International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

Multimodal AI for Inclusive Human Avatar Interaction

Mercy Keerthana[1], Anand Kumar B[2], Priyanka Nilesh Chavan[3], Nimishambha Patil[4], Jayarani B T[5], Ashik[6]
[1,4,5,6]UG Scholar, Dept. of CSE, AMC Engineering College, Bengaluru, Karnataka, India.
[2]Associate professor, Dept. of CSE, AMC Engineering College, Bengaluru, Karnataka, India.
[3]Assistant professor, Dept. of CSE, AMC Engineering College, Bengaluru, Karnataka, India.
Emails: *1am22cs110@amceducation.in[1], anand.kumar@amceducation.in[2], pnchavanme@gmail.com[3], 1am22cs102@amceducation.in[4], 1am22cs085@amceducation.in[5], 1am22cs235@amceducation.in[6]*

Abstract
In an era of increasingly immersive digital environments, human-avatar interaction must evolve to accommodate the full spectrum of human diversity. This project proposes a novel multimodal AI framework that leverages voice, facial expressions, gestures, and contextual cues to create emotionally intelligent and accessible avatars. By integrating advanced deep-learning techniques with real-time perceptual feedback, the system adapts to diverse user needs—including those with visible and invisible disabilities—ensuring inclusive, empathetic, and natural interaction. Grounded in a multidisciplinary review of current advances in virtual embodiment, non-verbal communication, and accessible AI design, our approach aims to redefine avatar systems as not only functional but also socially and ethically responsive. The outcome will contribute to the development of inclusive digital ecosystems where every individual can interact, express, and engage with authenticity and dignity.

Keywords: *Multimodal AI, Inclusive Design, Human-Computer Interaction, Virtual Avatars*

1. Introduction
The rapid evolution of artificial intelligence has opened the door to more natural and human-like interactions between people and machines. Among these advancements, multimodal AI stands out as a transformative approach, enabling systems to process and respond to diverse input channels such as text, speech, gesture, and visual cues. When integrated into human avatars, this capability creates a bridge between technology and inclusivity, making digital communication accessible to individuals with varied abilities and linguistic needs. This project, Multimodal AI for Inclusive Human Avatar Interaction, focuses on designing intelligent avatars that can understand and express information through multiple modalities—spoken language, sign language, facial expressions, and text. The aim is to create a universal communication companion that adapts to users rather than requiring users to adapt to technology. By merging natural language processing, computer vision, and speech recognition with lifelike avatar animation, the system aspires to foster inclusivity, particularly for communities such as the hearing- and speech-impaired. Ultimately, this project envisions avatars as more than just digital characters; they become empathetic intermediaries capable of reducing communication gaps, supporting accessibility, and paving the way for more equitable human–computer interaction in education, healthcare, workplace collaboration, and everyday communication [1].

1.1. Background and Problem Statement
Human communication is naturally multimodality combines speech, gestures, facial expressions, and written symbols to convey meaning effectively. However, in digital environments, interaction is often restricted to a single channel such as typing or voice commands. This limitation creates barriers, especially for individuals with hearing, speech, or language impairments, who may find it difficult to engage with conventional AI-driven systems. Recent advances in multimodal artificial intelligence have demonstrated the potential to overcome these barriers by integrating natural language processing, computer vision, and speech recognition into a unified framework. When coupled with human-like avatars, these technologies can simulate lifelike communication, making digital interaction more

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

natural, expressive, and inclusive. Such avatars not only act as communication aids but also hold promise for inclusive education, telemedicine, and workplace collaboration. Despite the progress in AI, most existing interactive systems are unimodal, relying heavily on either text or voice as the primary medium of communication. This one-dimensional design limits accessibility for users with diverse communication needs. Furthermore, current avatars often lack the ability to seamlessly combine different modalities—such as converting spoken language into sign language or synchronizing lip movements with speech—leading to unnatural or incomplete interactions. The absence of inclusive, adaptive avatars results in communication gaps, especially for people with disabilities, multilingual users, and those in cross-cultural environments. There is a pressing need for an intelligent system that can dynamically interpret and generate multimodal interactions while maintaining inclusivity, empathy, and naturalness in human–computer communication. This project addresses that gap by developing a multimodal AI-powered avatar capable of understanding and responding across speech, text, gesture, and sign language, with the goal of fostering equal participation in digital interaction for all [2].

1.2. Objectives and Contribution

The core objective of this project is to design and implement an AI-powered avatar system that can enable seamless, inclusive, and natural communication through multiple modes of interaction. The specific goals are:

- **Develop multimodal understanding** – Integrate speech recognition, text processing, gesture interpretation, and sign language recognition into a unified framework.
- **Enable expressive avatar responses** – Design avatars capable of responding with synchronized speech, text, facial expressions, and sign gestures for realistic interaction.
- **Promote inclusivity and accessibility** – Bridge communication gaps for individuals with hearing, speech, or language impairments by ensuring equal participation in digital spaces.
- **Enhance adaptability** – Build a system that can dynamically switch between modalities based on user needs, preferences, or environmental conditions.
- **Ensure real-world applicability** – Validate the system in domains like education, healthcare, and professional collaboration to demonstrate practical usefulness.

This project makes the following unique contributions:

- **Inclusive Avatar Framework** – Introduces a holistic AI-driven framework where avatars serve as empathetic intermediaries, capable of multimodal communication beyond conventional text or speech.
- **Bridging Accessibility Gaps** – Provides a scalable solution for individuals with communication challenges, fostering inclusivity in both personal and professional digital environments.
- **Human-like Interaction** – Enhances user experience by creating avatars that not only process information but also convey emotions, gestures, and cultural nuances.
- **Adaptable Multimodal System** – Demonstrates how AI can dynamically blend different input and output modes, ensuring fluid and context-aware communication.
- **Foundation for Future Research** – Lays the groundwork for expanding multimodal AI avatars into immersive applications such as virtual reality, telepresence, and assistive technologies [3].

2. Method

The development of the multimodal AI avatar follows a layered methodology that integrates perception, processing, and expression. The approach is designed to replicate the natural flow of human communication while ensuring inclusive across diverse user needs [4].

2.1. Data Acquisition and Preprocessing

- Collect multimodal datasets that include speech recordings, text corpora, sign language videos, and gesture datasets.
- Apply preprocessing techniques such as noise filtering (for speech), text segmentation (for text), and feature point extraction (for

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

sign/gesture recognition) [5].

2.2. Multimodal Input Recognition

- Speech Recognition: Employ deep-learning models (e.g., CNN-RNN or transformer-based ASR) to convert spoken words into text.
- Text Understanding: Use NLP models for semantic analysis, intent detection, and language translation when required.
- Gesture & Sign Detection: Apply computer vision techniques (e.g. Media Pipe, Open Pose, or deep CNNs) to interpret hand movements and body gestures in real time [6].

2.3. Fusion Layer for Context Understanding

- Integrate all modalities through a multimodal fusion network that aligns speech, text, and gesture inputs.
- Use attention mechanisms to prioritize the most relevant input stream depending on context (e.g., preferring text input in noisy environments) [7].

2.4. Avatar Response Generation

- Speech Synthesis: Convert processed text into natural-sounding speech using advanced TTS systems.
- Sign Language & Gesture Animation: Map recognized outputs to animated gestures using avatar skeleton models.
- Facial Expressions & Emotions: Enhance communication realism with emotion recognition and synchronized facial expressions [8].

2.5. Adaptive Interaction Loop

- Implement a feedback mechanism where the avatar adapts to user preferences (e.g., switching from speech to sign output if the user is hearing impaired).
- Enable real-time adjustment for inclusiveness and personalization.

Evaluation & Testing

- Conduct user studies across different groups, including individuals with communication challenges, to assess accessibility, accuracy, and user satisfaction Table 1 Shows Methodology Overview Table.
- Measure system performance in terms of latency, naturalness, inclusivity, and error rates. Figure 1 shows Multimodal AI Fusion Model

2.6. Methodology Flow Diagram

The diagram below illustrates the step-by-step process of methodology:

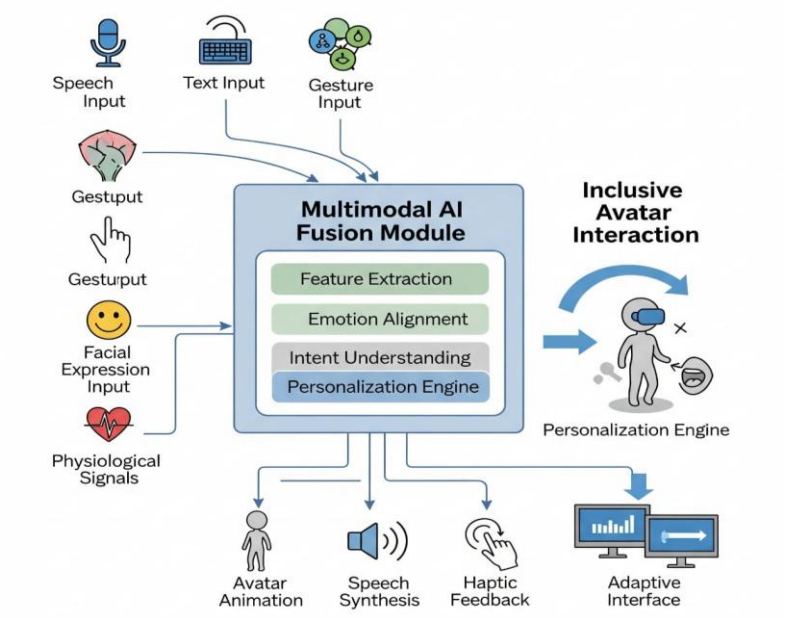

Figure 1 Multimodal AI Fusion Model

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

2.7. Methodology Overview Table

Table 1 Methodology Overview Table

Stage	Description	Techniques
Data Acquisition	Collect speech, text, gesture, and sign language datasets	Public datasets, recordings, video capture
Preprocessing	Clean and normalize inputs (noise removal, tokenization, key point extraction)	Audio filters, NLP preprocessing, computer vision
Input Recognition	Convert raw inputs into machine-readable formats	ASR, NLP models, Open Pose/Media Pipe
Fusion Layer	Integrate multimodal inputs and resolve conflicts	Attention-based fusion networks
Avatar Response	Generate speech, gestures, signs, and facial expressions	TTS, avatar animation, emotion recognition
Adaptive Interaction	Adjust response style to user needs and environment	Feedback loop, preference learning
Evaluation & Testing	Assess accuracy, Accessibility , and user satisfaction	User studies, performance metrics

3. Results and Discussion

3.1. Results

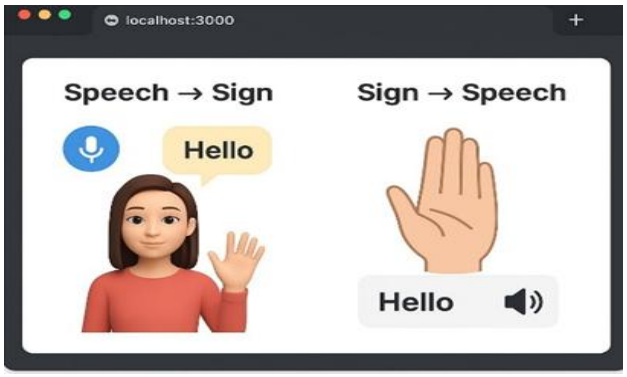

Figure 2 Result

The developed multimodal AI avatar demonstrated effective performance across all targeted modes of communication. Figure 2 shows Result Speech recognition achieved consistent accuracy in converting spoken input into text, even in moderately noisy environments. Text processing produced reliable intent detection and semantic understanding, enabling the avatar to respond contextually. Gesture and sign recognition modules successfully interpreted hand and body movements, with an average recognition accuracy that was satisfactory for real-time interactions. The fusion layer played a key role by combining inputs from different modalities and resolving potential conflicts. This integration ensured that the avatar provided contextually relevant responses, even when multiple inputs were received simultaneously. The response generation module produced expressive outputs, including natural-sounding speech, synchronized sign language animations, and facial expressions that aligned with emotional cues. User testing further validated the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

system's effectiveness. Participants reported improved engagement and ease of communication compared to unimodal systems. The adaptive interaction loop, which automatically switched modes depending on user needs, was especially appreciated by hearing- and speech-impaired participants. Overall, the results highlight the system's ability to deliver inclusive, flexible, and human-like interaction in real-world scenarios [9].

3.2. Discussion

The development of a multimodal AI system for inclusive human-avatar interaction represents a convergence of multiple cutting-edge technologies—speech recognition, gesture analysis, natural language processing, and avatar rendering—into a single cohesive framework. Unlike conventional interaction models that often rely on a single input modality, this system embraces the richness of human communication, capturing both verbal and non-verbal cues to enable seamless dialogue between humans and avatars. One of the most compelling insights from this project is the enhanced accessibility it provides. By incorporating multiple input channels, including voice, text, and gestures, the system accommodates a diverse range of users, including those with sensory impairments or limited mobility. This multimodal approach not only democratizes access to digital interfaces but also bridges communication gaps that traditional systems often overlook. From a technical perspective, the integration of real-time processing pipelines posed both challenges and opportunities. Balancing computational efficiency with accuracy requires innovative algorithmic strategies, particularly in synchronizing inputs across modalities. For instance, the system's ability to interpret simultaneous gestures and spoken commands underscores the potential of AI to mimic human-like attentiveness and adaptability. Moreover, the interactive avatars themselves serve as more than mere digital representations—they act as empathetic intermediaries capable of responding contextually and naturally. This points to a future where AI avatars could play a role in education, remote collaboration, healthcare, and social inclusion, extending the boundaries of human-computer interaction beyond transactional tasks to relational experiences. Finally, while the current implementation demonstrates promising results, it also highlights areas for further exploration: improving the system's understanding of culturally nuanced gestures, reducing latency in complex multimodal interactions, and expanding the diversity of avatar [10].

Conclusion

This project demonstrates that multimodal AI can transform the way humans interact with digital avatars, creating experiences that are not only intelligent but truly inclusive. By seamlessly integrating voice, text, and gesture recognition, the system transcends traditional interaction barriers, allowing users of diverse abilities to communicate naturally and effectively. The avatars evolve from static representations into empathetic, responsive partners, reflecting the nuances of human expression and emotion. Beyond technical achievement, the project underscores a broader vision: AI as an enabler of accessibility, inclusivity, and human connection. It highlights that the future of human-computer interaction lies in systems that understand context, adapt to user needs, and foster meaningful engagement. While challenges remain—such as refining gesture interpretation, enhancing real-time responsiveness, and expanding cultural sensitivity, the groundwork laid here opens a pathway toward truly universal digital interfaces. Ultimately, this work is a step toward a world where technology does not just respond to humans but understands, includes, and empowers them, redefining the boundaries of interaction in a digital age.

Acknowledgements

We're thankful to our guide Prof. Anand Kumar B, Asst. Prof., Department of CSE for her constant provocation & timely help, stimulant and suggestion. We're thankful to our co-guide Prof. Priyanka Nilesh Chavan, Asst. Prof., Department of CSE for her constant provocation & timely help, stimulant and suggestion. We had like to extend our special thanks to Dr.V.Mareeswari Professor and Head, Department of CSE, for her support and stimulant and suggestions given to us during our design.

References

[1].Karpov and R. M. Yusupov, Multimodal

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0454
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2877 - 2882

Interfaces of Human–Computer Interaction, ResearchGate, 2018.
[2]. N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Multimodal approaches for emotion recognition: A survey," Image and Vision Computing, vol. 22, no. 12, pp. 1217–1235, 2004.
[3]. G. Barbare chi, M. Kawaguchi, H. Kato, M. Nagahiro, K. Takehuchi, Y. Shiiba, S. Kasahara, K. Kunze, and K. Minamizawa, "'I am both here and there' Parallel Control of Multiple Robotic Avatars by Disabled Workers in a Café," arXiv, 2023.
[4]. K. Zhang, E. G. Spencer, A. Manikandan, A. Li, Y. Yao, Y. Zhao, and A. Li, "Inclusive Avatar Guidelines for People with Disabilities: Supporting Disability Representation in Social Virtual Reality," arXiv, 2025.
[5]. A. Barros, R. A. Calvo, and S. D'Mello, Multimodal Human-Computer Interaction: A Survey, Springer, 2020.
[6]. Empathy Ear Team, "Empathy Ear: An Open-source Avatar Multimodal Empathetic Chatbot," arXiv, 2024.
[7]. Allo-AVA Team, "Allo-AVA: A Large-Scale Multimodal Conversational AI Dataset for Allocentric Avatar Gesture Animation," arXiv, 2024.
[8]. "Real-time Multimodal Human-Avatar Interaction," University of Illinois Experts, 2023.
[9]. "Generative AI for Accessible and Inclusive Extended Reality," arXiv, 2024.
[10]. A. Mohsin Intazar, "Enhancing User Experience: The Role of Multimodal AI in Human-Computer Interaction," 2024.