 35 USD. Figure 7.1 shows an example of the parasocial interaction.

7.2.2 Results

At the end of the study, we collected 28 parasocial responses to each of the 8 speaker videos. Participants produced wide variety of behaviors including both generic feedback(e.g. head nod) and specific feedback (e.g. headshake and expressive facial expressions) (Bavelas et al., 2000). The specific feedback is always triggered by certain events mentioned in the conversation. In this study, we chose head nods, headshakes and smiles, which are the mostly occurred behaviors in our dataset, as the target behaviors. We will leave other common behaviors, such as frowns, to the future work.

7.2.3 Annotation

We are interested in three kinds of nonverbal behaviors: head nods, headshakes and smiles. A mix of manual and automatic annotation techniques were used to extract target behavioral responses from the recorded videos. To annotate head nods and headshakes, we recruited naïve annotators from Amazon Mechanical Turk. To facilitate the annotation work, we developed a web-based annotation tool (as shown in Figure 7.2) that helps annotators go through the videos and annotate target behaviors efficiently. Each annotator examined seven videos in sequence and only annotated only a single type of behavior at one time. Each video was annotated by two independent annotators and each were paid 3 USD.

Smiles were annotated automatically using the OKAO vision system. OKAO has been previously used with some success to measure facial expressivity (Gratch & Boberg, 2013). Briefly, it uses computer vision techniques to identify 16 facial landmarks. From this, it derives a variety of facial pose estimates including a smile intensity ranging from 0 (no smile) to 100 (full smile). By setting the threshold to 50, we can reliably determine whether the participant is smiling or not.

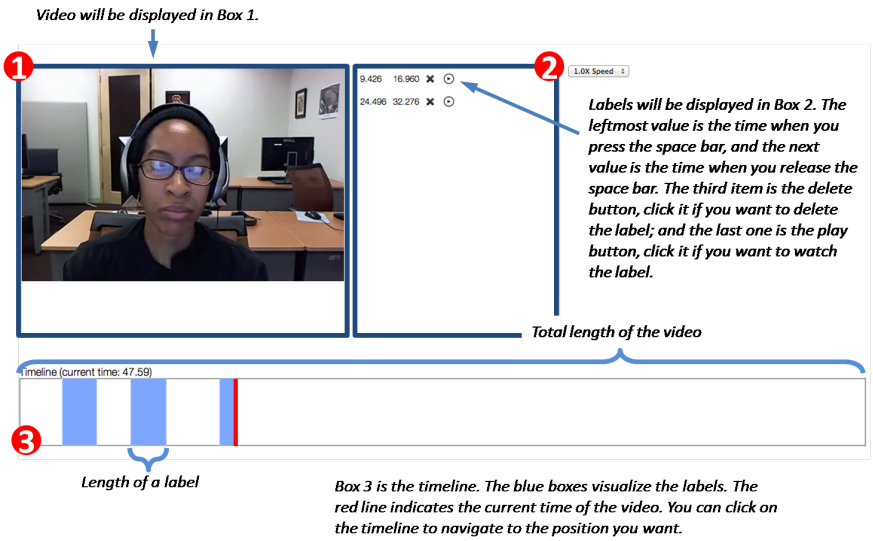

Figure 7.2: This is the annotation interface. Coders press the space bar to start loading a video, and the loading progress will be shown in Component 1. After the video is loaded, coders press the space bar to start playing the video. At the beginning of the target behavior, coders press the space bar and hold it, and release the space bar when the target behavior ends. After finish labeling the video, coders can adjust the labels by dragging on their boundaries.

7.3 Data Analysis

7.3.1 Comparison between different implementations of measurement channel

As mentioned before, changing measurement channel may affect the PCS data. In this section, we want to understand the differences between the data collected via keyboard and camera.

For each of the 8 speaker videos, we aggregate the nonverbal behaviors (head nods, headshakes, and smiles) from all 28 participants to build the consensus view. Figure 7.3 shows an example of the consensus of head nod, headshake and smile. The peaks found in both the consensus of head nod and headshake are potential backchannel opportunities. But headshakes occur a lot less than head nods do, and they are usually associated with semantically negative events in the speech. There is a noticeable jump

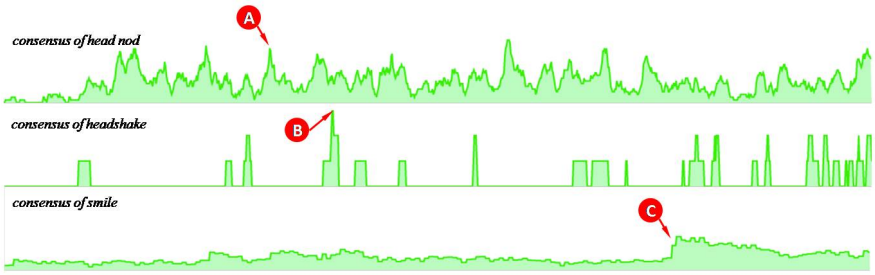

Figure 7.3: An example illustrates the consensus of head nod (top), headshake (middle) and smile (bottom). The speaker video is from a sexual harassment training course. At point A, the speaker said "It's from Rick in accounting or Rick in legal or something, and [pause], he said 'oh, no' ..." and the nod is most likely to occur during the pause; at point B, the speaker said "and she says, you know, 'I gave him a ride once when his car broke down, now he won't leave me alone, it's been five weeks, I always get these emails, and e-cards, and he won't leave me alone'..." and the shake is most likely to occur when the speaker described the fact that Rick kept bothering the lady; at point C, the speaker said "and then she says 'oh, and next, I am gonna need a foot massage', and then she shuts the blinds..." and the smile is most likely to occurs after mentioning the foot massage.

in the consensus of smile, where the speaker says the most dramatic part of the story. Interestingly, this phenomenon is observed in all 8 videos.

To compare the parasocial consensus data, we selected the same videos used in both studies (Chapter 6 and 7). Previously, we asked the participants to press a button whenever they felt like to nod. Here, we use the annotations of head nods as well. We call the consensus view built from the data measured by camera the "naturalistic consensus", and the consensus view built from the data measured by keyboard the "keyboard consensus". Besides the differences mentioned before between the naturalistic and keyboard data, there are two other main differences between the two consensus views: (1) When the measurement channel is keyboard, we collect data from 350 participants; while the number is only 28 when the measurement channel is camera so we can expect more variance in the naturalistic data. (2) When the measurement channel is keyboard, participants press a button to give feedback. Therefore, each feedback is associated with exactly one backchannel opportunity; however, with naturalistic measurement,

Interaction Index	Correlation Coefficient
Video 1	0.73
Video 2	0.64
Video 3	0.56
Video 4	0.71
Video 5	0.78
Video 6	0.50
Video 7	0.71
Video 8	0.62
Overall	**0.66**

Table 7.1: The correlation coefficient between the "naturalistic consensus" and "keyboard consensus" for each video.

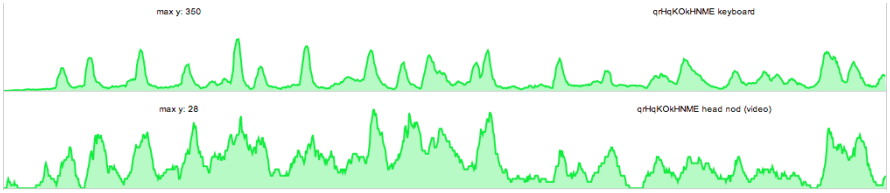

Figure 7.4: In this example, the top one is the consensus view of the head nods built from the "keyboard" channel; and the bottom one is the consensus view of the head nods built from the "naturalistic" channel.

participants give feedback by nodding, shaking or smiling, and these nonverbal behaviors usually last for a period of time (e.g. several seconds) which may be associated with several backchannel opportunities. Despite of these differences, we compute the correlation coefficients between the two consensus views over all interactions (as shown in Table 7.1). The average value is rather high (r=0.66) and the standard deviation is 0.09.

The numbers suggest that the consensus view not change significantly even if we change the measurement channel from keyboard to camera. Figure 7.4 shows an example illustrating the consensus view of the head nods built from the "keyboard" channel and the "naturalistic channel".

We then compare the speaker feature set that triggers listener backchannel feedback in both cases. We use the same approach as described in Section 6.3 to measure the

Keyboard (550 participants)		Camera (28 participants)	
Feature Name	The number of occurrence	Feature Name	The number of occurrence
Speaker eye gaze	546/550=0.99	Speaker eye gaze	28/28=1.0
Speaker head nod	404/550=0.73	Pause	13/28=0.46
Pause	358/550=0.65	Speaker head nod	10/28=0.36

Table 7.2: The top three speaker features that the listeners rely on most to decide when to give backchannel feedback.

importance of every speaker feature, that is, which speaker features do the listeners rely on most to decide when to give backchannel feedback. For each listener, we select the top five most important speaker features. Over all listeners, we count the number of occurrence of each speaker feature appeared in the top five list, and the result is shown in Table 7.2. This suggests participants rely on very similar features (speaker eye gaze, speaker's head nod and pause) to decide whether and when to give backchannel feedback regardless of the measurement channel.

7.3.2 How listener's personality traits influence the behavior

To examine the impact of personality on coder behavior, we calculated personality scores from the 90-item personality inventory as described in Section 6.1. For each personality subscale (e.g. extroversion), we performed a median split and grouped participants into a high and low scoring group. For example, those scoring below the median for the 8-item subscale of extroversion would be combined into an introverted group whereas those scoring above the median would be combined into an extroverted group (as in Section 6.3). We then contrasted the consensus of the behaviors of these two partitions. The results are shown as below.

Usually, when running t-test, we consider the difference is significant if p-value is less than 0.05. However, since we have a relatively large number of hypotheses in this

Trait	Low	High	p-value
Extroversion	**39.7**	**64.1**	**p=0.002**
Agreeableness	65.0	65.7	p=0.92
Conscientiousness	52.0	68.2	p=0.026
Neuroticism	**68.9**	**31.8**	**p=0.005**
Openness	**33.3**	**59.3**	**p=0.003**
Self-consciousness	103	42.8	p=0.011
Other-consciousness	**31.6**	**81.4**	**p=0.0001**
Shyness	74.5	65.8	p=0.17
Self-monitor	77.0	58.0	p=0.016

Table 7.3: Compare the average number of head nods between the *low_group* and *high_group* with respect to each attribute.

Trait	Low	High	p-value
Extroversion	2.2	1.2	p=0.017
Agreeableness	1.2	1.9	p=0.10
Conscientiousness	1.7	1.8	p=0.92
Neuroticism	2.0	1.2	p=0.12
Openness	0.375	1.375	p=0.027
Self-consciousness	0.7	3.5	p=0.21
Other-consciousness	2.25	0.46	p=0.11
Shyness	1.33	0.81	p=0.48
Self-monitor	2.41	1.21	p=0.058

Table 7.4: Compare the average number of headshakes between the *low_group* and *high_group* with respect to each attribute.

Trait	Low	High	p-value
Extroversion	0.13	0.21	p=0.018
Agreeableness	**0.15**	**0.26**	**p=0.001**
Conscientiousness	**0.13**	**0.40**	**p=0.0002**
Neuroticism	**0.40**	**0.11**	**p<0.0001**
Openness	0.14	0.12	p=0.3
Self-consciousness	**0.06**	**0.27**	**p=0.0001**
Other-consciousness	**0.08**	**0.04**	**p=0.006**
Shyness	0.29	0.16	p=0.03
Self-monitor	**0.02**	**0.21**	**p<0.0001**

Table 7.5: Compare the amount of smiles between the *low_group* and *high_group* with respect to each attribute. The number is calculated by dividing the duration of the listener's smiling by the duration of the whole interaction.

test, the likelihood of witnessing a rare event also increases. According to Bonferroni Correction (Bonferroni, 1935), we should test each individual hypothesis at a statistical significance level of 1/n times (i.e. here n = 9) what it would be if only one hypothesis were tested. Therefore, we only consider the difference is significant when p-value is less than 0.0056. The rows where significant differences are found are highlighted. Similar to the results shown in Section 6.3, listener's personality traits have significant influence on the number of head nods (as shown in Table 7.3). However, it is important to point out that the sets of personality traits where significant differences are found are different from those when the measurement channel is keyboard. We'll discuss it later. Besides head nods, we also examined headshake and smile. As Table 7.4 shows, headshake rarely happens during the interaction. In our data, we find that its occurrence is always associated with the semantically significant events (usually negative events) in the speech. Listener's personality traits don't have significant influence on headshake. However, smiling, although a kind of specific feedback, is significantly affected by the listener's personality traits (as shown in Table 7.5). This is similar to the results we found in literature. For example, (Shiota et al., 2006) showed that more extraverted, more conscientious, more agreeable, and less neurotic people are more likely to experience joy.

Although no significant difference is found (as described in Section 7.3.1) between the consensus views and the speaker feature sets that trigger listener backchannel feedback, different personality groups behave differently when the measurement channel changes from keyboard to camera (as shown in Table 7.3 and Table 6.2). For example, when the measurement channel is keyboard, there is no significant difference between the *low_group* and *high_group* of extroversion; while the significant difference shows up when the behaviors are measured in a naturalistic fashion. On the other hand, there is significant difference between the *low_group* and *high_group* of agreeableness when the channel is keyboard; whereas this difference fails to materialize with naturalistic measurement.

Correlation Coefficient	Extroversion	Agreeableness	Neuroticism	Shyness
Number of Headshakes	0.63	0.75	-0.87	-0.81

Table 7.6: The correlation coefficients between speaker personality traits and the number of headshakes of crowds.

These personality results illustrate that the different forms of measurement (deliberate keyboard presses vs. naturalistic parasocial enactments) can affect the results of the PCS approach and hints that different measurement channels may index different sources of knowledge about social interaction. For example, in that naturalistic measurement requires a performance on the part of the coder, it will likely be subject to the same biases that impact people's ability to perform in social situations (e.g., extroverts are generally more expressive than introverts).

7.3.3 How speaker's personality traits influence the behavior

Each speaker video was watched by all participants, and we call the aggregation of their behaviors the "crowds' behavior". The crowds' behavior is different when interacting with different speaker videos. A natural follow-up question is whether or not the speaker's personality traits affect the crowds' behavior. In a previous study (Gratch et al., 2007), we measured each speaker's personalities. We compute the correlation coefficients between the speaker's personality measurements and the crowds' behavior (i.e. the number of head nods, the number of headshakes, and the amount of smiles). The results suggest that speaker's personality does not affect the listeners' head nods or smiles. Listeners' smiles are highly correlated with the speakers' smiles (correlated coefficient = 0.80), indicating the listener was mimicking the speaker's smile. However, some of the speaker's personality measurements are highly correlated with the listeners' headshakes (as shown in Table 7.6).

The result shows that the number of headshakes is positively correlated with the speaker's extroversion and agreeableness measurements, and is negatively correlated

with the neuroticism and shyness measurements. In our task, headshake always indicates negative emotions towards what the speaker said. That is, if the speaker is more extroverted and agreeable, the listeners are more likely to express their negative emotions; however, if the speaker is more neurotic and shyer, the listeners tend to hide their negative emotions.

7.3.4 Predicting Personality from Parasocial Responses

There have been increasing interests in understanding how human beings encode personalities in and decode personalities from nonverbal behaviors (Biel et al., 2011) (Gifford, 1994). Our previous results have demonstrated that there is close relationship between personality and backchannel feedback. Here we investigate how well we can predict personality just from the listener backchannel feedback and how well we can explain the variability of listener backchannel feedback by only using the listeners' personality. We ran a stepwise linear regression analysis between backchannel feedback (including the number of nods, the number of shakes and the duration of smiles) and personality measurements (The analysis is done in SPSS using the "backward" method. That is, non-significant independent variables will be removed at each step. At the end, if there is no independent variables remaining, it means no significant relationship is found between the independent variables and the dependent variable).

First, we predict personality traits from the parasocial consensus. The dependent variable is each of the personality traits (e.g. extroversion), and the independent variables are the number of head nods, the number of headshakes, and the duration of smiles produced by PCS coders. We observed significant results for neuroticism and self-consciousness. Smile itself (correlation coefficient = -0.23) can predict about 12% of the variance of neuroticism (F=3.4, p=0.07); smile (correlation coefficient = 0.2) and nod (correlation coefficient = -0.17) together can predict about 20% of the variance of self-consciousness (F=3.11, p=0.062). Second, we run the same analysis reversely; that is, the dependent variable is the number of head nods, the number of headshakes and the

duration of smiles respectively, and the independent variables are the personality traits. We only observed significant result for smile. Self-consciousness (correlation coefficient = 0.868) and neuroticism (correlation coefficient = -0.658) can predict about 28% of the variance of smile (F=4.95, p=0.015). Together, this suggests we can intuit something about a speakers personality simply by looking at the responses of their conversation partner, although this relationship is rather modest.

7.4 Conclusion

In this chapter, we extended the PCS framework by adding camera as another implementation of measurement channel so that we can measure more than one target behavior at the same time. By grouping participants based on their personalities, we found that different personality groups behave significantly different in both quantities and types. By comparing the PCS data collected via keyboard with the data collected via camera, we showed that the consensus view keeps consistent, which demonstrates the flexibility of the PCS framework.

Chapter 8

Applying Parasocial Consensus Sampling: Improving Virtual Human System

In Chapter 4 and 5, we validated the PCS framework by applying it to model listener backchannel feedback and turn-taking behavior. In Chapter 6 and 7, we explored and extended the PCS framework and investigated how personality traits influence the individual's nonverbal behaviors in face-to-face interactions. The ultimate goal of PCS framework is to help build better nonverbal behavior models for virtual human and thus improve the virtual human systems. In this chapter, we integrate the PCS-data driven models into a virtual human system, and compared it with the Rapport Agent. Human subjects are asked to evaluate the agents regarding the correctness of their behaviors, the rapport they feel during the interaction and the overall naturalness.

8.1 Background: Rapport Agent

Inspired by the three-factor theory of rapport (Tickle-Degnen & Rosenthal, 1990), Rapport Agent (Gratch et al., 2007) was designed to establish rapport with human participants by providing contingent nonverbal feedback while the participant is speaking. The initial system focused on a "quasi-monolog" paradigm, where a human speaker (the narrator) retells some previously observed series of events (e.g., the events in a

Human Speaker Behavior	Rapport Agent Response
Lowering of pitch	Head nod
Raised loudness	Head nod
Speech disfluency	Posture/Gaze shift
Posture shift	Mimic
Gaze away	Mimic
Head nod/shake	Mimic

Table 8.1: Rapport Agent Behavior Mapping Rules

recently-watched video) to a non-speaking but nonverbally attentive agent. In designing the Rapport Agent, Gratch and colleagues extracted a small number of simple rules (as shown in Table 8.1) from social science literature.

To produce listening feedback, the agent first collects and analyzes the speaker's upper-body movement and voice. To detect features from the participants' movement, it uses Waston (Morency et al., 2005) to track the head position and orientation. With the head tracking data, it can detect head gestures, posture shifts and gaze direction. Acoustic features are derived from properties of the pitch and intensity of the speech signal, using a signal processing package, LAUN. The recognized speaker's features are then mapped to reactions through a set of authorable mapping rules. These reaction animations are passed to the SmartBody (Thiebaux et al., 2008) animation system using the Behavior Markup Language (BML); and finally, the animations are rendered by a commercial game engine and displayed to human users. The animations that the Rapport Agent can perform are relatively simple, such as two continuous nods with equal amplitude and posture shifts.

The agent has been applied in a series of empirical studies to investigate how people are influenced by such computer-generated behaviors. In these studies, human participants sit in front of the Rapport Agent and are prompted to either retell some previously experienced situation (monologue) or interviewed by the agent to answer some predefined questions (interview). After the interaction, participant's rapport is assessed by a variety of subjective and behavioral measures. These studies showed that by interacting

with the Rapport Agent, people have: greater feelings of self-efficacy (Kang et al., 2009), less tension (Wang & Gratch, 2010) and less embarrassment (Kang et al., 2009), greater feelings of rapport (Wang & Gratch, 2010), a greater sense of mutual awareness (Von der Putten et al., 2009), and greater feelings of trustworthiness (Kang et al., 2009). The contingent nonverbal feedback of the Rapport Agent also changes participants behavior. Behavioral effects include: more disclosure of information including longer interaction time and more words elicited (Gratch et al., 2007) (Gratch et al., 2006), more fluent speech (Gratch et al., 2007) (Gratch et al., 2006), more mutual gaze (Wang & Gratch, 2010) and fewer negative facial expression (Wang & Gratch, 2009).

Although it has been demonstrated effective in many studies, the current models and behaviors of the Rapport Agent have limitations with regard to the three-factor theory of rapport.

Mutual Attention and Coordination: Tickle-Degnen and Rosenthal emphasized that, with rapport, participants fall into a cohesive, unified pattern of behavior arising through close attention to and tight-coordination of nonverbal signals. The Rapport Agent attempts to realize these two factors by attending to human nonverbal cues (e.g., gestures and prosodic signals) and utilizing them to coordinate its responses (such as backchannel feedback). However, there are reasons to suspect the agents attention and coordination could be significantly improved. Like many virtual agents, the Rapport Agents behavior is driven by general rules derived from social science literature, in contrast to more recent approaches (Morency et al., 2008) (Lee & Marsella, 2009) (Jonsdottir et al., 2008) that attempt to learn behaviors directly from large datasets. Although based on human-behavioral studies, such "literature-based" rules are often intended to make general theoretical points rather than to directly drive virtual agents. Further, such rules are often generated in a variety of social contexts that may differ considerably from the situations to which the Rapport Agent has been applied. Consequently, such rules are unlikely to capture the subtlety in both timing and realizations of nonverbal behaviors.

Positive Emotion Communication: A third component of rapport relates to sense of emotional alignment and positivity that participants experience in the course of rapportful interactions. Nonverbally, this feeling arises from the positive and empathetic expression of emotion. Other research on rapport has emphasized the equal importance of verbal expressions of emotion, for example, through the reciprocal self-disclosure of hopes and fears (Moon, 2000). Thus, a clear limitation of the Rapport Agent is its inability to engage in emotional communication, both verbally and nonverbally, a point highlighted in some of the evaluations of the system (Wang & Gratch, 2009).

8.2 Virtual Rapport 2.0

Virtual Rapport 2.0 improves over the previous work by directly addressing the main limitations of the Rapport Agent. We enhance mutual attention and coordination by applying data-driven approaches to build context-specific (i.e. the same context where the virtual human is deployed) response models, which better model the subtlety of timing and realization of nonverbal behaviors. By integrating affective information and strengthening reciprocity, we also enable the virtual human to communicate positive emotions both verbally and nonverbally.

8.2.1 Enhanced Mutual Attention and Coordination: Data-driven Approach

To enhance mutual attention and coordination, we learn models to predict backchannel and turn-taking opportunity points from the human behavior observed in the same dyadic conversation settings in which the Rapport Agent is intended to be used. By using such contextually-appropriate data, and employing more sophisticated techniques than are typically used in the social sciences, we expect to better model the subtlety and variability in both timing and realizations of the nonverbal behaviors. To collect human behavior data for the response models to learn from, we apply Parasocial Consensus Sampling framework (as described in Chapter 4 and 5).

For the backchannel prediction model, we use the learning technique described in Section 4.2.3. Different from the conventional Conditional Random Field (CRF), which uses the forward-backward inference engine that requires the full sequence available (i.e. offline processing), we implement the real-time CRF model using the forward-only inference (Murphy, 2002) so that it can make predictions in real-time. The output of CRF indicates the likelihood of giving backchannel feedback. By setting a threshold, we predict the time of feedback by comparing the output with the threshold. The CRF model predicts when to give feedback and how to give such feedback is learned from actual listeners' behaviors. We found the typical styles of head nods from the listeners' behavior in the Rapport data set [1] in two steps. First, the listeners' head positions were tracked by Watson (Morency et al., 2005) and converted to frequency domain by Fast Fourier Transform. Then K-means (k=3) was applied to cluster all head nods to find typical styles, which are implemented in Behavior Markup Language (BML):

- Small and continuous head nod: four continuous small nods with decayed amplitudes and speed;
- Normal nod: two continuous head nods with decayed amplitudes and equal speech;
- Single nod: one slow head nod;

In the current implementation, we randomly choose one of the three styles when it is proper to give backchannel feedback.

We build the turn-taking model based on the analysis of the parasocial consensus data collected in Section 5.1. The model can be summarized as follows:

(1) When the pause duration is longer than 1.5s and the speaker has been looking at the virtual human for more than 1.0s, it is a turn-taking place;

(2) When the speaker is looking away, the virtual human will wait until the speaker looks back and then go to (1);

[1] Datasets are available for research purpose at rapport.ict.usc.edu

(3) When head nod co-occurs with a longer-than-1.5s pause, the virtual human will take the turn 200ms after the end of the head nod;

After the virtual human takes a turn, the human speaker is allowed to interrupt him. The virtual human will stop and yield his turn to the human speaker by saying "I'm sorry, keep going" with a regretful facial expression.

8.2.2 Enhanced Positive Emotion Communication: Affective Response and Reciprocity

The feeling of positivity, which is important in establishing rapport in initial encounters, can be enhanced by communicating positive emotions both nonverbally and verbally.

Facial expression is an important channel to convey positive emotion nonverbally. Our previous study (Section 7.3.3) showed that the listeners tend to mimic the speaker's smile. We implement the mimicking model as follows: the OKAO vision system by Omron Inc (Lao & Kawade, 2005) is used to track the facial feature points of the human speaker in real-time, from which it infers the level of smiling (continuous value from 0 to 100). By setting the threshold to 50, we can reliably determine whether the human speaker is smiling or not. When there is a backchannel opportunity and the human speaker happens to smile at the same time, the virtual human will display backchannel feedback with a smiling face.

Recent research by (Kang & Gratch, 2011) has emphasized some simple strategies for conveying positive feelings verbally. In her study, the interviewee discloses more intimate information if the interviewer (virtual human) discloses itself first. The mutual self-disclosure, or reciprocity, positively affects the human user's social attraction to the virtual human. In our system, we follow the same strategy of strengthening reciprocity. Before the virtual human asks its human partner questions, he will first disclose the information about himself, that is, sharing some of his autobiographical back story. For example, instead of simply asking "how old are you?", the virtual human says "I was created about three years ago. How old are you?".

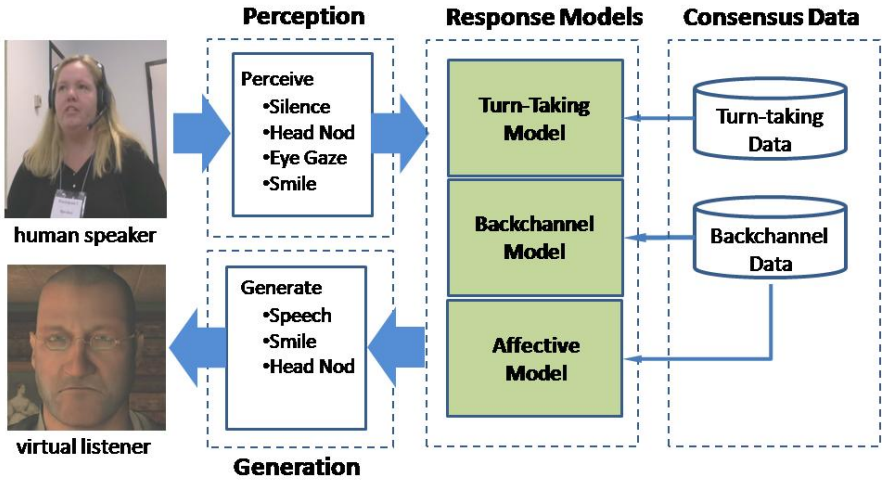

Figure 8.1: System Architecture of Virtual Rapport 2.0: The perception module detects human behavior (e.g. silence in speech, nod, gaze aversion, and smile) in real-time; then the data-driven based response models take these feature as input and predict the timing of backchannel feedback and turn-taking, and the affective response; finally, the generation module generates speech and animations (e.g. smile and nod) to display to the human speaker.

8.2.3 The Virtual Rapport 2.0 System

With the enhancement in the three factors of rapport, we build the new system as shown in Figure 8.1. The system consists of three main parts: (1) perception, which detects the audiovisual features of human speakers in real-time; (2) response models, which predict the timing of backchannel feedback and end-of-turn and the affective state (e.g. smile or not); and (3) generation, which animates the virtual human's behavior such as head movements and facial expression.

Perception: The four main audiovisual features extracted in real-time are silence, head nod, eye-gaze (looking at listener or not) and smile. The audio feature detector extracts intensity from the raw signal every 100ms using the signal processing package, Praat [2]. With intensity information, it outputs a binary feature, speech or silence, every 100ms. The visual feature detector (a confidential commercial product) tracks

[2]software available on http://www.fon.hum.uva.nl/praat/

the position of face and facial feature points, the direction of eye-gaze and the smile level. With this information, it outputs visual features indicating the human is nodding or not, looking away or not and smiling or not.

Response Models: Based on the perceived audiovisual features, the backchannel, turn-taking and affective models decide in real-time the most appropriate responses. All three models take advantage of the PCS data-driven approach. These responses also take into account the agent state (e.g. whether the virtual human is holding the turn or not). For example, if the virtual human is holding the turn, the output from backchannel model is ignored. The backchannel model takes silence and eye-gaze as input, the turn-taking prediction model uses features such as silence, eye-gaze, and head nod and the affective model takes smile as input.

Generation: The output from the response models drives the virtual human behaviors. For example, if the human speaker smiles, the virtual human will smile as well when giving the backchannel feedback. These animations are first implemented in BML and then sent to an action scheduler module, which keeps track of the duration of each animation. If the current animation has not completed yet, new animations will be ignored. The BMLs are passed to the animation system, Smartbody, which is a virtual human animation system designed to seamlessly blend animations and procedural behaviors. Finally, animations are rendered by a commercial game engine, Gamebryo, and displayed to users.

8.3 Evaluation

To evaluate the performance of our virtual human, we conducted a subjective evaluation to compare it with the Rapport Agent along four dimensions: rapport, overall naturalness, backchannel feedback and end-of-turn prediction.

We guided human participants to interact with both virtual humans one after the other, where the virtual human acts as an interviewer and steps through a series of questions one by one, and the human participant acts as the interviewee. For each

interaction, we used different question sets derived from (Moon, 2000). The order of virtual humans and question sets were randomized in the experiment. After each interaction, the human participant was asked to assess the virtual human's performance.

In a within-subject design, 21 participants were recruited to evaluate both virtual humans. Before the experiment started, the participant was required to read the instructions and ask questions about anything they do not understand. They were told "Your partner will ask you several questions and your task is to answer as best as you can. For each question, please try to answer in at least one or two sentences. You partner will listen when you answer. Please do not ask your partner questions. Your partner does not know who you are, your behavior will not be recorded and your identity will be kept anonymous". After interacting with each virtual human, participants were asked to assess the virtual human along four dimensions:

Rapport:

We measure rapport using the 5-item social presence scales suggested in (Bailenson et al., 2001).

- *I perceive that I am in the presence of another person in the room with me. (1(strongly disagree) - 7(strongly agree))*
- *I feel that the person is watching me and is aware of my presence. (1(strongly disagree) - 7(strongly agree))*
- *The thought that the person is not a real person crossed my mind often. (1(strongly disagree) - 7(strongly agree))*
- *The person appears to be sentient (conscious and alive) to me. (1(strongly disagree) - 7(strongly agree))*
- *I perceive the person as being only a computerized image, not as a real person. (1(strongly disagree) - 7(strongly agree))*

Overall Naturalness

- *Do you think the virtual agent's overall behavior is natural? (1(not natural at all) - 7(absolutely natural)).*

Backchannel Feedback

- *Precision: How often do you think the virtual human generated feedback at inappropriate time? (1(all the time) - 7(never inappropriate))*
- *Recall: How often do you think the virtual human missed feedback opportunities? (1(always miss) - 7(never miss))*

End-of-Turn Prediction

- *Correct time: How often do you think the virtual human ask the next question too early? (1(always) - 7(never))*
- *In time: How often do you think the virtual human ask the next question too late? (1(always) - 7(never))*

When the experiment was done, the participant was forced to choose the one s/he likes better.

The results are summarized in Figure 8.2. In each figure, the left bar is for Rapport Agent and the right bar is for Virtual Rapport 2.0. The star (*) means there is significant difference between the versions under the bracket.

Rapport: The answers of the five-item social presence scales are highly correlated with each other (the Cronbach's alpha is 0.9). Therefore, we average them into one scale. It is 2.6 for Rapport Agent and 3.84 for Virtual Rapport 2.0, and the difference is significant ($p<0.01$).

Overall Naturalness: The overall naturalness for Rapport Agent is 2.55, while it is 4.5 for Virtual Rapport 2.0, and the difference is significant ($p<0.01$).

Backchannel Feedback: For the precision question, the mean value of Rapport Agent is 3.6 while it is 5.25 for Virtual Rapport 2.0; for the recall question, the mean value

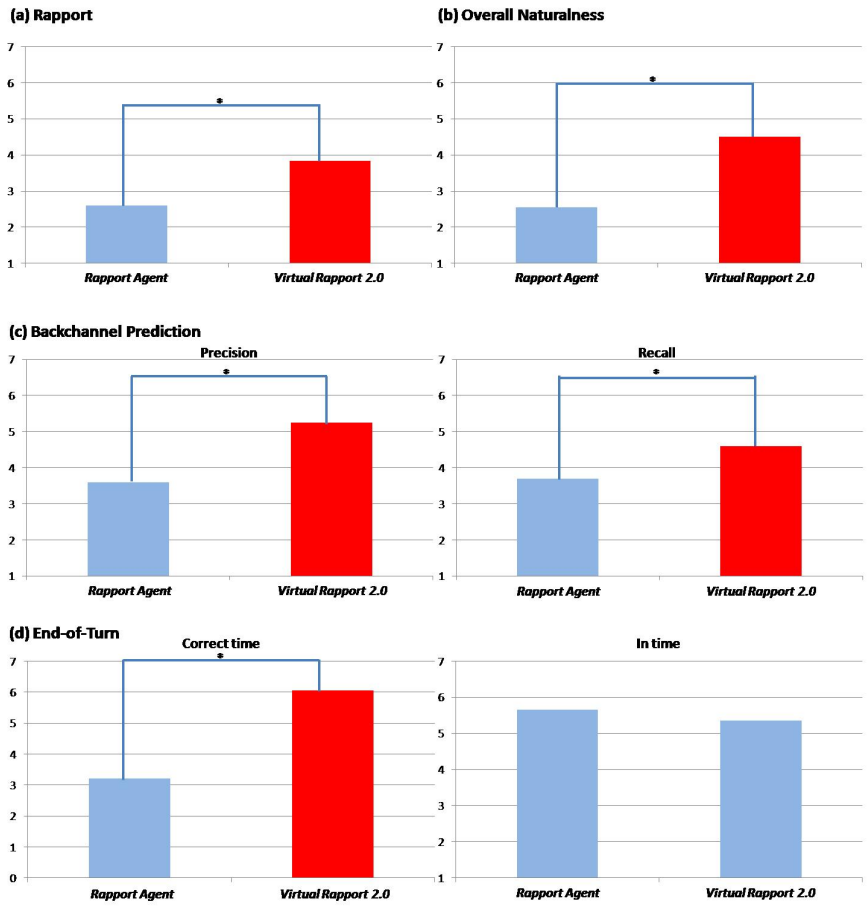

Figure 8.2: The comparison of subjective evaluation results between Rapport Agent and Virtual Rapport 2.0. Virtual Rapport 2.0 is significantly better than Rapport Agent in predicting the timing of backchannel feedback (c) and end-of-turn (d); it is also significantly better than Rapport Agent in overall naturalness (b). Therefore, Virtual Rapport 2.0 creates much stronger feeling of rapport (a) than the Rapport Agent does.

of Rapport Agent is 3.7, while it is 4.6 for Virtual Rapport 2.0. Virtual Rapport 2.0 is significantly better ($p<0.05$) than Rapport Agent in both questions.

End-of-Turn Prediction: For the correct time question, the mean value of Rapport Agent is 3.2, while it is 6.05 for Virtual Rapport 2.0, and there is significant difference ($p<0.01$) between the two; for the in time question, the mean value of Rapport Agent is 5.65 and it is 5.35 for Virtual Rapport 2.0, and the difference is not significant.

In the force-choice task, among all 21 participants, 19 (90%) participants preferred our virtual human to Rapport Agent. From the results we see a great improvement over the original Rapport Agent:

- Regarding creating rapport, our virtual human is significantly better than the Rapport Agent. One of the main advantages of our virtual human is that it is based on models learned from parasocial consensus data. This innovative approach is reflected in all the response models. The data-driven approach promotes the feeling of mutual attention and coordination. Besides, the strengthened reciprocity and affective response communicate positive emotions both verbally and nonverbally.

- Regarding the timing, our new backchannel prediction model significantly outperforms the Rapport Agent's in both precision and recall, which indicates that our virtual human is more "in sync" with the human speaker during the interaction. Rapport Agent tends to take a turn (ask the next question) too quickly. Such turn-taking strategy is likely associated with negative and strong personality (ter Maat et al., 2010), which is opposite to the goal of creating rapport.

- Regarding the behaviors, compared to Rapport Agent, our virtual human has a richer set of behaviors that is correlated with creating rapport. For example, the virtual human mimics the human speaker's smiles, it performs more natural head gestures and strengthens reciprocity by self-disclosure. All these improvements may explain the significant difference on the overall naturalness between our virtual human and the Rapport Agent.

8.4 Conclusion

This chapter describes our effort toward building a virtual human whose goal is to create rapport during interactions with human users. We applied data-driven approaches to build its response models based on parasocial consensus data, which enhances its mutual

attention to and coordination with human users. Besides, the strengthened reciprocity and proper affective responses improve its positivity toward the human users. By comparing with Rapport Agent, we found that our virtual human predicts the timing of backchannel feedback and end-of-turn more precisely, performs more natural behaviors and thereby creates much stronger feelings of rapport between users and virtual agents.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

Virtual humans are playing an important role in the advancement of today's immersive virtual worlds, including domains such as virtual training (Swartout et al., 2006), education (Rowe et al., 2010), and health care (Bickmore et al., 2010). One of the key challenges in creating virtual humans is giving them human-like nonverbal behaviors. However, modeling nonverbal behavior is hard. Regardless of years' research, it still poses problems. Although previous research mostly focused on the techniques for learning from data, there has been less attention to the data these systems learn from. Face-to-face interaction data is traditionally considered the gold standard, but it presents several drawbacks: (1) human behavior contains considerable variability and not all of them should be considered as the positive examples for the learning algorithms (2) face-to-face interactions are "co-constructed" so that it is hard to tease apart the causality (3) it is very expensive and time-consuming to collect and annotate face-to-face interaction data.

In order to solve the problems of the traditional face-to-face interaction data, this research work proposes a new methodology, called *Parasocial Consensus Sampling (PCS)*, to collect human nonverbal behavior data. The PCS framework is based on the theory of *parasocial interaction* introduced by Horton and Wohl (Horton & Wohl, 1956), in

which they argued that people exhibit a natural tendency to interact with media representation of people as if they were interacting with the actual person face-to-face. The basic idea of Parasocial Consensus Sampling is to have multiple independent participants experience the same social situation parasocially (i.e. act "as if" they were in a real dyadic interaction) in order to gain insight into the typicality of how individuals behave within face-to-face interactions. It has three advantages. (1) By aggregating the behavior data from multiple participants, we build the consensus view to describe the probability of how likely the behavior will occur over time, which better captures the variability in human behavior. (2) Since PCS allows participants to experience the same social situation, it breaks down the contingency of face-to-face interactions by holding one side of the interaction consistent, so that we are able to analyze the causality of the variability in human behavior. (3) The framework allows much larger scale and more efficient mechanism to collect human behavior data.

To validate this new methodology, we applied PCS to model listener backchannel feedback and turn-taking behavior in face-to-face interactions. We had participants interact with pre-recorded videos parasocially in order to achieve specific interactional goals (e.g. create rapport or take appropriate conversational turns). Their parasocial responses were used to drive the virtual human behavior directly and learn prediction models. The results suggested that (1) people are able to provide valid behavior data in parasocial interaction, (2) PCS data generates better virtual human behavior and (3) can be used to learn a better prediction model for virtual human.

The better performance of PCS data is probably due to the fact that it captures the variability of human behavior better than face-to-face interaction data does. We then continued to explore two other advantages of PCS framework. Specifically, we were crowdsourcing backchannel feedback to public via Amazon Mechanical Turk so that hundreds of participants can interact with the same speaker video parasocially. We grouped these participants based on their personality traits, and investigated how personalities can influence the nonverbal behavior in face-to-face interactions. PCS

facilitates such research because it helps tease apart the causality by holding some variable consistent (e.g. the speaker's behavior) and allows much more efficient data collection mechanism.

PCS framework consists of five elements: media, an interactional goal, target behavioral responses, target population, and a measurement channel. This thesis demonstrated the flexibility and generality of this framework by varying the interactional goal, target behavioral responses, target population and measurement channel.

- Interactional Goal: creating rapport (Chapter 4) and taking the conversational turns (Chapter 5);
- Target Behavioral Responses: Head nod, headshake, and smile (Chapter 4, 5, 6, 7);
- Target Population: Different personality groups (Chapter 6);
- Measurement Channel: Keyboard (Chapter 6) and camera (Chapter 7);

The ultimate goal of PCS framework is to help build better nonverbal behavior models for virtual human and thus improve virtual human systems. We integrated the PCS-data driven models into a virtual human system, and compared it with the Rapport Agent (Gratch et al., 2007) in real interactions. Human subjects were asked to evaluate the performance of each agent regarding the correctness of the agents behaviors, the rapport they feel during the interactions and the overall naturalness. The results suggest that the new agent predicts the timing of backchannel feedback and end-of-turn more precisely, performs more natural behaviors and thereby creates much stronger feeling of rapport between users and agents.

9.2 Limitations and Future Work

This dissertation intended to solve the problem of modeling human nonverbal behavior from the "data" perspective. Although we have already demonstrated the advantages

of our approach over the traditional face-to-face interaction data, it is clear that our work has some limitations:

1. Our work only focused on modeling generic nonverbal feedback. In fact, both generic and specific feedback (Bavelas et al., 2000) are important factors in creating a successful interaction. Especially, when virtual human researchers seek to realize more complex social scenarios, it is necessary for virtual humans to show specific nonverbal feedback to reflect its participatory role, interactional goal, and comprehension of the conversation. The PCS framework suggests a possible approach to help model such specific feedback, because it can conveniently create appropriate scenarios to collect behavioral data by simply customizing the framework in different ways (e.g. change the interactional goal).

2. Although we did not explicitly develop models of specific nonverbal feedback, Chapter 7 illustrates that the PCS framework can elicit specific feedback, for example, headshake. As mentioned in Section 7.3.2, the occurrence of headshake is always associated with semantically significant events (usually negative events) in the speech. By incorporating proper semantic features into the learning process, it should be possible to learn specific nonverbal feedback models. Although fully speech understanding is still an unsolved problem, previous work (DeVault et al., 2011) (Jonsdottir et al., 2007) has demonstrated that it is possible to extract useful semantic features (e.g. keyword spotting) from partial speech understanding. We should integrate such features into nonverbal behavior models in the future.

3. We did not take personality traits into account when building nonverbal behavior models. Chapter 6 and 7 clearly showed that different personality groups (e.g. extroverted versus introverted people) have significantly different behaviors. Ideally, by sampling from different populations, we can collect parasocial responses reflecting the characteristics of the corresponding group, which could help us build different nonverbal behavior models and thus create virtual humans with different

styles (e.g. different personalities).In future work, we will add personality as an extra parameter into the model. Subjective studies are needed to evaluate whether or not the virtual humans can successfully exhibit such different styles.

4. In Chapter 8, we compared our new agent, Virtual Rapport 2.0, with the Rapport Agent (Gratch et al., 2007) and the preliminary study showed that the new one performs significantly better than the old one. Compared with the old one, the new agent has a few differences, such as different nonverbal behavior prediction models, different appearance, and different animations. It is necessary to analyze how each of the differences contributes to the overall improvement in order to guide the development of future versions. Moreover, we will deploy the new agent in human studies, such as (Gratch et al., 2006) (Gratch et al., 2007) (Wang & Gratch, 2010) (Von der Putten et al., 2009), to find out whether it can induce similar or better social effects on human participants.

5. To induce good parasocial experiences, the videos we used in previous work were in line with the goal of the parasocial interaction. For example, to model listener backchannel feedback, we asked participants to create rapport with speakers by providing backchannel feedback; accordingly, the videos were derived from real interactions where the speaker retold a story and the listener provided backchannel feedback to him to create rapport (i.e. the same interactional goal). However, the non-contingent nature of parasocial interaction provides us with the flexibility to investigate new interactional goals, even if they are different from the nature of the original interactions, with pre-existing videos. For example, if the participant in parasocial interaction is given the goal to be disruptive, is it important that the speaker in a video acts similarly disrupted? In other words, do we need to bend over backwards to maintain some suspension of disbelief (i.e. create better parasocial experiences), or will coders provide reasonable responses in the face of

clear evidence that their responses are irrelevant to the trajectory of the interaction? It is interesting to explore the effects of media on parasocial interaction in future work.

Bibliography

Abrilian, S., Devillers, L., Buisine, S., & Martin, J. C. (2005). Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. *HCI International*.

Afzal, S., & Robinson, P. (2009). Natural affect data - collection and annotation in a learning context. *Proceedings of International Conference on Affective Computing and Intelligent Interaction*.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Seventh International Conference on Spoken Language Processing*.

Bailenson, J.N., Blascovich, J., Beall, A.C., & Loomis, J.M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and Virtual Environments*, *10*, 583–598.

Bailenson, J.N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, *16*, 814–819.

Bargh, J.A. (1988). Automatic information processing: Implications for communication and affect. *Communication, social cognition, and affect*, 9–32.

Bavelas, J.B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*, 941–952.

Bavelas, J.B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*, 566–580.

Bavelas, J.B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*.

Beattie, G.W. (1982). Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted. *Semiotica*, *39*, 93–114.

Bernieri, F.J., & Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. *Fundamentals of nonverbal behavior*, 401–432.

Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., & Sereika, S.M. (2010). Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers*, *22*, 276–288.

Biel, J.I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. *Proc. of International Conference on Weblogs and Social Media*.

Bonferroni, C.E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*, 645.

Brunner, L.J. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology*, *37*, 728–734.

Burgoon, J.K., Buller, D.B., & Woodall, W.G. (2010). *Nonverbal communication*. Allyn & Bacon.

Burgoon, J.K., & Hoobler, G.D. (1994). Nonverbal signals. *Handbook of interpersonal communication*, *2*, 229–285.

Burns, M. (1984). Rapport and relationships: The basis of child care. *Journal of Child Care*, 47–57.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, *41*, 181–190.

Cassell, J. (2000). More than just another pretty face: Embodied conversational interface agents. *Communications of ACM*, *43*, 70–78.

Cassell, J., et al. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, 1–27.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Proceedings of the 21st annual conference on Computer graphics and interactive techniques* (pp. 413–420).

Cassell, J., Vilhjalmsson, H.H., & Bickmore, T. (2001). Beat: the behavior expression animation toolkit. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (pp. 477–486).

Chartrand, T.L., & Bargh, J.A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893.

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks (pp. 371–388.).

Darwin, C., Ekman, P., & Prodger, P. (2002). *The expression of the emotions in man and animals*. Oxford University Press, USA.

de Kok, I., & Heylen, D. (2009). Multimodal end-of-turn prediction in multi-party meetings. *Proceedings of the 2009 international conference on Multimodal interfaces* (pp. 91–98).

de Kok, I., & Heylen, D. (2011). The multilis corpus–dealing with individual differences in nonverbal listening behavior. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, 362–375.

de Kok, I., Ozkan, D., Heylen, D., & Morency, L.P. (2010). Learning and evaluating response prediction models using parallel listener consensus. *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (pp. 3–10).

de Melo, C., Zheng, L., & Gratch, J. (2009). Expression of moral emotions in cooperating agents. *Intelligent Virtual Agents* (pp. 301–307).

Del Barrio, V., Aluja, A., & García, L.F. (2004). Relationship between empathy and the big five personality traits in a sample of spanish adolescents. *Social Behavior and Personality*, *32*, 677–682.

DeVault, David, Sagae, Kenji, & Traum, David (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, *2*, 143–170.

Drolet, A.L., & Morris, M.W. (2000). Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, *36*, 26–50.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, *23*, 283–292.

Duncan, S. (1974). On the structure of speaker–auditor interaction during speaking turns. *Language in society*, *3*, 161–180.

Duncan, S.F., & Fiske, D.W. (1985). *Interaction structure and strategy*. Cambridge University Press.

Ekman, P., & Friesen, W.V. (1977). Facial action coding system.

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, *66*, 398–412.

Gratch, J., Cheng L. Marsella S., & Boberg, J. (2013). Felt emotion and social context determine the intensity of smiles in a competitive video game. *10th IEEE International Conference on Automatic Face and Gesture Recognition*.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., & Morency, L.P. (2006). Virtual rapport. *Intelligent Virtual Agents* (pp. 14–27).

Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating rapport with virtual agents. *Intelligent Virtual Agents* (pp. 125–138).

Hartmann, T., & Goldhoorn, C. (2010). Horton and wohl revisited: Exploring viewers's experience of parasocial interactions. *the annual meeting of the International Communication Association*.

Heylen, D. (2005). Challenges ahead head movements and other social acts in conversations. *Artificial Intelligence and the Simulation of Behaviour*.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, *3*, 241–267.

Horton, D., & Wohl, R.R. (1956). Mass communication and para-social interaction. *Psychiatry*, *19*, 215–229.

Jonsdottir, Gudny, Gratch, Jonathan, Fast, Edward, & Thórisson, Kristinn (2007). Fluid semantic back-channel feedback in dialogue: Challenges and progress. *Intelligent Virtual Agents* (pp. 154–160).

Jonsdottir, G., Thorisson, K., & Nivel, E. (2008). Learning smooth, human-like turntaking in realtime dialogue. *Intelligent Virtual Agents* (pp. 162–175).

Kang, S.H., & Gratch, J. (2011). People like virtual counselors that highly-disclose about themselves. *The Annual Review of Cybertherapy and Telemedicine*.

Kang, S.H., Gratch, J., & Watt, J.H. (2009). The effect of affective iconic realism on anonymous interactants' self-disclosure. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems* (pp. 4021–4026).

Kendon, A. (2002). Some uses of the head shake. *Gesture*, *2*, 147–182.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kipp, M., Neff, M., Kipp, K., & Albrecht, I. (2007). Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. *Intelligent Virtual Agents* (pp. 15–28).

Krumhuber, E., Manstead, A.S.R., Cosker, D., Marshall, D., Rosin, P.L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*, 730–735.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML* (pp. 282–289).

Lao, S., & Kawade, M. (2005). Vision-based face understanding technologies and their applications. *Advances in Biometric Person Authentication*, 339–348.

Lee, J., & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. *Intelligent Virtual Agents* (pp. 243–255).

Lee, J., & Marsella, S. (2009). Learning a model of speaker head nods using gesture corpora. *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems* (pp. 289–296).

Levy, M.R. (1979). Watching tv news as para-social interaction. *Journal of Broadcasting*, *23*, 60–80.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (cert). *IEEE International Conference on Automatic Face and Gesture Recognition*.

Masters, R.D., & Sullivan, D.G. (1990). *Nonverbal behavior and leadership: Emotion and cognition in political information processing*. Institute of Governmental Studies, University of California, Berkeley.

McClave, E.Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*, 855–878.

McDougall, W. (2003). *An introduction to social psychology*. Dover Pubns.

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

Mehrabian, A. (1981). *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth.

Mehrabian, A. (2007). *Nonverbal communication*. Aldine.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, 323–339.

Morency, L.P., de Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. *Intelligent Virtual Agents* (pp. 176–190).

Morency, L.P., Quattoni, A., & Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).

Morency, L.P., Sidner, C., Lee, C., & Darrell, T. (2005). Contextual recognition of head gestures. *Proceedings of the 7th international conference on Multimodal interfaces* (pp. 18–24).

Murphy, K.P. (2002). *Dynamic bayesian networks: representation, inference and learning*. Doctoral dissertation.

Nishimura, R., Kitaoka, N., & Nakagawa, S. (2007). A spoken dialog system for chat-like conversations considering response timing. *Proceedings of the 10th International Conference on Text, speech and Dialogue* (pp. 599–606).

Novick, D.G., Hansen, B., & Ward, K. (1996). Coordinating turn-taking with gaze. *ICSLP 96* (pp. 1888–1891).

Ozkan, D., & Morency, L.P. (2011). Modeling wisdom of crowds using latent mixture of discriminative experts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 335–340).

Ozkan, D., & Morency, L.-P. (2010). Self-based feature selection for nonverbal behavior analysis. *workshop on Human Behavior Understanding, ICPR*.

Pelachaud, C. (1996). Simulation of face-to-face interaction. *Proceedings of the workshop on Advanced visual interfaces* (pp. 269–271).

Perse, E.M., & Rubin, R.B. (1989). Attribution in social and parasocial relationships. *Communication Research*, *16*, 59–77.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.

Rowe, J., Shores, L., Mott, B., & Lester, J. (2010). Integrating learning and engagement in narrative-centered learning environments. *Intelligent Tutoring Systems* (pp. 166–177).

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 696–735.

Selting, M. (2000). The construction of units in conversational talk. *Language in Society*, *29*, 477–517.

Shiota, M.N., Keltner, D., & John, O.P. (2006). Positive emotion dispositions differentially associated with big five personality and attachment style. *The Journal of Positive Psychology*, *1*, 61–71.

Smith, A., Sen, A., & Hanley, R.P. (2010). *The theory of moral sentiments*. Penguin Classics.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A.Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254–263).

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (pp. 1–8).

Strauss, M., & Kipp, M. (2008). Eric: a generic rule-based framework for an affective embodied commentary agent. *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems* (pp. 97–104).

Sundar, S.S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor? *Communication Research*, *27*, 683–703.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday Books.

Swartout, W.R., Gratch, J., Hill Jr, R.W., Hovy, E., Marsella, S., Rickel, J., & Traum, D. (2006). Toward virtual humans. *AI Magazine*, *27*, 96–108.

Tatar, D.G. (1997). *Social and personal effects of a preoccupied listener.* Stanford University.

ter Maat, M., Truong, K., & Heylen, D. (2010). How turn-taking strategies influence users impressions of an agent. *Intelligent Virtual Agents* (pp. 441–453).

Thiebaux, M., Marsella, S., Marshall, A.N., & Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems* (pp. 151–158).

Thorisson, K.R. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in language and speech systems*, 173–207.

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *1*, 285–293.

Von der Putten, A., Kramer, N.C., & Gratch, J. (2009). Who's there? can a virtual agent really elicit social presence. *Proceedings of the PRESENCE.*

Wang, N., & Gratch, J. (2009). Rapport and facial expression. *Affective Computing and Intelligent Interaction, 2009* (pp. 1–6).

Wang, N., & Gratch, J. (2010). Don't just stare at me! *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1241–1250).

Wang, Zhiyang, Lee, Jina, & Marsella, Stacy (2011). Towards more comprehensive listening behavior: Beyond the bobble head. *Intelligent Virtual Agents* (pp. 216–227).

Ward, N. (1997). Aizula: A back-channeling system for english and japanese. *Eurospeech*.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, *32*, 1177–1207.

Yngve, V. (1970). On getting a word in edgewise. *Sixth Regional Meeting of the Chicago Linguistic Society*. University of Chicago, Department of Linguistics.

Appendix A

90-item questionnaire assessing personality traits

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number (as shown in Table A.2) next to each statement to indicate the extent to which you agree or disagree with that statement.

I see myself as someone who ...

1. Is talkative
2. Tends to find fault with others
3. Does a thorough job
4. Is depressed, blue
5. Is original, comes up with new ideas
6. Is reserved
7. Is helpful and unselfish with others
8. Can be somewhat careless
9. Is relaxed, handles stress well
10. Is curious about many different things
11. Is full of energy
12. Starts quarrels with others
13. Is a reliable worker
14. Can be tense
15. Is ingenious, a deep thinker

Trait	Question Index
Extroversion	1, 6, 11, 16, 21, 26, 31, 36
Agreeableness	2, 7, 12, 17, 22, 27, 32, 37, 42
Conscientiousness	3, 8, 13, 18, 23, 28, 33, 38, 43
Neuroticism	4, 9, 14, 19, 24, 29, 34, 39
Openness	5, 10, 15, 20, 25, 30, 35, 40, 41, 44
Self-consciousness	45, 46, 47, 48, 49, 50, 51, 52, 53, 89, 90
Other-consciousness	54, 55, 56, 57, 58, 59, 60
Shyness	61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82
Self-monitor	83, 84, 85, 86, 87, 88

Table A.1: The table shows the set of questions measuring each of the personality traits.

Description	Score
Strongly Disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly Agree	5

Table A.2: 5-point Likert scale.

16. Generates a lot of enthusiasm
17. Has a forgiving nature
18. Tends to be disorganized
19. Worries a lot
20. Has an active imagination
21. Tends to be quiet
22. Is generally trusting
23. Tends to be lazy
24. Is emotionally stable, not easily upset
25. Is inventive
26. Has an assertive personality
27. Can be cold and aloof

28. Perseveres until the task is finished
29. Can be moody
30. Values artistic, aesthetic experiences
31. Is sometimes shy, inhibited
32. Is considerate and kind to almost everyone
33. Does things efficiently
34. Remains calm in tense situations
35. Prefers work that is routine
36. Is outgoing, sociable
37. Is sometimes rude to others
38. Makes plans and follows through with them
39. Gets nervous easily
40. Likes to reflect, play with ideas
41. Has few artistic interests
42. Likes to cooperate with others
43. Is easily distracted
44. Is sophisticated in art, music, or literature
45. I'm always trying to figure myself out
46. I think about myself a lot
47. I often daydream about myself
48. I never take a hard look at myself
49. I generally pay attention to my inner feelings
50. I'm constantly thinking about my reasons for doing things
51. I sometimes step back (in my mind) in order to examine myself from a distance
52. I'm quick to notice changes in my mood
53. I know the way my mind works when I work through a problem

54. I'm concerned about my style of doing things
55. I care a lot about how I present myself to others
56. I'm self-conscious about the way I look
57. I usually worry about making a good impression
58. Before I leave my house, I check how I look
59. I'm concerned about what other people think of me
60. I'm usually aware of my appearance
61. I feel tense when I'm with people I don't know well
62. I am socially somewhat awkward
63. I do not find it difficult to ask other people for information
64. I am often uncomfortable at parties and other social functions
65. When in a group of people I have trouble thinking of the right things to talk about
66. It does not take me long to overcome my shyness in new situations
67. It is hard for me to act natural when I am meeting new people
68. I fell nervous when speaking to someone in authority
69. I have no doubts about my social competence
70. I have trouble looking someone right in the eye
71. I feel inhibited in social situations
72. I do not find it hard to talk to strangers
73. I am more shy with members of the opposite sex
74. It's hard for me to work when someone is watching me
75. I get embarrassed very easily
76. I have no particular desire to avoid people
77. I often find social occasions upsetting
78. I usually feel calm and comfortable at social occasions
79. If the chance comes to meet new people, I often take it

80. I often feel nervous or tense in causal get-togethers in which both sexes are present
81. Being introduced to people makes me tense and nervous
82. I tend to withdraw from people
83. I am often able to read people's true emotions correctly (through their eyes)
84. In conversations, I am sensitive to even the slightest change in the facial expression of the person with whom I am conversing
85. My powers of intuition are quite good when it comes to understanding the emotions and motives of others
86. I can usually tell when others consider a joke to be in bad taste, even though they may laugh convincingly
87. I can usually tell when I've said something inappropriate by reading it in the listener's eyes
88. If someone is lying to me, I usually know it at once from that person's manner of expression
89. I consider myself to be a self-conscious person
90. People tell me I am too self-conscious