

Article

Control Modality and Accuracy on the Trust and Acceptance of Construction Robots

Daeguk Lee ¹, Donghun Lee ² , Jae Hyun Jung ³  and Tazoon Park ^{1,*}

¹ Department of Industrial & Information Systems Engineering, Soongsil University, Seoul 06978, Republic of Korea; daeguklee@soongsil.ac.kr

² Department of Mechanical Engineering, Soongsil University, Seoul 06978, Republic of Korea; dhlee04@ssu.ac.kr

³ Department of Chemical Engineering, Soongsil University, Seoul 06978, Republic of Korea; nfejjh@ssu.ac.kr

* Correspondence: tzpark@soongsil.ac.kr; Tel.: +82-2-828-7034

Abstract

This study investigates how control modalities and recognition accuracy influence construction workers' trust and acceptance of collaborative robots. Sixty participants evaluated voice and gesture control under varying levels of recognition accuracy while performing tiling together with collaborative robots. Experimental results indicated that recognition accuracy significantly affected perceived enjoyment (PE, $p = 0.010$), ease of use (PEOU, $p = 0.030$), and intention to use (ITU, $p = 0.022$), but not trust, usefulness (PU), or attitude (ATT). Furthermore, the interaction between control modality and accuracy shaped most acceptance factors (PE, $p = 0.049$; PEOU, $p = 0.006$; PU, $p = 0.006$; ATT, $p = 0.003$, and ITU, $p < 0.001$) except trust. In general, high recognition accuracy enhanced user experience and adoption intentions. Voice interfaces were favored when recognition accuracy was high, whereas gesture interfaces were more acceptable under low-accuracy conditions. These findings highlight the importance of designing high-accuracy, task-appropriate interfaces to support technology acceptance in construction. The preference for voice interfaces under accurate conditions aligns with the noisy, fast-paced nature of construction sites, where efficiency is paramount. By contrast, gesture interfaces offer resilience when recognition errors occur. The study provides practical guidance for robot developers, interface designers, and construction managers, emphasizing that carefully matching interaction modalities and accuracy levels to on-site demands can improve acceptance and long-term adoption in this traditionally conservative sector.

Keywords: human–robot interaction; collaborative robot; user acceptance; control modality



Academic Editors: Emilio Jorge González-Galván and Marco Mendoza

Received: 27 September 2025

Revised: 31 October 2025

Accepted: 4 November 2025

Published: 6 November 2025

Citation: Lee, D.; Lee, D.; Jung, J.H.; Park, T. Control Modality and Accuracy on the Trust and Acceptance of Construction Robots. *Appl. Sci.* **2025**, *15*, 11827. <https://doi.org/10.3390/app152111827>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With advances in sensor technologies, control systems, and safety-certified designs, robots have gradually worked safely side by side with human workers. Collaborative robots, or cobots, are equipped with force-limiting actuators, real-time vision systems, and smart feedback mechanisms that enable them to adapt to dynamic environments and safely interact with people. Cobots are increasingly used across various industries to enhance efficiency, improve safety, and support human workers. Some representative examples include their use in shipbuilding [1], logistics [2], manufacturing [3], food and beverage [4], and pharmaceutical industries [5]. As compared to other industries, the construction industry has been slow to adopt cobots because its work environments are unstructured, dynamic, and exposed to the outdoors, featuring irregular tasks that pose

significant challenges [6]. Unlike structured factory settings, dynamic construction sites require humans and cobots to coordinate complex tasks in real time. This makes designing effective interfaces particularly difficult. Based on the needs, this study investigated actionable guidance for selecting and configuring interfaces, thereby supporting near-term adoption and day-to-day usability of cobots in construction.

1.1. Characteristics of Construction Industry

The construction industry has traditionally been characterized by labor-intensive processes, volatile working conditions, and fragmented project delivery methods. These persistent challenges have prompted researchers and practitioners to seek automation solutions that can improve safety, productivity, and quality. Recent attention has turned toward collaborative robots to address labor shortages, reduce workplace hazards, and enhance construction efficiency. Unlike traditional industrial robots, which often require physical barriers or safety cages, cobots are equipped with sensors and intelligent controls that enable real-time interaction with human co-workers [7]. As such, cobots hold the promise of increasing productivity without replacing human labor outright, instead augmenting skilled workers' capabilities.

The construction industry presents a uniquely challenging context for the adoption of cobots, due to its unstructured environments, dynamic workflows, and reliance on manual labor. Unlike manufacturing environments, where cobots have been widely adopted in structured settings with repeatable tasks, construction sites are characterized by irregular terrains, variable lighting, unpredictable schedules, and constantly changing configurations [6]. These characteristics introduce substantial complexity in integrating cobots, particularly regarding mobility, perception, and task adaptability.

One major constraint is the low level of standardization across construction projects. Tasks often vary significantly by location, scale, and design, which makes pre-programmed automation less effective. Cobots, which excel in repetitive, semi-structured operations, require contextual awareness and real-time adaptability to operate safely and productively on-site [8,9]. Furthermore, construction work typically involves close human collaboration, requiring cobots to interact not only with physical materials but also with human co-workers. This amplifies the need for robust safety mechanisms, intuitive interfaces, and real-time interaction models.

1.2. Cobots in Construction Industry

Despite these challenges, there have been successful demonstrations of cobots in specific construction tasks. For example, the semi-autonomous rebar tying robot developed has been deployed on bridge construction projects, significantly reducing labor time and physical strain on workers [10]. Other successful applications include robotic systems for drywall installation, bricklaying, and on-site 3D concrete printing, which have demonstrated the potential for cobots to improve consistency, speed, and worker safety in repetitive or ergonomically taxing tasks [11]. These projects showcase how task-specific cobots can be effectively integrated into workflows when environmental variables are controlled and when human–robot task division is well-designed.

However, limitations remain in scalability and generalization. Many of the successful deployments have been in pilot or controlled demonstration settings, often with significant pre-site calibration or constrained operational areas. Moreover, interaction interfaces in these systems are often rudimentary, relying on tablet controls or manual overrides, which may not suit multitasking workers who are constantly shifting between roles. As a result, the full potential of cobots in construction is yet to be realized. To move forward, future research must address not only robotic autonomy and sensing, but also human–robot interaction (HRI) tailored to construction contexts.

Preferred mode of front-end control on construction sites would be gesture or voice because these natural modalities align with the site's practical constraints—workers' hands and gaze are often occupied, ambient noise is high, layouts change frequently, and safety demands minimal workflow interruption. Gesture interfaces are acoustically immune and increasingly robust thanks to vision/depth sensing, while depth sensors mitigate prior lighting issues and support intuitive, one-handed command vocabularies [12,13]. Voice interfaces offer hands-free, eyes-free control with strong performance when signal-to-noise is managed [14].

Taken together, a good combination of gesture/voice with proper recognition accuracy would facilitate the acceptance of cobots in a construction site. Recent advancements in machine learning improved multimodal interaction with robots [15,16], but the empirical accuracy of gesture/voice recognizers needs to be verified. Voice and gesture are natural, low-friction modalities but fail for different reasons: voice degrades with signal-to-noise ratio and reverberation; gesture degrades with camera field-of-view, occlusion, lighting, and personal protection equipment. Accordingly, we anticipate that recognition accuracy will moderate modality preferences, such that the relative acceptability of voice versus gesture-including interfaces varies across accuracy conditions. We include trust as a hypothesized outcome because HRI theory links trust to antecedents—reliability, transparency/observability, predictability/legibility, and workload—that differ by front-end modality. Recognition accuracy operationalizes reliability, a primary driver of trust calibration. Thus, we test modality, accuracy, and their interaction on trust alongside acceptance [17,18]. The following are the proposed research questions and hypotheses.

RQ1: *How do modality and recognition accuracy influence trust and acceptance?*

H1 (modality): *Control modality (gesture, voice, gesture + voice) affects users' trust and technology acceptance.*

H2 (accuracy): *Higher recognition accuracy increases users' trust and technology acceptance.*

H3 (interaction): *Recognition accuracy and control modality interact such that the impact of accuracy on attitude and trust varies across modalities.*

2. Methods

An empirical study was conducted to examine the effects of control modality (voice and gesture) and recognition accuracy on user trust and acceptance of a collaborative robot. Volunteer participants performed a tile-laying task, chosen to simulate direct human–robot interaction in a realistic work environment, using interfaces with varying accuracy levels. Following the interaction, participants' trust and acceptance were measured using a questionnaire survey. The study protocol was reviewed and approved by the Institutional Review Board (SSU-202202-HR-413-1), and all participants provided written informed consent before participating in the experiment.

2.1. Experimental Setup

A collaborative tiling scenario was implemented using a 6-DOF industrial cobot (UR10, Universal Robots, Odense, Denmark) equipped with a four-cup vacuum gripper, a vertical mounting panel lined with hook-and-loop fabric, and fifteen acrylic surrogate tiles of full size plus pre-cut edge pieces. The cobot, pre-programmed in RoboDK, placed full-size tiles in a specified sequence; the participant then seated each tile using a rubber mallet and installed the pre-cut pieces to close perimeter gaps. A floor-level pickup zone was marked to the robot's right for tile retrieval. Interaction hardware comprised an action camera

(for gesture capture), a close-talk microphone (for voice input), and a loudspeaker (for feedback prompts). To approximate site conditions, illuminance and ambient noise were reproduced from field measurements; audio recorded at an active tiling site was looped and level-matched to the measured sound-pressure level. An overview of the apparatus—including the robot, vacuum gripper, tiles, vertical panel, pickup zone, rubber mallet, camera–microphone–speaker set, and participant workspace—is provided in Figure 1.

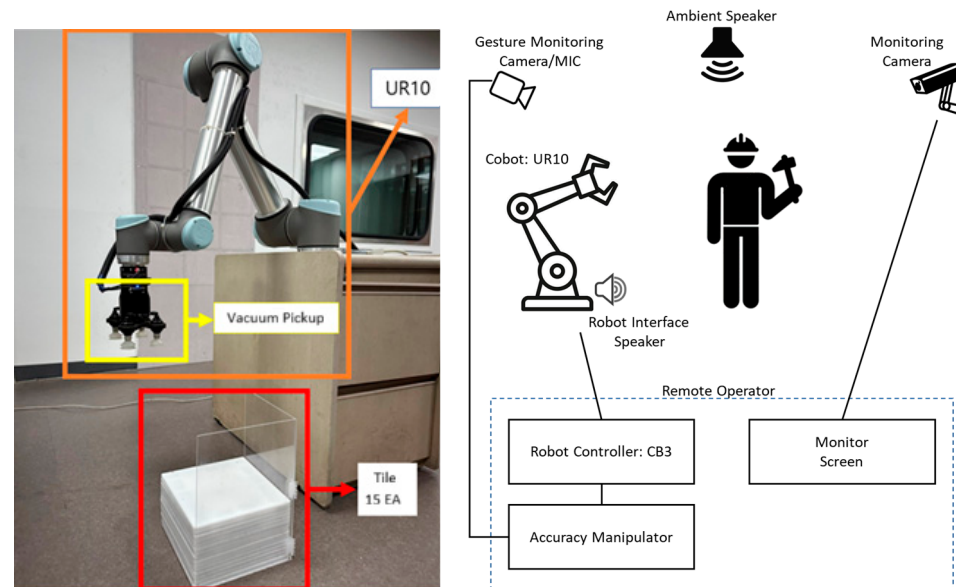


Figure 1. Experimental setup and schematic diagram of robot control systems.

2.2. Independent Variable Manipulations

2.2.1. Modality

Gesture and voice are widely regarded as natural, intuitive means of expression and thus enable fluid human–robot communication [19]. However, they have different characteristics in production, coding, environmental susceptibility, and operational implications. Gestures require hand/arm motion, while voice relies on vocal fold excitation and articulation; thus, gestures remain viable when speech is masked or disruptive, whereas voice works when hands are occupied or fine motor gestures are infeasible. Gestures are camera-dependent and directional—effective only within the sensor’s field of view and free of occlusion—whereas voice is effectively omnidirectional from the user’s standpoint, making it viewpoint agnostic but sensitive to signal-to-noise ratio. Encoding diverges as well: gesture recognition performs spatiotemporal coding of pose and trajectory into symbolic commands, while speech recognition performs verbal (phonetic to lexical) coding into words. In practice, gestures degrade with occlusion, poor lighting, clutter, and personal protection equipment; voice degrades with high noise, reverberation, and competing talkers but improves with close talk headsets and constrained grammars.

On construction sites, combining the two strengthens the control channel; augmenting gesture commands with voice makes the interaction closely resemble human–human communication [20]. Accordingly, the interface factor in this experiment included three levels: gesture-only, voice-only, and gesture + voice (multimodal).

Based on the prior pilot study, the command set comprised six words: Emergency Stop, Pre command, Start, Stop, Faster, and Slower. The primary control objective was to regulate the cobot’s motion speed. Emergency Stop had the highest priority and did not require a pre-command [21]. For all other commands, a pre-command was issued before the main instruction. In the gesture condition, participants delivered the pre-command

and the main command as a continuous sequence of movements. In the voice condition, participants first performed a “call” behavior—functionally akin to addressing the robot by name—before stating the command. In the combined condition, participants used both simultaneously, but any one input identified first was accepted as a valid command.

2.2.2. Accuracy

To operationalize channel quality in the human–robot interface, we manipulated recognition accuracy at two levels—high and low—informed by prior reports on gesture and voice performance. Published results place gesture recognition in the 93–99.16% range across static and dynamic command sets (e.g., [13]), whereas voice recognition spans 66–98%, with accuracy degrading as speaker–microphone distance increases and as ambient noise rises [22]. Consistent with these findings, voice was treated as more environmentally sensitive than gesture.

Accordingly, we specified the following target accuracies for the experimental manipulation. In the high-accuracy condition, both gesture and voice interfaces were set to 95% recognition. For the multimodal (gesture + voice) condition, we adopted a disjunctive fusion assumption—i.e., a command is accepted if either modality recognizes it—yielding a combined target of 99.75% ($1 - (1 - 0.95)^2$). In the low-accuracy condition, gesture and voice were each set to 80%, with the multimodal target set to 96% under the same fusion rule. These values reflect realistic single-modality ranges observed in the literature while capturing the expected robustness gains from redundant, complementary modalities. After each command, recognition success or failure was communicated to participants via an auditory cue played through the speaker, providing immediate feedback without interrupting task flow.

2.3. Dependent Variables and Measurements

The dependent variables were analyzed on the basis of trust and technology acceptance. In addition to defining each construct, we describe the measurement instruments employed. All measures were administered individually to participants immediately after each experimental session. The questionnaires were originally authored in English and were translated into Korean for administration.

Trust is a critical determinant of successful human–robot interaction and a prerequisite for user acceptance; accordingly, this study measured participants’ trust in the collaborative robot. Operationally, trust was defined as the participant’s belief in, and willingness to work with, the cobot under each control modality. To quantify trust across combinations of control modality and recognition-accuracy level, we adopted the trust in automation questionnaire, which is widely used in robotics and related fields [23]. The instrument comprises 12 items rated on a 7-point Likert scale.

We posit that users’ intention to adopt collaborative robots is positively associated with trust; consequently, acceptance may vary across experimental conditions. To assess the effects of control modality and recognition-accuracy level on acceptance, participants completed a questionnaire following each session based on the Technology Acceptance Model (TAM) [24] and the instrument proposed by Park and Del Pobil [25]. The survey included five constructs, each with four items: Perceived Enjoyment (PE), Perceived Ease of Use (PEOU), Perceived Usefulness (PU), Attitude toward Use (ATT), and Intention to Use (ITU). Individual items were adapted to the context of this study.

2.4. Experimental Design

To compare the relative efficacy of interaction modalities, it was essential that each participant experience all levels of the control-modality factor; therefore, a within-subjects design was used for control modality, and the order of modality conditions was randomized for each

participant to mitigate idiosyncratic bias and order effects. By contrast, recognition accuracy was manipulated between subjects: exposing a participant to both accuracy levels would confound factor effects and introduce unavoidable learning because the same tasks are repeated under each level. A between-subjects manipulation ensures that participants exposed to one accuracy level cannot influence outcomes under the other, and participants were randomly assigned to accuracy groups to minimize between-group differences. Thus, the resulting design is a mixed 3 (control modality: gesture, voice, gesture + voice) $\times$ 2 (recognition accuracy: high, low) factorial. Each participant completed only the three modality conditions within the assigned accuracy level (gesture, voice, gesture + voice)

2.5. Procedures

Participants attended individual sessions in the laboratory. Upon arrival, each participant was escorted to the experimental room and provided with a briefing on the study's procedures and safety precautions, after which they provided written informed consent. The experimenter controlled the cobot from an adjacent control room—a configuration known as a Wizard-of-Oz experiment—observing the session through a one-way mirror. Before the task, participants viewed an instructional video and were informed that the system's camera and microphone enabled gesture and voice recognition.

The primary task consisted of a collaborative tile-laying sequence as outlined in Figure 2. Upon the subsequent “start” cue, the cobot began picking tiles from the pickup zone and placing them on the wall, after which the participant secured each placement with a rubber mallet. During the sequence, scripted audio prompts elicited speed adjustments—after tiles 1 and 13, participants issued a faster command, and after tiles 7 and 15, they issued a slower command. A logistics event was introduced after tile 10, when a cue required the participant to pause the cobot, replenish the stock of large tiles at the pickup zone, and restart the operation on instruction. In addition, when the cobot placed tiles 1–3, the participant affixed a small tile at position 4 (left of tile 3), and the same small-tile procedure was repeated for the remaining positions. Throughout the task, small tiles were retrieved exclusively from the green basket, and participants were not permitted to pre-place or hold tiles in advance. During the task, participants responded to randomly issued emergency-stop prompts by issuing the emergency-stop command. They are also allowed to adjust the speed even without a prompt.

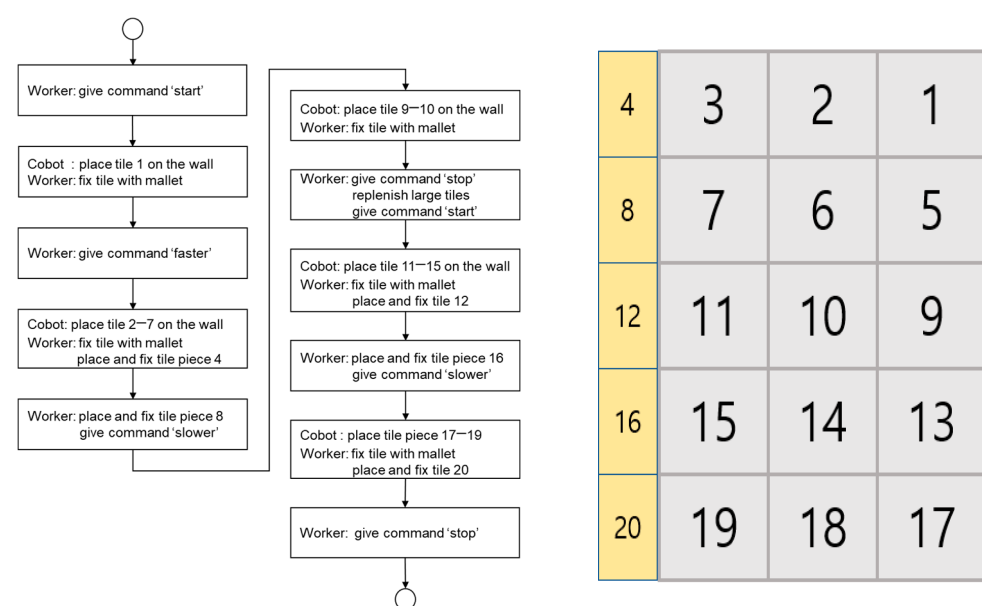


Figure 2. Sequence of experimental tasks and the order of the tiles being attached.

3. Results

3.1. Descriptive Statistics

A total of 60 participants conducted the experiment. In the high-accuracy group, 30 participants (17 men, 13 women; mean age = 23.40, SD = 2.67) and in the low-accuracy group, 30 participants (15 men, 15 women; mean age = 23.23, SD = 3.38)—comprising university staff, graduate students, and undergraduates—volunteered. Only five participants had more than 10 h of prior experience with collaborative robots; three had less than 5 h; and 52 had no prior experience.

The average and standard deviation of responsible variables are listed in Table 1. There were no large differences across control modalities: for example, trust averaged 5.32 (SD = 1.08) with gesture, 5.20 (1.10) with voice, and 5.36 (0.95) with gesture + voice; perceived enjoyment was likewise similar—5.52 (1.33), 5.55 (1.41), and 5.51 (1.49), respectively. By contrast, the accuracy manipulation showed a consistent pattern: the high-accuracy condition outperformed the low-accuracy condition on every measure—TRU 5.42 (0.85) vs. 5.16 (1.13), PE 5.94 (1.11) vs. 5.11 (1.30), PEOU 5.69 (0.88) vs. 5.14 (1.02), PU 5.81 (1.07) vs. 5.29 (0.97), ATT 5.46 (1.10) vs. 4.92 (1.14), and ITU 5.51 (1.10) vs. 4.84 (1.10). Reliability was high, with Cronbach's α ranging from 0.803 (PEOU) to 0.957 (PE), and all other scales ≥ 0.94 , indicating excellent internal consistency.

Table 1. Average and reliability of participants' responses according to modality and accuracy.

Variables		Modality **			Accuracy		Cronbach's Alpha
		G	V	G + V	High	Low	
TRU *	M (SD)	5.32 (1.08)	5.20 (1.10)	5.36 (0.95)	5.42 (0.85)	5.16 (1.13)	0.918
PE	M (SD)	5.52 (1.33)	5.55 (1.41)	5.51 (1.49)	5.94 (1.11)	5.11 (1.30)	0.957
PEOU	M (SD)	5.39 (1.16)	5.46 (1.20)	5.40 (1.16)	5.69 (0.88)	5.14 (1.02)	0.803
PU	M (SD)	5.46 (1.20)	5.62 (1.28)	5.57 (1.33)	5.81 (1.07)	5.29 (0.97)	0.944
ATT	M (SD)	5.17 (1.47)	5.15 (1.42)	5.24 (1.37)	5.46 (1.10)	4.92 (1.14)	0.941
ITU	M (SD)	5.24 (1.34)	5.01 (1.52)	5.27 (1.39)	5.51 (1.10)	4.84 (1.10)	0.940

* TRU: trust, PE: perceived enjoyment, PEOU: perceived ease of use, PU: perceived usefulness, ATT: attitude, ITU: intention to use, (Likert scale 1–7); ** G: gesture, V: voice, G + V: gesture and voice.

3.2. Inferential Statistics

To examine the interaction effects of the independent variables on participants' trust and technology acceptance after interacting with the collaborative robot, we conducted a mixed ANOVA. Assumption checks indicated no violations. For the within-subjects effects (modality) and the modality $\times$ accuracy interaction, Mauchly's test of sphericity was non-significant ($p = 0.21$). Levene's tests for homogeneity of variances and Shapiro–Wilk tests of residual normality revealed no departures from model assumptions.

Control modality did not have a statistically significant effect on trust ($F(2, 118) = 0.382$, $p = 0.683$). Similarly, technology acceptance measures are not significant for PE ($F(2, 118) = 0.011$, $p = 0.989$), PEOU ($F(2, 118) = 0.061$, $p = 0.941$), PU ($F(2, 118) = 0.233$, $p = 0.793$), ATT ($F(2, 118) = 0.075$, $p = 0.928$), and ITU ($F(2, 118) = 0.603$, $p = 0.548$). Recognition accuracy did not have a statistically significant effect on trust ($F(1, 58) = 1.029$, $p = 0.315$), PU ($F(1, 58) = 3.739$, $p = 0.058$), or ATT ($F(1, 58) = 3.446$, $p = 0.069$). In contrast, accuracy significantly affected PE ($F(1, 58) = 7.077$, $p = 0.010$), PEOU ($F(1, 58) = 4.971$, $p = 0.030$), and ITU ($F(1, 58) = 5.509$, $p = 0.022$). For all significant effects, higher recognition

accuracy yielded higher scores, i. e. greater perceived enjoyment (PE), greater perceived ease of use (PEOU), and stronger intention to use (ITU) relative to low accuracy. The interaction between control modality and recognition accuracy was not significant for trust (F(2, 116) = 0.186, $p = 0.812$). In contrast, significant interaction effects were found in the acceptance variables. Significant interactions were observed for PE (F(2, 116) = 3.134, $p = 0.049$), PEOU (F(2, 116) = 5.347, $p = 0.006$), PU (F(2, 116) = 5.454, $p = 0.006$), ATT (F(2, 116) = 6.500, $p = 0.003$), and ITU (F(2, 116) = 10.113, $p < 0.001$). This result indicates that voice commands are preferred when recognition accuracy is high, but as accuracy declines, users shift preference toward interfaces that include gestures (see Figure 3).

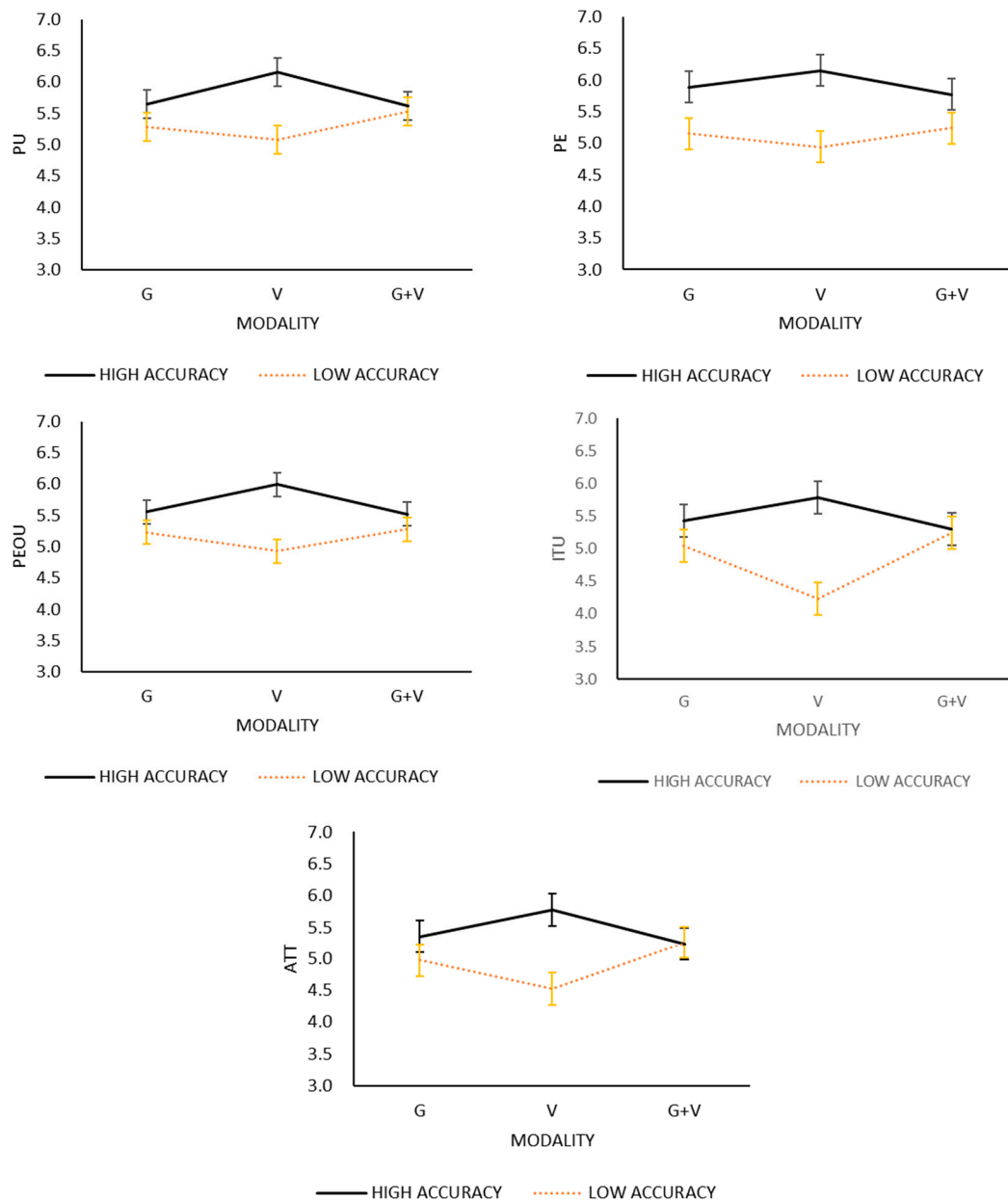


Figure 3. Interaction plot of modality $\times$ accuracy G: gesture, V: voice, G + V: gesture + voice.

4. Discussion

This study examined how control modality and recognition accuracy influence users' trust and acceptance when applying collaborative robots to construction. Interface design is central to human–robot interaction because it mediates how users express task intent. When formulating modality hypotheses, we posited that, on construction sites, the burden

of speaking loudly enough for voice commands to overcome noise would outweigh the inconvenience of two-handed work or holding tools; accordingly, we expected a preference for gesture. However, the main effect of control modality was not statistically significant for any dependent variable. This diverges from prior reports that gestures are preferred in noisy settings [21], and may reflect that some participants found gesture commands more difficult than voice, or that the sample size lacked power to detect modest effects.

Recognition accuracy is a design factor that directly affects communication with the robot, and we anticipated a preference for higher accuracy on construction sites. Results indicate that higher accuracy was indeed beneficial: perceived enjoyment, perceived ease of use, and intention to use were higher under high accuracy. In contrast, trust, perceived usefulness, and attitude toward use were not significantly affected by accuracy. The null effects on trust fit with trust scholarship: meta-analyses and frameworks emphasize that trust is driven by perceived reliability over time, transparency, and appropriate automation behavior, and it often changes slowly relative to short laboratory exposures. In other words, short exposure to collaborative work sessions may have been underpowered to shift trust, even while acceptance facets (PE, PEOU, ITU) were sensitive to recognition accuracy [26]. Longer-horizon or longitudinal designs, richer explanations/visualizations of robot intent, and repeated performance with few errors are typically needed to calibrate trust. Also, heterogeneity in participants' prior robot familiarity and exposure may dilute effects.

Given site noise, we further expected constraints on voice at low accuracy. Combining control modality with recognition accuracy yielded significant interactions for PE, PEOU, PU, ATT, and ITU. Voice with high recognition accuracy increased acceptance, consistent with participants' perception that accurate speech recognition did not interrupt workflow. Conversely, under high accuracy, gesture reduced acceptance because issuing gestures was perceived to break task flow; gesture + voice also reduced acceptance despite its higher nominal recognition rate, as simultaneous gesturing and speaking introduced cognitive load. Under low accuracy in noisy conditions, voice was perceived as more disruptive than gesture-including interfaces, indicating that these interfaces were more effective when recognition reliability was limited. These results echo prior work reporting that speech is often the most intuitive channel when signal-noise ratio and microphones are well-managed, whereas gesture is robust to acoustic noise and thus helpful in harsh shop-floor conditions [19]. Overall, voice outperformed gesture-including interfaces at high accuracy across enjoyment, ease of use, usefulness, attitude, and intention, aligning with the view that voice is more intuitive and natural than gesture [27]. At low accuracy, gesture-including interfaces outperformed voice, consistent with evidence that gestures are advantageous in noisy environments [21].

The gesture + voice interface was sometimes less acceptable due to the added cognitive load moderates the standard "multimodal is better" claim. Multimodal HRI typically assumes complementary or disambiguating fusion (e.g., speech "place there" + pointing), but simultaneous, redundant issuing can create dual-task costs. Future designs should favor semantic fusion (sequential or role-split inputs) and lightweight confirmations, as recommended by multimodal interface studies and AR-for-robotics work that emphasize transparency and operator focus [19]. The mixed results align with broad human-robot collaboration surveys showing that "no single interface wins everywhere"; instead, usability depends on task, environment, and system design. Industrial reviews consistently recommend natural user interfaces coupled with clear feedback and safety scaffolds, but also warn that environmental constraints strongly shape performance—precisely the pattern you observed across accuracy conditions [9].

In a representative scenario, the cobot's interface would adapt to the operational context. It would default to voice control in low-noise environments and autonomously

switch to gesture recognition when ambient noise increases, confirming the transition with audio-visual cues. Should noise levels drop while the user's speech is impeded by a mask or face shield, the system would adopt a hybrid model, reverting to voice for general commands but retaining gestures for safety-critical functions. This strategy ensures workflow continuity and user acceptance across dynamic conditions. These findings are likely to generalize to other domains, such as mining and agriculture, that share core characteristics with construction, like high ambient noise, variable lighting, frequent occlusions, and workers operating with personal protection equipment (PPE) while multitasking. In such settings, the same modality–accuracy tradeoff could apply. Accordingly, integrators in these domains can adopt a context-adaptive strategy, prioritizing microphone quality and constrained vocabularies, and shifting toward camera-robust gestures and semantic multimodal fusion when visibility and noise conditions fluctuate.

5. Conclusions

This study investigated how front-end control modality (gesture, voice, and gesture + voice) and recognition accuracy shape users' trust and technology acceptance when deploying collaborative robots to construction. Against the backdrop of robots' evolution from caged automation to human-proximate cobots, and the distinctive variability of construction sites, we focused on natural user interfaces that can sustain real-time coordination under noise, occlusion, and frequent task switching. Using a tile-laying scenario that approximated site constraints, we found no main effect of control modality on any outcome (Trust: $p = 0.683$; PE: $p = 0.989$; PEOU: $p = 0.941$; PU: $p = 0.793$; ATT: $p = 0.928$; ITU: $p = 0.548$), but clear effects of recognition accuracy and robust modality $\times$ accuracy interactions. Recognition accuracy significantly improved PE ($p = 0.010$), PEOU ($p = 0.030$), and ITU ($p = 0.022$), with no effects on Trust, PU, or ATT. Critically, modality and accuracy interacted for all acceptance variables except Trust—PE ($p = 0.049$), PEOU ($p = 0.006$), PU ($p = 0.006$), ATT ($p = 0.003$), and ITU ($p < 0.001$). Interaction effects showed that voice was preferred and more acceptable when recognition accuracy was high, whereas interfaces that include gestures were favored when accuracy was low, reflecting voice's sensitivity to signal-to-noise ratio and gesture's resilience to acoustic interference. Notably, simultaneous gesture + voice did not uniformly improve acceptance, as concurrent issuing introduced cognitive load despite higher nominal recognition rates.

These findings translate into practical design guidance for on-site cobot interfaces. First, prioritize accuracy: improvements in recognition quality directly elevate user experience and willingness to adopt. Second, adopt a context-adaptive, multimodal strategy rather than a single default: use voice as the primary channel where a good signal-to-noise ratio can be ensured (e.g., close-talk mics, beamforming, constrained grammars), and pivot to gesture-including interaction when noise or reverberation degrades speech. Third, implement semantic fusion instead of redundant simultaneous commands—e.g., quick voice intents with gesture for disambiguation—paired with lightweight confirmations and clear feedback to preserve workflow continuity. Finally, integrate these interfaces within a shared-autonomy control stack so the robot handles low-level precision and safeguards while humans retain strategic oversight, aligning with construction's demand for fast adjustments and high situational awareness.

This work has several limitations that bound the interpretation and generalizability of the results. First, the study was conducted in a controlled laboratory using a Wizard-of-Oz configuration and a single, stylized tile-laying task. Although the task approximated site constraints (noise, logistics, shared workspace), it cannot capture the full variability of construction. Second, exposure was brief, and outcomes were primarily self-reported (trust and TAM constructs). Trust, in particular, typically calibrates over repeated, reliable

interactions; short sessions may underdetect change and are vulnerable to common-method variance. Third, recognition accuracy was experimentally manipulated to target preset rates rather than measured from a production-grade ASR/vision stack in situ. This improves experimental control but reduces ecological validity because real systems exhibit nonstationary error profiles driven by microphone placement, accents, PPE, reverberation, lighting, occlusion, and operator pose. Fourth, our sample ($N = 60$)—mostly university-affiliated and relatively homogeneous in age—limits external validity across crafts, skill levels, and multilingual crews; the study was likely underpowered for small effects on trust and attitude. Finally, we did not collect objective performance or safety metrics, nor did we quantify workload or attention; thus, conclusions about productivity and cognitive burden rest on perceived evidence. Future work will assess workload and performance using multiple complementary measures, including the NASA-TLX, time-stamped logs such as cycle time and command corrections, and wearable signals such as heart-rate variability and electrodermal activity for a longer period. Using these data, we can model how modality and recognition accuracy shape productivity and cognitive load and test whether workload mediates the link between interface design and user acceptance.

Author Contributions: Conceptualization, T.P.; methodology, D.L. (Daeguk Lee) and T.P.; software, D.L. (Daeguk Lee); validation, D.L. (Daeguk Lee) and T.P.; formal analysis, D.L. (Daeguk Lee) and T.P.; investigation, T.P.; resources, T.P.; data curation, D.L. (Daeguk Lee); writing—original draft preparation, D.L. (Daeguk Lee) and T.P.; writing—review and editing, D.L. (Daeguk Lee), J.H.J., D.L. (Donghun Lee) and T.P.; visualization, D.L. (Daeguk Lee) and T.P.; supervision, T.P.; project administration, T.P.; funding acquisition, J.H.J., D.L. (Donghun Lee) and T.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Soongsil University Research Fund (Convergence Research No. 202210001475) of 2022 and by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government (MOTIE) (HRD program for industrial innovation No. P0017033).

Institutional Review Board Statement: The study protocol was reviewed and approved by the Institutional Review Board of Soongsil University (SSU-202202-HR-413-1 approved in September 2022).

Informed Consent Statement: All participants provided written informed consent before participating in the experiment.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, D.; Lee, S.; Ku, N.; Lim, C.; Lee, K.-W.; Kim, T.-W.; Kim, J.; Kim, S.H. Development of a mobile robotic system for working in the double-hulled structure of a ship. *Robot. Comput.-Integr. Manuf.* **2010**, *26*, 13–23. [\[CrossRef\]](#)
2. Wurman, P.R.; D’Andrea, R.; Mountz, M. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Mag.* **2008**, *29*, 9–19. [\[CrossRef\]](#)
3. Krüger, J.; Lien, T.K.; Verl, A. Cooperation of human and machines in assembly lines. *CIRP Ann.* **2009**, *58*, 628–646. [\[CrossRef\]](#)
4. Accorsi, R.; Tufano, A.; Gallo, A.; Galizia, F.G.; Cocchi, G.; Ronzoni, M.; Abbate, A.; Manzini, R. An application of collaborative robots in a food production facility. *Procedia Manuf.* **2019**, *38*, 341–348. [\[CrossRef\]](#)
5. Geersing, T.H.; Franssen, E.J.F.; Pilesi, F.; Crul, M. Microbiological performance of a robotic system for aseptic compounding of cytostatic drugs. *Eur. J. Pharm. Sci.* **2019**, *130*, 181–185. [\[CrossRef\]](#)
6. Delgado, J.M.D.; Oyedele, L.; Ajayi, A.; Akanbi, L.; Akinade, O.; Bilal, M.; Owolabi, H. Robotics and automated systems in construction: Understanding industry-specific challenges for adoption. *J. Build. Eng.* **2019**, *26*, 100868. [\[CrossRef\]](#)
7. Scholz, C.; Cao, H.-L.; Imrith, E.; Roshandel, N.; Firouzipouyaei, H.; Burkiewicz, A.; Amighi, M.; Menet, S.; Sisavath, D.W.; Paolillo, A.; et al. Sensor-Enabled Safety Systems for Human–Robot Collaboration: A Review. *IEEE Sensors J.* **2024**, *25*, 65–88. [\[CrossRef\]](#)

8. Liang, C.-J.; Wang, X.; Kamat, V.R.; Menassa, C.C. Human–Robot Collaboration in Construction: Classification and Research Trends. *J. Constr. Eng. Manag.* **2021**, *147*, 03121006. [\[CrossRef\]](#)
9. Villani, V.; Pini, F.; Leali, F.; Secchi, C. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* **2018**, *55*, 248–266. [\[CrossRef\]](#)
10. Tan, X.; Xiong, L.; Zhang, W.; Zuo, Z.; He, X.; Xu, Y.; Li, F. Rebar-tying Robot based on machine vision and coverage path planning. *Robot. Auton. Syst.* **2024**, *182*, 104826. [\[CrossRef\]](#)
11. Zhang, M.; Xu, R.; Wu, H.; Pan, J.; Luo, X. Human–robot collaboration for on-site construction. *Autom. Constr.* **2023**, *150*, 104812. [\[CrossRef\]](#)
12. Oudah, M.; Al-Naji, A.; Chahl, J. Hand gesture recognition based on computer vision: A review of techniques. *J. Imaging* **2020**, *6*, 73. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Sylari, A.; Ferrer, B.R.; Lastra, J.L.M. Hand Gesture-Based On-Line Programming of Industrial Robot Manipulators. In Proceedings of the 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 22–25 July 2019; pp. 827–834. [\[CrossRef\]](#)
14. Pires, J.N. Robot-by-voice: Experiments on commanding an industrial robot using the human voice. *Ind. Robot* **2005**, *32*, 505–511. [\[CrossRef\]](#)
15. Lv, M.; Feng, Z.; Yang, X.; Guo, Q.; Wang, X.; Zhang, G.; Wang, Q. AMCIU: An Adaptive Multimodal Complementary Intent Understanding Method. *Int. J. Hum.–Comput. Interact.* **2025**, 1–24. [\[CrossRef\]](#)
16. Rabiee, A.; Ghafoori, S.; Bai, X.; Farhadi, M.; Ostadabbas, S.; Abiri, R. STREAMS: An Assistive Multimodal AI Framework for Empowering Biosignal Based Robotic Controls. In Proceedings of the 2025 6th International Conference on Artificial Intelligence, Robotics and Control (AIRC), Savannah, GA, USA, 7–9 May 2025; pp. 61–68. [\[CrossRef\]](#)
17. Hancock, P.A.; Billings, D.R.; Schaefer, K.E.; Chen, J.Y.; De Visser, E.J.; Parasuraman, R. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* **2011**, *53*, 517–527. [\[CrossRef\]](#)
18. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [\[CrossRef\]](#)
19. Mautua, I.; Fernandez, I.; Tellaeche, A.; Kildal, J.; Susperregi, L.; Ibarguren, A.; Sierra, B. Natural multimodal communication for human–robot collaboration. *Int. J. Adv. Robot. Syst.* **2017**, *14*. [\[CrossRef\]](#)
20. Kaczmarek, W.; Panasiuk, J.; Borys, S.; Banach, P. Industrial robot control by means of gestures and voice commands in off-line and on-line mode. *Sensors* **2020**, *20*, 6358. [\[CrossRef\]](#)
21. Berg, J.; Lu, S. Review of interfaces for industrial human-robot interaction. *Curr. Robot. Rep.* **2020**, *1*, 27–34. [\[CrossRef\]](#)
22. Urban, M.; Bajcsy, P. Fusion of voice, gesture, and human-computer interface controls for remotely operated robot. In Proceedings of the 2005 7th International Conference on Information Fusion, Philadelphia, PA, USA, 25–28 July 2005; p. 8. [\[CrossRef\]](#)
23. Jian, J.-Y.; Bissantz, A.M.; Drury, C.G. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics. Int. J. Cogn. Ergon.* **2000**, *4*, 53–71. [\[CrossRef\]](#)
24. Davis, F.D.; Bagozzi, R.P.; Warshaw, P.R. User acceptance of computer technology: A comparison of two theoretical models. *Manag. Sci.* **1989**, *35*, 982–1003. [\[CrossRef\]](#)
25. Park, E.; Del Pobil, A.P. Users’ attitudes toward service robots in South Korea. *Ind. Robot* **2013**, *40*, 77–87. [\[CrossRef\]](#)
26. Kohn, S.C.; de Visser, E.J.; Wiese, E.; Lee, Y.-C. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Front. Psychol.* **2021**, *12*, 604977. [\[CrossRef\]](#)
27. Tsarouchi, P.; Makris, S.; Chrysosolouris, G. Human–robot interaction review and challenges on task planning and programming. *Int. J. Comput. Integr. Manuf.* **2016**, *29*, 916–931. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.