Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior

Lixing Huang
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina Del Rey, CA, 90292 USA
lhuang@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina Del Rey, CA, 90292 USA
morency@ict.usc.edu

Jonathan Gratch
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina Del Rey, CA, 90292 USA
gratch@ict.usc.edu

ABSTRACT

Virtual humans are embodied software agents that should not only be realistic looking but also have natural and realistic behaviors. Traditional virtual human systems learn these interaction behaviors by observing how individuals respond in face-to-face situations (i.e., direct interaction). In contrast, this paper introduces a novel methodological approach called parasocial consensus sampling (PCS) which allows multiple individuals to vicariously experience the same situation to gain insight on the typical (i.e., consensus view) of human responses in social interaction. This approach can help tease apart what is idiosyncratic from what is essential and help reveal the strength of cues that elicit social responses. Our PCS approach has several advantages over traditional methods: (1) it integrates data from multiple independent listeners interacting with the same speaker, (2) it associates probability of how likely feedback will be given over time, (3) it can be used as a prior to analyze and understand the face-to-face interaction data, (4) it facilitates much quicker and cheaper data collection. In this paper, we apply our PCS approach to learn a predictive model of listener backchannel feedback. Our experiments demonstrate that a virtual human driven by our PCS approach creates significantly more rapport and is perceived as more believable than the virtual human driven by face-to-face interaction data.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Measurement, Performance, Design, Experimentation

Keywords

Virtual Humans, Rapport, Backchannel Feedback, Parasocial

Cite as: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior, Lixing Huang, Louis-Philippe Morency, Jonathan Gratch, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010),* van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1265-1272

1. INTRODUCTION

Virtual humans are embodied software agents designed to simulate the appearance and social behavior of humans, typically with the goal of facilitating natural interaction between humans and computers. Previous psychological work [6][7] has emphasized that face-to-face interactions between people can be richly interactive, involving verbal and nonverbal synchrony and frequent feedback between interlocutors including nods, interjections and facial expressions. When present, these characteristics promote effective communication and have encouraged the development of virtual humans that can replicate this richness. Indeed, recent work has demonstrated that, through simulating such interactional behaviors, virtual humans can promote feelings of rapport [4][13][14][19], increase interactional fluency [18] and promote self-disclosure of intimate information [17].

In order to achieve those effects, virtual human researchers have turned to data-driven methods to automatically learn realistic interactional behaviors. Traditionally, virtual humans learn from annotated recordings of face-to-face interaction [1][3][5][15][22][23]. However, there are some drawbacks with such data. First, there is considerable variability in human behavior and not all human data should be considered a positive example of the behavior a virtual human is attempting to learn. For example, if the goal is to learn to produce feelings of rapport, it is important to realize that many face-to-face interactions fail in this regard. Ideally, such data must be separated into good and bad instances of the target behavior, but it is not always obvious how to make this separation. Second, a virtual human is attempting to learn a general behavior pattern that it could apply across social situations, yet each example in a face-to-face dataset is intrinsically idiosyncratic – illustrating how one particular individual responded to another. Such data gives us no insight on how typically the responses might be or how well they might generalize across individuals.

Although the common wisdom is that face-to-face interaction data is the gold standard and third party observers always have different feelings from people involved in an interaction, research into *parasocial interaction* [19] suggests that individuals can readily respond as if they were in a natural social interaction when they interact with pre-recorded media. In this paper, we present a data-collection paradigm called Parasocial Consensus Sampling (PCS) that exploits this characteristic of human behavior. Instead of recording face-to-face interactions,

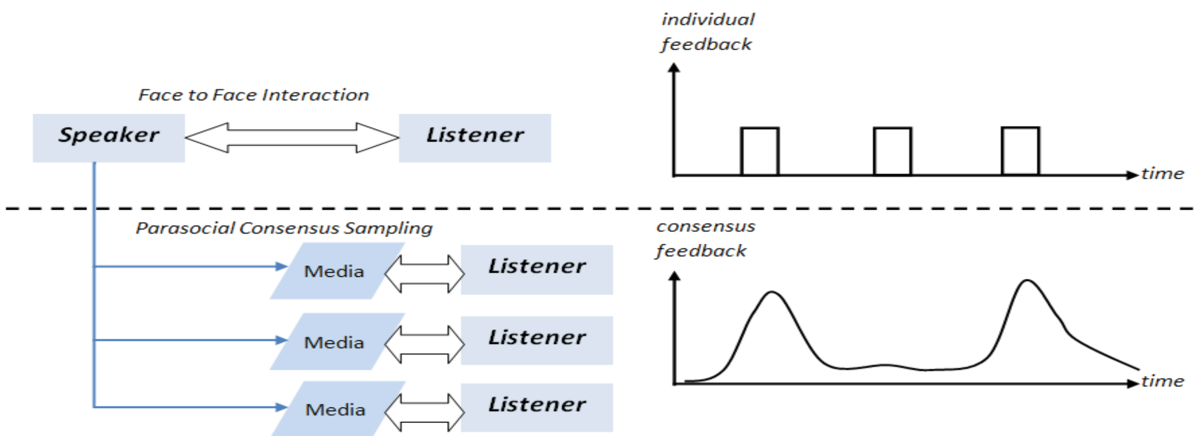

Figure 1. Comparison between Parasocial Consensus Sampling (PCS) and conventional Face-to-Face Interaction. Unlike face-to-face interaction, where interaction behaviors are deduced by observing how individuals respond in a social situation, parasocial consensus sampling allows multiple individuals to 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.

participants are guided through a parasocial interaction. Given some communicative goal (for example, convey to the person you are interested in what s/he is talking about), the participants are requested to achieve that by interacting with the mediated representation of a person. In this way, multiple participants can interact with the *same* media. This approach has several advantages over the traditional ways: (1) it allows multiple independent listeners to interact with the same speaker, (2) it associates probability of how likely feedback will be given over time, (3) it can be used as a prior to analyze and understand the face-to-face interaction data, (4) it substantially reduces the time and cost of data collection.

The following section describes the related work in virtual human non-verbal behavior generation and parasocial interaction. Section 3 explains the general framework of Parasocial Consensus Sampling paradigm. In section 4, we apply the Parasocial Consensus Sampling paradigm to collect listener backchannel data. Section 5 presents the subjective evaluation experiments and discusses the evaluation results. We conclude our work in Section 6.

2. Related Work

Prior research has produced a variety of virtual humans that can provide rich interactive feedback to human speakers. The bulk of this work has focused on techniques for analyzing or learning from large datasets of face-to-face interactions. For example, Ward et al. [3] examined natural face-to-face interactions to derive a rule-based model where backchannels are associated with a region of low pitch lasting 110ms during speech. Nishimura et al. [15] 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. Maatman et al. [23] combined Ward's algorithm with a simple method of mimicking head nods and subjective evaluations demonstrated the generated behaviors do improve feelings of rapport and speech fluency. Morency et al. [1] advanced Ward's work by proposing a statistical machine learning model, they developed an automatic feature selection strategy and trained Latent Dynamic Conditional Random Field based on multimodal features (lexical words, prosodic features, eye gaze) to learn the dynamic structure during interaction. Jonsdottir et al. [22] built a dialogue system which uses prosody features to learn turn-taking behaviors. They implemented a reinforcement learning model to learn this on the fly and the system is very close to human speakers with regards to speed.

Although this work has innovated techniques for learning from data, there has been less attention to innovating methods for collecting the data these systems use to learn. The implicit assumption in the above work is that the best results can be obtained from collecting lots of examples of face-to-face interactions. However, as discussed before, face-to-face interaction has problems, such as individual variability and less generalization.

An alternative way to collect data is to interact parasocially. *Parasocial Interaction*, first introduced by Horton and Wohl [19], occurs when people exhibit the natural tendency to interact with media representations of people *as if* they were interacting face-to-face with the actual person [24]. Fifty-years of research has documented that people readily produce such "parasocial" responses and these responses bear close similarity to what is found in natural face-to-face interactions, even though the respondents are clearly aware they are interacting with pre-recorded media [25]. For example, Levy [21] found people behave as if they were having a two-way conversation with a television news anchorperson while watching the person on TV. Parasocial interaction research suggests that participants could assume the role of one interaction partner in a previously recorded conversation and produce social responses similar to what they would exhibit if they were in the original face-to-face conversation. But there is no similar work, as far as we know, that shows whether the parasocial interaction works for human interaction data collection. This paper is the first one to apply the parasocial interaction theory in collecting human behavior data and generating virtual human behavior.

3. Parasocial Consensus Sampling

Parasocial consensus sampling 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, 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. 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 these moments. 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.

More concretely, we define parasocial consensus sampling as follows. Given:

- *An interactional goal:* this is the intended goal of the virtual human interactional behaviors. For example, Gratch et al [2] 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 promote 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 conversational partner, participants should be encouraged to produce this behavior. Candidate behavioral responses include backchannel feedback [2], turn taking [26], evaluative facial expressions or paraverbals such as "uh-huh"[3].
- *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. 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 conversation 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.

There are several differences between PCS and the traditional data collection approach as shown in Figure 1.

(1) In face-to-face interaction, speaker and listener are paired; while in PCS, multiple independent listeners interact with one speaker. The listeners actually do not interact with speakers directly; instead, the interaction is done through media, for example, through videos. In other words, what the listeners interact with is the mediated representation of the speaker. Therefore, it is possible to make multiple independent listeners interact with the same speaker, which is not typically possible in the traditional methods.

(2) In face-to-face interaction, for each speaker, only one listener's feedback data is collected. As shown in the upper part of Figure 1, what the data can provide 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

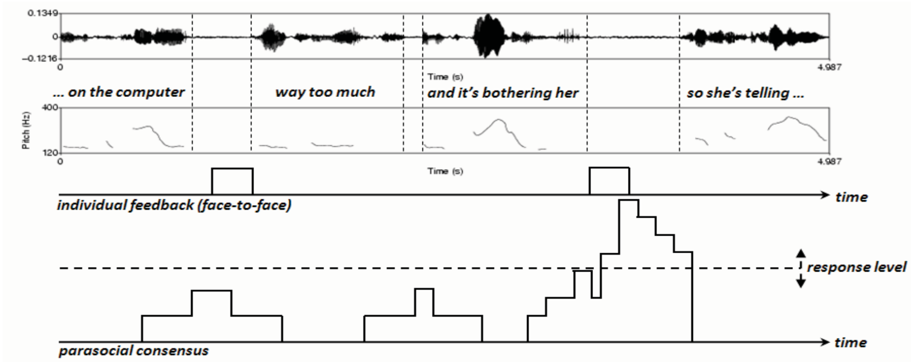

Figure 2. Example segment illustrating a parasocial consensus of listener backchannel (see Section 4) varies over time. While individual feedback (from the original face-to-face interaction) only gives discrete prediction; our parasocial consensus shows the relative importance of each feedback. By applying a response level to our parasocial consensus, we get only important feedback.

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, we call it parasocial consensus, over time, which shows how many participants agree to give feedback at time *t*. The more the number of participants agreeing to give feedback at time *t*, the higher probability the feedback has.

(3) In face-to-face interaction, the listener's feedback may contain outliers, or idiosyncratic ones. Those outliers can be excluded by applying the parasocial consensus as a mask on the original listener's feedback. A feedback is selected only if several participants all agree to give it.

4. Building Parasocial Consensus of Listener Backchannel Feedback (Experiment 1)

Parasocial consensus sampling (PCS) is a general framework for efficiently learning the typicality of human responses in social interactions. We now illustrate the utility of PCS by applying it to the problem of learning a predictive model of human backchannel feedback. Such feedback plays an important role in the establishment of rapport between people and learning when to provide this feedback has been a focus of prior research [1][3].

In this first experiment, we assess some basic questions about the methodology: can people provide parasocial responses? Do they believe their responses are meaningful? Does the resulting consensus have any correspondence to the interactional goal? In the next section we then assess if the resulting consensus can then be used to animate a virtual listener.

4.1 Method

As discussed in Section 3, parasocial consensus sampling is defined by five key elements: interaction goal, target behavioral response, media, target population and measurement channel. In our study, we targeted our parasocial sampling as follow:

- *Interactional goal:* Creating rapport.
- *Target behavioral response:* Backchannel feedback.
- *Media*: Pre-recorded videos.
- *Target population:* General public.
- *Measurement channel:* Keyboard.

The choices of interactional goal and target behavioral response are based on previous works howing the importance of creating rapport in human-human interaction [6][8][9] [10][11][12] and identifying backchannel feedback as one of the key behavioral cues [3] to create rapport. As our choice for media, we decided to use pre-recorded videos of human speakers retelling a story to another human listener. This paradigm was previously used for studying human behaviors, including rapport [4]. The most interesting design decision is the measurement channel: pressing a key to express feedback. We selected this challenging measurement channel to push the boundaries of conventional consensus sampling and find a more efficient method to model human behaviors.

4.2 Procedure

We recruited 42 participants over the web to watch pre-recorded videos. Each participant watched six randomly selected videos from a list of 30. The participants were adults from Asia, North America and Europe. Each pre-recorded video showed a different speaker retelling a story drawn from [4].

Participants 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 instructed to press the keyboard each time they felt like providing backchannel feedback such as head nod or paraverbals (e.g. "uh-huh" or "OK").

To assess participants' subjective impressions of the task, we included three questions after each video:

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

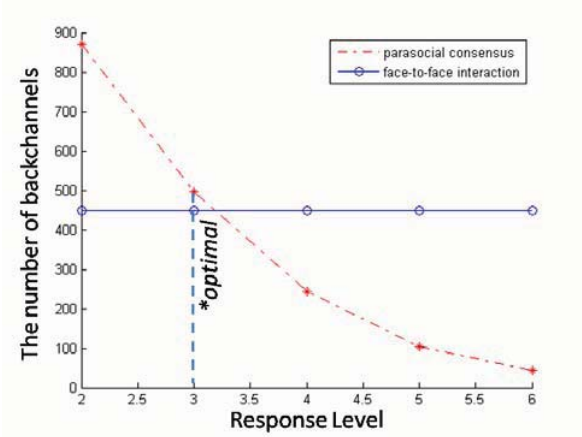

Figure 3. Selecting the response level. When the response level is set to 3, the number of backchannels from parasocial consensus data is closest to the number from face-to-face interaction data

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

4.3 Results

We bui lt the parasocial consensus by computing the histogram over time. As suggested in [1][3], the time line is converted into samples w ith s ample rate of 0. 1s and every backchannel fr om participants has a width of 1 second, that is , 10 samples. Whenever there i s a b ackchannel occur ring on a sample, the histogram of that sample increas es by 1. Th us, each sample is associated with a numb er indicating probability to give backchannel. Figure 2 shows an example of our paras ocial consensus and compares it to the backchannel feedback f rom the listener in the or iginal face-to-face interaction. By looking at the original listener's feedback, it is clear th at paus e is a good predictor of f eedback, but the r elative s trength of t his feature is not certain. On the other hand, the parasocial consensus shows the relative impor tance of e ach feedback. The las t one i s the most important. L ooking back on the interactio n data, the utterances before t he first t wo pauses are s tatements, while the last one expresses an opinion, suggesting that pauses a fter opinions may be stronger predictors o f lis tener feedback. Als o, the speaker expressed emp hasis on the third utterance. This result gives us a tool to better analyze features that predict backchannel feedback.

4.3.1 Self-assessment Questionnaire

By looking at the results from the three questions, we are able to know the p articipants' self-assessment about their feedback in the experiment.

Table 1. Self-assessment results

	Competence	Missed opportunities	Timing
Mean	4.0 1.3		1.2

It is clear that the part icipants think the tas k is eas y, and the number of missed opportunities and wrong feedback are s mall. In other words, they do feel l ike they can do such a tas k quite well. Some comments indicated tha t after watching the first video and being accustomed to the s pecial way to "interact" with the speakers in the video, it is easy to follow that routine later.

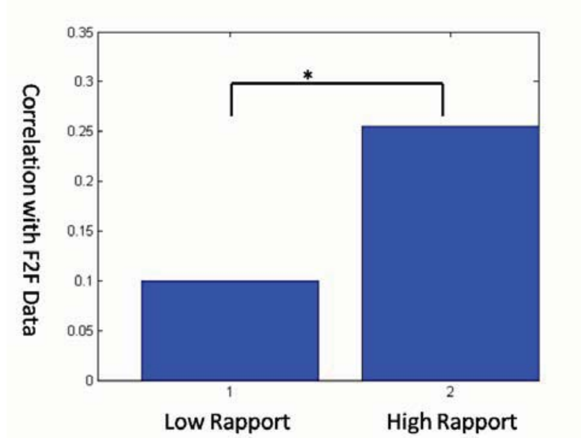

Figure 4. Correlation between PCS with face-to-face data for low-rapport set (left) and high-rapport set (right). The AVONA test on the two sets shows F = 6.32, p = 0.0184, which means parasocial consensus data correlates with high-rapport set significantly better than the low-rapport set.

4.3.2 Response Level

When predicting backcha nnels from paraso cial cons ensus, a threshold is set to filter out the backchannels whose probabilities are low. The pr obability i s determ ined by the num ber of participants agreeing to give that feedback. In the consensus data, different feedback is associated with different probability so that the higher the threshold is , the f ewer the backchannels are selected. In [1], the a uthors explained the thres hold as a way to make the virtual hum an have d ifferent expre ssiveness; the more frequent the feedback is, the more expressive the virtual human will be. We follow the concept here.

The threshold is selected to make the parasocial consensus data as expressive as the original listener's behavior. By testing different values for the threshold, as shown in Figure 3, the response level is s et to 3, whe re the number of backchannels from paras ocial consensus is closest to that from the face-to-face interaction data.

4.3.3 Objective Evaluation on Interaction Goal

Although the participants reported that they can do th is task quite well, it is necessary to find an obj ective wa y to measur e the quality of their consensus. Participants were instructed to create a sense of rapport, s o one way to as sess the quality of their consensus is to compare the cons ensus be haviors with the listeners' behaviors in the original da taset: if the beha vior of an original listener closely approximates the consensus behavior, we would predict that the l istener would be j udged as exhibiting high rapport; if they differed s ignificantly from the consens us, we would expect them to have low rapport. Indeed, this is what we show.

More specifically, we:

a) **Separated videos into a low-rapport set and a high-rapport set**: W e s ort the videos in ascending order based on the level of rapport that the original speaker felt in their f2f interaction, and group the first half into low-rapport set and the second half into high-rapport set.

b) **Predict backchannels**: As mentioned in 4.3.2, the response level is set to 3, the peaks in parasocial consensus whose values are larger than that are selected as the predicted backchannel time.

c) **Compute correlation**: The correlation is measured by computing the percentage of predicted backchannels that can find matches in the f2f interaction data for each video.

d) **Compare the correlation with low-rapport set and high-rapport set**: each video has a correlation measurement between parasocial consensus data and face-to-face data. ANOVA test is applied to find whether there is significant difference for the correlation measurement of videos in the two sets. The mean value of the correlation for low-rapport set is 0.1, and the mean value for high-rapport set is 0.26, $F = 6.32$, $p = 0.0184$. (As shown in Figure 4.)

Clearly, there is significant difference between the two video sets, which means the parasocial consensus correlates with the face-to-face interaction data much better when the speaker reported high rapport level. In other words, the parasocial consensus represents the listeners' backchannels that create more rapport. This is objective evidence that the participants can do this task well.

5. Subjective Evaluation of Parasocial Consensus (Experiment 2)

Experiment 1 demonstrated that participants feel comfortable producing parasocial responses and that their consensus is correlated with the desired interactional goal. In Experiment 2 we assess if the parasocial consensus can be used to naturally animate the behavior of virtual humans and if this behavior achieves the interactional goal.

Specifically, we construct videos illustrating a human interacting with a virtual listening agent (Figure 5) and assess the naturalness and perceived rapport of alternative methods for generating the virtual human's backchannel feedback to the human's speech. We hypothesize that PCS will be better in terms of rapport (given that the elicitations demonstrate the consensus view toward this interactional goal) and comparable in naturalness that a human listener experienced in the face-to-face interaction.

5.1 User Study

Five speaker videos are randomly selected from the 30 pre-recorded face-to-face interactions. For each speaker video, the virtual human [16] is driven by four kinds of backchannel data respectively:

- **PCS**: the backchannels from parasocial consensus where the response level is set to 3.
- **F2F**: the face-to-face interaction's backchannels.

Figure 5. Videos for subjective evaluation

- **PCS all**: the backchannels from parasocial consensus where the response level is set to 0.
- **Random**: random backchannels.

The four versions of virtual human's behavior are composed together with the corresponding speaker's video as shown in Figure 5.

In a within-subjects design, 33 participants were recruited to evaluate the quality of these different behavioral mappings. Each participant saw the four versions (presented in a random order) of one of the five videos. Before watching those videos, the participants are told that "In each video, there is a speaker telling a story and a virtual human trying to give feedbacks 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 to evaluate the virtual human's feedback 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))

5.2 Results

General Linear Model repeated measure [27] is used here to find whether there is significant difference among the four versions. The results are summarized in Figure 6.

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.

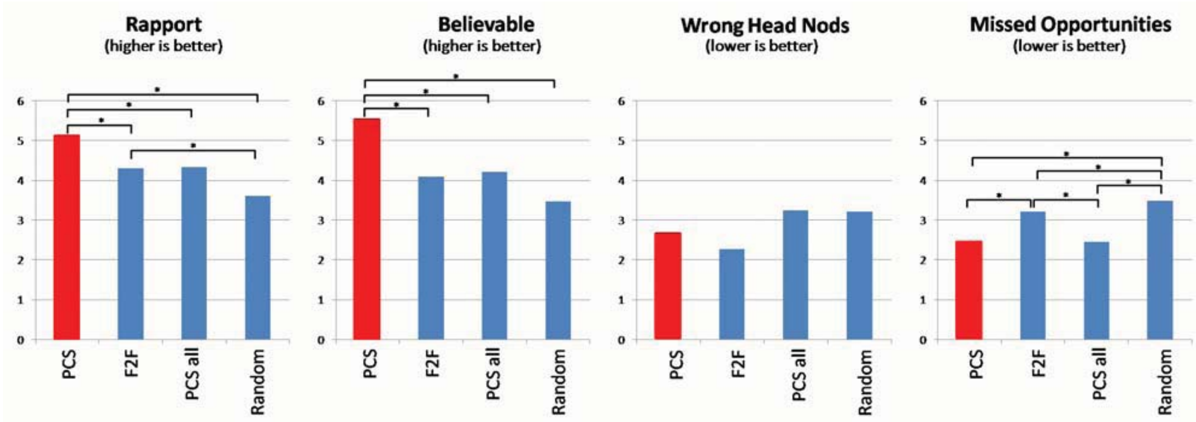

Figure 6. the subjective evaluation results for rapport, believable, wrong head nods, and missed opportunities of the four versions: PCS, F2F, PCS all, and Random. The star(*) means there is significant difference between the versions under the brackets.

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*.

5.3 Discussion

From the rapport and believable question (mentioned in section 5.1), 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 the parasocial consensus sampling learns a better model of listener backchannels than the conventional face-to-face interaction data. Not surprisingly, *random* head-nods produce the worst result, which matches the work in [2], where the authors found "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 response level, as described in section 4.3.2.

When looking at the wrong head nods and missed opportunities questions, we can see 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 opportunities to give backchannels, while by aggregating the feedback from multiple independent participants, we could get a more complete picture. Also it is worth noticing that the number of missed opportunities is identical for *PCS* and *PCS all*, showing that the response level did not filter important backchannel feedback.

In other words, the results from our subjective evaluation shows that the *PCS* data has the least false negative samples of backchannels, and the virtual human driven by *PCS* data creates the most rapport within the interaction, thus, it is the most believable one as well.

6. Conclusion and Future Work

In this paper, we presented a new paradigm called parasocial consensus sampling (PCS) which allows multiple individuals to vicariously experience the same situation to gain insight on the typical (i.e., consensus view) of human responses in social interaction. This approach helps tease apart what is idiosyncratic from what is essential and helps reveal the strength of cues that elicit social responses. Comparing with face-to-face interaction data, our PCS approach has several advantages: (1) it allows multiple independent listeners to interact with the same speaker, (2) it associates probability of how likely feedback will be given over time, (3) it can be used as a prior to analyze and understand the face-to-face interaction data, (4) it can collect data in a much faster and cheaper way. We applied parasocial consensus sampling to collect listener backchannel data, and the experiments showed the virtual human driven by our PCS approach creates significantly more rapport and is perceived as more believable than the virtual human driven by face-to-face interaction data.

The current work can be extended in several ways. We tested the new paradigm in the context of backchannel prediction, but there are many possible candidates which are potentially suited to this approach, such as turn-taking, eye gaze shift, facial expression. We want to run some similar experiments on other problems as well to testify the validation of our approach in advance.

7. ACKNOWLEDGMENTS

This m aterial is bas ed upon work s upported by the National Science Foundation under Grant No. 0729287 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no offic ial endorsement s hould be inferred.

8. REFERENCES

[1] Morency, L.-P., Kok I. de, Gratch, J. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In Proceedings of 8th International Conference on Intelligent Virtual Agents (Tokyo, Japan, 2008).

[2] Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R. 2007. Creating Rapport with Virtual Agents. In Proceedings of 7th International Conference on Intelligent Virtual Agents (Paris, France, 2007).

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

[4] Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Werf, R. J., and Morency, L.-P. 2006. Virtual Rapport. In Proceedings of 6th International Conference on Intelligent Virtual Agents (Marina del Rey, CA, 2006).

[5] Heylen D.K.J. 2008. Listening Heads. Modeling Communication with robots and virtual humans. (2008) 241 - 259

[6] Tickle-Degnen, L., Rosenthal R. 1990. The Nature of Rapport and its Nonverbal Correlates. Psychological Inquiry 1(4) 1990, 285 - 293.

[7] Burgoon, J.K., Dillman, L., Stern, L.A. 1993. Adaptation in Dyadic Interaction: Defining and Operationalizing Patterns of Reciprocity and Compensation. Communication Theory. (1993) 295 - 316.

[8] Drolet, A., Morris, M. 2000. Rapport in conflict resolution: accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Experimental Social Psychology 36 (2000) 26-50.

[9] Goldberg, S. 2005. The secrets of successful mediators. Negotiation Journal 21(3) (2005), 365-376.

[10] Tsui, P., Schultz, G. 1985. Failure of rapport: Why psychotherapeutic engagement fails in the treatment of asian clients. American Journal of Orthopsychiatry 55 (1985) 561 - 569.

[11] Fuchs, D. 1987. Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education 7 (1987) 90 - 104.

[12] Burns, M. 1984. Rapport and relationships: the basic of child care. Journal of Child Care 2 (1984) 47 - 57.

[13] Kang, S.-H., Gratch, J., Wang, N., Watt, J. 2008. Agreeable People like Agreeable Virtual Humans. In Proceedings of 8th International Conference on Intelligent Virtual Agents (Tokyo, Japan, 2008).

[14] Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., and Morency, L.-P. 2007. Can virtual humans be more engaging than real ones? In Proceedings of 12th International Conference on Human-Computer Interaction (Beijing, China, 2007).

[15] Nishimura, R., Kitaoka, N., Nakagawa, S. 2007. A spoken dialog system for chat-like conversations considering response timing. LNCS 4629 (2007) 599 - 606.

[16] Thiebaux, M., Marshall, A., Marsella, S., Kallmann, M. 2008. Smartbody: Behavior realization for embodied conversational agents. In Proceedings of 7th International Conference on Autonomous Agents and Multiagent Systems (Estoril, Portugal, May, 2008).

[17] Kang, S.-H., Gratch, J., Watt, J. 2009 . The Effect of Affective Iconic Realism on Anonymous Interactants' Self-Disclosure. In Proceedings of Interaction Conference for Human-Computer Interaction (Boston, 2009)

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

[19] Cassell, J., Gill, A.J., Tepper, P.A. 2007. Coordination in Conversation and Rapport. In Proceedings of ACL workshop on Embodied Natural Language. (Prague, CZ, 2007)

[20] Horton, D., Wohl, R.R. 1954. Mass communication and para-social interaction: Observation on intimacy at a distance. Psychiatry 19 (1956) 215-229.

[21] Levy, M.R. 1979. Watching TV news as para-social interaction. Journal of Broadcasting. 23 (1979) 60-80.

[22] Jonsdottir, G.R. 2008. A Distributed Dialogue Architecture with Learning. Master Thesis. Reykjavik University.

[23] Mattman, M., Gratch, J., Marsella, S. 2005. Natural behavior of a listening agent. In Proceedings of Interactional Conference on Intelligent Virtual Agents (Kos, Greece, 2005).

[24] Sundar, S.S., Nass, C. 2000. Source orientation in human-computer interaction: Programmer, networker, or independent social actor? Communication Research. 27 (6), 683 - 703

[25] Houlberg, R. 1984. Local television news audience and the para-social interaction. Journal of Broadcasting. 28 (1984) 423- 429.

[26] Jonsdottir, G.R., Thorisson, K.R., Nivel, E. 2008. Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In the Proceedings of International Conference on Intelligent Virtual Agents. (Tokyo, Japan, 2008)

[27] GLM. http://www.statsoft.com/textbook/general-linear-models/