Crowdsourcing Backchannel Feedback: Understanding the Individual Variability from the Crowds

Lixing Huang, Jonathan Gratch

Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, California, 90094
lhuang@ict.usc.edu, gratch@ict.usc.edu

Abstract

During conversation, listeners often provide so-called backchannel feedback (e.g., nods and filled pauses) during their partner's speech and these behaviors serve important interactional functions. For example, the presence of backchannels has been shown to cause increased rapport, speech fluency and speech intimacy, even when produced by computer-generated listeners. Prior work by us and others has shown that specific acoustic and visual features predict when backchannels are likely to occur, but there is also considerably individual variability not explained by such models. Here we explore a data collection framework known as Parasocial Consensus Sampling (PCS) to examine and characterize some of this individual variability. Our results indicate that common personality traits can capture much of this variability. This suggests we can build models that capture individual differences in backchannel "style" and possibly identify individual traits from observations of backchannel behavior.
Index Terms: Backchannel, crowdsourcing

1. Introduction

Face-to-face interaction is a cooperative process. While the speaker is talking, the listener's nonverbal reactions, such as head nod, paraverbal and facial expression, also provide moment-to-moment feedback that can alter and serve to co-construct subsequent speech. These nonverbal behaviors are called backchannels, and they play an important role in efficient social interactions [1,7,8].

Several research efforts have attempted to study and model the characteristics of a speaker's behavior that predict listener backchannel feedback [3-5,9]. Increasingly, these efforts employ data-driven approaches that automatically learn such models from large amounts of annotated face-to-face interaction data [3,9]. Although face-to-face interaction data is traditionally considered as gold standard, it presents several drawbacks. First, there is considerable variability in human behavior, and not all human data should be considered as positive examples of the behavior that we want to model. For example, if we want to find out how backchannel feedback helps establish the feeling of rapport, it is important to realize that many face-to-face interactions fail in this regard. Ideally, such data should be separated into good and bad instances of the target behavior, but it's not obvious how to make this separation. Second, face-to-face interactions are co-constructed in that the behaviors of individuals not only depend on their own specific characteristics, but also 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 [7]. This mutually contingent nature of social interactions amplifies the underlying variability of human behavior but also makes it difficult to tease apart causality (i.e. is this person a non-engaging speaker, or is he reacting to a disengaged listener). These issues are not insurmountable but they imply that we need collect large amounts of data to surmount them, which brings us to the third problem: the traditional way of recording face-to-face interaction data is expensive and time-consuming. It can take months to recruit participants, followed by an extensive period of recording and annotating the data.

To address these issues, we have previously proposed a data collection technology called Parasocial Consensus Sampling (PCS) [5]. It is inspired by the theory of parasocial interaction introduced by Horton and Wohl [2], where they argued that people exhibit a natural tendency to interact with media representation (e.g. video recordings) of people as if they were interacting with the actual person face-to-face. The basic idea of PCS 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 would behave within face-to-face interactions. This approach helps address the three issues of the traditional face-to-face interaction data.

Variability: By having different people experience the same social situation, we can aggregate their behaviors and count the probability of each possible behavior from the consensus view, which represents how likely the behavior should happen.

Contingency: The usage of media representation of people breaks down the contingency in face-to-face interaction, because we hold the behavior of one participant constant (e.g. a pre-recorded speaker cannot react to listener feedback). This can help to unpack the bidirectional causal influences that naturally occur in conversations, but it might destroy the very phenomenon we wish to study (i.e., by preventing speakers from reacting to listener feedback, it might change the nature of this feedback). Fortunately, this hasn't occurred in practice, at least for modeling backchannel and turn taking behaviors, as the learned models are similar to those learned with face-to-face data and produce similar social effects when used to drive conversational behaviors with virtual humans [5,6].

Efficiency: By using media representation, we can parallelize the data collection process (i.e. different people interact with the same social situation simultaneously), and the parasocial interaction also enables us to use more efficient way to measure human behavior. For example, in a previous study [12], 9 participants interacted with 45 speaker videos parasocially in just *one* day. They were asked to press a button whenever they felt like to give backchannel feedback so that we can record the time automati-

cally instead of having coders to annotate when the backchannel feedback happened.

In this paper, we apply the Parasocial Consensus Sampling on a much larger scale. Specifically, we are crowdsourcing backchannel using Amazon Mechanical Turk (AMT). This allows us to collect hundreds of responses to each video in a much faster and less expensive way, compared with traditional approaches. Due to the large amount of data we have, it is now possible to analyze and explain individual variability in backchannel feedback. The following section describes the data collection procedure and the visualization tool for exploring the crowdsourcing dataset. Section 3 illustrates the variability in backchannel production and presents our preliminary results in explaining such variability. Section 4 concludes the paper.

2. Data Collection and Visualization.

2.1. Data Collection

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 specify prices for completing them. The idea is to utilize people's (i.e. worker) 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 workers repeat the same task. Although the individual worker is usually not an expert for the task, one often can achieve expert-level results by relying on the wisdom of crowds [10].

We implemented a web application and integrated it with AMT. Workers can find our tasks on the marketplace and follow a link to our website. First, they finish a 90-item questionnaire that assesses several individual differences that we expect to influence backchannel behavior (listed in Section 3.3). They next watch an example video illustrating the process of interacting with a human speaker parasocially. Next, they watch 8 videos in sequence, each about 2 to 3 minute long. Each video features a human speaker telling two stories. Coders are instructed to pretend that they are very interested in what the speaker says. Whenever they think it is a good time to provide feedback, such as nodding or uttering "uh-hum", to the speaker, they press a button. The timestamp of each press is recorded and sent to our server by using JavaScript. After interacting with each video, coders answer a 6-item questionnaire regarding their parasocial experiences [11]. 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 dollars for their work. Following this procedure, we initially constructed a dataset of 350 coders providing backchannel data for 8 videos. To better understand speaker variability, we subsequently coded additional 16 videos (in two rounds) using 100 coders each. For the analysis that follows, we collapse these three data collections into a single dataset of 24 videos.

2.2. Visualization

For each video, we have N sets of parasocial responses (T_1, $T_2 \ldots T_N$) from N coders. Each set of parasocial responses T_i, contains the timestamps $T_i = \{t_1, t_2 \ldots\}$ representing when the coder gave a response. Each timestamp can be viewed as a window of opportunity where a backchannel feedback is likely. We create a one second time window centered around each timestamp, and the timeline is sampled at a frame rate of 10Hz. In this way, we convert the timestamps $\{t_1, t_2 \ldots\}$ into time series data, as shown in Figure 1, where 1 indicates that backchannel feedback occurs and 0 indicates that no backchannel feedback occurs.

Figure 1: *Example timeline of feedback from a single coder. Each line indicates a point where backchannel feedback occurs.*

A consensus view is created by summing together the data from each coder for a specific video. The result is equivalent to a histogram that indicates how many coders felt a particular point in the speech merited feedback. This is illustrated in Figure 2. Peaks in the consensus view indicate time points where there is high agreement for providing feedback.

Figure 2 also shows a visualization tool that helps explore the PCS data. By selecting a video ID from the video table (Part 4), the corresponding video (Part 1) and coders (Part 5) show up; the consensus view (Part 2) of that video will be computed and also show up. By selecting a coder ID from the coder table (Part 5), the parasocial response of the coder (as shown in Figure 1) will show up; if multiple coders are selected, a histogram will be computed by using the responses from those coders (Part 3). As described in Section 2.1, we measure several personality attributes of each coder using standard questionnaires. By selecting an attribute (Part 7), the coder table (Part 5) will be populated with the corresponding values of all coders. And a histogram (Part 6) will be displayed, indicating the distribution of coders along the selected attribute. This helps us group coders and investigate how individual difference (e.g. personality traits) affects backchannel feedback. As the video plays, a timeline (the red vertical line) will move correspondingly so that we can compare the consensus view with the human speaker's behavior.

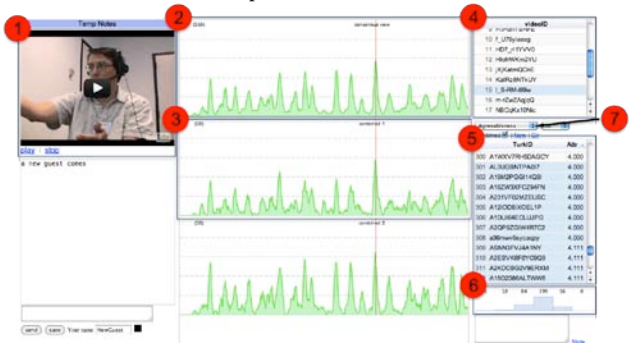

Figure 2: *The interface of the visualization tool. (2) represents the consensus of all coders for video (1), and the two histograms below (2) represent the consensus of two sub-groups of coders: those that are least and most agreeable, respectively.*

3. Analysis of Backchannel Consensus Data

3.1. Variability

PCS provides a unique tool to characterize some of the observed variability in backchannel production. Given that many coders are responding to the same speaker, we can ask if this variance arises from characteristics of speakers or aspects of listeners. At

one extreme, all the variance might reside in the listener: for example, coders might be providing feedback or their behavior could be governed by individually varying personality traits. If so, we should expect a uniform distribution of feedback across the speaker's story. At the other extreme, feedback might be solely determined by characteristics of the speaker, in which case we should expect perfect consensus amongst coders but variance across speakers. Figure 2 (which illustrates the consensus view), indicates that the answer is somewhere in the middle: there are many points of high coder agreement but also considerable variability within and across speakers.

A quick examination of this data suggests that peaks in the consensus data correspond to behavior of speakers that have been previously suggested as backchannel elicitors. For example, peaks often co-occur with to speaker non-filled pauses. Peaks also correspond to semantically significant events in the story. For example, in Figure 2, the highest peak corresponds to a climactic moment in the narrative.

However, there is also considerable variability across coders. For example, Figure 3 gives a histogram illustrating the amount of feedback provided by different coders for a given video. Feedback varied from no feedback at all to 64 responses by the most prolific coder. We now turn to a more formal analysis of listener variability.

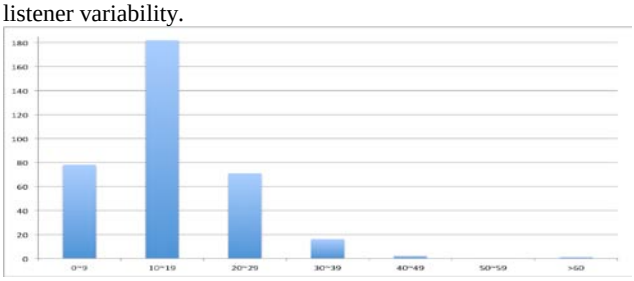

Figure 3: *We group coders into 7 bins (0-9, 10-19,…,50-59, >60) according to the number of feedback they provided. x-axis represents the bins, and y-axis represents the number of coders in the corresponding bin.*

3.2. Speaker Nonverbal Features

Clearly, some of the variance of this data 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 several features previously implicated in backchannel production. For each speaker feature, we have the starting time t_s and end time t_e that have been labeled by human coders. For listener backchannel feedback, we record the time (t_b) when the coder pressed the button. We count it as a match if

$$t_s \leq t_b \leq t_e + window$$

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 [9], we add the variable "*window*" to count the case where backchannel feedback is triggered by speaker features but with a certain delay. In this study, *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:

$$P = \frac{\#\ of\ cooccurrence}{\#\ of\ occurrence\ of\ a\ feature}$$

$$R = \frac{\#\ of\ cooccurrence}{\#\ of\ backchannel\ feedback}$$

And then, the importance I is calculated as the harmonic mean of P and R. By ranking the speaker features on the importance measure 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. The result suggests that coders use almost the same subset of speaker features to decide when to give feedback, which cannot explain the individual variability in backchannel feedback.

3.3. Individual Difference and Backchannel Feedback

Only some of the observed variance can be explained by speaker behavior and here we examine how personality traits might impact backchannel production. Table 1 lists several individual traits of coders that we are currently investigating.

Table 1. *The attributes of each coder we measured in Section 2.1*

Big Five Personality Traits	Extroversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Self-Consciousness	Self-directed, Other-directed
Parasocial Experience	Parasocial experience scale [11]
Other	Shyness, Self-monitoring, Gender

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, we group the coders whose values are the lowest into the *low_group*, and those with the highest values into the *high_group*[1]. We compute three numbers to represent each group as follows:

(1) For each video, a consensus is computed by using the data from all coders as described in Section 2.2;
(2) For each group, a histogram is computed by using the data from the coders in the corresponding group;
(3) We sum up the histogram computed in step (2) 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;
(4) We calculate the ***correlation coefficient*** between the consensus and the histogram of each group;
(5) We compute the ***entropy*** of the histogram of each group. This can be considered as a measurement of agreement among coders.

Finally, we have the three numbers for each group, which are: the average number of feedback, the correlation coefficient, and the entropy. And the three numbers are computed for every video. T-Test is used to find whether there is significant differ-

[1] We calculate mean μ and standard deviation σ from all coders. *low_group* has coders whose values are less than $\mu - \sigma$, while *high_group* has coders whose values are larger than $\mu + \sigma$.

ence between the *high_group* and the *low_group*. The results are summarized as follows.

Table 2. *The average amount of feedback*

	high	low	t-test
Extroversion	*173.51*	*172.61*	*p=0.88*
Agreeableness	*190.72*	*166.59*	*p<0.01*
Conscientiousness	*207.72*	*178.76*	*p<0.01*
Neuroticism	*180.12*	*182.02*	*p=0.67*
Openness	*203.58*	*171.52*	*p<0.01*
Self-consciousness	*210.17*	*173.71*	*p<0.01*
Other-consciousness	*205.76*	*171.04*	*p<0.05*
Shyness	*175.51*	*178.26*	*p=0.66*
Self-monitor	*206.03*	*166.97*	*p<0.01*
PSI	*203.06*	*163.50*	*p<0.01*
Gender (F/M)	*182.71*	*181.73*	*p=0.69*

Table 3. *Correlation Coefficient*

	high	low	t-test
Extroversion	*0.90*	*0.91*	*p=0.42*
Agreeableness	*0.89*	*0.87*	*p<0.05*
Conscientiousness	*0.90*	*0.88*	*p<0.05*
Neuroticism	*0.91*	*0.89*	*p<0.01*
Openness	*0.91*	*0.89*	*p<0.05*
Self-consciousness	*0.88*	*0.89*	*p=0.18*
Other-consciousness	*0.88*	*0.85*	*p<0.01*
Shyness	*0.91*	*0.88*	*p<0.05*
Self-monitor	*0.87*	*0.89*	*p<0.01*
PSI	*0.88*	*0.89*	*p<0.05*
Gender (F/M)	*0.98*	*0.97*	*p=0.08*

Table 4. *Entropy*

	high	low	t-test
Extroversion	*6.68*	*6.61*	*p<0.05*
Agreeableness	*6.65*	*6.68*	*p<0.05*
Conscientiousness	*6.74*	*6.73*	*p=0.62*
Neuroticism	*6.63*	*6.67*	*p=0.12*
Openness	*6.69*	*6.64*	*p=0.13*
Self-consciousness	*6.67*	*6.64*	*p=0.19*
Other-consciousness	*6.72*	*6.66*	*p<0.05*
Shyness	*6.61*	*6.60*	*p=0.48*
Self-monitor	*6.69*	*6.64*	*p<0.05*
PSI	*6.72*	*6.61*	*p<0.01*
Gender (F/M)	*6.77*	*6.80*	*p<0.01*

Table 2 shows the difference of the average feedback number between the *high_group* and the *low_group* for each attribute, Table 3 shows the difference of correlation coefficient, and Table 4 shows the difference of entropy. It is clear that individual difference has significant influences on backchannel feedback. For example, the *high_group* of agreeableness tends to provide more backchannel feedback than the *low_group*, and they have more agreement among each other than the coders in the *low_group*; the coders who have good parasocial experience tend to provide more backchannel feedback but have less agreement than the coders who have bad parasocial experience; there is no significant difference between male and female except female coders tend to have more agreement.

3.4. Discussion

There is significant individual variability in listener backchannel feedback. However, from the feature analysis (Section 3.2), we find that coders depend on almost the same subset of speaker features to provide backchannel feedback, indicating there may be consistent backchannel feedback rules. The correlation coefficient (≈0.9) between the histogram of groups of coders and the consensus also suggests the same thing. Our preliminary analysis (Section 3.3) suggests that the reason underlying the significant individual variability may be the individual differences, such as personality traits and parasocial experience, among coders.

4. Conclusions

In this paper, we presented our work in analyzing individual variability in backchannel feedback. Under the Parasocial Consensus Sampling (PCS) framework, we applied the crowdsourcing technique to collect hundreds of "listeners" backchannel feedback to one human speaker from the web. The results showed that there is significant individual variability in backchannel feedback; however, people depend on almost the same subset of human speakers' features to provide feedback. Our preliminary analysis suggests that the reason underlying such individual variability is the individual differences, such as personality traits and parasocial experience, among coders. Our work also demonstrates the advantage of Parasocial Consensus Sampling framework. By breaking down the contingency in face-to-face interaction, it is possible to run such analysis that cannot be done by using the traditional dataset.

5. References

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

[2] Horton, D. and Wohl, R.R. Mass communication and para-social interaction: Observation on intimacy at a distance. Psychiary 19: 215-229, 1956.

[3] Ward, N. and Tsukahara, W. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 32:1177-1207, 2000.

[4] de Kok, I.A. and Heylen, D.K.J. Appropriate and Inappropriate Timing of Listener Responses from Multiple Perspectives. Proceedings of 10th IVA, 248-254, 2011.

[5] Huang, L, Morency, L.-P. and Gratch, J. Parasocial Consensn Sampling: combining multiple perspectives to learn virtual human behavior. Proceedings of 9th AAMAS, 1265-1272, 2010.

[6] Huang, L., Morency, L.-P. and Gratch, J. A Multimodal end-of-turn Prediction Model: Learning from Parasocial Consensus Sampling. Proceedings of 10th AAMAS, 1289-1290, 2011.

[7] Bavelas, J.B., Coates, L. and Johnson, T. Listeners as co-narrators. Journal of Personality and Social Psychology, 79: 941-952, 2000.

[8] Bavelas, J.B. and Gerwing, J. The listener as addressee in face-to-face dialogue. International Journal of Listeneing. 2011.

[9] Morency, L.-P., de Kok, I. and Gratch, J. Predicting listener back-channels: A probabilistic multimodal approach. Proceedings of 8th IVA, 176-190, 2008.

[10] Surowiecki, J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies, and nations. Doubleday Books. 2004.

[11] Hartmann, T. and Goldhoorn, C. Horton and Wohl revisited: Exploring viewer's experience of parasocial interactions. Annual meeting of the International Communication Association. 2010.

[12] Huang, L., Morency, L.-P. and Gratch, J. Learning backchannel prediction model from parasocial consensus sampling: a subjective evaluation. Proceedings of 10th IVA, 159-172, 2010.