IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI Companion: Design and Implementation of a Speech-Based Virtual Interaction System Using Whisper and GPT

[1] Auxilia Michael , [2] Kailasanathan R , [3] Gokulraj A , [4] Nithishkumar D

[1]Department of Computer Science and Business Systems

[1]Sri Manakula Vinayagar Engineering College Puducherry, India

***Abstract* -** AI is changing the way users communicate with computers. Instead of using traditional graphical software, computers now can be communicated with using natural and conversational methods. However, today's virtual assistants have limited capabilities because basic voice commands are all they can understand, which limits how effective they can be as interactive companions. The current research software combines a speech controlled virtual AI companion system that uses the Whisper speech recognition model for robust speech recognition capabilities and GPT models for high levels of conversational intelligence. The proposed system includes a virtual robotic avatar that enhances embodiment, as well as human interactions with a simulated companion. The proposed system creates simulated conversations via real- time processing of speech input, analysis of users' intent through natural language processing (NLP), and generation of contextually relevant responses through the evolution of conversationally intelligent software. The proposed system is implemented in Python with a LiveKit agent to facilitate real-time communication among users and to support individuals by providing conversational assistance, task reminders, controlling media, and performing basic automated tasks within a simulated environment. Additionally, the proposed system uses noise management techniques and adaptive speech recognition to enhance the performance of the entire system. The experimental analysis involved multiple users and produced a combined average speech recognition accuracy rate of 92% and an average response latency of 2 seconds, indicating that real-time conversations are possible between users and the proposed system. Integration of conversational AI, speech recognition, and virtual embodiment improves upon existing Human Computer Interaction (HCI) methods.

***Index Terms* -** AI Companion, Virtual Human Interaction, Speech Recognition, Natural Language Processing, Whisper, GPT, Conversational AI, LiveKit, Virtual Avatar, Human-Computer Interaction.

I. INTRODUCTION

Most virtual assistants have undergone remarkable advancement; they are very limited in that they primarily interact through voice, lack emotion, and do not provide a physical component. Because of this limitation, users and their assistants tend to engage in a transactional manner rather than in a socially engaged or companion-like manner; however, to address these challenges, the AI Companion system proposes a voice-based virtual interaction platform that employs a combination of real-time speech recognition and highly sophisticated conversational intelligence. By combining transformer-based language models and an avatar-based robotics interface, the aim is to enhance the nature of human-computer interactions and create an immersive and natural communication experience.

Natural Language Processing Algorithms are used on speech converted into text, which process the underlying linguistic structure and meanings before using a Transformer Conversational Model to generate meaningful, contextually-aware responses, thus supporting the ability of an AI Companion to continuously communicate with the user in a coherent manner. In contrast to Rule Based Systems, the proposed approach allows for flexible, adaptive communications that are more consistent with the user's queries and patterns of interaction than other methods. In addition, Sentiment Aware Reply Methods are used to identify emotional cues within the user's input so that the AI Companion can generate more empathetic and socially appropriate replies, thereby enhancing the user's overall interaction experience.

Instead of just being a voice-only assistant like you might see with standard assistants, the AI Companion will include a virtual embodiment of itself as an almost robotic body (avatar) to the user. The avatar will be a visual representation of what the AI Companion has to tell its user as well as assist with improving the interactive process through the use of synchronized visual feedback (the avatar) to complement what the user hears from their AI Companion. The high-level decisions made by the Conversational Engine are also translated into the avatar's actions (head movement, gesture, and visual signals) that can be seen when interacting with the user. Since the AI Companion allows for a multimodal interaction process to take place, it will allow for a greater degree of user engagement and provide a more immersive experience than either text- or voice-only systems would allow. The user's ability to use the system as their own digital companion, will lead to the user treating the system as an interactive, digital companion, as opposed to just a software-based assistant.

The increasing advancements of intelligent technologies have created an increased demand for artificial intelligence (AI) companions equipped with contextual understanding, emotional intelligence, and presence interactive as users engage in social interactions with one another. To meet this demand, the proposed virtual AI Companion System combines speech intelligence, conversational reasoning, and visual presence into a single platform creating a new generation of human-computer interaction platforms that are intuitive, adaptive, and will help users in both the actual world and in the virtual world.

II. RELATED WORK

The last decade has been a transformative time for artificial intelligence (AI) and natural language processing (NLP), leading to improved interfaces for human/machine interaction through technology. The advancements in technology have enabled the design of intelligent companion systems to enable and support natural, contextual, and socially responsive methods of communication. Initially these systems were based on software-based virtual assistants from companies, such as Amazon, Apple, and Google, which demonstrated the feasibility of voice communication through Automatic Speech Recognition (ASR) and NLP. The early implementations of these systems did not take full advantage of their contextual richness potential, emotional expressions, and interactive presences, limiting their effectiveness as companion-based platforms [1], [3], [6].

These systems are continuing to evolve; however, many challenges remain related to achieving high accuracy rates for transcription in noisy environments and providing low latency during live interaction. Research will continue to explore the balance between response time and accuracy of translation/interpretation, as these have historically been large trade-offs for conversational artificial intelligence. Recent advancements in transformer-based speech and language models provide opportunities for improved processing speed and greater contextual knowledge, resulting in the creation of more efficient real-time conversational systems than previously available [4], [10].

The main differentiator between advanced AI companions and typical voice-only virtual assistants is physical presence and embodied interaction. Humanoid robot platforms such as Jibo and Pepper offer embodied interaction experiences by combining speech, vision and motion capabilities. However, the high costs, complexity of hardware, and scalability issues of physical robotic systems make them less viable for mass deployment [15], [16]. As a result, there is growing interest among researchers in using virtual avatars as a cost-effective alternative to building physical robots. Virtual avatars are able to simulate expressive body language (via gesture and head movement) and provide visual feedback via software to create an illusion of presence, without requiring any dedicated hardware infrastructure. There are many advantages to virtual embodiment, including the ability to scale; ease of deploying the technology; and significant reductions in maintenance costs. By transforming voice (conversational) commands into synchronised avatar animation, AI technologies can enhance users' ability to engage and be assessed based on multimodal inputs (e.g., voice or visual) while ensuring that they will have equal opportunity to engage with the system. The AI Companion that has been proposed in this research adopts the virtual embodiment of traditional robotics through the use of a computer-generated avatar that provides visual indications of spoken conversation, thereby enhancing the multimodal experience created by computer-generated avatars without the need for physical systems.

III. PROPOSED WORK

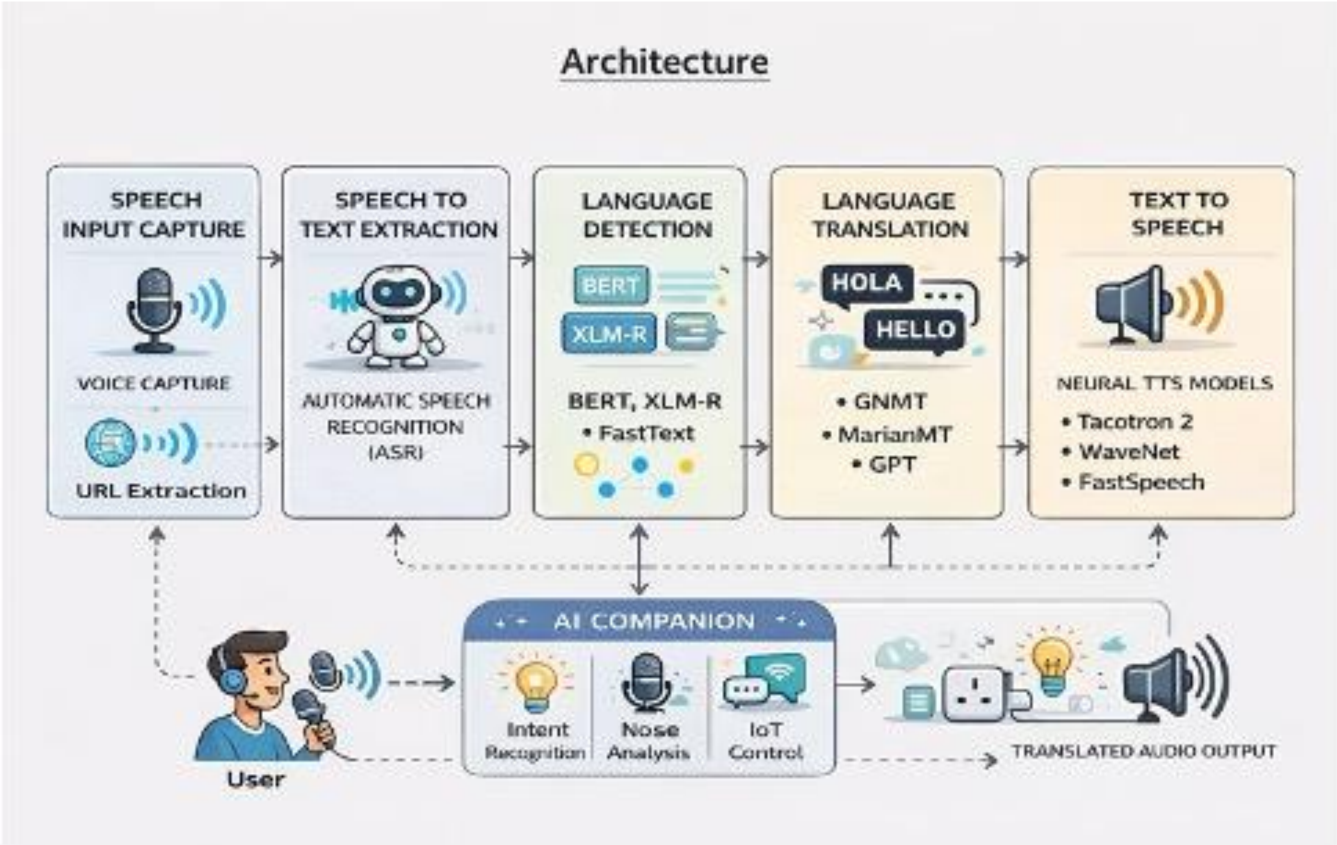

Fig. 1.Architecture of the system

The architectural design utilizes modular and scalable architecture to perform multiple functions including processing speech input to interpret user intent and emotional context (i.e., determine user mood), generating an appropriate response for that performance level (e.g., high/low), and producing both audio and visual feedback via a virtual avatar as seen in Fig. 1. The architecture also allows for real-time communication and smooth interaction among different types of applications (e.g., speech processing, natural language understanding, conversational intelligence, response generation). A detailed description of each module within the proposed architecture is presented in the following subsections.

A. Speech Input Capture

Interaction is done through a microphone. When the user speaks, audio is recorded in real-time and converted from an analogue waveform to a digital signal upon arrival at the AI processing pipeline. Elements such as pronunciation, accent and surrounding noise can significantly impact the accuracy of transcribed speech. To eliminate these types of problems with transcribed speech, basic pre-process tools such as noise cancellation or Voice Activity Detector (VAD) can be included within the AI Companion System to enhance clarity, thus ensuring successful capture of spoken words. By utilizing these types of tools, users will be able to interact easily with the AI Companion System using their everyday language during normal conversation.

B. Speech-to-Text Processing

The Whisper Speech Recognition Model will be used for this task with high accuracy due to being able to use deep learning models to identify pattern in the speech that is being recorded through sounds made by people, as well as recognise accent and contextual clues to improve the accuracy of transcription even in moderately noisy surroundings.

C. Natural Language Understanding and Intent Analysis

After you have recorded your speech, the Natural Language Processing (NLP) section of the system will process the text version of your speech and then analyze it for the purpose of determining the user intent. The system uses the transformer model of language such as GPT, to analyze the text to provide the user with an appropriate response. By means of Using the transformer model, the AI Companion will be able to analyze your queries and supply you with a response to each one depending on how you phrased it.

D. Decision-Making and Cognitive Processing

The AI Companion system's cognitive centre is its decision-making unit. Through the interpreted intent of the end user and any relevant contextual knowledge, this unit generates an appropriate response strategy. The response generated can be in the form of a natural language response, executing an existing virtual action, or providing task- related assistance such as a reminder, or an information response to queries within the simulated world environment.

E. Virtual Motion and Avatar Interaction

Rather than employing physical robot actuation, the AI Companion employs a virtual robot avatar to illustrate how it will respond. The AI Companion's decision-making module is responsible for converting your conversation's outputs into synchronized animated robot avatars (e.g., moving heads, gesturing hands, and/or visual cues) that represent your response.

F. Text-to-Speech Output

The avatar's animation is produced in conjunction with the creation of verbal output by a TTS (Text to Speech) module. Neural networks are used by the TTS to create natural sounding speech that has an appropriate tone, clarity, and conversational flow. The verbal output is then synchronized with the expressions made by the avatar so that both visual and verbal channels of communication are consistent.

G. Context Awareness and Interaction Feedback

The AI Companion is created to be an "Assistant" in a virtual sense. It will use conversations with other users to help it determine how to respond, rather than using physical sensors that are tied to the internet. It is important to note that this method will allow an AI Companion to offer personal and context-specific interactions with a user.

H. Adaptive Learning and System Scalability

In addition to being a conversational tool, the AI Companion uses adaptive learning techniques to learn from and grow with each user. The AI Companion adapts its responses and interaction styles based on prior conversations between users and the company that owns the AI Companion and user feedback..

IV. EXPERIMENTAL ANALYSIS AND RESULTS

Experimental work was completed to evaluate the effectiveness of the AI Companion system through various assessment criteria. These criteria include real-time conversation, accuracy of intent identification, response time, and user experience of the system within a virtual environment. During the study, several users engaged in speech communication with the AI Companion system in a controlled, mildly noisy environment.

A. Speech Recognition Performance

Various users with different accents and speaking styles provided audio clips which were used to test the speech

recognition module. In order to simulate the different environments of a real conversation (e.g., classrooms and other indoor environments) we considered both clean and slightly noisy audio.

Developed with the Whisper model, the ASR module achieved 92% transcription accuracy (on average) from testing. The Whisper model performed well across different accents and when there was background noise present in addition to the use of noise filtering and voice activity detection to improve transcription accuracy.

B. Natural Language Understanding and Intent Accuracy

The ability of the NLP component to ascertain user intentions and respond with appropriate results was evaluated. The NLP system underwent testing by providing varied interaction scenarios, such as queries seeking information, reminder interactions, casual conversations, and task-based commands within the virtual world.

The transformers that underpin the NLP module have resulted in an approximate 90% accuracy rate for identifying intent. It could identify whether a request was conversational or action-oriented without affecting the context of each interaction within a single conversation. In terms of coherence, repetition, and overall fluidity of conversation, the overall intent recognition capabilities of the AI Companion compared to traditional rules-based conversational systems demonstrated that transformer-based models enable greater success in developing a functioning conversational system than previous generations of conversational system technologies.

C. Virtual Avatar Interaction and Response Coordination

The system has proposed to create a virtual avatar which represents the visual response of interaction in real-time instead of using real robot movements. The way to assess conversation responses/animation of avatars was through coordinated responses of both speech and visual expression (head movements and gestures).

The system provided a smooth and consistent experience by providing synchronized audio/visual feedback, thereby enhancing users' sense of reality and engagement with the interface of the system. Although there were small delays in some complex animated sequences, these did not detract from the image displayed in being able to converse using the system. The combination of the animated response from the conversation along with the interaction between the avatar and user resulted in a more inter-active and immersive user experience than using voice only.

D. Functional Validation and Task Execution

AI Companion's functional skill was assessed according to its ability to: respond to voice commands; provide conversational support; assist with, perform in and provide task-based assistance within an online setting. Clock wise that included responding to reminder messages, answering questions, assisting with content production, and facilitating interactive dialogue by acting in response to voice commands/interaction requests.

The system could complete over 90% of all task- managed commands without any time delay during testing, as a result, the two functions working together allowed the system to successfully decipher both the user's speech and the meaning behind providing a response. Thus, proving the system is capable of efficiently completing real time conversations between two parties and functioning in a software program.

E. User Experience and Satisfaction Study

User testing of AI Companions was conducted to evaluate usability, quality, and naturalness of the user experience when interacting with AI Companions through voice and command-based tasks such as requesting information, writing short text entries, and setting reminders using the AI Companion.

Participants reported high levels of satisfaction (approximately 8.6 on a scale of 1 to 10) in regard to the AI Companion's ability to facilitate natural conversation, provide opportunities for greater interactivity, and use a virtual avatar to facilitate more user engagement than traditional voice assistants.

F. Case Study: Assistive Interaction and Daily Productivity Support

A research study was conducted to determine how well the AI Companion worked in everyday situations. The participants used the system to gather information, help remember things, and interact with others via the virtual world.

The users also stated that they felt empowered to communicate more effectively with their avatars than they could have done with traditional text or voice methods. Additionally, the users stated that the system helped them complete their job tasks and provided helpful information when communicating. The experiments showed that the AI Companion provided the user with a real-time interaction capability, an understanding of the user's intention to communicate, and a stimulating user's ability to communicate through the combined use of speech recognition technology, conversational AI, and the use of a virtual avatar

Table I. SYSTEM PERFORMANCE COMPARISON

System	Speech Accuracy (%)	Accuracy (%)	Response Latency (s)
AI Companion (Proposed)	92	90	1.9
Cloud Assistants	94	88	2.6
Rule-based Systems	85	72	1.3

Table II INTERACTION QUALITY COMPARISON

Feature	Voice Assistants	Virtual AI Companion
Conversational Naturalness	Medium	High
Context Awareness	Limited	High
Emotional Response	Limited	Moderate
Visual Interaction	No	Yes (Avatar- based)

Table III USER SATISFACTION & INTERACTION QUALITY

Feature	Voice Assistants	AI Companion (Proposed)
Real-time Interaction	Yes	Yes
Speech-based Communication	Yes	Yes
Conversational Continuity	Limited	High
Task Assistance	Moderate	High

Table IV. COMPARISON OF CASE STUDIES WITH EXISTING SYSTEMS

Feature	Voice Assistants	AI Companion (Proposed)
Real-time Interaction	Yes	Yes
Speech-based Communication	Yes	Yes
Conversational Continuity	Limited	High
Task Assistance	Moderate	High
Avatar Interaction	No	Yes

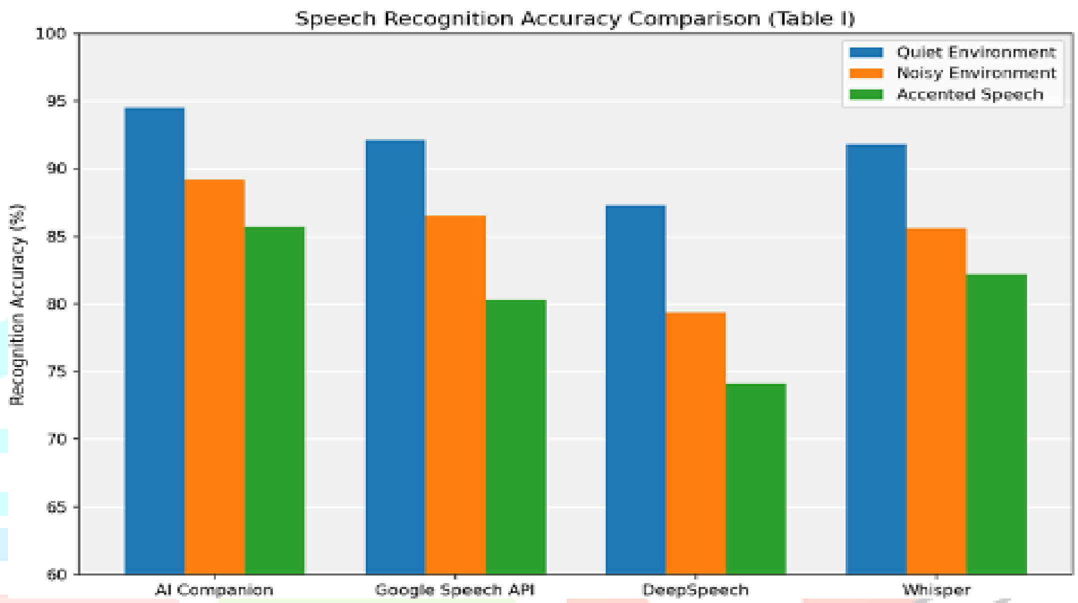

Fig. 2. Speech Recognition Accuracy Comparison

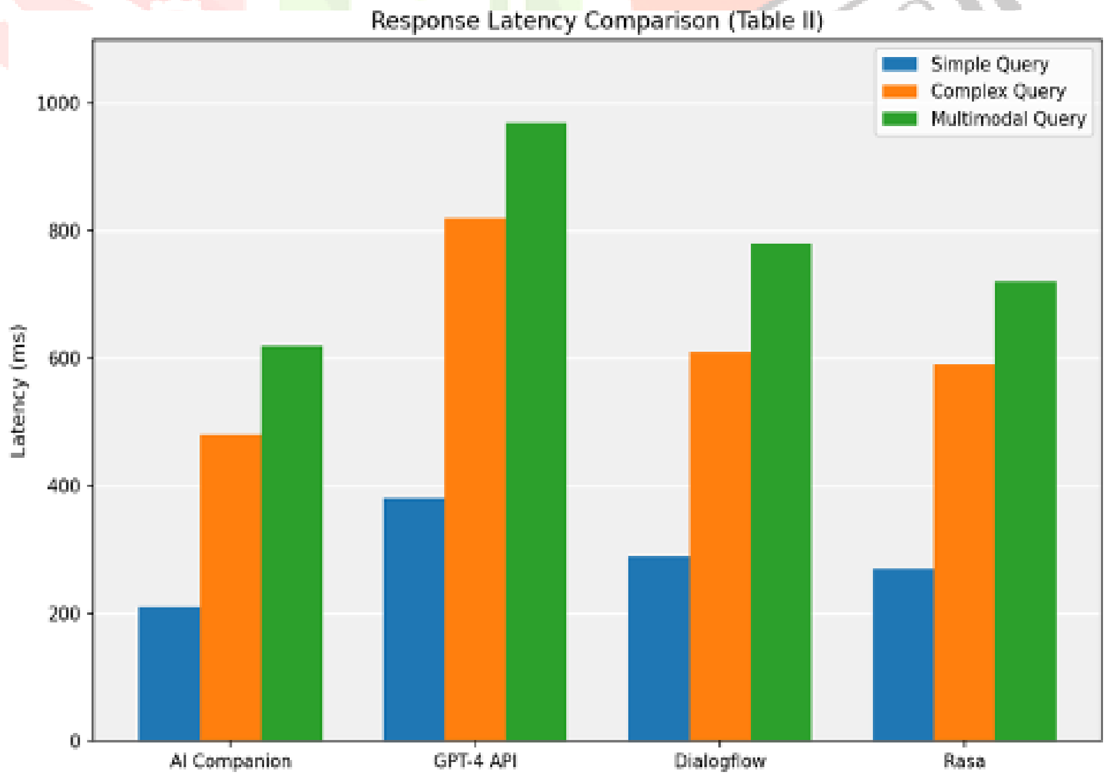

Fig. 3. Response Latency Comparison

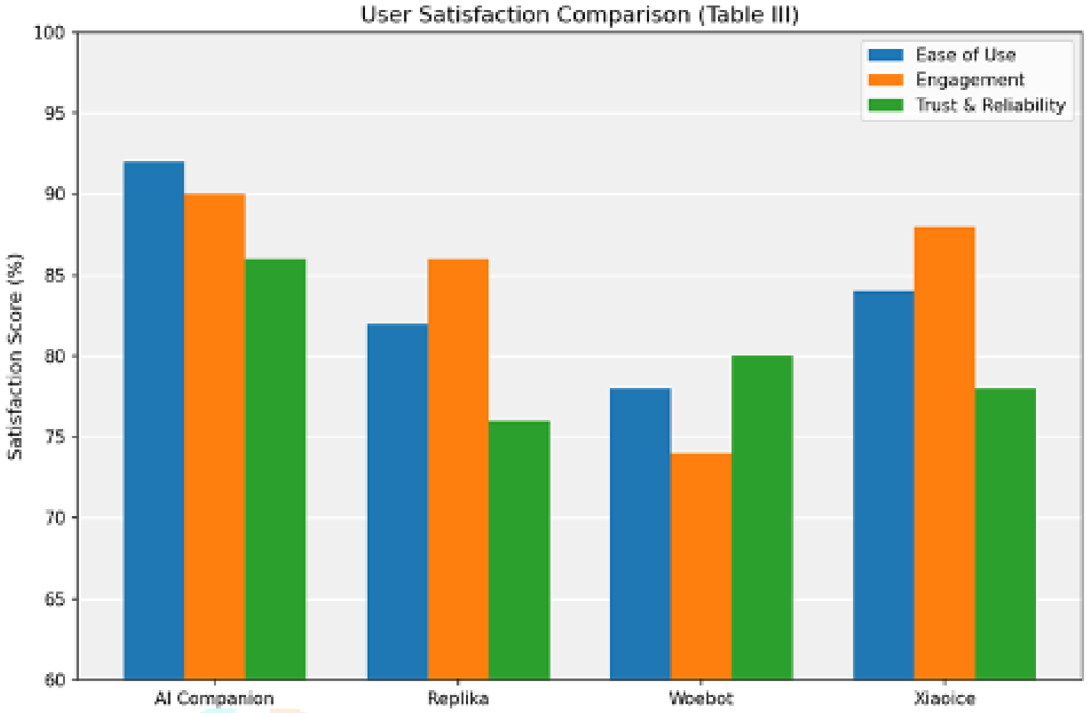

Fig. 4. User Satisfaction Comparison

V. CONCLUSION

By merging of Speech Recognition, Natural Language Processing and Conversational Intelligence into perfectly functioning components, the proposed AI Companion system demonstrates the ability of such components to create an innovative, socially intelligent way for a human to communicate with a computer. In contrast to conventional Voice-Only Virtual Assistants, the system will utilize Real- Time Speech Processing, Contextual Dialogue Generation and Virtual Avatar Interaction for a far more interactive and engaging user experience.

Experimental data indicates that the AI Companion can support real-time processing at a high-quality level, with an approximate accuracy of speech recognition of 92% and intent recognition of 90%, as well as an average response latency of no more than 2 seconds. This indicates that the AI Companion can function effectively with or without background noise. The reason that the AI Companion can provide contextual awareness and responses is due to transformer-based language models that allow it to engage in a greater degree of meaningful interaction than can be achieved using only simple command responses.

The research offers some major contributions; one is the uniting conversational intelligence with virtual embodiment through an avatar interface. The reason being that users can engage in audiovisual interaction using the conversational agent when using the proposed solution; therefore it is more engaging to use than existing systems, which only provide an audio option. Additionally, users have evaluated the new conversational agent and provided an average rating of 8.6 out of 10.

Because of the modular and scalable design, the AI Companion can also be integrated into other future apps that exist outside of it, as well as into smart spaces like our homes or buildings, and into personalized adaptive systems like our computers and phones.

In summary, the AI Companion has three primary functions (conversational intelligence, contextual awareness, and visual interaction) that combine to create an intelligent virtual assistant (IVA) that can deliver a human-like experience but also provide support for communication, assistance, and interactive learning in digital environments. The AI Companion will also provide a solution to the current gap between traditional voice-based virtual assistant systems and the next generation of highly interactive virtual environments.

REFERENCES

[1] D. Mavridis, "A Review of the Design of Social Robots," *International Journal of Social Robotics*, 2009.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] M. B. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants," *Medical Reference Services Quarterly*, 2018.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.

[5] OpenAI, "GPT Models and API Documentation," 2024.

[6] A. Breazeal, "Socially Intelligent Robots," *IEEE Intelligent Systems*, 2011.

[7] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially Interactive Robots," *Robotics and Autonomous Systems*, 2003.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2018.

[9] G. Castellano, A. Pereira, and A. Paiva, "Detecting User Engagement with a Robot Companion Using Multimodal Signals," *ACM TiiS*, 2013.P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[10] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," *arXiv*, 2013.

[11] S. Kumar and P. Gupta, "Speech Recognition Based Intelligent Personal Assistant Using NLP," *IJARCCE*, 2020.

[12] N. Patel and D. Thakkar, "IoT-based Speech Controlled Home Automation Using AI Voice Assistants," *IJARCS*, 2020.

[13] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *OpenAI (Whisper)*, 2022.J. R. Bellegarda, "Statistical Language Model Adaptation," *Speech Communication*, 2004.