

An Agentic AI-Based Voice Interface for Dynamic Website Interaction using Retrieval-Augmented Generation

**Dr. Gurrampally Kumar¹, P.Sai Shankar², B.Karthik², K.Madhav Rao², A.Rutvik
Yadav²**

¹Associate Professor, ²UG Student, ^{1,2}Department of Artificial Intelligence & Machine Learning

^{1,2}J. B. Institute of Engineering & Technology (UGC-Autonomous), Moinabad,
Hyderabad 500075, Telangana.

*Corresponding author: P.Sai Shankar (saishankarpunna@gmail.com)

ABSTRACT

With the rapid growth of web-based platforms and digital content, accessing relevant information from websites has become increasingly complex and time-consuming. Traditional website interaction methods rely heavily on manual navigation, which limits user efficiency and accessibility. Although chatbot systems have been introduced to improve interaction, most of them are rule-based and lack contextual understanding, while standalone artificial intelligence models often generate responses that are not grounded in actual website data. To overcome these limitations, this project proposes an AI Voice Agent for Websites that enables dynamic and intelligent interaction using Retrieval-Augmented Generation (RAG). The system extracts website content using automated scraping techniques, processes the data through text chunking and embedding generation, and stores it in a vector database for efficient semantic retrieval. When a user submits a query in text or voice form, the system retrieves relevant information and generates accurate responses using a language model. Additionally, the integration of Speech-to-Text and Text-to-Speech technologies enables seamless voice-based communication, allowing users to interact with websites in a natural and intuitive manner. The system also supports a multi-agent architecture, where separate AI agents can be created for different websites or knowledge bases. The proposed solution is scalable, efficient, and improves user experience by transforming static websites into intelligent conversational platforms.

Key Words: Retrieval-Augmented Generation, Conversational AI, Large Language Models, Semantic Search, Vector Embeddings, Website-Based Chatbot, Web Scraping, LangChain, OpenAI API, Chroma Vector Database, Natural Language Processing, Generative AI, Context-Aware AI, Knowledge Retrieval, AI Agents.

1. INTRODUCTION

The rapid growth of internet technologies and web-based platforms has significantly increased the amount of information available online. Modern websites contain large volumes of structured and unstructured data across various domains such as education, e-commerce, healthcare, and business services. However, interacting with this information is still largely dependent on manual navigation, where users must browse through multiple pages, menus, and links to find relevant content. This traditional approach is time-consuming, inefficient, and often leads to poor user experience, especially when dealing with complex or content-rich websites.

In recent years, advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have enabled the development of conversational systems such as chatbots and virtual assistants. These systems allow users to interact with digital platforms using natural language instead of traditional navigation methods. However, most existing chatbot systems are rule-based or limited to predefined datasets, which restricts their ability to handle dynamic website content. On the other hand, large language models (LLMs) can generate

human-like responses but often produce inaccurate or misleading information when they are not connected to real-time or domain-specific data sources.

To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as an effective approach that combines information retrieval with generative AI models. In this approach, relevant data is first retrieved from a knowledge source and then used to generate accurate and context-aware responses. This significantly improves the reliability of conversational systems by reducing incorrect or unrelated outputs. RAG-based systems are particularly useful in applications where real-time access to specific data is required.

At the same time, voice-based interaction technologies have gained popularity due to their convenience and accessibility. Speech-to-Text (STT) and Text-to-Speech (TTS) technologies allow users to communicate with systems using voice, making interaction more natural and user-friendly. While voice assistants such as Alexa and Google Assistant have demonstrated the effectiveness of this approach, they are generally limited to predefined functionalities and do not support dynamic interaction with arbitrary website content.

This project proposes an AI Voice Agent for Websites that integrates web scraping, semantic search, vector databases, and generative AI models to enable intelligent interaction with website content. The system allows users to query website information using text or voice and receive accurate, context-aware responses. By combining Retrieval-Augmented Generation with voice interaction, the proposed system transforms static websites into dynamic conversational platforms, improving accessibility, efficiency, and overall user experience.

2. LITERATURE SURVEY

The development of intelligent

conversational systems and information retrieval techniques has gained significant attention in recent years due to the increasing demand for efficient interaction with digital platforms. Traditional chatbot systems were primarily rule-based, where predefined responses were mapped to specific user queries. Although these systems were effective for handling simple and repetitive tasks, they lacked the ability to understand complex queries and adapt to dynamic data sources. As a result, their performance was limited in real-world applications.

With the advancement of machine learning and deep learning techniques, more sophisticated conversational models have been developed. Transformer-based models and large language models (LLMs) have demonstrated strong capabilities in natural language understanding and generation. These models can produce human-like responses and handle a wide range of queries. However, despite their effectiveness, they often generate responses without grounding in actual data sources, leading to inaccuracies and hallucinations. This limitation reduces their reliability in applications that require precise and domain-specific information.

To overcome these challenges, Retrieval-Augmented Generation (RAG) has been introduced as a hybrid approach that combines semantic search with generative models. In this method, relevant information is first retrieved from external knowledge sources and then used to generate responses. This significantly improves the accuracy and contextual relevance of the output. Vector databases have further enhanced this approach by enabling efficient similarity search using high-dimensional embeddings, making them suitable for real-time conversational systems.

In addition to text-based interaction, voice-based systems have also been widely explored. Technologies such as Speech-to-

Text (STT) and Text-to-Speech (TTS) have enabled the development of voice assistants that allow users to interact with systems using natural speech. Popular voice assistants demonstrate the effectiveness of this approach in improving accessibility and user convenience. However, most of these systems are limited to predefined domains and do not support dynamic interaction with arbitrary or real-time web content.

Recent research has also focused on web data extraction and automation techniques to build domain-specific knowledge systems. Web scraping tools and automated data processing pipelines enable the collection of large amounts of information from websites. While these approaches provide dynamic data access, they are often not integrated with conversational AI systems, limiting their ability to provide interactive responses.

Despite these advancements, there is still a lack of a unified system that combines web data extraction, semantic retrieval, generative AI, and voice interaction into a single scalable solution. Existing systems either focus on conversational AI without real-time data grounding or on data retrieval without natural language interaction. This project addresses these limitations by developing an integrated AI Voice Agent for Websites that combines Retrieval-Augmented Generation with voice-based interaction to provide accurate and dynamic communication with website content.

3. PROPOSED SYSTEM

The proposed system is an AI Voice Agent for Websites designed to transform static web platforms into intelligent conversational systems. The system enables users to interact with website content using natural language in both text and voice formats. It utilizes Retrieval-Augmented Generation (RAG) to provide accurate, context-aware responses by combining semantic search with generative AI models. The overall system is designed to be scalable, efficient, and capable of handling

multiple websites through a modular architecture.

The system begins with the data acquisition process, where website content is extracted using automated web scraping techniques. The extracted data includes textual information such as headings, paragraphs, and relevant content from the website. This data is then cleaned and processed to remove unnecessary elements like navigation menus, scripts, and redundant HTML tags, ensuring that only meaningful information is retained.

After preprocessing, the content is divided into smaller segments using text chunking techniques. This step helps in maintaining semantic context while enabling efficient processing of large amounts of data. Each chunk is then converted into a numerical representation using embedding models. These embeddings capture the semantic meaning of the content and are stored in a vector database, which allows fast and efficient similarity-based retrieval.

When a user interacts with the system by entering a query or speaking through the voice interface, the input is first converted into text (in case of voice input) using Speech-to-Text technology. The query is then transformed into an embedding and compared with the stored vectors in the database to retrieve the most relevant content. This retrieved information is passed to a generative language model, which produces a meaningful and context-aware response.

The generated response is then converted into speech using Text-to-Speech technology, enabling the system to communicate with the user in a natural and interactive manner. This creates a complete conversational loop, where users can continuously interact with the system using voice or text.

An important feature of the proposed system

is its multi-agent capability. The system allows users to create separate AI agents for different websites or knowledge bases. Each agent maintains its own data and embeddings, ensuring accurate and domain-specific responses without interference from other sources. This modular design improves scalability and makes the system suitable for real-world applications.

The proposed system incorporates a robust Retrieval-Augmented Generation (RAG) pipeline to ensure accurate and context-aware responses. When a user submits a query, the input is first converted into a vector embedding using an embedding model. This query embedding is then compared with precomputed document embeddings stored in a Qdrant vector database using semantic similarity search. The system retrieves the top relevant content chunks based on similarity scores and a predefined threshold, ensuring that only meaningful and contextually aligned information is selected. These retrieved chunks are then assembled into a structured context and passed to a GPT-based language model, which generates a coherent and grounded response.

To efficiently handle large-scale data and concurrent user interactions, the system utilizes an asynchronous queue architecture built with BullMQ and Redis. The architecture separates chat processing and agent ingestion workflows into independent queues. The chat queue is configured to handle multiple user requests concurrently with rate limiting to ensure system stability, while the agent ingestion queue manages resource-intensive tasks such as web scraping and embedding generation. This design enables the system to scale effectively while maintaining low latency and high throughput.

The system also includes a Puppeteer-based website crawler that automates content extraction from user-provided URLs. The crawler operates within domain constraints,

filtering out irrelevant or restricted pages such as login screens, media files, and administrative routes. It systematically navigates through multiple pages of a website and extracts meaningful textual content, which is then cleaned and processed for further analysis. This ensures that the knowledge base of each AI agent is both relevant and high-quality.

For semantic processing, the extracted content is divided into chunks of optimal size with overlapping regions to preserve contextual continuity. These chunks are embedded using a pre-trained embedding model and stored in the vector database, enabling efficient retrieval during query processing. Additionally, the system maintains conversational context using a memory mechanism, allowing it to retain previous interactions and provide more coherent and personalized responses across sessions.

The overall system is built using a modern full-stack architecture, including Node.js and Express for backend services, Next.js for the frontend interface, MongoDB for structured data storage, and Qdrant for vector-based semantic search. Redis and BullMQ handle asynchronous processing, while Docker ensures consistent deployment across environments. This integrated architecture enables the system to deliver a scalable, efficient, and real-time conversational experience for website interaction.



Figure 1: Retrieval Augmented Generation (RAG) Architecture

4. RESULT DESCRIPTION

The proposed AI Voice Agent for Websites provides an interactive platform that enables users to communicate with website content

using both text and voice. The system is designed with a user-friendly interface that allows seamless interaction and efficient information retrieval. The main functionalities of the system include website data processing, conversational query handling, and voice-based response generation.

The interface of the system includes a dashboard where users can create and manage multiple AI agents for different websites. Each agent represents a specific website or knowledge base and allows users to interact with its content through a chat interface. The chat interface is designed similar to a messaging platform, where users can input queries and receive responses in real time. The system processes user queries using the Retrieval-Augmented Generation pipeline and provides accurate and context-aware answers based on the extracted website data.

The voice interaction feature enables users to communicate with the system using speech. When the user activates the microphone, the system converts voice input into text using Speech-to-Text technology. The processed query is then sent to the AI model, which generates a response based on the retrieved information. The response is then converted into audio using Text-to-Speech technology, allowing the system to provide spoken output. This feature enhances accessibility and provides a more natural interaction experience.

The system also demonstrates efficient data retrieval and response generation. The use of vector databases enables fast similarity search, ensuring that relevant information is retrieved quickly. The integration of generative AI models ensures that responses are coherent and contextually accurate. The system is capable of handling multiple queries efficiently and provides real-time responses without significant delay.

Overall, the results show that the proposed

system successfully transforms static websites into dynamic conversational platforms. The integration of text-based and voice-based interaction improves user accessibility, reduces navigation effort, and enhances the overall user experience.

Additionally, the system demonstrates strong performance in handling real-time conversational queries with minimal latency. The integration of the Retrieval-Augmented Generation pipeline ensures that responses are not only relevant but also grounded in the actual content of the website, thereby improving reliability and reducing incorrect outputs. The use of vector-based similarity search allows the system to efficiently retrieve contextually related information even for complex or indirectly phrased queries.

The multi-agent functionality further enhances the system by enabling the creation of separate agents for different websites, allowing users to switch between domains without losing contextual accuracy. This ensures that each agent maintains its own knowledge base and provides domain-specific responses

Moreover, the voice interaction feature significantly improves user accessibility by enabling hands-free communication, which is particularly beneficial for users with accessibility needs or those interacting in dynamic environments. The seamless transition between voice input and audio output creates a natural conversational experience, similar to interacting with a human assistant.

Overall, the experimental observations indicate that the system is robust, scalable, and capable of providing an efficient and user-friendly solution for intelligent website interaction. The combination of semantic retrieval, generative AI, and voice technology makes the system highly effective in transforming traditional web interfaces into interactive conversational platforms.

Final Observation

The proposed AI Voice Agent for Websites successfully demonstrates the effectiveness of integrating Retrieval-Augmented Generation, semantic search, and voice-based interaction into a unified system. By transforming static website content into an interactive conversational interface, the system significantly improves the way users access and interact with information. The use of vector databases enables efficient and accurate retrieval of relevant data, while the generative AI model ensures that responses are context-aware and meaningful.

The integration of Speech-to-Text and Text-to-Speech technologies enhances user accessibility by enabling natural voice-based communication, making the system more intuitive and user-friendly. The multi-agent architecture further strengthens the system by allowing independent agents for different websites, ensuring scalability and domain-specific accuracy.

Overall, the system reduces the need for manual navigation, minimizes user effort, and provides faster access to relevant information. The proposed solution offers a practical and scalable approach for modern web interaction and highlights the potential of combining AI-driven retrieval systems with conversational and voice technologies to build intelligent and accessible digital platforms.

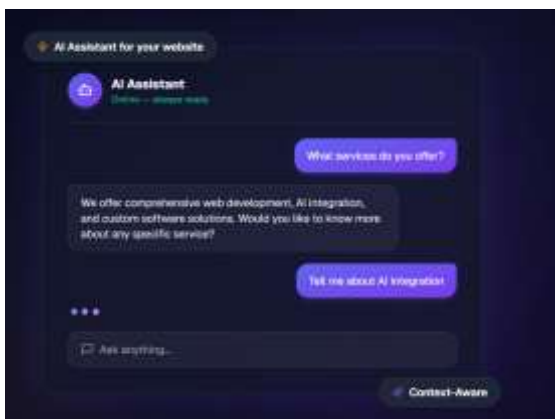


Figure 2 Dashboard Interface of AI Voice Agent System (Source: Available at:

<https://ask-agent.saishankar.me/>)

5. CONCLUSION

This project successfully demonstrates the design and implementation of an AI Voice Agent for Websites that enables intelligent and dynamic interaction with web-based content. By integrating Retrieval-Augmented Generation (RAG) with semantic search and vector databases, the system provides accurate and context-aware responses based on real-time website data. This approach effectively overcomes the limitations of traditional navigation methods and rule-based chatbot systems.

The incorporation of Speech-to-Text and Text-to-Speech technologies further enhances the system by enabling natural voice-based interaction, making it more accessible and user-friendly. The multi-agent architecture allows the system to support multiple websites independently, ensuring scalability and domain-specific response accuracy. The overall system provides a seamless conversational experience, reducing user effort and improving efficiency in accessing information.

The results of the system demonstrate that combining web scraping, semantic processing, generative AI, and voice interaction can significantly enhance user engagement and accessibility. The proposed solution provides a practical and scalable framework for transforming static websites into intelligent conversational platforms, paving the way for more advanced AI-driven web interaction systems in the future.

6. REFERENCES

[1]. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information*

Processing Systems.

[2]. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems.*

[3]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention is All You Need. Advances in Neural Information Processing Systems.*

[4]. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). *Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP.*

[5]. Johnson, J., Douze, M., & Jégou, H. (2019). *Billion-scale Similarity Search with GPUs. IEEE Transactions on Big Data.*

[6]. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of EMNLP.*

[7]. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL.*

[8]. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training. OpenAI.*

[9]. OpenAI. (2024). *GPT Models and API Documentation. Available: <https://platform.openai.com>*

[10]. LangChain. (2024). *LangChain Framework Documentation. Available: <https://www.langchain.com>*

[11]. Qdrant. (2024). *Vector Database for Semantic Search. Available: <https://qdrant.tech>*

[12]. Firecrawl. (2024). *Web Data Extraction and Crawling Tool. Available: <https://www.firecrawl.dev>*

[13]. Mozilla. (2023). *Web Speech API for Speech Recognition and Synthesis. Available: <https://developer.mozilla.org>*

[14]. Amazon Web Services. (2023). *Text-to-Speech and Speech Recognition Systems. Available: <https://aws.amazon.com>*

[15]. Google. (2023). *Speech-to-Text and Text-to-Speech APIs. Available: <https://cloud.google.com>*