Perception of Virtual Audiences: Influence of Gender and Nonverbal Behavior

Sarah Saufnay
QuantOM, HEC Management School, University of Liège
Liège, Belgium
sarah.saufnay@uliege.be

Michaël Schyns
QuantOM, HEC Management School, University of Liège
Liège, Belgium
m.schyns@uliege.be

ABSTRACT

Virtual Reality (VR) has demonstrated its potential as a training and education tool through its immersive capabilities. Combined with Intelligent Virtual Agents (IVAs), applications for acquiring public speaking skills have notably been developed to replicate distressing social situations, yet crucial in one's life. The training experience can be tailored to the users' needs by manipulating the IVAs' behavior. However, building challenging and immersive experiences requires in-depth knowledge of how virtual agents are perceived. In this context, the present article investigates the perception of medium-sized virtual audiences' valence and arousal in VR. A total of 35 audience conditions were considered, varying in IVAs' gender and nonverbal behavior, and were evaluated by 70 participants to assess the influence of these characteristics on user perception. Specifically, 7 attitude types were designed, assigning distinctive behaviors to each IVA to convey the desired levels of valence and arousal. Additionally, 5 variations in audience gender composition were included. Overall, this study provides valuable guidelines for designing virtual audiences by identifying validated attitudes associated with specific valence and arousal levels. Notably, 4 out of the 7 designed attitudes successfully elicited the intended perception ratings. The misperception of some attitudes reflects the complexity of designing virtual audiences using a priori characteristics drawn from the literature. Importantly, the attitude types remained valid regardless of the audience gender composition, as no significant influence of gender on perception was observed.

CCS CONCEPTS

• **Human-centered computing → User studies**.

KEYWORDS

Virtual audience, Perception, Nonverbal behavior, Gender

ACM Reference Format:
Sarah Saufnay and Michaël Schyns. 2025. Perception of Virtual Audiences: Influence of Gender and Nonverbal Behavior. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3717511.3747085

1 INTRODUCTION

In a multitude of domains, Intelligent Virtual Agents (IVAs) have been leveraged to substitute human presence or enhance experiences. Depending on their assigned objectives, IVAs can indeed assume a plethora of roles, such as virtual hotel and shopping assistants [3, 5], or public speaking coaches [16]. Their behavior and appearance can be tailored to suit the context, resulting in a wide range of potential applications. However, the design of IVAs should be carefully considered as their characteristics are likely to influence user perception and acceptance of the system [10, 23, 26, 28].

In particular, public speaking applications have greatly benefited from the development of IVAs to simulate human presence. Whether relying on PC [2, 7, 13, 25] or Virtual Reality (VR) technology [11, 12, 17, 21, 24], training applications have been developed to enable users to practice their communication skills in front of a virtual audience, which can vary in terms of size and behavior to challenge the user. This has motivated the design of IVAs capable of adjusting their attitude toward the speaker depending on the training objectives and the user's specific needs. Consequently, the system may integrate a supportive audience to build user confidence or, conversely, a critical one to help them acquire stress management skills [13, 14, 17, 22]. Additionally, enabling the audience to react to the speaker's performance is worth considering to provide highly realistic simulations that offer immediate performance feedback to users through the IVAs' behavior [7].

To achieve such results and guide the design of IVAs, perceptual studies have been conducted, investigating the influence of their characteristics on both users' emotions and perception. However, most of these studies rely on 2D-based systems [10]. The use of VR in perceptual studies remains relatively limited, although VR has great potential as a training tool, fostering both skills acquisition and user motivation through its immersive capabilities [20]. This highlights the need to deepen knowledge of how IVAs are perceived in VR, particularly in public speaking settings, such that agents can adopt clear and interpretable attitudes toward the speaker. More precisely, the perception of virtual audiences (*i.e.,* groups of IVAs) deserves particular attention, since speaking in front of a group is common in numerous public speaking situations, but has not been sufficiently explored in the literature.

This study addresses these gaps by examining how medium-sized virtual audiences are perceived in VR, focusing on the influence of gender-related cues and the nonverbal behavior of IVAs. Gender can indeed bias the interpretation of IVAs' attitudes [1, 25], eventually leading to perceptions that differ from expectations. Therefore, understanding these effects is key to designing virtual audiences with clear and appropriate attitudes, especially for public speaking applications.

IVA '25, September 16–19, 2025, Berlin, Germany

ACM ISBN 979-8-4007-1508-2/2025/09
https://doi.org/10.1145/3717511.3747085

2 THEORETICAL BACKGROUND

In the context of public speaking training applications, the presence of IVAs is of prime importance, as they greatly contribute to skills acquisition [7, 17]. However, given the complexity of multimodal factors that influence their perception [10], IVAs should be carefully designed.

In 2002, within one of the first public speaking applications that was developed, three distinct virtual audiences were integrated, and designed as *positive* (*i.e.,* friendly and appreciative), *neutral*, or *negative* (*i.e.,* hostile and bored) [24]. Depending on the scenario, variations in posture, facial expressions, head movements, and gaze were implemented in the system. Even though VR technology was still in its early stages and the environment was less realistic than what can be developed nowadays, the negative audience successfully elicited a stronger anxiety response among users, paving the way for the development of dedicated training solutions. The accurate perception of such audience types was further supported by a 2023 study, which found that a positive emotional response was elicited when users were confronted with a positive audience in VR [25]. While these latter results are noteworthy, the behaviors implemented were not highly diverse. Positive and neutral audiences were both defined by a single combination of posture, facial expression, and head movement, whereas the negative audience was associated with specific events involving phone calls and chatting, along with a negative facial expression and head shakes.

Alternative approaches have been proposed to model IVAs' behavior. *Personality traits* have been notably assigned, including emotional stability and extroversion [19], or agreeableness [14]. By observing the agents' nonverbal behavior, these characteristics were accurately perceived by users. Furthermore, specific *attitudes* were attributed to IVAs and virtual audiences, either toward the speaker or the presentation itself. In [12] and [14], audiences were designed to appear as indifferent, critical, bored, or enthusiastic. Both studies resulted in the correct identification of the IVAs' attitude, using VR and PC-based systems, respectively.

In [14, 15], the *mood* of the IVAs was integrated within their proposed audience model and described in terms of valence, arousal, and dominance. Valence and arousal are two commonly used dimensions that guide the design of agents' nonverbal behavior. These dimensions are notably employed in [12] to delineate the agents' attitudes (*i.e.,* indifferent, critical, bored, or enthusiastic). As defined in [6], *emotional valence* refers to the positive or negative emotions felt by the IVA toward the speaker or the presentation, while *arousal* corresponds to its level of alertness during the presentation. Virtual agents can effectively convey different levels of valence and arousal through their nonverbal behavior [6, 9, 12, 14]. More specifically, [9] proposed a classification of 40 behavioral combinations, each consisting of a specific posture (*i.e.,* backward, forward, and upright variations), facial expression (*i.e.,* smile, frown, raised eyebrows, or none), and head movement (*i.e.,* nod, shake, tilt, or none). These combinations were evaluated using individual agents in VR, based on the levels of valence and arousal elicited, and ranked accordingly. Using these findings, the combinations were then grouped into 7 distinct categories, according to the associated perceived levels of emotional valence (*i.e.,* negative, neutral, or positive) and arousal (*i.e.,* low, neutral, or high). Although 9 valence–arousal pairings are theoretically possible, only 7 were represented among the combinations considered, with low arousal being exclusively associated with neutral valence. This categorization enabled the construction of a structured library of behaviors, from which each group can be used to assign IVAs animations that convey the desired levels of valence and arousal. The 40 behavioral combinations were however not evenly distributed across the categories, with certain groups containing fewer behaviors. This study also highlighted that the realism of IVAs plays a key role in user perception, as photorealistic avatars increase users' confidence in assessing their attitudes compared to cartoon ones, without leading to uncanny effects [9].

In addition to the nonverbal behavior and realism of IVAs, gender has been identified as a factor shaping perception. [1] has pointed out that gender stereotypes that occur in human-human interactions are replicated in human-agent interactions. For example, male virtual agents are considered more competent [1, 27], while female IVAs are perceived as friendlier and more positive [25, 27]. Gender also impacts users' emotions, with preferences for female assistive agents [8] and responding more positively to a female presence in a two-person audience [25]. These studies provide valuable insight regarding the influence of gender on perception, but focus on individual agents and small groups. To the authors' knowledge, no study has yet explored the influence of gender distribution within larger virtual audiences on the perception of their nonverbal behavior. Although previous studies have controlled for gender effects by balancing the audience composition [6, 12], further investigation is needed to inform the design of virtual audience behaviors, with consideration for gender. This study aims to address this gap.

There is also a predominance of perceptual studies focusing on the nonverbal behavior of individual agents rather than groups [10]. On one hand, [6] and [14] have investigated the perception of virtual audiences composed of 10 agents, but using PC-based systems, thus corroborating the gap in the literature regarding the limited use of immersive devices for perceptual studies. Moreover, the validity of these results is not guaranteed in the context of VR applications. On the other hand, VR devices were used in [12] to study virtual audiences composed of 10 agents, but with different behavioral combinations than [9] and without considering gender aspects in the analysis. While the IVA in [9] was seated behind a desk, the agents in [12] were simply seated on chairs, assuming the role of spectators within an audience. This divergence in audience design resulted in the implementation of distinct behavioral cues in both studies, thereby underscoring their complementarity.

In short, studies investigating the nonverbal behavior of audiences in VR remain scarce, and additional insights are needed to support the development of effective VR-based training systems. The present study will therefore examine how virtual audiences' nonverbal behavior impacts user perception, however, using larger audiences composed of 20 agents, which are common in public speaking contexts. The behavioral sequences proposed by [9] will be used, as they include more arousal and valence variations as well as different behaviors than those already studied by [12]. This is worth investigating whether these variations are similarly perceived by users when facing virtual audiences in VR. This would allow for the identification of additional behavioral sequences that could be used to design recognizable attitudes.

3 EXPERIMENT SETTINGS

The present study aims to address the aforementioned gaps through several research objectives. The behavioral combinations from [9] will be distributed within virtual audiences of 20 photorealistic agents to assess whether similar valence and arousal ratings are observed at the group level, and to define audience types that consistently convey the desired perceptions. In addition, various gender distributions of IVAs within the audiences will be displayed to examine the influence of gender on user perception. In total, 35 audience conditions will be assessed.

3.1 Virtual environment design

A team of 3D game artists and game developers working at the authors' university was responsible for the development of the virtual environment used in the present study, using *Blender* and *Unity* respectively. The created environment consists of three rows of benches on which the virtual audience is seated. Although the room resembles a classroom, it remains neutral such that it can be used to simulate a variety of public speaking situations, thereby ensuring the general applicability of the results. The virtual audience is composed of 20 IVAs, which vary in terms of nonverbal behavior and gender across the experiment conditions. The user is standing in front of them (see Figure 1).

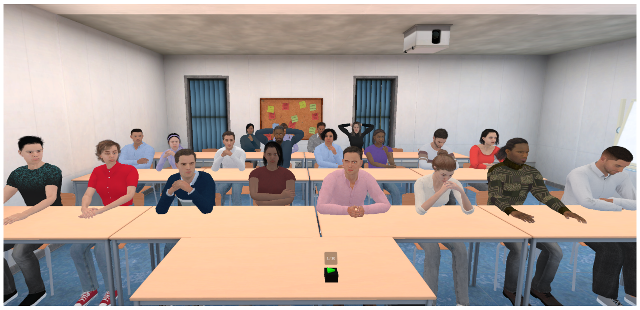

Figure 1: Screenshot of the virtual environment, displaying an audience composed of 10 male and 10 female agents.

To design the audience, 40 photorealistic virtual agents have been created using the *Character Creator* solution of the *Reallusion* suite. To reinforce the feeling of presence while immersed in VR, both the agents and the virtual environment were developed to be highly realistic. More specifically, 20 adult female and 20 adult male agents were conceived, representing various ethnic origins. These 40 virtual agents were used to shape 5 groups of 20 IVAs, each with a different distribution of male and female agents to create gender variations. The gender distributions that were included in the study are presented in Table 1. In the virtual audience G2 *(resp. G4)*, the 5 female *(resp. 5 male)* agents were grouped and seated on the left side of the room. This configuration was chosen to make the presence of the under-represented gender more noticeable to users. When it comes to the condition G3, male and female agents were mixed and distributed among the three rows (see Figure 1).

In addition, the combinations of posture, head movements, and facial expressions presented in [9] were integrated into the present system to create additional audience variations, based on the IVAs' nonverbal behavior. The 7-group categorization of the 40 different animations proposed in that study was used as a basis. Each group corresponds to a specific level of perceived valence (*i.e.,* negative, neutral, or positive) and arousal (*i.e.,* low, neutral, or high), and includes a set of animations that reflect distinct nonverbal behaviors, previously associated with these perception ratings in [9] at the individual level. For each designed attitude condition, the animations from the corresponding group are assigned across the IVAs in the audience, creating diversity in their behavior while aiming to produce a coherent overall perception of valence and arousal.

Table 1: Description of gender and attitude conditions.

Gender conditions		Attitude conditions		
	Description		**Description**	**# sequences**
G1	20 male agents	**A1**	Low arousal, neutral valence	3
G2	15 male, 5 female agents	**A2**	Neutral arousal, negative valence	2
G3	10 male, 10 female agents	**A3**	Neutral arousal, neutral valence	4
G4	5 male, 15 female agents	**A4**	Neutral arousal, positive valence	2
G5	20 female agents	**A5**	High arousal, negative valence	20
		A6	High arousal, neutral valence	3
		A7	High arousal, positive valence	6

This process led to the creation of 7 distinct audience attitude types (see Table 1). Moreover, as the gaze behavior of IVAs is a key cue for perceiving arousal [6], their percentage of eye contact with the user was modulated based on the intended arousal level (*i.e.,* 20%, 50%, and 80% for the low, neutral, and high arousal levels respectively). These values were adapted from those used in [6] and slightly revised to strengthen the distinction between the designed arousal levels, while avoiding extreme gaze behaviors (*e.g.,* 0% or 100%) that are rarely observed in natural audience settings. This results in dynamic audience behavior, where each IVA maintains a specific posture and facial expression, while displaying its assigned head and gaze movements at different times, making the audience appear more realistic. Finally, the behavioral combinations were attributed to the IVAs following specific rules. First, based on the valence and arousal rankings proposed by [9], the best-recognized behaviors in each category were first assigned to the virtual agents that were closest to the user. Then, for categories containing only a few animations (*i.e.,* all except A5 and A7), behaviors were alternated between IVAs to keep the experience as realistic as possible.

As a result, the virtual environment includes 35 audience conditions (*i.e.,* 5 gender × 7 attitude). To minimize variability within the observed audiences, the 5 gender distributions were kept identical across participants. Each time a specific gender condition was observed, the same agents were used and placed in the same positions within the audience rows. This was applied to the attitude manipulation as well, with identical behavioral sequences assigned to the same positions within the rows, regardless of the gender condition.

3.2 Methodology

The experimental conditions were distributed among 70 participants (*i.e.,* 35 women and 35 men), in a way that ensured each condition was assessed by an equal number of male and female participants. Given the study's focus on the impact of IVAs' gender on user perception, particular attention was paid to maintaining gender balance both in the recruitment process and the conditions distribution to minimize potential bias. With a final sample size of

70 participants, the required sample sizes indicated by the power analyses were exceeded (N = 45 for Wilcoxon signed-rank tests; N = 62 for linear mixed-effects models).

Most participants were young adults (M = 28.01, SD = 10.26), from Western Europe (N = 68) and French-speaking (N = 67). All were above the age of 18. A majority (N = 51) had already tried virtual reality, mostly for entertainment purposes, though typically only once or twice. Participants were recruited through a dedicated call and were invited within the authors' university to take part in the experiment. A quiet room was used to favor both immersion and participants' focus during the experience. Before anything else, the conditions of the study were presented to the participants, and informed consent was requested.

A short questionnaire covering demographic questions as well as information about the participants' previous experience with VR was first completed. Then, the definitions of both *valence* and *arousal*, as described by [7], were provided. Participants had to confirm their understanding of these concepts before the start of the second part of the experiment. However, no illustrative animations of valence and arousal levels were presented to the participants before the experiment, as the aim was to remain close to the end-user experience, in which no such information would be provided.

Participants were then immersed in the VR environment using Meta Quest 3 headsets and were asked to observe 10 virtual audiences, one after the other. Each of the observed audiences corresponds to a specific gender and a specific attitude condition, as described in Section 3.1. The number of sequences was limited to 10 to keep the immersion brief and maintain participant engagement throughout the experiment. Ultimately, each of the 35 virtual audience conditions was assessed by the same number of participants, with an equal number of male and female respondents. In addition, each participant assessed all gender conditions, whatever the attitude type associated with the audience, and all attitude conditions, whatever the gender distribution within the audience. They are, however, presented to users in a different order to avoid potential bias. Finally, while no participant was exposed to the same gender-attitude pair more than once, they were still exposed to multiple audiences sharing the same gender or attitude category. This introduced within-subject dependencies in the perception ratings. A repeated measures design was therefore adopted.

For each audience, the participants should observe its behavior for at least 30 seconds, since IVAs eventually display head movements and gaze variations over time. After this time limit, a sound was triggered to notify the participant that the audience can be evaluated in terms of valence and arousal. The participants could however take as long as necessary before giving their answer. Both emotional valence and arousal were rated using 7-point Likert scales, where 1 corresponded to *"Very negative"* and *"Very low"* respectively, and 7 to *"Very positive"* and *"Very high"*. In addition, participants were asked to rate their level of confidence in each of these two answers on separate 7-point Likert scales, ranging from *"Very low"* to *"Very high"*. Once the observed audience has been evaluated, participants could move to the next one by pressing a dedicated button located on the desk in front of them (see Figure 1). This process was repeated for all 10 virtual audiences to be evaluated, after which the VR experience was concluded.

Finally, participants were asked to complete a Presence Questionnaire. For French-speaking participants, the validated French-Canadian version of the Presence Questionnaire was used [4, 18], while non-French-speaking participants completed the original version developed by Witmer and Singer [29]. Before concluding the experiment, participants were also allowed to provide written comments about their experience.

4 RESULTS

4.1 Nonverbal behavior perception

To determine how the designed attitudes were perceived in terms of valence and arousal, regardless of the gender of the audience, Wilcoxon signed-rank tests were conducted. This non-parametric approach was chosen given the repeated measures design of the study and the violation of normality assumptions. To control for multiple comparisons, p-values were adjusted using the Holm correction method.

Each test assessed whether the distribution of participant ratings for a given attitude was significantly different than 4, thereby indicating if the attitude was perceived as neutral or not. This was done for each dependent variable, *i.e.,* valence and arousal ratings. Moreover, for significant tests, the median was used to better understand how the attitude was perceived. In short, an attitude is considered *positive* (*resp. negative*) in terms of valence when the test is statistically significant and the median is greater than 4 (*resp. less than 4*). For arousal, the same logic is applied to determine high and low arousal levels. In all other cases, the attitude is interpreted as neutral. To complete the analysis, effect sizes were computed to assess how strongly attitudes influenced perceived valence and arousal. This three-step process allows for an identification of how each attitude is perceived, which helps to define a library of nonverbal behaviors, as in [9], that maintain consistent perception, without being altered by potential group effects.

Valence. The analysis indicated that the designed attitudes effectively conveyed different levels of emotional valence (see Table 2a). Attitudes A4 and 7 were perceived as particularly positive, with a median rating of 5 and 6, respectively ($p_{\text{adj}} < .001$). Moreover, attitude A7 presents a large effect size ($r = .82$), which indicates that the attitude was perceived as strongly positive in terms of valence. The effect was more moderate for A4 ($r = .39$). These results confirm that the positive valence of these two attitudes was successfully perceived by participants. Conversely, attitude A5 was rated as significantly negative ($p_{\text{adj}} < .001$), with a median of 3 and a large effect size ($r = .54$). These results were in line with expectations (see Table 1), indicating that these attitudes were successfully perceived in terms of valence.

For attitude A1, the Wilcoxon signed-rank test was significant ($p_{\text{adj}} = .002$), indicating a non-neutral perception of this attitude. However, since the median rating is equal to 4, a one-sided Wilcoxon signed-rank test was conducted to clarify the direction of this deviation. The results revealed that ratings tended to be higher than 4 ($p < .001$, $r = .35$), suggesting a slight positive bias in the perception of this attitude. Although the attitude was generally rated as neutral and therefore classified as such, participants tended to evaluate it more positively than expected. This should be taken into account when using this attitude in future applications.

The remaining attitudes (*i.e.,* A2, A3, and A6) did not significantly differ from the neutral point ($p > .05$) and presented median valence scores of 4 and small effect sizes ($r < .22$). These attitudes were then perceived as neutral. This perception was expected for both attitudes A3 and A6. In contrast, attitude A2 was designed based on behavioral sequences previously described as negative [9], showing a gap between how the attitude was designed and perceived.

Table 2: Summary of perceived valence and arousal by audience attitude

(a) Valence ratings

Attitude	Median	p_{adj}	Effect Size (r)	Perception
A1	4	< .01	.349	Neutral
A2	4	.107	.212	Neutral
A3	4	.259	.154	Neutral
A4	5	<.001	.387	Positive
A5	3	<.001	.536	Negative
A6	4	.259	.153	Neutral
A7	6	<.001	.816	Positive

(b) Arousal ratings

Attitude	Median	p_{adj}	Effect Size (r)	Perception
A1	4	.232	.157	Neutral
A2	5	< .01	.346	High
A3	4	.232	.160	Neutral
A4	5	< .001	.540	High
A5	5	< .001	.725	High
A6	5	< .001	.540	High
A7	7	< .001	.862	High

Arousal. The analysis of arousal ratings revealed that only two levels of arousal were perceived by the participants (see Table 2b). Indeed, five attitudes, namely A2, A4, A5, A6, and A7, were associated with high levels of arousal, as indicated by significant Wilcoxon signed-rank tests ($p_{\text{adj}} < .01$) and median scores larger than 4. Among them, attitude A7 stands out from the analysis with an extremely low p-value ($p_{\text{adj}} < .001$), a median of 7, and a very large effect size ($r = .862$), indicating that the audiences associated with this attitude were perceived as highly arousing. This aligns with its intended design. Attitudes A2, A4, A5, and A6 were all associated with high arousal levels, showing significant results ($p_{\text{adj}} < .01$) and median scores of 5. The effect sizes for A4 ($r = .54$), A5 ($r = .725$), and A6 ($r = .54$) were large, indicating that these attitudes were perceived as clearly arousing. This was expected for A5 and A6, which were explicitly designed to reflect high arousal. However, A4 was initially intended to appear neutral, suggesting a mismatch between design and perception. Regarding A2, the effect size was moderate ($r = .346$), indicating that it was perceived as fairly arousing rather than neutral, as originally intended.

Only two attitudes, A1 and A3, were classified as neutral in terms of arousal. Both had non-significant Wilcoxon test results and median ratings of 4. While A3 was designed to reflect a neutral arousal level and was perceived accordingly, A1 was designed to depict a low arousal level but was instead perceived as neutral, suggesting a mismatch between its intended and perceived effect.

Correlation between valence and arousal. Spearman's rank-order correlations were computed to better understand the relationship between valence and arousal ratings, as well as the possible mismatches between the intended and perceived audience attitudes. The analysis revealed a positive overall correlation between arousal and valence across all attitudes, indicating that higher arousal levels were generally associated with higher valence ratings ($\rho = .46$, $p < .001$). When examining this relationship separately for each attitude, significant correlations between valence and arousal were also observed ($p < .05$), with coefficients ranging from $\rho = .24$ for the attitude A5 to $\rho = .50$ for the attitude A7. The strength of this relationship was therefore particularly high for attitude A7. This suggests that, although the strength of the correlations varied slightly across behaviors, the overall pattern remained consistent, with participants tending to associate higher arousal levels with more positive perceptions of valence.

Confidence levels. Confidence ratings were consistently high across all attitudes, with mean scores ranging from 5.36 to 6.18. This suggests that participants generally felt sure about how they rated the different attitudes. Among them, attitude A7 stood out with the highest average confidence scores for both valence (M = 6.06) and arousal (M = 6.18), which may reflect how clearly it was perceived by participants. Given the overall similarity in confidence levels, no further analysis was considered necessary.

4.2 Gender influence on perception

To analyze the influence of the audience's gender on perception, linear mixed-effects models were conducted. This approach was selected given the repeated measures design of the study. Four models were first included in the analysis, each examining the effect of audience gender on perceived valence, perceived arousal, and the confidence ratings associated with these dimensions, while taking into account the within-subject variability through a random effect for the participant. For all models, no significant effect of audience gender on ratings was observed (all $p > .19$). A slight trend towards less confident arousal ratings was however observed in the G2 condition ($\beta = -0.17$, $p = .099$), even if not significant. Although small variations in participants' answers were observed, no systematic influence of gender was found in the results, indicating that the gender composition of the virtual audiences did not significantly impact participants' ratings.

To deepen the analysis, four additional linear mixed-effects models were performed, including the interaction between audience gender and attitude. The objective is to examine whether their combined effect influenced perceived valence, arousal, and confidence ratings. These models still did not reveal any significant interaction (all $p > .05$). Nevertheless, a few combinations approached significance. Attitude A3 indeed presented increased confidence in valence ratings when displayed by an all-male audience (*i.e.,* G1; $\beta = 0.74$, $p = .051$). Similarly, the interaction between attitude A6 and the same gender condition was associated with a positive yet non-significant effect on confidence in valence ratings ($\beta = 0.41$, $p = .278$). In contrast, a negative interaction trend was observed between attitude A6 and the G1 condition for arousal confidence ratings ($\beta = -0.57, p = .123$). These results then reveal variations in confidence when attitudes are displayed by male audiences. However, since none of these effects reached significance, no conclusion can be drawn regarding their influence on perception.

Since the study deals with gender-related aspects, the impact of the gender of the participants was considered as well. More precisely, the goal was to check whether it interacted with the gender composition of the audience to determine perception ratings. To do so, linear mixed-effects models were again conducted, including a fixed interaction effect between the participant's gender and the audience's gender. Across all models, no significant interaction was observed (all $p > .15$), indicating that participants reported similar perception ratings for a given attitude, regardless of both their gender and the gender of the virtual audience.

In short, no significant or consistent effects of audience gender on perception were found in any of the models. Whether tested in interaction with attitude type or participant gender, the audience gender did not significantly influence ratings.

4.3 Presence Questionnaire

The virtual environment was also evaluated in terms of the feeling of presence, to ecologically validate both the room design and the virtual audiences. Overall, the participants reported a high sense of presence within the environment, with an average score of 72.97%, which is relatively high. Among the dimensions included in the presence questionnaire, the realism of the solution reached a score of 70.75% on the associated scale. For instance, participants reported that the observed environment was similar to what they might observe in real-life settings (66%). In addition, they felt highly involved in the experience (82%). The realism of the designed virtual agents, as well as their behavior, therefore did not hinder the feeling of presence. These results are highly promising given the objective to use the system for public speaking training. Participants were also able to examine the virtual environment clearly (71.29%) and found it interactive (68.48%), although only a few interactions were implemented. The quality of the experience received particularly good ratings (79.57%), and the performance dimension was rated the highest (82.30%). The scores were very high across all dimensions and consistent with other studies [4, 9, 18], which reported similar scores. This confirms that the experience was immersive and credible to the participants.

5 DISCUSSION

In the context of this study, a set of 7 attitudes for virtual audiences was developed, each corresponding to an expected level of valence and arousal (see Table 1). An overview of these attitudes, summarizing the expected and perceived levels for both valence and arousal, is shown in Table 3. Overall, 4 out of 7 attitudes were perceived as expected, namely A3, A5, A6, and A7. The most recognizable attitudes were A5 and A7, which are of particular interest in the context of public speaking training. They represent a critical but attentive audience and an interested and approving audience, respectively. The recognition of such audience types has been demonstrated in other studies [12, 14], thus reinforcing their relevance in public speaking training systems. They are indeed easily recognizable and can therefore be used to provide users with immediate feedback on their speaking performance, thereby inducing positive or negative emotions accordingly. Furthermore, attitude A6, which represents a compromise between the two aforementioned attitudes, has also been effectively recognized. This neutral yet arousing audience can thus serve as an indicator to the speaker that the audience's attitude is gradually shifting. This particular audience type has not previously been examined using VR systems, thereby underscoring the users' capacity to perceive variations in valence in the presence of highly aroused audiences.

Table 3: Actual perception and expectations for the designed attitudes.

	Valence		Arousal	
Attitude	**Perceived**	**Expected**	**Perceived**	**Expected**
A1	Neutral	Neutral	Neutral	Low
A2	Neutral	Negative	High	Neutral
A3	Neutral	Neutral	Neutral	Neutral
A4	Positive	Positive	High	Neutral
A5	Negative	Negative	High	High
A6	Neutral	Neutral	High	High
A7	Positive	Positive	High	High

Additionally, the neutral audience (*i.e.,* neutral valence and neutral arousal) was accurately perceived in this study. Having a neutral state is particularly relevant in public speaking training, as it can be used as a default nonverbal behavior, thereby enabling speakers to track how audience attitudes change over time. In contrast, the two other moderately arousing audiences were not perceived as such, although the A4 audience correctly conveyed the associated level of valence. This suggests that, to be properly designed, this attitude should have been perceived as slightly less arousing. The correlation between valence and arousal may provide a potential explanation for this misperception, as positive valence ratings are associated with higher perceptions of arousal. However, given its median value of 5 and its moderate associated effect size, the attitude could potentially still be employed in this context as neutral. It is indeed perceived as less arousing than attitude A7, even though it is associated with the same perceived valence and arousal levels. Similarly, attitude A1 was perceived as neutral rather than low in arousal, highlighting that the designed attitudes were generally perceived as more arousing than expected. As can be observed in [9], arousal ratings for the observed behavioral sequences were typically high, with only a few rated as low. In particular, the sequence rated as least arousing received 57% negative ratings, compared to 90% for the most negative sequence in terms of valence. This disparity suggests that the range of perceived arousal was narrower and closer to neutral, potentially making these sequences more blurred and difficult to interpret. The inclusion of specific events, such as IVAs sleeping or looking at their phones instead of listening, may then be necessary to convey low arousal levels. This observation is consistent with participants' comments, who explained that they expected more pronounced audience reactions, closer to what can be observed in a classroom. Finally, this tendency for increased arousal perception ratings may be the result of a group effect. Indeed, there were always a few agents in motion due to head movements and gaze shifts. At the group level, this may have increased the overall sense of movement and led to higher perceived arousal than for individual agents.

Valence variations within the audience were more accurately recognized than arousal ones, with only one attitude misperceived. Specifically, attitude A2 was expected to convey negative valence but was perceived as neutral. One potential solution could be to favor specific behaviors, eliciting stronger negative valence perceptions, in the distribution of animations within the audience, rather than distributing them equitably across IVAs. However, this requires enough behavioral variety within the considered attitude group. For instance, while this could help for attitude A1, which was associated with neutral instead of low arousal, this attitude group contains only three different behavior combinations associated with only two postures (see Table 1), which may then reduce the realism. In such cases, integrating behaviors from other attitude groups may help, as the accuracy of perception improves when more clearly defined agents are incorporated [6]. For instance, to obtain the negative perception expected with A2, some behaviors from A5 could be added, as it successfully elicited negative valence ratings. This trick can therefore shift a neutral perception to a negative one, but this should be done carefully to keep the overall arousal perception moderate, as A7 is also linked to high arousal.

Participants also expressed that, without a baseline, it was difficult to assess the audience since the range of potential responses was unknown. This approach was preferred to better understand how potential users of the training system, who may not be familiar with the full range of possible attitudes, would perceive them. However, by observing audience reactions at runtime during the presentation, variations may become more salient to the user. Consequently, it may be valuable to investigate whether changes in valence or arousal help users detect, in real time, shifts in audience attitudes toward more positive or negative states. These questions represent promising areas of research to further support the development of clearly identifiable virtual audiences.

In addition, it is important to note that one particular posture, namely P3 (*i.e.,* backward posture with arms behind the head) from [9], generated comments from the participants. This posture was described as not natural, as it does not occur frequently in real audiences, according to the participants. They also reported that they felt they correctly identified the meaning of this posture and answered the questionnaire accordingly, while still being disturbed. This posture should therefore be displayed sparingly when creating virtual audiences, in order not to reduce the realism of the audience and the level of presence, which is essential for training. In addition, although particularly high presence scores were obtained, these may be explained by the fact that most of the participants had already tried VR (*i.e.,* 51 participants) and therefore felt at ease with the system. These results still remain particularly promising given that most participants had only used VR once or twice, and those with no prior experience at all reported similar presence scores.

These results regarding the perception of the nonverbal behavior of virtual audiences constitute a strong contribution to this paper, providing guidelines for their design. This study offers a set of validated audience attitude types and provides a comprehensive library of virtual audience configurations, shown to elicit consistent perception ratings for groups of IVAs in VR environments. These results also underscore the complexity of group effects, which can significantly alter perception. An additional contribution is the identification of the absence of gender effects in the evaluation of virtual audiences. While previous research has shown that gender influences perception [1, 10, 25], none of the gender conditions in this study had a significant impact. No interaction with the user's gender has been identified either. This suggests that the gender of IVAs within virtual audiences can be adjusted without leading to a misinterpretation of attitudes. Nevertheless, other aspects of public speaking may still be influenced by gender, such as the speaker's stress level, confidence, or personal preferences, which would require further investigation. Additionally, investigating how the gender and nonverbal behavior of IVAs might influence the speaker's gaze during an actual speaking task is worth considering. This aspect was not examined in the present study, as participants were explicitly instructed to observe the entire audience.

Nevertheless, significant work remains to be done in order to develop realistic and responsive virtual audiences. This includes the development of IVAs that can adapt their behavior based on the speaker's performance. Previous studies have identified relevant performance indicators, making the creation of such adaptive audiences feasible [30]. However, it is still necessary to define how audience attitudes should evolve during the presentation so that the changes feel realistic and serve the training objectives. The development of virtual audiences and, more broadly, of VR training environments tailored to the users' needs, remains therefore a challenging yet necessary field of research, given the importance of communication skills in both personal and professional contexts.

6 CONCLUSION

The virtual audiences designed for this experiment have been partially validated in terms of attitude, even though the most characteristic ones were correctly perceived. The proposed attitudes should therefore be slightly adjusted in future studies to ensure accurate perception of both valence and arousal. A potential solution suggested within the scope of this study is the inclusion of additional behaviors that correspond to specific events observed in real audiences, such as IVAs talking to their neighbors, looking at their phones, or even falling asleep. Such behavioral cues may reinforce the intended design of the audience types. Similarly, the gender of the audience did not affect perception within the scope of this experiment. However, it may have an impact on other user-related aspects, such as the user's stress, confidence, or preferences, which were not considered here. This study also allowed the validation of the virtual environment in terms of presence, as it was associated with particularly high scores across all dimensions of the questionnaire [4, 29]. Moreover, the experience was described by the participants as enjoyable and realistic, providing a strong impression of speaking in front of a real crowd.

Future work should now aim to integrate the obtained results into a dedicated VR-based public speaking training system, to investigate how the characteristics of the virtual audience influence users' emotions and perception during speaking tasks, and how this impacts skills acquisition under different audience conditions.

REFERENCES

[1] Marjorie Armando, Magalie Ochs, and Isabelle Régner. 2022. The Impact of Pedagogical Agents' Gender on Academic Learning: A Systematic Review. *Frontiers in Artificial Intelligence* 5 (2022). https://doi.org/10.3389/frai.2022.862997

[2] Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. Cicero - Towards a multimodal virtual audience platform for public speaking training. In *13th International Conference on Intelligent Virtual Agents (IVA 2013)*. Scottish Informatics and Computer Science Alliance. https://doi.org/10.1007/978-3-642-40415-3_10

[3] Sihem Ben Saad. 2024. The digital revolution in the tourism industry: role of anthropomorphic virtual agent in digitalized hotel service. *International Journal of Contemporary Hospitality Management* 36, 11 (2024), 3751–3773. https://doi.org/10.1108/IJCHM-09-2023-1485

[4] Stéfan Bouchard and G. Robillard. 2019. Validation canadienne-française du Gatineau Presence Questionnaire auprès d'adultes immergés en réalité virtuelle. In *7e Congrès de l'ACFAS*.

[5] Veena Chattaraman, Wi-Suk Kwon, and Juan E. Gilbert. 2012. Virtual agents in retail web sites: Benefits of simulated social interaction for older users. *Computers in Human Behavior* 28, 6 (2012), 2055–2066. https://doi.org/10.1016/j.chb.2012.06.009

[6] Mathieu Chollet and Stefan Scherer. 2017. Perception of Virtual Audiences. *IEEE Computer Graphics and Applications* 37, 4 (2017), 50–59. https://doi.org/10.1109/MCG.2017.3271465

[7] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *3rd ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*. Association for Computing Machinery, Inc. https://doi.org/10.1145/2750858.2806060

[8] Anna Esposito, Terry Amorese, Marialucia Cuciniello, Maria Teresa Riviello, Antonietta M. Esposito, Alda Troncone, Maria Inés Torres, Stephan Schlögl, and Gennaro Cordasco. 2021. Elder user's attitude toward assistive virtual agents: the role of voice and gender. *Journal of Ambient Intelligence and Humanized Computing* 12, 4 (2021), 4429–4436. https://doi.org/10.1007/s12652-019-01423-x

[9] Elodie Etienne, Anne-Lise Leclercq, Angélique Remacle, Laurence Dessart, and Michaël Schyns. 2023. Perception of avatars nonverbal behaviors in virtual reality. *Psychology and Marketing* 40, 11 (2023). https://doi.org/10.1002/mar.21871

[10] Elodie Etienne, Marion Ristorcelli, Sarah Saufnay, Aurélien Quilez, Rémy Casanova, Michaël Schyns, and Magalie Ochs. 2024. A Systematic Review on the Socio-affective Perception of IVAs' Multi-modal behaviour. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA 2024, 2)*. 1–10. https://doi.org/10.1145/3652988.3673943

[11] Matteo Girondi, Ivana Frigione, Mariapia Marra, Milena Stefanova, Margherita Pillan, Angelo Maravita, and Alberto Gallace. 2024. Decoupling the role of verbal and non-verbal audience behavior on public speaking anxiety in virtual reality using behavioral and psychological measures. *Frontiers in Virtual Reality* 5 (2024). https://doi.org/10.3389/frvir.2024.1347102

[12] Yann Glémarec, Jean-Luc Lugrin, Anne-Gwenn Bosser, Aryana Collins Jackson, Cédric Buche, and Marc Erich Latoschik. 2021. Indifferent or Enthusiastic? Virtual Audiences Animation and Perception in Virtual Reality. *Frontiers in Virtual Reality* 2 (2021). https://doi.org/10.3389/frvir.2021.666232

[13] Arno Hartholt, Sharon Mozgai, and Albert Rizzo. 2019. Virtual job interviewing practice for high-anxiety populations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA 2019)*. Association for Computing Machinery, Inc, 238–240. https://doi.org/10.1145/ivade780

[14] Ni Kang, Willem-Paul Brinkman, Birna Van Riemsdijk, and Marc Neerickx. 2016. The design of virtual audiences: Noticeable and recognizable behavioral styles. *Computers in Human Behavior* 55 (2016), 680–694. https://doi.org/10.1016/j.chb.2015.10.008

[15] Ni Kang, Willem-Paul Brinkman, M. Birna van Riemsdijk, and Mark A. Neerincx. 2013. An Expressive Virtual Audiencewith Flexible Behavioral Styles. *IEEE Transactions on Affective Computing* 4, 4 (2013), 326–340. https://doi.org/10.1109/TAFFC.2013.2297104

[16] Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. 2019. You'll be Great: Virtual Agent-based Cognitive Restructuring to Reduce Public Speaking Anxiety. In *8th International Conference on Affective Computing and Intelligent Interaction* (Cambridge, UL) *(ACII 2019, 8925438)*. IEEE. https://doi.org/10.1109/ACII.2019.8925438

[17] Leon O. H. Kroczek and Andreas Mühlberger. 2023. Public speaking training in front of a supportive audience in Virtual Reality improves performance in real-life. *Scientific Reports* 13, 1 (2023). https://doi.org/10.1038/s41598-023-41155-9

[18] Mylène Laforest, Stéphane Bouchard, Ana-Maria Crétu, and Olivier Mesly. 2016. Inducing an Anxiety Response Using a Contaminated Virtual Environment: Validation of a Therapeutic Tool for Obsessive–Compulsive Disorder. *Frontiers in ICT* 3, 18 (2016). https://doi.org/10.3389/fict.2016.00018

[19] Jean-Luc Lugrin, Jessica Topel, Yann Glémarec, Birgit Lugrin, and Marc Erich Latoschik. 2023. Posture Parameters for Personality-Enhanced Virtual Audiences. *23rd ACM International Conference on Intelligent Virtual Agents* 46. https://doi.org/10.1145/3570945.3607311

[20] Guido Makransky, Stefan Borre-Gude, and Richard E. Mayer. 2019. Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning* 35, 6 (2019), 691–707. https://doi.org/10.1111/jcal.12375

[21] Diego Monteiro, Airong Wang, Luhan Wang, Hongji Li, Alex Barrett, Austin Pack, and Hai-Ning Liang. 2024. Effects of audience familiarity on anxiety in a virtual reality public speaking training tool. *Universal Access in the Information Society* 23, 1 (2024). https://doi.org/10.1007/s10209-023-00985-0

[22] Prasanth Murali, Ha Trinh, Lazlo Ring, and Timothy Bickmore. 2021. A Friendly Face in the Crowd: Reducing Public Speaking Anxiety with an Emotional Support Agent in the Audience. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents (IVA 2021)*. Association for Computing Machinery, Inc, 156–163. https://doi.org/10.1145/3472306.3478364

[23] Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *10th International Conference on Intelligent Virtual Agents (IVA 2010, 6356)*. Springer Verlag, 222–235. https://doi.org/10.1007/978-3-642-15892-6_24

[24] David-Paul Pertaub, Mel Slater, and Chris Barker. 2002. An experiment on public speaking anxiety in response to three different types of virtual audience. *Universal Access in the Information Society* 11, 1 (2002), 68–78. https://doi.org/10.1162/105474602317343668

[25] Marion Ristorcelli, Emma Gallego, Kévin Nhuy, Jean-Marie Pergandi, Rémy Casanova, and Magalie Ochs. 2023. Investigating the Impact of a Virtual Audience's Gender and Attitudes on a Human Speaker. In *25th International Conference on Multimodal Interaction (ICMI 2023)*. Association for Computing Machinery. https://doi.org/10.1145/3610661.3616128

[26] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *16th International Conference on Intelligent Virtual Agents* (Los Angeles, USA) *(IVA 2016, 10011)*. Springer Verlag, 40–50. https://doi.org/10.1007/978-3-319-47665-0_4

[27] Silke ter Stal, Monique Tabak, Harm op den Akker, Tessa Beinema, and Hermie Hermens. 2020. Who Do You Prefer? The Effect of Age, Gender and Role on Users' First Impressions of Embodied Conversational Agents in eHealth. *International Journal of Human-Computer Interaction* 36, 9 (2020), 881–892. https://doi.org/10.1080/10447318.2019.1699744

[28] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. In *ACM Conference on Motion, Interaction, and Games (MIG 2019)*. Association for Computing Machinery, Inc. https://doi.org/10.1145/3359566.3360065

[29] Bob G. Witmer and Michael J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (1998), 225–240. https://doi.org/10.1162/105474698565686

[30] Torsten Wörtwein, Louis-Philippe Morency, and Stefan Scherer. 2015. Automatic assessment and analysis of public speaking anxiety: A virtual audience case study. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 187–193. https://doi.org/10.1109/ACII.2015.7344570