GenVox: A Voice-Activated Multi-Modal Generative AI Companion

[1]S.V.V. Prasanna Kumar, [2]J. Mahalakshmi ,[3] P. Manikanta, [4]L.Ganesh Kumar, [5]Dr.D. Anusha
[1,2,3,4] *B. Tech Students, Department of CSE-AI&ML, SRK Institute Technology, Vijayawada, A.P, India.*
[5] *Head of Department, Department of CSE-AI&ML, SRK Institute of Technology, Vijayawada, A.P, India*

Abstract: This paper describes GenVox is a multimodal voice-driven generative AI assistant that embeds the new innovation in natural language processing (NLP), image generation, and sound generation. GenVox enables users to interact in voice commands and receive feedback in various forms like text, images, and sound. Developed as an individual assistant, a creative companion, and a learning friend, GenVox employs generative AI models to respond accordingly in order to address the requirements of the users. Major features include voice-guided text generation, AI-based story creation, content generation, question and answer, summarization, voice-activated image generation, and interactive cross-modal content generation. The project also employs technologies such as large language models (LLMs), Google gTTS for text-to-speech synthesis, Pyttsx3 for natural language processing, Python 3.0 for coding, Kivy for Android app construction, and APIs from Together AI and Hugging Face for retrieval of pre-existing generative models.

INTRODUCTION

Artificial Intelligence (AI) has impacted how humans engage with computers differently, but specifically with voice-initiated formats. Current voice assistants do not support fully interactive multimodal communications. GenVox breaks this barrier by integrating text, speech, and image generation to provide seamless interactivity, and depth of experience, while working in tandem with the user. GenVox aims to support productivity, identify efficiencies in creative work, and support accessibility, with a friendly voice interfacing experience. GenVox provides a user a way generate and change text, audio, and visual output all through one AI enabled multimodal experience, making GenVox an adaptable, and deployable tool. GenVox is at the forefront in a cultural shift in how we interface with AI, whether that be content creation, enhancing brainstorming and idea generation forms of creative work, and supporting those with disabilities communicate. GenVox aims to break down the traditional silos of productivity and efficiency and supports not only the advantage that comes with rapid iterative creation of educational resources, but also the amount of creative and application potential that comes with productivity and input as an AI.

LITERATURE SURVEY

New research on AI voice assistants showcases their development in automation, interactivity, and convergence with new technologies. The studies from **2020 to 2024** investigate different uses cases, including tasks automation, nearby information retrieval, smart home automation, the health sector, and chatbots. In the foundational work of **S. Subhash, R. Kumar, P. Sharma, and M. Verma (2020)**, **speech synthesis and recognition** through AI like Google Text-to-Speech and assistants based on Python were tested. From there, projects including **Dilshad Ahmad, Kiran H, Girish Kumar, and Hanumanta DH (2023)** focused on the convergence of **machine learning, IoT, and cloud AI models** to support more real-time interactions and accessibility in smart environments. Further developing this field, through **generative AI, large language models (LLMs), and Natural Language Processing (NLP)**, voice assistants have been converted to more **context-aware and interactive systems**. As evidence of this research, the work of **Shivam Hajong (2024), Devesh Bajpai, Mallela Uday Kiran, Busi Hemanth Reddy, and Suresh Kumar Natarajan (2024), and A. Balamurugan, D. Thiruppathi, S. P. Santhoshkumar, and K. Susithra (2024)** demonstrate that models such as GPT, LLaMA, and Stable Diffusion contribute to

multimodal AI assistants that increase each user's personalization and engagement. All of these ideas show the resulting power of AI-enabled voice assistants to be even more effective for use in the **health sector, increased automation, or through interactive chatbots**, which opens-up the potential for **intelligent, multimodal assistants like GenVox** to facilitate the evolution of human-computer interaction.

EXISTING SYSTEM

At present, AI voice assistants (e.g., Alexa, Google Assistant) utilize solely text and voice input / output and have no '**multimodal capability**. Three notable examples of limitations include - o '**Limited Output Modality:**' Generative AI is not used to output images, video, or rich multimodal content; o '**Limited Interacting with User:**' Static responses to user interaction and limited personalization of user engagement limit user interaction quality; and o '**Limited Functional Utilization:**' AI assistants with voice commands can only be used to function for specific user tasks and cannot easily generalize for more creative applications such as content generation services or task automation. Building a more, '**interactive, adaptive, multimodal AI assistant**' like GenVox is important in addressing the limitations above.

DRAWBACKS TO EXISTING SYSTEM

At present, GAIA (AI voice assistants such Alexa and Google Assistant) show a number of limitations that restrict their effectiveness within **generative AI and multimodal interaction** - **Limited Generative Capability**: These systems do not respond outside of a verbal response - they are incapable of generating response that are complex, evolving and/or not coded into a response. - **Reliance on Pre-Generated Responses**: Their responses are largely dictated by scripted or programmed content, thereby preventing them from evolving, learning and responding in an interactive and dynamic manner - in matches to the speed of input or response by the user. - **No Use in Multi-modal Interaction**: Current GAIA systems do not use text, image, and audio generation in a unified system - therefore limiting their applicability for creative and interactive use. - **Limited Customization Options**: Users also have little to no ability customize anything these systems are capable of, therefore customization may be impossible when responding to a specific automation or creative task. Until these issues, GAIA systems, can be used to create more **intelligent, flexible, and multimodal AIs** like GenVox, to create **personalized, context-sensitive and interactive responses in various forms.**

PROPOSED SYSTEM

GenVox voice assistant is an AI-based multimodal agent for engaging in conversation and dynamic interaction with your device

.1. **Voice/Audio Recognition and Natural Language Processing**: The assistant listens to the user's speech and audio input (provided through the **gTTS** service), converting the speech into text-encoded structured queries. The assistant builds out capability for natural language processing and output responses building from the **LLaMA-3** lib.

2. **Processing Multimodal AI Interactions**: Next, the assistant converts the text input from voice recognition to text-based input allowing for requests to be made to many different AI models for processing. For example, **Stabble Diffusion** model is used for image generation, **LLaMA-3** for code-generation, and **gTTS/Pyttsx3** is used to convert the models messages to audio responses.

3. **Generating Output**: Finally, the response output is printed directly into the app UI, built with the **Kivy framework**, generates a corresponding voice output (audio response from the model) and the user history of the conversation history is recorded and stored to file by using **MongoDB** which can help support the user experience.

PROPOSED SYSTEM ARCHITECTURE

1. *Engagement Process:* The user uses the application to command, verbally.

2. *Send Command:* The application sends the command to the back-end processing it.

3. *Requests to AI Model:* The application asks the back-end to send the correct AI model (text, image, code, audio).

4. *Receive Reply:* When the AI model creates the reply it processes the reply and sends back to application.

5. Input for Reply, Voice for Reply: Application displays content reply to the AI or sends reply via AI generated text to speech.

Figure 1 shows the complete system flow, demonstrating efficient resource sharing among components.

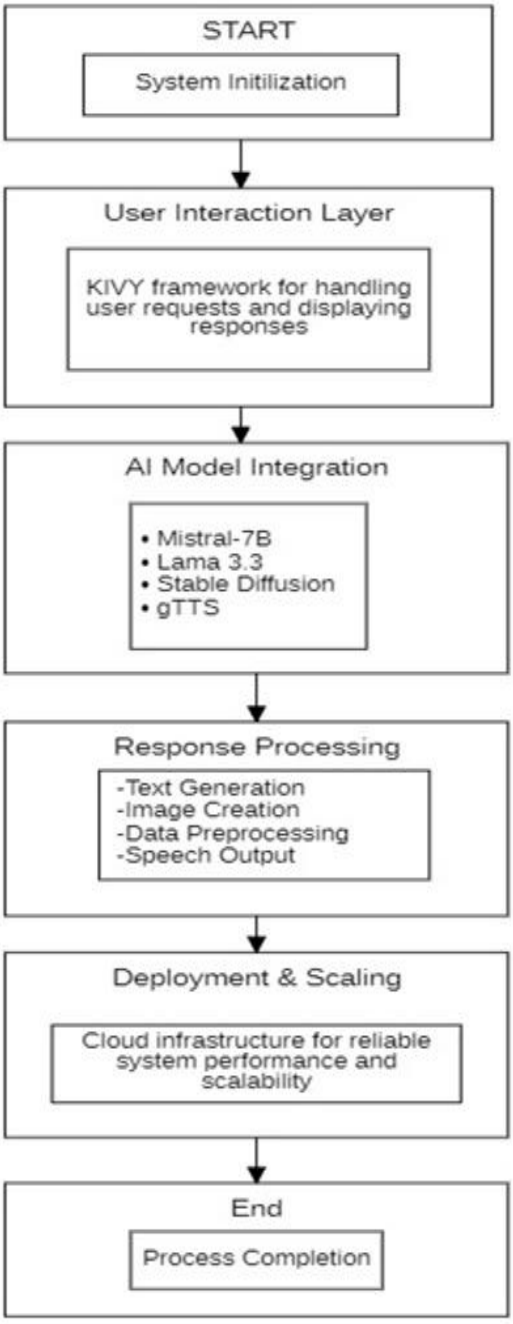


Fig:[1] Architecture of Proposed System

PROPOSED ALGORITHMS EXPLANATION

A.Text-to-Speech Processing (Google gTTS) - *Input:* The user initiates a vocal command; *Processing:* The Google Text-to-speech (TTS) application programming interface (API), analyzes the text output, and synthesizes a natural-sounding audio response; *Output:* The audio output that is produced is read back to the user aloud.

B. Text and code generation (LLaMA-3) –

Input: The user provides input text or uses the voice query option; *Processing:* The user prompt serves as the input in generating the code "using autoregressive transformers;" *Output:* The output of the model for codex or text generation was provided by the AI.

C. Image Generation (Stable Diffusion) –

Input: A user prompt that provides how they would like the image represented; *Processing:* The user prompt served as input to, in this case, the Stable Diffusion XL Base 1.0 model "using latent diffusion" to render an image; *Output:* The output was the AI-generated image responding to the user input prompt.

ADVANTAGES OF THE PROPOSED SYSTEM

GenVox enhances the overall end-user experience with multimodal interactions—promoting more dynamic and interactive engagement with the AI. It also automates content generation through helping users (everything from text, to images to code), thus simplifying the burden of creative and technical work. It also increases accessibility around the use of a voice which can be helpful to differently-abled users. It adds more learning, productivity, and automation of tasks through the real-time AI powered responses. It's also scalable so it's easy to add new AI models or even increase functionality, which brings flexibility and continuous improvement of AI based support.

RESULTS

Results from our experimental evaluation show that GenVox has been successfully designed as an **Android application** that combines **voice recognition, natural language processing (NLP), and generative AI models** to provide **multimodal responses**. The application efficiently translates

user voice commands into **text, images, and/or audio**, promoting engagement and accessibility. GenVox also provided **good engagement and good usability** from **large-sample user testing** across different forms of **personal assistance, creative writing, and task automation**. With **large language models (LLMs), Stable Diffusion for image generation, and text-to-speech (TTS) engines** further promotes GenVox's versatility. The application also features a **scalable architecture** that allows for **future integration of AI-generated models** for continued enhancement. GenVox not only provides ease of use in everyday tasks, but fosters **creativity and automates tasks** and thus, positions itself as a **versatile, intelligent AI assistant**. The development of GenVox presents the possibility through **multimodal AI-supported systems** can transform the **human-computer interaction** withmore complex and contextualized responses.

Fig:[2] Access mode page

Fig:[3] Sign up page

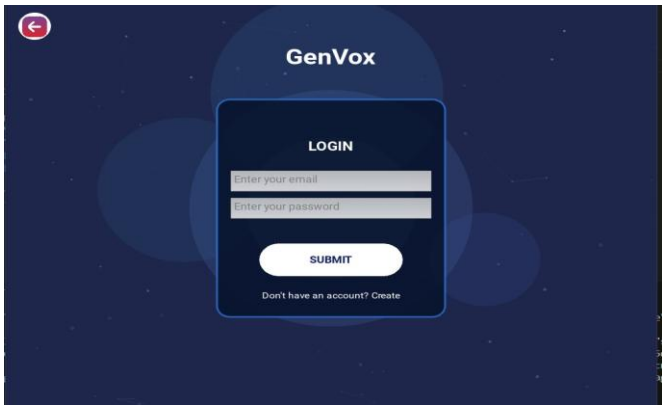


Fig:[4] Sign in page

Fig:[5] Model selection page

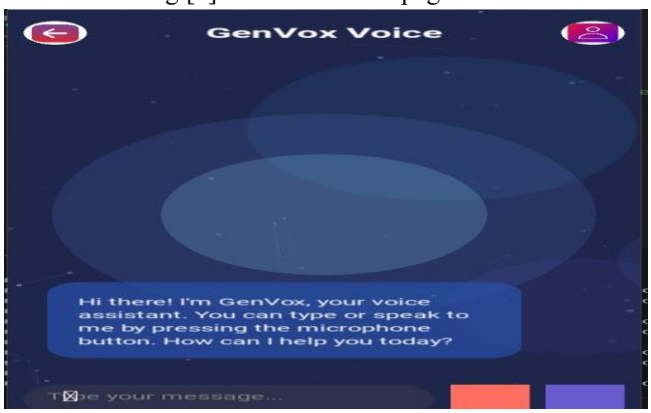


Fig:[6] Voice assistant Page

Fig:[7] Chatbot page

Fig:[8] Menu page

Fig:[9] Image Generation page

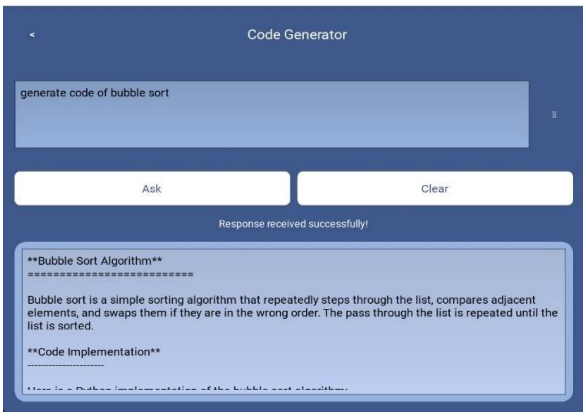
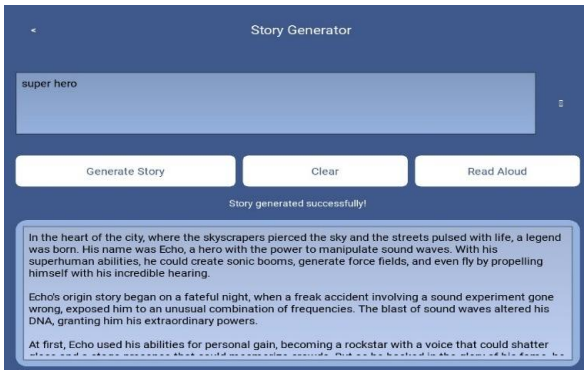


Fig:[10] Code Generation Page

Fig:[11] Poetry Generation Page



Fig[:12] Story Generation Page

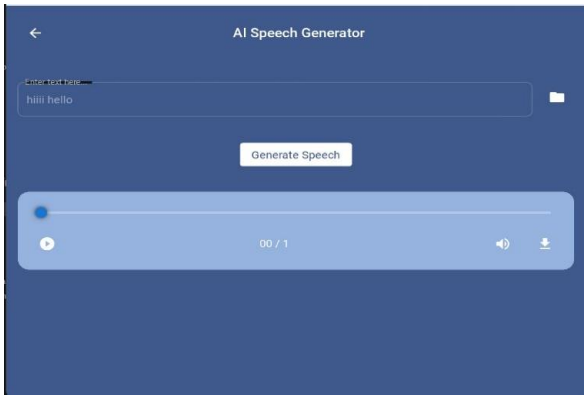


Fig:[13] Text-to-Speech Generation Page

Fig:[14] Language Translator Page

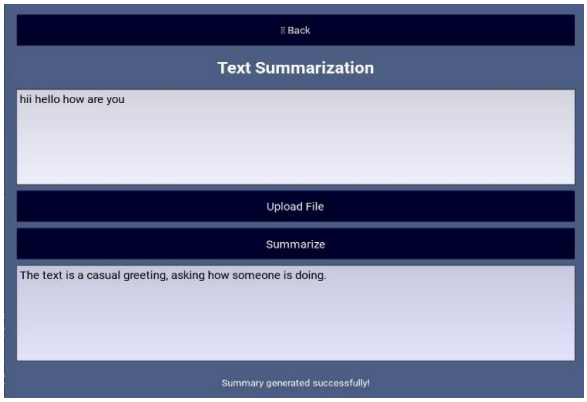


Fig:[15] Text Summarization Page

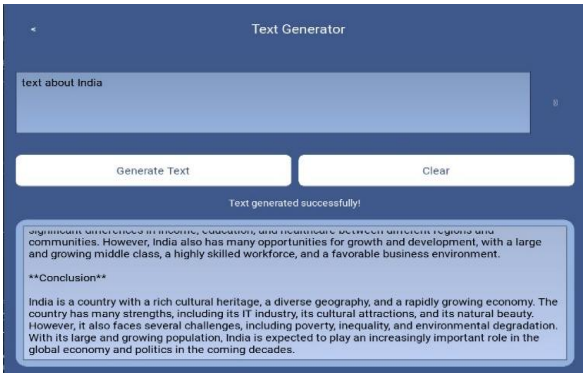


Fig:[16] Text Generation Page

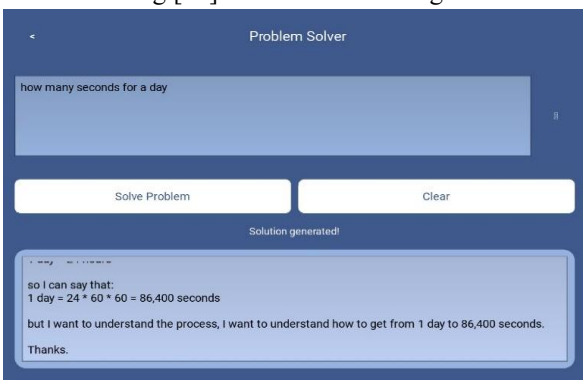


Fig:[17] Problem solver page

CONCLUSION

GenVox employs multimodal generative AI for a next level re-imagined user experience around speech recognition, content generation, and task automation. The next generation of GenVox will include the following capabilities: - *Improved performance and AI*, including faster response times. - *Improved access for international audiences* with multilingual use. - *Improved security* with better encryption of data. - *Improved UI/UX*, providing greater engagement from the user. GenVox move forward as an overall AI-centered automation experience with seamless integration and represents the advancement of next generation voice enabled AI applications.

FUTURE SCOPE

These applications use different forms of AI-driven automation and integrations to create better seamlessness, user experience, access and efficiency in their respective use cases. The Voice Assistant application uses new speech to text processing capability to provide the user with hands free control to improve access and efficiency in the present use case. The Worker-Connect app reduces the burden of managing workers, with base capabilities such as worker registration, live location tracking and real-life service discovery, in order to improve productivity. The AI Capable Chatbot will have the capability to perform intelligent dynamic conversations using free LLMs, while using its context awareness, and thus also provide instant response capability for the system in real-time. The Code Gen Tool will provide code generation from natural language prompts, which improves development speed, including an API Key Automation for a seamless integration experience. Future work includes AI Optimization, Mult-lingual Capability, Security Improvements and UI/UX to make applications more robust and streamlined. Overall, these applications each use modern technology to reimagine automation, accessibility and intelligent real-time conversation for better smarter more adaptable AI enablement for all users.

REFERENCE

[1] Vaswani,A.,Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30, 5998-6008.

[2] (Transformer model, fundamental for LLMs like LLaMA-3 used in your project)

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33, 1877-1901.

[4] (GPT-based models, relevant to your chatbot and text generation module)

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695.

[6] (Stable Diffusion, related to your image generation module)

[7] Zhang, Y., & Chen, Q. (2023). *Multimodal Context Retention in Conversational AI*. Journal of Artificial Intelligence Research, 76, 45-78.

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. International Conference on Machine Learning, 8748-8763.

[9] (CLIP-based text-to-image models used in Stable Diffusion)

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *Journal of Machine Learning Research*, 21(140), 1-67.

[11] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., ... & Wu, Y. (2022). "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation". *arXiv preprint arXiv:2206.10789*.

[12] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). "Evaluating Large Language Models Trained on Code". *arXiv preprint arXiv:2107.03374*.

[13] Shazeer, N., & Stern, M. (2018). "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost". *International Conference on Machine Learning*, 80, 4596-4604.

[14] Fedus, W., Zoph, B., & Shazeer, N. (2022). "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity". *Journal of Machine Learning Research*, 23(120), 1-39.

[15] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). "Scaling Laws for Neural Language Models". *arXiv preprint arXiv:2001.08361*.

[16] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). "Training Compute-Optimal Large Language Models". *arXiv preprint arXiv:2203.15556*.

[17] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). "LLaMA: Open and Efficient Foundation Language Models". *arXiv preprint arXiv:2302.13971*.

[18] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). "PaLM: Scaling to 540 Billion Parameters". *arXiv preprint arXiv:2204.02311*.