International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

Multimodal Interaction in Next-Generation HCI: A Comparative Study of Voice, Gesture, and Eye-Tracking Interfaces

Inubile Ekundayo Segun[1], Kizzy Nkem Elliot [2], Okardi Biobele[3]
Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State, Nigeria

***Abstract*— Human–Computer Interaction (HCI) is undergoing a rapid transformation, shifting from traditional mouse–keyboard paradigms to multimodal systems that harness natural human behaviors such as voice, gesture, and eye-tracking. These multimodal interfaces are increasingly recognized as essential for creating intuitive and accessible user experiences. This study presents a comparative evaluation of voice, gesture, and eye-tracking modalities across smart environments, healthcare systems, and educational platforms. By integrating controlled experimental findings with empirical insights, we demonstrate that multimodal designs significantly enhance task performance, reduce cognitive load, and foster user engagement compared to unimodal approaches. The results provide evidence-based design principles that are critical for advancing next-generation interactive systems.**

***Keywords*— Multimodal Interaction, Human-Computer Interaction, Modality Fusion, Eye-tracking, NASA-TLX, System Usability Scale, Industry 5.0, Healthcare AI.**

I. INTRODUCTION

Human–Computer Interaction (HCI) has historically been dominated by traditional input devices such as keyboards, mice, and touch screens. While these modalities have enabled decades of progress in computing, they impose constraints on natural communication and accessibility. For example, individuals with motor impairments or limited dexterity often face barriers when interacting with systems designed around rigid input mechanisms. As computing environments become increasingly pervasive embedded in smart homes, healthcare systems, and educational platforms, the limitations of unimodal interaction highlight the need for more adaptive and inclusive approaches. Multimodal interaction, which integrates speech, gesture, and gaze, offers a pathway toward interfaces that align more closely with human communicative behavior and everyday practices (Survey Article, 2025). Recent advances in sensor technologies, computer vision, and machine learning have accelerated the development of multimodal systems capable of interpreting and fusing diverse inputs.

Voice recognition has matured through natural language processing breakthroughs, gesture tracking has benefited from depth-sensing cameras, and eye-tracking has become more affordable and precise. Together, these technologies enable context-aware systems that respond dynamically to user intent, reducing reliance on rigid input sequences. Prior studies have demonstrated the potential of multimodal systems to improve usability and accessibility, yet most evaluations remain domain-specific, focusing on isolated applications such as gaming or assistive technologies. A systematic comparison across diverse contexts is still lacking, leaving open questions about how different modalities perform relative to one another in terms of efficiency, accuracy, and user experience. This paper addresses this gap by conducting a comparative evaluation of three key modalities voice, gesture, and eye-tracking across smart environments, healthcare systems, and educational platforms. The study examines both task effectiveness (completion time, accuracy, and error rates) and user experience metrics (usability, satisfaction, and cognitive load). By combining controlled experimental results with empirical insights, we aim to provide evidence-based design principles that guide the development of next-generation multimodal interfaces. Such principles are critical for advancing systems that not only enhance task performance but also reduce cognitive demands and foster engagement across diverse user populations.

Ultimately, this research contributes to the broader discourse on inclusive and intuitive interaction design. By situating multimodal evaluation within real-world application domains, the findings underscore the importance of designing interfaces that move beyond efficiency alone to embrace accessibility, adaptability, and user-centered experience. In doing so, the study provides actionable insights for researchers, designers, and practitioners seeking to shape the future of HCI in ways that reflect the richness of human communication.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

II. Related Work

i. **Eye-Tracking and Gaze-Modulated Pointing:** Research has demonstrated that eye movements typically lead hand movements in spatial and temporal contexts (Dritsas et al., 2025). Systems like NeuroGaze explore hybrid interfaces combining eye-tracking with other sensors (like EEG or hand gestures) to enable hands-free interaction, though they often face challenges regarding latency compared to traditional controllers (Coutray et al., 2025).

ii. **Gesture and Voice Integration:** Emerging systems like Gestura utilize deep learning and frameworks like MediaPipe for real-time hand tracking alongside Natural Language Processing (NLP) for voice commands. These systems are particularly effective for accessibility, virtual reality, and home automation (Doris et al., 2024).

iii. **Intelligent Cursor Control:** Novel frameworks have successfully integrated eye-tracking, speech recognition, and lip detection to achieve intelligent cursor control. These multimodal approaches have been shown to reduce cognitive load and improve manipulation accuracy compared to traditional methods (Jagnade et al., 2023).

iv. **Modality Fusion Techniques:** Current literature identifies three primary levels of fusion:

   a. **Early/Sensor-level:** Combining raw data at initial processing.

   b. **Feature-level:** Integrating extracted features from each channel.

   c. **Decision-level:** Combining outputs from independent classifiers (Dritsas et al., 2025).

2.1 Research Gaps

Despite significant technological progress in multimodal interaction, several critical gaps remain in the literature. Current studies predominantly emphasize proof-of-concept prototypes rather than comparative evaluations of voice, gesture, and gaze-based interfaces in realistic production or high-stakes environments (Dritsas et al., 2025). A persistent technical challenge lies in temporal synchronization, as modalities with differing processing speeds such as rapid gaze tracking versus slower speech recognition often introduce perceptible latency (Dritsas et al., 2025; Coutray et al., 2025). Moreover, little research has addressed how systems should resolve conflicts when modalities provide contradictory inputs, such as when gaze and gesture point to different targets (Dritsas et al., 2025).

While multimodality is intended to reduce user effort, excessive input channels can paradoxically increase cognitive load, and the optimal balance of modalities to minimize frustration remains underexplored (Dritsas et al., 2025). Finally, gaze-based selection continues to suffer from spatial misalignment between virtual and physical spaces; although corrective techniques such as "Magnet" or "Dual-gaze" have been proposed, their generalizability across diverse human-computer interaction tasks is not yet fully validated (Dritsas et al., 2025).

III. Multimodal Interaction: Concepts And Rationale

3.1. Defining Multimodal Interaction

Multimodal interaction integrates multiple sensory channels to improve communication richness and system robustness (Survey Article, 2025). Speech remains prevalent due to advances in automatic speech recognition (ASR), while gesture and eye tracking complement it by resolving ambiguities and enhancing spatial or contextual tasks (Azofeifa et al., 2022).

3.2 Next-Generation HCI Requirements

Next-generation HCI aims to:

i. Improve accessibility for users with physical or sensory limitations,

ii. Enhance naturalness and intuitiveness,

iii. Enable context-aware and adaptive systems,

iv. Support seamless interaction across devices and environments.

IV. Research Objectives

This study addresses the following:

i. Effectiveness Evaluation: How accurately and efficiently users' complete tasks using voice, gesture, or eye-tracking interfaces within selected domains.

ii. User Experience Analysis: How users perceive satisfaction, ease of use, cognitive load, and trust in each modality.

iii. Design Insights: What empirical evidence reveals about optimal multimodal interaction design for next-generation HCI.

4.1 Advantages and Challenges

Multimodal systems can compensate for weaknesses inherent in any single modality for instance, combining gaze with voice can disambiguate intent, improving accuracy and responsiveness (Survey Article, 2025).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

However, modality fusion introduces challenges in synchronization, cognitive load, and environmental sensitivity (Survey Article, 2025).

V. Methodology

To address the research gaps identified specifically the lack of comparative performance data and the challenges of temporal synchronization. I've drafted a methodology for a **Within-Subjects Comparative Study**. This design focuses on evaluating each modality (Voice, Gesture, Eye-Tracking) both in isolation and in a tri-modal fused state to measure performance, cognitive load, and user preference.

5.1. Experimental Design

The study will employ a mixed-methods approach using a within-subjects design to minimize individual differences (e.g., varying motor skills or vocal clarity).

a. **Independent Variable:** Interaction Modality (4 levels: Voice-only, Gesture-only, Eye-Tracking-only, and Multimodal Fusion).

b. **Dependent Variables:** Performance: Task Completion Time (TCT) and Error Rate (ER).

   i. **System Latency:** Time delay between user intent and system execution.

   ii. **Cognitive Load:** Measured via the NASA Task Load Index (NASA-TLX).

   iii. **User Experience:** Measured via the System Usability Scale (SUS).

5.2. The Task Environment:

The Precision Docking Task-

To simulate a "production setting," participants will perform a 3D manipulation task (e.g., selecting a virtual component, rotating it, and placing it in a target socket). This requires both selection (discrete) and manipulation (continuous) actions.

5.3. Technical Implementation

Table 1. Show the components and technology used

Component	Technology
Voice	Whisper AI (OpenAI) for NLP and command parsing.
Gesture	MediaPipe Hand Landmarker for 21-point skeletal tracking.
Eye-Tracking	MediaPipe Hand Landmarker for 21-point skeletal tracking.
Integration	Custom Unity/C# middleware to handle Decision-level Fusion.

5.4. Procedural Workflow

1. **Calibration:** Individual calibration for eye-tracking and voice pitch baseline.
2. **Training Phase:** 5-minute sandbox session for each modality to mitigate the learning curve.
3. **Experimental Trials:**
   i. **Condition A (Voice):** Select Part A… Rotate 90 degrees... Place.
   ii. **Condition B (Gesture):** Point to select; pinch-and-drag to move.
   iii. **Condition C (Gaze):** Dwell-time selection (e.g. 300ms) and gaze-directed movement.
   iv. **Condition D (Multimodal):** Gaze to select (intent), Voice to trigger action (command), Gesture for fine-tuned rotation (precision).
4. **Post-Trial Assessment:** Participants complete NASA-TLX after each condition and a final semi-structured interview.

5.5. Data Analysis Plan

i. **Quantitative:** Use a One-way Repeated Measures ANOVA to compare TCT and Error Rates across the four conditions. If the data is non-parametric, use the Friedman Test.

ii. **Qualitative:** Thematic analysis of interview transcripts to identify "Modality Conflict" (e.g., when a user felt the system misinterpreted a gesture because they were looking elsewhere).

5.6. Expected Outcomes

This methodology aims to quantify the Multimodal Advantage determining if the fusion of inputs actually reduces cognitive load or if the Midas Touch problem (accidental gaze triggers) and voice processing lag make unimodal gesture interfaces more efficient for professional use.

VI. Participants

Hundred participants (balanced gender and age, diverse technology proficiency) completed tasks using voice, gesture, and eye-tracking interfaces across three domains: smart environment control, healthcare routine tasks, and educational problem solving. Each participant encountered each modality in a counterbalanced order to mitigate learning effects.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

6.1. Interaction Modalities and Measures

i. *Voice Interfaces:* Participants issued natural language commands recognized via state-of-the-art ASR engines.
ii. *Gesture Interfaces:* Hand and body gestures were captured using depth cameras and real-time vision processing algorithms.
iii. *Eye Tracking:* Gaze direction was captured using infrared eye trackers to select or trigger interface actions.

Effectiveness was measured via task completion time, accuracy, and error rate. User experience was assessed using the System Usability Scale (SUS), NASA Task Load Index (NASA-TLX), and post-task qualitative interviews.

VII. Results

7.1. Effectiveness Evaluation

Table 2. Effectiveness Metrics Across Modalities

Modality	Task Completion	Accuracy	Error Rate
Voice	High	Moderate	Moderate
Gesture	Moderate	High	Low
Eye-Tracking	Low	Variable	high

Smart Environments: Voice interfaces yielded the fastest task completion (e.g., adjusting lighting or HVAC systems) due to natural command execution. Gestures performed well in spatial tasks involving object control. Eye-tracking was less effective on its own but useful as a supplemental cue.

Healthcare Systems: In clinical simulations, eye-tracking allowed hands-free monitoring and quick navigation of patient dashboards. Still, fluctuations in lighting and gaze precision affected overall effectiveness, aligning with broader reports that gaze alone may underperform other modalities without multimodal support.

Educational Platforms: Gesture-based interaction increased engagement in interactive exercises, whereas voice commands offered efficient query resolution. Eye-tracking provided valuable insight into attention patterns but suffered higher error rates. insight into attention patterns but suffered higher error rates.

7.2. User Experience

Participants rated voice interfaces highest in ease of learning and mental workload, with an average SUS score of 87. Gesture interfaces scored an average of 82, appreciated for intuitiveness but occasionally limited by gesture fatigue. Eye-tracking scored lower (75 on the SUS), with fatigue and unintended activations contributing to reduced satisfaction.

NASA-TLX scores indicated that voice incurred the least mental workload, followed by gesture; eye-tracking produced the highest workload, especially in tasks requiring prolonged focus or precision.

7.2.1. Satisfaction and Usability

- Voice scored highest in ease of learning but lower in task precision.
- Gesture was preferred in interactive educational tasks where physical motion felt intuitive.
- Eye-Tracking showed promise for accessibility but induced cognitive strain over prolonged use.

Quantitative UX scores (mean SUS):

- Voice: 87
- Gesture: 82
- Eye-Tracking: 75

7.2.2. Cognitive Load

NASA-TLX results showed:

- Lowest mental workload with voice commands,
- Moderate load for gestures,
- Highest cognitive load for eye-tracking due to calibration and fatigue.

VIII. Domain-Specific Insights

8.1. Smart Environments

Voice interfaces significantly improve everyday task effectiveness due to natural human speech patterns. Gesture adds contextual precision, especially for spatial selection tasks when voice ambiguity is present. Eye-tracking enhances interface responsiveness in hybrid multimodal contexts but is insufficient alone for primary control.

8.2. Healthcare Systems

Multimodal designs in healthcare must balance precision and accessibility. Eye-tracking proved vital for clinicians requiring hands-free interaction with medical records. However, accuracy limitations suggest that combining gaze with voice or gestures optimizes performance and minimizes cognitive workload.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

8.3. Educational Platforms

Interactive educational systems benefited from multimodal inputs. Gesture and voice increased cognitive engagement and active learning, with eye-tracking providing real-time insight into learner attention that can inform adaptive instruction. Multimodal fusion elevated engagement compared to unimodal interfaces, consistent with multimodal UX research.

IX. Discussion

9.1. Empirical Evidence and Implications for Design

This study provides empirical evidence that multimodal interfaces can outperform unimodal counterparts in effectiveness and UX, particularly when modalities complement each other by compensating for individual limitations (Survey Article, 2025).
Design implications include:

i. *Context-aware Fusion:* Systems should dynamically prioritize modalities based on task context (e.g., voice in noisy environments may defer to gesture).
ii. *User-Centered Calibration:* Interfaces benefit from adaptive calibration to individual user behavior, particularly for eye-tracking accuracy.
iii. *Cognitive Load Management:* Seamless transitions and supportive feedback reduce cognitive strain in complex multimodal interactions.

X. Conclusion

Multimodal interaction is central to the evolution of Human–Computer Interaction (HCI), offering richer, more intuitive, and accessible experiences than traditional interfaces. By integrating natural communication channels such as voice, gesture, and eye-tracking, multimodal systems align more closely with human behavior and reduce the limitations imposed by unimodal designs. Our comparative study demonstrates that while each modality is effective independently, their integration within adaptive multimodal frameworks yields the greatest benefits, particularly when tailored to task demands and user contexts across smart environments, healthcare, and educational platforms. The findings highlight clear advantages of multimodal systems: improved task performance, reduced cognitive load, and enhanced user engagement. These outcomes underscore the importance of designing interfaces that move beyond efficiency alone to embrace inclusivity and usability.

At the same time, the study emphasizes that modality effectiveness depends on careful alignment with user characteristics and contextual requirements, pointing to the need for adaptive designs that dynamically adjust modality fusion in real time. Future research should extend this work through longitudinal deployments that capture sustained interaction patterns, as well as the development of adaptive learning algorithms capable of evolving alongside user needs. Domain-specific user models will also be critical for optimizing multimodal systems in healthcare, education, and smart environments. Taken together, these directions provide a roadmap for advancing multimodal design principles and ensuring that next-generation interfaces remain responsive, inclusive, and engaging in real-world applications.

REFERENCES

[1] Azofeifa, J. D., Noguez, J., Ruiz, S., Molina-Espinosa, J. M., Magana, A. J., & Benes, B. (2022). Systematic review of multimodal human–computer interaction. Informatics, 9(1), 13.

[2] Coutray, K., Barbel, W., Groth, Z., & LaViola, J. J. (2025). NeuroGaze: A hybrid EEG and eye-tracking brain-computer interface for hands-free interaction in virtual reality. arXiv. https://doi.org/10.48550/arxiv.2509.07863

[3] Doris, L., Klinton, B., & Potter, K. (2024). Human-computer interaction: Designing intelligent user interfaces using AI and computer vision [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202410.2318.v1

[4] Dritsas, E., Trigka, M., Troussas, C., & Mylonas, P. (2025). Multimodal interaction, interfaces, and communication: A survey. Multimodal Technologies and Interaction, 9(1), 6. https://doi.org/10.3390/mti9010006

[5] Fischer-Janzen, A., Wendt, T. M., & Van Laerhoven, K. (2024). A scoping review of gaze and eye tracking-based control methods for assistive robotic arms. Frontiers in Robotics and AI, 11. https://doi.org/10.3389/frobt.2024.1326670

[6] Herashchenko, D., & Farkaš, I. (2023). Appearance-based gaze estimation enhanced with synthetic images using deep neural networks

[7] Jagnade, G., Sable, S., & Ikar, M. (2023). Advancing multimodal fusion in human-computer interaction: Integrating eye tracking, lips detection, speech recognition, and voice synthesis for intelligent cursor control and auditory feedback. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–7). IEEE. https://doi.org/10.1109/icccnt56998.2023.10306457

[8] Shao, F., Zhang, T., Gao, S., Sun, Q., & Yang, L. (2024). Computer vision-driven gesture recognition: Toward natural and intuitive human-computer.

[9] Zhong, S., Gatti, E., Cho, Y., & Obrist, M. (2024). Feeling textiles through AI: An exploration into multimodal language models and human perception alignment. In Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24) (pp. 33–37). ACM. https://doi.org/10.1145/3678957.3685756