BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 74(3), 2026, Article number: e158302
DOI: 10.24425/bpasts.2026.158302

Gestures for attraction, gestures for communication: A field study of social robot interaction modalities at an educational fair

Tomasz GRZEJSZCZAK*

Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science, Department of Automatic Control and Robotics, Akademicka 16, 44-100 Gliwice, Poland

Abstract. This paper investigates human-robot interaction (HRI) within the challenging, high-noise environments of educational fairs, advocating for an HCI-centered approach to evaluate social robots in unpredictable public spaces. Utilizing the advanced social robot Furhat, the study compares the effectiveness of speech-based interaction against gesture-based recognition to overcome the limitations of traditional voice systems in loud settings. Five interaction programs were evaluated: a Large Language Model (LLM) variant relying on voice, and two gesture-controlled games (Rock-Paper-Scissors and Blocks guessing game) tested both with and without passive robot gestures. The results, drawn from field tests, demonstrate that the gesture recognition module is a highly effective alternative to speech recognition in noisy environments. While the voice-based LLM program struggled with a 33% success rate and high idle times due to environmental noise, gesture-based interactions achieved significantly higher success rates, ranging from 77% to 96%. Furthermore, the study confirms that a gesturing social robot is significantly more effective at attracting attention. The inclusion of passive gestures reduced the robot idle time from an average of 141.7–143.7 seconds to 105.6–106.5 seconds, while increasing participant engagement by 16% to 21%. These findings underscore the importance of non-verbal communication and multimodal perception in fostering reliable and engaging HRI in dynamic, high-social environments.

Keywords: human-robot interaction; HRI; social robot; gestural interaction; engagement and communication.

1. INTRODUCTION

Educational fairs are dynamic hubs connecting academia, industry, and future students. The atmosphere is vibrant and sensory, filled with colorful banners, digital displays, and interactive booths designed to attract visitors. The space is loud with discussions and presentations. For brands, these fairs are crucial for visibility. Research shows that interactive booth designs are particularly effective at capturing attention [1–3].

Social robots are ideally suited for this environment. Research indicates that humanoid robots with human-like traits, such as Furhat, are more successful at capturing user attention and fostering engagement [4, 5].

However, the loud and chaotic atmosphere of a trade fair presents a significant challenge for Human-Robot Interaction (HRI). While speech-based HRI is well-documented, its failure in high-noise, unpredictable public environments (like trade fairs) remains a major barrier. Speech-based interaction works in quiet settings, High Social Environments (HSEs) introduce noise that hinders data acquisition. The issue is that current HRI models often make users adapt to the system, not vice versa. This study addresses that gap by evaluating gesture-based interaction as a necessary redundant modality for maintaining interaction success in noisy, unstructured environments. Deploying Furhat in such a crowded setting revealed practical difficulties, such as the robot inability to identify who was speaking. To overcome this, a series of measures were implemented, including the use of a designated microphone, clear user instructions, and predefined protocols for the robot responses [6].

1.1. Contribution

The study was carried out during two events. The first one was an educational fair aimed at high school graduates, organized by the Department of Automatic Control, Electronics, and Computer Science at the Silesian University of Technology, which took place on September 19–20, 2024, in Gliwice, Poland (Fig. 1). This part of the research has been described in a separate paper, currently in press, serving as a foundation for the present study. The second event was the 9th Silesian Science Festival Katowice, held on December 6–8, 2024, at the International Congress Centre in Katowice, Poland, where the research was extended and further validated. The main objective of this research was to assess human-robot interaction (HRI) in dynamic and noisy environments typically found at large educational fairs. A Furhat social robot was stationed at the departmental exhibition stand, equipped with specialized software designed to enable natural and intuitive interactions with human attendees.

For this occasion, two separate interaction games were created and implemented to investigate the effect of robotic gestures on user involvement. These games were conducted in two

*e-mail: tomasz.grzejszczak@polsl.pl

Manuscript submitted 2025-10-09, revised 2026-03-01, initially accepted for publication 2026-03-03, published in May 2026.

Fig. 1. Education fair. September 19, 2024, Gliwice, Poland

experimental conditions: one featuring the robot gesturing capabilities activated and the other with gesturing turned off. The games were crafted to promote voluntary participation, utilizing the robot human-like traits to draw attention and encourage interaction. Comprehensive logs of the robot control programs, participant interactions, and system responses were carefully documented for later analysis.

The collected data provided a strong basis for testing the following hypotheses:

1. In environments with elevated ambient noise levels, an effectively functioning HRI workstation can be established using a gesture recognition module as a suitable alternative to a speech recognition module. This hypothesis addresses the difficulties posed by noisy environments, which frequently hinder the efficiency of speech-based communication.
2. A gesturing social robot captures more attention during fairs and exhibitions compared to a robot that does not exhibit passive gestures during interaction. This hypothesis examines the degree to which non-verbal signals, such as gestures, improve the robot capability to engage participants and sustain their interest.

This research highlights the promise of gesture-based interaction modules in surpassing the constraints of conventional speech recognition systems in noisy settings. Additionally, it underscores the importance of non-verbal communication in cultivating meaningful and engaging interactions between humans and robots. The results shared in this article have ramifications for the wider HRI field, providing practical insights for the use of social robots in educational, commercial, and public contexts.

2. LITERATURE REVIEW

The field of Human-Computer Interaction (HCI) has evolved significantly. This analysis classifies HRI literature and advocates for an HCI-centered perspective to evaluate interactive systems in dynamic settings like educational fairs. It also outlines key HRI frameworks for designing social robot interactions.

The interaction loop separates the human from the robot (Fig. 2). A simple analogy likens the human body to the robot physical structure, human senses to its perceptual algorithms, and the brain to its data processing functions.

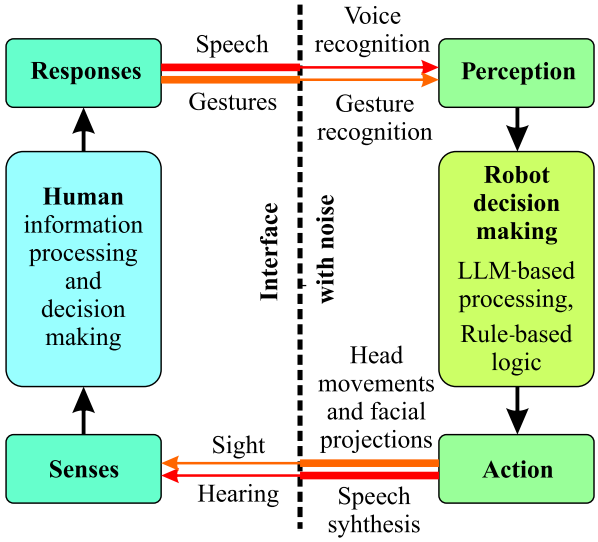

Fig. 2. Interface as a link between human and robot in HRI loop, with processed interaction channels

The efficacy of gestural HRI is rooted in the framework of embodied cognition, which posits that cognitive processes are fundamentally shaped by bodily experiences and physical engagement with the environment. In Human-Robot Interaction, this is operationalized through motor resonance: the activation of shared representations in the human brain when observing robotic actions that mirror human motor chains [7]. This resonance is essential for achieving interaction primacy, a phenomenon where non-verbal behaviors supply a wealth of information to onlookers in the opening minutes of an encounter, often preceding and framing the interpretation of subsequent verbal messages [8]. Because the human brain processes non-verbal cues faster than language, a gesturing social robot like Furhat can effectively get in the first word through physical embodiment, leveraging phylogenetic primacy - an evolutionary predisposition to prioritize nonverbal signals that has developed over 150 million years [2, 8].

Recent research distinguishes between Attributed Social Presence (ASP) – the user's subjective perception of a robot – and Intrinsic Social Presence (ISP), which represents the inherent capacity of a social agent to exhibit behaviors like non-verbal social cues that foster a sense of presence independently of verbal scripting [9]. While previous systems like ChildBot have successfully integrated multimodal perception in controlled edutainment tasks, they often rely on spontaneous data but remain constrained by laboratory spatial installs [10]. This work exceeds this state of knowledge by empirically testing the transition from

ASP to ISP in a real-world field study, showing that the robotic gestures are fundamental to creating co-experiance in social interaction [11].

2.1. Perception, action, and decision-making in social robotics

Effective human-robot interaction in social robotics requires a tightly integrated pipeline of perception, action, and decision-making. Perception enables robots to interpret their environment through techniques such as image processing, voice recognition, and sensor fusion. While Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) facilitate natural interaction, they struggle with noise and overlapping speech in crowded settings. Progress in beamforming and deep learning offers mitigation, and sensor fusion combining cameras, microphones, and LiDAR enhances situational awareness at the cost of increased computational complexity [12–14]. Adaptive multimodal approaches integrating auditory and visual data show promise for robustness in noisy environments [15], with complementary solutions such as wearable IMU-based controllers and touch sensors providing reliability when conventional sensing is impaired [16]. High Social Environments (HSEs) like educational fairs demand situated multimodal interpretation, as ASR performance degrades significantly with background noise, often leading to user frustration due to interaction failures [2, 17].

Action and actuation translate robot intentions into physical engagement, utilizing motor-driven joints, soft actuators, and hybrid mechanisms that balance precision, expressiveness, and safety. These hardware components define the robot action space and are essential for conveying emotions and functions during interaction (Fig. 2).

Planning languages like PDDL enable logical action sequencing but struggle with dynamic environments. Reinforcement Learning (RL) learns optimal policies through reward maximization, though it demands substantial data and computation. Imitation Learning (IL) acquires skills from human demonstrations, proving useful when rewards are ambiguous. Other paradigms such as transfer learning, meta-learning, and Model Predictive Control further expand the methodological landscape [18].

3. MATERIALS AND METHODS

This section details the instruments, technologies, and strategies utilized in the research, highlighting their contributions to fostering a vibrant and interactive atmosphere for participants. The objective of this chapter is to elucidate the procedural and technical elements of the deployed solutions. Among those discussed in the literature review, the interaction is based on methods depicted in Fig. 2.

3.1. Social robot

Furhat is an advanced social robot designed for natural human-robot interaction. It combines sophisticated hardware, software, and a flexible design to operate in dynamic settings. Its core feature is a unique projection system that displays dynamic, lifelike facial expressions on a 3D-printed head, allowing it to convey a wide range of emotions. This visual experience is enhanced by motors that enable human-like head movements such as nodding and turning.

For this study, Furhat was programmed using its RemoteAPI and a corresponding Python library, allowing it to function in a semi-open environment. It encompasses a collection of predefined functions, such as `furhat.gesture(name="Nod")`, which triggers a gesture change on the projector and maneuvers the head actuators in the appropriate sequence, `furhat.say(text=x, lipsync=True)` to articulate a sentence with lip synchronization and the specified voice, or `furhat.attend(user="CLOSEST")` to utilize the camera for individual detection and to maintain eye contact.

3.2. Voice recognition

Voice recognition is an essential element of social robots, enabling them to comprehend and react to human speech. The primary component of the implemented Python code utilizes the speech_recognition library in Python, which offers a flexible and user-friendly interface for incorporating voice recognition functionalities.

The library supports Google's Web Speech API as a streaming recognizer, permitting real-time transcription of spoken language. Additionally, the library features built-in capabilities for reducing ambient noise, which improves recognition accuracy in noisy settings typical of educational fairs. These characteristics make the speech_recognition library a practical option for designing voice-enabled social robots.

3.3. Gesture recognition

Initial studies on gesture recognition algorithms concentrated on rule-based approaches [19, 20], but with the evolution of deep learning, numerous high-performance algorithms are now accessible.

MediaPipe, an open-source framework created by Google, has become a premier solution for real-time gesture recognition. It offers a comprehensive set of tools for accurately detecting and tracking human poses, hands, and facial landmarks. MediaPipe's structure is crafted to be lightweight and efficient, making it well-suited for use on devices with limited computational power.

The framework harnesses advanced machine learning models to process video input and extract essential features. MediaPipe Hands utilizes a combination of palm detection and hand landmark localization to ascertain the position and orientation of hands in real time. This capability is crucial for enabling the robot to interpret a variety of non-verbal signals.

Open-source models that are pre-trained are utilized to identify hand gestures based on landmarks recognized by MediaPipe. Depending on the specific use case, the primary model categorizes gestures into [$Thumb\_Up$, $Thumb\_Down$, $Other$], while a secondary model identifies gestures as [$Rock$, $Paper$, $Scissors$]. An example of gesture recognition through the localization and classification of hand feature points is illustrated in Fig. 3.

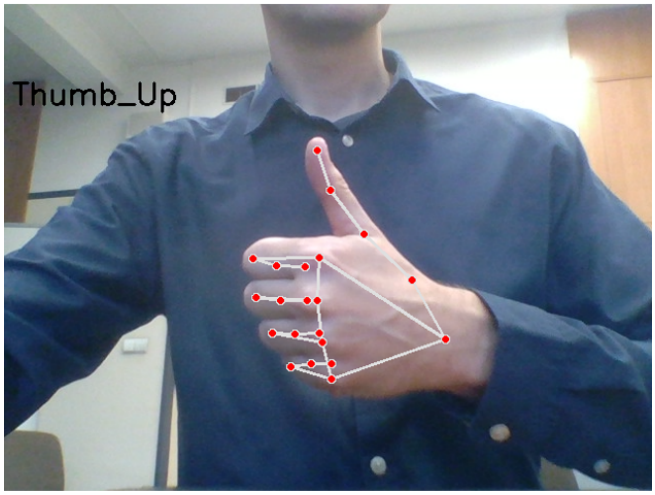

Fig. 3. Gesture recognition using MediaPipe

3.4. Interactions

To summarize this chapter, the interaction programs developed integrate various robot perception mechanisms and actions within the robot action space. The perception includes both voice recognition and gesture recognition, whereas the robot actions encompass speech synthesis and gesture execution. The interaction algorithms implemented primarily utilize rule-based planning techniques.

The amalgamation of these interaction abilities forms the foundation for creating three distinct interaction programs.

The simplest program leverages a Large Language Model (**LLM**), which detects spoken words, converts them into text, uses the text as a prompt for a chat-based language model, and directs the model output to the robot speech synthesizer.

3.4.1. Rock Paper Scissors game

The first game is the well-known Rock Paper Scissors game (**RPS**). In this game, the user can present one of the three gestures representing rock, paper, or scissors. Simultaneously, a random gesture is displayed on a screen adjacent to the robot (Fig. 4). The user's gesture, along with its corresponding classification, is also shown in real time. The robot engages with the user by speaking sentences from Table 1. Each gesture has a winning and losing counterpart, providing both players with an equal $1:1:1$ chance of a win, loss, or tie.

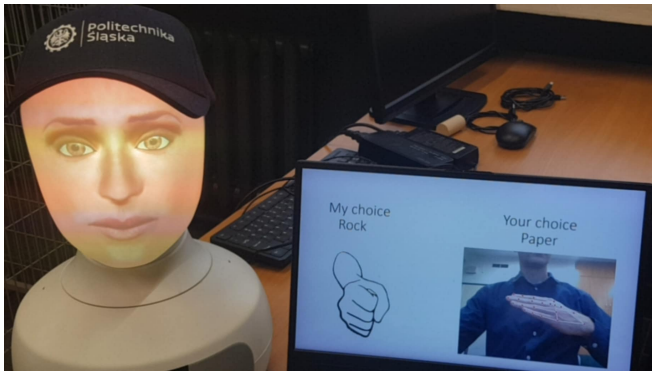

Fig. 4. Rock Paper Scissors interaction game with a robot

Table 1
RPS game sentences

Hi, would you like to play a rock paper scissors game? Let me count to 3 and let's show a gesture.
3, 2, 1, go!
Congratulations, you win.
Hurray, I won.
Would you like to try again?
Sorry, I didn't see any gesture. Please be sure to point it out in front of the camera and let's try again.

3.4.2. Blocks guessing game

The upcoming game is a block arrangement guessing game (**Blocks**), where participants are presented with 3 blocks and can freely arrange them using 3 available positions on the table, either stacking them or choosing to omit some blocks from the setup. This method of arranging blocks is inspired by the Blocks World, a well-known planning problem domain [21]. Throughout the game, the robot goal is to ask questions that help deduce the arrangement of the blocks. Once successfully guessed, the arrangement is displayed on the screen.

For 3 blocks, there are 51 possible arrangements ($a = 51$). Listed in Table 2 are the sentences spoken by the robot, which include 7 parameterized questions (marked with *). The variables X and Y are substituted with appropriate colors from the list $[red, yellow, blue]$. Figure 5 showcases examples of these arrangements. At the start of the game, a matrix A is generated, where the number of rows corresponds to each arrangement, and the number of columns corresponds to each question. There are 24 potential questions ($q = 33$) that can be formulated as combinations of elements from the X and Y lists. The matrix A retains true or false values that represent responses to a specific

Table 2
Blocks arrangement guessing game sentences

Hi, would you like to play a game? You can see some blocks in front of me. Please arrange them as you wish.
Show me a thumb up if you are ready to start.
Ok, let me start with first question. To answer, show thumb up or thumb down.
Is there a X block in your set? *
Is the X block on the table? *
Is the X block on Y block? *
Is the X block on position Y? *
Are there X blocks directly on the table? *
Are there X blocks in your set? *
Is there anything on top of X block? *
I think I know the arrangement. Is it correct?
Oh, something must have gone wrong.
Hurray, I guessed it.
Would you like to try again?
** This is a parametrized question. X and Y are the names of blocks*

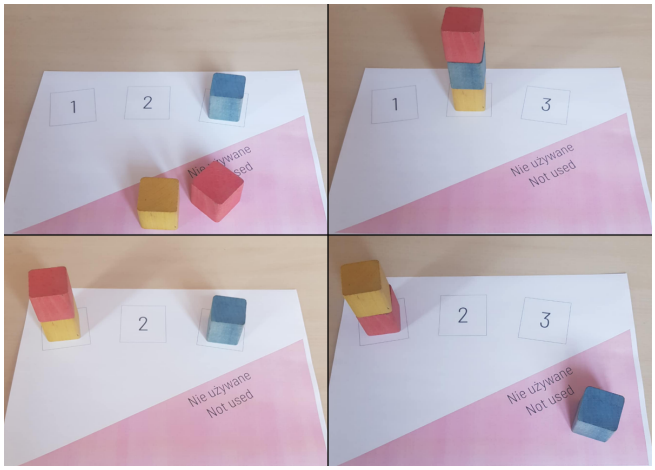

Fig. 5. Example arrangements from the blocks guessing game

question q for each arrangement a. Therefore, the dimensions of the matrix A_{aq} for the set of 3 blocks is 51×33.

During the game, for each guessing round, the selected question is the one that most evenly splits the remaining possible arrangements, approximately in half. The sum

$$S_q = \sum_{a=1}^{n} A_{aq} \tag{1}$$

of all true values is calculated for each column in A. Question to be asked index i points the question for which the sum of true answers is closest to half of arrangements count a

$$\underset{i}{\operatorname{argmin}} \left| S_q - a/2 \right|, \tag{2}$$

where $i \in \{1, 2, \ldots, q\}$.

When user provides the answer ans as true or false, the rows from A for which A_{ai} differs from ans are deleted from matrix A. This creates new matrix

$$A' = \{a_{rj} \mid a_{ri} = ans, \forall j\} \tag{3}$$

with r rows indicating possible arrangements. For the next question, matrix $A \leftarrow A'$. Now there are less possible arrangements, as $a \rightarrow r$. The formulas (1)–(3) are executed in loop as long as there is only one possible arrangement $a = 1$.

3.4.3. Performing gestures

Furhat is capable of performing specific gestures, including nodding, winking, smiling, expressing sadness, looking in a chosen direction, and maintaining eye contact. Both games are expanded to include a gesture variant (+G), wherein an appropriate gesture is executed after each statement from Tables 1 or 2.

In this variant, the facial tracking feature is utilized for the individual nearest to the robot. When initiating the game, the robot winks in an encouraging manner after posing questions. A gesture of sadness along with a nodding of the head follows a failure or loss, while winning elicits a smile and a head lift. Prior to posing the next question, the robot adopts a thoughtful pose after delivering an answer.

The compilation of three developed programs, along with their two variants, results in a total of five interaction programs that were trialed at the educational fair. The capabilities of the interaction programs are detailed in Table 3.

1. **LLM** – Large Language Model voice-based conversation
2. **RPS** – Rock Paper Scissors game with gesture recognition
3. **RPS + G** – Rock Paper Scissors with robot gestures
4. **Blocks** – Block configuration guessing game
5. **Blocks + G** – Block guessing game with robot gestures

Table 3
Robot interaction programs

	Recognition		Actions	
	voice	gestures	speech	gestures
LLM	✓		✓	
RPS		✓	✓	
RPS+G		✓	✓	✓
Blocks		✓	✓	
Blocks+G		✓	✓	✓

4. TESTS

As outlined in the test section (3.4), five types of interaction programs were executed on the Furhat social robot. Each program operated for a total of 3 hours (with a minimal standard deviation of $\sigma = 103[s]$). The logs of program operations were recorded, and the following metrics were acquired:

- N – Number of played games
- P – Number of participants
- S – Number of successful interactions
- $t_b(p)$ – Break time between participants (after p'th person), when the robot was idle.
- $t_I(p)$ – Interaction time for a person p. The time that one person interacts with robot, which is the game duration.

To assess the statistical significance of the differences between two independent groups, the unpaired (independent) t-test is utilized. This method is suitable for comparing the means of two groups when the samples do not affect one another. There is no need to examine the correlation of t_I, as it is clear that each game involves different interaction times; however, to evaluate **Hp2**, the statistical significance of t_b is analyzed. To assess significance, the t value of the test statistic is computed using the formula

$$t = \frac{\overline{t_{b1}} - \overline{t_{b2}}}{\sqrt{\frac{s_p}{P_1} + \frac{s_p}{P_2}}}, \tag{4}$$

where s_p is Cohen's polled standard deviation. Next, the t-distribution table is used to find the p-value associated with the computed t-statistic and degrees of freedom. By comparing the p-value to the significance level ($\alpha = 0.05$), if the p-value is

smaller than α, the null hypothesis should be rejected, suggesting a significant difference between the means of the groups. Otherwise, the null hypothesis is not rejected.

To address the need for user-centered data and participant demographics, a structured questionnaire was administered following the interactions at the educational fair. The questionnaire result is presented in Table 4. It is organized into four primary sections: Part 1 (Demographics) collects data on participant age and gender to account for the diverse spectrum of attendees, such as high school students and academic researchers; Part 2 (Interaction Modality & Usability) employs a 5-point Likert scale to evaluate the effectiveness of the input methods, specifically focusing on the clarity of communication and user frustration in the fair noisy environment; Part 3 (Perception of Robot Gestures & Personality) assesses the subjective impact of the robot's facial expressions and head movements on engagement and its perceived "aliveness"; and Part 4 (Qualitative Feedback) provides open-ended queries regarding interaction challenges and general suggestions for improvement. This multidimensional assessment ensures that the technical logs of interaction time (tI) and success rates (S) are balanced with empirical evidence of user perception.

Table 4
The most important and highest results of a survey conducted on 86 participants

Category	Key metric (from Questionnaire)	Result (Mean/%)
Demographics	High school students (< 18)	51%
	University students (18–25)	27%
Usability	"Robot understood my inputs clearly" (1–5)	Gesture: 4.4 Voice: 2.1
Usability	"Ease of communication in noise" (1–5)	Gesture: 4.2 Voice: 1.8
Engagement	"Robot gestures made it more enjoyable" (1–5)	+G: 4.7 Basic: 3.2
Perception	"Robot seemed engaged" (1–5)	+G: 4.5
User state	"Felt frustrated during interaction" (1–5)	LLM: 4.1 Games: 1.5
Personality	Description: "Friendly" or "Intelligent" (%)	88% "Friendly" 56% "Intelligent"

5. RESULTS AND DISCUSSION

5.1. Performance metrics and hypothesis testing

To evaluate hypothesis **Hp1**, one should examine Table 5 alongside the graph in Fig. 6, which is derived from it. The analysis compares two variables: the number of games played, denoted as N, and the number of participants, P, who engaged with the robot during a $3[h]$ operational period. The **LLM** interaction program employed voice recognition technology, while other programs relied on gesture recognition assuming the robot was unable to hear. The most significant distinction is seen in the number of games played; however, it is essential to consider that these elevated figures stem from individual users repeating interactions multiple times. Participants of the **LLM** typically posed around $4-5$ prompts, while users of **RPS** played approximately 3–4 games, whereas those using **Blocks** generally engaged in just one game, which was longer in duration. Hence, this information cannot be utilized to validate hypothesis 1.

Table 5
Mean times and numbers of interaction with robot

	$\overline{t_b}$ [s]	$\overline{t_I}$ [s]	N	P	S
LLM	205.6	104.5	218	52	17
RPS	141.7	52.9	307	81	78
RPS+G	105.6	55.8	327	94	89
Blocks	143.7	126.9	61	57	44
Blocks+G	106.5	122.9	73	69	54

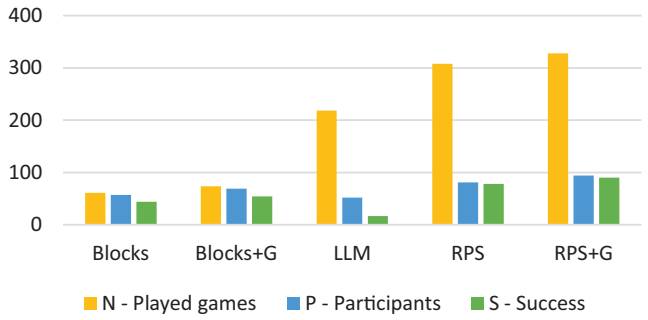

Fig. 6. Comparison of number of played games N, interacting participants P and successfully ended interactions S for 5 types of interaction programs

Two noteworthy relationships emerge from the data presented. The **LLM** displays both the fewest number of participants and the lowest rate of successfully completed interactions. By contrasting P with S, it is possible to determine the success rate of the interaction programs. For **RPS**, the success rate is roughly 96%, while for **Blocks**, it stands at around 77%; in contrast, the **LLM** only achieved a successful interaction rate of 33%, with users often leaving before completion. Adequate comparative data would suffice to assess hypothesis 1, but in this scenario, the results favor gesture-based solutions. The findings suggest that not only can gesture recognition serve as a feasible alternative to speech recognition, it may even outperform it in certain cases.

Statistical analyses, as outlined in Section 4, are conducted on the data shown in Table 5 and presented in Table 6. This analysis examines the break intervals between interactions, specifically the downtime necessary for the robot to capture the attention of the subsequent user. The observations drawn from Table 6 lead to the following conclusions:

1. There is a significant difference between **LLM** and other programs, with extremely significant difference between gesture variants (**+G**).

2. There is no significant difference between **RPS** and **Block**, comparing both basic and gesture variants (+**G**).
3. There is a significant difference between an interaction program and its gesture variant (+**G**).

Table 6
Statistical test data (p and t value from (4))

	1. Blks+G		2. Blocks		3. RPS+G		4. RPS		5. LLM	
	p	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*
5.	0.00	7.42	0.01	3.62	0.00	8.45	0.01	3.52	1	0
4.	0.03	2.33	0.88	0.16	0.07	3.02	1	0		
3.	0.71	0.36	0.01	2.66	1	0				
2.	0.07	2.06	1	0						
1.	1	0								

Following the observations mentioned above, it can be concluded that voice interaction (like **LLM**) is distinct from gesture-based interaction in terms of capturing attention (1). This difference arises primarily from the nature of the interaction (voice versus gesture). It is possible that individuals gesturing in front of the robot might draw the attention of others in the crowd, though this hypothesis cannot be confirmed in this study. It is important to point out that significantly fewer attendees came to the stand to engage. This was due to the lengthy idle gaps between user interactions. Both gesture-based games exhibited the same level of attention attraction. The main distinction between them is that one encourages participation in multiple games, while the other focuses on a single longer game; however, this difference does not affect the statistical analysis, as noted in (2).

The most significant finding, as indicated in (3), is that incorporating gesture capability into the robot results in a notable decrease in idle time during interactions. Analyzing the data presented in Table 5, it can be estimated that 16%–21% more participants engaged with the robot that utilized gestures. Moreover, the interactions tended to last longer, as users played approximately 6%–19% more games. This data supports the hypothesis 2, suggesting that a gesturing social robot is more capable of drawing attention during trade shows and exhibitions than one that does not utilize gestures during interaction.

5.2. User perception and qualitative feedback

A striking disparity 33% success for voice based LLMs versus 95-96% for gestures validates the Technology Acceptance Model (TAM) [22] in social robotics Multimodal feedback including gestures and sound boosts Perceived Usefulness (PU) and Behavioral Intention (BI) by fostering emotional connection and reducing effort [22, 23]. To advance HRI design for public spaces I propose modality failover in noisy HSEs like fairs systems should prioritize gesture recognition to avoid ASR related hallucinations and frustration ISP calibration embed Intrinsic Social Presence behaviors such as nodding eye contact and winks as ostensive cues to establish interaction primacy and epistemic trust before speech and co-experience scaffolding design robotic games as open laboratories that enable co-experiences with bystanders positioning the robot as a social actor rather than a tool.

5.3. Design guidelines for social robots in HSEs

Based on the empirical findings of this field study and the supporting HRI literature, I propose four actionable guidelines for the design of social robots intended for deployment in unstructured public spaces:

1. Prioritize gestural input in noisy HSEs: When ambient noise exceeds 75 dB, ASR reliability often drops below 35%. Gestural recognition should be a mandatory fail-over primary modality, maintaining 95–96% success and reducing frustration.
2. Establish interaction primacy through ISP: Robots should use non-verbal cues (e.g., eye contact, nodding) to establish interaction primacy before speech. This phylogenetic approach positions the robot as a social agent .
3. Implement visual feedback redundancy: Robots should visually communicate their perceptual state (e.g., displaying recognized gestures). This transparent feedback fosters epistemic trust—users reported high enjoyment (4.7/5) even during lags.
4. Design for co-experience, not just task completion: In HSEs, prioritize shared play over pure efficiency. Games like Rock-Paper-Scissors transform the robot into a social actor, sustaining engagement and converting bystanders into participants.

6. CONCLUSIONS

This study investigates whether gesture recognition can effectively replace speech for human-robot interaction (HRI) in noisy environments like educational fairs. The Furhat robot, and three interaction programs were used to test two hypotheses: (1) if gesture recognition is a viable alternative to speech in loud settings, and (2) if gesturing robots are more appealing to users.

Three interaction programs were created, with one version enabling robot gestures and another disabling them for comparison. Results showed the gesticulating robot attracted more participants and sustained greater engagement. The gesture recognition module proved to be a robust substitute for speech, sometimes outperforming it in user attraction within the chaotic fair environment.

First-hand observations confirmed that the speech recognition system frequently failed due to crowd noise, leading to user frustration and abbreviated interactions. In contrast, occasional errors in gesture recognition were more tolerable to users, fostering a more conducive environment for sustained dialogue.

This research provides empirical evidence for the superiority of gestures in noisy, dynamic HRI, highlighting the need for robust, multimodal interaction systems.

The findings confirm reports from the design guideline for social robots. To maximize user engagement and trust, systems must intentionally provide ostensive cues (explicit signals like eye contact or directed gesturing) that lead users to feel recognized as subjects and establish epistemic trust in the robot as a reliable communication partner [7, 24].

ACKNOWLEDGEMENTS

This research was funded by the Silesian University of Technology (SUT) through the subsidy for maintaining and developing the research potential grant in 2026.

REFERENCES

[1] T. Bauer and V. Hantel, "Built to attract: Evaluating trade show booth designs using attention analysis in a live communication context," *J. Conv. Event Tour.*, vol. 23, no. 3, pp. 240–268, 2022, doi: 10.1080/15470148.2021.1988022.

[2] J.A. Duncan, F. Alambeigi, and M.W. Pryor, "A survey of multimodal perception methods for human–robot interaction in social environments," *ACM Trans. Hum.-Robot Interact.*, vol. 13, no. 4, pp. 1–50, 2024.

[3] C.H. Park, R. Ros, S.S. Kwak, C.-M. Huang, and S. Lemaignan, "Towards real world impacts: Design, development, and deployment of social robots in the wild," *Front. Robot. AI*, vol. 7, p. 600830, 2020.

[4] E. Giannitzi, "When eyes meet laughter: Exploring non-verbal cues in human-robot interaction with furhat," Master's thesis, University of Gothenburg, Department of Philosophy,Lingustics and Theory of Science, 2024. [Online]. Available: https://gupea.ub.gu.se/bitstreams/be0caa4a-0c86-4071-bf11-d71c99606242/download

[5] D.E. Kulathunga, "Exploring the experience of different robot personalities in enhancing university students' learning," Master's thesis, Tampere University, Faculty of Information Technology and Communication Sciences, 2024. [Online]. Available: https://trepo.tuni.fi/bitstream/handle/10024/157790/KulathungaDakshika.pdf?sequence=2

[6] S. Al Moubayed *et al.*, "Furhat goes to robotville: A large-scale multiparty human-robot interaction data collection in a public space," in *International Workshop on Multimodal Corpora, Tools, and Resources.*, Istanbul, Turkey, 2012.

[7] C. Fang, L. Peternel, A. Seth, M. Sartori, K. Mombaur, and E. Yoshida, "Human modeling in physical human-robot interaction: A brief survey," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5799–5806, 2023.

[8] J.K. Burgoon, V. Manusov, and L.K. Guerrero, *Nonverbal communication*. Routledge, 2021.

[9] N.H. Wijesinghe, Y. Peng, S. Konrad, M. Jayasuriya, J.B. Grant, and D. Herath, "Reframing social presence for human robot interaction," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 1716–1721.

[10] N. Efthymiou *et al.*, "Childbot: Multi-robot perception and interaction with children," *Robot. Autonom. Syst.*, vol. 150, p. 103975, 2022.

[11] D.C. Herath, E. Jochum, and E. Vlachos, "An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings," *IEEE Trans. Cogn. Dev. Syst.*, vol. 10, no. 4, pp. 1096–1105, 2017.

[12] R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[13] K. Pondel-Sycz, A.P. Pietrzak, and J. Szymla, "End-to-end deep neural models for automatic speech recognition for polish language," *Int. J. Electron. Telecommun.*, vol. 70, no. 2, pp. 315–321, 2024, doi: 10.24425/ijet.2024.149547.

[14] D.P. Nguyen *et al.*, "Socially aware motion planning for service robots using lidar and rgb-d camera," *arXiv preprint arXiv:2410.09803*, 2024, doi: 10.48550/arXiv.2410.09803.

[15] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Soc. Robot.*, vol. 11, pp. 555–573, 2019, doi: 10.1007/s12369-019-00524-z.

[16] E.Y. Zhang, Z. Pan, and A. D. Cheok, "Emotion recognition using affective touch: A survey," *IEEE Trans. Affect. Comput.*, vol. 17, pp. 1–20, 2026, doi: 10.1109/TAFFC.2025.3592197.

[17] S. Ojha, F. Gervits, and C. Espy-Wilson, "Speaking with robots in noisy environments," in *20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 1057–1061.

[18] M. Lapan, *Deep reinforcement learning hands-on*. Packt Publishing Ltd, 2024.

[19] T. Grzejszczak, M. Kawulok, and A. Galuszka, "Hand landmarks detection and localization in color images," *Multimed. Tools Appl.*, vol. 75, pp. 16 363–16 387, 2016, doi: 10.1007/s11042-015-2934-5.

[20] T. Grzejszczak, R. Molle, and R. Roth, "Tracking of dynamic gesture fingertips position in video sequence," *Arch. Control Sci.*, vol. 30, no. 1, pp. 101–122, 2020, doi: 10.24425/acs.2020.132587.

[21] A. Galuszka and A. Swierniak, "Planning in multi-agent environment using strips representation and non-cooperative equilibrium strategy," *J. Intell. Robot. Syst.*, vol. 58, pp. 239–251, 2010, doi: 10.1007/s10846-009-9364-4.

[22] R. Tutul, I. Buchem, A. Jakob, and N. Pinkwart, "Technology acceptance in university robot-supported quiz-based learning: Verbal-only versus multimodal feedback with sound input," *IEEE Access*, vol. 14, pp. 956–965, 2025.

[23] B. Panigrahi, A.H. Raj, M. Nazeri, and X. Xiao, "A study on learning social robot navigation with multimodal perception," *arXiv preprint arXiv:2309.12568*, 2023.

[24] M. Saito and S. Yamada, "How adaptation and user engagement affect trust in audio guide agents," *IEEE Access*, vol. 13, pp. 81 553–81 568, 2025, doi: 10.1109/ACCESS.2025.3567067.