Integration of Artificial Intelligence into Virtual Reality Training Systems for Public Speaking

Sarah Saufnay and Michaël Schyns
HEC Liège – Management School of the University of Liège, LabXR, Belgium
Corresponding author: sarah.saufnay@uliege.be

Keywords: Public speaking, virtual reality, artificial intelligence, training

Introduction

Although speaking in public is common, many people struggle with it due to limited practice opportunities, which can cause significant anxiety (Bodie, 2010; Kroczek & Mühlberger, 2023). Virtual Reality (VR) training helps address this by simulating social interactions, which are costly and difficult to reproduce in real life (Hirzle et al., 2023). Instead, once wearing a VR headset, users are immersed in a realistic presentation setting, facing an audience of Intelligent Virtual Agents (IVAs), i.e., computer-controlled virtual humans that display social cues to mimic human presence (Kyrlitsias & Michael-Grigoriou, 2022). Moreover, Artificial Intelligence (AI) has expanded VR's interaction capabilities across various training fields (Hirzle et al., 2023). By analyzing user behavior, AI indeed enables the development of responsive Virtual Environments (VEs) and the generation of associated content. These advances have been especially relevant for Public Speaking Training (PST) environments, where AI can help assess speakers' performance, provide feedback, and customize training sessions (Chang et al., 2026).

Carefully designing the virtual audience's behavior in response to the speaker's performance is essential. Both verbal and nonverbal cues of IVAs significantly influence audience perception (Glémarec et al., 2021), which, in turn, can affect speakers' real-world self-confidence and performance (Kroczek & Mühlberger, 2023). Consequently, these design decisions are critical in training contexts. In addition, incorporating AI into VR is essential to achieve enhanced interactivity and shift from static to dynamic training simulations, but it presents challenges, including limited computational power, latency, data privacy concerns, and potential dependence on external services.

In this context, a PST application in VR has been developed, allowing users to choose specific training configurations and interaction modalities based on their objectives. The system provides diverse presentation contexts for subsequent use in school programs or companies, and contributes at the technical level by integrating AI into standalone VR applications. Currently, this application is employed at the authors' University across multiple faculties to train students in public speaking and prepare them for future careers.

Platform Framework

The platform features a web-based companion interface alongside a standalone VR application. The web interface supports the preparation and management of training sessions, while the VR application provides the actual immersive training experience. Before entering VR,

users can upload presentation materials such as slides and notes, which are then available in the VE. After each training session, recorded data and performance feedback can also be accessed via the web interface to support reflection and skill improvement.

Training sessions can be customized by selecting a room, an associated role, and a scenario defining the audience's attitudes and training events. Once immersed in the chosen room, users can interact with their slides and virtual notes during practice, thereby reproducing key elements of a real presentation setting. Sessions can also be recorded and replayed from a third-person perspective, either from the website or directly within the VE, allowing users to review and reflect more objectively on their performance (Zhou et al., 2021).

Training configurations

Several rooms, corresponding to traditional public speaking contexts and associated with specific roles and training scenarios, have been developed to cover a broad range of communication situations.

The Auditorium, which replicates an existing university auditorium, represents a traditional large-scale public speaking context, whereas the Boardroom and Meeting Room provide more formal settings for business presentations. These environments feature configurable and validated virtual audiences, whose attitudes can range from attentive and supportive to disengaged or negative (Etienne et al., 2023; Saufnay and Schyns, 2025). More specialized environments target specific communication contexts. The Courtroom familiarizes future lawyers with courtroom procedures and speaking constraints, while the Thesis Defense reproduces an academic jury using photorealistic avatars of actual professors, enabling the study of familiarity effects on stress, social presence, and performance. The Pitch environment similarly simulates entrepreneurial presentations. In the latter two VEs, jury members can ask either prerecorded generic questions or specific ones, generated at runtime using the embedded AI system. Other environments prepare users for professional situations before internships. The Classroom focuses on teacher training and classroom management through pedagogical and disciplinary events occurring during immersion, such as students talking, using their phones, sleeping, or rocking their chairs. Speech recognition and intention detection enable the system to identify appropriate interventions from the users' speech and dynamically adapt the virtual students' behavior. The Psychology Internship scenario exposes future psychologists to challenging interactions, including patients questioning their credibility because of their age, remaining silent, or displaying strong emotional reactions. Finally, the Office supports job interview training from both the applicant and interviewer perspectives and can also be used to investigate recruitment biases by using IVAs of different genders and ethnicities. Interactions may rely on predefined audio recordings or on conversational virtual agents for more diversified discussions.

Intelligent and Adaptive Virtual Audiences

A key component of the platform is the relationship between IVAs' behaviors and users' perception of audience attitudes. These attitudes are modeled along the dimensions of valence and arousal, representing the positive or negative attitude towards the speaker and activation level of the audience response (Etienne et al., 2023).

Attitudes are designed as combinations of nonverbal cues, including posture, gaze, head movements, and facial expressions. The objective was to use combinations that consistently convey the intended attitude. However, this becomes more complex when moving from individual agents to groups, as an audience's overall attitude cannot be reduced to the perceptions of each member in isolation. Factors such as the proportion of agents displaying specific behavior, its recurrence, or the IVAs' position within the room may influence group perception. The platform therefore investigates how models validated at the individual level can be transferred to virtual audiences.

Beyond pre-scripted events, the application's long-term objective is to make audience reactions directly responsive to speakers' performance. Using openSMILE (Eyben et al., 2016), acoustic features can be extracted from the speaker's voice and used to estimate performance- and emotion-related dimensions, including valence, arousal, stress, and vocal dynamics. Speech alone provides valuable information for evaluating performance, yet it can be further enriched by tracking the user's nonverbal behavior. These features can then be used to influence the virtual audience. For example, a monotonous delivery could reduce engagement, whereas a more dynamic presentation could elicit attentive or positive responses. This creates an interaction loop in which speaker behavior is analyzed and translated into changes in the VE, making the audience a key component of the training.

Embedded Artificial Intelligence and Interaction Management

In addition to the agents' nonverbal behavior, the audience is often brought to ask questions and interact with the speaker during the presentation. This implementation requires specific AI components for general contextual interactions. A major technical objective of the platform is to investigate whether these components can run locally on standalone VR headsets to preserve user privacy, minimize response latency, and avoid dependence on external services (Wang et al., 2025). Lightweight models are therefore preferred.

Automatic Speech Recognition (ASR), such as Vosk, is used first to convert the user's speech into text. Depending on the scenario's complexity, the resulting text can then be processed either by a dedicated intention-detection model or by a Large Language Model (LLM). Intention detection is particularly useful when a restricted and predefined set of intents must be recognized in the speaker's speech. In the Classroom environment, for example, the system primarily needs to determine whether the user is asking the students to stop a specific behavior or to do something. A restricted set of intentions can then be detected to trigger the relevant IVA behavior. More flexible scenarios, including job interviews, thesis defenses, and entrepreneurial pitches, can instead rely on small LLMs, such as a 3B-parameter Llama model deployed through the UndreamAI framework, to generate context-specific questions or responses based on the current interaction and the chosen scenario. Generated responses can then be converted into speech through a Text-To-Speech component (e.g., Piper). Response latency is an important constraint in this pipeline, as excessive delays between user speech and virtual agent responses can reduce the naturalness of the interaction. Whenever computational resources allow, components are executed locally on the VR headset.

Research Perspective

The platform is intended both as a training application and as a research platform for investigating social interactions in VR. Research objectives span the entire development pipeline, from the socio-affective perception of multimodal IVAs' behavior to the automatic assessment of users' speaking performance and the evaluation of the training effectiveness. The project also enables comparisons between predefined and AI-generated interactions. Prerecorded content offers greater experimental control, whereas generative AI may increase relevance and variability. Comparing these approaches can help determine the best design to achieve ecological validity and a strong sense of presence.

Overall, the proposed platform aims to move beyond static VR exposure toward an adaptive PST system. By combining configurable training environments, intelligent virtual audiences, performance metrics, as well as recording and replay functionalities, the system is designed not only to reproduce public-speaking contexts but also to adjust the virtual environment dynamically to offer a more realistic training experience and support skills transfer to real-world public-speaking situations.

References

Bodie, G. D. (2010). A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education, 59*(1), 70–105. https://doi.org/10.1080/03634520903443849

Chang, E., Che, X., & Zhang, J. (2026). Designing and evaluating LLM-driven coaching agents for VR public speaking practice. *Frontiers in Psychology, 17*, 1886780. https://doi.org/10.3389/fpsyg.2026.1886780

Etienne, E., Leclercq, A.-L., Remacle, A., Dessart, L., & Schyns, M. (2023). Perception of avatars' nonverbal behaviors in virtual reality. *Psychology & Marketing, 40*(11), 2464–2481. https://doi.org/10.1002/mar.21871

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Glémarec, Y., Lugrin, J.-L., Bosser, A.-G., Collins Jackson, A., Buche, C., & Latoschik, M. E. (2021). Indifferent or enthusiastic? Virtual audiences animation and perception in virtual reality. *Frontiers in Virtual Reality, 2*. https://doi.org/10.3389/frvir.2021.666232

Hirzle, T., Müller, F., Draxler, F., Schmitz, M., Knierim, P., & Hornbæk, K. (2023). When XR and AI meet: A scoping review on extended reality and artificial intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581072

Kroczek, L. O. H., & Mühlberger, A. (2023). Public speaking training in front of a supportive audience in virtual reality improves performance in real-life. *Scientific Reports, 13*(1), 13968. https://doi.org/10.1038/s41598-023-41155-9

Kyrlitsias, C., & Michael-Grigoriou, D. (2022). Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality, 2*. https://doi.org/10.3389/frvir.2021.786665

Saufnay, S., & Schyns, M. (2025). Perception of virtual audiences: Influence of gender and nonverbal behavior. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents (IVA '25)* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3717511.3747085

Wang, Y., Guizani, M., & Hossain, M. S. (2025). Artificial intelligence for virtual reality: State of the art, challenges, and future perspectives. *ACM Transactions on Multimedia Computing, Communications, and Applications, 21*(11). https://doi.org/10.1145/3769090

Zhou, H., Fujimoto, Y., Kanbara, M., & Kato, H. (2021). Virtual reality as a reflection technique for public speaking training. *Applied Sciences, 11*(9), 3988. https://doi.org/10.3390/app11093988