*Human Interaction and Emerging Technologies (IHIET-FS 2026)*, Vol. 235, 2026, 293–303
https://doi.org/10.54941/ahfe1007283

# An Adaptive XR Platform for Multimodal Public Speaking Training and Performance Assessment

**Lia Cardoso[1], Hugo Correia[1,2], Bernardo Marques[1,2], Paulo Dias[1,2], Samuel Silva[1,2], and Beatriz Sousa Santos[1,2]**

[1]Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Portugal
[2]IEETA, University of Aveiro, Portugal

## ABSTRACT

Public speaking is a fundamental competence across academic, professional, and social contexts, yet an estimated 77% of the population experiences public speaking anxiety (PSA), manifesting through psychological symptoms – such as fear of judgment and avoidance behaviours – and physiological responses including increased heart rate and electrodermal activity. Traditional training approaches face limitations in scalability, reproducibility, and objective performance assessment, often relying on subjective rubrics and offering limited opportunities for repeated practice in realistic settings. Extended Reality (XR) technologies have emerged as a promising alternative, as immersive virtual environments can reliably elicit anxiety responses comparable to real audiences while providing safe, controllable, and repeatable practice scenarios. However, most existing XR platforms focus primarily on exposure, without systematically capturing or leveraging the multimodal signals that reflect a speaker's internal state and communicative behaviour. Incorporating behavioural data such as gaze patterns, gesture dynamics, and spatial movement, alongside psychophysiological signals and speech acoustics, is essential to move beyond subjective evaluation toward objective characterization of user states, enabling personalized feedback and real-time training adaptation. This paper presents the design, implementation, and preliminary assessment of an adaptive XR platform for public speaking training that integrates these multimodal data streams within a five-layer architecture (Environment Generation, Data Gathering, Analysis, Feedback, and Visualization), instantiated through a modular microservices approach. A preliminary usability assessment with ten participants demonstrates the platform's learnability and task completion effectiveness, advancing beyond exposure-only systems toward intelligent, data-driven, and personalized skill development.

**Keywords:** Extended reality, Public speaking anxiety, Multimodal data analysis, Adaptive feedback, Virtual reality training

## INTRODUCTION

Effective communication is a cornerstone of success in professional, academic, and personal contexts. In higher education, the ability to present ideas clearly during oral assessments, thesis defenses, and collaborative discussions is essential for student development (Al-Madani, 2015). In

Received April 10, 2026; Revised April 25, 2026; Accepted May 12, 2026; Available online June 22, 2026

professional settings, public speaking enables persuasion, leadership, and knowledge dissemination (Sumaiya et al., 2022). Yet, public speaking anxiety (PSA), commonly known as glossophobia, affects an estimated 77% of the population, inducing physiological symptoms such as rapid heartbeat, vocal tremors, and diaphoresis, alongside psychological manifestations including fear of judgment, avoidance behaviors, and impaired concentration (Jim et al., 2025; Dănescu & Romăşcanu, 2024).

Traditional approaches to public speaking training exhibit several structural limitations. Classroom-based courses typically allow only one student to present at a time, constraining practice opportunities (Bertotti, 2022). Assessment relies predominantly on subjective rubrics such as the NCA Competent Speaker evaluation form (Morreale et al., 2007) and the rubric proposed by Van Ginkel et al. (2017), which introduce inter-rater variability and lack the granularity needed to capture nuanced performance dimensions. Furthermore, practicing in isolation (without a live audience) fails to replicate the social pressure that characterizes real speaking situations, limiting both anxiety management gains and skill transfer (Zuardi et al., 2013).

XR technologies offer a promising avenue for addressing these limitations by providing immersive, controllable environments where users can practice before configurable virtual audiences that elicit realistic psychological and physiological responses (Kang, 2016; Kroczek & Mühlberger, 2023). VR-based training has been shown to reduce PSA scores (Rodero & Larrea, 2022; Sarpourian et al., 2022) and improve confidence through supportive audience design (Kroczek & Mühlberger, 2023). However, a critical gap remains: most existing systems emphasize subjective outcomes and post-session ratings without exploiting the full potential of XR for collecting synchronized, objective behavioral, and physiological metrics. This objectivity is key since it enables more consistent performance assessment, reduces inter-rater variability, and supports longitudinal comparability across training sessions.

This paper addresses this gap by presenting an adaptive XR platform that integrates speech acoustics, kinematic behaviors, and psychophysiological signals. The contributions are: (1) an XR platform capturing multimodal data for public speaking training; (2) metric families mapped to cognitive and behavioral constructs relevant to PSA; (3) a layered architecture with plugin-based real-time analysis and adaptive feedback; and (4) a preliminary usability assessment demonstrating learnability and effectiveness.

## BACKGROUND

### Public Speaking Training and Assessment

PSA is a pervasive challenge that requires both skill development and emotional regulation through practice (Ristorcelli et al., 2025). Traditional training methods combine instructor-led lectures, prepared in-class speeches with peer and instructor feedback, and rehearsal activities such as impromptu talks and articulation drills (Meadows, 2019). When addressing PSA specifically, psychological interventions, including systematic desensitization

through gradual exposure, cognitive restructuring of negative automatic thoughts, and targeted skills training modules, are often incorporated (Sonata & Tetelepta, 2024). Assessment relies on standardized rubrics such as the NCA Competent Speaker Speech Evaluation Form (Morreale et al., 2007), which defines eight core competencies spanning preparation and delivery, and the Van Ginkel et al. (2017) rubric comprising eleven items covering content quality, structural coherence, audience interaction, and delivery aspects. Despite their widespread adoption, these instruments introduce considerable evaluator subjectivity and inter-rater variability, and the training itself is constrained by limited practice time and the absence of objective, data-driven feedback.

## XR for Public Speaking

VR creates fully immersive environments with complete environmental control, enabling realistic audience simulation while maintaining a safe, repeatable practice space (Rostami et al., 2025). Research has demonstrated that virtual audiences with expressive body language and gaze behavior can affect perceived social presence and influence speaker confidence (Kang, 2016), that supportive virtual audiences improve both self-assessed and instructor-assessed performance in subsequent real-life presentations (Kroczek & Mühlberger, 2023), and that VR training with distractors significantly reduces PSA compared to non-VR training (Rodero & Larrea, 2022). Educational deployments, such as Harvard's VR public speaking modules, have shown that VR practice increases student confidence and engagement while reducing logistical constraints (Bertotti, 2022). Recent work has also explored AR for public speaking, with Jim et al. (2025) presenting an augmented reality tool that overlays virtual audience members into the user's physical environment.

Despite these advances, current XR public speaking platforms exhibit important gaps. Most systems emphasize subjective outcomes such as self-reported anxiety, confidence, or instructor ratings, without exploiting multimodal, synchronized objective metrics (Kang, 2016; Rodero & Larrea, 2022). Few integrate detailed logging of gaze, gesture, or spatial behavior, and even fewer incorporate psychophysiological signals as part of the assessment pipeline (Sarpourian et al., 2022). Feedback is typically static (post-session ratings or generic tips) rather than dynamically adaptive based on measured user state. Finally, existing tools rarely align their metrics with established public speaking rubrics, limiting comparability with traditional assessment frameworks.

## Measuring Behavior and Cognition in XR

Research converges on four complementary metric families for evaluating public speaking performance. Speech-based metrics capture how a person speaks – pitch range, speaking rate, pause patterns, and dysfluency ratio – and approximate both charisma (PASCAL score) and cognitive load (Niebuhr, 2025; Niebuhr et al., 2020). Behavior-based metrics cover kinematic and non-verbal dimensions, including motion energy, gesture strike zone adherence, audience coverage, and gaze aversion rate (Wörtwein et al., 2015; Carstens,

2019). Psychophysiology-based metrics – heart rate, HRV, and electrodermal activity – serve as objective markers of arousal and anxiety (Zuardi et al., 2013; Gallego et al., 2022). Self-report measures such as the PSSE and PRCS complement sensor data by capturing perceived confidence and anxiety levels (Marshall-Wheeler et al., 2022). Together, these four families approximate the cognitive and affective state of the speaker within XR environments, enabling a richer and more objective basis for performance assessment than traditional rubrics alone.

## VISION AND REQUIREMENTS

To guide the research and an understanding of potential users and their motivations, three personas were developed during the design process (student, therapist, and researcher) along with several scenarios. Four design drivers, identified through the literature review, persona analysis, and scenario development, guided the overall requirements for the platform (as listed in Table 1): (a) the platform must induce responses comparable to real-life speaking situations, as virtual audiences have been shown to elicit anxiety similar to real audiences (Zuardi et al., 2013); (b) training must accommodate varying skill and anxiety levels, supporting systematic desensitization through progressive challenge (Sonata & Tetelepta, 2024); (c) performance data from all metric families must be organized and presented clearly, avoiding cognitive overload while retaining analytical depth, adhering to coherence and signaling principles from instructional design research (Yang et al., 2020); and (d) the platform must be intuitive for users with varying XR expertise, from novices to experienced practitioners, while minimizing cybersickness risk through adherence to VR UI best practices (Mehmedova et al., 2025).

**Table 1**: Key functional (F) and non-functional (NF) requirements.

| Requirement | Description |
|---|---|
| FR1 | Support configurable virtual environments with adjustable audience sizes and difficulty-scaled behavior patterns |
| FR2 | Allow capture and analysis of vocal characteristics (rate, pitch, volume, filler words, pauses) |
| FR3 | Enable tracking of kinematic behaviors (head orientation, gaze, gestures, spatial movement) |
| FR4 | Support wearable sensor integration for physiological data acquisition (HR, EDA) |
| FR5 | Allow synchronization of multimodal data streams with session-level timestamps |
| FR6 | Provide post-session summaries with actionable feedback and longitudinal progress tracking |
| FR7 | Support integration with an analytics platform and data export in standardized formats |
| NFR1 | Maintain ≥72 fps; speech analysis latency ≤500 ms |
| NFR2 | Secure storage with encryption and GDPR-compliant data controls |
| NFR3 | Modular extensibility for additional analysis plugins without architectural changes |

## DEVELOPMENT

The platform is organized into five conceptual layers: (a) **Environment Generation & Customization** Layer handles virtual scenario creation (audience composition, behaviors, settings); (b) **Data Gathering** Layer captures raw streams from the VR headset, microphone, and external sensors; (c) **Analysis Layer** processes data through specialized plugins for voice, behavior, physiology, and cognition; (d) **Feedback** Layer adjusts difficulty and audience reactions based on computed metrics; and (e) **Visualization** Layer presents results and trends through in-VR and web-based interfaces.

These layers are instantiated through the technical architecture shown in Figure 1. A Data Collection Service captures HMD data via C#/Unity and Python services. A Session Logging Service persists timestamped data to MongoDB. A RabbitMQ message queue enables asynchronous delivery to four types of analysis plugins, each focusing on a different dimension: Voice Analysis (using librosa and Parselmouth for acoustic feature extraction), Physiological Analysis (processing HR, HRV, and EDA from wearables), Behavioral Analysis (processing motion and gaze data from HMD and controllers), and Cognitive Analysis (inferring higher-level state indicators). A Metrics Computation Service aggregates plugin outputs into composite indicators – an engagement index and an anxiety index – which the Context and Feedback System translates into environmental responses (e.g., reducing audience size when anxiety exceeds a threshold).

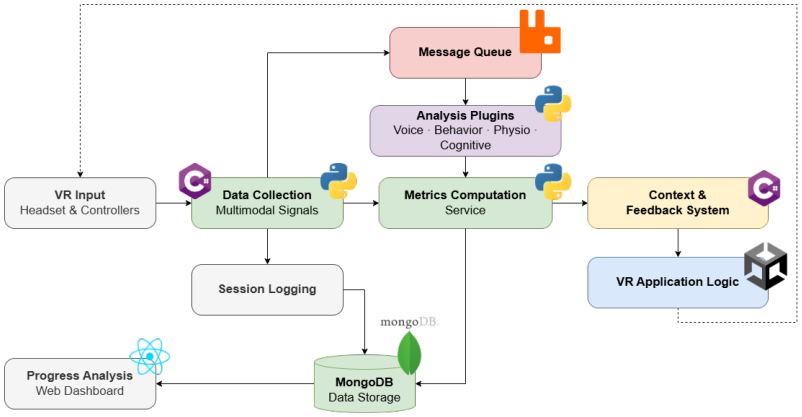


**Figure 1:** Technical architecture showing the project's pipeline: data collection from VR hardware, session logging to MongoDB, asynchronous plugin-based analysis via RabbitMQ, metrics computation, context-driven feedback, and the Unity VR application.

### VR Environment and Interaction

At this stage, the prototype supports four core implemented features: (1) Selection Menus for level/difficulty/session configuration; (2) the Audience Behavior Layer for NPC spawning and management; (3) the Presentation Layer for loading and navigating slides in VR; and (4) the Speech Analysis Layer for real-time vocal processing. A central Session Manager drives the system through a state machine: scene loading, difficulty selection, NPC

spawning, countdown, active session (all analytics and recording active), and results display.

The VR environment is built with Unity targeting Meta Quest headsets, using the Meta XR All-in-One SDK and Meta Building Blocks. The auditorium scene (Figure 3) features a 324-seat capacity populated by configurable NPCs driven by a shared Animator Controller with a blend tree mapping a continuous parameter to behavioral states (idle, asking questions, conversational gestures, distracting behaviors), enabling smooth transitions between supportive and challenging audience configurations.

The Presentation Layer converts user-provided PowerPoint or PDF files to images via LibreOffice, Ghostscript, and ImageMagick, displaying them as textures on an in-scene SlideBoard (visible in Figure 2). Users navigate slides using the controller and can point a laser at the virtual screen, replicating familiar presentation interaction patterns.

**Figure 2:** The virtual auditorium environment with NPC audience members exhibiting difficulty-scaled behavioral patterns. The speaker's view includes the presentation screen, laser pointer, and audience seating.

## Data Collection and Speech Analysis

The Speech Analysis Layer captures audio from the headset microphone and transmits it to a Python Flask backend that extracts pitch, speaking rate, volume, pause patterns, and filler word frequency. Results are presented to the user at session end through an in-VR dashboard (Figure 3) displaying scores for clarity, pitch, volume, rate, and pauses, along with personalized recommendations.


**Figure 3:** In-VR speech analysis results dashboard displaying scores for clarity, pitch, volume, speaking rate, and pauses, with personalized recommendations for improvement.

For behavioral data, the platform captures head orientation, gaze direction, controller-based gesture activity, and spatial movement from the HMD and controller tracking streams, providing the inputs for the Behavioral Analysis plugin described above.

## Analytics and Visualization

For spatial and gaze analytics, the platform integrates Cognitive3D[1] – an XR analytics platform that enriches raw session data with spatial heatmaps, gaze timeline visualizations, and session event logging. Within Unity, a Cognitive3D Manager handles event and sensor data, an Avoidance Analytics module tracks gaze aversion, and a Presentation Analytics module logs slide timing. Integrating Cognitive3D enables research-grade insight into spatial behavior patterns that would be difficult to implement from scratch, and its event-driven API aligns naturally with the platform's Session Manager state machine.

The preliminary assessment revealed that the full Cognitive3D analytics dashboard provides comprehensive data suitable for detailed performance analysis. From this set of choices, we decided to implement a progressive disclosure of the data for clinical users who are less familiar with analytics platforms. In this regard, a custom SpeakEase Analytics web dashboard (Figure 4) was developed as a role-appropriate complement. It presents high-level session summaries and cross-session performance trends in a simplified format, allowing therapists to access key indicators directly while retaining the option for deeper exploration through Cognitive3D.

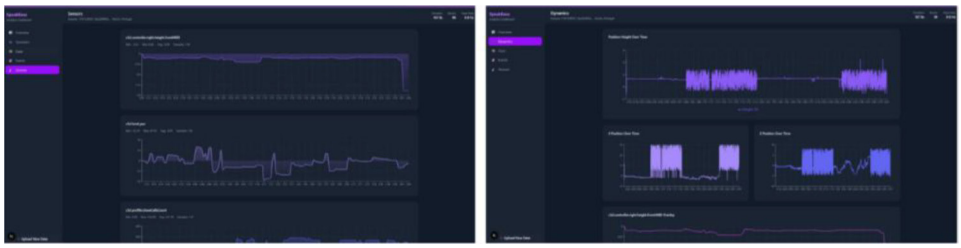

**Figure 4**: The custom SpeakEase Analytics web dashboard showing session summaries and cross-session performance trends.

## Multimodal Metrics

The analysis layer operationalizes the different metric families from the background section and at this stage we adopted two: speech and behavioural metrics. Speech metrics approximate charisma (PASCAL components), fluency (dysfluency ratio), and cognitive load (pause frequency, reduced pitch variability), extracted automatically from in-headset audio. Behavioral metrics include motion energy, stage usage, gesture strike zone adherence, and gaze-based audience coverage, computed from HMD and controller tracking. These are combined into composite indicators: an engagement

[1]Cognitive3D XR Analytics Platform: https://cognitive3d.com

index (motion energy, gesture activity, audience coverage, pitch variability) and an anxiety index (gaze aversion, physiological arousal, speech rate acceleration, dysfluency). The Feedback Layer consumes these to drive real-time adaptations – reducing audience size or switching to supportive NPC behaviors when anxiety exceeds a threshold.

## PRELIMINARY ASSESSMENT

At this stage of the prototype, with the four core features implemented, we were interested in understanding whether users could independently operate the platform to complete representative speaker and therapist tasks, and whether the interface design supported learnability without prior training. To this end, a preliminary usability assessment was conducted with ten participants: five university students enrolled in a VR/AR course acting as speakers, and five faculty members or graduate students acting as therapists. Participants ranged from 21 to 23 years of age, with varying prior experience with VR (from none to regular use). Each session was observed by a researcher, and participants were encouraged to think aloud during task completion. Completion times, task correctness, and need for assistance were recorded.

### Speaker Tasks

Speakers completed six tasks: (1) start the application and select a level, (2) select difficulty and start the session, (3) select a file to present, (4) navigate the presentation, (5) point a laser to the screen, and (6) identify the clarity score. Table 2 summarizes results.

**Table 2:** Speaker task completion times and outcomes.

| | T1 | T2 | T3 | T4 | T5 | T6 |
|---|---|---|---|---|---|---|
| P1 | 3 s | 10 s | 57 s; help | 5 s | 7 s | 11 s; correct |
| P2 | 6 s | 14 s | 35 s; correct | 8 s | 12 s | 12 s; correct |
| P3 | 12 s | 22 s | 1m18s; help | 12 s | 10 s; help | 9 s; correct |
| P4 | 4 s | 10 s | 25 s; correct | 6 s | 8 s | 6 s; correct |
| P5 | 5 s | 10 s | 55 s; help | 8 s | 79 s | 10 s; correct |

Tasks 1, 2, 4, and 5 were typically completed under 15 seconds. Task 3 (file selection) proved most challenging – three participants requested help, with times up to 78 seconds – suggesting the file browser workflow requires refinement. Task 5 was also slow for P5 (79 s), attributed to initial unfamiliarity with controller button mapping; this pointed to a need for improved in-scene controller tooltips. Task 6 (clarity score) was answered correctly by all participants, indicating the feedback dashboard is readable and interpretable.

### Therapist Tasks

Therapists completed three tasks on the analytics dashboard: (1) identify presence level, (2) determine speech duration, and (3) count Avoidance Alert triggers. Table 3 summarizes the results.

**Table 3:** Therapist task completion times and outcomes.

| | Task 1 | Task 2 | Task 3 |
|---|---|---|---|
| P1 | 1 min; correct; help | 42 s; correct | 27 s; correct |
| P2 | 48 s; correct | 35 s; correct | 40 s; correct |
| P3 | 1m22s; help; correct | 55 s; incorrect→corrected | 50 s; correct |
| P4 | 35 s; correct | 28 s; correct | 22 s; correct |
| P5 | 58 s; help; correct | 45 s; correct | 1m5s; help |

Task 1 (presence level) was most challenging – three participants asked for help, with times up to 82 seconds – suggesting that this metric requires more prominent placement in the dashboard. Tasks 2 and 3 were generally completed within 30–55 seconds with correct answers.

### Discussion

The assessment demonstrates that the platform is learnable for both roles, with most tasks completed within acceptable time frames and without critical errors, and no participant reporting cybersickness symptoms. For speakers, file selection (Task 3) was the primary pain point, with three of five participants requiring assistance, pointing to a need for a refined file browser workflow and improved controller tooltips. For therapists, the presence level metric (Task 1) proved hardest to locate, suggesting it requires more prominent placement in the dashboard; Tasks 2 and 3 were generally completed correctly within 30–55 seconds. Participants also noted that distractor behaviors felt excessive at higher difficulty levels, indicating a need for more naturalistic NPC patterns. These findings informed concrete refinements, namely a simplified analytics dashboard, improved tooltips, and recalibrated distractor frequency, already incorporated into the current prototype.

## CONCLUSION

This paper presented the design, implementation, and preliminary assessment of an adaptive XR platform for public speaking training that integrates synchronized speech, behavioural, and physiological metrics to drive real-time adaptive feedback. The five-layer architecture with extensible plugin-based analysis pipelines, combined with four metric families mapped to PSA-relevant constructs, establishes a systematic framework for multimodal performance assessment within XR environments. The preliminary usability assessment with ten participants confirms that the platform is learnable and supports effective task completion, while identifying concrete interface improvements – particularly the need for a simplified analytics dashboard and more naturalistic audience behaviours – that have already informed a first round of prototype refinements.

Future work will focus on a more comprehensive evaluation with a larger and more diverse participant pool, longer multi-session studies at different difficulty levels, and validation of the composite metric models. Configurable

deployment profiles – a "lite" mode using only headset-native sensors for educational contexts, and a "full" mode incorporating wearables for research and clinical applications – will be explored to balance analytical depth with practical accessibility. The framework's applicability to AR and MR modalities, as well as broader communication training contexts, also presents promising directions for extending the approach.

## ACKNOWLEDGMENT

This work was supported by the Foundation for Science and Technology (FCT), contract doi.org/10.54499/UID/00127/2025. The authors would like to thank Cognitive3D for providing an academic license granting full access to the platform's premium features, which supported the spatial analytics capabilities presented in this work.

## REFERENCES

Al-Madani, F. M. (2015). Relationship between teachers' effective communication and students' academic achievement. European Journal of Educational Research, 4(2), 90–96.

Bertotti, C. (2022). Tapping the power of virtual reality to enhance public speaking. Harvard VPAL.

Carstens, A. (2019). Advice on the use of gestures in presentation skills manuals. Image & Text, 33.

Dănescu, D. F., & Romăşcanu, M. C. (2024). Social anxiety and speech anxiety. In Psychological Applications and Trends 2024, 616–620.

Gallego, A., McHugh, L., Penttonen, M., & Lappalainen, R. (2022). Measuring public speaking anxiety. Behavior Modification, 46(4).

Jim, M. E., Yap, J. B., Laolao, G. C., Lim, A. Z., & Deja, J. A. (2025). Speak with confidence: Designing an AR training tool for public speaking. arXiv:2504.11380.

Kang, N. (2016). Public speaking in virtual reality. PhD dissertation, Delft University of Technology.

Kroczek, L. O. H., & Mühlberger, A. (2023). Public speaking training in front of a supportive audience in VR improves performance in real-life. Scientific Reports, 13(1), 13968.

Marshall-Wheeler, N., Meng, Y., & Worker, S. (2022). Exploring public speaking self-efficacy. Journal of Extension, 60(4).

Meadows, S. (2019). Introduction to teaching public speaking. National Speech & Debate Association.

Mehmedova, E., Berrezueta-Guzman, S., & Wagner, S. (2025). Virtual reality user interface design. arXiv:2508.09358.

Morreale, S., Moore, M., Surges-Tatum, D., & Webster, L. (2007). The competent speaker speech evaluation form. NCA.

Niebuhr, O. (2025). On the cross-modal makeup of charisma. In Proc. Interspeech 2025, 4548–4552.

Niebuhr, O., Brem, A., Michalsky, J., & Neitsch, J. (2020). What makes business speakers sound charismatic? Cadernos de Linguística, 1(1), 1–40.

Ristorcelli, M. et al. (2025). Evaluating multimodal behavioral features for public speaking assessment in VR. In Proc. ACM IVA 2025.

Rodero, E., & Larrea, O. (2022). Virtual reality with distractors to overcome public speaking anxiety. Comunicar, 30, e007.

Rostami, M. et al. (2025). A comprehensive review of extended reality. Progress in Aerospace Sciences, 157, 101118.
Sarpourian, F. et al. (2022). The effect of VR therapy on students' public speaking anxiety. Health Science Reports, 5, e816.
Sonata, B., & Tetelepta, M. (2024). Psychological interventions to overcome public speaking anxiety. BIJ, 7(4), 14–21.
Sumaiya, B. et al. (2022). The role of effective communication skills in professional life. World Journal of English Language, 12(3), 134–141.
Van Ginkel, S. et al. (2017). Assessing oral presentation performance. Journal of Applied Research in Higher Education, 9(3).
Wörtwein, T. et al. (2015). Multimodal public speaking performance assessment. In Proc. ACM ICMI 2015.
Yang, K., Zhou, X., & Radu, I. (2020). XR-Ed framework. In Proc. ACM CHI 2020, 1–13.
Zuardi, A. W. et al. (2013). Human experimental anxiety: Actual public speaking induces more intense physiological responses. Revista Brasileira de Psiquiatria, 35(3), 248–253.