Explaining the Variability of Human Nonverbal Behaviors in Face-to-Face Interaction

Lixing Huang and Jonathan Gratch

Institute for Creative Technologies, University of Southern California
lxhuang1984@gmail.com, gratch@ict.usc.edu

Abstract. Modeling human nonverbal behaviors is a key factor in creating a successful virtual human system. This is a very challenging problem because human nonverbal behaviors inherently contain a lot of variability. The variability comes from many possible sources, such as the participant's interactional goal, conversational roles, personality and emotions and so on, making the analysis of the variability hard. Such analysis is even harder in face-to-face interactions since these factors can interact both within and across the participants (i.e. speaker and listener). In this paper, we introduce our initial efforts in analying the variability of human nonverbal behaviors in face-to-face interactions. Specifically, by exploring the Parasocial Consensus Sampling (PCS) framework [13], we show personality has significant influences on listener backchannel feedback and clearly demonstrate how it affects backchannel feedback. Moreover, we suggest that PCS framework provides a general and effective approach to analyze the variability of human nonverbal behaviors, which would be difficult to perform by using the traditional face-to-face interaction data.

Keywords: Parasocial Interaction, Nonverbal Behaviors, Variability, Personality.

1 Introduction and Background

Today, we have seen a few virtual human systems with natural and realistic behaviors in interactive scenarios such as training [1], health care [2] and education [3]. One of the key factors that makes these systems successful is that virtual humans can provide contingent and appropriate feedback to their interactional partners in real time. There have been many efforts in building nonverbal behavior models to predict when and how the virtual human should respond to his interactive partner accordingly. A lot of progress has been made. Originally, researchers depend on the findings from the social psychology literature. They [4] [5] [6] [7] usually derived a set of rules from the literature to drive the virtual human's behavior. However, such descriptive rules are more helpful as general theoretical points than to directly drive a virtual human's behavior as they typically describe general findings and do not precisely characterize the specific circumstance and timing information for when such behaviors should

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 275–284, 2013.

be employed. Recently, researchers [8] [9] [10] start to explore more advanced machine learning techniques to learn behavior models from large amounts of annotated human behavior data. Such approach usually generates quantitative models which can directly be used to drive the virtual human's behavior. More importantly, by changing the dataset that the algorithm learns from, it is feasible to train models that are in line with the context where the virtual human will be applied. However, there are still challenging and unsolved problems.

First, most of the virtual humans can only provide generic feedback. Bavelas et al. [11] proposed that nonverbal feedback can be classified into two classes. One is generic feedback, which is not closely connected to what is being said. Such generic behaviors don't convey any specific meanings, and would be appropriate in different scenarios. The other one is specific feedback, which is tied to a deeper understanding of, and reaction to, the personal relevance of what is being said. Such specific behaviors usually depend not only on the understanding of the semantic meanings but also on our own role and participatory goals, which may change as the conversation unfolds [15]. Currently, most of the virtual human systems address the first type of behavior. For example, the Rapport Agent [7] relies on low level analysis of the nonverbal signals of the human speaker and provides contingent feedback, such as head nod, accordingly. In order to apply the virtual human technology in more complex secnarios, it is inevitable that we need start building models for specific nonverbal behaviors. Second, virtual human needs not only respond to his interactional partner but also be able to reflect his own emotion, personality, and interactional goal. For example, the recent SEMAINE project built the Sensitive Artificial Listener [12]. By exhibiting different styles of audiovisual listener feedback, the listener is able to express four different personalities. However, it is still difficult to perform formal analysis on how personalities can affect nonverbal behaviors in face-to-face interactions.

Because of these reasons, nonverbal behaviors inherently contain a lot of variability. They are affected by many internal factors, such as emotion and personality, and external factors, such as the presence of others and others' responses. These factors interact both within and across participants - for example the emotions of one participant in a conversation can spill over and alter the behavior of other actors - making it difficult to isolate and model the variability of nonverbal behaviors. The problem is not insurmountable but it implies that we will have to collect large amounts of behavioral data. But the traditional way of recording face-to-face interaction data is very expensive and time-consuming. It usually takes months to recruit pairs of participants, followed by an extensive period of manually-annotating the resulting recordings.

To solve these problems, Huang et al. [13] proposed a new approach called Parasocial Consensus Sampling (PCS), where 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. Since multiple participants can now interact with the same social situation, usually pre-recorded videos (e.g. speaker video), we hold one side of the interaction consistent. This helps unpack the bidirectional causal influences that naturally occur in conversations. Moreover, by using pre-recorded speaker videos, we can dramatically increase the efficiency of the data collection process by having multiple participants interact with the same speaker simultaneously.

In this paper, we extend the original Parasocial Consensus Sampling work in two ways. First, we examine a more naturalistic approach (i.e. videotaping) to measure participants' behavior. In the original work, participants were guided to press a button whenever you feel like to respond. Although efficient, it has several limitations. For example, pressing a button demands an explicit conscious decision from a participant and it is difficult to measure multiple behaviors at the same time since pressing different buttons for different behaviors is likely to place too much cognitive load on the participants. In our study, we ask participants to interact with pre-recorded speaker videos and act as if they were in a real conversation (e.g. smile if they feel like smiling) and videotape the participants' nonverbal responses. Second, we take advantage of the efficiency of this framework to increase the number of participants. Our goal is to analyze how participant's personality affects their nonverbal behaviors in the interaction. As mentioned before, there are a lot of possible sources for the variability of human behavior. By exploring the PCS framework, it is possible to examine how each of these sources can affect human behavior independently (e.g. by assigning different interactional goals to different participants, we can investigate how interactional goal affects nonverbal behaviors). We examine personality first and use it as an example to demonstrate how PCS can help us tease apart the causalities.

The following section describes the data collection and data annotation process. Section 3 discusses the results. We conclude our work in Section 4.

2 Data Collection and Annotation

2.1 Data Collection

In the study, we recruited 28 participants via www.craigslist.com from the general Los Angeles area. Before beginning the study, the participants were required to read the instructions and ask questions about anything they do not understand. They were informed beforehand that they would be videotaped and instructed to pretend to show interest and create a sense of rapport with the

Table 1. The attributes of each coder we measured before they started interacting with the speakers parasocially

Big Five Personality Traits	Extroversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Self-Consciousness	Self-directed, Other-directed
Parasocial Experience	Parasocial experience scale [18]
Other	Shyness, Self-monitoring, Gender

speaker in the video[1] by showing backchannel feedback such as head nod, head shake, and smile and so on. They first finished a 90-item personality inventory to measure their personality traits. Table 1 lists several individual traits that we are currently investigating. Next, they watched 8 speaker videos in sequence in a random order. Their nonverbal responses to the speakers were videotaped. At the end of the study, the participants were debriefed and each was paid 35 USD. Figure 1 shows an example of the parasocial interaction.

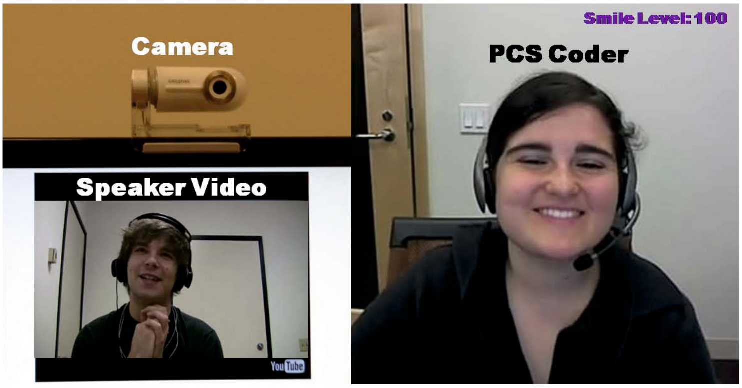


Fig. 1. An example of the parasocial interaction. The participant (right side) interacted with the speaker video (left side) parasocially, and her nonverbal behaviors were recorded by a camera. In this example, the speaker paused and tried to remember the details of the story he was supposed to tell. He had an embarrassed smile because it took him a relatively long time, and the participant smiled back, probably to reassure him. Although the participant was aware that the interaction was not real, she displayed such facial expressions seemingly automatically. We use the OKAO vision system from Omron Inc [14] to detect smiles, which can infer the level of smiling (continuous value from 0 to 100).

2.2 Results

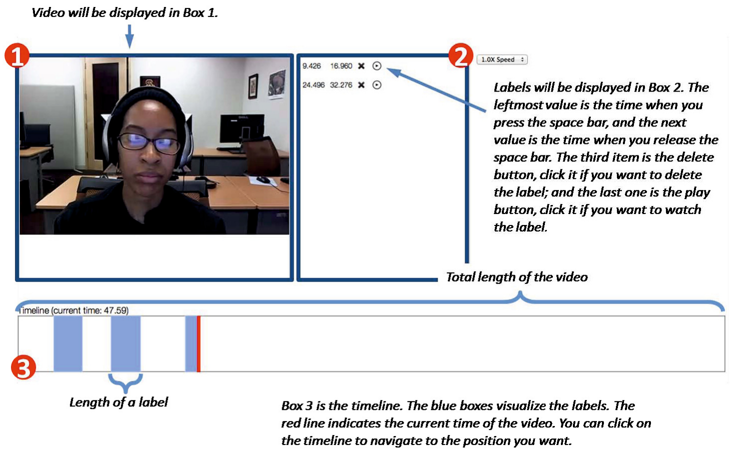
At the end of the study, we collected parasocial responses from all 28 participants to each of the 8 speaker videos. Participants produced wide variety of behaviors including both generic feedback (e.g. head nod) and specific feedback (e.g. headshake and expressive facial expressions) [11]. The specific feedback is always triggered by certain events mentioned in the conversation. In this study, we chose head nods, headshakes and smiles, which are the mostly occurred behaviors in our dataset, as the target behaviors. We will leave other common behaviors, such as frowns, to the future work.

[1] The video set used in this study was previously collected and used for studying how humans create rapport during face-to-face interactions. Each video records a human speaker retold a story to another human listener. The dataset is available at `rapport.ict.usc.edu`.

2.3 Annotation

We are interested in three kinds of nonverbal behaviors: head nods, headshakes and smiles. A mix of manual and automatic annotation techniques were used to annotate these behaviors from the recorded videos. To annotate head nods and headshakes, we recruited native annotators from Amazon Mechanical Turk. To facilitate the annotation work, we developed a web-based annotation tool (as shown in Figure 2) that helps annotators go through the videos and annotate behaviors efficiently. Each annotator examined seven videos in sequence and only annotated a single type of behavior at one time. Each video was annotated by two independent annotators and each was paid 3 USD.

Smiles were annotated automatically using the OKAO vision system [14]. Briefly, it uses computer vision techniques to identify 16 facial landmarks. From this, it derives a variety of facial pose estimates including a smile intensity ranging from 0 (no smile) to 100 (full smile). By setting the threshold to 50, we can reliably determine whether the participant is smiling or not.



Video will be displayed in Box 1.

Labels will be displayed in Box 2. The leftmost value is the time when you press the space bar, and the next value is the time when you release the space bar. The third item is the delete button, click it if you want to delete the label; and the last one is the play button, click it if you want to watch the label.

Box 3 is the timeline. The blue boxes visualize the labels. The red line indicates the current time of the video. You can click on the timeline to navigate to the position you want.

Fig. 2. This is the annotation interface. Coders press the space bar to start loading a video, and the loading progress will be shown in Component 1. After the video is loaded, coders press the space bar to start playing the video. At the beginning of the target behavior, coders press the space bar and hold it, and release the space bar when the target behavior ends. After finish labeling the video, coders can adjust the labels by dragging on their boundaries.

3 Data Analysis

For each of the 8 speaker videos, we aggregate the nonverbal behaviors (head nods, headshakes, and smiles) from all 28 participants to build the consensus view. Figure 3 shows an example of the consensus of head nod, headshake and smile. The peaks found in both the consensus of head nod and headshake are potential backchannel opportunities. But headshakes occur a lot less than head nods do, and they are usually associated with semantically negative events in the speech. There is a noticeable jump in the consensus of smile, where the speaker says the most dramatic part of the story. Interestingly, this phenomenon is observed in all 8 videos.

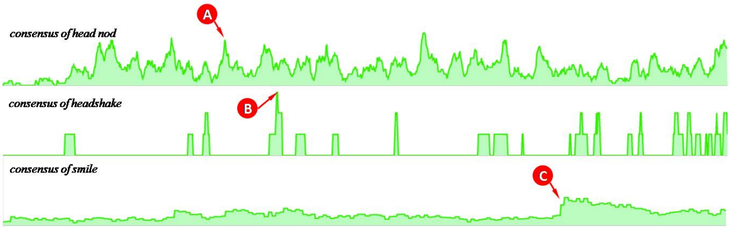


Fig. 3. An example illustrates the consensus of head nod (top), headshake (middle) and smile (bottom). The speaker video is from a sexual harassment training course. At point A, the speaker said “It’s from Rick in accounting or Rick in legal or something, and [pause], he said ‘oh, no’ ...” and the nod is most likely to occur during the pause; at point B, the speaker said “and she says, ’you know, I gave him a ride once when his car broke down, now he won’t leave me alone, it’s been five weeks, I always get these emails, and e-cards, and he won’t leave me alone’...” and the shake is most likely to occur when the speaker described the fact that Rick kept bothering the lady; at point C, the speaker said “and then she says ‘oh, and next, I am gonna need a foot massage’, and then she shuts the blinds...” and the smile is most likely to occur after mentioning the foot massage.

3.1 How Listener’s Personality Traits Influence the Behavior

To examine the impact of personality on nonverbal behavior, we calculated personality scores from the 90-item personality inventory as described in Table 1. For each personality subscale (e.g. extroversion), we performed a median split and grouped participants into a high and low scoring group. For example, those scoring below the median for the scale of extroversion would be combined into an introverted group whereas those scoring above the median would be combined into an extroverted group. We then contrasted the consensus of the behaviors of these two partitions with Bonferroni Correction [16]. The results are shown as below.

The rows where significant differences are found are highlighted. Table 2 shows that listeners personality traits have significant influence on the number of head nods. The result is in line with previous research. For example, Chartrand and Bargh [19] found empathic individuals exhibit more mimicry behavior during the interaction to a greater extent than not empathic individuals; accordingly, our data suggests that extroversion, openness and consciousness all have similar influence on the number of head nods.

Besides head nods, we also examined headshake and smile. As Table 3 shows, headshake rarely happens during the interaction. In our data, we find that its

Table 2. Compare the average number of head nods between the *low_group* and *high_group* with respect to each attribute

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

occurrence is always associated with the semantically significant events (usually negative events) in the speech. Listeners personality traits dont have significant influence on headshake. However, smiling, although a kind of specific feedback, is significantly affected by the listeners personality traits (as shown in Table 4). This is similar to the results we found in literature. For example, Shiota et al. [17] showed that more extraverted, more conscientious, more agreeable, and less neurotic people are more likely to experience joy.

Table 3. Compare the average number of headshakes between the *low_group* and *high_group* with respect to each attribute

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

3.2 How Speaker's Personality Traits Influence the Behavior

Each speaker video was watched by all participants, and we call the aggregation of their behaviors the "crowds' behavior". The crowds' behavior is different when interacting with different speaker videos. A natural follow-up question is whether or not the speaker's personality traits can influence the crowds' behavior. In a previous study [7], we measured each speaker's personalities. We compute the correlation coefficients between the speaker's personality measurements and the crowds' behavior (i.e. the number of head nods, the number of headshakes, and the amount of smiles). The results suggest that speaker's personality does not affect the listeners' head nods or smiles. Listeners' smiles are highly correlated with

Table 4. Compare the amount of smiles between the *low_group* and *high_group* with respect to each attribute. The number is calculated by dividing the duration of the listener's smiling by the duration of the whole interaction.

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

Table 5. The correlation coefficients between speaker personality traits and the number of headshakes of crowds

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

the speakers' smiles (correlated coefficient = 0.80), indicating that the listener was mimicking the speaker's smile. However, some of the speaker's personality measurements are highly correlated with the listeners' headshakes (as shown in Table 5).

The result shows that the number of headshakes is positively correlated with the speakers extroversion and agreeableness measurements, and is negatively correlated with the neuroticism and shyness measurements. In our task, headshake always indicates negative emotions towards what the speaker said. That is, if the speaker is more extroverted and agreeable, the listeners are more likely to express their negative emotions; however, if the speaker is more neurotic and shyer, the listeners tend to hide their negative emotions.

3.3 Predicting Personality from Parasocial Responses

We investigate how well we can predict personality just from the listener backchannel feedback and how well we can explain the variability of listener backchannel feedback by only using the listeners' personality. We ran a stepwise linear regression analysis between backchannel feedback (including the number of nods, the number of shakes and the duration of smiles) and personality measurements.

First, we predict personality traits from the parasocial consensus. The dependent variable is each of the personality traits (e.g. extroversion), and the independent variables are the number of head nods, the number of headshakes, and the duration of smiles produced by PCS coders. We observed significant results for neuroticism and self-consciousness. Smile itself (correlation coefficient = -0.23) can predict about 12% of the variance of neuroticism (F=3.4, p=0.07); smile (correlation coefficient = 0.2) and nod (correlation coefficient = -0.17) together can predict about 20% of the variance of self-consciousness (F=3.11, p=0.062). Second, we run the same analysis reversely; that is, the dependent variable is the number of head nods, the number of headshakes and the duration of smiles respectively, and the independent variables are the personality traits. We only observed significant result for smile. Self-consciousness (correlation coefficient = 0.868) and neuroticism (correlation coefficient = -0.658) can predict about 28% of the variance of smile (F=4.95, p=0.015). Together, this suggests we can intuit something about a speakers personality simply by looking at the responses of their conversation partner, although this relationship is rather modest.

4 Conclusion and Future Work

In this paper, we introduced our initial efforts in analyzing and modeling the variability of human nonverbal behaviors in face-to-face interaction. We extended the Parasocial Consensus Sampling (PCS) framework to make such analysis possible. The results showed that personality has significant influences on backchannel feedback and clearly demonstrated how it affects backchannel feedback. In the future, we will integrate the results into nonverbal behavior models, and test whether the virtual human driven by such models can exhibit the corresponding personalities or not. Moreover, we will further explore the PCS framework to investigate how other factors, such as interactional goal, roles in the conversation and emotions, can influence nonverbal behaviors in face-to-face interaction.

References

1. Swartout, W.R., Gratch, J., Hill Jr., R.W., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward virtual humans. AI Magazine 27(2), 96–108 (2006)
2. Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., Sereika, S.M.: Maintaining reality: Relational agents for antipsychotic medication adherence. Interacting with Computers 22(4), 276–288 (2010)
3. Rowe, J.P., Shores, L.R., Mott, B.W., Lester, J.C.: Integrating learning and engagement in narrative-centered learning environments. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 166–177. Springer, Heidelberg (2010)
4. Pelachaud, C.: Simulation of face-to-face interaction. In: Proceedings of the Workshop on Advanced Visual Interfaces, pp. 269–271 (1996)
5. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, pp. 413–420 (1994)
6. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)

7. 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)
8. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, pp. 289–296 (2009)
9. 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)
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. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology, 941–952 (2000)
12. Bevacqua, E., Mancini, M., Pelachaud, C.: A listening agent exhibiting variable behaviour. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 262–269. Springer, Heidelberg (2008)
13. Huang, L., Morency, L.-P., Gratch, J.: Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, pp. 1265–1272 (2010)
14. Lao, S., Kawade, M.: Vision-based face understanding technologies and their applications. In: Li, S.Z., Lai, J.-H., Tan, T., Feng, G.-C., Wang, Y. (eds.) Sinobiometrics 2004. LNCS, vol. 3338, pp. 339–348. Springer, Heidelberg (2004)
15. Wang, Z., Lee, J., Marsella, S.: Towards more comprehensive listening behavior: Beyond the bobble head. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 216–227. Springer, Heidelberg (2011)
16. Bonferroni, C.E.: Il calcolo delle assicurazioni su gruppi di teste. In: Tipografia del Senato
17. Shiota, M.N., Keltner, D., John, O.P.: Positive emotion dispositions differentially associated with Big Five personality and attachment style. The Journal of Positive Psychology, 61–71 (2006)
18. Hartmann, T., Goldhoorn, C.: Horton and wohl revisited: Exploring viewerss experience of parasocial interactions. In: The Annual Meeting of the International Communication Association
19. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 893–910 (1999)