Sound-Driven Turn-Taking in Robot-Assisted Quizzes: How Non-Verbal Buzzing Boosts and Challenges Motivation⋆

Rezaul Tutul[1,†] and Ilona Buchem[2,†]

[1] Humboldt University of Berlin, 12489 Berlin, Germany
[2] Berliner University of Applied Sciences, 13353 Berlin, Germany

Abstract

This study investigates how sound-driven first-responder detection influences learner motivation and affect in a robot-assisted quiz moderated by the Pepper robot. In each round, Pepper presents a multiple-choice question, both teams press physical buzzer buttons, and a Random Forest–based classifier identifies which team buzzed first from overlapping mono-channel audio. Only the selected team is then allowed to answer verbally, after which Pepper provides immediate multimodal feedback through speech, gestures, and music before continuing to the next question. Sixty university students participated in a between-subject experiment comparing this sound-driven competitive mode with a sequential response mode. Post-intervention results from the Game Experience Questionnaire (GEQ) show that the sound-driven condition yields higher perceived challenge and tension but lower positive affect and slightly lower perceived competence, indicating increased arousal and excitement at the cost of emotional comfort. Qualitative feedback confirms that while the interaction is experienced as engaging and lively, recognition errors can lead to frustration and reduced trust. The findings demonstrate both the motivational potential and the emotional risks of competitive, sound-based interaction in educational human–robot interaction and highlight that technical reliability and transparent feedback are not merely engineering concerns but core enablers of sustainable engagement.

Keywords

Educational Robotics, Human-Robot Interaction, Motivation, Sound Recognition, Pepper Robot

1. Introduction

The use of social robots in education has expanded significantly over the past decade, providing new opportunities for interactive and motivating learning experiences [1, 2]. Robots such as Pepper can act as social facilitators that support collaboration, feedback, and engagement through speech, gesture, and emotional expression [3]. At the same time, gamification has become a central strategy in technology-enhanced learning, using points, badges, and competition to strengthen goal-oriented motivation and participation [4, 5]. When combined, social robots and gamified interaction offer powerful opportunities to create engaging, socially rich learning environments.

In this work, we investigate a robot-moderated quiz game in which Pepper orchestrates a competitive team-based interaction using non-verbal sound input. In each round, Pepper presents a multiple-choice question both verbally and on its tablet. Two teams then press physical buzzer buttons, each emitting a distinct sound. A machine-learning–based sound classifier analyzes the overlapping audio in real time and determines which team buzzed first. Only the selected team is then allowed to answer verbally, after which Pepper immediately responds with multimodal feedback using speech, gestures, and short music cues, updates the score display, and continues to the next question. This interaction loop creates a fast-paced, competitive, and socially expressive learning experience that differs fundamentally from conventional turn-taking or touch-based input.

⋆ EduRobotX 2025: 3rd International Workshop of the EATEL SIG, September 16, 2025, Newcastle, UK, co-located with the 20th ECTEL2025
1 buchem@bht-berlin.de.
† These authors contributed equally.
tutulrez@student.hu-berlin.de (R. Tutul); buchem@bht-berlin.de (I. Buchem)
0000-0002-8604-8501(R. Tutul); 0000-0002-9189-7217 (I. Buchem)

Sound-based interaction, such as buzzing or clapping, offers a spontaneous and physically engaging alternative to touch or speech input. It can intensify participation and competition but may also elevate arousal and tension, especially when technical errors occur. Prior work has shown that multimodal feedback combining gestures, music, and speech can improve user enjoyment, social presence, and engagement [6, 7]. However, few studies have examined how sound-driven first-responder detection under real-time conditions affects learners' motivational and emotional experience, particularly when decisions must be made from low-latency, overlapping mono-channel audio input where technical precision and reliability are critical [8].

From a psychological perspective, such interactions are inherently ambivalent. Competition and time pressure can increase excitement and challenge, but they can also induce stress, frustration, and a loss of perceived control when the system behaves unpredictably. In educational contexts, these emotional dynamics are especially important because motivation, perceived competence, and trust in the system strongly influence learning engagement. Yet, the motivational consequences of sound-driven, competitive turn-taking in robot-assisted learning remain underexplored.

To address this gap, this study develops and evaluates a sound-driven quiz system for the Pepper robot that detects the first buzzer sound using a Random Forest classifier and provides multimodal gamified feedback. Sixty university students participated in a between-subject experiment comparing a sound-driven competitive mode with a sequential response mode in which teams answer in a fixed order. Motivation and affect were assessed using the Game Experience Questionnaire (GEQ) alongside qualitative feedback. The main research question guiding this work is:

RQ: How does sound-driven first-responder detection in robot-assisted quizzes affect engagement and motivation compared to sequential interaction?

By linking motivation theory with concrete system design and empirical evaluation, this paper contributes to a deeper understanding of how competition, technical reliability, and emotional feedback interact in sound-driven educational human–robot interaction.

2. Theoretical Framework

This study draws on three complementary perspectives to interpret the motivational and affective mechanisms in sound-driven HRI: Self-Determination Theory (SDT), the Octalysis Gamification Framework, and the Yerkes-Dodson Arousal Law. Self-Determination Theory [9] explains intrinsic motivation through the fulfillment of autonomy, competence, and relatedness. In a quiz game, autonomy emerges when learners decide when to buzz; competence is reinforced through correct answers and rewards; and relatedness arises from team collaboration and interaction with the robot. When recognition errors occur, the perceived loss of competence or technical precision can undermine motivation. The Octalysis Framework [10, 14] conceptualizes eight core drives of motivation, including achievement, social influence, and unpredictability. The present system activates positive drives competition, feedback, and social collaboration but excessive unpredictability from misrecognition may introduce negative emotions, reducing engagement. Finally, the Yerkes-Dodson Law [11] posits that performance peaks at a moderate level of arousal and declines when arousal becomes too low or too high. Sound-driven competition can push arousal beyond the optimal range, increasing tension and frustration despite its stimulating effects [15]. Together, these frameworks clarify that sound-based interaction can simultaneously enhance and challenge motivation: it heightens engagement through competition and novelty yet risks emotional overload when reliability and technical precision are compromised.

3. System Design and Architecture

The proposed system integrates sound-driven turn-taking with multimodal gamified feedback in a Pepper robot-moderated quiz. Its architecture (Figure 1) comprises three layers Input, Processing, and Output forming a closed perception–action loop for interactive learning.

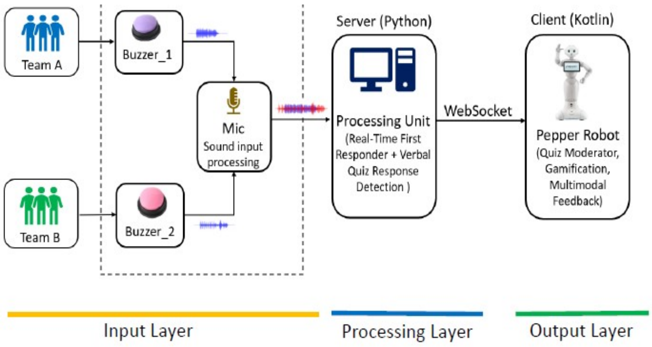

Figure 1: System architecture integrating sound input, Random Forest detection, and multimodal robot feedback.

The input layer captures buzzer sounds and verbal responses. Each team used a recordable buzzer button emitting a distinct sound. Sounds were received via a Bluetooth microphone receiver placed equidistantly between the two teams to ensure first responder detection technical precision. In parallel, Pepper's built-in microphones captured verbal responses through SoftBank QiSDK's Automatic Speech Recognition (ASR) module, which matched transcribed answers against a predefined answer list.

The core component is a Random Forest–based first-responder detection module trained to identify which team pressed first when sounds overlap [13]. The model uses three acoustic features extracted from 250-sample audio frames: (1) Onset Strength Difference: measures sound intensity over time, (2) Spectral Centroid Difference: captures brightness or frequency distribution, and (3) Onset Time Difference: estimates temporal gaps between detected sound peaks. These features were selected for real-time performance and robustness to classroom noise. The classifier, implemented in Kotlin and integrated with Android's audio APIs, outputs the detected team ID to Pepper's game controller, which coordinates quiz flow and scoring.

The output layer delivers multimodal gamified feedback, including speech, gestures, tablet visuals, and short music cues. When a team answers correctly, Pepper performs positive animations with "happy" music and verbal praise ("Correct! Great job!"). Incorrect answers trigger "sad" music and empathetic phrases ("Oops! Try next one!"). Points and badges are displayed on the tablet to reinforce motivation and competence.

Figure 2 illustrates the main game loop. Pepper welcomes players, explains rules, and presents quiz questions both verbally and on the tablet. A "*ding*" confirms buzzer detection, followed by verbal response collection. Scores update immediately after each question, and the robot concludes with an end speech and celebratory gesture for the winning team. The tablet interface provides real-time feedback, including color-coded buzz indicators, question progress, and total points. This interface supports both the sound-driven and sequential modes, enabling controlled comparison across conditions.

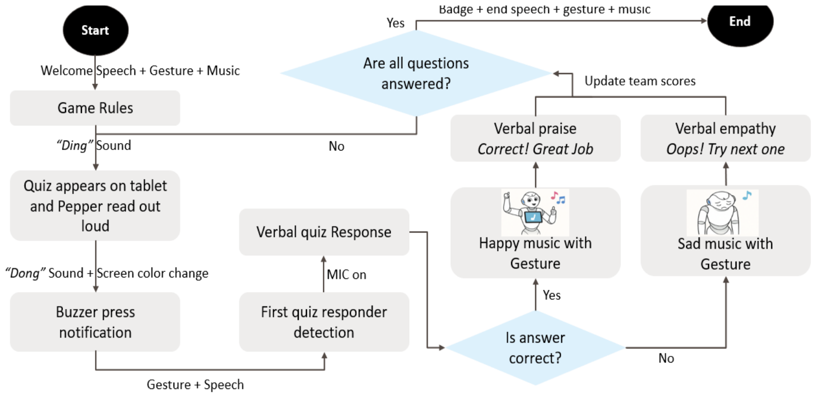

Figure 2: Game flow diagram illustrating sequential steps from question delivery to first responder detection, verbal answer, and feedback loop.

4. Methodology

4.1. Participants

The study involved 60 undergraduate students (aged 19–25) enrolled in introductory computer science courses at a German university. Participants had no prior experience interacting with humanoid robots. All participants provided informed consent prior to participation. Students were randomly assigned to either the experimental group (n = 30) or the control group (n = 30). Each experimental session consisted of two teams of five students competing against each other.

4.2. Materials and System

The experiment used a Pepper humanoid robot to moderate a team-based quiz game. The robot presented questions verbally and, on its tablet, coordinated turn-taking, collected answers, provided multimodal feedback, and displayed scores. Each team was equipped with a physical recordable buzzer button emitting a distinct sound. Buzzer sounds were captured via a Bluetooth microphone receiver positioned equidistantly between the two teams to minimize systematic bias in detection latency.

Verbal answers were captured using Pepper's built-in microphones and processed through SoftBank's QiSDK automatic speech recognition module, which matched transcribed answers against a predefined answer list.

The core technical component of the experimental system was a Random Forest–based sound classifier that identified which team buzzed first from overlapping mono-channel audio input. The classifier used three acoustic features extracted from short audio frames: onset strength difference, spectral centroid difference, and onset time difference. These features were selected for real-time performance and robustness in classroom environments. The classifier was implemented in Kotlin and integrated into the Android-based Pepper control system.

The robot provided multimodal feedback in both conditions, including speech, gestures, tablet visuals, and short music cues. Correct answers triggered positive animations and encouraging phrases, while incorrect answers triggered empathetic responses and negative feedback animations. Points and badges were displayed on the tablet in both conditions.

4.3. Conditions

The experiment followed a between-subject design with two interaction conditions that differed only in how the responding team was selected.

4.3.1. Experimental Condition: Sound-Driven First-Responder Mode

In the experimental condition, both teams were allowed to press their buzzer buttons simultaneously after each question was presented. The Random Forest classifier analyzed the overlapping audio signals and selected the team that buzzed first. Only the selected team's microphone channel was then enabled for verbal answering. If the non-selected team attempted to answer, their input was ignored by the system and Pepper prompted them to wait. After the selected team responded, Pepper provided feedback and the game continued with the next question.

4.3.2. Control Condition: Sequential Response Mode

In the control condition, teams answered questions in a fixed alternating order without any competition for response priority. No buzzer buttons were used. Pepper directly prompted the designated team to answer each question. Apart from the response selection mechanism, all other aspects of the interaction including robot behavior, feedback, scoring, timing, and quiz content were identical to the experimental condition. This design ensured that any observed differences in motivation or affect could be attributed specifically to the sound-driven competitive turn-taking mechanism.

4.4. Measures

Motivation and affect were assessed using the Game Experience Questionnaire (GEQ) core module [12], which consists of 33 items across seven subscales: Competence, Sensory and Imaginative Immersion, Flow, Tension/Annoyance, Challenge, Negative Affect, and Positive Affect. Items were rated on a 5-point Likert scale (1 = not at all, 5 = extremely).

In addition, participants answered three open-ended questions addressing (1) what they liked most about the system, (2) what they found frustrating or difficult, and (3) what they would suggest for improvement.

4.5. Procedure

Each experimental session lasted approximately 60 minutes and followed a standardized protocol:

1. Introduction (10 minutes): Pepper greeted participants, explained the rules of the game, and demonstrated either the buzzer-based interaction (experimental condition) or the sequential answering procedure (control condition).
2. Gameplay (40 minutes): Participants played a quiz consisting of ten multiple-choice questions on basic machine learning concepts. In the experimental condition, the Random Forest–based sound detection module identified the first responding team. In the control condition, teams answered in a fixed sequence. After each answer, Pepper provided multimodal feedback and updated the team scores displayed on the tablet.
3. Post-survey (10 minutes): Participants completed the GEQ questionnaire and the open-ended questions.

4.6. Data Analysis

Quantitative data were analyzed using independent-samples t-tests to compare GEQ subscale scores between the experimental and control groups. Effect sizes were calculated using Cohen's d. Qualitative responses were analyzed using thematic analysis. Responses were coded into recurring themes related to enjoyment, engagement, tension, collaboration, and technical reliability. Coding was performed by two researchers, and disagreements were resolved through discussion to reach consensus. The qualitative findings were used to contextualize and interpret the quantitative results.

5. Results

Post-intervention data from the Game Experience Questionnaire (GEQ) were analyzed to compare the experimental (sound-driven) and control (sequential) conditions. Descriptive statistics and independent-samples t-tests are presented in Table 1.

Table 1

Descriptive statistics and t-test results of GEQ subscales (N = 60)

Subscale	Experimental Mean (SD)	Control Mean (SD)	t	p	Cohen's d
Competence	3.70 (0.66)	4.04 (0.76)	-1.77	.08	-0.48
Immersion	3.12 (0.57)	3.41 (1.01)	-1.38	.17	-0.35
Flow	2.83 (0.81)	3.10 (0.71)	-1.34	.18	-0.35
Tension/Annoyance	2.61 (1.09)	1.69 (0.95)	3.33	.00	0.90
Challenge	2.66 (0.68)	2.34 (0.68)	1.80	.07	0.47
Negative Affect	2.42 (0.95)	1.99 (0.85)	1.77	.08	0.47
Positive Affect	3.38 (0.81)	4.05 (0.89)	-2.95	.00	-0.79

Results indicate significant differences in Tension/Annoyance ($p = .00$, $d = 0.90$) and Positive Affect ($p = .00$, $d = -0.79$), with the sound-driven group reporting higher tension and lower positive affect than the control group. Slightly higher Challenge and Negative Affect scores in the experimental group suggest that the competitive nature of buzzing introduced excitement but also stress. No significant differences were found in Flow, Immersion, or Competence, although means trended higher for the sequential condition. Overall, these results show that sound-based turn-taking enhanced challenge-driven engagement but also increased arousal and emotional strain, aligning with the Yerkes-Dodson prediction that excessive stimulation may reduce enjoyment. Open-ended responses were analyzed thematically, revealing six major themes summarizing participants' affective and motivational experiences (Table 2).

Table 2

Thematic summary of qualitative responses (experimental condition)

Theme	**Key Observations**
Fun & Enjoyment	Participants described the quiz as *"entertaining"* and "*interactive.*" They highlighted Pepper's gestures, music, and animations as "*fun and motivating.*"
Learning Motivation	Students felt the system "*made learning more exciting*" and "*encouraged them to pay attention and respond quickly.*" The points and badges increased engagement.
Tension & Challenges	Many found the competitive buzzer mode "*exciting but stressful.*" The fast pace and uncertain recognition created time pressure and anxiety.
Collaboration and Teamwork	Participants enjoyed working in teams, strategizing to buzz first, and celebrating together after correct answers.
Negative Affect (Recognition Errors)	Some students expressed frustration when "*the robot picked the wrong team*" or "*failed to recognize a buzz.*" Such errors were due to misclassification.
Human-Like Behavior	The integration of gestures, speech, and music was often described as making the robot "*feel alive*" and "*emotionally responsive.*"

Participants valued the competitive and multimodal features for making the quiz engaging and lively. However, many reported heightened tension and occasional frustration due to sound recognition errors, which diminished trust in the system. The coexistence of enjoyment and stress reflects a dual motivational effect: the sound-driven competition stimulated excitement but also risked emotional overload when recognition reliability was compromised. These qualitative results complement the quantitative findings, reinforcing that technical reliability and precision are essential for maintaining positive motivation in sound-driven educational HRI.

6. Discussion

This study examined how sound-driven first-responder detection influences motivation, affect, and engagement in a Pepper robot–moderated quiz game. The results reveal a characteristic dual effect: the sound-based competitive condition increased perceived challenge and tension while simultaneously reducing positive affect and slightly lowering perceived competence. This pattern indicates that the proposed interaction paradigm heightens arousal and excitement but at the cost of emotional comfort and motivational stability when technical reliability is not perfect.

From the perspective of Self-Determination Theory (SDT) [9], these findings can be interpreted primarily through the psychological need for competence. In the sound-driven condition, participants were given greater behavioral freedom: they could decide when to buzz, compete actively with the other team, and attempt to secure the right to answer. Under ideal technical conditions, this would be expected to strengthen both autonomy and competence. However, when the system occasionally misidentified the first responder, participants experienced a breakdown between their action and the system's response. Such mismatches undermine the perception that outcomes are contingent on one's own performance, thereby directly threatening perceived competence and fairness. This mechanism is reflected in both the quantitative trend toward lower competence scores and the qualitative reports of frustration when "the robot picked the wrong team." In other words, technical unreliability does not merely reduce usability; it disrupts a core psychological need that sustains intrinsic motivation.

The observed increase in tension and annoyance in the sound-driven condition can be further explained through the Yerkes–Dodson law [11], which posits an inverted-U relationship between arousal and performance or enjoyment. The sound-based competition introduces several arousal-amplifying factors simultaneously: strict time pressure, social competition between teams, public performance in front of peers, and uncertainty about whether the system will correctly recognize the buzzer input. Together, these factors likely push arousal beyond the optimal zone for many participants. While moderate arousal can enhance engagement and focus, excessive arousal manifests as stress, tension, and reduced positive affect exactly the pattern observed in the experimental condition. In contrast, the sequential control condition imposes a slower, more predictable rhythm that limits time pressure and uncertainty, thereby maintaining arousal at a more comfortable level and supporting higher positive affect and perceived competence.

Importantly, this interpretation also explains why the sound-driven condition was simultaneously experienced as more challenging and less pleasant. The interaction design succeeds in intensifying engagement through competition and urgency, but without sufficient technical transparency and reliability, this intensity turns into emotional overload rather than productive excitement. The findings therefore illustrate that arousal is not inherently beneficial; its motivational value depends critically on whether the system preserves a sense of control and fairness.

From the perspective of the Octalysis Gamification Framework [10], the system strongly activates several positive motivational drives, particularly achievement, social influence, and competition. Points, badges, team play, and multimodal robot feedback all contribute to a lively and socially engaging experience, as confirmed by participants' qualitative reports. At the same time, recognition errors introduce negative drives related to unpredictability, loss aversion, and perceived unfairness. This combination helps explain why enjoyment and frustration coexisted in the sound-driven condition: the same mechanics that amplify excitement also amplify disappointment when outcomes appear arbitrary.

More broadly, these results contribute to the HRI and educational robotics literature by demonstrating that technical reliability and transparency are not merely engineering qualities but psychologically central components of motivational interaction design. In competitive, time-critical scenarios, even small recognition errors can have disproportionate emotional and motivational consequences because they break the perceived link between effort and outcome. This finding aligns with emerging work emphasizing fluency, trust, and predictability as core dimensions of successful human–robot interaction [6, 8].

Taken together, the present findings suggest that sound-driven, competitive interaction is a powerful but fragile design choice. It can significantly increase challenge and engagement, but only if arousal is carefully regulated and the system preserves users' sense of competence, control, and fairness. Design strategies to achieve this balance may include explicit confirmation cues, transparent visualization of detection decisions, short response windows to regulate tempo, and error-handling behaviors in which the robot acknowledges uncertainty or ambiguity in recognition. Such mechanisms could help transform high-arousal excitement into sustainable, motivating engagement rather than emotionally costly tension.

7. Conclusion and Future Work

This paper investigated how sound-driven first-responder detection shapes motivation and affect in a Pepper robot–moderated quiz game. By comparing a competitive, sound-based interaction mode with a sequential response mode, the study showed that sound-driven competition increases perceived challenge and arousal but also elevates tension and reduces positive affect when technical reliability is not perfect. These results demonstrate that non-verbal, competitive interaction can be highly engaging but also emotionally fragile, especially in time-critical educational settings.

From a design perspective, the findings highlight that technical reliability and transparency are not merely implementation details but core psychological enablers of motivation, perceived competence, and trust in human–robot interaction. Sound-driven competition can only realize its motivational potential when users experience a stable and fair link between their actions and the system's responses.

Future work will focus on two concrete directions. First, we will improve robustness and trust by introducing multimodal redundancy (e.g., combining audio with tablet or visual confirmation cues) and explicit confidence visualization of detection decisions. Second, we will investigate adaptive interaction pacing and arousal regulation mechanisms, such as short response windows or dynamic tempo adjustment, to keep users within an optimal engagement zone. Together, these steps aim to transform high-arousal competition into a sustainable and emotionally supportive form of robot-assisted learning interaction.

8. Declaration on Generative AI

The authors declare that no generative artificial intelligence (AI) tools or systems were used in the conception, design, implementation, analysis, or writing of this work. All content, including the study design, software development, data analysis, and manuscript preparation, was produced solely by the authors.

References

[1] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka. Social robots for education: A review. Science Robotics, 3(21), 2018.

[2] A. Riedmann, P. Schaper, B. Lugrin. Integration of a social robot and gamification in adult learning and effects on motivation, engagement and performance. AI & Society, 39(1), 369–388, 2022. https://doi.org/10.1007/s00146-022-01514-y

[3] J. Qu, M. Jarosz, B. Sniezynski. Robot Control Platform for Multimodal Interactions with Humans Based on ChatGPT. Applied Sciences, 14(17), 8011, 2024. https://doi.org/10.3390/app14178011

[4] J. Hamari, J. Koivisto, H. Sarsa. Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. In: Proc. 47th Hawaii Int. Conf. on System Sciences (HICSS), pp. 3025 −3034, IEEE, 2014. https://doi.org/10.1109/hicss.2014.377

[5] R. Tutul, N. Pinkwart. Design and Evaluation of a Sound-Driven Robot Quiz System with Fair First-Responder Detection and Gamified Multimodal Feedback. Robotics, 14(9), 123, 2025. https://doi.org/10.3390/robotics14090123

[6] G. Hoffman. Evaluating Fluency in Human–Robot Collaboration. IEEE Transactions on Human-Machine Systems, 49(3), 209–218, 2019.

[7] R. Tutul, I. Buchem, A. Jakob, N. Pinkwart. Enhancing Learner Motivation, Engagement, and Enjoyment Through Sound-Recognizing Humanoid Robots in Quiz-Based Educational Games. In: Lecture Notes in Networks and Systems, pp. 123–132, Springer, 2024. https://doi.org/10.1007/978-3-031-66594-3_13

[8] A. Schäfer, R. Esterbauer, B. Kubicek. Trusting robots: a relational trust definition based on human intentionality. Humanities and Social Sciences Communications, 11(1), 2024. https://doi.org/10.1057/s41599-024-03897-3

[9] E. L. Deci, R. M. Ryan. The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268, 2000.

[10] Y.-K. Chou. Actionable gamification: Beyond points, badges, and leaderboards. Octalysis Media, 2015.

[11] R. M. Yerkes, J. D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurology and Psychology, 18, 459–482, 1908.

[12] W. A. IJsselsteijn, Y. A. W. de Kort, K. Poels. The Game Experience Questionnaire. Technische Universiteit Eindhoven, 2013.

[13] R. Tutul, I. Buchem, A. Jakob, N. Pinkwart. Millisecond-Level Sound Order Classification via Amplitude-Filtered Random Forest for Human-Robot Interaction. In: Proc. ACDSA, pp. 1–6, IEEE, 2025. https://doi.org/10.1109/acdsa65407.2025.11166423

[14] R. Tutul, & N. Pinkwart. Design and Evaluation of a Sound-Driven Robot Quiz System with Fair First-Responder Detection and Gamified Multimodal Feedback. Robotics, 14(9), 123, 2025. https://doi.org/10.3390/robotics14090123

[15] R. Tutul, I. Buchem, A. Jakob, & N. Pinkwart. Evaluating Intrinsic Motivation in Robot-Supported Quiz-Based Learning: A Comparative Study of Verbal-Only and Multimodal Feedback with Sound Input. Acta Informatica Pragensia, 15(1), 2026. https://doi.org/10.18267/j.aip.292