An Expressive Virtual Audience with Flexible Behavioral Styles

Ni Kang, Willem-Paul Brinkman, M. Birna van Riemsdijk, and Mark A. Neerincx

Abstract— Currently, expressive virtual humans are used in psychological research, training, and psychotherapy. However, the behavior of these virtual humans is usually scripted and therefore cannot be modified freely at run time. To address this, we created a virtual audience with parameterized behavioral styles. This paper presents a parameterized audience model based on probabilistic models abstracted from the observation of real human audiences (*n* = 16). The audience's behavioral style is controlled by model parameters that define virtual humans' moods, attitudes, and personalities. Employing these parameters as predictors, the audience model significantly predicts audience behavior. To investigate if people can recognize the designed behavioral styles generated by this model, 12 audience styles were evaluated by two groups of participants. One group (*n* = 22) was asked to describe the virtual audience freely, and the other group (*n* = 22) was asked to rate the audiences on eight dimensions. The results indicated that people could recognize different audience attitudes and even perceive the different degrees of certain audience attitudes. In conclusion, the audience model can generate expressive behavior to show different attitudes by modulating model parameters.

Index Terms—Expressive listening behavior, parameterized audience model, public speaking, virtual agents

1 Introduction

Like a human audience, an audience of virtual humans has the ability to elicit responses in humans, e.g., [1], [2]. This ability makes a virtual audience beneficial when it comes to training, psychotherapy, or psychological stress testing. For example, it can help musicians to practice performing in front of an audience [3]. Virtual audiences are also being used as part of exposure therapy for individuals with social anxiety disorder [4] by exposing them to situations they fear. Instead of learning to cope with anxiety, some studies (e.g.,[5]) suggest that virtual audiences may also be used in the Trier Social Stress Test (TSST) [6] to induce stress in an individual with the aim of studying the effect of stress.

Besides the logistic advantage of not needing to arrange audience members and a suitable location, a virtual audience also offers the ability of control over the audience. For example, although the procedure for the standard TSST aims for a neutral audience, some have also explored variations with supportive or non-supportive audiences [7]. For exposure therapy, control of the fear stimuli is also desirable, as therapists aim to gradually expose patients to more fear-eliciting situations. Besides switching between different situations, e.g., an audience of fewer or more people [8], Emmelkamp [9] also suggests that treatment of social anxiety can also benefit from control over the fear stimuli within a virtual reality session, e.g., the behavior of the audience, as patients need to experience a certain amount of anxiety. Some treatment manuals [8] even give specific instructions on the desired anxiety level. Therefore, these manuals [8], [10] suggest using the attitude of an audience (e.g., negative or positive audience) as an effective means of controlling anxiety in a public speaking scenario. Currently virtual audiences are often represented by 3D models animated by a predefined script, e.g., [2], or videos of actual people embedded in Virtual Environments (VE)s, e.g., [11]. To control the audience's behavior, different animations or videos should be prepared so that operators can switch between these clips. However, the preparations may require considerable effort and thus are usually made in advance because explicit behaviors need to be scripted along the timeline for each audience member. Due to the effort involved, these pre-scripted animations and videos are often relatively short, causing the virtual audience to behave in repeating loops. This repetition may reduce the behavioral realism, thereby lowering the desired effect, e.g., lowering treatment efficacy [12]. From an engineering perspective, a more flexible and efficient system can be developed by applying software agents for the virtual audience to generate expressive behaviors automatically. Instead of specifying individual audience behavior, operators can adjust the agent parameters, e.g., attitude or personality, at run time to change the audience behavioral style. Controlling the audience on this higher level of abstraction reduces workload, as low-level audience behaviors no longer need to be controlled manually.

We therefore propose to use a statistical model, i.e., a model based on a corpus of audience behavior instead of theories of audience behaviors, to generate expressive behavior of virtual audience members. This method allows a human operator (e.g., researcher, therapist, or trainer) to control the virtual audience's behavioral styles by setting the agents' attributes (e.g., attitude and mood) and environment settings (e.g., interrupting events). This paper describes the creation of such a virtual audience, set within a public speaking scenario, as this is a commonly used laboratory procedure to elicit stress, e.g., as part of TSST, and as this is also one of the most common social situations that people with

- *N. Kang, W.-P. Brinkman, M.B. van Riemsdijk, and M.A. Neerincx are with Department of Intelligent Systems, Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands. E-mail: conniekdl@gmail.com; W.P.Brinkman@tudelft.nl; M.B.vanRiemsdijk@tudelft.nl.*
- *M.A. Neerincx is also with TNO Human Factors, Kampweg 5, 3769 DE Soesterberg, the Netherlands. E-mail: Mark.Neerincx@tno.nl.*

social anxiety fear [13]. Since the audience in public speaking situations usually shows their attitudes through body expressions, the design focuses on the generation of bodily responses of the virtual audience. To create such an audience, the main contributions of this study are: (1) a parameterized audience model which generates expressive behaviors based on statistical models, and (2) a corpus of audience behavior in public speaking situations.

2 Design of the Virtual Audience

As already mentioned, the behavior of the virtual audience should be realistic, flexible and expressive to display different attitudes. Thus, this paper proposes a parameterized agent model for the audience to generate expressive listening behavior. The behavioral styles can be modulated by adjusting agent attributes such as mood, attitude, personality, and energy level. The models for behavior generation are realized through a statistical approach.

2.1 Realistic and Flexible Expressive Behavior

Behavioral realism can be achieved by using autonomous agents. Although few studies have reported on the behavior of an autonomous audience, the potential for natural behavior has already been shown in recent studies of Embodied Conversational Agents (ECA). For example, a speaking agent can generate natural head movements [14], and a listening agent with simulated backchannel (head nod and smile) can improve the rapport in the human-agent interaction [15].

Adjustable expressive behavior can be implemented by a parameterized agent model. The parameters should affect the virtual humans' behavior so that they can behave expressively. For example, a model of listeners' feedback behavior in a multiparty conversation [16] was able to take into account several factors which affect agents' behavior, such as their conversational roles and goals, understanding, and attitudes. Furthermore, Busso et al. [17] shows the possibility of computational models to predict a speaker's head motion for different emotions, and their evaluations suggested that these models successfully emphasized the emotional content and improved the virtual speaker's behavioral naturalness.

Using such a parameterized model, operators can adjust the virtual audience's behavioral style by modifying its parameters. To convey affective connotations via body language, the parameters were selected from attributes that can affect and can be expressed in a person's nonverbal behavior. These attributes include moods, attitudes, personality, and physiological states [18].

2.2 Behavior Generation

Behavior generation of autonomous agents is often implemented by two main approaches: the theoretical and the statistical approach. The theoretical approach is to craft the rules that specify which behavior should be generated in a certain context based on psychological knowledge and literature. Examples using this approach include the listener model by Bevacqua et al. [19]. The statistical approach has also been widely used. It uses statistical models taken from observations or corpora of human behavior to predict virtual agents' behavior. For example, a speaking agent [14] was developed using a machine learning approach, and listener's backchannel behavior (head nod and smile) [15] was generated by a probabilistic prediction model. Whereas the statistical approach needs real-life observations of a certain phenomenon to build a model, the theoretical approach requires more broad and general knowledge of the phenomena. At this moment, complete, coherent, and formal specifications of audience behavior cannot be derived from current theories; hence the statistical approach was applied in this study to generate the virtual audience's behavior.

3 System Overview

This section describes the high-level design of the audience model based on the implementation methods discussed in Section 2.1. Fig. 1 illustrates the framework of the integrated system and the architecture of the members of the autonomous audience. The overall structure of the agent architecture is based on common components of autonomous agents that should be able to perceive and act in the environment in which they are operating (see [20]). That is, the agent model includes a mind module for making decisions, a behavior module for translating the input from the mind into actions in the VE, and a perception module for perceiving the world (consisting of the VE and the user). Percepts work as input for the agent's decision making. In this way, the architecture implements a sense-reason-act cycle. This structure has also been widely used in ECAs, e.g., [15], [19]. Each module of this system is described in more detail later on, illustrating how they are composed to target the application of generating virtual audience behavior.

The *mind module* stores the values of the agent attributes. These attributes affect agent behavior and can be set by the operator. The agent attributes comprising personality, attitude (i.e., whether the agent is interested and positive towards the speech), mood, and energy level are assigned to two categories: the exogenous parameters and the endogenous parameters. Attributes regarded as static factors in a scenario such as personality and attitude belong to the exogenous parameters. These parameters can be set directly by the operator and remain constant unless they are modified by the operator. Dynamic attributes such as energy and mood belong to the endogenous parameters. As some ECA studies show, humans can perceive [21] and be affected by [22] virtual humans' affective states. Manipulation of the dynamic factors over time may be needed to regulate the user's state. Therefore, the endogenous parameters not only need initial values but are also influenced by the agent's *mental resource manager*. The mental resource manager stores emotional and physiological models, which can be defined by the operator and specify how the parameters change over time. According to the setting, the parameters will change automatically during the audience simulation. The parameters then feed into the *decision module*. Together with the perceived events such as a phone ringing or a fellow virtual human talking, the decision module will decide whether or not to react to these events. When the decision is made, the decision module will pass on the parameters to the *behavior module* to generate behavior.

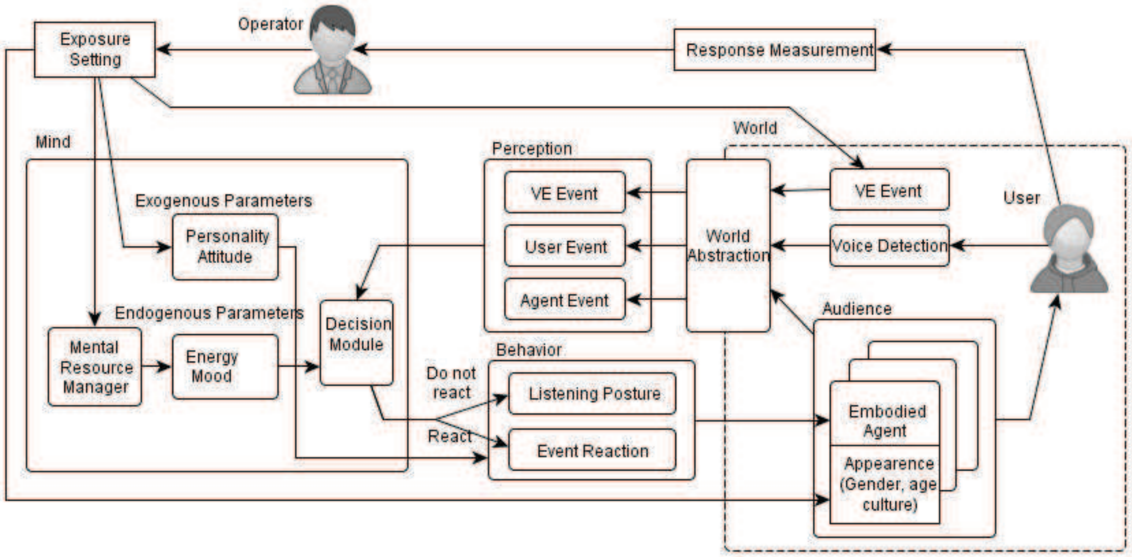

Fig. 1. The framework of the virtual audience simulator. The arrows (→) in the diagram illustrate the direction of information flow.

The behavior module has two sub-modules: a *listening posture module* and an *event reaction module*. If the agent decides to react to a perceived event, it will output this decision to the event-reaction sub-module of the behavior module. The event-reaction sub-module will then generate an event response. If it decides not to respond or no event is perceived, it will directly pass the parameter values to the listening posture sub-module of the behavior module to generate a listening posture. The behavior module will generate a posture or movement every two seconds using one of its sub-modules (i.e., event reaction or listening posture). This posture or movement will then be used to animate the embodied agent in the VE.

While the user is giving a talk and the virtual agents are being animated, the agent perceives the *world* through the *perception module*. The world consists of the VE and the user, who is immersed in the VE. The perception module acquires the information from the *world abstraction module*, which works as an interface between the world and the agent. It provides abstracted information about the world such as a door slam (*VE event*), agent-agent interactions (*agent event*), and *user events*, e.g., a user's performance obtained by evaluating the user's speech using voice detection technology [23]. These events are passed on as percepts to the mind module, which can use the event information in its decision making. For example, when an agent perceives a door slam, it can decide to turn around and look at the door.

Besides the perception-mind-behavior model in this system, the operator has direct control over certain aspects of the VE, in particular, the appearances of the virtual agents and the occurrence of VE events to meet the audience simulation requirements.

4 Data Collection for the Agent Model

As the autonomous audience's behavior was generated based on statistical models abstracted from real-life observations, we observed and analyzed real audiences' behavior in different conditions.

Pertaub et al. [24] found that the speakers' anxiety levels differ when they faced respectively a neutral static audience, a positive audience, and a negative audience (which exhibited bored and hostile expressions). To create an audience with more flexibility, besides the positive and neutral audiences, our observation data included two additional negative types: a critical audience and a bored audience. The critical audience was concerned about the speech topic but also critical of the talk. The bored audience was impatient and tired due to the boring speech. In summary, the virtual audience was designed to show at least four attitudes: positive, neutral, bored, and critical.

Additionally, audiences' personality and mood data were included in the model to add realism and variety. Personality affects listening behavior [25] and can be perceived from virtual human's behavior (e.g., [19]). Studies (e.g., [26]) have also shown that mood can be expressed and perceived in several ways, e.g., postures and facial expressions.

To achieve such a design, real audience's behavior was observed, and data of their personality, mood, attitude, and energy level was collected in four conditions: positive presentation, neutral presentation, boring presentation, and critical presentation. The audience behavior was then coded for the preparation of modeling the behavioral patterns.

4.1 Observation

4.1.1 Measures

The following measures were used to assess the personality, mood, attitudes, and energy levels of each audience member.

--*International Personality Item Pool* (IPIP-NEO [27]). The IPIP-NEO is a public domain collection of items for personality tests of adults. This study used a short inventory containing 120 items measuring the Big Five personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Each trait is scored on a scale of 0 to 99.

--*Self-Assessment Manikin* (SAM [28]). SAM is a nonver-

TABLE 1
QUESTIONNAIRE ITEMS FOR MEASURING ENERGY LEVEL AND AUDIENCE ATTITUDES

Questionnaire		Item	Label	
			0	6
ME	Energy	How is your current physical state?	Tired	Energetic
MA	Interest	What do you think of the topic?	Boring	Interesting
	Approval	How is your attitude towards the content?	Negative	Positive
	Eagerness for information	I was eager to get information and remember facts from the speech.	Extremely disagree	Extremely agree
	Criticism	I was critical to the speech and wanted to find flaws.	Extremely disagree	Extremely agree
	Impatience	I was impatient and hoped to finish as soon as possible during the speech.	Extremely disagree	Extremely agree

bal pictorial assessment technique that directly measures three dimensions of mood: valence, arousal, and dominance. Each dimension has a 5-point rating scale.

--*Measure of Energy Levels (ME) and Measure of Attitudes (MA)*. Self-designed questionnaires were used to assess the audience members' energy levels and attitudes towards the presentations. The items of ME and MA were all rated on a scale of seven points, specified in Table 1.

4.1.2 Procedure

16 participants (seven females, nine males) were recruited from fellow PhD students studying computer science. The participants' ages ranged from 24 to 33 years (*M* = 27.6, *SD* = 2.9). All the participants signed the informed consent form. None of them knew the speaker beforehand. The participants were split into two eight-person groups. The behavior of participants in one group was video-recorded while they acted as an audience listening to four different presentations. Right before the first presentation, the audience was asked to complete the SAM questionnaire to assess their emotional states and the ME for energy levels. After each presentation, lasting around seven minutes, they were asked again to fill in the SAM and ME to track their emotional and energy states, and MA to acquire audience's attitudes towards the presentation. In each presentation, an interrupting event (i.e., a door slam or a telephone ring) was arranged randomly.

The whole process was repeated with the other group of participants, but this time the presentations were given in a reverted order to avoid the potential order effects. Thus, a set of videos consisting of four conditions for 16 participants was obtained. All 16 participants completed the IPIP-NEO personality inventory afterwards. Ethical approval for this study was obtained from the university ethics committee.

4.1.3 Materials

The settings of the four presentations designed to evoke the four attitudes were as follows.

--*Positive presentation*. To obtain the audience's interest, the audience was told at the beginning that they would win a small prize if they listened carefully and got a high score in the quiz afterwards. The topic was a novel invention of a robot gripper which was much more advantageous than traditional ones to evoke a positive attitude.

--*Neutral presentation*. The topic was a software design method and there were no additional instructions for the audience.

--*Boring presentation*. The speaker read aloud some text from the book of Nicomachean Ethics by Aristotle. However, the order of paragraphs had been rearranged so that the talk no longer contained a clear story line and therefore was no longer understandable for the listeners.

--*Critical presentation*. The presentation criticized all the PhD students in the audience, saying that they did worst comparing with PhD students in other departments and in computer science departments at other universities. To address this, statistics were shown that only four out of 108 PhDs graduated on time over the last eight years and the average time needed for a PhD in the department to graduate was 5.5 years, which was a half year more than the average time needed for a PhD in computer science to graduate in the Netherlands. Additionally, a number of provocative policies were argued for, e.g., working hours from 9:00am to 6:00pm with only a half hour for lunch, and salary reduction if the research progress was slow. The speaker said that this presentation would also be given to the head of the department.

4.1.4 Condition Verification

To confirm that the audience attitudes were respectively positive, neutral, bored, or critical towards the presentations, the MA questionnaire results reflecting their attitudes were analyzed. The results (Table 2) show that the positive audience was significantly more interested ($t(15)$ = 10.02, $p <$ 0.001), more positive (i.e. high ratings in Approval, $t(15)$ = 7.12, $p <$ 0.001), more eager to get information ($t(15)$ = 4.37, p = 0.001), and less impatient ($t(15)$ = -7.01, $p <$ 0.001) than the bored audience. The critical condition was similar to the positive condition but significantly less positive ($t(15)$ = -5.00, $p <$ 0.001) and more critical ($t(15)$ = 4.79, $p <$ 0.001) than the positive condition. The questionnaire results of the neutral condition were always between the high-level and low-level results. Therefore, the audience was respectively positive, neutral, bored, and critical in the corre-

TABLE 2
MA QUESTIONNAIRE RESULTS, MEAN±SD

Questionnaire	Presentation condition			
item	Positive	Critical	Boring	Neutral
Interest	5.06 ± 1.18^{H}	4.19 ± 1.68^{H}	0.69 ± 1.08^{L}	3.94±1.34
Approval	4.94 ± 0.77^{H}	2.44 ± 1.86^{L}	1.44 ± 1.75^{L}	4.50±1.03
Eagerness for information	4.75 ± 1.61^{H}	4.50 ± 1.37^{H}	1.88 ± 1.71^{L}	3.00±1.63
Criticism	0.75 ± 1.00^{L}	3.50 ± 2.37^{H}	0.88 ± 1.93^{L}	1.63± 1.67
Impatience	1.00 ± 1.03^{L}	0.75 ± 1.24^{L}	4.12 ± 2.13^{H}	1.25±1.24

Note: a mean with H indication is significantly ($p <$ 0.01) higher than a mean with L indication within one questionnaire item.

TABLE 3
STATISTICS OF AUDIENCE DATA USED AS AGENT ATTRIBUTES

Dimension/ measures	Parameters	Measuring scale	Raw data			
			Min	Max	Mean	SD
Mood/ SAM	Valence	1 - 5	1	5	2.46	0.93
	Arousal	1 - 5	1	5	3.55	1.01
	Dominance	1 - 5	1	5	2.86	0.92
Energy/ ME	Energy	0 - 6	0	6	3.61	1.38
Attitude/ MA	Interest	0 - 6	0	6	3.47	2.12
	Approval	0 - 6	0	6	3.58	1.99
	Eagerness for information	0 - 6	0	6	3.53	1.94
	Criticism	0 - 6	0	6	1.69	2.09
	Impatience	0 - 6	0	6	1.78	1.99
Personality/ IPIP-NEO	Openness to Experience	0 - 99	0	77	36.81	24.02
	Conscientiousness	0 - 99	1	99	60.56	29.78
	Extraversion	0 - 99	0	96	46.00	28.08
	Agreeableness	0 - 99	7	88	58.13	25.41
	Neuroticism	0 - 99	1	64	28.50	20.37

sponding conditions.

The questionnaire results were used as agent attributes in the agent modeling. As the attributes were rated on different scales (Table 3), all the attribute data was normalized by subtracting the minimum of each attribute from the raw value and then dividing the difference by the difference between the maximum and the minimum of the raw value. Hence the data for each attribute covered the whole scale. This result was then multiplied by 10 so that these attributes were rated on a common scale ranging from 0 to 10.

4.2 Coding Postures

To annotate the recorded video and characterize the audience's behavioral patterns, a posture-coding scheme was developed. The coding scheme describes how a certain part of the body moves in three-dimensional space using both anatomical and external reference frames [29]. A posture code consists of five sub-codes to convey position or movement information of the following parts: head, gaze, arms and hands, torso, and legs. Table 4 shows the posture-coding scheme for each part. That is, a sitting posture can be noted as a combination of the five sub-codes, i.e., HxGxAxTxLx. For example, if a person turns the head to the right (H2) and looks at the right side (G4), sitting up straight (T1) with a hand tapping on the desk (A17) and twisted ankles (L2), the posture will be annotated as H2G4A17T1L2. Although the coding scheme includes the audience's gaze information, the gaze has not been implemented in the behavioral model currently. Hence the postures that are mentioned below are combinations of positions or movements of the four parts: head, arms and hands, torso, and legs.

Fig. 2 shows a video screenshot of an observed audience. To determine the sampling interval [30], the codable behaviors in the videos were analyzed. The shortest duration of a codable state was 2 seconds. Thus, with an interval of two seconds, a coder coded each audience member's postures with the posture-coding scheme. The postures were coded by recording four position variables: head, arms and hands, torso, and legs. Taking the coding of leg positions for example, the coder had a choice out of three position codes (i.e., L1, L2, and L3) for each coding unit. Moreover, when an interrupting event occurred, additional information was recorded, specifically the reaction of each participant (i.e., turning angle of the head and reaction duration) and the event information (i.e., the direction and the duration of the event). To assess the coding reliability, an additional coder was trained and independently coded an eight-minute video sample of one audience member according to the coding scheme. The sample consisting of 240 units was coded out of a total length of 320 minutes (i.e., 9600 units). To avoid a biased sample, a representative sample was selected with similar frequencies of posture shifts as observed on average in the whole corpus, i.e., 1.25, 1.50, 0.63, and 0.38 behavior shifts per minute for head, arms and hands, torso, and legs respectively for the sample, and 1.26 (*SD* = 1.66), 1.19 (*SD* = 1.27), 0.40 (*SD* = 0.76), and 0.34 (*SD* = 0.88) for the whole corpus. Like the first coder, the second coder coded the four position variables at two-second intervals from the same starting point. The coding agreement between the two coders was assessed by computing Cohen's kappa [31] for each variable. The agreement coefficients for the four position variables were respectively 0.85, 0.85, 0.94, and 0.93, which shows an acceptable level of agreement [32]. Note that the combination of relatively few behavioral shifts and the relatively short two-second sampling intervals created large sequences without variations, which might cause relatively high agreement level.

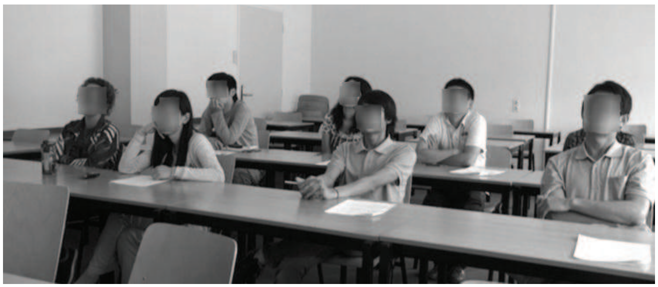

Fig. 2. A video screenshot of an observed audience.

From the recorded videos lasting around 320 minutes, over 300 unique postures (presented by unique combinations of the four sub-codes, HxAxTxLx) were observed. To simplify the analysis and system implementation, the postures which occurred less than 6 times (an occurrence of a posture was counted only when the posture changed to another) in the whole observation were removed, resulting in 59 postures. The remaining coding still accounted for 80% of the 9600 coding units.

The collected data was used to build statistical models of audience behavior. The next two sections explain the statistical models and how the *mind module* and the *behavior module* use these models to generate the virtual agent's behavior.

5 The Mind Module

The *mind module* stores the agent attributes that affect the virtual agent's behavioral style and includes a *decision module* for event response. The agent attributes, including personality, attitude, energy and mood, are presented by parameters

TABLE 4
THE POSTURE CODING SCHEME

Part of body	Behavior category	Short description
Head	H1: Head up	The head keeps the neutral position.
	H2: Head turn	The head turns right or left.
	H3: Head down	The head is lowered.
	H4: Head tilt	The head tilts right or left.
	H5: Head nod	Nod the head: the head moves down and then up again quickly.
	H6: Head shake	Shake the head: the head turns from side to side.
Gaze	G1: Towards the speaker	The gaze is directed towards the speaker.
	G2: Upwards	The gaze is directed above the speaker position.
	G3: Downwards	The gaze is directed below the speaker position.
	G4: Averted sideways	The gaze is directed away from the speaker
Arms and hands	A01: Hands on legs	Both hands are on the legs.
	A02: Open arms on desk	Both arms rest on the desk without touching each other.
	A03: Arms crossed	The arms are crossed in front of the body.
	A04: Catapult	The hands are holding behind the head like a catapult.
	A05: Hands steeple	The fingers of one hand lightly press against those of the other hand to form a church steeple.
	A06: Hands clenched	The hands are clenched, and the elbows rest on the desk.
	A07: Chin/ cheek touch	One hand touches the chin or cheek, and the other arm rests in front of the body.
	A08: Supporting head	One or two arms support the head and the elbows rest on the desk.
	A09: Desk and chair back	One arm rests on the desk and the other rests on the chair back.
	A10: Self-hold	One arm swings across the body to hold or touch the other arm.
	A11: Desk and leg	One arm rests on the desk and the other rests on one leg.
	A12: Nose touch	One hand touches the nose, and the other arm rests on the front torso.
	A13: Eye rub	One hand rubs the eyes, and the other arm rests on the front torso.
	A14: Ear touch	One hand touches the ear, and the other arm rests on the front torso.
	A15: Neck touch	One hand touches the neck, and the other arm rests on the front torso.
	A16: Mouth touch	One hand touches the mouth, and the other arm rests on the front torso.
	A17: Hand tap	One or two hands tap the desk continuously.
Torso	T1: Torso upright	The torso keeps upright.
	T2: Torso forward	The torso leans forward and the spine keeps straight.
	T3: Torso backward	The torso leans backward and the spine keeps straight.
	T4: Torso back in the chair	The torso leans back in the chair and the spine is relaxed.
	T5: Torso bent forward	The torso leans forward, and the spine is bent forward.
Legs	L1: Standard position	The knees are bent at a right angle with both feet flat on the ground, and the legs are not crossed.
	L2: Legs crossed/ twisted	The legs are crossed or the ankles are twisted.
	L3: Leg joggle/ tap	The upper leg joggles or the lower leg taps the floor when the legs are crossed, or one or two feet tap the floor when both feet rest flat on the floor.

listed in Table 3. Before passing the parameters to the *behavior module*, the *decision module* will first decide whether or not the agent should respond in case of interrupting events.

To mimic the probabilities with which real persons would respond to the events, the decision model was trained using the observed data. A supervised classification method, logistic regression, was applied. The agent parameters (Table 3) and event information (i.e., event duration and event location) were used as predictors. The training and test data used the normalized questionnaire results and the coding information of the observations. For the endogenous parameters, such as energy level and mood, data was collected before and after each presentation. To simplify the model, the values of these parameters were assumed to change linearly during the presentation. Hence the missing data during each talk for mood and energy level were linearly interpolated using the results of ME and SAM before and after each presentation.

A sample of 39 cases was drawn from the original data set with an almost equal number of cases where a person did or did not respond to an event. This avoided a biased function caused by an imbalanced data set where the sizes of classes are not similar. The logistic regression model can be expressed by the following formulae:

$$p = 1/\{1 + \exp[-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)]\}$$
$$Y = \begin{cases} 1 & p \geq 0.5 \\ 0 & p < 0.5 \end{cases} \qquad (1)$$

where b_0, b_1, b_2, …, and b_k are regression coefficients for predictor variables, x_1, x_2, …, and x_k, p presents the probability for the agent to respond, and Y is the prediction output. The model selected a cutoff point of 0.5, i.e., the prediction is to respond to events (Y = 1) when p is no less than 0.5, otherwise the prediction is not to respond (Y = 0). A test of the full model versus a model with intercept only was statistically significant, $\chi^2(3, N = 39) = 18.92$, $p < 0.001$, with an overall correct prediction of 81.2% (85.0% for non-response and 78.9% for response cases). This model was also tested on 11 cases (five response and six non-response cases) that had not been used for training. The overall correct prediction was 90.9% (i.e., five response and five non-response cases were correctly classified). This result was significantly (binomial test, p = 0.01) above a case allocation of 54.54% (i.e., 6 out of 11 cases).

6 The Behavior Module

The *behavior module* generates listening postures using the parameters from the *mind module* and generates head turns as an event reaction if the mind has decided to react.

6.1 Generation of Listening Postures

To derive listening postures from agent parameters, a relationship between the parameters and listening behavior was established in the module. To do this, the 59 observed postures (see Section 4.2) were first categorized so that the agent attributes could be used to predict a category, from which a posture would be selected afterwards.

The 59 postures were clustered according to their transition probabilities to other postures. The posture sequences were transformed to a 59×59 transition matrix, i.e.,

$$P = \begin{pmatrix} p_{1,1} & \cdots & p_{1,j} & \cdots & p_{1,59} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i,1} & \cdots & p_{i,j} & \cdots & p_{i,59} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{59,1} & \cdots & p_{59,j} & \cdots & p_{59,59} \end{pmatrix} \quad (2)$$

where $p_{i,j}$ is the probability of transitioning from posture i to posture j within two successive observations, i.e., every two seconds. The postures with similar transition probabilities were clustered into one category. Take Table 5 for example, each row presents the probabilities for one posture to transition to Posture 1, 2, and 3 respectively. As the probabilities for Posture 1 and 2 are very similar, i.e., all around 0.71, 0.29, and 0.00, the two postures were clustered together. This also means that the postures in one category were always followed by postures from a certain posture set (here in this example the set consists of Posture 1 and 2). Therefore, each row of the transition matrix shown in (2), presenting the probabilities of transition from one posture to others, was used as a clustering feature. An agglomerative hierarchical clustering method with Ward linkage was then employed to group the postures using a Euclidean distance measurement. A distance threshold of 20 on a scale from 0 to 25 was set to seek an optimum in the maximum number of categories and a maximum number of similar postures within each category. Thus, 15 categories were identified with each containing 3 to 7 similar postures, which often differ only in one of four coding parts, e.g., H1A08T5L3 and H4A08T5L3.

The second step was to create a logistic regression prediction model using agent parameters (i.e., mood, energy, attitude, and personality) as predictors. Since logistic regression predicts a dichotomous outcome, i.e., whether the parameter set belongs to a certain category or not, the models were trained separately to predict each category. Like the training data for event reaction, this training data was also randomly sampled so that the data for each category was distributed equally. 15 prediction functions were established using the following form:

$$p_i = 1/\{1+\exp[-(b_{0i}+b_{1i}x_1+b_{2i}x_2+\cdots+b_{ki}x_k)]\} \quad (3)$$

where p_i represents the probability of being category i, x_1, x_2, …, and x_k are the predictor variables, and b_{0i}, b_{1i}, …, and b_{ki} are the regression coefficients for category i. To predict the exact category using a set of agent attributes, x_1, x_2, …, and x_k, one probability was calculated for each of the 15 categories by (3) respectively, and eventually based on this set of probabilities the predicted category n would be selected that satisfies

$$p_n = \max\{p_1, p_2, \cdots, p_{15}\}\,. \quad (4)$$

The overall correct prediction of the training set was 66.4%, ranging from 48.2% to 88.6% for individual categories.

As 20% of the balanced observed data had not been used for training, it was possible to conduct a holdout validation for the model. The test data included 50 cases of each category, i.e., 750 cases in total. The results of the test set showed an overall correct prediction of 64.4% for the 15 categories, which was significantly (binomial test, $p < 0.001$) above the random allocation threshold of 6.7% (i.e. 1 out of 15). For the 15 individual categories, correct prediction ranged from 48.0% to 94.0%, which were all significantly (binomial test, $p < 0.001$) above the random allocation threshold of 6.7%. Interestingly, we compared the effect of exogenous parameters (i.e., personality and attitude) and endogenous parameters (i.e., energy and mood) on the prediction of a category by analyzing the odds ratios of the regression coefficients, i.e., $\exp(b_{ki})$. The absolute value of b_{ki} was examined so that odds ratios $\exp(b_{ki})$ and $\exp(-b_{ki})$ can reflect the same impact on the category selection. For each function, the sums of the absolute coefficients for exogenous and endogenous parameters were calculated, i.e., $\sum_{k=1}^{m}|b_{(exo)ki}|$ and $\sum_{k=1}^{n}|b_{(end)ki}|$ where m and n are the numbers of exogenous and endogenous parameters in the ith function. By comparing the sums of all 15 functions, the effect of exogenous parameters ($M_{\sum|b(exo)|} = 19.22$, $SD_{\sum|b(exo)|} = 25.32$) was found to be greater than that of endogenous ones ($M_{\sum|b(end)|} = 1.98$, $SD_{\sum|b(end)|} = 2.82$), $t(14) = -2.87$, $p = 0.01$. Compared with the endogenous parameters, the exogenous had more effect on the category selection on average. Still all parameters contributed to the model significantly ($p < 0.05$) according to the Wald statistics.

TABLE 5
AN EXAMPLE OF THE POSTURE TRANSITION MATRIX WITH ACCUMULATIVE PROBABILITY INTERVALS

		Category 1		Category 2
		Posture 1	Posture 2	Posture 3
Category 1	Posture 1	0.73	0.27	0.00
		[0, 0.73)	[0.73, 1]	(1, 1]
	Posture 2	0.69	0.31	0.00
		[0, 0.69)	[0.69, 1]	(1, 1]
Category 2	Posture 3	0.02	0.03	0.95
		[0, 0.02)	[0.02, 0.05)	[0.05, 1]

After a behavior category is determined, a posture will be selected within this category. Since the *behavior module* updates the embodied agents' behavior every two seconds, the category is very likely to remain unchanged. When the category does not change, the posture will be selected according to the transition matrix to keep the sequential pattern of the behavior. For example, supposing that the current posture category is 1 and the current posture code is 1, the following posture should be selected within Category 1. According to the transition probabilities in Table 5, the next posture has a chance of 0.73 to be Posture 1 and 0.27 to be Posture 2. To select a posture, a random number between 0 and 1.00 is generated as the accumulative probability. If the number is 0.96, which is within the range of [0.73, 1], Posture 2 will be selected.

Accepted version of the paper: Kang, N., Brinkman, W.-P., van Riemsdijk, M.B., Neerincx, M.A. (2014). An expressive virtual audience with flexible behavioral styles. IEEE Transactions on Affective Computing. doi 10.1109/TAFFC.2013.2297104

6.2 Generation of Event Responses

Since turning one's head was the only event response considered in the observation data, the module only determines how many degrees an agent turns its head (turning angle *TA*) and how many seconds the response takes (response duration *DUR*). In the audience observation, information of interrupting events and audience responses was recorded. This information as well as the audience attributes was used to predict the turning angle of a virtual human's head *TA* and its response duration *DUR*.

Linear regression results indicated that only the event direction, which is *a* degrees ($-180° < a \leq 180°$) relative to the front direction of the virtual agent, significantly predicted the agent's turning angle *TA* (R^2 = 0.78, $F(1, 18)$ = 63.03, β = 0.88, p < 0.01) by the following model:

$$TA = 0.517a \,. \tag{5}$$

The event direction also significantly predicted the agent's response duration *DUR* in seconds (R^2 = 0.80, $F(1, 18)$ = 73.65, β = 0.90, p < 0.01), in the following form:

$$DUR = 0.013\,|\,a\,| \,. \tag{6}$$

As indicated by the two models, when an event occurs right behind the virtual human (i.e., $a = 180°$), the virtual human would turn its head about 90 degrees, and the response duration lasts about 2.3 seconds at maximum. Contrarily, if an event is in front of the agent, assuming that a<20°, the reaction duration would be very short according to the rules, i.e. *DUR* < 0.26s. In practice, this short head turn would not be acted out, which coincides with our observation: when an event occurs in front, the head turn was unnecessary because the event was still in the audience member's field of view.

7 The World and the Perception Module

As explained in Section 3, the agent perceives the world through the *perception module*. The information about what is happening in the world needs to be abstracted into events to be usable for the *decision module*. The interrupting events in the *world* include *VE events*, *agent events*, and *user events*. The *VE event* is evoked by the "physical" objects in the VE, such as door slam and telephone ring, and can be set in the therapy settings. The *agent event* refers to the agent's behavior which may evoke interaction with another agent. For example, a head turn is generated by the *behavior module* so that an agent look at another agent for a while, which may cause another agent to turn the head back. The *user event* relates to the user's performance which may evoke the agents to change their behavior, e.g., the user stops talking for a moment, which may result in the distraction of the agents.

The way in which information is abstracted depends on what is needed at the level of decision making. For example, concerning the user speaking, it may be enough to generate an event to indicate whether she is speaking or not. However, if this system enables speech interaction with the human speaker, e.g., asking questions about the talk, the *world abstraction module* needs to pass on more detailed information about it, such as the topic and key words. Thus, the *world abstraction module* also works as an information provider and makes the system easy to be extended.

While the agents perceive the world, the operator should be aware of the information from the world too. The user's response feedback may include information such as gaze direction [33] and anxiety or stress level by measuring anxiety such as subjective unit of discomfort and psychophysiological data [34]. Additionally, information on the interrupting events in the VE could be recorded with response measurement in a log file so that the user's response can be analyzed afterwards.

Moreover, the operator has direct control over certain aspects of the VE, e.g., the virtual human's appearances which can be determined by static appearance parameters like gender and age, and the occurrence of VE events defined by parameters such as event location and event duration. Other controllable appearance elements could also be added such as ethnicity and clothes to construct a more realistic environment [35].

8 Perception Evaluation of a Virtual Audience

This study proposes a framework for a public speaking simulation system in which an operator can control the behavioral styles of an autonomous virtual audience. Among the components of this system, this study mainly focused on the audience model and the creation of such an audience. Since the previous sections already show that the audience model fits well with the corpus data, a next step was to examine how people perceive the audience. To do this, an autonomous audience in a public speaking situation was created using this model so that individuals could evaluate the model by watching the audience's behavior.

8.1 Method

8.1.1 Hypotheses and Experiment Design

The hypotheses for this evaluation were that people could perceive the different audience's attitudes (H1), moods (H2), and personalities (H3) from the behavioural styles modulated by corresponding parameters.

To test H1, the evaluation mainly examined the perception of the four designed attitudes: critical, positive, neutral, and bored. To further investigate whether people can recognize the different degrees of a certain attitude, the positive and bored attitudes respectively included two conditions: an extremely positive condition and a positive condition, and, an extremely bored condition and a bored condition. Hence there were six attitude conditions: a critical attitude, an extremely positive attitude, a positive attitude, a neutral attitude, a bored attitude, and an extremely bored attitude.

Concerning H2 and H3, the study only explored some of the mood and personality dimensions, namely, valence, arousal, and extraversion. These dimensions were selected because these dimensions may be perceived more easily than others, e.g., neuroticism [36]. Thus, the study also includes six additional audience conditions labeled as follows: extrovert, introvert, high arousal, low arousal, positive valence, and negative valence.

To test these hypotheses, the evaluation was conducted in two ways with different participants. One group of partici-

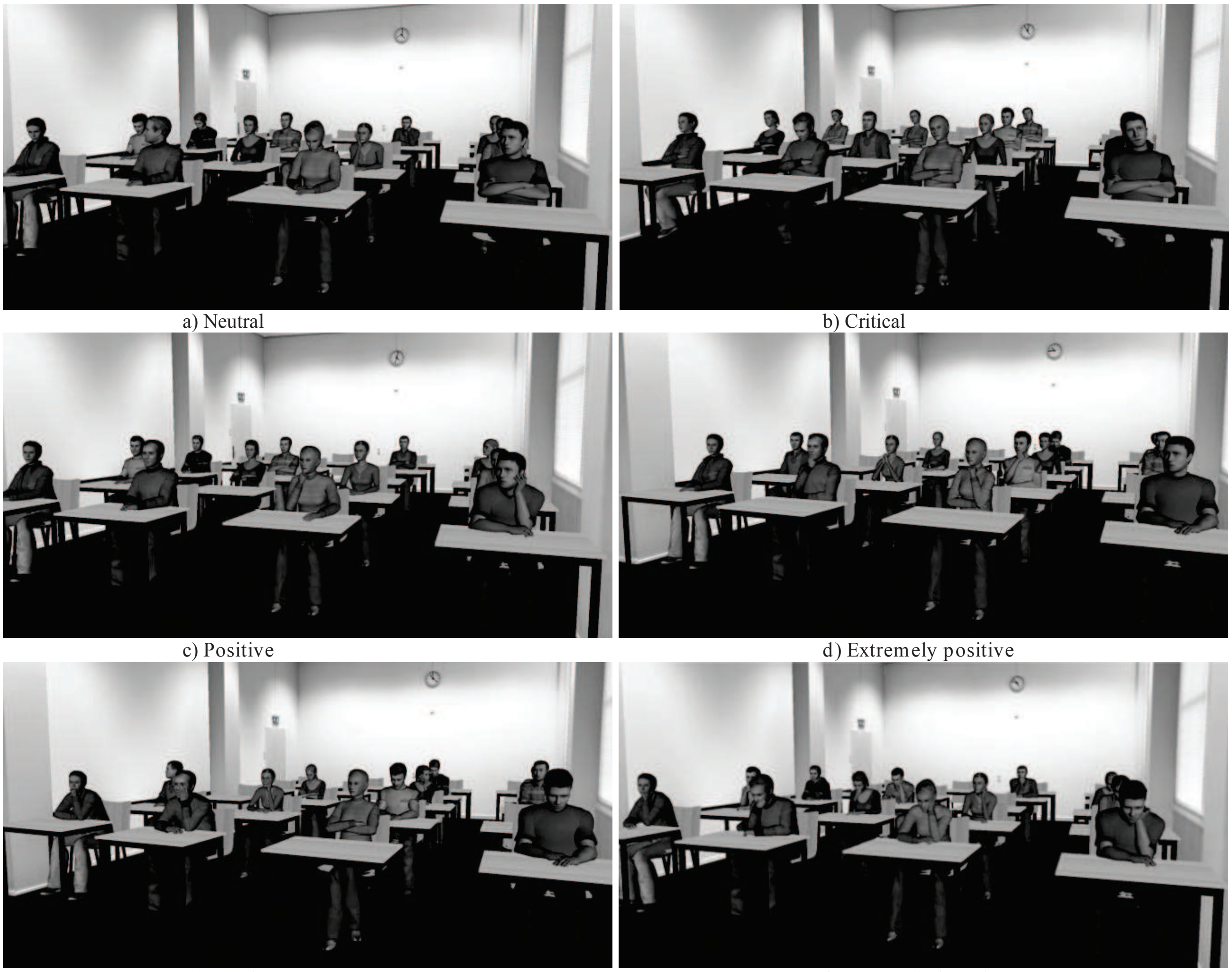

Fig. 3. Snapshots of the autonomous audience in six attitude conditions.

pants was asked to describe the audience's state freely. This open-question avoided framing their observations or biasing the participants' response towards a specific factor. Therefore, the audience description should reflect their natural thoughts. The second group was asked to rate their observation with a questionnaire to obtain information on the factors to be examined.

8.1.2 Materials

An executable program was made to display the simulation of a public speaking situation in which a 12-person audience was seated in a classroom. The executable program generated the audience's behavior in real time so that participants watched different audience animations due to the random element in the simulation. The viewpoint was set from the perspective of an outsider in front of this audience slightly on the right (Fig. 3). The speech the audience listened to was selected from the news report in Uygur language so that participants could not understand the speech and therefore would not be affected by the speech content.

12 audience conditions of one minute each were created to show different attitudes, moods, and personalities. These 12 conditions were created by setting each agent's attributes as Table 6. The attitude conditions were made by modulating the attitude parameters. Specifically, the four conditions, namely critical, extremely positive, neutral, and extremely bored, were created according to the results of attitude questionnaire (MA) obtained in the four observed situations (see Table 2). The positive and the bored conditions employed moderated settings of the extremely positive and the extremely bored conditions by adjusting one or two parameters from the extremities to a medium level. Fig. 3 shows the snapshots of the audience in six attitude conditions.

Instead of using the observed audience conditions, the effects of the mood and personality parameters were explored by setting extremities in the examined dimensions, e.g., the extrovert condition only set the parameter Extraversion as high.

8.1.3 Measures

A questionnaire about the virtual audience was designed to quantitatively measure the perceived audience's attributes, including attitude, mood, and personality. The attitude and mood questions were adapted from MA and SAM (section 4.1.1) to refer to the virtual audience's state. For example, the criticism item in the questionnaire became:

> *The audience was critical towards the speech and wanted to find flaws.*

Additionally, the questionnaire did not include the dimensions that were not evaluated in the 12 conditions, e.g., the dominance item in SAM. Thus, the questionnaire included five attitude items and two mood items. It also included a personality item, formulated as follows:

Most audience members scored high on extraversion.

This item was rated on a 7-point scale, from “extremely disagree” (i.e., 0) to “extremely agree” (i.e., 6).

Therefore, the hypotheses could be tested by comparing these measurements with the parameter settings of the evaluation conditions.

8.1.4 Procedure

The evaluation included two parts: free description and factor rating. For the first part, 22 participants (10 females, 12 males) were recruited throughout the university campus to evaluate the virtual audience. Their ages ranged from 22 to 38 years with a mean of 28.1 (*SD* = 3.2) years. Each participant was asked to watch a 12-minute audience simulation using a Sony HMZ-T1 head-mounted display (HMD) with an orientation tracker to track the participant’s head orientation. The HMD displayed a virtual image comparable to viewing a 720-inch display at 20 meters and the visual field spanned 45 degrees diagonally. The resolution of the right and left display was 1280*720 (horizontal*vertical) pixels with a refresh rate of 60Hz.

While watching the simulation, the participant was asked to observe and describe orally the state of the audience. To avoid framing or biasing the participants’ description, no examples of audience description were given to the participants. Their description was audio recorded. Among the 22 participants, 16 participants were Chinese and reported in standard Chinese in the experiment, and six other participants were Dutch and Iraqi and they reported in English. The order of those 12 conditions was randomly given to each participant to avoid the potential order effects.

For the factor rating part of the evaluation, another 22 participants (13 females, 9 males) were recruited. Their ages ranged from 23 to 44 years with a mean of 28.5 (*SD* = 5.5) years. The participants included four therapists and four psychology master students who all had experience in using virtual reality exposure system [37] to treat patients with social anxiety disorder. The other 14 participants, with no such experience, were recruited throughout the university campus. Like the first part, each participant was asked to watch the 12 audience conditions using the HMD in a random order. However this time, the participants were asked to rate the factors with the questionnaire after watching each condition.

TABLE 6
PARAMETER SETTINGS FOR THE AUDIENCE CONDITIONS

Condition	Parameter setting
Critical	Attitude parameter: Interest (H), Approval (L), Eagerness for information (H), Criticism (H), Impatience (L).*
Extremely Positive	Attitude parameter: Interest (H), Approval (H), Eagerness for information (H), Criticism (L), Impatience (L).
Positive	Attitude parameter: Interest (H), Approval (H), Criticism (L).
Neutral	*This condition was set as baseline: all parameters were set at the medium level.*
Bored	Attitude parameter: Interest (L), Eagerness for information (L), Criticism (L), Impatience (H).
Extremely Bored	Attitude parameter: Interest (L), Approval (L), Eagerness for information (L), Criticism (L), Impatience (H).
Extrovert	Personality parameter: Extraversion (H).
Introvert	Personality parameter: Extraversion (L).
Positive Valence	Mood parameter: Valence (H).
Negative Valence	Mood parameter: Valence (L).
High Arousal	Mood parameter: Arousal (H).
Low Arousal	Mood parameter: Arousal (L).

**Each condition was set up by attitude (MA), mood (SAM), and personality (IPIP-NEO) parameters. This table only specifies the parameters which were set as High (9) or Low (1). All the other parameters were set at the neutral level (5).*

8.2 Analysis and Results

8.2.1 Free Description

To statistically investigate whether the participants could recognize the different conditions, a coding scheme for their description was developed, shown in Table 7. Each participant’s comments for each minute were analyzed afterwards by a coder. The coder recorded whether or not the comments in a condition included terms that would fall into one or more of the eight description categories. In this way, a set of eight binary digits was obtained per participant per condition.

TABLE 7
THE CODING SCHEME FOR THE RECORDED DESCRIPTION AND RELIABILITY ASSESSMENT

Description category	Short description and utterance examples	Cohen’s kappa
Attentive	The audience pay attention to the speech and may be attracted: “paying attention”, “attentive”, “interested”, “thoughtful about the speech content”, “curious”, “concentrate”, “focus”, “pleased with the speech”, “positive attitude”	0.94
Neutral	The audience do not show any interest or negative attitude towards the speech: “neutral”, “neutrality”	*
Distracted	The audience do not pay attention to the speech: “distracted”, “looking away”, “inattentive”, “day-dreaming”	0.88
Bored	The audience are bored with the speech and impatient to wait to the end: “bored and tapping the desk/ floor”, “frustrated about being there”, “impatient”, “lose interest”, “unhappy to be there”	0.87
Critical	The audience show disapproval towards the speech: “disagree”, “critical”, “try to find flaws”, “ angry with what is talking about”, “ skeptical”	1.00
Active	The audience are active and aware of the surroundings: “active”, “awake”, “conscious”, “alert”	*
Sleepy	The audience almost fall asleep or show a state of low energy level: “sleepy”, “slouching”, “tired”	1.00
Relaxed	The audience are physically relaxed, instead of sitting straight up: “relaxed”	0.87

**No statistics were computed because no rater had coded the sample recordings using this category, resulting in two constant coding sequences.*

TABLE 8
NUMBER OF PARTICIPANTS WHO USED CERTAIN DESCRIPTION CATEGORIES AND RESULTS OF WILCOXON SIGNED RANK TESTS (*N* = 22)

		Description category							
Condition		Attentive	Neutral	Distracted	Bored	Critical	Active	Sleepy	Relaxed
	Neutral	12	1	16	8	2	1	5	2
Comparison with Neutral condition	Critical	11	0	7*	4	2	0	4	1
	Extremely Positive	18*	0	7*	4	0	2	5	4
	Positive	15	0	12	4	1	2	5	2
	Bored	8	0	9	12	1	1	3	2
	Extremely Bored	4*	0	16	15*	0	0	6	0
Comparison with its opposite condition	Extrovert	13	2	6	6	4	0	3	3
	- Introvert	16	2	6	8	3	2	0	2
	Positive Valence	14	2	3	10	2	3	3	1
	- Negative Valence	13	0	9	7	4	1	4	3
	High Arousal	13	3	9	6	1	0	7	2
	- Low Arousal	15	0	9	6	0	2	4	1

*$p < 0.05$

To assess the reliability of the coding, another coder was trained to code the audio recordings according to the coding definitions. The additional coder coded independently a sample of 36 minutes out of a total length of 264 minutes. The coding agreement between the two coders was assessed by computing the Cohen's kappa [31]. Table 7 also presents the agreement coefficients varying from 0.87 to 1.00, showing an acceptable agreement level [32].

After coding all the recordings, 22 sets of binary data were obtained from all the participants for each condition. The data sets were then added up to count how many participants have mentioned a certain category in one condition (Table 8) to establish an overview of the differences across those conditions.

To examine if participants' utterance responses differed among the conditions, Wilcoxon signed rank tests for two related samples were conducted on each description category. The attitude conditions were respectively compared with the neutral condition, which was regarded as the baseline condition. Other conditions were respectively compared with their opposite conditions, e.g., Positive Valence versus Negative Valence. The results are also presented in Table 8.

The participants described the audience as distracted significantly (z = -2.32, p = 0.02) more often in the Neutral condition than the Critical condition. Furthermore, they described the audience as attentive more often (z = -2.12, p = 0.03) and as distracted less often (z = -2.32, p = 0.02) in the Extremely Positive condition than in the Neutral condition. Finally, compared with the Neutral condition, the audience in the Extremely Bored condition was described as less attentive (z = -2.14, p = 0.03) and more bored (z = -2.65, p = 0.008). This suggested that the participants could significantly differentiate an extremely positive or extremely bored audience from a neutral audience.

No significant difference was found in comparisons between Positive and Neutral and between Bored and Neutral conditions. However, when positioning the description results in the order of Extremely Positive, Positive, Neutral, Bored, and Extremely Bored condition, a trend seems to appear for the Attentive and Bored categories, as shown in Fig. 4. This suggested that the participants may even perceive the different degrees of a certain attitude, e.g., differentiating extremely positive attitude from positive attitude.

For the exploratory conditions, it seems that the participants did not make any reference with regard to the moods and personalities.

8.2.2 Factor Rating

Nonparametric tests were also conducted on rated items for the audience conditions because the ratings were not normally distributed. The analysis results of ratings are shown in Table 9. To investigate whether the perceptions of therapists and non-therapists were consistent with each other, Spearman's correlations were calculated between the medians of eight (students-)therapists and 14 non-therapists for the different items. The correlations ranged from weak positive (0.23) to very strong positive relationships (0.77) with an average strong positive relationship (0.52), which suggests a reasonable level of agreement between the two groups across the items. Therefore, the analysis was conducted on the data from all participants.

To verify whether the differences across the conditions correspond to the condition settings as hypothesized, all the conditions were compared with baseline conditions. The Extremely Bored condition was selected as the baseline for attitude conditions, hypothesized to receive a lower score for

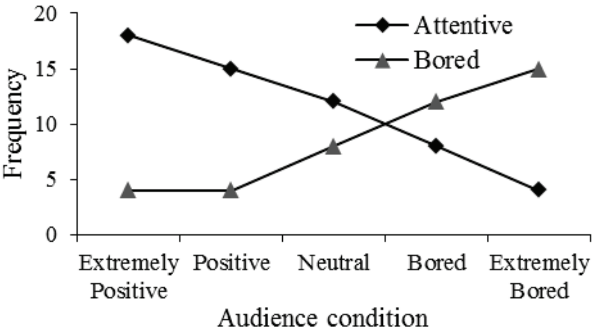

Fig. 4. The frequency of Attentive and Bored in audience description against audience attitude conditions.

TABLE 9
MEDIAN AND TEST RESULTS OF FACTOR RATINGS FOR THERAPISTS (T, N=8), NON-THERAPISTS (NT, N=14), AND ALL PARTICIPANTS (A, N=22)

Condition	Questionnaire item: Interest			Criticism			Approval			Eagerness for information			Impatience			Valence			Arousal			Extraversion		
	NT	T	A	NT	T	A	NT	T	A	NT	T	A	NT	T	A	NT	T	A	NT	T	A	NT	T	A
Neutral	3.0	3.5	3.0	3.0	3.0	3.0	3.0	3.5	3.0	3.0	2.5	3.0	3.0	2.5	3.0	4.0	4.0	4.0	4.0	2.0	4.0	3.0	3.0	3.0
Critical	4.0	4.0	4.0^{H}	3.5	3.0	3.0^{H1}_{H2}	3.5	4.0	4.0^{H}	4.0	4.0	4.0^{H1}	3.0	2.5	3.0^{L}	4.0	4.0	4.0^{H}	3.5	2.5	3.0	3.0	3.5	3.0^{H}
Extremely Positive	4.5	4.0	4.0^{H}	3.0	1.0	1.5^{L2}	3.5	4.0	4.0^{H}	5.0	4.0	4.0^{H1}_{H2}	1.0	2.0	1.0^{L}	4.0	4.0	4.0^{H}	4.0	3.0	3.0	3.0	3.0	3.0^{H}
Positive	4.0	2.5	4.0^{H}	2.0	1.0	2.0	3.0	2.5	3.0	3.0	2.5	3.0	2.0	2.5	2.0^{L}	4.0	4.5	4.0^{H}	3.0	3.0	3.0	3.0	3.5	3.0^{H}
Bored	4.5	3.5	4.0	3.0	2.0	3.0	3.0	3.5	3.0	4.0	2.5	3.0^{L2}	2.5	2.5	2.5	3.5	4.0	4.0	3.5	2.0	2.0	3.0	3.0	3.0
Extremely Bored	1.0	0.5	1.0^{L}	1.5	1.0	1.0^{L1}	1.5	1.0	1.0^{L}	1.0	1.0	1.0^{L1}	4.5	5.0	5.0^{H}	2.0	2.5	2.0^{L}	2.0	1.5	2.0	2.0	1.5	2.0^{L}
Extrovert	4.0	3.0	4.0	2.5	2.0	2.0	3.5	3.0	3.0	4.0	3.0	4.0	2.0	2.5	2.0	4.0	4.0	4.0	4.0	2.0	2.5	4.0	3.5	4.0
Introvert	2.5	2.5	2.5	3.0	1.5	3.0	2.5	2.5	2.5	3.0	1.5	2.0	3.0	3.5	3.5	2.5	4.0	3.0	3.5	2.0	2.0	3.0	2.5	3.0
Positive Valence	5.0	4.5	5.0	3.5	1.0	3.0	4.0	4.0	4.0	5.0	3.5	4.5	1.5	2.5	2.0	4.0	4.0	4.0	4.0	3.0	4.0^{H}	3.0	3.5	3.0
Negative Valence	4.5	4.0	4.0	3.0	1.5	2.5	3.5	3.5	3.5	4.0	3.0	4.0	2.0	2.0	2.0	4.0	4.0	4.0	2.5	2.0	2.0^{L}	3.5	1.5	3.0
High Arousal	3.0	3.0	3.0	3.0	1.5	2.5	3.0	3.0	3.0	3.0	2.0	3.0	3.0	4.0	3.5	4.0	4.0	4.0	3.0	2.0	2.5	3.0	2.5	3.0
Low Arousal	4.0	2.5	4.0	4.0	2.0	3.0	3.5	2.0	3.0	4.0	1.5	4.0	1.5	3.5	2.0	4.0	3.0	4.0	4.0	2.0	2.0	3.0	3.0	3.0
Corrrelation between NT and T	0.76**			0.38			0.63*			0.77**			0.59*			0.40			0.36			0.23		

*$p<0.05$; **$p<0.01$.

Note: a median with H or Hx indication is significantly ($p < .05$) higher than a median with L or Lx indication in the same column whereby indices refers to the specific pair which was compared. For example, the median of Eagerness for Information in the Critical condition was 4.0^{H1} which was significantly higher than 1.0^{L1}, the median in the Extremely Bored condition.

the five attitude items with the exception of the impatience item which was hypothesized to get a higher score in the Extremely Bored condition (Table 6). The mood and personality conditions were compared with their opposite conditions, e.g., Positive Valence versus Negative Valence. To investigate these differences, Wilcoxon signed rank tests were conducted on each questionnaire item.

The extremely positive audience was perceived to be significantly more interested (z = -3.41, p = 0.001) in the talk, more positive (z = -2.25, p = 0.03) towards the talk, more eager to get information (z = -3.12, p = 0.002), and less impatient (z = -3.10, p = 0.002) than the extremely bored one. This finding completely matches with the parameter settings of the Extremely Positive and Extremely Bored conditions (Table 6). The Critical condition was perceived to be similar to the Extremely Positive condition, with one exception that the audience was significantly more critical (z = -2.18, p = 0.03) than the extremely bored audience while the extremely positive audience was not. This result also matches with the parameter settings of Critical condition, except for the Approval item. The Positive condition was also perceived to be similar to the Extremely Positive condition, but the positive audience was not significantly more positive or more eager to get information than the extremely bored audience was. This result is consistent with the parameter settings of the Positive condition except again for the Approval item.

To further investigate the attitude conditions, the items of Interest, Approval, Eagerness for Information, and Impatience in these conditions were compared with those in the Extremely Positive condition that were all set at high levels. The Criticism items in these conditions were compared with that in the Critical condition which was also set high. The bored audience was found to be significantly less eager (z = -2.13, p = 0.03) to get information than the extremely positive audience. This suggests that the Bored condition differentiated from the Neutral condition, which showed no significant difference from this high-level condition. Additionally, the extreme positive audience was found to be significantly less critical (z = -2.33, p = 0.02) than the critical audience.

For the mood and personality conditions, no significant difference was found to support the second and third hypotheses. Still, the audience with a positive valence parameter setting was perceived to be more aroused (z = -1.97, p = 0.049) than the audience with a negative valence parameter setting. Also the audience in Extremely Bored condition was rated as less extrovert than audience in the Critical (z = -2.54, p = 0.01), Extremely Positive (z = -2.28, p = 0.02), and Positive (z = -2.12, p = 0.03) conditions.

9 DISCUSSION AND CONCLUSIONS

This study has built an audience model that significantly predicts audience behavior using the agent parameters of mood, attitude, and personality. The audience model can generate expressive behavior by setting the model parameters. Both results of free-description and factor rating evaluation show that people can perceive variations in the attitude of the virtual audience that are caused by manipulation of the corresponding agent attitude parameters, which supports the first hypothesis. The results did not find similar matching between agent parameter manipulation and individual's perceptions of the virtual audience when it came to audience's mood (*H2*) and personality (*H3*). However, manipulation of the agent's valence parameter had an effect on perceived level of audience's arousal. This may be caused by a correlation between valence and arousal, suggested by many studies on the affective space, e.g., [38]. Furthermore, the virtual audience's expressiveness of moods might have been limited. Adding facial expressions might enhance this as various reports confirm that virtual characters express emotion better when using multi-modal expressions, e.g., [26]. Likewise, variation in audience's personality, in this case extraversion, was observed in the extremely bored, positive and critical attitude conditions. In other words, personality trait variation was only perceived in the attitude conditions that were more complex, i.e., created by multi-parameter manipulation. These conditions might enable more behavioral variations of the audience, which exhibited the personality traits. This is essential as the ability to express and perceive a person's individuality is situation dependent [39]. For example, someone's personality might be easier to assess when observed at a neighborhood party than as a soldier in a military parade. Furthermore, the expressiveness of the personality parameters could also have been constrained by the scope of the corpus. Although already extensive with 9600 coding units, the corpus was obtained by observing 16 individuals with personalities that did not cover the entire spectrum of personality trait combinations.

This study can be extended in many directions. First, control over the simulation environment can be added to easily construct different scenarios. For example, the classroom could be adjusted to a business meeting with fewer people sitting around a table or to a large podium with a larger audience, or the scenario can be changed from public speaking to musical performance. Second, the model could be extended by including social influence among individuals, as suggested by Poeschl and Doering [40]. Third, the functionality of the perception module in the model could be extended by adopting the Perception Markup Language (PML) [41] standard so that new perception technology by other researchers can be integrated. Fourth, the effect of changes in the settings of individuals' parameters on the output behavior and perceptions of this could be studied. Besides providing insight into which changes result in noticeable behavioral changes, this might also inform theories about audience behavior. Finally, potential operators (e.g., therapists) could be involved in the design of specific audiences to meet their needs.

Although the behavioral model for the autonomous agents was designed to create a virtual audience, a similar model might also be applied to VR systems that need autonomous virtual humans in other social situations. This could be for psychotherapy concerning disorders such as paranoia [42] and agoraphobia [43].

The evaluation presented in this paper focuses only on how people perceive the virtual audience behavior. This is an important validation step before claims can be made that a specific virtual audience setting (e.g., critical attitude) has a specific effect on users' emotional state in the future.

In conclusion, the main contributions of this work are as follows: (1) an audience model for public speaking simulation systems that generates expressive behavioral styles flexibly by adjusting agent parameters of mood, attitude, and personality, and (2) a corpus[1] of audience behavior showing different attitudes in public speaking situations and a coding scheme for posture observation. This audience model was built using a statistical approach based on observations of real audiences in public speaking situations. Using the parameters of attitude, mood, and personality as predictors, the audience model significantly predicted the audience behavior. This model was applied to an audience simulation, and the evaluation results showed that the virtual audience can behave expressively with regard to their attitude, and the behavioral styles can be controlled by modifying the model parameters. This is an important step towards providing users with a flexible and dynamic virtual environment in which they can be exposed to a virtual audience, for example, as part of a psychological stress test procedure, training, or psychotherapy.

Acknowledgment

This research is supported by the Netherlands Organization for Scientific Research (NWO), grant number 655.010.207, and China Scholarship Council (CSC), grant number 2010609042. Special thanks to Mark J. C. Hendrikx who carefully prepared and gave several talks for the audience observation.

References

[1] C. Zanbaka, A. Ulinski, P. Goolkasian, and L.F. Hodges, "Social Responses to Virtual Humans : Implications for Future Interface Design," in *Proc. SIGCHI Conf. Human Factors in Computing Syst*, 2007, pp. 1561–1570.

[2] M. Slater, D.-P. Pertaub, C. Barker, and D.M. Clark, "An Experimental Study on Fear of Public Speaking Using a Virtual Environment," *Cyberpsychology & Behavior : the Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, vol. 9, no. 5, pp. 627–33, Oct. 2006.

[3] J. Bissonnette, F. Dubé, and M.D. Provencher, "The Effect of Virtual Training on Music Performance Anxiety," in *Int. Symp. Performance Science*, 2011, pp. 585–590.

[4] M. Powers and P.M.G. Emmelkamp, "Virtual Reality Exposure Therapy for Anxiety Disorders: A meta-analysis," *J. Anxiety Disorders*, vol. 22, no. 3, pp. 561–569, 2008.

[5] O. Kelly, K. Matheson, A. Martinez, Z. Merali, and H. Anisman, "Psychosocial Stress Evoked by a Virtual Audience: Relation to Neuroendocrine Activity," *CyberPsychology & Behavior*, vol. 10, no. 5, pp. 655–62, Oct. 2007.

[6] C. Kirschbaum, K.-M. Pirke, and D.H. Hellhammer, "The 'Trier Social Stress Test'- a Tool for Investigating Psychobialogical Stress Responses in a Laboratory Setting," *Neuropsychobiology*, vol. 28, no. 1–2, pp. 76–81, 1993.

[7] S.E. Taylor, T.E. Seeman, N.I. Eisenberger, T.A. Kozanian, A.N. Moore, and W.G. Moons, "Effects of a Supportive or an Unsupportive Audience on Biological and Psychological Responses to Stress," *J. Personality and Social Psychology*, vol. 98, no. 1, pp. 47–56, Jan. 2010.

[8] S.F. Hofmann and M.W. Otto, *Cognitive Behavioral Therapy for Social Anxiety Disorder: Evidence-Based and Disorder-Specific Treatment Techniques*, 1st ed. Routledge, 2008.

[9] P.M.G. Emmelkamp, "Behavior Therapy with Adults," in *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*, 6th ed., no. 1, M.J. Lambert, Ed. Wiley, 2013.

[10] R.G. Heimberg and R.E. Becker, *Cognitive-Behavioral Group Therapy for Social Phobia: Basic Mechanisms and Clinical Strategies*. Guilford Publications, 2002, p. 334.

[11] M.J. Gallego, P.M.G. Emmelkamp, M. Van der Kooij, and H. Mees, "The Effects of a Dutch Version of an Internet-Based Treatment Program for Fear of Public Speaking: A Controlled Study," *Int. J. Clin. Health Psychol.*, vol. 11, no. 3, pp. 459–472, 2011.

[12] V. Vinayagamoorthy, A. Steed, and M. Slater, "Building Characters: Lessons Drawn from Virtual Environments," in *Proc. Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, pp. 119–126.

[1] The corpus will be available at http://ii.tudelft.nl/~nikang/

[13] C.A. Pollard and J.G. Henderson, "Four Types of Social Phobia in a Community Samply," *J. Nerv. Ment. Dis.*, vol. 176, no. 7, pp. 440–445, 1988.

[14] J. Lee and S.C. Marsella, "Predicting Speaker Head Nods and the Effects of Affective Information," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 552–562, 2010.

[15] L. Huang, L.-P. Morency, and J. Gratch, "Virtual Rapport 2.0," in *Proc. 10th Int. Conf. Intelligent Virtual Agents*, 2011, pp. 68–79.

[16] Z. Wang, J. Lee, and S. Marsella, "Towards More Comprehensive Listening Behavior: Beyond the Bobble Head," in *Proc. 10th Int. Conf. Intelligent Virtual Agents*, 2011, vol. 6895, pp. 216–227.

[17] C. Busso, Z. Deng, and M. Grimm, "Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1075–1086, 2007.

[18] J. Allwood, "Bodily Communication Dimensions of Expression and Content," in *Multimodality in Language and Speech Systems*, B. Granstrom, D. House, and I. Karlsson, Eds. Kluwer Academic Publishers, 2002, pp. 7–26.

[19] E. Bevacqua, E. Sevin, S.J. Hyniewska, and C. Pelachaud, "A Listener Model: Introducing Personality Traits," *J. Multimodal User Interfaces*, vol. 6, no. 1–2, pp. 27–38, Apr. 2012.

[20] S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. New Jersey: Prentice Hall, 2003.

[21] C. Qu, W.-P. Brinkman, Y. Ling, P. Wiggers, and I. Heynderickx, "Human Perception of a Conversational Virtual Human: an Empirical Study on the Effect of Emotion and Culture," *Virtual Reality*, online first, Aug. 2013.

[22] C.N. Moridis and A. A. Economides, "Affective Learning : Empathetic Agents with Emotional Facial and Tone of Voice Expressions," *IEEE Trans. Affective Computing*, vol. 3, no. 3, pp. 260–272, 2012.

[23] N. ter Heijden and W.-P. Brinkman, "Design and Evaluation of a Virtual Reality Exposure Therapy System with Automatic Free Speech Interaction," *J. CyberTherapy and Rehabilitation*, vol. 4, no. 1, pp. 41–55, 2011.

[24] 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 & Virtual Environments*, vol. 11, no. 1, pp. 68–78, 2002.

[25] W.A. Villaume and G.D. Bodie, "Discovering the Listener Within Us: The Impact of Trait-Like Personality Variables and Communicator Styles on Preferences for Listening Style," *Int. J. Listening*, vol. 21, no. 2, pp. 102–123, Aug. 2007.

[26] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," *IEEE Trans. Affective Computing*, vol. 4, no. 1, pp. 15–33, Jan. 2013.

[27] "Short form for the IPIP-NEO (International Personality Item Pool Representation of the NEO PI)." [Online]. Available: http://www.personal.psu.edu/j5j/IPIP/ipipneo120.htm.

[28] M. Bradley and P.J. Lang, "Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential," *J. Beh. Ther. & Exp. Psychiat.*, vol. 25, no. I, pp. 49–59, 1994.

[29] H.M. Rosenfeld, "Measurement of Body Motion and Orientation," in *Handbook of Methods in Nonverbal Behavior Research*, K.R. Scherer and P. Ekman, Eds. Cambridge: Cambridge University Press, 1982, pp. 199–286.

[30] C. Robson, "Coding Sequences of Behavior," in *Real World Research*, 2nd ed., Blackwell Publishing.

[31] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[32] K. Krippendorff, *Content Analysis: an Introduction to its Methodology*. Beverly Hills: Sage Publications, 1980.

[33] H. Grillon, F. Riquier, B. Herbelin, and D. Thalmann, "Virtual Reality as a Therapeutic Tool in the Confines of Social Anxiety Disorder Treatment," *Int. J. Disability and Human Development*, vol. 5, no. 3, pp. 243–250, Jan. 2006.

[34] D. Hartanto, N. Kang, W.-P. Brinkman, I.L. Kampmann, N. Morina, P.M.G. Emmelkamp, and M.A. Neerincx, "Automatic Mechanisms for Measuring Subjective Unit of Discomfort," in *Annu. Review of Cybertherapy and Telemedicine*, B. K. Wiederhold and G. Riva, Eds. Interactive Media Institute and IOS Press, 2012, pp. 192–196.

[35] A.J. Cowell and K.M. Stanney, "Manipulation of Non-Verbal Interaction Style and Demographic Embodiment to Increase Anthropomorphic Computer Character Credibility," *Int. J. Human-Comput. Stud.*, vol. 62, no. 2, pp. 281–306, Feb. 2005.

[36] L.K. Kammrath, D.R. Ames, and A.A. Scholer, "Keeping Up Impressions: Inferential Rules for Impression Change Across the Big Five," *J. Exp. Soc. Psychol.*, vol. 43, no. 3, pp. 450–457, May 2007.

[37] W.-P. Brinkman, D. Hartanto, N. Kang, D. de Vliegher, I.L. Kampmann, N. Morina, P.G.M. Emmelkamp, and M. Neerincx, "A Virtual Reality Dialogue System for the Treatment of Social Phobia," in *Proc. CHI 2012: extended abstracts on Human Factors in Computing Syst.*, pp. 1099–1102.

[38] P.J. Lang, M.M. Bradley, and B.N. Cuthbert, "International Affective Picture System (IAPS): Technical Manual and Affective Ratings," 1997.

[39] D. Pennington, *Essential Personality*. Arnold, 2003, p. 284.

[40] S. Poeschl and N. Doering, "Designing Virtual Audiences for Fear of Public Speaking Training - an Observation Study on Realistic Nonverbal Behavior," in *Annu. Review of Cybertherapy and Telemedicine*, B. K. Wiederhold and G. Riva, Eds. Interactive Media Institute and IOS Press, 2012, pp. 218–222.

[41] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A.S. Rizzo, and L. Morency, "Perception Markup Language : Towards a Standardized Representation of Perceived Nonverbal Behaviors," in *Proc. 12th Int. Conf. Intelligent Virtual Agents*, 2012, pp. 455–463.

[42] W.-P. Brinkman, W. Veling, E. Dorrestijn, G. Sandino, V. Vakili, and M. van der Gaag, "Using Virtual Reality to Study Paranoia in Individuals with and Without Psychosis," *J. CyberTherapy and Rehabilitation*, vol. 4, no. 2, pp. 249–251, 2011.

[43] K. Meyerbröker, N. Morina, G. Kerkhof, and P.M.G. Emmelkamp, "Virtual Reality Exposure Treatment of Agoraphobia: a Comparison of Computer Automatic Virtual Environment and Head-Mounted Display," *Stud. in Health Technology and Informatics*, vol. 16, pp. 51–56, 2011.

Ni Kang received a B.S. degree and a M.Sc. degree in instrumentation science and engineering from Southeast University, Nanjing, China, in 2007 and 2010. She is currently pursuing the Ph.D. degree in computer science at Delft University of Technology, Delft, The Netherlands. Her research interests include modeling expressive virtual characters and human-computer interaction.

Willem-Paul Brinkman received his PhD degree in 2003 at Eindhoven University of Technology, The Netherlands. Currently he is an assistant professor in the Intelligent Systems Department at Delft University of Technology. His main research interests lie in the area of human-computer interaction and mental health computing, such as virtual reality exposure therapy systems and virtual health coaching.

M. Birna van Riemsdijk received a Ph.D. degree in the Intelligent Systems group at Utrecht University, The Netherlands, in 2006. She is currently an assistant professor in the Department of Intelligent Systems, Delft University of Technology. Her research interests are techniques for intelligent support systems with a focus on the use and development of agent programming languages.

Mark A. Neerincx received a Ph.D. degree in psychology from University of Groningen, The Netherlands. He is currently a Senior Research scientist at TNO and a Professor at Delft University of Technology, The Netherlands. His research interests include cognitive engineering, electronic Partners, social robots, and cognitive taskload modeling for adaptive interfaces.