Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch

Institute for Creative Technologies, University of Southern California,
13274 Fiji Way, Marina del Rey, CA 90292, USA
{lhuang,morency,gratch}@ict.usc.edu

Abstract. Backchannel feedback is an important kind of nonverbal feedback within face-to-face interaction that signals a person's interest, attention and willingness to keep listening. Learning to predict when to give such feedback is one of the keys to creating natural and realistic virtual humans. Prediction models are traditionally learned from large corpora of annotated face-to-face interactions, but this approach has several limitations. Previously, we proposed a novel data collection method, Parasocial Consensus Sampling, which addresses these limitations. In this paper, we show that data collected in this manner can produce effective learned models. A subjective evaluation shows that the virtual human driven by the resulting probabilistic model significantly outperforms a previously published rule-based agent in terms of rapport, perceived accuracy and naturalness, and it is even better than the virtual human driven by real listeners' behavior in some cases.

Keywords: Parasocial Interaction, Virtual Human, Backchannel Prediction.

1 Introduction

When people interact face-to-face, actions often speak louder than words. A speaker's facial expressions, gestures and postures can dictate the meaning of an utterance; whereas a listener's nonverbal reactions provide moment-to-moment feedback that can alter and serve to co-construct subsequent speech [21,22,23]. Beyond its impact on meaning, nonverbal signals communicate emotion and personality, enhance the persuasiveness of speech, express social status and regulate conversational flow. Not surprisingly, considerable effort has been directed at endowing virtual humans with the ability to recognize, understand and exploit the nonverbal channel [16,17,18].

Virtual humans that produce such nonverbal signals can induce desirable social changes in their human interaction partners. Synthetic nonverbal behaviors can enhance the persuasiveness of virtual human speech [7], encourage people to take their medicine [12], and promote more cooperation in economic games [11]. Our own studies with the Rapport Agent [3] suggest that nonverbal behavior plays a causal role in achieving these effects. As a result of its contingent nonverbal feedback, human speakers speak more fluently with the Rapport Agent [6], disclose more intimate information about themselves [13] and may better remember recent events [14]. Indeed,

J. Allbeck et al. (Eds.): IVA 2010, LNAI 6356, pp. 159–172, 2010

these and related studies suggest that a virtual human's behavior may be more important than its appearance in achieving social effects [8].

Although early research on virtual humans relied on hand-crafted algorithms to generate nonverbal behaviors, informed by psychological theories or personal observations of face-to-face interaction [4], recent scholarship has seen an explosion in interest in data-driven approaches that automatically learn virtual human behaviors from annotated corpora of human face-to-face interactions. Several systems now exist that automatically learn a range of nonverbal behaviors including backchannel feedback [2], conversational gestures [9,15] and turn-taking cues [10].

It is widely assumed that natural human-to-human interaction constitutes the ideal dataset from which to learn virtual human behaviors, however, there are drawbacks with such data. First, natural data can be expensive and time-consuming to collect. Second, human behaviors contain variability so that some of the behavior samples may conflict with the social effect that we want the virtual human to produce. Finally, each instance in face-to-face interaction only illustrates how one particular individual responds to another, yet such data fails to give us insight on how well such responses generalize across individuals. Rather than simply exploring more powerful learning algorithms that might overcome these drawbacks, we argue that attention should also be directed at innovative methods for collecting behavioral data.

Recently, we proposed a novel data collection approach called *Parasocial Consensus Sampling* (PCS) [1] to inform virtual human nonverbal behavior generation. Instead of interacting face-to-face, participants were guided through a "parasocial" interaction in which they attempted to produce natural nonverbal behaviors to prerecorded videos of human interaction partners. Through this method we were able to quickly collect large amounts of behavioral data, but more importantly, we were able to assess how multiple individuals might respond to the identical social situation. These multiple perspectives afford the possibility of driving virtual humans with the consensus view on how one should respond, rather than simply concatenating many idiosyncratic responses. A test of this approach, applied to the problem of generating listener nonverbal feedback, showed that 1) participants felt comfortable producing behavior in this manner and 2) the resulting consensus perceived more accurate and more effective than natural feedback (i.e., feedback from the natural listener in face-to-face conversation). Although this was a promising first step, it remains to demonstrate that consensus data can be used to train an effective predictive model.

In this article, we take this next logical step in demonstrating the power of the PCS: using consensus data, we train a predictive model of listener backchannel feedback. We compare the performance of this model against our previous Rapport Agent that generated behaviors according to a hand-crafted mapping. Our subjective evaluation shows the virtual human driven by this probabilistic model performs significantly better than the Rapport Agent [6] in terms of rapport, perceived accuracy and naturalness, and it is even better than the virtual human driven by real listener's behavior in some cases.

2 Background: Parasocial Consensus Sampling

Horton and Wohl [19] first introduced the concept of parasocial interaction. This describes people's natural tendency to interact with media representations of people as if

they were interacting face-to-face with the actual person. Many researchers [20,29,30] have documented that people readily produce such "parasocial" responses and these responses bear similarity to what is found in natural face-to-face interactions, even if the respondents are clearly aware they are interacting with pre-recorded media. By exploiting this characteristic of humans, we proposed the parasocial consensus sampling framework [1].

Parasocial Consensus Sampling is a new methodological framework that collects typical human responses in social interactions.

Unlike the traditional way to collect human behavioral data, where participants' behaviors are recorded during the social interaction, *parasocial* consensus sampling guides multiple independent individuals to vicariously experience the same media representation of social interaction in order to gain the typicality (i.e., consensus view) of human response.

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 participants smile at certain points 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 media representation of an interaction. It 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 responses.

Although we can never know how every person will respond to a given situation, *sampling* is a way to estimate the consensus by randomly selecting individuals from a given 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 observers). 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.

2.1 Definition

We define parasocial consensus sampling as a composite of five elements:

(1) *Interactional Goal*: this is the intended goal of the virtual human interactional behaviors. Before participating in parasocial consensus sampling, participants should be explicitly or implicitly encouraged to behave in a manner which is consistent with this goal, for example, creating rapport.

(2) *Target behavioral response*: this is the particular response or set of responses that the virtual human is going to generate in order to create a specific interactional goal. Participants should be encouraged to produce such behaviors when they are participating in the parasocial interaction. Candidate behavioral responses include backchannel feedback, turn-taking, evaluative facial expressions and paraverbals such as "uh-huh".

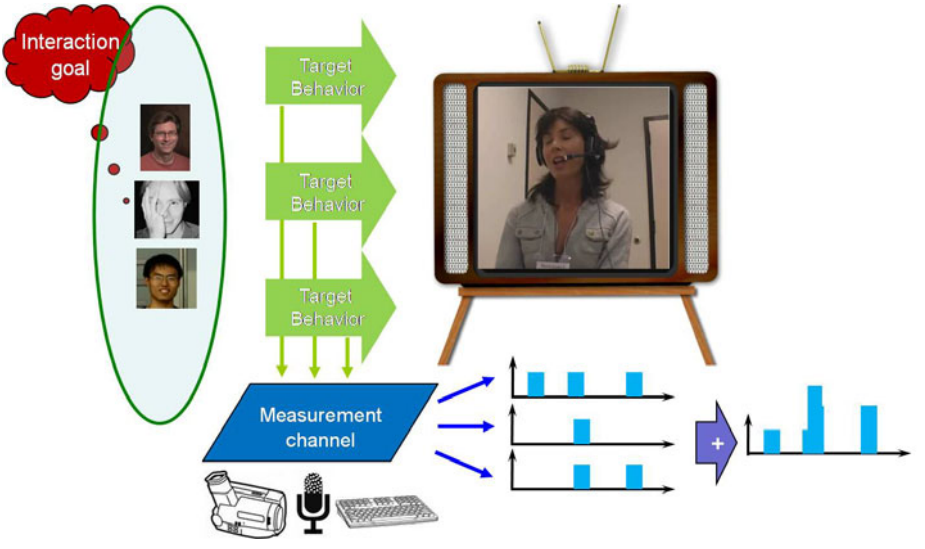

Fig. 1. Parasocial Consensus Sampling (PCS) works as follows: we first recruit participants from some *population*, and then encourage them to give *particular responses* (e.g. backchannels, facial expressions, and so on), measured via some *channel* (i.e. visual channel, audio channel, and mechanical channel), in order to create the *interactional goal* within the parasocial interaction with the *media* representation of social interaction.

(3) *Media*: this is the set of stimuli that will be presented to the 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.

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

(5) *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 participating in face-to-face interaction and record their responses. However, to take advantage of the imaginary nature of parasocial interaction, participants might be encouraged to elicit responses in a more easily measured way. For example, if we are interested in the consensus of when to smile in an interaction, we can ask participants to exaggerate the behavior or even to press a button whenever they feel the response is appropriate. Candidate measurement channels include the visual channel (e.g. videotaping), audio channel (e.g. voice recording) or mechanical channel (e.g. press a button).

2.2 PCS in Action: Collect Listener Backchannel Feedback

Prior research [2,4] has suggested that backchannel feedback [31] plays an important role in establishing rapport between interactants and this paper is going to learn a probabilistic model to predict the backchannel feedback. First, we illustrate how to apply parasocial consensus sampling framework to collect listener backchannel feedback data.

Parasocial consensus sampling consists of five key elements: interactional goal, target behavioral response, media, target population and measurement channel. We customized the parasocial consensus sampling in our work as follows:

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

We recruited 9 fluent English speakers (2 female, 7 males) from a local temporary employment agency to participate in the parasocial interactions with the human speaker videos from our previously collected corpus of face-to-face interactions [5]. The average age of the participants is 45.2 years old, and the standard deviation is 12.6. 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 nods 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, which is much more efficient than the original approach that collecting behavioral data from face-to-face interaction. They gave about 18000 backchannel feedback in total; on average, it is about 7 or 8 backchannels per minute. In next section, we are going to show how to learn a probabilistic model from the parasocial consensus sampling data.

3 Learning a Probabilistic Model from PCS

To learn probabilistic models from parasocial consensus sampling data, we must build a consensus model from the individual parasocial coders and then uses this consensus data to learn a probabilistic model. One advantage of learning from a consensus is it separates what is idiosyncratic from what is essential. Our goal is to learn a probabilistic model which will generalize the PCS data to new sequences (or live interactions) not seen in the training set. The probabilistic model is trained from the speaker's actions (e.g., pause, eye gaze, and specific lexicon words) to predict the listener backchannel feedback (i.e., head nods).

3.1 Building Consensus

The backchannel PCS dataset described in Section 2.2 consists of *N* sets of parasocial responses: T_1, T_2, ..., T_N , where N is the number of participant. For each parasocial

interaction T_i, the PCS dataset contains the response timestamps $T = \{t_1, t_2, \ldots\}$ indicating when the participant gave a response. These response timestamps are combined to create the 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. Following the work of Ward and Tsukahara [4], we create a one second time window centered about each timestamp. The timeline is then sampled at a constant frame rate of 10Hz [4]. Figure 2 illustrates this approach.

Fig. 2. t1, t2, t3 are the time spots when a participant gives backchannel feedback in a parasocial interaction. A 1.0s window of opportunity is put around each timestamp so that the time spot is in the middle of the window. The samples within the window are set to 1 to indicate the presence of feedback, while others are set to 0.

(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 [27,28]. Therefore, the quality of consensus data can be improved if we first factor out these individual differences before combining response timestamps into a consensus. We can estimate this delay by comparing the parasocial interaction with the face-to-face interaction. We follow the approach in [2,4] to count how often PCS matches the real listener's behaviors and find the time offset that maximizes this score. This process was repeated independently on the nine participants of the PCS data. The reaction time values varied from 600ms to 1200ms, with average of 970ms. The 10 video sequences used for our subjective evaluation described in Section 4 were not part of the video sequences used to select the reaction times.

(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 3 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.

By applying a threshold, the *consensus level*, to the parasocial consensus, feedback with less importance can be filtered out. Following the work in [1], we select a

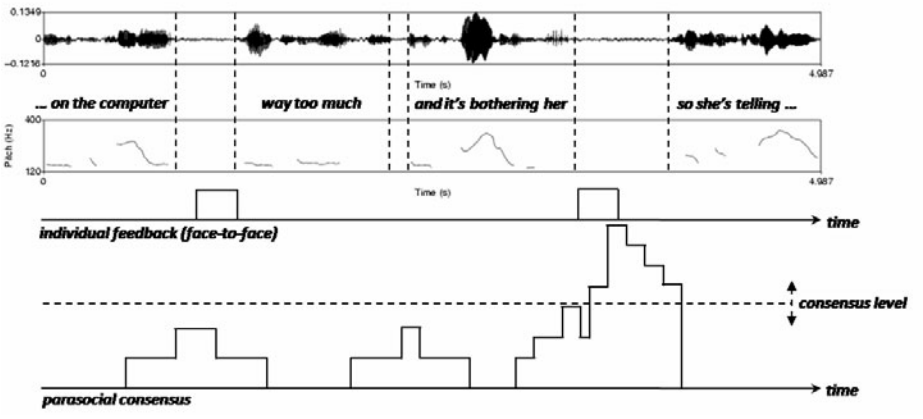

Fig. 3. Example segment showing a parasocial consensus of listener backchannel varies over time. While individual feedback (from the original face-to-face interaction) only gives discrete prediction, our parasocial consensus shows the relative importance of each feedback. By applying a consensus level to the parasocial consensus, we get only important feedback.

consensus level that makes the number of backchannels from parasocial consensus closest to that from the original face-to-face interaction data.

3.2 Learning Probabilistic Model

To build the predictive model for virtual humans, we find the relationship between speaker's features and the consensus. Recently, there has been seen an explosion in interest in data-driven approaches that automatically find such patterns using machine learning methods [2,9,10,15]. Given the time-series nature of human behavior, sequential model is a good one to learn the internal dynamic structure existing in human behavior. We apply a similar strategy as [2] to learn a *Conditional Random Field* (CRF) model from parasocial consensus sampling data. This method takes as input a sequence of human speaker's features and returns a sequence of probabilities to give backchannel feedback.

Although semantic information is an important feature in predicting backchannel feedback, it has been mentioned in other work [2,4] that non-verbal information itself also provides lots of clues in backchannel prediction. In this paper, we try to push the state of the art of non-verbal feature based models. Four speaker features are selected as suggested in [2]:

- *Pause* using binary encoding
- *Speaker looking at the listener* using ramp encoding with a width of 2 seconds and a 1 second delay
- *'and'* using step encoding with a width of 1 second and a delay of 0.5 seconds
- *Speaker looking at the listener* using binary encoding

All the features mentioned above were hand labeled by coders. While training, we split the data set (the videos used for evaluation in Section 4 are not included) into

training set and validation set. This is done by N-fold cross validation. This means N-1 folders are used for training, and the remaining folder is used as validation data for testing the model. This process is repeated N times, and then the best model is selected based on the performance of our models. The performance is measured by F_1 score, which is the harmonic mean of precision and recall. Precision is the probability that predicted backchannels correspond to actual listener behavior; recall is the probability that a backchannel produced by an actual listener was predicted by the model.

Given new test sequence, CRF outputs probability over time to indicate the likelihood of giving backchannel feedback. The local maximum of the probability are selected as the candidates. In order to generate the final backchannel feedback, we have to pick up a feedback level as shown in Figure 4. In this paper, we set the feedback level so that the number of feedback from CRF model is closest to that from the training set.

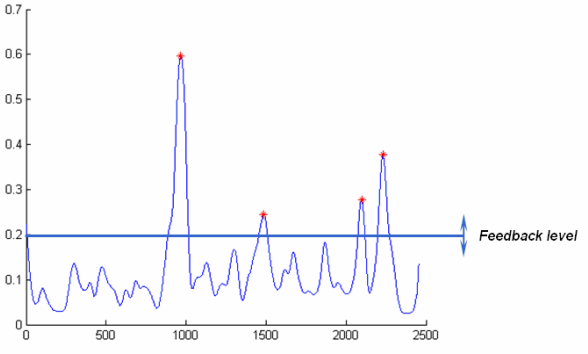

Fig. 4. Generate the final backchannel feedback by applying the feedback level to the output of CRF model. The stars (*) are the final backchannels.

4 Subjective Evaluation

In evaluating the performance of the probabilistic model, we conduct a subjective evaluation experiment to assess whether the virtual human driven by the CRF model can be used to achieve the interactional goal: creating rapport, when compared against the Rapport Agent and the original human listener. Specially, we compose videos illustrating a human speaker interacting with the virtual human (Figure 5) and contrast subjective impressions of different models for generating the virtual human's behavior.

We claim that a potential advantage PCS over traditional training methods is that the consensus data better reflects the intended interactional goal than typical face-to-face data. To better assess this claim we assess the approach against three classes of face-to-face interactions: high-rapport interactions where the original human listener exhibited high rapport; low-rapport interactions where the original human listener exhibited low rapport, and "typical" interactions that contain a mixture of both.

Fig. 5. Videos for subjective evaluation

4.1 Backchannel Prediction Models

We selected 10 speaker videos not used in training the CRF model. When these face-to-face interactions were originally conducted, speakers were asked to assess the rapport they felt with their conversation partner. Five videos were those from our corpus with the lowest rapport score and 5 were those with the highest rapport score. We created three variants of each of these videos, replacing the human listener with a virtual human whose behavior was driven by one of three different prediction models:

(1) *PCS-CRF*: the virtual human is driven by the CRF model trained on parasocial consensus. The training set doesn't include the 10 videos used for evaluation.

(2) *Natural*: the virtual human is driven by the real listener's backchannel feedback from the original face-to-face interaction.

(3) *Rapport Agent*: Gratch et al. [6] built the Rapport Agent by applying a rule-based model to predict when to give backchannel feedback. The backchannels were predicted from two rules: (a) If the speaker nods, the listener should nod back, (b) if there are backchannel opportunities in the speaker's speech, the listener should nod back. The Rapport Agent uses Watson [26] to detect head nods and LAUN [6] to detect backchannel opportunities using the approach of Ward and Tsukahara [4]. We replicate the Rapport Agent's behavior by using the same two tools to extract features from human speaker videos and applying the same rules for backchannel prediction.

4.2 User Study

We recruited 17 participants to evaluate the quality of the virtual human's behavior. Before watching videos, 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 agent's head nods." After watching each video, participants evaluated the virtual human's behavior by answering 7 questions:

Rapport Scale:

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

Perceived accuracy:

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

Naturalness:

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

4.3 Results

ANOVA test is applied to find whether there is significant difference among the three versions. The four items related to rapport are averaged into a single scale that showed good reliability (Cronbach's alpha = 0.98).

The results are summarized from Figure 6 to 9. 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 start (*) means there is significant difference between the versions under the bracket.

4.3.1 Rapport Scale

Overall, the virtual human driven by the CRF model (PCS-CRF) is significantly better than the Rapport Agent [6]. 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 listener's behavior, we don't see significant difference overall, but there is significant

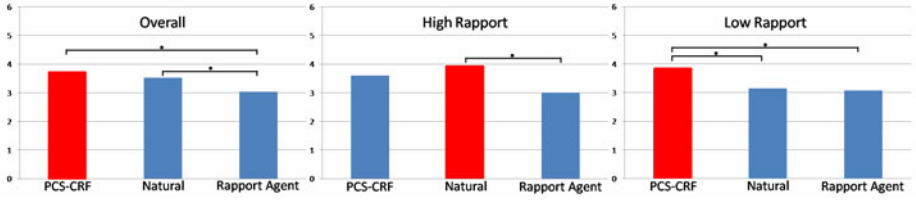

Fig. 6. *Rapport Scale*. Overall, the virtual human driven by CRF is significantly better than Rapport Agent. For low-rapport videos, the virtual human driven by CRF is significantly better than the one driven by real listener's behavior.

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 and do better than those who fail to.

4.3.2 Perceived Accuracy

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 (Fig. 8), while the precision is low (Fig. 7). 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 give enough appropriate backchannel responses 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 create rapport at the same time.

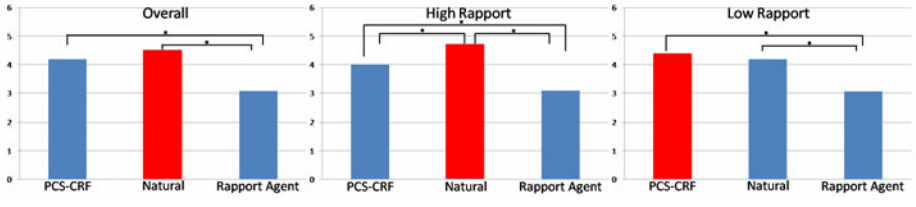

Fig. 7. *Precision.* The virtual human driven by CRF provides backchannel feedback more precisely than the Rapport Agent.

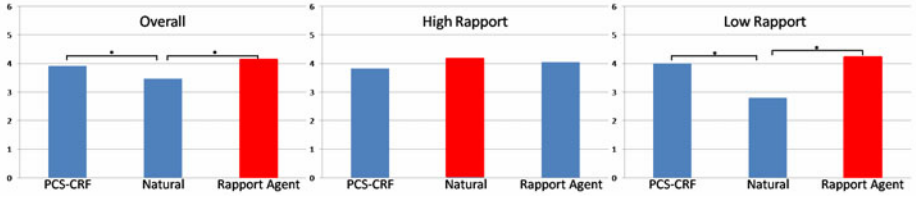

Fig. 8. *Recall*. The virtual human driven by real listener's behavior misses more opportunities to provide backchannel feedback than the other two versions do.

4.3.3 Naturalness

By comparing the Natural question (Fig. 9) with the Rapport Score question (Fig. 6), we find the virtual human is perceived more natural when it creates more rapport

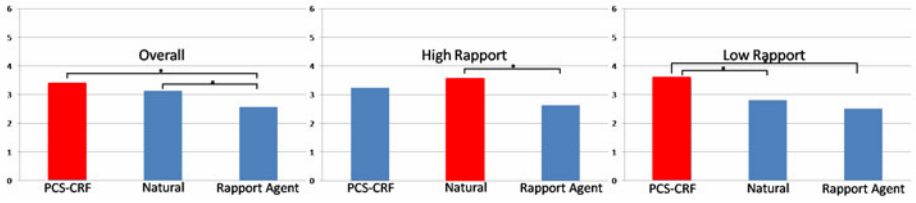

Fig. 9. *Natural.* Overall, The virtual human driven by CRF is more natural than Rapport Agent. For low-rapport videos, the virtual human driven by CRF is more natural than the one driven by real listener's behavior.

within the interaction, which confirms previous finding that creating rapport does lead to positive social effects.

5 Conclusion and Future Work

In this paper, we learned a probabilistic model for predicting listener backchannel feedback from parasocial consensus sampling data. By comparing the virtual humans driven by (1) CRF model trained on PCS data, (2) real listener's behavior and (3) Rapport Agent's behavior, we found that the virtual human driven by CRF model is significantly better than the one driven by Rapport Agent's behavior, and it has almost the same performance as the one driven by real listener's behavior. The result demonstrated we could learn a better prediction model from PCS data and proved the validity of this data collection framework in advance. As future work, we are planning to assess the power of the PCS framework with other conversation cues such as smiling, turn-taking and interruptions, other interactional goals besides rapport, and other measurement channels, such as vision-based methods.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 0729287. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Huang, L., Morency, L.-P., Gratch, J.: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In: Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems (2010)
2. Morency, L.-P., de Kok, I., Gratch, J.: Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 176–190. Springer, Heidelberg (2008)
3. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating Rapport with Virtual Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)

4. Ward, N., Tsukahara, W.: Prosodic features which cue backchannel responses in English and Japanese. J. Pragmatics 23, 1177–1207 (2000)
5. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Werf, R.J., Morency, L.-P.: Virtual Rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
6. Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., Morency, L.-P.: Can Virtual humans be more engaging than real ones? In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 286–297. Springer, Heidelberg (2007)
7. Bailenson, J.N., Yee, N.: Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science 16, 814–819 (2005)
8. Bailenson, J.N., Yee, N., Merget, D., Schroeder, R.: The Effect of Behavioral Realism and Form Realism of Real-Time Avatar Faces on Verbal Disclosure, Nonverbal Disclosure, Emotion Recogition, and Copresence in Dyadic Interaction. PRESENCE: Teleoperators and Virtual Environments 15(4), 359–372 (2006)
9. Lee, J., Marsella, S.: Learning a Model of Speaker Head Nods using Gesture Corpora. In: 8th International Conference on Autonomous Agents and Multiagent Systems (2009)
10. Jonsdottir, G.R., Thorisson, K.R., Nivel, E.: Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 162–175. Springer, Heidelberg (2008)
11. de Melo, C., Gratch, J.: Expression of Moral Emotions in Cooperating Agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 301–307. Springer, Heidelberg (2009)
12. Bickmore, T., Puskar, K., Schlenk, E., Pfeifer, L., Sereika, S.: Maintaining Reality: Relational Agents for Antipsychotic Medication Adherence. J. Interacting with Computers special issue on Mental Health (2010)
13. Kang, S.-H., Gratch, J., and Watts, J. The Effect of Affective Iconic Realism on Anonymous Interactants' Self-Disclosure. In: Proceedings of Interaction Conference for Human-Computer Interaction (2009)
14. Wang, N., Gratch, J.: Can a Virtual Human Build Rapport and Promote Learning? In: Proceedings of 14 International Conference on Artificial Intelligence in Education (2009)
15. Kipp, M., Neff, M., Kipp, K.H., Albrecht, I.: Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 15–28. Springer, Heidelberg (2007)
16. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F.: Embodied Conversational Agents. MIT Press, Cambridge (2000)
17. Gratch, J., Rickel, J., Andre, E., Badler, N., Cassell, J., Petajan, E.: Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems, 54–63 (July/August 2000)
18. Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., Slater, M.: Building Expression into Virtual Characters. In: Eurographics 2006 (2006)
19. Horton, D., Wohl, R.R.: Mass communication and parasocial interaction: Observation on intimacy at a distance. Psychiatry 19, 215–229 (1954)
20. Levy, M.R., Watching, T.V.: News as parasocial interaction. J. Broadcasting 23, 60–80 (1979)
21. Heylen, D.: Understanding Speaker-Listener Interactions. In: Proceedings of 10th Annual Conference of the International Speech Communication Association (2009)
22. Bavelas, J.B., Coates, L., Johnson, T.: Listener Responses as a Collaborative Process: The Role of Gaze. J. Communication 52(3), 566–580 (2006)

23. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. J. Personality and Social Psychology 79(6), 941–952 (2000)
24. Bernieri, F.J., Gillis, J.S., Davis, J.M., Grahe, J.E.: Dyad Rapport and the Accuracy of Its Judgment Across Situations: A Lens Model Analysis. J. Personality and Social Psychology 71(1), 110–129 (1996)
25. Gifford, R.: A Lens-Mapping Framework for Understanding the Encoding and Decoding of Interpersonal Dispositions in Nonverbal Behavior. J. Personality and Social Psychology 66(2), 398–412 (1994)
26. Morency, L.-P., et al.: Contextual Recognition of Head Gestures. In: Proceedings of 7th International Conference on Multimodal Interactions (2005)
27. Montare, A.: The simplest chronoscope: group and interindividual differences in visual reaction time. J. Perceptual and motor skills 108(1), 161–172 (2009)
28. Reaction time, http://en.wikipedia.org/wiki/Reaction_time
29. Houlberg, R.: Local television news audience and the para-social interaction. J. Broadcasting 28, 423–429 (1984)
30. Rubin, A.M., Perse, E.M., Powell, R.A.: Loneliness, para-social interaction, and local television news viewing. Human Communication Research 12, 155–180 (1985)
31. Yngve, V.: On Getting a Word in Edgewise. In: 6th Regional Meeting of the Chicago Linguistic Society, pp. 567–577.