Manipulating the Perception of Virtual Audiences using Crowdsourced Behaviors

Mathieu Chollet[1], Nithin Chandrashekhar[1], Ari Shapiro[1], Louis-Philippe Morency[2], and Stefan Scherer[1]

[1]Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Playa Vista, CA, USA
[2]Language Technology Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA
\{mchollet,nithinch,shapiro,scherer\}@ict.usc.edu,morency@cs.cmu.edu

Abstract. Virtual audiences are used for training public speaking and mitigating anxiety related to it. However, research has been scarce on studying how virtual audiences are perceived and which non-verbal behaviors should be used to make such an audience appear in particular states, such as boredom or engagement. Recently, crowdsourcing methods have been proposed for collecting data for building virtual agents' behavior models. In this paper, we use crowdsourcing for creating and evaluating a nonverbal behaviors generation model for virtual audiences. We show that our model successfully expresses relevant audience states (*i.e.* low to high arousal, negative to positive valence), and that the overall impression exhibited by the virtual audience can be controlled my manipulating the amount of individual audience members that display a congruent state.

Keywords: Virtual Audience, Crowdsourcing, Non-verbal Behaviors

1 Introduction

Modern professional and personal life often involve situations where we are required to speak in public, such as when performing a professional presentation or when making a toast at a wedding. The ability to speak in public proficiently can greatly influence a person's career development, help build relationships, resolve conflict, or even gain the upper hand in negotiations. While there is no such thing as a best style of public speaking, every efficient public speech requires the mobilization of varied skills, ranging from the selection and arrangement of appropriate and convincing arguments to the efficient vocal and non-verbal delivery of the speech. This very desirable set of skills is not innate to most of us, and many people actually dread the prospect of speaking in public: it is actually one of the most commonly reported phobias [2]. Fortunately, public speaking ability can be improved through training and public speaking anxiety can be reduced through a number of methods, including exposure to virtual audiences [6, 14]. Virtual audiences are collections of virtual agents situated in 3D environments designed to reproduce a public speaking situation [3, 9].

Multimodal interactive systems for social skills training have recently been proposed in domains such as job interview training [4] or public speaking training [5,15]. While virtual audiences have been used since fifteen years for the mitigation of public speaking anxiety [6,14], they have only recently been proposed for training public speaking skills [1,3]. Training systems using such audiences could hold several advantages over traditional public speaking training methods, such as training workshops and rehearsals with colleagues or friends [7]: they are always available, whereas audiences of friends or public speaking experts are not; whilst some people could be reluctant to training their public speaking ability with real people out of fear of being judged, virtual audiences do not pose such a threat [10]; in addition, they can be finely controlled, allowing training to be standardized. Finally, virtual humans are excellent in captivating individuals' attention, in creating rapport and engaging the learner [16], which are essential prerequisites for successful learning outcomes.

Designing virtual audiences for public speaking training requires understanding how they are perceived and how this perception can be manipulated through their behavior, layout or appearance. However, this has only started to be investigated recently [8,9], and many aspects of this question remain unanswered. In particular, it is still unclear how combinations of behaviors from different modalities (*e.g.* postures, head movements, facial expressions, gaze) are perceived when expressed by a virtual audience character. Additionally, to the best of our knowledge, the overall perception of audiences containing characters displaying disparate states has not been studied yet. In this paper, we set out to study these research questions by using crowdsourcing methods.

In the next section, we begin by reviewing related works on social skills training using multimodal interfaces and on the design and usage of virtual audiences. We then present in Section 3 a study on the relationship between virtual characters behaviors and perceivable audience states. We then realized another experiment, outlined in Section 4, in order to validate that the overall perception of a virtual audience can be finely controlled by adjusting the amount of characters that display a target state.

2 Related Work

Recently, public speaking training with multimodal interfaces providing direct feedback mechanisms has been a popular topic. The Rhema system uses Google Glass to provide the speaker with feedback on speaking volume and speaking rate [15]. Logue [5] is a similar system that provides realtime feedback to presenters on their speech rate body openness and body energy using functional icons displayed on Google Glass. Barmaki and Hugues presented a system for training teachers to adopt better body postures, using a virtual classroom populated with manually controlled virtual students [1]. A particular paradigm for interfaces for public speaking training is the virtual audience. Such a system aims at reproducing a public speaking situation with high fidelity, using an environment that is typical of public speaking situations (*e.g.* a conference room) and populating it with virtual characters acting as the user's audience. Virtual audiences have first been investigated to treat public speaking anxiety. North *et al.* found in a series of studies that virtual audiences were effective in inducing stress and reducing public speaking anxiety [6, 12]. Researchers also investigated the effect of three different types of virtual audiences, namely a neutral, a positive, and a negative audience, consisting of eight virtual characters [14]. Virtual audiences have only been recently used for specifically improving public speaking ability, and not solely reducing their anxiety. In previous work, we introduced the Cicero public speaking training framework [3], which uses an interactive virtual audience to deliver natural feedback using the virtual characters' non-verbal behavior. In a preliminary study, the audience used head nods (resp. shakes) and forward (resp. backward) leaning postures for positive (resp. negative) feedback; however we did not systematically study the effect of these behaviors.

The perception of virtual audiences' behaviors has only recently been systematically studied by Kang *et al.* [8,9]. In [9], two real audiences were recorded while listening to presentations designed to elicit certain states in the audience, *e.g.* a speech advocating for pay cuts was used in order to elicit a negative reaction from the audience. Participants' behaviors were coded every 2 seconds and the resulting dataset was used to build models for choosing full body postures (*i.e.* head, arms, hands, torso and feet) according to given input state variables (attitude, personality, mood, energy). Facial expressions and gaze were not considered, and head nods and shakes were rarely found in the annotations. An evaluation by users describing the audience displayed in a virtual reality headset seems to indicate that the model adequately portrays attentiveness or boredom, but not valence. In [8], further studies were realized using this framework in order to evaluate which behavior types and states are recognized by participants observing a virtual audience. In particular, they found that study participants could differentiate between audiences with different levels of valence and arousal (or interest), and described five main recognizable audience types. While their work shows many insights about the design of virtual audiences, one main limitation of their model is that it considers only full body postures, and not other crucial listening behaviors such as head nods and head shakes, and it is difficult to grasp from their results the particular influence of a certain modality compared to the others. They also did not investigate the perception of audiences consisting of virtual agents with mixed states (*e.g.* an audience with 2 engaged characters and 3 bored characters).

In this paper, we build on Kang *et al.*'s work, adopting valence and arousal as two underlying dimensions that drive our virtual audience behaviors and evaluate the influence of homogeneity of virtual audience members' behaviors on the perception of the overall audience (*i.e.* what happens when 2 characters are bored, 3, 4, *etc.*). In a first study, we investigate the role of non-verbal modalities for expressing these states. Then, we explore in a second study participants' overall impressions of virtual audiences when manipulating the amount of virtual characters that express a target state. In a nutshell, we try to answer the following questions:

Q_1: Which non-verbal signals make a virtual audience character appear critical or supportive of the speech? Bored or engaged?

Q_2: Can we control users' perception of a virtual audience's level of arousal and valence by manipulating individual audience members' behaviors?

In the next section, we set out to answer Q_1 by using a crowdsourcing method to collect audience members' behaviors.

3 Crowdsourcing Audience Behaviors

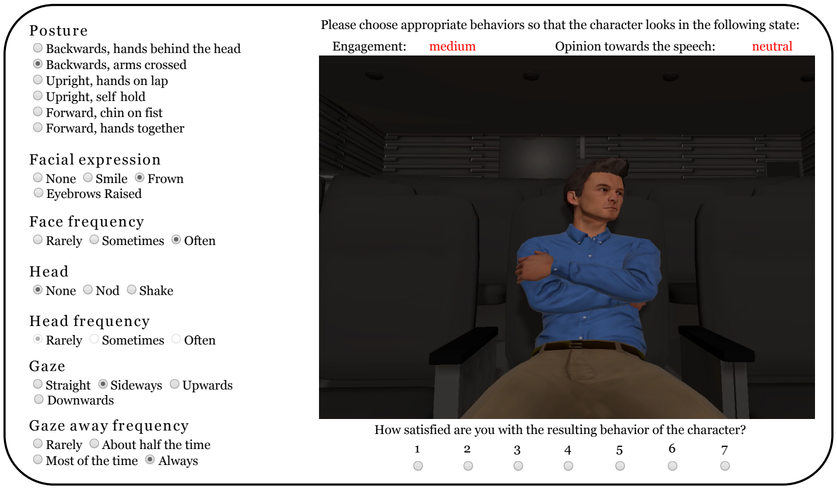
Our first research goal was to identify the link between different non-verbal signals of various modalities and the relevant audience state dimensions we defined in the previous section, *i.e.* valence and arousal. To achieve this goal, we used a methodology introduced by Ochs *et al.* that consists of asking users to select combinations of behaviors (and/or parameters of these behaviors, *e.g.* duration or intensity) that adequately portray a studied socio-emotional phenomenon [13]. For instance, Ochs *et al.* used this method to collect a repertoire of amused, polite and embarassed smiles by letting users create their own virtual smiles, choosing what they thought to be adequate intensities, durations and combination of facial movements involved for the chosen smile category.

3.1 Crowdsourcing Interface

We adopted this methodology for our first research question Q_1, producing a web interface shown in Figure 1. This interface consists of a task description, a panel containing a number of possible behavior choices, a video panel displaying a virtual character (male or female) enacting the chosen behaviors, and a 7-point scale to indicate how well the participant thinks the resulting video conveys the input condition. The different parameters that could be chosen by the users were the following:

- Amount of time with an averted gaze: 0%, 25%, 50%, 75%, 100%
- Direction of the averted gaze, if any: Sideways, Down, Up.
- Posture: 6 different choices (5 of them visible in Figure 3).
- Facial expression, if any: smile, frown, eyebrows raised.
- Facial expressions frequency (if applicable): 25%, 50%, 75% of the time.
- Head movements, if any: nod, shake.
- Head movements frequency (if applicable): 1/2/3 times every 10 seconds.

These parameters allow us to cover most modalities of human communication. We only discarded gestures and vocal behavior which we do not consider in our framework, meaning the audience only produces listening behavior. We also chose a variety postures, allowing us to explore different underlying dimensions of postural behavior, in particular proximity to the speaker (lean backward/forward) and openness (hands behind head *vs* arms crossed) [11]. Some heuristics were introduced in order to make sure that no clashes between behaviors would happen in the videos (*e.g.* no head shake while the gaze direction is



Posture

○ Backwards, hands behind the head
● Backwards, arms crossed
○ Upright, hands on lap
○ Upright, self hold
○ Forward, chin on fist
○ Forward, hands together

Facial expression

○ None ○ Smile ● Frown
○ Eyebrows Raised

Face frequency

○ Rarely ○ Sometimes ● Often

Head

● None ○ Nod ○ Shake

Head frequency

● Rarely ○ Sometimes ○ Often

Gaze

○ Straight ● Sideways ○ Upwards
○ Downwards

Gaze away frequency

○ Rarely ○ About half the time
○ Most of the time ● Always

Please choose appropriate behaviors so that the character looks in the following state:

Engagement: medium Opinion towards the speech: neutral

How satisfied are you with the resulting behavior of the character?

1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

Fig. 1: The crowdsourcing interface.

changing) and to introduce some variability in behavior timings. We defined five values for both the arousal (resp. valence) states: “very low”, “low”, ”medium”, “high”, “very high” (resp. negative/positive). We created 10s videos corresponding to all of the possible different combinations of the behavior parameters for a male and a female character, resulting in 10920 videos.

3.2 Experimental Results

We recruited 72 participants using the Amazon Mechanical Turk website[1] to create combinations of behaviors for the states we considered. Using our web interface, we collected 1045 combination of behaviors, an average of 20.9 combinations of behaviors per input state.

In order to explore the data and answer our research question Q_1, we tested the following hypotheses about how behaviors were chosen by participants.

H1 Arousal and expressions: More frequent facial expressions, head movements and less frequent gaze aversions lead to higher arousal.
H2 Valence and expressions: Smiles and nods lead to positive valence, frowns and head shakes to negative valence, while eyebrow raises are mostly neutral.
H3 Arousal and postures: Postures chosen for high arousal involve leaning closer to the speaker than postures chosen for lower arousal.
H4 Valence and postures: Open postures lead to higher valence compared to more closed postures.

The distributions of behaviors per valence and arousal states regarding these hypotheses are displayed in Figure 2. We conducted statistical tests to ensure

[1] `https://www.mturk.com`

that these behavior distributions were statistically significant. Prior to conducting these tests, we transformed our arousal and valence data into numerical values (very low → 1 to very high → 5), and we created numerical variables for proximity (backward → 1 to forward → 3) and openness (arms crossed and self-hold → 1, arms behind the head → 3, the rest → 2).

For **H1**, **H3** and **H4**, the data being of ordinal nature, we realized Kruskal-Wallis tests. For H1, we set the arousal as the independent variable (IV) and conduct tests with the face, head and gaze frequencies as dependent variables (DV). The three tests are significant, for facial expressions ($H(3) = 49.88$, $p < 0.001$), head movements ($H(3) = 101.09$, $p < 0.001$) and gaze ($H(4) = 347.32$, $p < 0.001$). For H3, we set arousal as the IV and proximity as the ordinal DV. The results confirm our hypothesis: higher arousal leads to higher postural proximity ($H(3) = 334.82$, $p < 0.001$). Similarly for H4, we set valence as the IV and openness as the DV and confirm our hypothesis ($H(3) = 73.59$, $p < 0.001$). For **H2**, the data being of categorical nature and not ordinal, we performed a Chi-squared test, which also showed statistical significance ($\chi^2(12) = 1559.8, p < 0.001$). These results confirm the four hypotheses we presented earlier. We found that higher arousal leads to more frequent expressions and to postures that are closer to the speaker, while valence affects the type of expressions used (*e.g.* smiles and nods for positive valence, frowns and shakes for negative valence) and leads to less open postures.

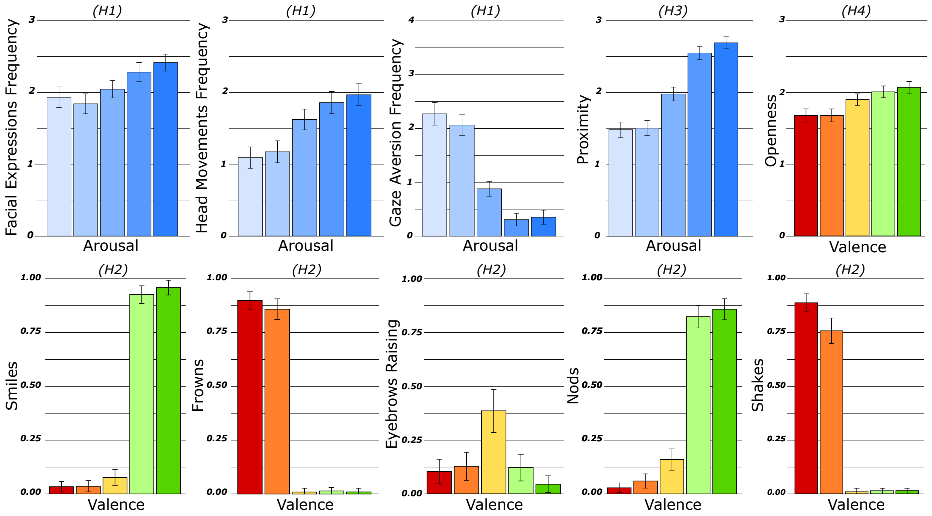


Fig. 2: Distribution of behaviors per state levels for the investigated hypotheses. From left to right in the subfigures, very low to very high arousal (resp. valence).

Using the data we collected, we then built a probabilistic model that reflect how often certain behaviors are chosen in a given state. In effect, it models the $P(Behavior|Arousal, Valence)$ behavior distributions for the different modalities and states used in the crowdsourcing study. At runtime, we can select appropriate behaviors for expressing one character's state by querying our model. For each modality, a random number (in the $[0, 1)$ range) is generated and compared to the cumulative distribution function (CDF) of that modality's behavior distribution. The behavior corresponding to that level of the CDF is then returned to be displayed by the character. This model allows us to select appropriate behaviors so that each virtual character reflects its current state, while still exhibiting variability in behaviors. Our second research question Q_2 was to validate that the perception of a virtual audience' overall state can be incrementally manipulated by adjusting the amount of characters that display the chosen state. To that end, we realized a second experiment, presented in the next section.

4 Overall Perception of Virtual Audiences

In order to investigate the perception of complete audiences, we realized a second study. The goal was to verify that by manipulating the expressed state of one virtual character at a time, the overall perceived state of the audience can be changed continuously.

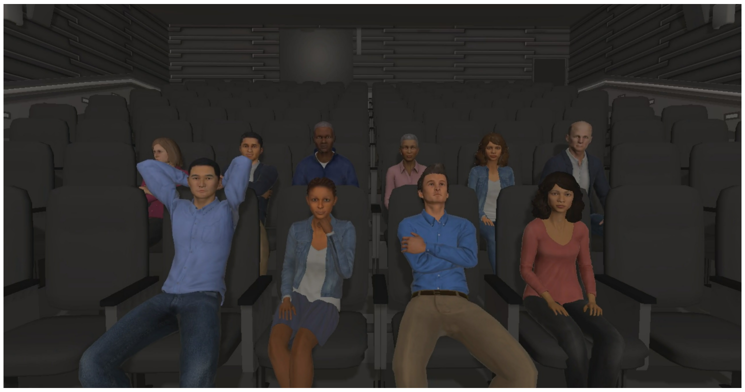


Fig. 3: Screenshot of the full audience.

For this study, we defined two independent variables: the target state **S**, consisting of a value of valence and arousal, and the number of manipulated characters **N**. We used a fixed audience configuration, displayed in Figure 3. In order to reduce the amount of tested conditions, we considered only three levels of valence and arousal, *i.e.* low, medium or high arousal and negative, neutral and positive valence, randomly selecting between a very low/low and very high/high level for generating a character's state when creating a video. The audience consisted of 10 characters and thus **N** could take 11 values, from 0 to 10. The (**N**) manipulated characters would be assigned behaviors according to their state using the probabilistic model built after the previous experiment.

For the other (10-**N**) non-manipulated characters, a random state was selected, meaning that they could display congruent, neutral or contradictory behaviors compared to the input condition. We created 4 video variants for every condition, meaning we evaluated a total of 396 videos.

We created another web interface for this study. The participants' task was to watch the video and to indicate their overall perception of the audience's level of arousal and valence, using 5-point scales. The participants were also recruited from Amazon Mechanical Turk. We collected 2643 answers for both dimensions from 105 participants, for an average of 7,1 answers per video, or 26,7 answers per input condition. For compiling the results, we used a majority voting to determine the perceived state of a particular video, *e.g.* if the audience of one video was rated with an arousal level of 5 by 4 participants and with an arousal level of 3 by 2 participants, then the video receives a score of 5. When a tie occurs, the scores are averaged for the video. The results, averaged over all input videos, are presented in Figure 4.

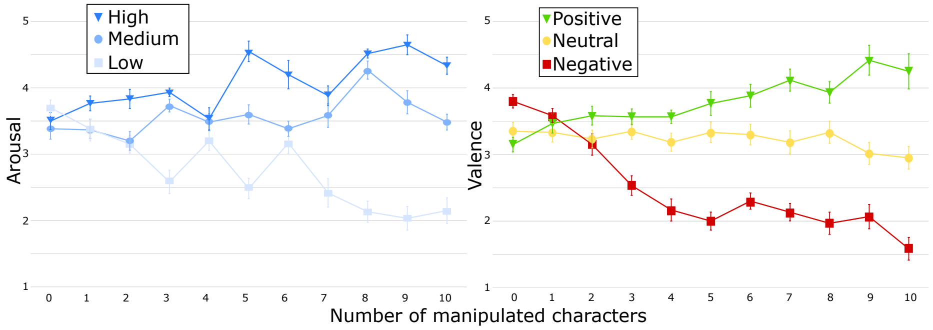


Fig. 4: Perception of arousal (resp. valence) for audiences of 10 characters, depending on the target state and the number of manipulated characters.

We can observe from Figure 4 that the perceived state of the audience gets more clearly recognized as the number of manipulated characters expressing the input condition increases. Our model can successfully express low, medium and high arousal as well as negative, neutral and positive valence. We conducted a linear regression analysis in order to further analyze the impact of manipulating individual characters. Specifically, we studied a regression model of the following form $y = \alpha + \beta_S * \mathbf{N}$, with y corresponding to the participants' rating of the arousal (resp. valence) for the video's audience, S being the input arousal condition (resp. valence) of the video, *i.e.* high, medium or low, and β_S the corresponding regression coefficient. Finally, **N** corresponds to the amount of manipulated characters. The results of our linear regression analysis are the following:

- $y_{Arousal} = 3.485 + 0.109 * N_{High} + 0.025 * N_{Medium} - 0.146 * N_{Low}$ $(F(3, 392) = 53.21, p < 0.001.\ R^2 = 0.29,\ StdErr = 0.99)$

- $y_{Valence} = 3.362 + 0.092 * N_{Positive} - 0.027 * N_{Neutral} - 0.181 * N_{Negative}$ $(F(3, 392) = 95.49, p < 0.001.$ $R^2 = 0.42$, $StdErr = 0.79)$

For medium arousal (resp. neutral valence), we find that the slope is not statistically significantly different from a flat line ($p > 0.05$ in both cases). We find that the slope coefficients for high and low arousal (resp. positive and negative valence) are significant ($p < 0.001$ in all 4 cases), *i.e.* the slope in these cases is significantly different from a flat line. This validates our second research question Q_2, meaning that it is possible to incrementally alter the arousal and valence manifested by the virtual audience by changing the state of one virtual character at a time. Another interesting result is that the slope for negative valence seems to be twice as strong as for positive valence. This suggests that users might perceive negative behaviors as more salient than positive behaviors.

5 Conclusion and Future Work

In this paper, we investigated virtual audience behaviors with the goal of understanding which non-verbal behaviors are relevant and recognizable for producing feedback for public speaking training. We used a crowdsourcing method to gather a user-created corpus of virtual characters' behaviors corresponding to audience states, consisting of valence and arousal dimensions, validated in previous work [8, 9]. We found that higher arousal leads to more frequent expressions and to postures that are closer to the speaker, while valence affects the type of expressions used and leads to less open postures. We then investigated whether the overall perception of audiences can be controlled by manipulating the number of characters displaying a target state. We observed that our virtual audience model successfully conveys both low, medium and high arousal levels as well as negative, neutral and positive valences. We also found that the perceived level of arousal (resp. valence) of our audience is proportional to the amount of characters that display it. This means that we can continuously vary the impression given by the virtual audience, by changing the expressed state of one virtual character at a time towards the target feedback state.

In future work, we will investigate the link between the placement of individual characters and their influence on the overall perception of the audience: indeed, it could be that the front row characters are more salient to the users than the back row characters. Understanding this effect, if it exists, could allow us to control the virtual audience impression even more precisely. Additionally, we will study the perception of our virtual audience during actual public speaking training sessions with participants.

Acknowledgments This material is based upon work supported by the National Science Foundation under grant No. IIS-1421330 and U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government, and no official endorsement should be inferred.

References

1. R. Barmaki and C. E. Hughes. Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 531–537, New York, NY, USA, 2015. ACM.
2. G. D. Bodie. A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education*, 59(1):70–105, 2010.
3. M. Chollet, T. Wortwein, L.-P. Morency, A. Shapiro, and S. Scherer. Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In *Proceedings of UbiComp 2015*, Osaka, Japan, 2015. ACM.
4. I. Damian, T. Baur, B. Lugrin, P. Gebhard, G. Mehlmann, and E. André. Games are better than books: in-situ comparison of an interactive job interview game with conventional training. In *International Conference on Artificial Intelligence in Education*, pages 84–94. Springer, 2015.
5. I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 565–574, New York, NY, USA, 2015. ACM.
6. S. R. Harris, R. L. Kemmerling, and M. M. North. Brief virtual reality therapy for public speaking anxiety. *Cyberpsychology and Behavior*, 5:543–550, 2002.
7. J. Hart, J. Gratch, and S. Marsella. *How Virtual Reality Training Can Win Friends and Influence People*, chapter 21, pages 235–249. Human Factors in Defence. Ashgate, 2013.
8. N. Kang, W.-P. Brinkman, M. B. van Riemsdijk, and M. Neerincx. The design of virtual audiences: Noticeable and recognizable behavioral styles. *Computers in Human Behavior*, 55:680–694, 2016.
9. N. Kang, W.-P. Brinkman, M. B. van Riemsdijk, and M. A. Neerincx. An expressive virtual audience with flexible behavioral styles. *Affective Computing, IEEE Transactions on*, 4(4):326–340, 2013.
10. G. Lucas, J. Gratch, A. King, and L.-P. Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.
11. A. Mehrabian. *Nonverbal communication.* Transaction Publishers, 1977.
12. M. M. North, S. M. North, and J. R. Coble. Virtual reality therapy: An effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, 3:2–6, 1998.
13. M. Ochs, B. Ravenet, and C. Pelachaud. A crowdsourcing toolbox for a user-perception based design of social virtual actors. In *Computers are Social Actors Workshop (CASA)*, 2013.
14. D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments*, 11(1):68–78, Feb. 2002.
15. M. Tanveer, E. Lin, and M. E. Hoque. Rhema: A real-time in-situ intelligent interface to help people with public speaking,. In *Proceedings of the 20th ACM Conference on Intelligent User Interfaces*, pages 286–295, 2015.
16. N. Wang and J. Gratch. Don't Just Stare at Me! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1241–1250, Chicago, IL, 2010.