Perception of Virtual Audiences

Mathieu Chollet and Stefan Scherer
Institute for Creative Technologies, University of Southern California, Playa Vista, CA 90094 USA.
{mchollet, scherer}@ict.usc.edu.

Abstract—Virtual audiences have been found to be useful tools in the treatment of public speaking anxiety, and recent results also show they could be used for training public speaking skills for non-anxious individuals. However, relatively little research has investigated how virtual audiences are perceived. Understanding how virtual audiences are perceived depending on their non-verbal behavior is crucial to create relevant, controllable stimuli. In this paper, we present our virtual audience behavior model based on crowdsourced data, allowing us to inexpensively and rapidly author virtual audience behaviors. We used this model to create a collection of virtual audience videos where different states, such as boredom or engagement, are expressed through varying behaviors, such as postures, gazes, or facial expressions. We used this dataset to investigate the perception of virtual audiences, and present our results according to three perspectives of analysis: the recognition of overall audience states, the influence of behavioral parameters on virtual audience perception, and the recognition of individual audience members' states.

Keywords: Nonverbal behavior synthesis, Affective computing applications

I. INTRODUCTION

VIRTUAL characters modeling has progressed tremendously both in the rendering domain and in the behavioral domain. Those advances allow virtual characters to be used in an ever increasing number of applications. In particular, interpersonal skills training systems have been burgeoning recently with varied application domains such as job interview training, public speaking training, negotiation skills training, and many more [1]. Indeed, virtual characters' behaviors can be finely controlled and developers can make them assume a large variety of roles, allowing users to be exposed to standardized, repeatable simulations of interpersonal encounters.

A particular type of such a simulation is the virtual audience, a collection of virtual characters situated in an environment mimicking that of a public speaking situation. Virtual audiences were first proposed as a tool in the treatment of public speaking anxiety. Since then, a number of studies found that they were indeed effective in reducing public speaking anxiety: patients that underwent Cognitive Behavioral Therapy (CBT) including sessions of exposure to virtual audiences significantly reduced their anxiety levels, even a year after treatment [2]. In addition, no difference was found between these patients and another group undergoing regular CBT with group therapy instead of exposure to virtual audience; the patients exposed to virtual audiences exhibited superior adherence to treatment with much lower rates of attrition compared to regular CBT. Recently, we proposed to use virtual audiences not only for mitigating public speaking anxiety but also as a tool for training public speaking skills, regardless of one's anxiety in public speaking situations [3]. We found that participants rehearsing with virtual audiences improved their public speaking proficiency, as judged by experts and objective behavior annotations on a variety of aspects, *e.g.* avoidance of pause fillers, confidence, overall performance, *etc*.

Those very positive results demonstrate that virtual audiences are indeed a very valuable tool. However, the impact of its members' behaviors on how a real or a virtual audience is perceived is a topic that has so far received very little attention [4]. Indeed, while many systems using virtual audiences have been proposed, many either used static audiences, or manually controlled audiences. However, the behavior of virtual character influences critically the kind of stimuli that it produces. This was shown for virtual audiences in an early study by Pertaub *et al.* [5], where positive, neutral and negative audiences were compared, and the negative audience provoked higher anxiety responses than the other two, regardless of the normal level

of confidence of the subject. It is therefore clear that the behaviors of virtual audiences affect the experience of the user, and understanding this effect seems crucial to be able to adapt the experience of an interaction with a virtual audience to the needs of users. Two recent studies by Kang *et al.* give us some insight into how audiences are perceived [6], they recorded real audiences primed to display certain states, *e.g.* bored by having a speaker recite a speech from Aristotle with the paragraphs randomly shuffled. They coded the audiences' behaviors every 2 seconds and used this data to create a model for selecting full body postures of virtual audience characters. After that, they investigated which behavior types and states are recognized by participants observing a virtual audience. They found that two main dimensions were recognized by participants, arousal and valence. Those two dimensions constitute a common framework in psychology, notably in the domain of emotion and affect; arousal can here be seen as the level of alertness of an audience member, and valence corresponds to how positively or negatively they feel towards the speaker or the presentation. While this work brings many valuable insights into the perception of virtual audience, one important limitation is that their coding strategy did not allow to study gaze patterns, head nods and head shakes, or facial expressions. Additionally, it is difficult to analyze the role of the different non-verbal modalities from their results, and how audiences consisting of virtual agents with mixed states (*e.g.* an audience with 2 engaged characters and 3 bored characters) are perceived.

In this article, we extend our previous work on how non-verbal behaviors influence the way a virtual audience is perceived [7]. We started by collecting data on virtual audience behaviors through a crowdsourcing study. Using this data, we built a virtual audience behavior model for the expression of various levels of arousal and valence. We then used this model to generate a collection of virtual audience videos for expressing different states with varying behaviors. This dataset was used in [7] to validate our virtual audience model. The novelty of this contribution is twofold. First, we studied the relative contribution of different audience non-verbal behaviors on the overall perception of an audience; this is presented in Section III.B. Second, we investigated how subjects perceive individual audience members, and how this perception is influenced by the behavior of the rest of the audience; these results are reported in Section III.C.

II. CROWDSOURCING VIRTUAL AUDIENCE BEHAVIORS

Multiple methodologies are available when designing a behavior model for virtual characters. One is to refer to the available literature and to handcraft a model based on existing results, however in the case of audience behavior, it appears that there is a significant lack of research on the topic [4]. A second approach consists in recording a corpus representative of the considered situation. This corpus is then annotated manually or processed automatically for occurrences of multimodal behaviors (e.g. postures, facial expressions). Statistical analysis can then be realized to find interesting patterns which will be encoded in a model: that is the approach chosen in [4] and [6]. A third method, which was recently proposed by Ochs *et al.* [8], consists in using crowdsourcing to let users design the relevant behaviors. For instance, Ochs et al. studied the differences between embarrassed, polite and amused smiles: on a web interface, users were asked to create one specific type of smile. To this end, they could manipulate different parameters of a virtual character's smile which was displayed on the interface: smile symmetry, size, duration, *etc*. We adopted this last methodology for collecting data on virtual audience behaviors. The task of the users was to select behaviors so that an audience member appears in a particular state of arousal and valence. To this end, we created the web interface shown in Figure 1. This interface consisted 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 "satisfaction" scale to indicate how well the participant thinks the resulting video conveys the input condition.

We reviewed the literature on bodily communication in order to identify relevant communicative non-verbal signals [8-10]. Even though, these works did not investigate audiences but single individuals, we assumed that signals relevant while listening in conversations would be relevant as well in an audience context. We defined 7 different parameters that could be chosen by the users:

- Posture: 6 different poses (descriptions below)
- Amount of averted gaze: 0%, 25%, 50%, 75%, 100%
- Direction of averted gaze (if applicable): Sideways, Down, Up.

- Facial expressions, 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 per 10 seconds interval.

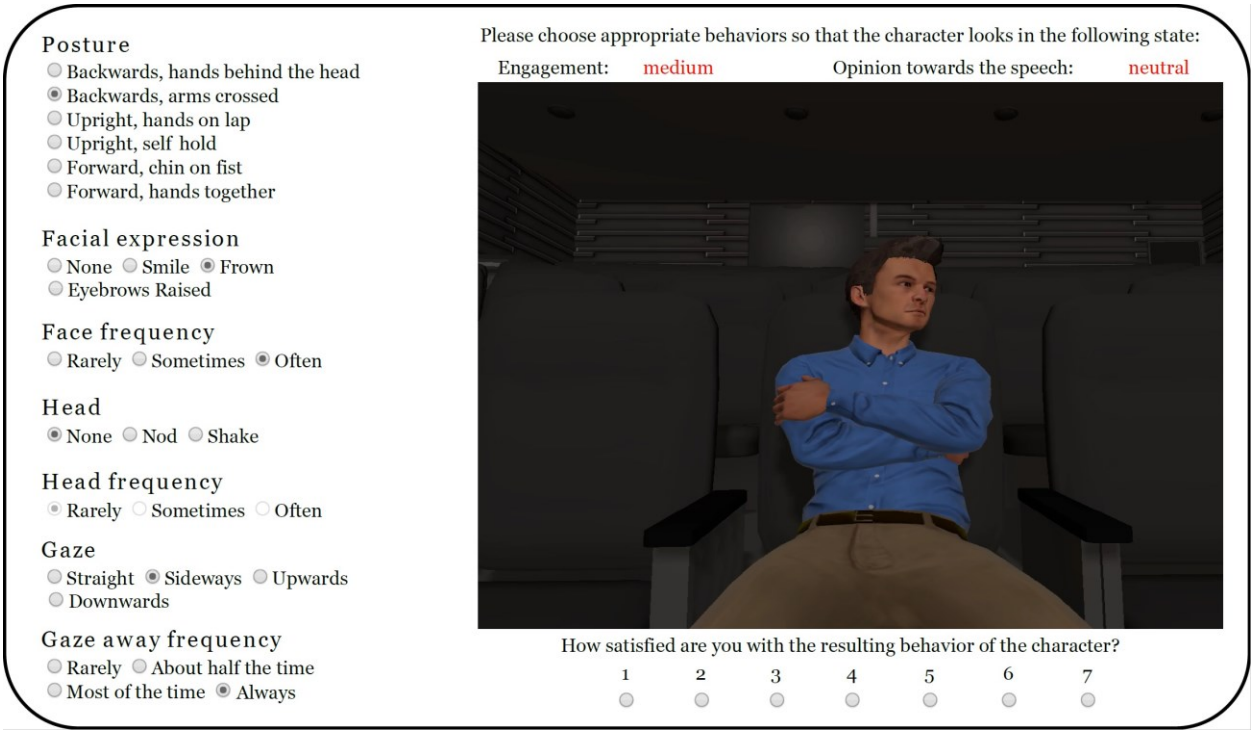

Fig. 1: Screenshot of the crowdsourcing interface.

We chose to limit the quantity of available postures to a subset which would allow us to study different aspects of postural behavior. We chose to vary the postures according to parameters identified in previous work as relevant for communicating various attitudes [9]: proximity to the speaker and relaxation. We chose the 6 following postures:

- Lean back with hands behind the head
- Lean back with arms crossed
- Upright with hands on legs
- Upright with one hand supporting the chin
- Leaning sideways with right hand holding the left arm
- Lean forward with hands joined

We created 20s videos corresponding to all of the possible different combinations of the behavior parameters for a male and a female character, resulting in 10920 videos. Example videos can be found in the following links: https://youtu.be/vRO4kPpJ5KM and https://youtu.be/dTE14IPBO98. Some checks 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 changing) and to introduce some variability in behavior timings. We defined five values for both the arousal (resp. valence) states that users would need to create behaviors for: “very low”, “low”, ”medium”, “high”, “very high” (resp. negative/positive).

A. Data Analysis

We recruited 72 participants using the Amazon Mechanical Turk website (https://www.mturk.com) 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 which behaviors are relevant for audience members to express arousal and valence, we tested hypotheses about how behaviors were chosen by participants. We defined these hypotheses following findings in the literature on non-verbal listening behavior [9-11].

H1 Arousal and expressions - Higher arousal leads to more feedback: *i.e.* more facial expressions, more head movements, and more gaze directed at the speaker [10-11].
H2 Valence and expressions - Smiles, nods are associated with positive valence; frowns, head shakes with negative valence; eyebrow raises are mostly neutral [11].
H3 Arousal and postures - Postures chosen for high arousal involve leaning closer to the speaker than postures chosen for lower arousal [10].
H4 Valence and postures - Relaxed postures lead to higher valence compared to more closed postures [9].

The distributions of chosen behaviors per valence and arousal states regarding these hypotheses are displayed in Figure 2. We conducted statistical tests to ensure 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 (backwards: 1 to forwards: 3) and relaxation (arms crossed and self-hold: 1, arms behind the head: 3, the rest: 2).

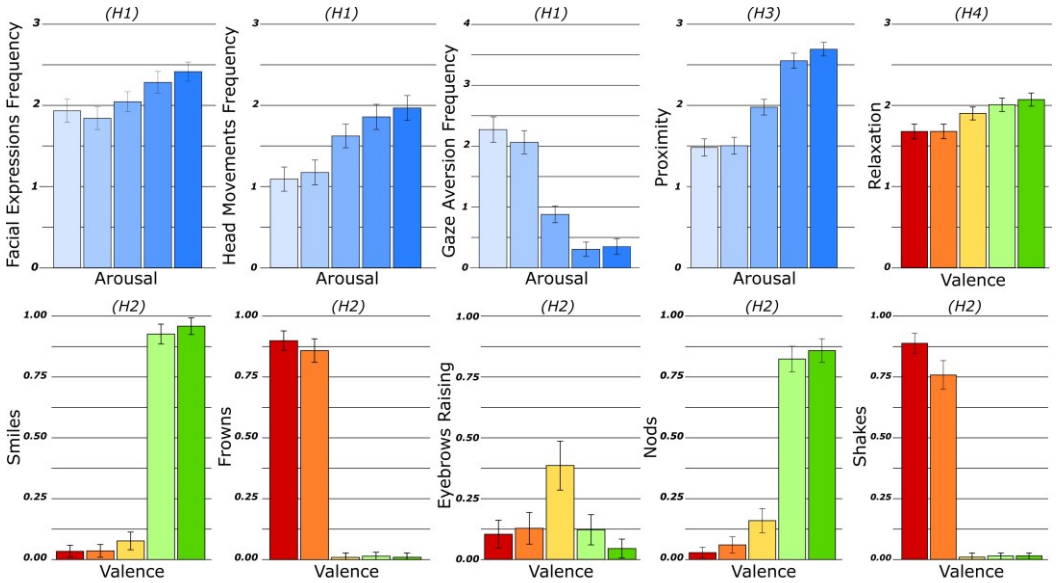

Fig. 2: Distribution of behaviors per state levels for the investigated hypotheses. From left to right in the subfigures, very low to very high arousal (resp. valence).

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 relaxation 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 provide support for 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 (*i.e.* smiles and nods for positive valence, frowns and shakes for negative valence) and leads to less relaxed postures.

Using the collected data, we proceeded to build a model for generating non-verbal behavior for members of a virtual audience. We defined a probabilistic model which models the relationship between two input states, arousal and valence, and the 7 output behavioral parameters defined in the previous section. In effect, it models the $P(Behavior|Arousal, Valence)$ behavior distributions for the different modalities and states used in the crowdsourcing study. The model parameters were learned using the crowdsourced data, using the participants' satisfaction scores as weights, therefore assigning more importance to behavior combinations that were deemed by participants as being better representations of the corresponding state.

This model can be queried in the following way. First, the manipulated character's arousal and valence values are set. Then, one random number is sampled from the uniform distribution in the $[0; 1)$ range for every behavioral parameter. Those numbers are then 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 among characters.

III. Perception of Virtual Audiences

With our virtual audience behavior model completed, we designed a study in order to investigate various aspects of the perception of virtual audiences. The first goal of this study was to validate the correct expression of audience states by our model. Secondly, we wanted to quantify the importance of different parameters of virtual audience behavior on how they are perceived. Finally, our third goal was to examine the identification of individual audience members' states.

For this study, we generated a dataset of audience videos in which we systematically varied the states of a number of audience characters using our model. We varied two independent variables: a target state S, consisting of a value of valence and arousal, and the number of manipulated characters N which are configured to display that state S. 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 ranged 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 adopt congruent, neutral or contradictory states compared to the input condition. We created 4 video variants for every condition, for a total of 396 videos.

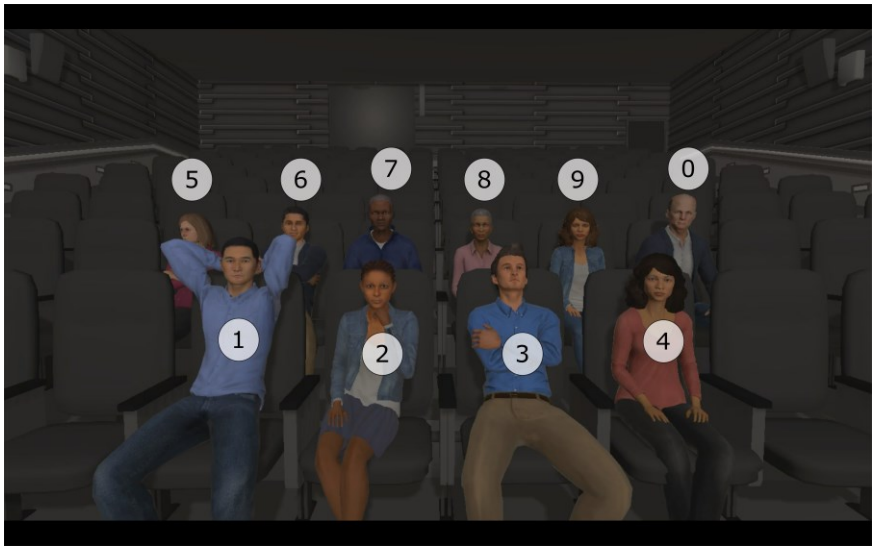

Fig. 3: Screenshot of the full audience. The numbers were not shown in the study videos, but are used for describing audience parts in section III-C1.

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. Additionally, participants were asked which characters displayed a particular state, *e.g.* "*Which characters appear particularly engaged?*". A grid was displayed with images of the 10 virtual characters along with checkboxes that participants could check whether that virtual character displayed that particular state. 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.

A. Recognition of Virtual Audience States

The first analysis we conduct is on how an audience is perceived overall, depending on the amount of characters that display a particular state. In a nutshell, this is an analysis on how our virtual audience non-verbal behavior model performs. The results, averaged over all input videos, are presented in Figure 4.

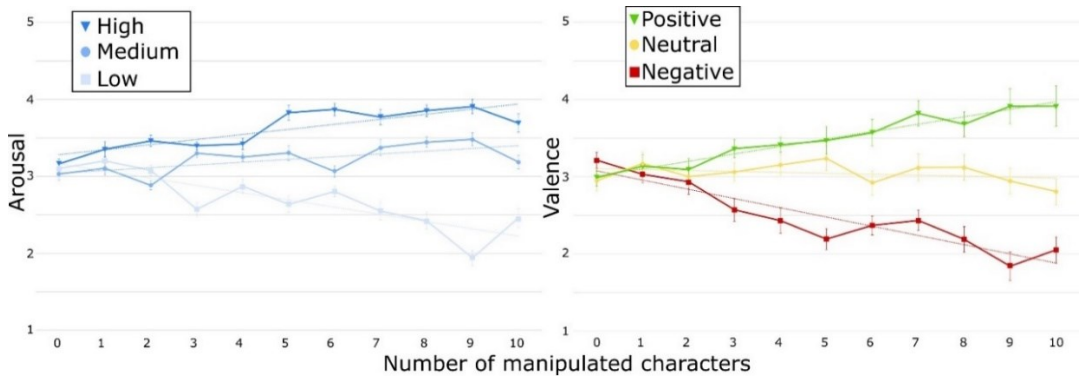

Fig. 4: Perception of arousal (resp. valence) for audiences of 10 characters, depending on the target state and the number of manipulated characters. In dotted lines are shown the trends for the average scores of audiences.

We can observe 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, provided that a sufficient number of audience members are configured to express that states. We can see that the three states start to be clearly distinguishable from 4 manipulated characters in the case of arousal, 3 in the case of negative valence and 6 for positive valence. We conducted a linear regression analysis in order to further analyze the impact of individual characters' states on the overall perception of an audience. Specifically, we studied a regression model of the following form: $y = \alpha + \beta_{Low} * N_{Low} + \beta_{High} * N_{High}$, with y corresponding to the participants' rating of the arousal (resp. valence) of the video's audience, N_{Low} (resp. N_{High}) the number of audience members displaying low arousal (resp. low valence), and β_{Low} (resp. β_{High}) the corresponding regression coefficient. The results of our linear regression analysis are the following (note that the dotted lines shown in Figure 4 are not representations of these regression equations; those are the trends of the audience average ratings):

$$Arousal = 3.30 + 0.067 * N_{High} - 0.109 * N_{Low}$$
$$(F(2;\ 393) = 111.6;\ p < 0.001: R^2 = 0.36;\ StdErr = 0.57)$$
$$Valence = 3.09 + 0.09289 - 0.116 * N_{Negative}$$
$$(F(2;\ 393) = 209.5;\ p < 0.001: R^2 = 0.51;\ StdErr = 0.48)$$

Other regression analyses were performed to investigate the possible role of interaction terms between variables (e.g. interaction between the number of low vs. high arousal characters), but those terms were found to be not significant.

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 means 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. This is an interesting finding for interactive virtual audiences, allowing us to understand how the overall perception of a virtual audience changes when some of its members change states. 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.

B. Behavioral Variables Influencing Virtual Audience Perception

In the previous section, we validated that we can manipulate how a virtual audience is perceived by modifying the states of individual audience members and using our probabilistic non-verbal behavior selection model to choose appropriate behaviors according to arousal and valence inputs, whilst previous work only succeeded in expressing arousal [6]. We now want to study whether we can predict the engagement and arousal scores of a virtual audience from the non-verbal signals they display. This will also allow us to investigate the importance of different non-verbal signal types on overall audience perception. To that end, we re-use the same video ratings as in the previous section, this time parsing the behavior logs to retrieve the data of audience behaviors for each video and comparing those to user ratings.

In a first analysis, we created one feature for each type of non-verbal signal, consisting of the number of characters that display that non-verbal signal in the considered video. For instance, if 7 characters perform a head nod at some point in a video v, then we would define the following feature $X_{Nod}(v) = 7$. We then realized a linear regression using these features to predict the MTurk participants' arousal and valence ratings for each video. The results are presented in Table I.

	Arousal. $R^2 = 0.186$ $F(14; 2628) = 44.12, p < 0.001$	
	Coefficient	*p*-value
(Intercept)	3.942	**0.079**
Lean back, arms crossed	-0.123	0.586
Lean back, arms behind the head	-0.266	0.242
Upright, hand on chin	-0.067	0.766
Lean forward, hands joined	0.001	0.997
Upright, hands on lap	-0.040	0.860
Lean sideways, self-hold	-0.090	0.692
Gaze Up	-0.092	**<0.001**
Gaze Down	-0.083	**0.001**
Gaze Sideways	-0.045	**0.005**
Nod	0.073	**<0.001**
Shake	0.058	**0.046**
Smile	0.023	**0.0252**
Frown	0.057	0.052
Eyebrows Up	0.030	0.205

	Valence. $R^2 = 0.275$ $F(14; 2628) = 72.662, p < 0.001$	
	Coefficient	*p*-value
(Intercept)	5.100	**0.007**
Lean back, arms crossed	-0.208	0.275
Lean back, arms behind the head	-0.315	0.100
Upright, hand on chin	-0.185	0.331
Lean forward, hands joined	-0.180	0.341
Upright, hands on lap	-0.201	0.289
Lean sideways, self-hold	-0.204	0.285
Gaze Up	-0.003	0.889
Gaze Down	-0.034	0.107
Gaze Sideways	-0.021	0.114
Nod	0.119	**<0.001**
Shake	-0.152	**<0.001**
Smile	0.005	0.759
Frown	0.022	0.367
Eyebrows Up	0.001	0.944

TABLE I: Linear regression results for predicting audience state scores from the amount of characters displaying each type of non-verbal behavior.

From this first experiment we can already draw some conclusions. For arousal, we see that the more characters are looking away the more disengaged the audience looks, and that the more head movements are produced (regardless whether they are nods or shakes) the more engaged the audience looks. For

valence, we also see a strong influence of head movements, with nods making the audience appear more positive and shakes appear more negative.

However, we found surprising that postures and facial expressions seemed to have no effect on the perception of non-verbal behavior. We thus conducted a second analysis where we tried to reduce the amount of features. We engineered a new set of features that encapsulate important dimensions of the audience's non-verbal behaviors. We reduced the postural features to the Relaxation and Proximity dimensions introduced in Section II-A, averaged over all 10 audience members and then normalized to be between 0 and 1. We joined the gaze behaviors (Up, Down and Sideways) in one single feature measuring the average proportion of time that audience members look away, also normalized between 0 and 1. We keep one feature for each type of facial expression and head movement as they showed significant importance in the previous experiment and as they differ largely in meaning from one another. However, we do not simply count the amount of characters that display these signals, but the average amount of signals displayed per second over the whole audience. We perform a new linear regression analysis, whose results are presented in Table II. We observe this time that postures and facial expressions indeed have a significant effect on audience perception.

	Arousal. $R^2 = 0.183$ $F(8; 2628) = 44.12, P < 0.001$	
	Coefficient	*p*-value
(Intercept)	2.798	**<0.001**
Average Proximity	1.006	**<0.001**
Average Relaxation	-0.636	**0.002**
Smiles per second	0.312	0.056
Frowns per second	0.436	**0.030**
Eyebrow raises per second	0.679	**0.002**
Nods per second	0.539	**0.004**
Shakes per second	-0.399	0.064
Average time looking away	-1.774	**<0.001**

	Valence. $R^2 = 0.273$ $F(8; 2628) = 72.66, P < 0.001$	
	Coefficient	*p*-value
(Intercept)	3.005	**<0.001**
Average Proximity	0.435	**<0.001**
Average Relaxation	-0.586	**<0.001**
Smiles per second	0.212	0.123
Frowns per second	-0.127	0.454
Eyebrow raises per second	0.202	0.262
Nods per second	0.919	**<0.001**
Shakes per second	-0.944	**<0.001**
Average time looking away	-0.604	**<0.001**

TABLE II: Linear regression results to predict arousal and valence scores from engineered behavioral features.

For arousal, the strongest non-verbal signals seem to be gaze, where more gaze directed to the speaker leads to higher arousal, and postures, where more distant and relaxed postures signal low arousal compared to forward leaning or tight, self-holding postures. Head nods and eyebrow movements also signal arousal. In comparison, smiles and head shakes seem to be the least important signals, although they still contribute to the overall audience impression as their corresponding *p*-values approach significance.

For valence, we observe that head movements are the most important signals, with nods being positive and shakes negative. Postures and gaze also contribute significantly to valence, in the same manner as they do for arousal. In comparison, facial expressions are less important. Whilst facial expressions are strong signals for signaling positive or negative valence in other situations (*e.g.* facial expressions of emotion in face-to-face interactions), it seems that they may be too subtle in virtual audience settings.

These results give us a better picture of the different contributions of various aspects of virtual audiences' non-verbal behaviors. In addition, those models can be used to compute a prediction of the expressed state of a virtual audience based on the behavior it displays, without having to know the actual underlying states of its members.

C. Perception of Individual Characters States

Finally, we analyze the participant answers to the questions related to identifying individual audience member states. In this study, a question was randomly chosen for each video out of 6 possible questions,

corresponding to the three possible values of arousal and valence: negative, neutral, positive. For instance, for positive valence, the question was "*Which characters had a positive opinion of the speech?*", while for negative arousal, the question was "*Which characters looked bored/uninterested?*". We first present our results in Table III. We can observe that identification rates are quite poor, with F-measures averaging at 0.35. Those results seem indeed to indicate that the task of identifying individual audience members was quite difficult.

We wanted to investigate further if we could identify factors that affected participants' recognition rates. First, we look at the influence of character placement on the correct identification of their state. Then, we investigate if the identification of individual audience members will be influenced by how the other audience members behave. We analyze this in two ways: first, we investigate how the level of congruence of an audience towards a question (*e.g.* if the audience looks positive and the question asks which characters express positivity, then we label the audience as congruent) influences correct identification rates of individual characters. For instance, if 8 characters out of 10 are expressing positive states, the audience will be perceived as globally positive, and the 8 congruent characters will be more likely to be identified as positive because of the large proportion of congruent characters. Our second analysis perspective is to look at how much an individual character's behavior contrasts with the rest of the audience (*e.g.* if one character is the only one to display negative behaviors while the rest of the audience shows positive behaviors, then there is a high contrast).

	Arousal			Valence		
	Negative	Medium	Positive	Negative	Medium	Positive
True positives	457	465	519	377	503	449
False positives	893	665	942	637	674	649
False negatives	911	1197	856	913	1181	956
True negatives	2399	1923	2153	2283	2032	2396
Accuracy	0.61	0.56	0.60	0.63	0.58	0.64
Precision	0.34	0.41	0.36	0.37	0.43	0.41
Recall	0.33	0.28	0.38	0.29	0.30	0.32
F-measure	0.34	0.33	0.37	0.33	0.35	0.36

TABLE III: Results of individual identification study

1) Impact of character position on recognition: One aspect we investigate that could influence the correct identification of individual audience members' states is their placement in the virtual environment. We hypothesize that characters located in the center and in the front row will be more correctly identified than characters on the side and in the back row. To analyze this effect, we compare the distribution of correct participants' answers to individual questions for different sections of the audience using Student t-tests. For presentation purposes, we numbered the audience members on Figure 3, and refer to those number to describe different sections. We compared the following audience sections:

- Front row (1 through 4) vs back row (5 through 10): $p = 0.579$
- Front row center (2 and 3) vs front row sides (1 and 4): $p = 0.697$
- Back row center (7 and 8) vs back row sides (5 and 0): $p = 0.643$

As it can be observed, we do not find any effect on the placement of a character on its rate of identification. Whilst front row characters occupy a larger screen space than back row characters, this did not improve significantly how well their states were recognized. Still, we only laid out audience members on two rows in our experiment, and it could be that on larger audiences, the effect of a character's distance to the camera plays a role.

2) Congruent vs incongruent characters: We then look at how each question's identification rate of individual character states vary according to the rest of the audience. We differentiate between two cases: whether a character's state is congruent or incongruent to the asked question. For instance, if the question is "*Which character look interested or engaged?*", characters with high arousal are considered congruent and

characters with low or medium arousal are considered as incongruent to that question. What we observe is that correct identification rates are influenced by how much the rest of the audience hold congruent states, with the exception of medium arousal states. For instance, when asking *"Which characters look negative?"*, then if many characters are indeed displaying negative states, then the other neutral and positive characters are most likely to be wrongly identified as negative.

Results are shown in Figure 5. With the exception of the medium arousal question (*i.e. "Which characters looked neither particularly bored nor particularly engaged?"*) where there doesn't seem to be an influence of the level of audience congruence on individual identification rates, we see that there is a clear effect on identification rates. Additionally, this effect seems to be stronger for valence than for arousal. Characters that display a congruent state to the asked question are more likely to be correctly identified if other characters also show congruent states. Respectively, if a character displays a state incongruent to the question, then it is less likely to be correctly identified as more characters hold states that are congruent to the question.

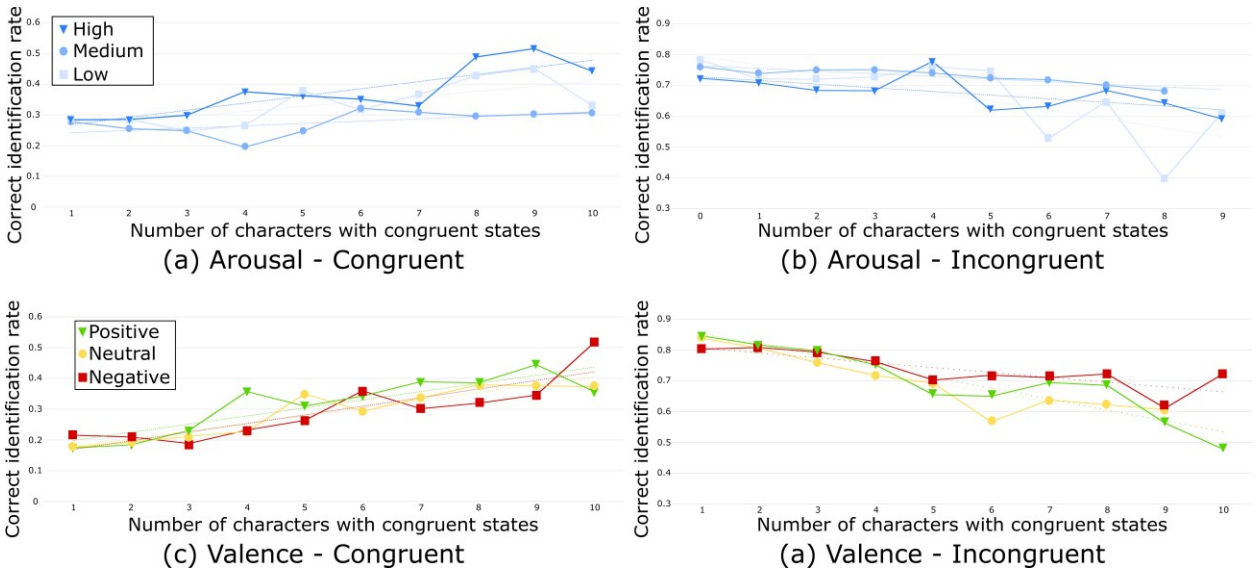

Fig. 5: Correct identification rates of audience members for arousal (top) and valence (bottom) questions, with respect to the total amount of audience members expressing a state congruent to the question. In the left figures are plotted correct identification rates for congruent characters. The right figures correspond to incongruent characters, *i.e.* characters whose states do not correspond to the question.

3) Audience contrast: We then investigate how recognition rates of individual characters are affected by how many other audience members contrast with them, regardless of the question. If the contrast is equal to 0, it means that all the other characters show a similar state. Respectively, if it is equal to 9, it means that all the other audience members are showing a different state. Our hypothesis is that the higher the contrast is, the lower the correct identification rates will be. Results are shown in Figure 6. We indeed see a clear trend of reduced correct identification rates as the number of contrasting characters increase.

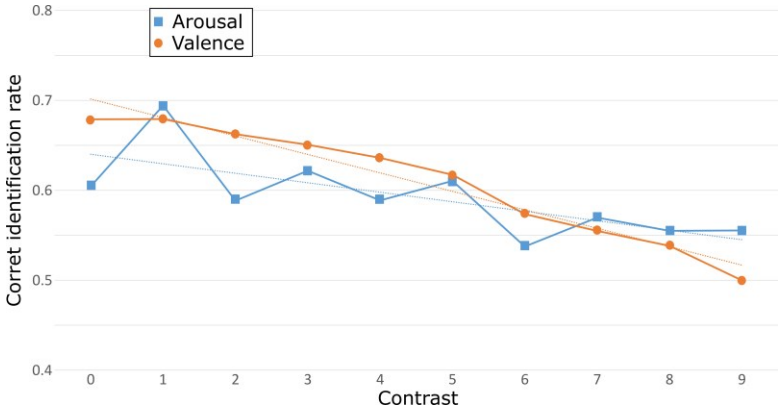

Fig. 6: Individual characters recognition rates for arousal and valence states depending on how many other audience members display contrasting states.

IV. CONCLUSION

A growing body of evidence shows that virtual audiences are a valuable tool in the treatment of social anxiety, and recent works show that it could also be used for training non-anxious individuals. However, there has been little research on how such audiences are perceived and on how the behavior of virtual audiences can be manipulated to create various types of stimuli. In this article, we used a crowdsourcing methodology for creating a virtual audience non-verbal behavior model, and then used that model to create a dataset of videos with virtual audiences containing varying behaviors. We then conducted a study using this dataset in order to find out how virtual audiences are perceived and which factors affect this perception. To this end, we adopted three different perspectives.

First, we investigated the overall impressions that a virtual audience creates when different proportions of its members express a particular state, *e.g.* half of the audience displays positive behaviors. What we observed is that the overall perceived state of an audience seems to be proportionately related to the individual states of its constituents. This means that the type of stimuli that is created by a virtual audience can be precisely controlled by adjusting the states of individual characters.

Then, we studied how particular types of non-verbal signals affect the perception of a virtual audience. We found that a reduced state of behavioral features can be used to predict the overall rating of a virtual audience. Using this result, a score for the level of arousal and valence can be computed for any virtual audience. An interesting result was that facial expressions, which can be used to display strong signals such as emotions, seem to be of less importance than other signals for the expression of valence in a virtual audience setting. It could be that facial signals are too subtle and that in such an environment, bodily signals are simply more salient because they are easier to perceive.

Finally, we investigated how individual characters are perceived. While participants did not seem to have trouble recognizing the overall state of an audience, we found that individual identification rates were much lower. Out of the possible explanations for such a phenomenon, we found that distance was not a factor while the behavior of the rest of the audience played an important role. Additionally, the state of other audience members was observed to strongly affect how well a particular individual character's state is identified.

In future work, we will investigate further how to leverage these effects to create relevant virtual audience stimuli. In particular, we intend to realize an eye-tracking study in order to study which behaviors are most salient, especially in larger audiences. We also plan to integrate our virtual audience system with techniques for automatically assessing public speaking behavior, such as [12]. The audience will then be configured to display stimuli corresponding to the public speaker's performance, allowing users to receive immediate feedback on their performance. Alternatively, the audience will react to the user's level of stress and provide positive or negative stimuli to maintain the user in a challenging, non-threatening state which will promote learning.

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] M. E. Hoque and R. W. Picard, "Rich Nonverbal Sensing Technology for Automated Social Skills Training," in *Computer*, vol. 47, no. 4, pp. 28-35, 2014.

[2] B. K. Wiederhold and S. Bouchard (2014). "Social anxiety disorder: Efficacy and virtual humans," in Advances in Virtual Reality and Anxiety Disorders (pp. 187–209). Springer.

[3] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro and S. Scherer (2015). Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In Proceedings of UbiComp (pp. 1143-1154). ACM.

[4] S. Poeschl and N. Doering (2011). Designing virtual audiences for fear of public speaking training-an observation study on realistic non-verbal behavior. In Annual Review of Cybertherapy and Telemedicine: Advanced Technologies in the Behavioral, Social and Neurosciences, ser. Studies in health technology and informatics, vol. 181, pp. 218–222.

[5] D.-P. Pertaub, M. Slater and C. Barker (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. Presence: Teleoperators and virtual environments, vol. 11, no. 1, pp. 68–78.

[6] N. Kang, W.-P. Brinkman, M. B. van Riemsdijk and M. Neerincx (2016). The design of virtual audiences: Noticeable and recognizable behavioral styles. Computers in Human Behavior, vol. 55, pp. 680–694.

[7] Chollet, M., Chandrashekhar, N., Shapiro, A., Morency, L. P., & Scherer, S. (2016). Manipulating the Perception of Virtual Audiences Using Crowdsourced Behaviors. In *International Conference on Intelligent Virtual Agents* (pp. 164-174). Springer International Publishing.

[8] M. Ochs, B. Ravenet and C. Pelachaud (2013). A crowdsourcing toolbox for a user-perception based design of social virtual actors. In Computers are Social Actors Workshop (CASA).

[9] A. Mehrabian, Non-verbal communication. Transaction Publishers, 1977.

[10] M. Argyle and R. Ingham (1972). Gaze, mutual gaze, and proximity. Semiotica, vol. 6, no. 1, pp. 32–49.

[11] M. L. Knapp, J. A. Hall and T. G. Horgan, Non-verbal communication in human interaction. Cengage Learning, 2013.

[12] Wörtwein, T., Chollet, M., Schauerte, B., Morency, L. P., Stiefelhagen, R. and Scherer, S. (2015). Multimodal public speaking performance assessment. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 43-50). ACM.

AUTHOR BIOGRAPHIES

Mathieu Chollet (*mchollet@ict.usc.edu*) is a post-doctoral researcher at the USC Institute for Creative Technologies. He received his PhD in 2015 from Telecom Paristech. His current research focuses on multimodal interaction techniques enabling social skills training, such as social signal processing to assess users' interpersonal skills, and building behavior models for embodied agents to simulate relevant social situations.

Stefan Scherer (scherer@ict.usc.edu) is a Research Assistant Professor and Associate Director of Neural Information Processing at the University of Southern California. He received the degree of Dr.rer.nat. from Ulm University in Germany. His research aims to automatically identify, model, and synthesize individuals' multimodal nonverbal behavior. He was awarded a number of best paper awards in renowned international conferences.