frontiers | Frontiers in Virtual Reality

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/frvir.2022.876433

Controlling the Stage: A High-Level Control System for Virtual Audiences in Virtual Reality

Yann Glémarec [1,2], Jean-Luc Lugrin [2], Anne-Gwenn Bosser [1], Cédric Buche[3] and Marc Erich Latoschik[2]*

[1]Lab-STICC, CNRS UMR 6285, ENIB, Brest, France, [2]Chair of Human-Computer-Interaction (Informatik IX), Julius-Maximilians-Universität Würzburg, Würzburg, Germany, [3]IRL Crossing, CNRS, ENIB, Adelaide, SA, Australia

This article presents a novel method for controlling a virtual audience system (VAS) in Virtual Reality (VR) application, called STAGE, which has been originally designed for supervised public speaking training in university seminars dedicated to the preparation and delivery of scientific talks. We are interested in creating pedagogical *narratives*: narratives encompass affective phenomenon and rather than organizing events changing the course of a training scenario, pedagogical plans using our system focus on organizing the affects it arouses for the trainees. Efficiently controlling a virtual audience towards a specific training objective while evaluating the speaker's performance presents a challenge for a seminar instructor: the high level of cognitive and physical demands required to be able to control the virtual audience, whilst evaluating speaker's performance, adjusting and allowing it to quickly react to the user's behaviors and interactions. It is indeed a critical limitation of a number of existing systems that they rely on a Wizard of Oz approach, where the tutor drives the audience in reaction to the user's performance. We address this problem by integrating with a VAS a high-level control component for tutors, which allows using predefined audience behavior rules, defining custom ones, as well as intervening during run-time for finer control of the unfolding of the pedagogical plan. At its core, this component offers a tool to program, select, modify and monitor interactive training narratives using a high-level representation. The STAGE offers the following features: i) a high-level API to program pedagogical narratives focusing on a specific public speaking situation and training objectives, ii) an interactive visualization interface iii) computation and visualization of user metrics, iv) a semi-autonomous virtual audience composed of virtual spectators with automatic reactions to the speaker and surrounding spectators while following the pedagogical plan V) and the possibility for the instructor to embody a virtual spectator to ask questions or guide the speaker from within the Virtual Environment. We present here the design, and implementation of the tutoring system and its integration in STAGE, and discuss its reception by end-users.

Keywords: virtual reality, virtual agent, behavior perception, public speaking, education

OPEN ACCESS

Edited by:
Funda Durupinar, University of Massachusetts Boston, United States

Reviewed by:
Meredith Carroll, Florida Institute of Technology, United States
Uğur Güdükbay, Bilkent University, Turkey

****Correspondence:***
Yann Glémarec
yann.glemarec@uni-wuerzburg.de

Specialty section:
This article was submitted to Virtual Reality and Human Behaviour, a section of the journal Frontiers in Virtual Reality

Received: *15 February 2022*
Accepted: *31 March 2022*
Published: *04 May 2022*

Citation:
Glémarec Y, Lugrin J-L, Bosser A-G, Buche C and Latoschik ME (2022) Controlling the Stage: A High-Level Control System for Virtual Audiences in Virtual Reality. Front. Virtual Real. 3:876433. doi: 10.3389/frvir.2022.876433

1 INTRODUCTION

Virtual Humans are now tremendously used in virtual reality (VR) systems. They are used for entertainment, training, or therapeutic purposes either alone or as a group of virtual agents. Several VR applications have a need for being populated with groups of virtual agents, whether for VR exposure therapies to mitigate the fear of public speaking (Wallach et al., 2009; Anderson et al., 2013; Kahlon et al., 2019), for public speaking training systems (Batrinca et al., 2013; Chollet et al., 2014), or even for audience management training (Hayes et al., 2013; Lugrin et al., 2016; Fukuda et al., 2017; Shernoff et al., 2020; Delamarre et al., 2021). Such virtual agents groups are defined as virtual audiences when the agents watch an activity without taking part in them or are mimicking virtual spectators.

The usefulness of virtual audiences lies in our ability to modify their behavior so that they convey emotions that are then perceived by the user. These perceived emotions come from non-verbal behaviors as well as from the various social signals emitted such as backchannels or interactions between agents and users (Kang et al., 2016; Chollet and Scherer, 2017). It is therefore through the control of the attitude towards the user that social applications in VR benefit from adequate teaching and therapeutic environments.

Consequently, the reason why virtual audiences are used as a fear stimulus during VR exposure therapies (which consists of repeatedly exposing a patient to varying degrees of a feared stimulus to modify a behavioral or cognitive response) (Rothbaum et al., 2000; Anderson et al., 2005), is that they provide the virtual environments with fine-tuning of this stimulus with virtual audiences which can elicit a fear response (Owens and Beidel, 2015). In fact, a dynamically controlled environment is mostly unfeasible or unsafe during an *in vivo* simulation, *e.g.,* a classroom or a lecture room. Similarly, virtual training simulations require controllable environments to supply instructors with training scenarios and plausible environments which can elicit emotional responses from the trainee (Lugrin et al., 2016). Hence, fine control of the audience behavior is paramount for rooting the user in the virtual scene and providing training and therapeutic adaptive environments.

In order to become efficient, virtual training or therapeutic systems rely on the phenomenon of Presence which is known as the feeling of "being here" or to be the moment when "there is successful substitution of real sensory data by virtually generated sensory data" (Slater et al., 2009). For instance, it seems that when the feeling of Presence is achieved, an interactive virtual environment significantly improves learning effectiveness (Messinis et al., 2010). Yet, to feel immersed in a virtual environment a user needs to embody an avatar assuming that movement tracking, latencies, the field of view, audio, and haptic feedback are issued. The feeling of presence and the performances for different tasks in VR are enhanced by a low latency with head tracking and a wide stereoscopic field of view (Arthur et al., 1993; Hale and Stanney, 2006; Lee et al., 2010; Lugrin et al., 2013) as well as good body tracking (Cummings and Bailenson, 2016).

By extension, the feeling of social presence, defined as the "sense of being with another" (Biocca et al., 2003) has to be considered in social VR applications. Therefore, if virtual audiences provide interpersonal interaction as well as a sensory awareness of the agents or other users the feeling of "being with others" or co-presence can be elicited (Slater et al., 2000). Thus, by controlling the audience's attitude, therapy and training systems can provide a strong sense of co-presence. This has the effect of reinforcing the feeling of immersion for the VR user. If the aforementioned technical prerequisites are fulfilled and added to a better feeling of immersion and realistic interactions with the virtual environment such systems can significantly enhance not only performances related to VR tasks but also communications between users (Narayan et al., 2005). Recent studies on social VR which exploit rich social signals and behavior patterns explored how to leverage these VR requirements for the feeling of co-presence as well as the interactions and immersion by adding co-located agents and an embodied avatar for the user to interact with the virtual environment (Latoschik et al., 2019).

In this paper, we describe and discuss STAGE (Speaking To an Audience in a digGital Environment), a high-level control system constructed around a state-of-the-art virtual audience simulation. The STAGE allows to leverage the potential of co-presence in finely controlled and tutor-led training for public speaking, through the creation of pedagogical narratives: narratives encompass affective phenomenon, and rather than organizing events changing the course of a training scenario, pedagogical plans using our system focus on organizing the affects it arouses for the trainees.

1.1 Related Works

There are many possible applications for virtual audiences and they all share the same needs for believable agents despite the fact that these systems are meant to be used in various domains, *e.g.*, public speaking training (Pertaub et al., 2002; Chollet et al., 2014), and therapeutic (Kahlon et al., 2019) or educational applications (Lugrin et al., 2016; Fukuda et al., 2017; Delamarre et al., 2021; Lindner et al., 2021) as well from the industry [Ovation [1] (VRSpeaking, 2022), VirtualSpeech [2] (VirtualSpeech, 2022)].

1.1.1 Virtual Audiences Behavior Models

Virtual Audience systems are based on virtual agents' non-verbal behavior. It is the non-verbal behaviors of each agent that overall produce the virtual audience attitude perceived by the user (Kang et al., 2016; Chollet and Scherer, 2017; Glémarec et al., 2021). Therefore, different approaches to investigate nonverbal behavior perception are used and it is often due to the chosen behavior model.

For instance, cognitive models such as Pleasure-Arousal-Dominance (Heudin, 2007), Appraisal (Marsella and Gratch,

[1]Ovation Application, https://www.ovationvr.com/ [Accessed 29 March 2022].
[2]Virtual Speech Application, https://virtualspeech.com/ [Accessed 29 March 2022].

2002), or Valence-Arousal (Chollet et al., 2014) are used in many cases and have led to different implementations. A way to represent these models is that continuous models (Valence-Arousal, Pleasure-Arousal-Dominance) map an individual's emotional states along dimensions (Mehrabian, 1996) whereas discrete models describe fixed emotions like the basic emotions from (Ekman, 1999). While in appraisal theories, models state the importance of the evaluation and the interpretation of an event to explain an individual's emotions (Roseman, 1991).

Hence, Chollet and Scherer (2017) used crowd-sourcing to get large samples of users who designed the agents' behaviors themselves built on a Valence-Arousal model. Other models are built by analyzing video records to get a representative corpus and identifying patterns with a statistical approach (Kang et al., 2016), or with user evaluations and past results from the literature (Pelachaud, 2009; Fukuda et al., 2017; Hosseinpanah et al., 2018). Still, some systems are not based on a cognitive model but on domain experts' knowledge. These systems are often related to a specific context such as the behavior of a classroom (Lugrin et al., 2016; Kahlon et al., 2019; Delamarre et al., 2021).

As a result, all these models provide a set of non-verbal behaviors which are used to display various audience attitudes. These models include facial expressions, postures, head movements, and gaze patterns, from very limited sets (Kang et al., 2016; Fukuda et al., 2017) to wider models also taking into account interactions between each non-verbal behavior (Chollet and Scherer, 2017). However, it is often pointed out that the sole use of non-verbal behaviors is not enough to fully simulate human behavior (Glémarec et al., 2021) and that this task requires a wider variety of social cues. For instance, backchannels, defined as "non-intrusive acoustic and visual signals provided during the speaker's turn" by Yngve (1970), can convey the interest of a virtual agent given to a conversation or its opinion towards it (Bevacqua et al., 2010). A straightforward example of a backchannel is a head nod with a para verbal "mmhmm" signifying agreement (Bevacqua et al., 2010). Finally, other context-specific behavioral cues can be used in such systems but they are closely related to a context and used for training or therapeutic purposes, e.g., disruptive behavior in a classroom.

1.2 Virtual Audience Control

In cognitive-behavioral therapy, PSA is defined as a social anxiety disorder expressed by the fear of negative evaluation of others in social situations and feeling embarrassed or humiliated (American Psychiatric Association, 2013). The use of virtual agents as a media to mitigate PSA (Wallach et al., 2009; Anderson et al., 2013) or to improve public speaking skills (Batrinca et al., 2013) became commonly used with VR. The reason is that Virtual audiences can elicit stress or anxiety similar to a real audience and can be used in a training system Kelly et al. (2007). Therefore, virtual training and therapeutic systems are provided with advanced behavior models to elicit a user behavioral response during the simulation allowing successful therapeutic or training outcomes.

Yet, there are several challenges in the design of social skills training or PSA treatment systems including interactive virtual agents. A critical one is to control a virtual audience to follow a training plan, whilst allowing it to react to the user's behaviors and interactions. An important limitation of existing systems is the fact that they mostly relied on a Wizard of Oz approach to drive the audience in reaction to the user's performance (Chollet et al., 2014; Harris et al., 2002; Pertaub et al., 2002; Lugrin et al., 2016; Fukuda et al., 2017).

For instance, in *Breaking Bad Behaviors* (Lugrin et al., 2016) a classroom behavior management system, each virtual agent behavior is driven by an instructor who adapts it on the fly according to the trainee's actions. Such systems seem to elicit a heavy cognitive load for the instructors when it comes to following a classroom strategy and manually authoring each virtual agent (Mouw et al., 2020). Delamarre et al. (2021) proposes another classroom behavior management training system to drive the VA behavior according to scenarios designed by experts in classroom management. As a result, the VA is scripted and has no scenario flexibility, and might suffer from simulation repetitiveness, but it provides a high-level authoring tool for users without knowledge in scripting languages. Hence, VA systems such as *Breaking Bad Behaviors* relying on a tutor-in-the-loop could benefit from higher-level user control to manipulate the audiences like in Delamarre et al. (2021). This could benefit VR training systems, concerning the trade-off between a fully autonomous simulation and a Wizard of Oz system where each spectator would be individually controlled. For instance, when replacing tutor expertise with an autonomous component is not desirable, e.g., VR therapy and training could need real-time adjustments and temporarily fine control of the environment.

Our proposal consists of a novel pedagogical narratives control tool that aims at solving specific requirements for training or therapeutic VR systems. Our approach is to make use of a VA behavior model to create pedagogical narratives relying on the affect it arouses in the users. Unlike training scenarios, these narratives do not rely on a sequence of actions and choices that makes the scenario branching but rather focus on the affective experience: during the course of the presentation, the students' affects are modulated by the audience's attitude changes.

The contribution is twofold: we first describe how we used a user-centered development process to develop a VR training system for bachelor students and then how we solved the aforementioned trade-off between a fully autonomous and a Wizard of Oz system. In doing so, we extended an existing VAS with non-verbal behaviors, backchannel, and affective cues based on the instructor's feedback and provide a high-level control interface allowing the instructors to design pedagogical narratives *via* a high-level application programming interface (API). Our system and its novel development process provide insights into the successful integration of VR-based formative educational tools into an existing university curriculum.

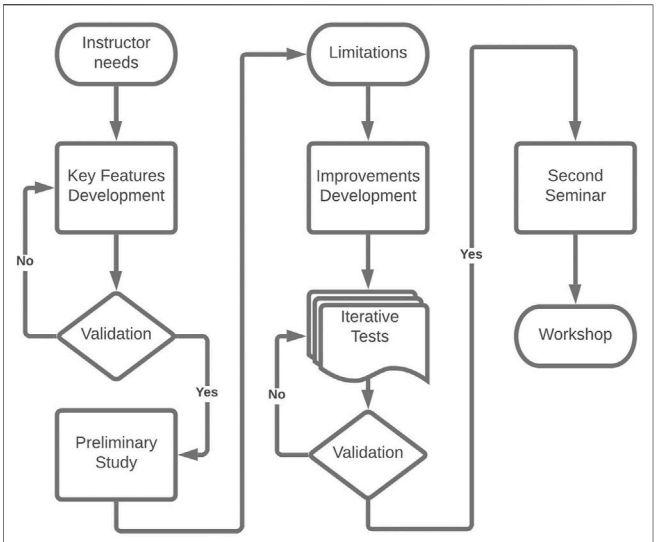

FIGURE 1 | The methodology used for the design of the system.

2 SYSTEM OVERVIEW

The VAS has been developed for *Scientific Writing and Presentation* seminars for postgraduate and undergraduate at the University of Würzburg in Germany. The system was used for two semesters, one in which volunteers participated in a preliminary study that helped us to develop the first VAS prototype and one in which all the students were able to practice in VR to prepare their final presentation. Before the VAS was integrated into the seminar, the students had no compulsory training and were mostly preparing their exams based on the lectures. According to the lecturers, only a minority of students were contacting the professors to get feedback on their presentations. Hence, we proposed the VAS as a VR training tool to let the students practice their public speaking skills, especially those which cannot be learned with online presentations like how to react to the audience behaviors, or how to use the space on stage. Thus, the VAS provided a learning tool that could be used to let the students be exposed to different situations they could experience during a real presentation.

The VAS was then designed to fit this seminar and provides both a safe learning environment for the students and a flexible educational tool. The training sessions were designed to give the students a chance to practice in front of a virtual audience with the professors watching it. On the one hand to help during the presentation and on the other hand to give a personalized review of the student's slides and presentation quality right after it.

In a desire to focus the system's development on the needs of the different users (i.e., the instructors and students), we first targeted the critical functionalities making it possible to provide a functional virtual training environment. Then we iteratively added different software improvements providing better control of the environment and the best experience for the students.

2.1 Development Methodology

In order to provide the most suitable system to the lecturers from the University of Wuerzburg, we followed a user-centered development driven by the lecturers in charge of the seminar. **Figure 1** shows the development process we followed.

Our first milestone was the identification of the critical features required by the instructors, based on their pedagogical needs. Hence, after breaking the instructor's and the student's tasks down we defined the list of features required for the seminar to happen in VR. The instructors needed to be able to listen to the presentation while watching the slides and the student's movements. As for the students they needed to be able to display their slides in VR, to have control over these slides with a remote controller, and to have feedback on the current state of their presentation, i.e., current slide displayed and time remaining. On top of these features, the system itself requires a plausible and believable virtual environment populated with a controllable virtual audience to expose the user to various public speaking situations. Moreover, the system had to allow the application to be used in mixed reality (MR) with a projected virtual audience with Kinect-based speaker tracking to accommodate students uncomfortable in VR and provide a more natural conference-like situation.

With these key features in mind, we started to develop a prototype with an iterative process in which each feature implemented was then tested and validated by the instructors. After the prototype was functional and validated we ran a preliminary study in which undergraduate students volunteers participated in training sessions for 8 weeks. The training session was structured as follows: the student was sending their slides in advance to test them before the presentation, then on the day of the training they had a training session in which they were able to test their slides in both VR and MR. After this training, the students had to choose between VR and MR and get ready for the presentation (**Figure 2**). The presentation was 10 min long with questions from the instructor at the end. It was followed by a semi-structured interview and a briefing between the students and the instructor who gives feedback about the slides' quality, the presentation content, and the public speaking skills.

From this first preliminary study, we gathered the first students' impressions of the system. The instructors also provided a list of additional requirements from their use of the system during this first seminar, namely regarding the controls and the cognitive load when it comes to both following the presentation and handling the VA. Thus, a second iterative process started with Ph.D. students volunteers, and instructors to test each improvement requested. Ph.D. Students were rehearsing their presentations for incoming research meetings and were able to provide further feedback for each iteration. Some other Ph.D. students were also asked to test specific aspects of the system such as the slides controls or the training instructions. Regarding the improvements made to the VAS in terms of audience behavior and controls, the instructors asked for new attitudes and behaviors as well as a new control interface to widen the possibilities for designing pedagogical narratives.

After the instructors validated the second prototype, a second seminar used the training system to let students practice VR before their final exam. In parallel, a workshop with lecturers was

FIGURE 2 | The three possible roles during the seminar: **(A)** Embodied Teacher with controls, **(B)** the virtual reality speaker, and **(C)** the mixed reality speaker.

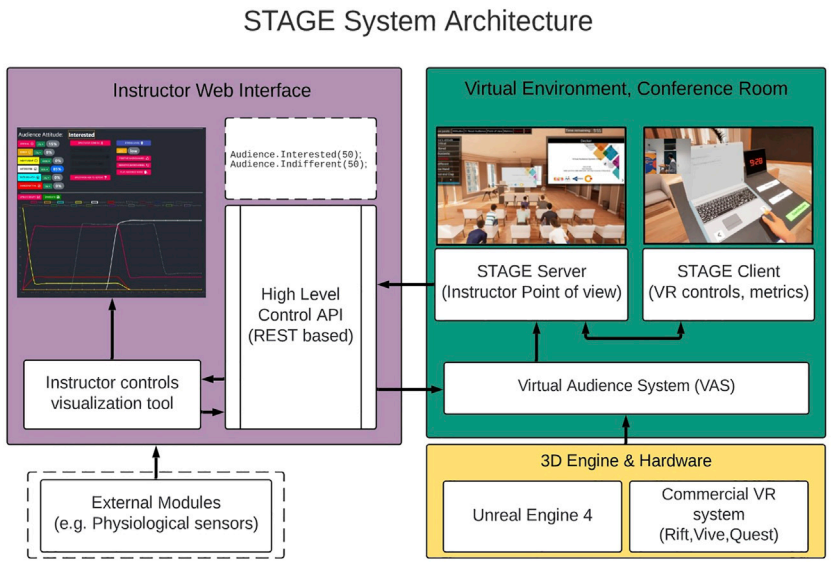

FIGURE 3 | Speaking To an Audience in a diGital virtual Environment (STAGE) system architecture.

organized to get more insight into its possible use in subsequent lectures and seminars.

2.2 Application Architecture

The prototype was based on the pedagogical needs of the instructors, namely: being able to listen to the student's presentation, watch the slides, to be able to observe his movements provided that the presentation takes place in a believable scientific conference environment.

The prototype was developed with *Unreal Engine 4* [3] (Epic Games, 2022) as a VR application for students and as a desktop application for the instructors (**Figure 3**). This application follows a client-server architecture where the client is here only responsible for the controls from the student, and the server is responsible for the virtual audience attitude and the instructor controls. This network architecture allows instructors to attend the presentation remotely: for instance during the first seminar students and instructors were in two different rooms but connected *via* the university network.

In order to meet the students' requirements, we created a first virtual environment allowing students to control their slides. To do so, we used Decker [4] (Latoschik et al., 2022), an open-source slide creation tool based on the *Markdown* language which is interpreted into *HTML* in a web browser. This tool is in use within seven German universities and is often used by lecturers at the University of Wuerzburg. Decker was already used in the seminar and presentation templates were given to the students. Thus, we created an interaction metaphor with the slides as naturally as possible. As visible in **Figure 2**, we implemented a virtual remote slide presenter with a laser pointer, appearing in the user's virtual hands, and simply controlled using the VR controllers buttons and thumb-sticks.

[3]Unreal Engine, https://www.unrealengine.com/en-US/ [Accessed 29 March 2022].

[4]Decker sources repository, University of Wuerzburg https://gitlab2.informatik.uni-wuerzburg.de/decker/decker,[Accessed 29 March 2022].

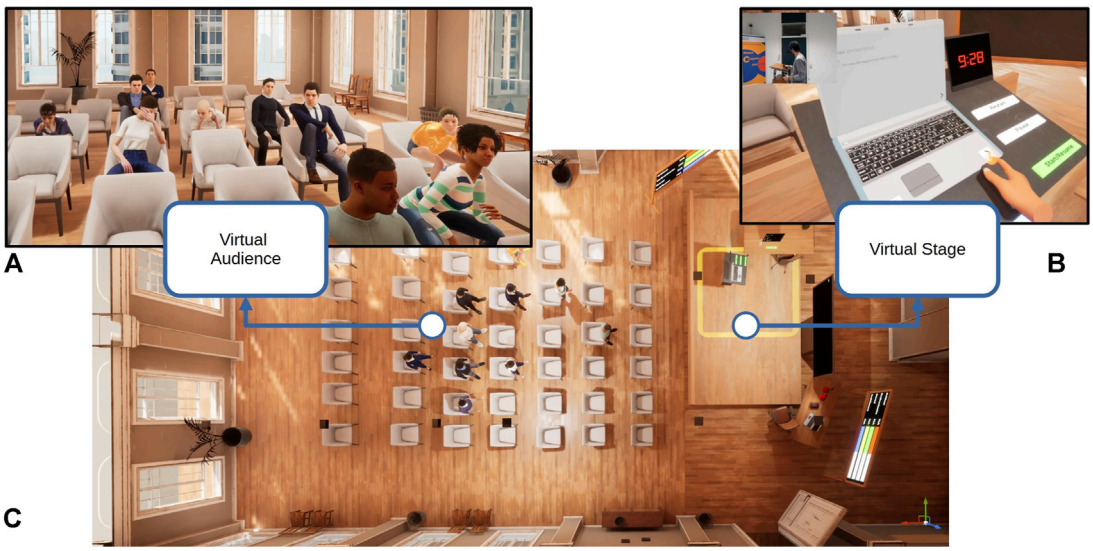

FIGURE 4 | System Overview, from a top-down perspective: With **(A)** the virtual audience, **(B)** the virtual stage with the laptop and the student menu, **(C)** the virtual conference room from a top-down perspective.

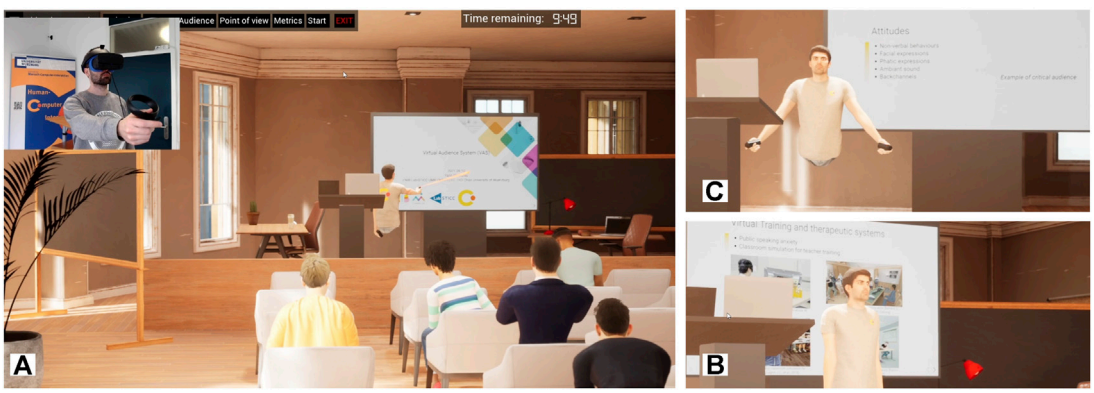

FIGURE 5 | Example of scanned avatar embodiment for the speaker **(A)** with the point of view from virtual spectators **(B)** and **(C)**.

Students could therefore use the controllers to both interact with slides by clicking on them and highlight certain elements of their slides with the laser. The controller buttons were also used to move to the next or previous slides. The slides were displayed *via* two web browsers, on a large panel that represents the projected slides and also on a laptop which allows the user to have feedback on the current slide while facing the VA. To keep track of the time spent during the presentation a timer displays the remaining time as soon as they start the presentation by pressing a button next to the virtual laptop (**Figure 4**). Some presentation-quality metrics were possible to visualize and export for the students such as the percentage of time looking at the audience, the time on each slide, which agents the user looks at the most, or the time talking. Besides the slides and laser pointer interactions, the user could embody an avatar composed of two virtual hands holding the laser controllers, a head-mounted display, and transparent footprints on the ground to locate the user. The head-mounted display was not visible from the student presenter's point of view. The hands were therefore animated and moving according to the capacitive sensors of the VR controllers which provide the location of the thumb and the index of the student. To improve the menu buttons' usability, we created a press interaction with the user's hands which when the student's hand gets closer to a button triggers a visual cue and a hand animation to encourage the user to press it with the index. Then when the student press the button, it triggers a visual effect as well as a smooth vibration in the controller to notify the student of the ongoing interaction. Finally, some objects in the virtual environment can be grabbed to increase the interactivity with the environment.

The STAGE can also use scanned avatars (**Figure 5**) which allows the users to embody their photo-scanned avatar using

FIGURE 6 | Instructor Graphical User Interface with controls and slides preview **(A)** and embodied virtual spectator by the instructor for the end questions **(B)**.

inverse kinematic to partially track the body movements based on the head and the controllers' location (this feature has not been used during the seminars yet).

The audience was also a critical prerequisite for the instructors. As stated above, the usefulness of VR training simulations lies in their ability to expose users to particular situations while controlling the degree of exposure, in our case the virtual audience and the behavior of virtual spectators who compose it. In the first prototype, the VA was entirely based on an existing non-verbal behavior model from (Glémarec et al., 2021) which made it possible to generate four audience attitudes, *i.e.,* bored, enthusiastic, indifferent, critical. Through the iterative development process used we extended this model with further attitudes, new context related behaviors and social interactions such as backchannels. A small set of behavioral cues was also added in order to support the narratives, *e.g,*. spectator leaving or coming in the room.

Finally, a graphical user interface (GUI) was added to extend the instructor desktop application to let them have high-level controls on the virtual audience's attitude (**Figure 6**). This GUI also lets the instructors make use of a live question system where they can embody a virtual spectator to raise a hand and talk through a microphone. It also shows the slides and has a camera system to get different points of view of the virtual environment, *e.g.*, from the back of the room, from the front row, or the stage. With the instructors' feedback given after the preliminary study we extended the desktop application with a web graphical user interface providing a high-level control API used for both controlling the VA at run-time, and pre-scripting fixed pedagogical narratives beforehand. This second web GUI is accompanied by a visualization tool to keep track of the ongoing narrative.

The training session was structured as follows: the students were sending their slides made with Decker a bit in advance to test them before the presentation, then they had a training session in which they were able to test their slides in both VR and MR. After this training, the student had to choose between VR and MR and get ready for the presentation. The presentation was 10 min long with questions from the instructor at the end. It was followed by a semi-structured qualitative interview and a briefing between the student and the instructor who gives feedback about the slides' quality, the presentation content, and the public speaking skills.

The first preliminary results led us to develop better interaction techniques, extend the VAS, and add a high-level GUI allowing us to launch previously established pedagogical narratives while providing graphical feedback on the state of the audience. Hence, all of the developments described in the following sections were made based on successive iterations. Following these developments, a second seminar took place as well as a workshop with university professors and lecturers who tried out the system. Thus, the following sections first introduce the VAS that we have extended and the behavioral cues requested by the instructors to let them create more suitable audiences for the narratives (**section 3**). We then describe the instructor interface allowing the design of pedagogical narratives with a high-level API which grants a fine control of the VA: this GUI also provides a visualization of the narratives and by extension of the current state of the virtual audience (**section 4**) Finally, all the feedback from students and lecturers who participated in the second seminar and a workshop are given and discussed to provide guidelines on the development of a similar training system for university curriculum (**section 5**).

3 VIRTUAL AUDIENCE SYSTEM

From the instructor's point of view, the VA is a teaching aid by which the student can experience simulated scientific talk. This implies a believable VA in terms of the audience's behaviors and reactions toward the presentation. The challenge of such a system is to provide an audience whose behavior allows the speaker to perceive its attitude. This has the effect of arousing affects in the users, either positive or not. The aim is to supply the student with an environment that will provide the best possible experience of a scientific talk.

3.1 Virtual Audience Implementation

In order to create these believable virtual audiences which could be used to populate a virtual seminar environment, we integrated an audience behavior model based on a Valence-Arousal cognitive model which was evaluated in VR (Glémarec et al.,

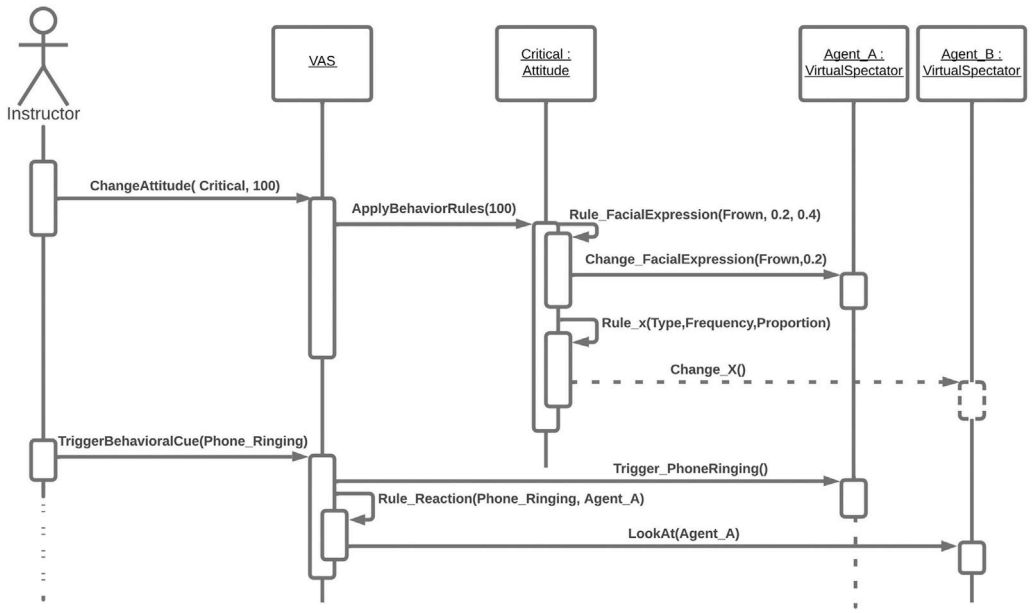

FIGURE 7 | Sequence diagram describing an attitude change into a critical one, with an example behavioral cue triggered with its reaction.

2021). This implementation was integrated by including an open-source *Unreal Engine* plugin [5] (Glémarec et al., 2022). The model provides sets of non-verbal behaviors to use to display an audience attitude the VR users can perceive, *e.g.,* critical, interested, bored. The non-verbal behavior sets include facial expressions, postures, head movements, gaze directions, and to what frequencies they should be displayed. Therefore, the model provides rules to generate audience attitudes according to how the users would perceive this non-verbal behavior in terms of opinion and engagement toward the speaker or the presentation, with the opinion being related to the valence and the engagement with the arousal. This rule-based model allows us to easily adjust the virtual agent behavior to the desired audience attitude. Hence, an attitude is a sum of non-verbal behavior rules which change the virtual agents' behavior over time, with the following format (**Eq. 1**):

$$Attitude = \sum rule_x\left(Type, Frequency, Proportion\right) \quad (1)$$

Where x is the nonverbal behavior category of the rule (e.g., posture, gaze), *Type* a pre-defined parameter characterizing the nonverbal behavior in the category, *Frequency* how often the behavior is displayed for each active agent, and *Proportion* the number of agents in the audience which will be actively displaying the behavior. An example of a rule would be (**Eq. 2**):

$$rule_{facialExpression}\left(Frown, 0.1, 0.2\right) \quad (2)$$

This can be read as 20% of the agents frown 10% of a given period. The implementation of this rule-based system allows us to directly manipulate the rules and extend existing ones or even create new ones to add new attitudes based on the experts' knowledge. **Figure 7** shows the logic of the VAS implementation.

Moreover, because the perceived attitude is composed of various behaviors, sometimes opposed in terms of opinion and engagement toward the speaker, a virtual audience might be displaying a mixed attitude, e.g., some bored spectators and few others interested. It is then the dominant attitude that is perceived by the user (Chollet and Scherer, 2017; Glémarec et al., 2021), *e.g.,* if 80% of the agents are displaying a critical attitude and 20% an enthusiastic one, the user is significantly more likely to perceive the overall audience's attitude as critical. Therefore, the model implementation allows us to design the most suitable audiences for the training plan thanks to its rules-based system and allows for instance to smoothly transit from a Critical to an Interested attitude by progressively decreasing the number of critical agents and increasing the number of interested ones.

3.2 Behavioral Cues

The previously described model does not include backchannels or behavioral cues because it was not related to a specific context and is only relying on non-verbal behaviors. Thus, the model had to be extended with specific behaviors to fit our context. Based on the feedback from the first seminar, the instructors gave us a list of behaviors they needed in their pedagogical narratives to design the audience's attitudes. They designed two narratives that use variations in the displayed attitude to let the students experiment with different types of audiences and different phases of a presentation that could happen in real life, *e.g.,* an interested audience at the beginning of the talk which get bored and

[5]Virtual audience project page, Chair of Human-Computer-Interaction University of Wuerzburg, http://hci.uni-wuerzburg.de/projects/virtual-audiences/[Accessed 28 January 2022].

indifferent and that finally becomes critical because the spectators did not appreciate the presentation. These narratives were also punctuated with contextual affective cues used as disturbing or supportive events during the presentation such as a spectator coming into the room during the presentation, someone yawning loudly, or playing a supportive backchannel. The two narratives were very different, one was meant to be supportive and less stressful as possible while the other one was meant to be challenging with regular disruptive behaviors and a majority of negative attitudes displayed. The main purpose of such narratives is to let the students face the different situations they cannot experience training individually and allow them to get personalized feedback from the professors.

Therefore, with the intent of providing plausible narratives to students to let them face these public speaking situations, we added several behaviors. To be specific we first extended the number of postures and variations of the existing model implementation to get rid of the looping behaviors which were often spotted by the students. Posture variations are changes in some parts of the body like crossing the legs differently or resting on the opposite hand. Then, because the instructors felt limited with the four attitudes from the model used we created new ones with the same rule-based system. One guideline for creating new attitudes was to use attitude-related behaviors to let the users perceive the difference when two attitudes are close in terms of perceived valence and arousal (Glémarec et al., 2021), e.g., to differentiate a bored audience from an indifferent one. The new model was then extended with two new attitudes, interested and disrespectful. The interested attitude is defined by the model with a high level of arousal and positive valence, i.e., with a positive opinion and engagement toward the speaker. Hence based on the model rules this attitude triggers frequent nodding and smiles with virtual agents leaning forward and mostly staring at the speaker or the slides. As for the specific behaviors related to the interested attitudes, we added two behaviors that were representative according to the instructor. The supplemental behaviors were taking notes and leaning sideways to look at the slide when the sight is obstructed. Conversely, the disrespectful attitude displays less frequent head movements and facial expressions while the virtual agents are leaning backward. In this case, the specific behaviors were agents texting, chatting together, or putting their arms behind the head. The new set of attitudes now includes around seventy postures, four different head movements, and four different facial expressions including specific behaviors like yawning for the bored attitude, texting for the disrespectful attitude, or taking notes for the interested one. These new animations were created from motion capture data which were then applied to the different virtual agents so that all of them can display the new behaviors.

Then to support the pedagogical narratives and improve the overall audience believability we added some affective cues and social interactions also made with motion capture, e.g., a phone ringing. The first prototype was already including moving spectators and whispering. Along with the specific behaviors like yawning or texting, we added contextual behaviors in which spectators were asked to repeat, with German voice lines depending on the virtual agent's gender or a phone ringing followed by apologies from the virtual agents. As for the social interactions, we added reactions to these contextual behaviors or the attitude-related ones, based on a proxemic awareness of what is happening around them. For instance when a phone rings the surrounding agents will look at the virtual spectator trying to switch off its phone. It works the same with spectators coming in late, the virtual agents close the newcomer stare at it because they are distracted.

FIGURE 8 | Example of audience reaction: when a new virtual spectator enters the conference room.

If the affective cues are periodic either manually triggered by the instructor or automatically by the narrative, these reactions are conditioned and can rely on pre-established rules such as the distance between virtual agents, the current attitude displayed, or some user metrics. These utility-based rules were implemented from audiences' accounts in which the instructors precisely described what type of behavior happens when they occur and in which circumstances. Consequently, all these reactions are based on heuristics from the instructors' knowledge of audiences' behaviors. For instance, in the situation where the virtual spectator's phone rings some others can look at it and frown if there are displaying an interesting attitude and might even whisper to it to switch it off if the user looks at them (**Figure 8**).

In order to improve the interactions between the student presenting and the VA, we added backchannels which increase the VA's engagement with the talk. However, backchannels are often used in conversation, and in this situation, there is only the presenter who is talking and a group of agents which are listening. Moreover, the system is not capable to analyze the content of the presentation and cannot guess when and how to interact. Hence, two types of backchannel that do not involve analyzing the talk were added, one supportive and one negative. The supportive ones notify the user that an agent better understands what is currently being said with the agent nodding and emitting a long "mmh", while the negative ones were notifying a miss-understanding with the agent frowning and emitting a specific negative or even rude backchannel specific to the German language that could be compared to long "what" in English. As for the moment when to trigger these backchannels we based them on heuristics as well. A utility-based function uses the user metrics gathered during the

TABLE 1 | STAGE Scalability Performances measured over a 2-min long period.

Number of virtual agents	Number of Frames per seconds (FPS)	Game Thread (ms)	Rendering thread (ms)	GPU thread (ms)
13	83	$\mu = 11$, $\sigma = 0, 2$	$\mu = 6.8$, $\sigma = 0.3$	$\mu = 12$, $\sigma = 0.2$
26	41	$\mu = 24$, $\sigma = 1.1$	$\mu = 12$, $\sigma = 0.9$	$\mu = 25$, $\sigma = 0.3$
36 (full room)	23	$\mu = 32$, $\sigma = 3.4$	$\mu = 24$, $\sigma = 1.4$	$\mu = 42$, $\sigma = 5.0$

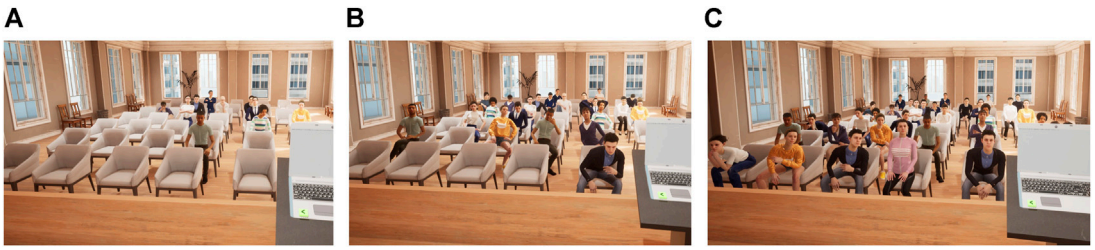

FIGURE 9 | The three virtual audiences were used to run the scalability performance evaluation: **(A)** the 13 agents used during the seminar, **(B)** 26 agents, and **(C)** 36 agents which correspond to a full conference room.

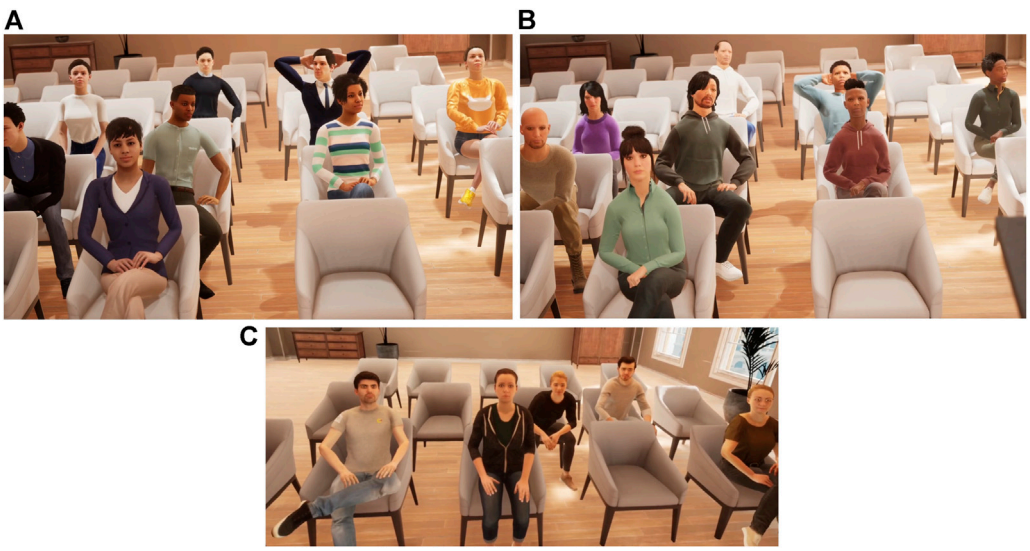

FIGURE 10 | Three different types of virtual agents used within the STAGE: **(A)** the characters used during the seminar from Mixamo®, **(B)** the Meta-Humans from Epic®Games, and **(C)** Photo-scanned avatars from the University of Wuerzburg.

presentation to decide when the backchannel should be supportive or not. So, for instance, when the speaker has a regular pace and the student is often looking at the audience it is more likely that the audience will trigger a supportive backchannel while if the student always looks at the notes and needs to go back to previous slides it is even more likely that a negative backchannel is triggered. Finally, to avoid long absences of noise coming from the virtual audience we added some noisy behaviors which do not affect any pedagogical plan or attitudes, *e.g.*, a virtual spectator picking up a pen, or others who cough.

All these behaviors were based on the instructors' requirements and the feedback from the first seminar was each time evaluated by Ph.D. students during their training through structured interviews and then approved by the instructors. Thus, we included new behavior rules similarly to **Eq. 2**. Similarly, as for the backchannels, the social interactions and the reactions were both only triggered by the instructors, the narrative, or autonomously activated by utility functions we designed from heuristics.

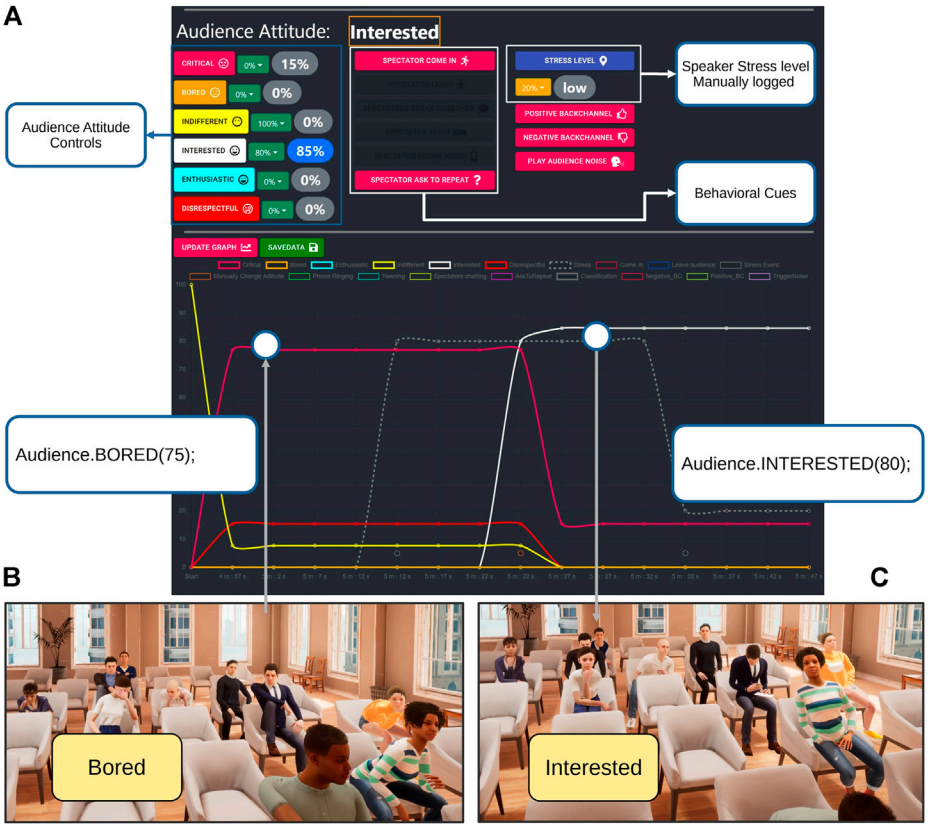

FIGURE 11 | Example of manual change of the virtual audience attitude: with **(A)** the initially bored audience, **(B)** the instructor web interface with the controls, and **(C)** the resulting interested audience obtained with a high-level instruction.

3.3 Performances and Scalability

In the STAGE the virtual audience is composed of 13 different virtual characters from *Adobe Mixamo* [6] (Adobe Systems, 2022). During the seminar the student were given an *Oculus®Rift S* with a constant 80 *Hz* refresh rate which is bounded to the hardware. However, to provide the most plausible environment such a virtual conference room should be able to issue larger crowds. Thus, we evaluated the STAGE performances and ran a scalability benchmark. We first measured the VAS performances within the seminar setup over 5 min and as expected the system VA behavior model implementation itself does not have a huge impact on the performance (μ = 1.68 ms, σ = 0.2). To perform the evaluation we used a computer running with Windows 10 64 bits, Intel®Core™i7-9700K central processing unit (CPU) at 3.60GHz, and NVIDIA®GeForce RTX 2080 SUPER Graphics Processing Unit (GPU) 8 GB GDDR5. An Oculus®Rift S was used to carry out the VR evaluation.

Nonetheless, a load test shows that the STAGE is quickly GPU bounded due to the virtual agent rendering (**Table 1**). Further investigations have shown that shaders used to render some agents' hair had a significant impact on the frame rate, it is a common issue that the use of some translucency is harmful to VR performances. With regards to the number of agents and the resulting frame rate, the system can handle 19 virtual agents from Mixamo. As an example, if we double the number of virtual agents we are already under 45 frames per second which is not suitable enough for VR uses (**Figure 9**).

Even if our system already uses different levels of details for the virtual agents' meshes and props in the environment, these measures testify to a need for a VA with a mix of highly detailed characters and less detailed ones. For instance, a mix between photo-scanned avatars, the Meta-Humans from Epic®Games, the ones we are using from Mixamo, and others using a simplified mesh structure could solve such limitations: with the most detailed virtual agents close to the speaker for the facial expressions to be visible so that the further the agent is the less detailed it is. The STAGE system can already be used with these different virtual agents (**Figure 10**).

4 STAGE CONTROL INTERFACE

After the first prototype was evaluated, instructors reported limitations due to the complexity of manually and

[6]Mixamo character and animation library, https://www.mixamo.com/#/Accessed 12 February 2022.

continuously controlling the virtual audience, even though the instructor only had to change the overall attitude or trigger specific behavior accordingly to the presentation. Hence we proposed to design a web interface in which they could monitor pedagogical narratives. In doing so they only have to listen to the presentation and monitor the ongoing narrative or eventually adopt the VA's attitude if the current training session does not fit with the pre-established narrative.

Therefore, we designed a graphical user interface allowing us to both control the VA by changing the attitude and triggering specific behavior like backchannel or context-related behaviors, and by controlling and monitoring the ongoing narrative. Thus, we created a web GUI based on the REACT framework using a REST API to communicate with the STAGE application. Regarding the design of the narrative, we implemented a high-level control API that can directly be used to control the VAS.

4.1 Audience Controls

If the instructors require a high-level control interface, they also needed to keep fine control over the audience to adapt it. The application has a central component that is responsible for providing a simple control interface with the behavior model and the different rules. Thanks to the attitude model, instructors can directly change the attitudes in percentages without taking care of each agent individually. This component simply provides direct access to the VAS. To do so we developed a high-level API to drive the virtual audience with simple instructions. **Figure 11** shows how simple instructions can be used to quickly adapt the audience to the presentation. However, the VAS cannot always follow the instructions and tries to get as close as possible to it. For instance, two agents speaking together can only be displayed under certain conditions, they have to be next to each other and close enough.

With the heavy cognitive load, such controls can elicit the GUI had to ease its use. Hence, based on the instructor's audience accounts we linked the different behavioral cues to each attitude. It allows us to dynamically adapt the displayed buttons of each specific behavior so that only those linked to the current attitude can be used. This nudges the instructors to only used attitude-related behaviors. Aside from the audience controls the interface provides some controls over the virtual environment such as the student's timer, a reset of the slide, or a logging system for the instructor to add information within the visualization tool, e.g., the speaker perceived stress.

4.2 Virtual Audience Control API

The aforementioned high-level API is also used to provide the instructors with a simple narrative editing tool. In addition to it, a state machine is provided to let the instructors use successive states containing the high-level instructions adapting the VA attitude and behavioral cues. In doing so the audience's attitude can be timed or conditioned to events, user metrics, or even external tools which can communicate through the REST API such as physiological sensors.

Instructors have tested physiological driven narratives in which the VA attitude was changed according to the student data obtained through an *Empatica E4* [7] wristband (Empatica, 2022), however, this pedagogical narrative was never used in the seminar to avoid any unfairness between students since none of these narratives have been evaluated. For the same reason, none of the narratives were linear during the seminar and were not branching to create alternative ones, so that all the students had the same narrative. Nonetheless, such features could provide major pedagogical help in terms of stress monitoring. Some students might suffer from fear of public speaking and could benefit from training sessions taking into account their stress where the audience could change its attitude to lower the students' stress. Moreover, branching narratives could lead to adaptive narratives and personalized sessions. Such narratives could already be used with the control API we provide by conditioning the audience behavior on user behaviors or physiological data (Snippet 1).

Listing 1. Example of a simple training plan using the Virtual Audience Control API in Javascript? for training on maintaining the visual contact with the audience.

```
OnStartNarrative(){
    this.Audience.Interested(60);
    this.Audience.Enthusiastic(40);
}

NarrativeLoop(float deltatime){
   if(this.Intructor.SpeakerEstimatedStressLevel>75){
   {
        // try to calm down speaker if too stressed
        this.Audience.Interested(60);
        this.Audience.Enthusiastic(40);
   }else{
        if(this.Speaker.TimeLookingAtSlides>50){
            this.Audience.Bored(60);
            this.Audience.Indifferent(30);
        }
        else {
            this.Audience.Interested(70);
            this.Audience.Indifferent(30);
        }
}

void OnEndNarrative(){
    if(this.Speacker.TimeLookingAtSlides>50) {
        this.Audience.Applaude(90);
    }
    else {
        this.Audience.Leave(90); // delay in seconds
    }
}
```

Again, to ease the use of the STAGE system the VA affective modulation automatically begins when the student press the *Start* button on the virtual laptop menu. Thus, it helps the instructors to directly focus on the starting presentation without over-monitoring the training settings.

[7]Empatica E4 wristband technical page, https://www.empatica.com/en-eu/research/e4/Accessed 12 February 2022.

4.3 Visualization Module

Despite being autonomous, the narratives and the affective modulations of the audience still need to be monitored by the instructor to be adapted to the presentation. Hence, we used the control API to log each attitude and behavioral cues changes in the ongoing narrative and draw it on a graph continuously updated. The different attitudes are drawn with different curves based on the percentage of affected agents. The behavioral cues or events are represented by colored circles (**Figure 11**). These data can then be saved under CSV format.

Such training data could be used for post-training briefing or to replay the presentation with the students to provide a formative evaluation of the presentation. Moreover, it could feed a performance analysis system provisioned with the simulation data and the user metrics or physiological data to provide a qualitative report about the presentation, *e.g.*, students could see when they were stressed and on what slide they were, but could also know how much time they spend per slides or when did they needed to look at their notes.

5 PRELIMINARY USER STUDY

In this section, we provide the feedback from the 16 students who participated in the seminars that we gathered from questionnaires and semi-structured interviews. We exclude the PhD students who participated in the iterative tests. We also report a review of the STAGE made by the lecturers in charge of the seminar and three researchers in computer science.

5.1 Methods

During the two seminars, we had the opportunity to gather the students' feedback about different aspects of the STAGE, namely the system acceptability, its usability and the VA believability.

To do so, in the first seminar, we mainly focused on semi-structured interviews since we needed specific details that can be harder to get with a questionnaire, for instance with the slides interactions or the simulation controls with the instructors. The second seminar was less about the system development but more about evaluating it. Thus, we were able to put the main items from the first set of interviews into questionnaires. Hence, we added questionnaires focusing on the acceptance and usability of the system. Along with these questionnaires, we added a public speaking anxiety scale (Bartholomay and Houlihan, 2016) to measure the students' public speaking anxiety (PSA) and compare it to their self-estimated stress and performances. In both seminars, students had a short training in VR to get used to the controls and the virtual environment, then they had the presentation, followed by the briefing with the instructor, and finally they had the questionnaires and a semi-structured interview. We chose not to give any questionnaires before the presentation to let the students keep their focus on the training since it was part of a lecture and not only a study.

Regarding the workshop, all participants had the opportunity to play both roles, i.e., the student role in VR and the instructor role. They were first doing a presentation of their own, for instance, a lecture or a scientific presentation, and then evaluating the presentation of their colleagues. Hence, they provided feedback on all the different aspects of the system. The discussion following the trial was two-part, with a conversation between the participants and then smaller guided discussions based on the different aspects we wanted to explore, such as the usability of the controls, the virtual audience behavior believability, or the system acceptance for further uses in other seminars.

5.2 Results

5.2.1 Students Feedback

Regarding these three lines of research, we got promising results. All students agreed in our questionnaires and interview that the STAGE could help improve their presentation skills and also agreed on the usefulness of such VR training system at the university, *e.g.*, *"[it could help] to get more confident with the presentation itself"*, *"I noticed where I had problems in finding words"*, *"Especially in times of Covid, it makes practicing easy"*. Regarding the feeling of engagement during the presentation the results are mixed and 50% of them still believe a real audience is much more engaging for a training session. However, they all agreed that using the STAGE for a practice session is "funnier" compare to what they usually do. In fact, 50% of the students declared practicing their presentation alone, the others prepare notes or ask other students to help them. Some comments highlight the reason why students agreed on it and it is probably due to the narratives instructors designed, *e.g.*,, *"I feel like it can be helpful to practice with distraction sounds, although they were very surprising when I first noticed," "It felt almost like a real experience and it helped me a lot during the presentation because I could notice how the audience was behaving and I could adapt a bit the way of presenting."*. With respect to the system usability, a frequent comment is the difficulty to read some figures or slides on the small laptop's screen especially when the color contrast is weak.

For the PSA score we obtained with a questionnaire, we were able to first identify students stressed about their presentation and who might also suffer from PSA while we were interested to know if there were a possible correlation. These preliminary results seems to show a correlation between the public speaking anxiety score and the reported stress during the presentation. We ran a correlation test on the students' PSA and the self-estimated stress from the second seminar, but we removed the students who had issues with there slides or with the VR application which might have induce some stress, e.g., video not playing in the slides or tracking issues that implied a restart of the VR device. Since we have a small sample with ties and a distribution which does not follow a normal one we used a Kendall's Tau correlation test and adjusted the p-value when ties occurred. Thus, the PSA and the self-estimated stress seem positively correlated (Kendall's $\tau_c = 0.796$, $p - value = 0.048$). However, these results are preliminary ones and only include eight students

from the second seminar. Moreover, it is worth mentioning that none of the students declared not being stressed but at least stressed "as normal" which correspond to the middle value in the provided scale. Moreover, it does not seem that the self-estimated performance is linked to the anxiety level measured with the questionnaire.

Finally, regarding the virtual environment, all students agreed it was a believable environment. Concerning the VA their behavior and their reactions were considered believable for a conference, *e.g.*, *"In comparison to a real audience at a conference or a similar event the virtual audience was probably very realistic," "Looked a little bored at the end, I think this could also be in reality"*. As for the impact on the presentation, 70% stated the audience behavior impacted their behavior, *e.g.*, *"I felt shortly distracted when a phone in the audience rang," "I looked more towards the audience and pointed out details."*, *"It made me feel a bit unsure about how my presentation was going when people were leaving the room. A ringing phone also made me lose focus for a bit."*. However, only 50% declared adapting their presentation to the VA, *e.g.,"I tried to refer to them directly for example as "all of you," which I probably would not have done if I was talking to just one person"*. Eventually, almost all students were able to recognize the audience attitude displayed and remember after the presentation when a specific attitude was displayed according to our questionnaires. They all remembered that the audience started interested and then became bored, only one student did not remember any specific attitudes.

We believe it is worth mentioning that a student got a very high score of public speaking anxiety (75/85) and stated not having paid attention to the virtual audience at all moreover, the student declared being stressed by the fact that real persons were listening to the presentation, i.e., the instructor. This student also stated to be disturbed by all the noises coming from the VA. Knowing that the PSA seems to be considered as a subgroup of social anxiety disorders in the literature (Blöte et al., 2009), it might be interesting when using such systems to detect students who might suffer from it. Adapting the narrative to them and thus providing a less stressful training session could be a solution, either with specific narratives or with dynamical ones adjusting the VA attitude to measure the anxiety with physiological sensors. Yet, such a hypothesis would need further investigations.

5.2.2 Virtual Audience Believability

The comments from the workshop's participants regarding the VA believability seem similar to the students: the different attitudes are noticeable and the different behavioral cues are even more noticeable. However, the virtual audience needs a better audio system with more sounds from it. It seems to have a lack of *"ambient noise,"* e.g., when a virtual agent changes its posture, its chair should sometimes creak. The room in which the seminar was running might also play a role in it. The room produces an echo when the students talk and the fans from the computer can be heard while the sounds from the virtual environment are played on the HMD speakers and seem to be easily covered. Students too reported this issue despite the sounds being spatialized, *"One noise, I could not identify what it was supposed to be. The noise being directly in your ear makes it seem a bit unrealistic"*. A solution for this would be to use headphones that do not cover the student's voice or speakers in the seminar room which would play the sound coming from the VA.

A proposition was to improve the narrative with agents displaying a certain *"personality"*, meaning that instead of letting the model freely change the virtual agent's behavior, it would take into account its past behaviors. The virtual agents could avoid displaying an opposite attitude or only display specific behavior, e.g., an agent with a disrespectful attitude would not suddenly become interested.

With respect to the feeling of social presence, participants from the workshop proposed to add some VR interaction with a human embodying an avatar before the beginning of the presentation, for instance, the training session in VR could be held with the student and instructors embodying their avatar, who explain the controls and directly show how to use them in the virtual environment. Such rich interaction between co-located agents and embodied avatar seems to reinforce the feeling of co-presence as well as the possibility for interaction with the virtual environment (Latoschik et al., 2019).

5.2.3 Stage Control Interface

As for the control interface, the visualization graph, as well as the user metrics logs, seems to be of great interest when it comes to looking at the students' performances afterward or using it as a replay tool. Hence, instructors could cross the narrative and the metrics to provide even more personalized feedback. Such metrics visualization would also be a first step for the system to be used alone by the students without the instructors. For instance, it would allow the students to watch their presentation with a quantitative assessment of their performance, e.g., with the time spent per slide and how long they looked at their notes.

Nonetheless, there are some areas of improvement in terms of usability: reading the graph while trying to stay focused on the presentation is to complicated at some point as well as reading the current percentage of agents displaying a specific attitude, *e.g.,* a pie chart might have been more suitable to read the audience attitude. As for the rest of the GUI, due to the iterative tests we ran to prepare for the seminar, almost everything was automated. So that the instructors could focus on the presentation and not on starting the narrative or on manually changing the attitude.

To follow with the narrative, the high-level API seems promising, it provides high-level instructions to design simple pedagogical narratives by modulating the VA's affective cues. Still, a graphical representation of the state machine would ease the design of states, at least to see the following state and the transition, similarly as in a graph. However, it can be used to author the VA without being bound to a specific system with both high-level controls and direct changes on the behaviors, provided to have some knowledge in computer science.

5.2.4 Integration in University Curriculum

Participant all agreed on the potential the STAGE represent in terms of ecological environment for a formative evaluation. Such VR training system like the *Breaking Bad Behaviors* system are used to practice classroom management skill through successive

training sessions either by using the system or by watching your peers practicing (Lugrin et al., 2016). Lecturers participants to the workshop recommended to let the students practice more than once in VR similarly as in *Breaking Bad Behaviors*. For example, according to the lecturers the students who do not remember their slides keep reading at it and look less at the audience. Hence, having multiple training sessions with specific focus could improve the training process, e.g., a first session could just be dedicated to the slides without VR while the following could use the STAGE to focus on the public speaking skills. In addition to such repeated training session, peer review sessions could be organized in which students could help each others improving their presentation.

The STAGE could also be used during hybrid sessions in which other students could join the presentation and embody a virtual spectator. This feature is already existing in the STAGE but would need further controls allowing the spectators to have a partial control over the avatar behavior or at least to participate to the overall audience attitude, similarly as in online conferences in which attendees can use emojis to interact or share their mood. Such features echo with the aforementioned recommendation for adding social interactions with humans in order to increase the feeling of co-presence.

6 DISCUSSION AND GUIDELINES

We describe how we used a user-centered development for the STAGE which is used in a scientific presentation seminar. The system is driven by pedagogical narratives relying on the affects the virtual audience aroused in the users. The virtual audience system provides a high-level API for controlling the overall attitude and the behavioral cues needed for the design of the narratives. This approach could partially solve the compromise to find between a fully autonomous system, where instructors cannot adapt their scenario during the training session, and a Wizard-of-Oz system in which the instructors have to manually author each virtual agent.

The results from the preliminary user study we ran during two seminars seem promising. All participants agreed on the potential pedagogical interest the STAGE has for university seminars and concurred on the audience behavior believability. Yet, the system may benefit from more sounds and audio feedback from the VA to improve the users' feeling of immersion.

As for the STAGE's control interface, the workshop we held with professors and lecturers from the university highlighted possible improvements in the visualization tool which can improve and ease the monitoring of the training, whilst the current visualization tool already has some value for post-training feedback and for the high-level controls it provides. The current audience attitude should be easy to read and the interface should only display relevant information. The same for the current narrative state, which is currently hidden in the main graph, a simple state machine graph could solve this problem.

The seminar could be improved as well by providing repeated training sessions and could be used as a hybrid system for formative evaluation in which other students could join the session to embody virtual spectators and participate in the presentation with non-verbal behavior controls for the embodied avatars. This could lead to peer review training sessions where students assist each other on the condition the STAGE provides a qualitative data visualization tool from the user metrics in the case an instructor would not attend the presentation.

The STAGE could now profit from a longitudinal study regarding the learning outcomes it provides. Previous studies underline the need for further research regarding what contributes to the success of VR public speaking training systems (Poeschl, 2017), even though recent studies at least show a good user acceptance for such training systems (Palmas et al., 2019). The new VAS model should also be evaluated in terms of perceived attitudes even if the instructors validated the audience behaviors and that students seem able to recognize the current attitude. Because the VA behavior was designed by the lecturers the resulting attitude might be biased and the students' attitude perception might differ from the instructor one. If the pedagogical narratives seem to affect students it would be interesting to further test adaptive narratives based on user metrics or physiological data. Such interactions may better suit students suffering from PSA by adjusting the audience's attitude to elicit a positive affect and decrease the anxiety induced by the VA.

With these preliminary results in mind, we can provide guidelines regarding the control of virtual audiences in the context of a VR training system.

High-level controls: High-level controls: make use of high-level behavior controls along with an evaluated behavior model to guarantee that the users are going to perceive the displayed behavior as intended, and to avoid the instructors keep focusing on editing the virtual agents' behavior.

Pedagogical Narratives: the design of plausible storytelling is essential to root the users in the training context. Like in role-playing games, instructors can author the ongoing narrative, and in our case, it can even be based on users' metrics to provide specific interest exercises.

Repeated Exposure:such training systems should be used on a repeated basis to let users get familiar with it, and to let them face exercises focusing on specific skills. By doing so, the trainee can improve from one session to another similarly to therapeutic systems.

Hence, in our future works, we first plan to formalize and evaluate the new VA behavior model we used to design the narratives and testified in favor of the new behavioral cues added to the model. Then we will continue to use the STAGE in a seminar at the University of Wuerzburg to refine the training sessions and propose efficient and engaging training sessions for the students. This involves refactoring the visualization tool we provide in the STAGE to find an equilibrium between what the instructor sees and the available controls needed to adapt the ongoing narrative. Such improvements would probably need further user usability evaluation. In the near future, we plan to make the STAGE available for free for evaluation purposes subject to the high-level control API completion. Finally, further investigations will be made regarding the use of interactive storytelling techniques which mitigate the issue of providing causally coherent narrative experiences, where user interaction is taken into account during the unfolding of a story,

e.g., with logical and rules-based perspective (Martens et al., 2013) or plan based perspectives (Young, 1999).

7 CONCLUSION

In this paper, we introduced the STAGE, a virtual reality public speaking training system for undergraduate and postgraduate students to practice Scientific Writing and Presentation seminars. We describe how we used a user-centered development process to create the STAGE system. The main contribution of this paper lies in the proposed trade-off between a fully autonomous simulation and a Wizard of Oz system where each virtual agent would be individually controlled by the instructor. This compromise is solved by the design of a high-level API integrated into a web GUI for the instructors to control and design pedagogical narratives. This API relies on a virtual audience behavior model used to influence the students' affects modulated by the displayed audience attitude changes. The preliminary results obtained from the students and experts who participated in the seminars and the workshop we ran, give some insight into the STAGE potential for providing VR formative educational tools. Finally, we provided guidelines in the discussion to help similar VR training systems be integrated into education or therapeutic curriculum.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YG, A-GB, J-LL contributed to conception and design of the system. YG developed the software and performed the interviews. A-GB, J-LL, CB and ML were the project administrators and supervised the work. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

YG is partly by the ENIB, French Ministry of Higher Education, Research and Innovation. This publication was supported by the Open Access Publication Fund of the University of Wuerzburg.

REFERENCES

Adobe Systems (2022). *Animated 3d Characters Library*.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 10. Washington, DC: American Psychiatric Association. 800 Maine Avenue, S.W., Suite 900, Washington, DC 20024.

Anderson, P. L., Price, M., Edwards, S. M., Obasaju, M. A., Schmertz, S. K., Zimand, E., et al. (2013). Virtual Reality Exposure Therapy for Social Anxiety Disorder: A Randomized Controlled Trial. *J. Consulting Clin. Psychol.* 81, 751–760. doi:10.1037/a0033559

Anderson, P. L., Zimand, E., Hodges, L. F., and Rothbaum, B. O. (2005). Cognitive Behavioral Therapy for Public-Speaking Anxiety Using Virtual Reality for Exposure. *Depress. Anxiety* 22, 156–158. doi:10.1002/da.20090

Arthur, K. W., Booth, K. S., and Ware, C. (1993). Evaluating 3d Task Performance for Fish Tank Virtual Worlds. *ACM Trans. Inf. Syst.* 11, 239–265. doi:10.1145/159161.155359

Bartholomay, E. M., and Houlihan, D. D. (2016). Public Speaking Anxiety Scale: Preliminary Psychometric Data and Scale Validation. *Personal. Individual Differences* 94, 211–215. doi:10.1016/j.paid.2016.01.026

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. (2013). "Cicero - towards a Multimodal Virtual Audience Platform for Public Speaking Training," in International Workshop on Intelligent Virtual Agents (Springer), 116–128. doi:10.1007/978-3-642-40415-3_10

Bevacqua, E., Pammi, S., Hyniewska, S. J., Schröder, M., and Pelachaud, C. (2010). "Multimodal Backchannels for Embodied Conversational Agents," in International Conference on Intelligent Virtual Agents (Springer), 194–200. doi:10.1007/978-3-642-15892-6_21

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators & Virtual Environments* 12, 456–480. doi:10.1162/105474603322761270

Blöte, A. W., Kint, M. J. W., Miers, A. C., and Westenberg, P. M. (2009). The Relation between Public Speaking Anxiety and Social Anxiety: A Review. *J. Anxiety Disord.* 23, 305–313. doi:10.1016/j.janxdis.2008.11.007

Chollet, M., and Scherer, S. (2017). Perception of Virtual Audiences. *IEEE Comput. Grap. Appl.* 37, 50–59. doi:10.1109/mcg.2017.3271465

Chollet, M., Sratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. (2014). "An Interactive Virtual Audience Platform for Public Speaking Training," in Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems) (Richland, SC: AAMAS '14), 1657–1658.

Cummings, J. J., and Bailenson, J. N. (2016). How Immersive Is Enough? a Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychol.* 19, 272–309. doi:10.1080/15213269.2015.1015740

Delamarre, A., Shernoff, E., Buche, C., Frazier, S., Gabbard, J., and Lisetti, C. (2021). The Interactive Virtual Training for Teachers (Ivt-t) to Practice Classroom Behavior Management. *Int. J. Human-Computer Stud.* 152, 102646. doi:10.1016/j.ijhcs.2021.102646

Ekman, P. (1999). Basic Emotions. *Handbook Cogn. Emot.* 98, 45–60. doi:10.1002/0470013494.ch3

Empatica, I. (2022). *Empatica e4 wristband*.

Epic Games, I. (2022). *Unreal Engine*.

Fukuda, M., Huang, H.-H., Ohta, N., and Kuwabara, K. (2017). "Proposal of a Parameterized Atmosphere Generation Model in a Virtual Classroom," in Proceedings of the 5th International Conference on Human Agent Interaction (New York, NY, USA: Association for Computing Machinery), 11–16. doi:10.1145/3125739.3125776

Glémarec, J.-L., and Lugrin, Y. (2022). *Virtual Audience Project page, Chair of Human-Computer-Interaction university of Wuerzburg*.

Glémarec, Y., Lugrin, J.-L., Bosser, A.-G., Collins Jackson, A., Buche, C., and Latoschik, M. E. (2021). Indifferent or Enthusiastic? Virtual Audiences Animation and Perception in Virtual Reality. *Front. Virtual Real.* 2. doi:10.3389/frvir.2021.666232

Hale, K. S., and Stanney, K. M. (2006). Effects of Low Stereo Acuity on Performance, Presence and Sickness within a Virtual Environment. *Appl. Ergon.* 37, 329–339. doi:10.1016/j.apergo.2005.06.009

Harris, S. R., Kemmerling, R. L., and North, M. M. (2002). Brief Virtual Reality Therapy for Public Speaking Anxiety. *Cyberpsychology Behav.* 5, 543–550. doi:10.1089/109493102321018187

Hayes, A. T., Hardin, S. E., and Hughes, C. E. (2013). Perceived Presence's Role on Learning Outcomes in a Mixed Reality Classroom of Simulated Students. In Virtual, Augmented and Mixed Reality. Systems and Applications: 5th International Conference, VAMR 2013, Held as Part of HCI International 2013. Proceedings, Part II. Las Vegas, NV, USA, Berlin, Heidelberg: Springer Berlin Heidelberg, 142–151. chap. Perceived Presence's Role on Learning Outcomes in a Mixed Reality Classroom of Simulated Students. doi:10.1007/978-3-642-39420-1_16

Heudin, J.-C. (2007). "Evolutionary Virtual Agent at an Exhibition," in Proceedings of the 13th International Conference on Virtual Systems and Multimedia (Berlin, Heidelberg: Springer-Verlag), 154–165. VSMM'07.

Hosseinpanah, A., Krämer, N. C., and Straßmann, C. (2018). "Empathy for Everyone? the Effect of Age when Evaluating a Virtual Agent," in Proceedings of the 6th International Conference on Human-Agent Interaction, 184–190.

Kahlon, S., Lindner, P., and Nordgreen, T. (2019). Virtual Reality Exposure Therapy for Adolescents with Fear of Public Speaking: a Non-randomized Feasibility and Pilot Study. *Child. Adolesc. Psychiatry Ment. Health* 13, 47–10. doi:10.1186/s13034-019-0307-y

Kang, N., Brinkman, W.-P., Birna van Riemsdijk, M., and Neerincx, M. (2016). The Design of Virtual Audiences: Noticeable and Recognizable Behavioral Styles. *Comput. Hum. Behav.* 55, 680–694. doi:10.1016/j.chb.2015.10.008

Kelly, O., Matheson, K., Martinez, A., Merali, Z., and Anisman, H. (2007). Psychosocial Stress Evoked by a Virtual Audience: Relation to Neuroendocrine Activity. *CyberPsychology Behav.* 10, 655–662. doi:10.1089/cpb.2007.9973

Latoschik, M.Team (Uni Würzburg) (2022). *Decker*.

Latoschik, M. E., Kern, F., Stauffert, J.-P., Bartl, A., Botsch, M., and Lugrin, J.-L. (2019). Not alone Here?! Scalability and User Experience of Embodied Ambient Crowds in Distributed Social Virtual Reality. *IEEE Trans. Vis. Comput. Graphics* 25, 2134–2144. doi:10.1109/tvcg.2019.2899250

Lee, C., Bonebrake, S., Bowman, D. A., and Höllerer, T. (2010). "The Role of Latency in the Validity of Ar Simulation," in 2010 IEEE Virtual Reality Conference (VR) (IEEE), 11–18. doi:10.1109/vr.2010.5444820

Lindner, P., Dagöö, J., Hamilton, W., Miloff, A., Andersson, G., Schill, A., et al. (2021). Virtual Reality Exposure Therapy for Public Speaking Anxiety in Routine Care: a Single-Subject Effectiveness Trial. *Cogn. Behav. Ther.* 50, 67–87. doi:10.1080/16506073.2020.1795240

Lugrin, J.-L., Latoschik, M. E., Habel, M., Roth, D., Seufert, C., and Grafe, S. (2016). Breaking Bad Behaviors: A New Tool for Learning Classroom Management Using Virtual Reality. *Front. ICT* 3, 26. doi:10.3389/fict.2016.00026

Lugrin, J.-L., Wiebusch, D., Latoschik, M. E., and Strehler, A. (2013). "Usability Benchmarks for Motion Tracking Systems," in Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology (New York, NY: VRST '13), 49–58. doi:10.1145/2503713.2503730

Marsella, S., and Gratch, J. (2002). A Step toward Irrationality: Using Emotion to Change Belief. *Proc. first Int. Jt. Conf. Autonomous Agents Multiagent Syst. part* 1, 334–341.

Martens, C., Bosser, A.-G., Ferreira, J. F., and Cavazza, M. (2013). "Linear Logic Programming for Narrative Generation," in International Conference on Logic Programming and Nonmonotonic Reasoning (Springer), 427–432. doi:10.1007/978-3-642-40564-8_42

Mehrabian, A. (1996). Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament. *Curr. Psychol.* 14, 261–292. doi:10.1007/bf02686918

Messinis, I., Saltaouras, D., Pintelas, P., and Mikropoulos, T. (2010). "Investigation of the Relation between Interaction and Sense of Presence in Educational Virtual Environments," in 2010 International Conference on e-Education (e-Business, e-Management and e-Learning), 428–431. doi:10.1109/IC4E.2010.137

Mouw, J. M., Fokkens-Bruinsma, M., and Verheij, G.-J. (2020). "Using Virtual Reality to Promote Pre-service Teachers' Classroom Management Skills and Teacher Resilience: A Qualitative Evaluation," in Proceedings of the 6th International Conference on Higher Education Advances (HEAd'20) (Valencia, Spain: Universitat Politècnica de València), 325–332. doi:10.4995/head20.2020.11049

Narayan, M., Waugh, L., Zhang, X., Bafna, P., and Bowman, D. (2005). "Quantifying the Benefits of Immersion for Collaboration in Virtual Environments," in Proceedings of the ACM Symposium on Virtual Reality Software and Technology (New York, NY: ACM), 78–81. doi:10.1145/1101616.1101632

Owens, M. E., and Beidel, D. C. (2015). Can Virtual Reality Effectively Elicit Distress Associated with Social Anxiety Disorder? *J. Psychopathol Behav. Assess.* 37, 296–305. doi:10.1007/s10862-014-9454-x

Palmas, F., Cichor, J., Plecher, D. A., and Klinker, G. (2019). "Acceptance and Effectiveness of a Virtual Reality Public Speaking Training," in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (IEEE), 363–371. doi:10.1109/ismar.2019.00034

Pelachaud, C. (2009). Studies on Gesture Expressivity for a Virtual Agent. *Speech Commun.* 51, 630–639. doi:10.1016/j.specom.2008.04.009

Pertaub, D.-P., Slater, M., and Barker, C. (2002). An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence: Teleoperators & Virtual Environments* 11, 68–78. doi:10.1162/105474602317343668

Poeschl, S. (2017). Virtual Reality Training for Public Speaking-A QUEST-VR Framework Validation. *Front. ICT* 4, 13. doi:10.3389/fict.2017.00013

Roseman, I. J. (1991). Appraisal Determinants of Discrete Emotions. *Cogn. Emot.* 5, 161–200. doi:10.1080/02699939108411034

Rothbaum, B. O., Hodges, L., Smith, S., Lee, J. H., and Price, L. (2000). A Controlled Study of Virtual Reality Exposure Therapy for the Fear of Flying. *J. consulting Clin. Psychol.* 68, 1020–1026. doi:10.1037/0022-006x.68.6.1020

Shernoff, E. S., Von Schalscha, K., Gabbard, J. L., Delmarre, A., Frazier, S. L., Buche, C., et al. (2020). *Evaluating the Usability and Instructional Design Quality of Interactive Virtual Training for Teachers (Ivt-t)*. Boston: Educational Technology Research and Development, 1–28.

Slater, M., Lotto, B., Arnold, M. M., and Sanchez-Vives, M. V. (2009). How We Experience Immersive Virtual Environments: the Concept of Presence and its Measurement. *Anuario de Psicología* 40, 193–210.

Slater, M., Sadagic, A., Usoh, M., and Schroeder, R. (2000). Small-group Behavior in a Virtual and Real Environment: A Comparative Study. *Presence: Teleoperators & Virtual Environments* 9, 37–51. doi:10.1162/105474600566600

VirtualSpeech, L. (2022). *Virtualspeech*.

VRSpeaking, L. (2022). *Ovation*.

Wallach, H. S., Safir, M. P., and Bar-Zvi, M. (2009). Virtual Reality Cognitive Behavior Therapy for Public Speaking Anxiety. *Behav. Modif* 33, 314–338. doi:10.1177/0145445509331926

Yngve, V. H. (1970). *On Getting a Word in Edgewise*. 6th Meeting. Chicago: Chicago Linguistics Society, 567–578.

Young, R. M. (1999). "Notes on the Use of Plan Structures in the Creation of Interactive Plot," in AAAI Fall Symposium on Narrative Intelligence (Palo Alto, CA: AAAI Press Menlo Park), 164–167.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.