ADYPU'S
INTERNATIONAL JOURNAL OF INNOVATION & RESEARCH
VOLUME 1
ISSUE 1

DESIGN AND DEVELOPMENT OF AN INTERACTIVE HUMANOID ROBOT

[1]Vidhey Tailor, [2]Satish Patil, [3]Ipsita Swain
School of Engineering / Ajeenkya D Y Patil University, Pune[1,2]
vidhey.tailor@adypu.edu.in[1], ipsita.swain@adypu.edu.in[2]

ABSTRACT

This project focuses on the design and development of an interactive humanoid robot that closely mimics human facial features and communication abilities to enhance human-robot interaction. The humanoid incorporates advanced 3D-printed facial structures, voice recognition, and speech synthesis to engage users in a natural and intuitive manner. Key aspects include creating anatomically accurate facial expressions, employing sophisticated text-to-speech (TTS) and speech-to-text (STT) systems, and integrating a modular hardware framework that supports adaptability and expansion. Each component, from the mechanical design of the face to the electronic and software systems, was meticulously crafted to facilitate smooth, lifelike interactions. The robot's ability to respond to voice commands and convey facial expressions positions it as a versatile platform for applications in fields such as healthcare, education, and customer service. The modular approach enables future enhancements, such as adding conversational AI and gesture recognition for deeper interaction. This work represents a significant step toward creating humanoid robots that are more relatable and effective in real-world environments, contributing to the advancement of human-centric robotics.

Keyword: Interactive Humanoid Robot, Human-machine interaction, Integration of text-to-speech

INTRODUCTION

Since humans have been fascinated by the idea of creating machines that mimic their own form and behavior. This fascination has fueled decades of innovation, leading to the emergence of modern robotics [1]. Among the various types of robots, humanoid robots stand out for their striking resemblance to humans, both in appearance and functionality [2].

Our research focuses on the design and development of an interactive humanoid robot, with particular emphasis on its human-like head. The head, often considered the focal point of human interaction, allowing the robot to convey responses effectively [3]. By programming the robot with human-like features, we aim to enhance its acceptability and usability in various social and functional contexts [4]. Creating such a robot is a complicated task as it requires expertise in mechanical engineering, electronics, programming, and cognitive science [5]. Each aspect of the robot's design and development presents its own set of challenges, from ensuring smooth and lifelike movements to integrating sophisticated sensors and artificial intelligence algorithms [6]. By embodying human-like characteristics, they can evoke empathy and understanding, paving the way for deeper and more meaningful interactions between humans and machines [7]. The present work aims to contribute to the field of robotics, pushing the boundaries of what is possible and bringing us closer to a future where humanoid robots are not just tools or gadgets but genuine companions and collaborators in our journey through life [8].

In recent years, natural language processing (NLP) models like ChatGPT have significantly advanced human-machine interaction, enabling more intuitive and conversational interactions [9]. However, the unidirectional text-based communication of traditional ChatGPT models poses limitations for users who prefer voice-based

inputs or outputs. Integrating TTS and STT capabilities into ChatGPT writing robots addresses these limitations by enabling bidirectional communication, thus enhancing accessibility and user experience [10].

The whole system is structured into three phases: Design, Development, and Integration of components. Beginning with an in-depth analysis of human anatomy and facial structure to conceptualize and refine the robot head's aesthetics and functionality iteratively [11]. Meanwhile, the development process involved the fabrication of intricate parts using 3D printing technology, employing PLA as the primary material. Concurrently, electronic components such as the ESP8266 microcontroller, speaker module, servo motors, and microphone module were used and tested for circuit and prototyping [12].

DESIGN PROCESS

The design process of the humanoid robot began with the analysis of human anatomy and facial structure in order to bring accurate human resemblance to the robot [1]. This initial phase involved a comprehensive study of key features and proportions of the human face [2]. By understanding and visually examining these aspects, a conceptual design was crafted [3].

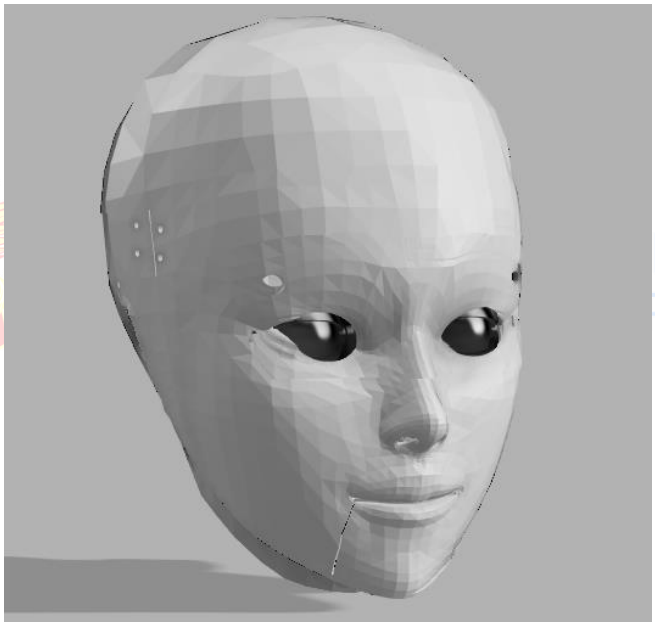

Figure 1: Final design of the humanoid robot

With the design finalized (Figure 1), preparations were made for manufacturing using 3D printing technology. Leveraging the capabilities of 3D printing, the components of the robot head could be fabricated with precision and efficiency [4]. The design created in Fusion 360 seamlessly translated into STL files, ensuring continuity and accuracy throughout the manufacturing process [5].

The development of the head's components, like the face plate, jaw, eyes, mechanism for facial expressions, forehead, and parietal, transitioned from design to assembly parts using 3D printing [1]. In the development, Polylactic Acid (PLA) was used to create the parts we designed earlier [2]. Simultaneously, while the 3D printing process was underway (Figure 2(a)), efforts were dedicated to the development of the electronic components and circuitry essential for the functionality of the robot head (Figure 2(b)) [3].

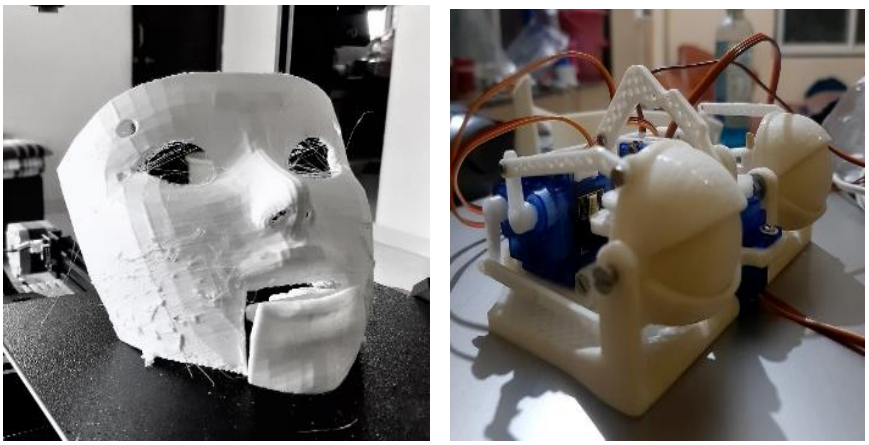

Figure 2: (a) 3D printed Face and (b) 3D printed Eyes mechanism

Integration of Components

Integrating all the different parts of the humanoid robot head was a critical step in bringing it to life. First off, we carefully put together all the 3D printed pieces, making sure they fit just right and leaving space for the electronic components (Figure 3(a)) [4]. Once the mechanical parts were assembled, we focused on getting the electronic side up and running. The programming of the ESP8266 microcontroller, which served as the brain of the robot head, to manage facial expressions, synthesizing speech, and processing audio inputs [5].

Each component was interfaced with the microcontroller according to the predetermined specifications, establishing a network of electronic pathways essential for the robot head's operation [6]. Through careful wiring and soldering, the electronic system was integrated into the mechanical framework of the head, creating a unified platform ready for testing and validation (Figure 3(b)). Any identified issues or shortcomings were meticulously addressed through iterative refinements [7]. These refinements encompassed adjustments to the software code, fine-tuning of hardware configurations, and optimization of mechanical components [8]. This iterative process was instrumental in ensuring that the final product adhered to the desired specifications and upheld stringent quality standards, thereby delivering a seamless and immersive user experience.

Figure 3: (a) integration of components, (b) final assembly of humanoid robot's head

SPEECH-TO-TEXT AND TEXT-TO-SPEECH PROCESS:

Speech Recognition for Robotic Control

The main goal of this paper is to introduce a "hearing sensor" and the speech synthesis to the mobile robot such that it is capable of interacting with humans through spoken natural language [1]. The context of speech recognition refers to a system where a person can speak via a microphone to a computer, and the computer translates the spoken words into either text or commands to execute functions in the computer (Figure 4) [2]. The intelligent speech recognition system enables the robot to understand spoken instructions [3]. The speech recognition system is trained in such a way that it recognizes defined commands, and the designed robot will navigate based on the instruction through the speech commands [4]. The complete system consists of three subsystems: the speech recognition system, a central controller, and the robot [5]. The results prove that the proposed robot can understand the meaning of speech commands and will act autonomously in a natural environment, communicating in a natural way with those people they are supposed to support [6].

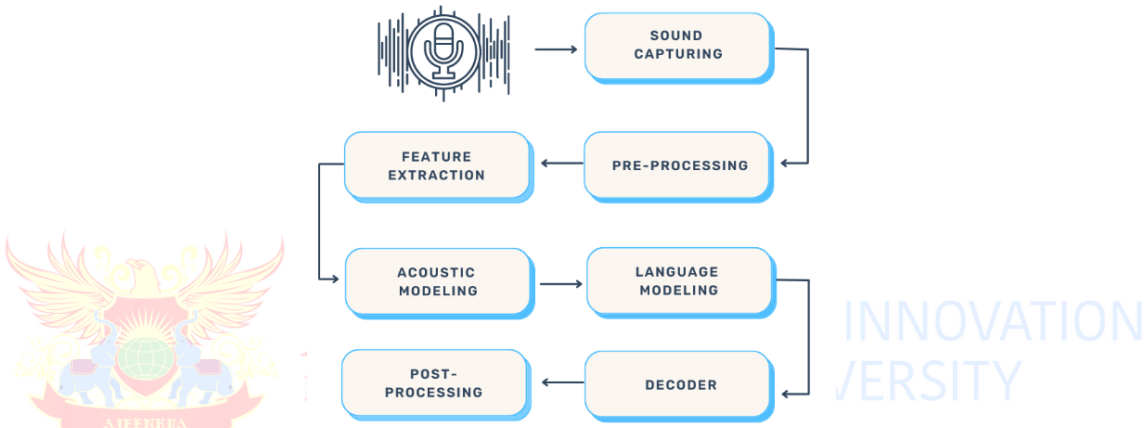

Figure 4: Algorithm for speech to text

The size of the recognition vocabulary

One important factor that directly affects the accuracy and performance of the system is the amount of the recognition vocabulary in the integration of text-to-speech (TTS) and speech-to-text (STT) capabilities in ChatGPT writing robots [7]. The collection of terms and expressions that the STT component can reliably translate from speech input into text is referred to as the recognition vocabulary [8]. Small vocabulary systems provide recognition capability for up to 100 words, medium vocabulary systems provide recognition capability for from 100 to 1000 words, and large vocabulary systems provide recognition capability for over 1000 words [9].

The knowledge of the user's speech patterns

Understanding user voice patterns is crucial for enhancing the accuracy, effectiveness, and overall user experience of ChatGPT writing robots that integrate text-to-speech (TTS) and speech-to-text (STT) features [10]. Analyzing and simulating different facets of users' verbal communication, such as their pronunciation, vocabulary usage, syntax, and conversational style, is necessary to understand their speech patterns [11]. Speaker-dependent systems are custom tailored to each individual talker, while speaker-independent systems work on broad populations of talkers, most of which the system has never encountered or adapted to [12].

The degree of dialogue between the human and the machine

In one-way communication, each user spoken input is acted upon. In system-driven dialog systems, the system is the sole initiator of a dialog, requesting information from the user via verbal input [1]. In natural dialogue systems, the machine conducts a conversation with the speaker, solicits inputs, acts in response to user inputs, or even tries to clarify ambiguity in the conversation [2].

Sounds and background interferences

The sound of human speech causes vibrations in the atmosphere, and a computer must go through several difficult steps to translate speech to text that appears on the screen [3]. Enough vectors in the form of acoustic speech vectors representing utterances via the communication channel are extracted from the client's inquiry during the speech conversion procedure (Figure 5). Subsequently, the digital audio is filtered by the system to eliminate undesirable signals and noise, and occasionally, it is divided into distinct frequency bands [4].

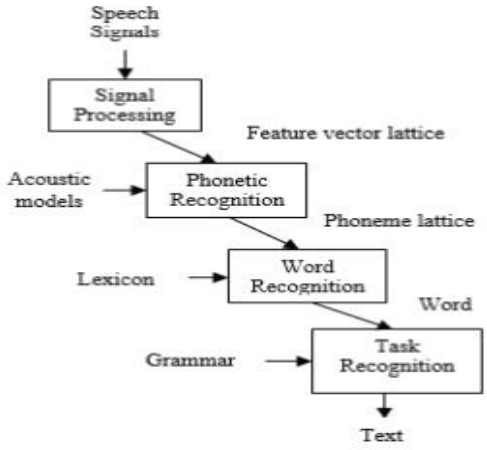

Figure 5: Algorithm for speech to text and text to speech

METHODOLOGY

In the present project, an interactive voice-driven system where users communicate with ChatGPT, a sequence of advanced components works together to create a seamless, conversational experience. This process begins with the Speech Input stage. The user speaks their query or command into a microphone or other audio input device, initiating a multi-step interaction. This initial audio input captures the user's voice, which could include a wide range of questions or commands. For instance, a user might ask, "What is the weather today?" or "Tell me a fun fact about space." This spoken language is the foundational input that drives the entire interaction with ChatGPT. Once the audio input is captured, it moves to the Speech Recognition (STT) component. STT stands for Speech-to-Text, a technology that translates spoken words into written text. By processing audio signals, this component effectively transcribes the user's voice into readable text format. For example, when a user says, "What is the weather today?" the STT system converts the audio into the text form "What is the weather today?" This transcription serves as the bridge between the user's speech and the system's text-based processing capabilities. Speech recognition must be accurate, as errors here can lead to misunderstandings in subsequent steps, so it often uses robust machine learning algorithms trained on extensive datasets to improve accuracy across different accents, speech patterns, and languages.

After transcription, the text is passed to the Natural Language Understanding (NLU) component. This part of the system is crucial for understanding the user's intent and the context of their query. NLU breaks down the text,

identifying keywords and phrases to interpret the user's underlying needs. In this weather example, the NLU would identify that the text "What is the weather today?" is a request for current weather information. The NLU component is designed to understand various types of questions, commands, and conversational language, extracting meaning from the input text to determine what the user is actually asking. NLU can identify intentions beyond literal meaning, allowing it to handle casual language or complex queries accurately.

Once the NLU has processed the query, the structured information is passed to the ChatGPT model. ChatGPT is a large language model designed to generate relevant, context-aware responses based on the input it receives. Using the information from the NLU component, ChatGPT crafts an appropriate response. For example, in the case of a weather query, ChatGPT might produce a response like, "Today's weather is sunny with a high of 75 degrees Fahrenheit." This response generation step is central to making the conversation feel natural, as ChatGPT can handle a wide range of topics and generate answers in a conversational tone.

The response text from ChatGPT is then sent to the Text-to-Speech Synthesis (TTS) component. TTS is the final stage in this process, responsible for converting the response text into synthesized speech. Using advanced neural network models, TTS creates audio that closely mimics natural human speech, with a cadence and tone suitable for the response context. For instance, it might voice a weather response in a calm, informative tone. This synthesized speech is then played back through a speaker or audio output device, allowing the user to hear the response.

This voice-based interaction pathway—from speech input to text output—is designed to make communicating with ChatGPT feel like a natural, human-like conversation (Figure 6). Each component in the chain plays a specialized role in understanding, processing, and responding to the user's query, working together to deliver accurate, real-time answers. Through this carefully orchestrated series of steps, users can experience hands-free, conversational interactions with ChatGPT that feel both intuitive and informative.

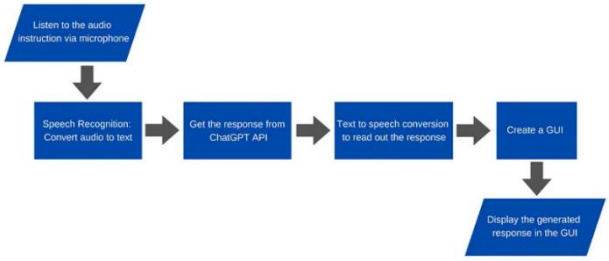

Figure 6: Schematic showing the workflow

CONCLUSION

This project successfully demonstrates the design, development, and integration of an interactive humanoid robot with human-like facial features, voice recognition, and speech synthesis capabilities. By closely emulating human facial anatomy and expressions and incorporating advanced text-to-speech (TTS) and speech-to-text (STT) technologies, the robot offers an engaging and natural interaction experience for users. Each stage of the development process—from conceptual design and 3D printing of components to electronic integration and programming—was carefully optimized to achieve fluid, life-like interactions, thereby enhancing human-machine engagement.

The implementation of speech recognition and synthesis enables the robot to process voice commands and respond interactively, addressing a significant challenge in human-robot interaction. This interactive capability

not only makes the robot suitable for various social and functional environments but also sets the stage for further enhancements in NLP and gesture-based communication. The modular approach used in this project ensures adaptability, enabling future expansions in both hardware and software to accommodate new features, such as advanced conversational AI or emotion recognition.

In summary, the project contributes valuable insights to the field of humanoid robotics by demonstrating a scalable framework for creating relatable, functional robots. This work aligns with the broader goal of advancing robotics technology to be more accessible and intuitive in everyday applications, whether in healthcare, education, or customer service. Future research could explore additional sensory capabilities, enhanced mobility, and AI-driven personality modeling to further enrich human-robot interaction.

REFERENCES:

[1] R. Tadeusiewicz, "Speech in human system interaction," in *Proc. 3rd Conf. Human System Interactions (HSI)*, 2010, pp. 2–13.

[2] F. Jelinek, Statistical Methods for Speech Recognition, Massachusetts Institute of Technology, 1997.

[3] J.-C. Junqua, Robust Speech Recognition in Embedded Systems and PC Applications, Kluwer Academic, 2000.

[4] L. Prina Ricotti and C. Becchetti, Speech Recognition: Theory and C++ Implementation, Wiley, 1999.

[5] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 1, pp. 35–45, 1990.

[6] J. P. Olive and M. Y. Liberman, "Text to speech—An overview," The Journal of the Acoustical Society of America, vol. 78, no. S1, 1985.

[7] D. Kalla and N. Smith, "Study and analysis of Chat GPT and its impact on different fields of study," International Journal of Innovative Science and Research Technology, vol. 8, no. 3, pp. 827–833, 2023. M. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[8] H. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Springer, 1993.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[10] D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer, 2015.

[11] J. A. Benesty, S. Makino, and J. Chen, Speech Enhancement, Springer, 2005.

[12] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in Proc. Eurospeech 97, 1997, pp. 371–374.

[13] K. Nakamura, K. Kinoshita, M. Delcroix, and T. Nakatani, "Speech dereverberation," in Handbook of Signal Processing in Acoustics, M. Havelock, S. Kuwano, and M. Vorlander, Eds., Springer, 2008, pp. 1–18.

[14] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," Communications of the ACM, vol. 57, no. 1, pp. 94–103, 2014.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.