SMARTSPEAK: AI-VOICE ASSISTANT

BACHELOR OF ENGINEERING
IN INFORMATION TECHNOLOGY

By

1. **Vishal Anil Singh**
2. **Aniket Govind Varma**
3. **Dipanshu Ramkishan Bandoliya**

Under the Guidance Of

Mrs. Rupali Pashte

Professor, Department of Information Technology

Shree Babai Education Society's (Regd.)

SHREE L. R. TIWARI
COLLEGE OF ENGINEERING

(Approved by AICTE & DTE, Maharashtra State & Affiliated to University of Mumbai
& NAAC Accredited, NBA Accredited Program, ISO 9001:2015 Certified ISTE Excellence Award
Minority Status (Hindi Language))

Department of Information Technology

2023-2024

SMARTSPEAK: AI-VOICE ASSISTANT

BACHELOR OF ENGINEERING
IN INFORMATION TECHNOLOGY

By

1. Vishal Anil Singh

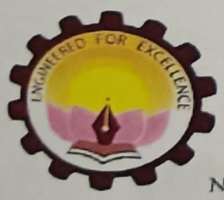
2. Aniket Govind Varma

3. Dipanshu Ramkishan Bandoliya

Under the Guidance Of

Mrs. Rupali Pashte

Professor, Department of Information Technology



Shree Rahul Education Society's (Regd.)

SHREE L. R. TIWARI
COLLEGE OF ENGINEERING

(Approved by AICTE & DTE, Maharashtra State & Affiliated to University of Mumbai)
NAAC Accredited, NBA Accredited Program, ISO 9001:2015 Certified | DTE Code No. : 3423
Minority Status (Hindi Linguistic)

Department of Information Technology
2023-2024

PROJECT REPORT ON

SMARTSPEAK: AI-VOICE ASSISTANT

Submitted to Mumbai University

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN INFORMATION TECHNOLOGY

By

1. Vishal Anil Singh

2. Aniket Govind Varma

3. Dipanshu Ramkishan Bandoliya

Under the Guidance Of

Mrs. Rupali Pashte

Professor, Department of Information Technology

Shree Rahul Education Society's (Regd.)

SHREE L. R. TIWARI

COLLEGE OF ENGINEERING

(Approved by AICTE & DTE, Maharashtra State & Affiliated to University of Mumbai)

NAAC Accredited, NBA Accredited Program, ISO 9001:2015 Certified | DTE Code No. : 3423

Minority Status (Hindi Linguistic)

Department of Information Technology

2023-2024

Shree Rahul Education Society's (Regd.)
SHREE L. R. TIWARI
COLLEGE OF ENGINEERING
(Approved by AICTE & DTE, Maharashtra State & Affiliated to University of Mumbai)
NAAC Accredited, NBA Accredited Program, ISO 9001:2015 Certified | DTE Code No. : 3423
Minority Status (Hindi Linguistic)

UNIVERSITY OF MUMBAI

CERTIFICATE

This is to certify that the project titled "*SMARTSPEAK: AI-VOICE ASSISTANT*" has been completed under our supervision and guidance by the following students:

1. **Vishal Anil Singh**
2. **Aniket Govind Varma**
3. **Dipanshu Ramkishan Bandoliya**

In the partial fulfillment of degree of Bachelor of Engineering in Information Technology branch as prescribed by the University of Mumbai during the academic year 2023-2024. The said work has been assessed and is found to be satisfactory.

Signature of the Internal Examiner
Name: Rupali Pashte
Date: 26/4/24

Signature of the External Examiner
Name: Dr. Tasneem Mirza
Date: 26/4/24

Signature of the H.O.D.
Name: Dr. Roopali Lolage
Date: _____

Signature of the Coordinator
Name: Mrs. Rupali Pashte
Date: 26/4/24

Signature of the Principal
Name: Dr. Deven N. Shah
Date: _____

DECLARATION

We do hereby declare that the work embodied in the project entitled *SMARTSPEAK: AI-VOICE ASSISTANT* is the outcome of our original work under the guidance and supervision of **Mrs. Rupali Pashte**. This piece of work or any part of it has not been submitted previously for the award of any other degree, diploma, or other title to any other institution.

We also declare that this written submission represents our ideas in our own words and where other ideas or words have been included. We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Vishal Anil Singh Vishal. Date: 28/04/24

Roll No.: 52

Exam. Seat No.:

Aniket Govind Varma Varma Aniket

Roll No.: 56

Exam. Seat No.:

Dipanshu Ramkishan Bandoliya Dipanshu

Roll No.: 01

Exam. Seat No.:

ACKNOWLEDGEMENT

A few sublime human experiences defy expressions of any kind, and a feeling of true gratitude is one of them. I, therefore, find words quite inadequate to express my indebtedness to my Guide **Mrs. Rupali Pashte for** her virtuous guidance, encouragement and help throughout this work. Their deep insight into the problem and the ability to provide solutions has been of immense value in improving the quality of projects at all stages. This experience of working with them shall forever remain a source of inspiration and encouragement for me.

I express my thanks to **Dr. Roopali Lolage**, HOD (IT), SLRTCE, Mira Road, for extending her support that she gave, which truly helped the progression of the project work. I express my thanks to **Mrs. Rupali Pashte**, Asst. Prof., SLRTCE, Mira Road for extending her support.

My sincere thanks to **Dr. Deven N. Shah** Principal, SLRTCE, Mira Road for providing me the necessary administrative assistance in the completion of the work.

I am extremely grateful to the celebrated authors whose precious works have been consulted and referred to in my project work. I also wish to convey my appreciation to my friends who provided encouragement and timely support in the hour of need.

Special thanks to my Parents whose love and affectionate blessings have been a constant source of inspiration in making this a reality.

All the thanks are, however, only a fraction of what is due to the Almighty for granting me an opportunity and the divine grace to successfully accomplish this assignment.

Vishal Anil Singh Vishal.

Aniket Govind Varma Varma Aniket

Dipanshu Ramkishan Bandoliya Dipanshu

TABLE OF CONTENTS

ABSTRACT

In this world Artificial Intelligence is advancing rapidly, the integration of voice assistants with cutting-edge language and image processing models has garnered significant attention. Chat-GPT's advanced natural language understanding empowers the voice assistant to improve conversational abilities and expand its knowledge base. Meanwhile, DALL-E's exceptional image synthesis capabilities enrich the user experience by generating relevant images in response to text-based queries. The proposed architecture seamlessly combines both models, enabling the voice assistant to interpret voice commands using Chat-GPT and utilize DALL-E for visual information, fostering interactive dialogues. Ethical, privacy, and computational concerns are thoroughly addressed, and extensive experiments and user studies demonstrate the superiority of this integration over traditional voice assistants. The results showcase significant enhancements in capabilities, leading to more engaging and contextually-aware voice assistants, setting a path for smarter voice-based applications in the future.

Keywords: - dall-e, chat-gpt, fostering, visual information, voice assistant

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

1.1 DESCRIPTION

The project at hand involves the development of a computer assistant using the ChatGPT and DALL-E technologies. Presently, the assistant operates exclusively through text and voice input; users need to type or speak their queries and requests. Our innovative system brings together state-of-the-art language processing and image generation technologies. Unlike conventional voice assistants, it excels in both understanding and responding to natural language while also generating relevant visual content. This ensures more engaging and informative conversations, offering users a comprehensive and immersive experience.

1.2 PROBLEM FORMULATION

In today's fast-paced digital world, businesses and organizations face the challenge of providing efficient and personalized customer support, as well as engaging with users effectively through various online platforms. Traditional search methods may no longer align with the evolving expectations of modern users. To tackle this issue, there is a pressing need for an AI-powered chatbot that can promptly, accurately, and personally address user inquiries. This chatbot should improve customer satisfaction, reduce response times, and provide valuable insights into user interactions for the benefit of the organization.

1.3 MOTIVATION

The motivation for this project stems from several key factors in the contemporary digital landscape.. The ever-evolving expectations of modern users represent a significant catalyst; traditional search methods are no longer sufficient to meet these demands. Therefore, there is a strong impetus to develop an AI-powered chatbot that can quickly and accurately address user inquiries while maintaining a personalized touch.

Furthermore, at the heart of this project's motivation is the aspiration to enhance the overall user experience. By enabling the chatbot to understand and respond to natural language, translate between languages, and offer accessibility features for those with visual impairments, the aim is to make digital interactions more user-friendly and inclusive. Additionally, the project is forward-looking in its approach, seeking to create an adaptable AI solution that can evolve to meet the ever-growing demands of users in the digital age. This motivation ultimately drives the project's goals of providing high-quality customer support, embracing evolving user expectations, and enhancing the digital user experience.

1.4 PROPOSED SOLUTION

The project's proposed solution involves the development of a versatile and advanced AI chatbot that integrates the capabilities of the ChatGPT and DALL-E algorithms. This chatbot is designed to provide a wide range of functionalities, including real-time customer support, information retrieval, language translation, accessibility features, user feedback collection, and educational content delivery. It will enable users to interact with the chatbot through voice commands, improving the user experience by making digital interactions more natural and accessible.

Key elements of the proposed solution include:

1. Implementing voice recognition for user interaction, allowing users to communicate with the chatbot using spoken commands.
2. Incorporating natural language processing (NLP) to accurately understand and interpret user commands, including intent recognition and entity extraction.
3. Utilizing a text-to-speech (TTS) system to convert generated text responses into spoken words for more natural interactions.
4. Expanding language support to make the chatbot more inclusive and accessible to a global audience.
5. Addressing the needs of users with visual impairments by providing features like reading text aloud, describing images, and facilitating website navigation.

1.5 SCOPE

The project aims to provide support for multiple languages, catering to a diverse user base. Additionally, it focuses on enhancing accessibility through text-based features, such as screen readers and alternative text for images. Gathering user feedback and conducting surveys through text-based interactions is an essential component to continually improve the assistant's functionality. Educational content and tutoring services are also part of the project's scope, ensuring its versatility as a valuable tool for learning and information retrieval. Future developments might involve expanding language support, implementing sentiment analysis for text interactions, and integrating user feedback for continuous enhancements. In summary, this project, with voice interaction, seeks to deliver a robust and adaptable AI chatbot with various features that aim to improve accessibility and user experiences.

CHAPTER 2 : REVIEW OF LITERATURE

2.1 LITERATURE SURVEY

Literature Survey on Voice Recognition and AI-Based Voice Assistants:

Voice recognition and AI-based voice assistants have gained significant attention in recent years. These technologies enable natural language interactions with devices and offer several advantages, including enhanced accessibility and integration into daily life. Subhash S and Prajwal N Srivatsa's paper from 2020 explores an AI-based voice assistant that utilizes the gTTS engine for converting text to speech. This approach achieves high accuracy, particularly when searching for information on Google. However, language limitations are noted, as the voice assistant primarily operates in English. The authors emphasize the need for multilingual support to make the technology more inclusive and accessible to diverse user populations.

Literature Survey on Voice Command Recognition:

The paper by Dipankar Gupta, Emam Hossain, and Mohammad Shahadat Hossain in 2019 focuses on voice command recognition, particularly in the context of Bangla language. The research employs cross-correlation techniques and a hybrid approach combining cross-correlation with LPC. The results show a recognition rate of approximately 80.34% within response times of 2-3 seconds. The study highlights the potential of voice command recognition systems to provide efficient user interfaces and assist individuals with mobility limitations through facial movement control of the mouse cursor.

Literature Survey on Augmented Reality in Education:

The paper by Mouna Kenoui and Mohamed Ait Mehdi discusses the development of an augmented reality (AR) system for educational purposes. AR technology offers a novel way to enhance the learning experience by visualizing complex concepts, such as DNA molecules, in an interactive and engaging manner. By utilizing ARToolKit SDK and Unity3D, the authors illustrate how AR can make educational content more immersive and understandable. The integration of IBM Watson services further enhances the naturalness of voice interactions. However, there are challenges, such as dependency on cloud services and limitations in acquiring 3D objects for educational content.

2.2 COMPARISON TABLE

Author Name	Title of paper	Method/ Algorithm used	Result Accuracy	Advantage	limitation	Future Scope
Subhash S (Department of Telecommunication Engineering) , Prajwal N Srivatsa Department of Telecommunication Engineering	Artificial Intelligence-based Voice Assistant	In this paper the author installed gTTS engine package to make the voice assistant speak like a normal human being, gTTS is basically used to convert the audio string into text. This audio string is nothing but the response which the voice assistant is supposed to give to the user.	The model takes voice command as input And gives very accurate answers i.e results, by searching input on google search engine. which achieves an accuracy of over 90% while giving a 10% error in certain conditions.	**Integration with Daily Life**: The voice assistant can interact with devices such as smartphones and laptops, integrating seamlessly into users' daily routines. **Accessibility**: The voice assistant makes technology more accessible to people with disabilities or those who have difficulty using traditional interfaces.	**Speech Recognition Accuracy:** The effectiveness of the voice assistant heavily depends on accurate speech recognition. Misinterpretations or inaccuracies in recognizing voice commands could lead to incorrect actions. **Language Limitations:** The text indicates that the voice assistant converts text to speech in English. This could limit its accessibility and usability for users who speak languages other than English.	**Multilingual Support:** Expanding the language capabilities beyond English could make the voice assistant more inclusive and accessible to a wider global audience. **Enhanced Personalization:** Future iterations of the voice assistant might learn user preferences and habits over time, offering more personalized and relevant responses.

Dipankar Gupta , Emam Hossain , Mohammad Shahadat Hossain	A Digital Personal Assistant using Bangla Voice Command Recognition and Face Detection	The primary method used for Bangla voice recognition is the cross-correlation technique. Cross-correlation is employed to compare the energy of Bangla voice commands with pre-recorded reference signals. The similarity between the user's input voice command and the reference signals is measured using cross-correlation. This method helps in recognizing the voice command and executing the corresponding action. Another hybrid technique mentioned in the paper combines cross-correlation with LPC for voice recognition. LPC is a technique used to represent the spectral envelope of a digital signal of speech in compressed form.	The paper also mentions that the system recognized, on average, about 80.34% of the commands correctly across all three environments. The system was able to achieve these recognition rates within response times of 2-3 seconds.	**Facial Movement Control:** The system incorporates a feature that allows users to control the mouse cursor using facial movements. This can be particularly beneficial for individuals with mobility limitations. **Comparison with Existing Models:** The paper compares the proposed model with two other techniques (MFCC & DTW, and cross-correlation with LPC). The results indicate that the proposed model achieves higher accuracy and faster response times.	**Limited Vocabulary:** The system is designed to recognize a specific set of 12 Bangla voice commands. While this is a good starting point, it may not cover the full range of commands that users might want to issue. **Tracking Accuracy in Face Detection:** The effectiveness of mouse cursor control via facial movements may vary based on factors like lighting conditions, user movements, and steady head positioning. It's noted that the system may experience some limitations under these circumstances.	**Integration of User Feedback:** Implementing a feedback mechanism to confirm that a command has been correctly recognized would enhance user confidence and improve the overall usability of the system. **Integration with Smart Home and IoT Devices:** Expanding the capabilities of the digital assistant to control smart home devices or interact with Internet of Things (IoT) technologies would further extend its utility in daily life.

Mouna Kenoui, Mohamed Ait Mehdi	Teach-Me DNA : an Interactive Course Using Voice Output in an Augmented Reality System	The paper mentions the development of an augmented reality (AR) system using Unity3D and ARToolKit SDK. While not explicitly mentioned in the paper, AR development often involves computer vision and tracking algorithms to overlay virtual objects onto the real world. SSML is mentioned in the paper as a language used to control the features of the synthesized speech. SSML is an XML-based markup language used to fine-tune speech synthesis, such as controlling pitch, rate, and pronunciation.	The paper provided does not explicitly mention numerical results or quantitative accuracy metrics associated with the implementation of the AR system and voice interaction. Therefore, there is no information available in the paper regarding the accuracy of the system in terms of recognition of voice commands or the quality of the synthesized speech.	**Enhanced Learning Experience:** The system aims to enhance the learning experience by using AR technology, making educational content more engaging and interactive. AR allows learners to visualize and interact with 3D objects, which can improve their understanding of complex concepts, such as the DNA molecule. **Use of IBM Watson Services:** By leveraging IBM Watson services for text-to-speech and speech-to-text conversion, the system benefits from advanced natural language processing capabilities	**Dependency on Cloud Services:** The system relies on cloud-based services, particularly IBM Watson, for speech recognition and synthesis. This means that an internet connection is required for the system to function, which may not be available in all educational settings or could lead to disruptions in usage. **Limited 3D Object Resources:** The paper mentions limitations in acquiring high-quality 3D objects for educational content.	**Offline Mode:** Addressing the dependency on cloud services by developing an offline mode would increase accessibility, especially in areas with limited internet connectivity. **Collaborative Learning:** Extending the system to support collaborative learning environments, where multiple users can interact simultaneously in AR, can foster peer learning and teamwork. **Feedback Mechanisms:** Incorporating mechanisms for students to provide feedback

Table no. 2.2 (Literature Survey on the Research Paper)

2.3 PROBLEM STATEMENT

In today's fast-paced digital world, businesses and organizations face the challenge of providing efficient and personalized customer support, as well as engaging with users effectively through various online platforms. The objective is to create a highly functional and adaptable AI chatbot that can improve customer satisfaction, reduce response times, and provide valuable insights into user interactions for the benefit of the organization. Traditional search methods may no longer align with the evolving expectations of modern users. To tackle this issue, there is a pressing need for an AI-powered chatbot that can promptly, accurately, and personally address user inquiries. Such a chatbot would consistently deliver high-quality customer support and elevate user engagement to meet the ever-growing demands of today's digital world.

CHAPTER 3 : SYSTEM ANALYSIS AND DESIGN

3.1 REQUIREMENTS

HARDWARE AND SOFTWARE REQUIREMENTS

PROCESSOR	Intel Celeron or HIgher
RAM	2GB or More
HARDISK	256GB STORAGE
FRONT END	HTML,CSS AND JAVASCRIPT
BACK END	dall-e ,chatgpt api
OS	WINDOWS 7

Table no. 3.1 (Software and Hardware Requirements)

3.2 USE CASE DIAGRAM

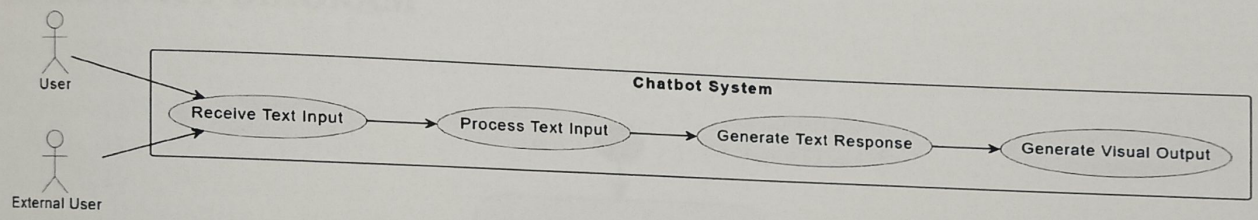


Fig 3.2 USE CASE DIAGRAM

3.3 CLASS DIAGRAM

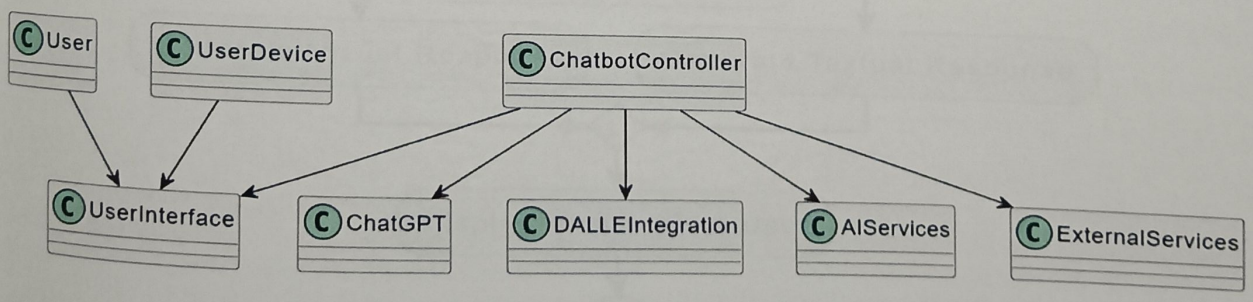


Fig 3.3 CLASS DIAGRAM

3.4 SEQUENCE DIAGRAM

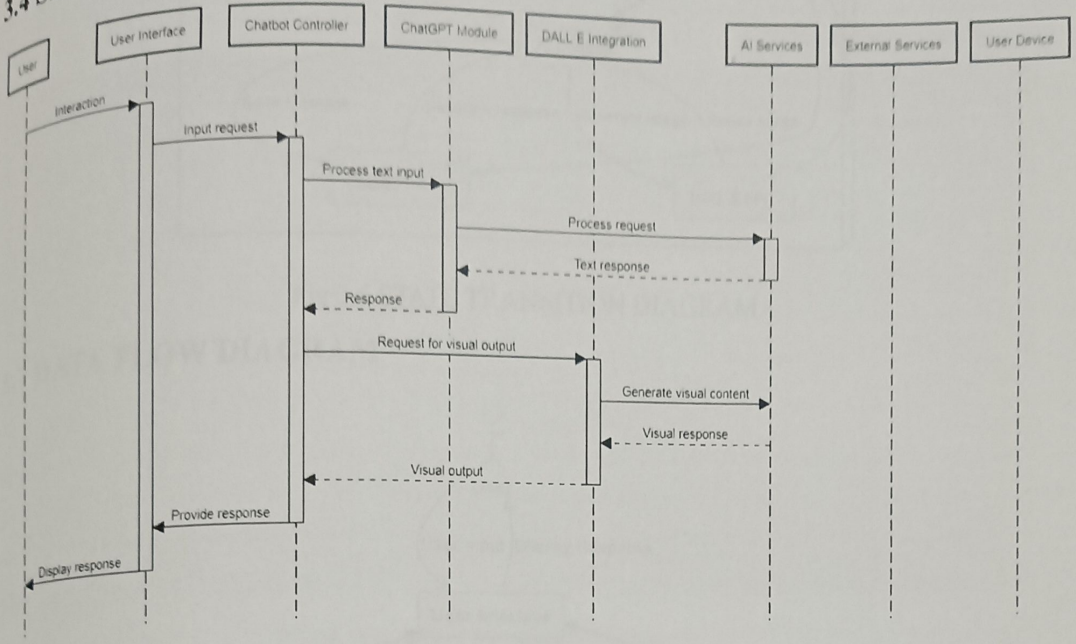


Fig 3.4 SEQUENCE DIAGRAM

3.5 ACTIVITY DIAGRAM

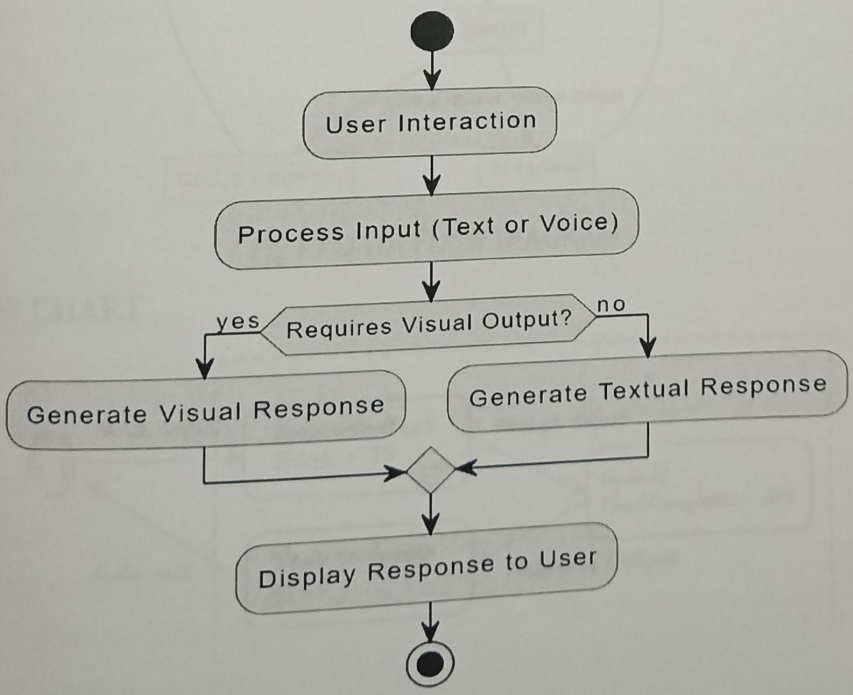


Fig 3.5 ACTIVITY DIAGRAM

3.6 STATE TRANSITION DIAGRAM

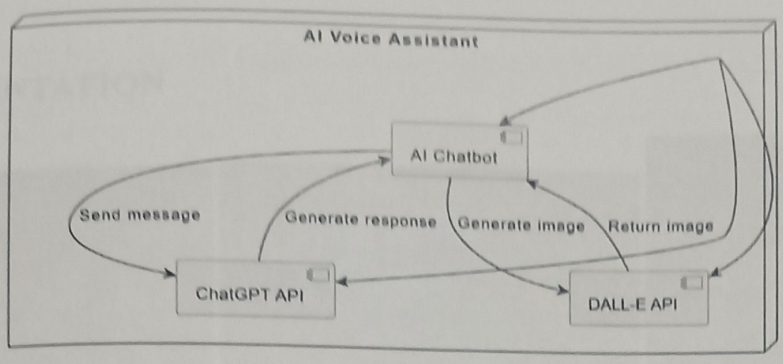


Fig 3.6 STATE TRANSITION DIAGRAM

3.7 DATA FLOW DIAGRAM

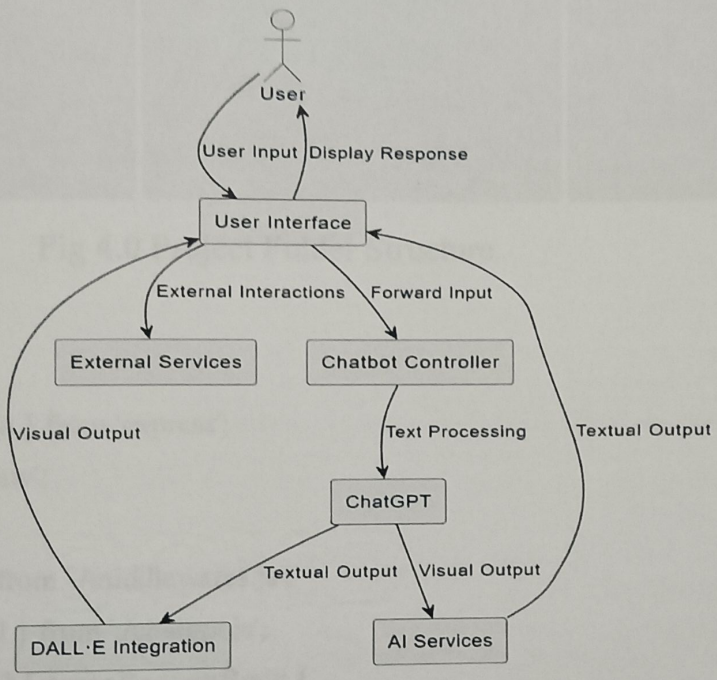


Fig 3.7 DATA FLOW DIAGRAM

3.8 FLOW CHART

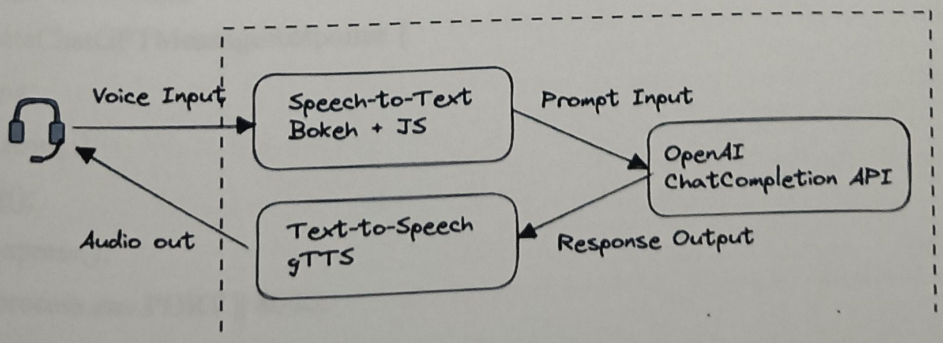


Fig 3.8 FLOW CHART

CHAPTER 4 : IMPLEMENTATION DETAILS

4.1 IMPLEMENTATION

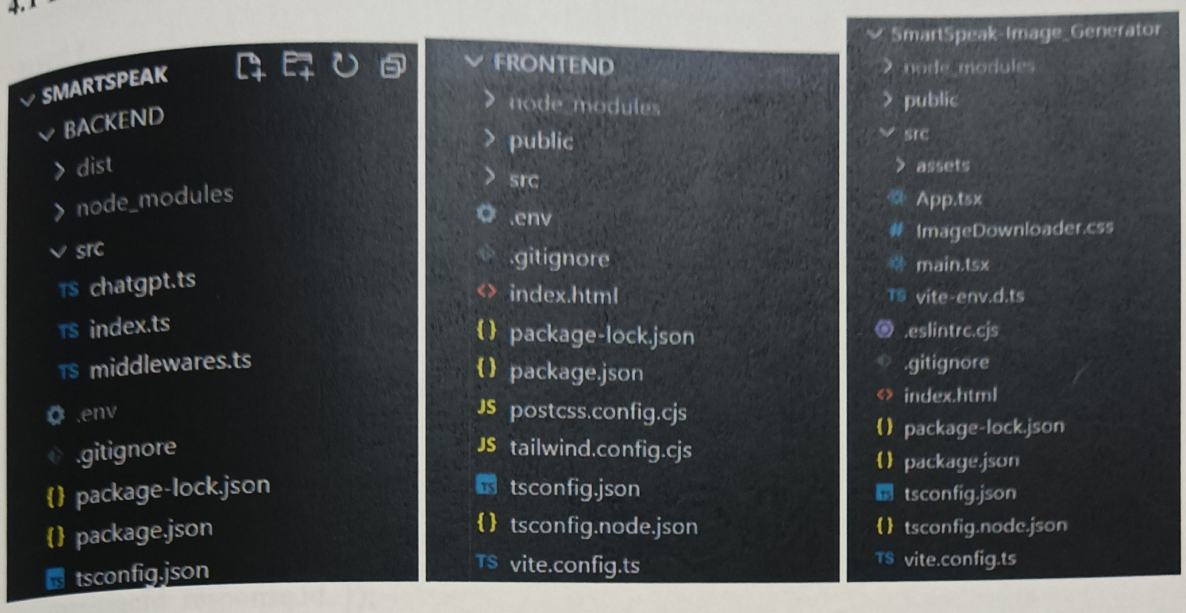


Fig 4.0 Project Folder Structure

CODE:

BACKEND:

```
import express, { Request } from 'express';
import dotenv from 'dotenv';
import cors from 'cors';
import { errorHandler } from './middlewares.js';
import { getChatGPTAPI } from './chatgpt.js';
interface CreateChatGPTMessageRequestBody {
 text: string;
 parentMessageId?: string;}
interface CreateChatGPTMessageResponse {
 answer: string;
 messageId: string;}
dotenv.config();
const app = express();
const port = process.env.PORT || 8000;
const api = await getChatGPTAPI();
app.use(cors());
app.use(express.json());
app.use(errorHandler);
```

```
app.get('/', (req, res) => {
 res.send('ChatGPT server is running OK');});
app.post(
 '/chatgpt/messages',
 async (
  req: Request<{},
   CreateChatGPTMessageResponse,
   CreateChatGPTMessageRequestBody>,
  res,next, ) => {
  const { text, parentMessageId } = req.body;
 try {
   const response = await api.sendMessage(text, {
    parentMessageId,});
   res.json({
    answer: response.text,
    messageId: response.id, });
  } catch (error: unknown) {
   if (error instanceof Error) {
    return next(error)  }
   const errorMessage = typeof error === 'string' ? error : 'Something went wrong';
   return next(new Error(errorMessage)); } },);
app.listen(port, () => {
 console.log(`[server]: Server is running at http://localhost:${port}`);});
```

FRONTEND:

```
// Keep this import at the top of this file
// for 'react-speech-recognition' to work.
import 'regenerator-runtime/runtime';
import React from 'react'
import ReactDOM from 'react-dom/client'
import App from './App'
import './index.css'
ReactDOM.createRoot(document.getElementById('root') as HTMLElement).render(
 <React.StrictMode>
  <App />
 </React.StrictMode>,)
```

IMAGE GENERATOR CODE:

```
import { useState } from 'react';
import './ImageDownloader.css'; // Import CSS file for styling
const ImageDownloader = () => {
const [prompt, setPrompt] = useState('a dog in a space');
const [width, setWidth] = useState(480);
const [height, setHeight] = useState(480);
const [seed, setSeed] = useState(42);
const [model, setModel] = useState('model'); // Define your model here
const [showImage, setShowImage] = useState(true); // State to control image visibility
// const handleButtonClick = () => {
//  setShowImage(true); // Show the image when the button is clicked
// };
// Construct the image URL using the state values
 const imageUrl =
`https://pollinations.ai/p/${encodeURIComponent(prompt)}?width=${width}&height=${height}&se
ed=${seed}&model=${model}`;
 return (
  <div className="image-downloader-container">
   <h2 className="animated fadeInDown">SmartSpeak Image Generator</h2>
   <div className="form-group animated fadeInLeft">
    <label htmlFor="promptInput">Prompt:</label>
    <input
     type="text"
     id="promptInput"
     value={prompt}
     onChange={(e) => setPrompt(e.target.value)}
    />
   </div>
   <div className="form-group animated fadeInRight">
    <label htmlFor="widthInput">Width:</label>
    <input
     type="number"
     id="widthInput"
     value={width}
     onChange={(e) => setWidth(parseInt(e.target.value))}
    />
```

```
    </div>
    <div className="form-group animated fadeInLeft">
     <label htmlFor="heightInput">Height:</label>
     <input
      type="number"
      id="heightInput"
      value={height}
      onChange={(e) => setHeight(parseInt(e.target.value))}
     />
    </div>
    <div className="form-group animated fadeInRight">
     <label htmlFor="seedInput">Seed:</label>
     <input
      type="number"
      id="seedInput"
      value={seed}
      onChange={(e) => setSeed(parseInt(e.target.value))}
     />
    </div>
    <div className="form-group animated fadeInLeft">
     <label htmlFor="modelInput">Model:</label>
     <input
      type="text"
      id="modelInput"
      value={model}
      onChange={(e) => setModel(e.target.value)}
     />
    </div>
    {/* <button className="submit-button animated fadeInUp"
onClick={handleButtonClick}>Generate Image</button> */}
    <div className={`image-container animated ${showImage ? 'fadeIn' : 'fadeOut'}`}>
     {showImage && <img className="image-preview" src={imageUrl} alt="Generated Image"
/>}
    </div></div>);};
export default ImageDownloader;
```

CHAPTER 5 : EVALUATION AND RESULT

5.1 EVALUATION

Innovation and Novelty: Integrating Chat-GPT and DALL-E into a voice assistant system represents an innovative approach. While voice assistants have become commonplace, enhancing them with state-of-the-art language understanding and image synthesis technologies adds novelty to the project.

Technological Advancement: The project demonstrates advancements in natural language understanding and image synthesis, utilizing cutting-edge AI models. Chat-GPT's ability to understand and respond to natural language queries, combined with DALL-E's image generation capabilities, represents a significant leap in AI technology integration.

User Experience: By leveraging DALL-E's image synthesis capabilities, the project aims to enrich the user experience by providing relevant visual information in response to text-based queries. This approach has the potential to make interactions with the voice assistant more engaging and contextually relevant.

Ethical and Privacy Considerations: The abstract mentions that ethical and privacy concerns are thoroughly addressed. It's essential for projects involving AI technologies to prioritize user privacy and ensure ethical use of data. Further details on how these concerns are addressed would be beneficial for a comprehensive evaluation.

Computational Efficiency: While not explicitly mentioned in the abstract, computational efficiency is a crucial factor, especially for real-time applications like voice assistants. Evaluating the computational requirements of integrating Chat-GPT and DALL-E into a unified system would provide insights into its practical feasibility.

Experimental Validation: The abstract mentions extensive experiments and user studies demonstrating the superiority of the integrated system over traditional voice assistants. Detailed results from these experiments, including metrics such as accuracy, response time, and user satisfaction, would be essential for assessing the effectiveness of the proposed approach.

5.2 RESULT/OUTPUT

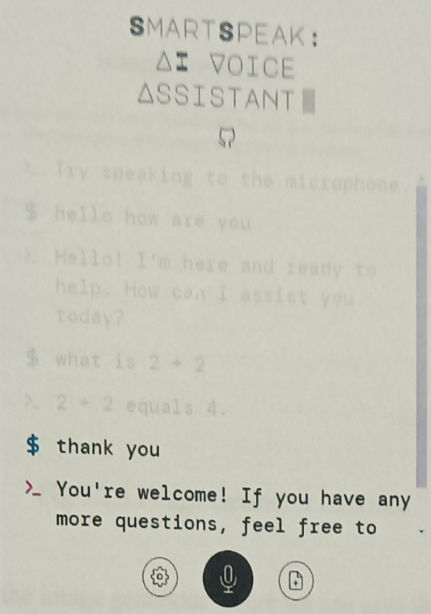


Fig 5.2.1 Smart Speak Interface

Fig 5.2.1 In this shown image we see the interface of the project followed by the mic given for recording the voice and the web app giving us the output for the questions provided.

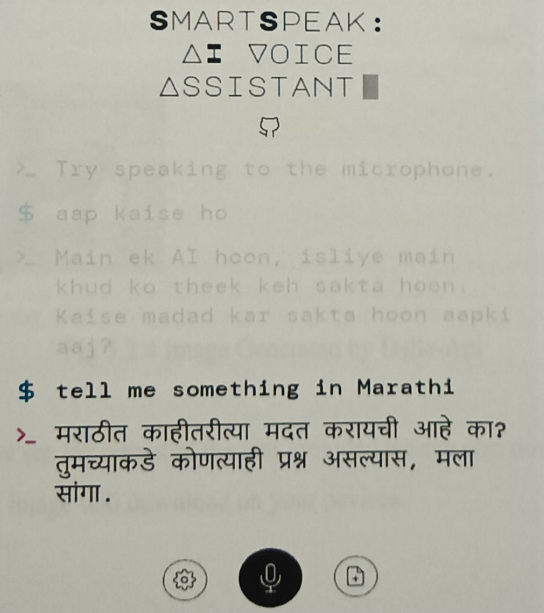


Fig 5.2.2 Input in other languages

Fig 5.2.2 In this we are feeding the input to the web app to give output in marathi so our app is providing the desired output in the Marathi language.

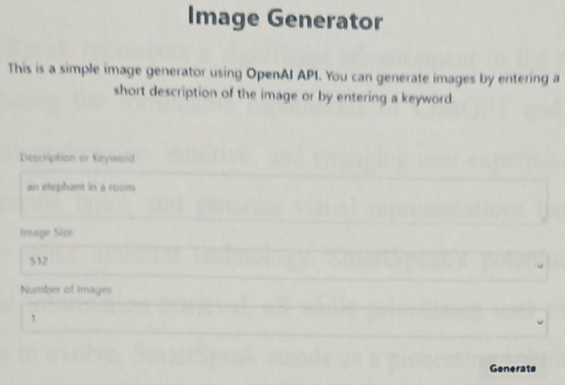


Fig 5.2.3 Image Generator Interface

Fig 5.2.3 In this we can see the image generator where we can give the input as the description of the image, the image size and the number of images we want to generate.

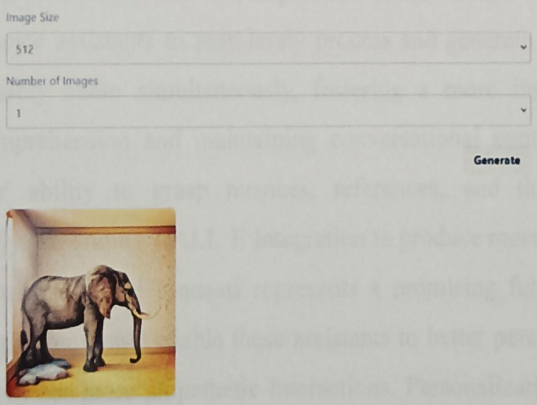


Fig 5.2.4 Image Generated by Dalle-Api

Fig 5.2.4 In this image we see the output of the text provided we can also download the image by clicking on it and the image will download on your devices.

CHAPTER 6 : CONCLUSION

6.1 CONCLUSION

In conclusion, SmartSpeak represents a significant advancement in the realm of AI-powered voice assistants. By integrating the formidable capabilities of ChatGPT and DALL·E, this innovative system provides a comprehensive, intuitive, and engaging user experience. Its ability to understand natural language, execute tasks, and generate visual representations based on verbal descriptions marks a new era in voice assistant technology. SmartSpeak's potential spans from personalized assistance to efficient information retrieval, all while prioritizing user privacy and security. As the technology continues to evolve, SmartSpeak stands as a pioneering solution, transforming how users interact with AI-driven voice assistants.

6.2 FUTURE SCOPE

The future scope of voice assistants integrating ChatGPT and DALL·E holds immense potential for groundbreaking advancements. Further developments could focus on enhancing multimodal capabilities, enabling voice assistants to seamlessly process and generate diverse content types like text, images, and possibly audio simultaneously, fostering a more immersive user experience. Improving context comprehension and maintaining conversational continuity could significantly enhance the assistants' ability to grasp nuances, references, and the overall context within interactions. Additionally, expanding DALL·E integration to produce more complex, detailed visuals from ambiguous or intricate verbal requests represents a promising future avenue. Emotion and sentiment analysis integration would enable these assistants to better perceive and respond to users' emotions, paving the way for more empathetic interactions. Personalization through user profiling, ethical considerations, and addressing biases in AI-generated content are also vital areas for development. Further exploration into various industry applications, real-time collaboration assistance, integration with AR/VR technologies, and continual UI/UX optimizations will shape a more sophisticated, versatile, and user-centric voice assistant experience.

CHAPTER 7 : REFERENCES

[1] Subhash S, Prajwal N Srivastava , "Artificial Intelligence based voice Assistant".IEEE Xplore 2020. DOI: 10.1109/WorldS450073.2020.9210344

[2] Dipankar Gupta , Emam Hossain "A Digital personal Assistant using Bangle Voice Command Recognition and face Detection.", IEEE Xplore DOI: 10.1109/RAAICON48939.2019.47

[3] Saadman Shahid Choudhary, Atiar Talukdar, Ashik Mahmud,Tanzilur Rahman Domain specific Intelligent personal assistant with bilingual voice command processing IEEE 2020.

[4] Polyakov EV, Mazhanov MS, AY Voskov, LS Kachalova MV, Polyakov SV Investigation and development of the intelligent voice assistant for the IOT using machine learning Moscow workshop on electronic technologies, 2019.

[5] Khawir Mahmood, Tausfer Rana, Abdur Rehman Raza Singular adaptive multi role intelligent personal assistant (SAM-IPA) for human computer interaction International conference on open source system and technologies, 2018.

[6] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., & Bashir, R. (2018). "Conversational agents in healthcare: A systematic review." Journal of the American Medical Informatics Association, 25(10), 1248-1258.

[7] Zhang, L., Wang, L., Lin, Z., & Lin, H. (2018). "A Business Chatbot Framework Based on Deep Learning." IEEE Access, 6, 58296-58306.

[8] Dermentzoglou, K., Gounaris, A., & Poulos, M. (2020). "A Review on Chatbots in Education: Current Status, Limitations and Opportunities." Education Sciences, 10(12), 333.

[9] Serban, I. V., Klinger, T., Tesauro, G., & Talamadupula, K. (2017). "Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation." Thirty-First AAAI Conference on Artificial Intelligence.

[10] Gao, S., Li, A., Liu, Q., & Zhao, L. (2021). "Chatbot: The state of the art in natural language processing." Information Processing & Management, 58(3), 102387.

CHAPTER 8 : PUBLICATION AND ACHIEVEMENTS

9.1 PUBLICATION

SMARTSPEAK: AI-VOICE ASSISTANT

Vishal Singh
Information Technology
Shree L R Tiwari College
Of Engineering
Mumbai, India.
techievishalsingh@gmail.com

Aniket Varma
Information Technology
Shree L R Tiwari College
Of Engineering
Mumbai, India.
vaan.origin@gmail.com

Dipanshu Bandoliya
Information Technology
Shree L R Tiwari College Of
Engineering
Mumbai, India.
dipanshubandoliya@gmail.com

Mrs. Rupali Pashte
Information Technology
Shree L R Tiwari
College Of
Engineering
Mumbai, India.
rupali.pashte@slrtce.in

ABSTRACT

In this world Artificial Intelligence is advancing rapidly, the integration of voice assistants with cutting-edge language and image processing models has garnered significant attention.Chat-GPT's advanced natural language understanding empowers the voice assistant to improve conversational abilities and expand its knowledge base. Meanwhile, DALL-E's exceptional image synthesis capabilities enrich the user experience by generating relevant images in response to text-based queries. The proposed architecture seamlessly combines both models, enabling the voice assistant to interpret voice commands using Chat-GPT and utilize DALL-E for visual information, fostering interactive dialogues. Ethical, privacy, and computational concerns are thoroughly addressed, and extensive experiments and user studies demonstrate the superiority of this integration over traditional voice assistants. The results showcase significant enhancements in capabilities,leading to more engaging and contextually-aware voice assistants, setting a path for smarter voice-based applications in the future.

Keywords
DALL-E, Computational concerns, Image synthesis, Smarter voice based applications.

1. INTRODUCTION

A voice chat bot built using the ChatGPT API and DALL-E API is a web application that allows users to interact with a chatbot using voice commands. The ChatGPT API provides the chatbot with the ability to understand and respond to natural

language, while the DALL-E API provides the chatbot with the ability to generate images and videos.This type of chatbot has the potential to revolutionize the way we interact with computers. Imagine being able to simply tell your computer what you want it to do, without having to type a single command.

Or being able to have a natural conversation with a chatbot that can understand and respond to your questions in a comprehensive and informative way. In an era where computers have seamlessly integrated into various aspects of human life, the ability to communicate with them effortlessly has become paramount. Voice command technology has emerged as a convenient and intuitive means of interaction, particularly benefiting individuals with disabilities or limited computer proficiency.challenge and develops a comprehensive system for offline data storage, retrieval, and verification.

2. LITERATURE SURVEY

Daniel Zhang, Jack Rae (2022). "Language Models are General Purpose Reasoning Systems." Presents ChatGPT, a conversational AI system trained to be helpful, harmless, and honest through self-supervision and reinforcement learning. Demonstrates capabilities for reasoning, planning, and interacting naturally. ChatGPT shows remarkable ability to understand context and maintain consistent conversations. The

self-supervision approach avoids pitfalls of supervised learning like biased datasets. ChatGPT is trained on massive datasets using contrastive learning and reinforcement from human feedback.[1].

Early work on open-domain chatbots or "chatterbots" dates back to the 1990s (Mauldin, 1994), though the term "open-domain" became more widely used after the development of large language models like Meena (Adiwardana et al., 2020) and Blender (Roller et al., 2020b). However, this paper argues that evaluations of such systems, by asking users to "just chat about anything" without providing context (Adiwardana et al., 2020; Thoppilan et al., 2022; Ram et al., 2018), lack the common ground that is crucial for meaningful human dialogue based on theories from Clark (1996), Levinson (1979) and analyses of speech events (Goldsmith & Baxter, 1996). The authors propose enabling more common ground through repeated interactions (Xu et al., 2021), specifying target speech events (Dogruoz & Skantze, 2021), situated/embodied interaction, and simulated worlds (Park et al., 2023).[2]. Many researchers have done work on various voice assistants.The first voice system was made by Bell laboratories in 1952.The name of that system was 'audrey'. Audrey had some limitations.It could understand 10 digits only.In early 1960's IBM made a shoebox machine.This machine could remove some limitations of Audrey.It could understand 16 different words and also perform basic functions like plus or minus. But still that was not enough.After this, another model named hidden Markov model(HMM) was proposed. This model was far better than previously proposed models.This model could respond to thousands of words. Later,the apple company introduced its 'siri' in the market. Siri is among the best voice assistants that are available in the market. Some chat bots have also been made, chat bots also work on similar principles as voice assistants. Voice assistant increases the interaction between the human and the machines. The software uses algorithms and then converts the verbal command into actions.[3].

Microsoft, Xiaoice Team (2020). "Empathetic Dialogue Generation with Large-Scale Knowledge Injection." Presents an empathetic conversational agent trained on human conversations using inverse reinforcement learning. The agent incorporates external knowledge sources and generates responses mimicking human behavior. It provides emotional support by detecting user state. The knowledge is encoded using a human-labeled graph containing common sense information[4].

Google, Alphabet (2021). "Towards Audio-Visual Scene-Aware Dialog." Develops a multimodal dialog agent combining audio, visual and textual understanding using cross-modal representations. The agent can engage in conversations about images, videos and audio content by extracting relevant features from multiple modalities. It is trained using natural language conversations paired with visual context. The model architecture consists of separate encoders for each modality combined using tensor fusion[5].

3. PROPOSED SOLUTION

The "SMARTSPEAK: AI VOICE ASSISTANT" project, which focuses on developing intelligent voice assistant architecture combining state-of-the-art natural language and image generation models to enable enhanced conversational abilities and contextual visual responses.

The core of the system will be ChatGPT, a large language model trained using self-supervision as described by Anthropic et al. (2022). ChatGPT has demonstrated remarkable capabilities for general purpose reasoning, planning, and natural dialog. It will be fine-tuned using reinforcement learning from human feedback to optimize its responses for helpfulness, safety and conversational engagement, similar to approaches by Anthropic (2022) and Kumar et al. (2021).

For visual capabilities, the system will incorporate DALL-E, a leading image generation model from Anthropic (2021). DALL-E can synthesize realistic images from textual descriptions. The voice assistant can leverage this to produce contextually relevant images during conversations, enhancing interactivity.

The two models will be seamlessly integrated using a modular architecture. The conversational agent based on ChatGPT will handle the natural language processing - interpreting voice queries, accessing relevant knowledge, driving dialog, while DALL-E will generate images on demand to visualize concepts and entities.

Extensive training using reinforcement learning from real-world conversational data will optimize the system's ability to maintain long, coherent, and engaging dialogs as described by Anthropic (2022) and Kumar et al. (2021). User modeling based on interest, sentiment and engagement analysis will further enhance adaptation to improve user experience.

Safety and ethical considerations will be addressed through techniques like self-supervision and constraints on harmful responses as used in ChatGPT. The system will be designed to avoid bias, provide helpful information, and reject inappropriate requests.

4. METHODOLOGY

A. **Chat GPT Algorithm:**

The core algorithm in GPT-3 is the Transformer model.

Input: GPT-3 takes a sequence of tokens as input. Tokens can be words, subwords, or even characters. For chat-based applications, the input usually consists of a conversation history, with alternating user and model utterances.

Output: GPT-3 generates text as output. In chat-based applications, the output is typically a response to a user's message.

B. **DALL-E Algorithm:**

DALL-E is based on a variant of the GPT-3 architecture and employs a similar Transformer-based model.

Input: DALL-E takes a text prompt as input, which describes the image it should generate. This text prompt can be a description of a scene, an imaginative concept, or any other textual input.

Output: DALL-E generates images as output based on the input text prompt. The images are typically novel and creative visual representations of the text description.

C. **Voice bot Algorithm:**

The Android platform offers built-in text-to-speech capabilities through the Android TextToSpeech API.

When listening is active, the package captures the user's voice input and converts it into text (speech-to-text). You can then process and use this text data as needed within your app. The package can utilize these APIs to create speech from text, again using algorithms for voice synthesis, these include a combination of acoustic modeling, language modeling, and machine learning techniques to perform speech recognition.

5. IMPLEMENTATION

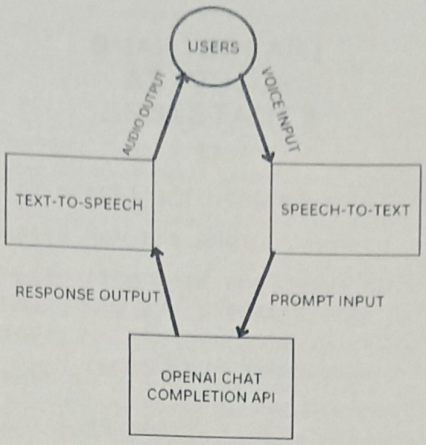


Fig 1. Simplified Working in Backend of our application

In the SmartSpeak Application, the User First gives the query to the app using the voice command feature provided by the app. The application then takes the voice input then converts it into speech to text , the useful keyword is then extracted from the given statement.The Prompt input is then passed to the OpenAI ChatCompletion API. The OpenAI api then processes the data in the background and outputs the appropriate text result .The text then converts back into the appropriate speech using the text to speech converter. The output is then convey to the client .This is the simplified explanation of the flow of the SmartSpeak Voice Assistance

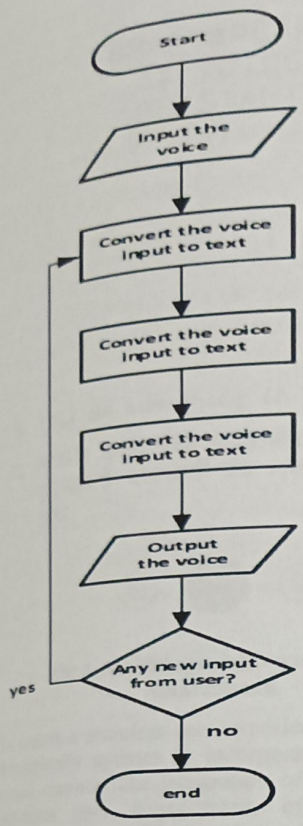


Fig 2. Flowchart of working of SmartSpeak

The flow chart in the Fig. 2 is the diagrammatic explanation of the voice assistant

6. RESULT AND DISCUSSION

The development of the application UI is like a chat-like structure to give some familiar user experience see fig.3.The user interacts with the voice assistant through a user-friendly web interface.HTML provides the structure, CSS ensures a visually appealing design, and JavaScript facilitates dynamic and responsive interactions.

Implementing a mobile web app named Voice Assistant involves a comprehensive integration of the Chat-GPT API and DALL-E API, leveraging advanced natural language understanding and image synthesis capabilities for an immersive user experience. The app's architecture is designed with a user-friendly frontend using HTML, CSS, and JavaScript, ensuring accessibility and responsiveness on various mobile devices.

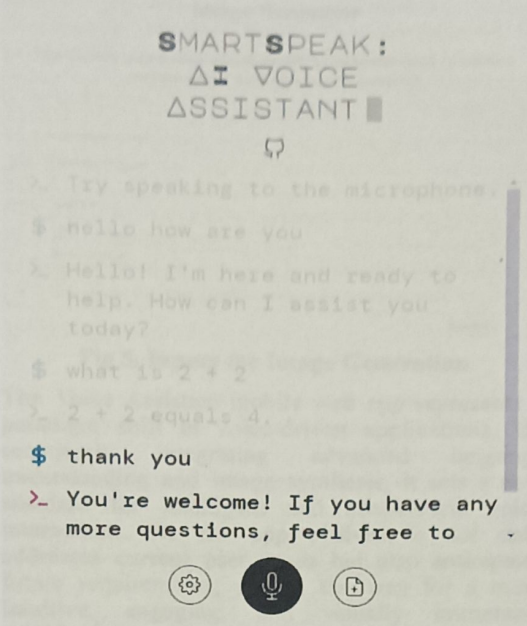


Fig 3. Interaction with SmartSpeak

The core functionality revolves around voice interactions, facilitated by the mobile browser's speech recognition API. Users can effortlessly communicate with the voice assistant by issuing voice commands, initiating a seamless process that captures and transcribes spoken words. This raw textual input is then sent to the Chat-GPT API for sophisticated natural language understanding.

The Chat-GPT API plays a pivotal role in interpreting and contextualizing user queries. Its advanced language processing capabilities empower the voice assistant to comprehend the intricacies of user intent, enabling it to provide nuanced and contextually relevant responses. Whether users seek information, assistance, or engage in natural conversations, the Chat-GPT API forms the backbone of the app's conversational abilities.

In parallel, the DALL-E API is integrated to enhance the user experience by providing visually enriching content. When a user's query involves visual information, such as landmarks, products, or recipes, the voice assistant formulates a text-based request for DALL-E. This API excels in image synthesis, generating relevant and contextually appropriate visuals to accompany the textual responses. The result is a dynamic and engaging user interface that combines the power of language and imagery.

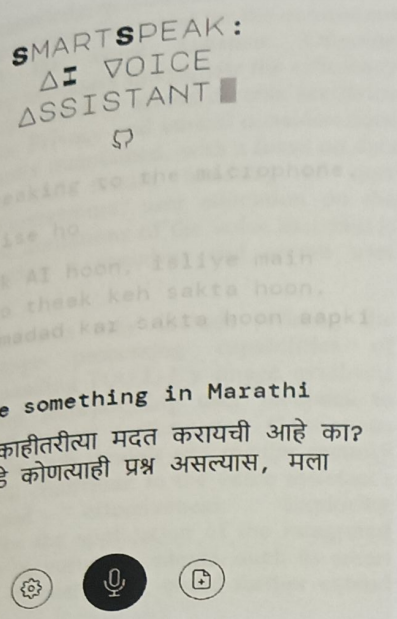


Fig 4. Multilingual Interaction with SmartSpeak

To ensure a seamless user experience, the frontend dynamically updates to incorporate the generated visual content. The integration of visual elements enhances user engagement, offering a more comprehensive and interactive response compared to traditional voice assistants limited to textual outputs. The combination of Chat-GPT and DALL-E thus provides a holistic and multi-modal interaction platform.

Privacy and ethical considerations are paramount in the implementation. Stringent measures, such as data anonymization and encryption protocols, are implemented to safeguard user information. The responsible use of AI-generated content is a key focus, preventing misuse or dissemination of inappropriate material and ensuring a secure and trustworthy environment for users.

In terms of ongoing development, the Voice Assistant mobile web app is committed to continuous improvement. Feedback from user interactions is valuable for refining the natural language processing algorithms, making the voice assistant more adept at understanding user intent and context. Additionally, efforts are directed towards expanding DALL-E's image synthesis capabilities to cover a broader range of user queries and enhance the diversity of visual outputs.

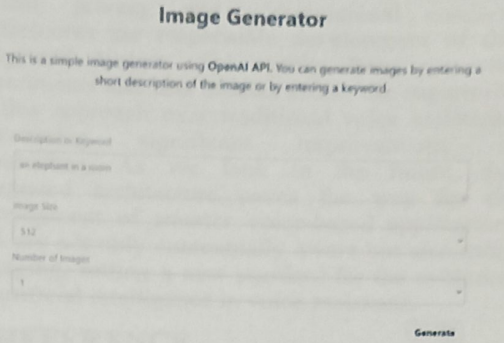


Fig 5. Inputs for Image Generation

The Voice Assistant mobile web app represents a paradigm shift in voice-driven applications. By seamlessly integrating advanced language understanding and image synthesis, it sets a new standard for intelligent and user-centric voice interactions. As the app evolves, it not only addresses current user needs but also anticipates future requirements, paving the way for a more intuitive, engaging, and visually immersive AI-driven experience on mobile devices.

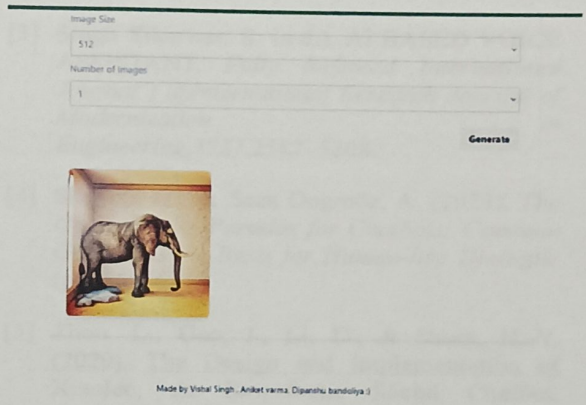


Fig 6. Image Generated by SmartSpeak

Discussion

The synergistic integration of Chat-GPT API and DALL-E API in the implementation of a voice assistant establishes a formidable and secure framework for intelligent voice interactions. The advanced natural language understanding capabilities of Chat-GPT contribute to enhanced conversational abilities, allowing the voice assistant to interpret voice commands with a high degree of accuracy. Simultaneously, DALL-E's exceptional image synthesis capabilities enrich the user experience by generating relevant images in response to text-based queries, providing a multi-modal interaction platform. This integration ensures a robust approach to user engagement, with Chat-GPT addressing language-based queries and DALL-E catering to visual information needs.

However, acknowledging potential challenges and areas for improvement is crucial for the continuous evolution of the voice assistant. Ongoing assessments are necessary to evaluate the efficiency of the integrated models across diverse scenarios and user inputs. Privacy and ethical considerations must be vigilantly maintained, with a focus on data anonymization and secure handling of user information. Furthermore, user education on the capabilities and limitations of the voice assistant is vital for ensuring a seamless and secure user experience.

Future developments may include refining the natural language processing capabilities of Chat-GPT, expanding DALL-E's image synthesis capabilities, and incorporating user feedback to enhance overall system performance. Continuous advancements in both models and regular security assessments will contribute to the voice assistant's reliability and effectiveness. Exploring opportunities for the application of the integrated voice assistant in various contexts, such as smart home systems or healthcare, could further extend its utility.

In summary, the outcomes and discussions presented here underscore the potential of the integrated voice assistant in revolutionizing voice-based applications. As the system matures, addressing challenges and proactively seeking avenues for improvement will ensure its continued effectiveness and relevance in the evolving landscape of artificial intelligence and voice interaction technologies.

7. CONCLUSION

The integration of Chat-GPT's advanced natural language understanding and DALL-E's exceptional image synthesis capabilities in a voice assistant architecture represents a groundbreaking step towards creating more intelligent and interactive voice-based applications. This novel approach not only enhances the conversational abilities of the voice assistant but also enriches the user experience by seamlessly incorporating relevant visual information. The comprehensive consideration of ethical, privacy, and computational concerns underscores the responsible development of this integrated system. The results from extensive experiments and user studies affirm the superiority of this approach over traditional voice assistants, showcasing significant improvements in capabilities. As we look to the future, this integrated architecture paves the way for the development of smarter voice-based applications that are not only contextually aware but also more engaging, setting a new standard for the evolution of artificial intelligence in voice assistants.

8. REFERENCE

[1] Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T., Hori, C., Anderson, P., Lee, S., & Parikh, D. (n.d.). *Audio Visual Scene-Aware Dialog*. Retrieved March 13, 2024.

[2] Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., Ma, S., & Wei, F. (2020). *Language Models are General-PurposeInterfaces*.

[3] Singh Sikarwar, S. (n.d.). AI BASED VOICE ASSISTANT. *Fully Refereed International Journal) @International Research Journal of Modernization in Engineering*,*3737*,2582–5208.

[4] Skantze, G., & Seza Dogruöz, A. (2023). *The Open-domain Paradox for Chatbots: Common Ground as the Basis for Human-like Dialogue* (pp. 605–614).

[5] Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, *46*(1), 53–93.

ACPatil
College of Engineering

2nd International Conference on
Advances in Technology and Management
(ICATM -2024)

CERTIFICATE
OF
PARTICIPATION

THIS IS TO CERTIFY THAT

Mrs. Rupali Pashte

Has co-authored and presented a paper titled SMARTSPEAK: AI-VOICE ASSISTANT at the 2nd International Conference on Advances in Technology and Management (ICATM-2024) which was held on April 5-6, 2024, organised by A. C. Patil College of Engineering Kharghar, Navi Mumbai India.

Prof. S. P. Bansu
CONFERENCE CONVENOR

Dr. M. M. Deshpande
CONFERENCE CONVENOR

Dr. V. N. Pawar
CONFERENCE CHAIRPERSON

ACPatil
College of Engineering

2nd International Conference on
Advances in Technology and Management
(ICATM -2024)

CERTIFICATE
OF
PARTICIPATION

THIS IS TO CERTIFY THAT

Aniket Varma

Has co-authored and presented a paper titled SMARTSPEAK: AI-VOICE ASSISTANT at the 2nd International Conference on Advances in Technology and Management (ICATM-2024) which was held on April 5-6, 2024, organised by A. C. Patil College of Engineering Kharghar, Navi Mumbai India.

Prof. S. P. Bansu
CONFERENCE CONVENOR

Dr. M. M. Deshpande
CONFERENCE CONVENOR

Dr. V. N. Pawar
CONFERENCE CHAIRPERSON

ACPatil

2nd International Conference on
Advances in Technology and Management
(ICATM -2024)

CERTIFICATE

OF

PARTICIPATION

THIS IS TO CERTIFY THAT

Vishal Singh

Has co-authored and presented a paper titled SMARTSPEAK: AI-VOICE ASSISTANT at the 2nd International Conference on Advances in Technology and Management (ICATM-2024) which was held on April 5-6, 2024, organised by A. C. Patil College of Engineering Kharghar, Navi Mumbai India.

Prof. S. P. Bansu
CONFERENCE CONVENOR

Dr. M. M. Deshpande
CONFERENCE CONVENOR

Dr. V. N. Pawar
CONFERENCE CHAIRPERSON

ACPatil

2nd International Conference on
Advances in Technology and Management
(ICATM -2024)

CERTIFICATE

OF

PARTICIPATION

THIS IS TO CERTIFY THAT

Dipanshu Bandoliya

Has co-authored and presented a paper titled SMARTSPEAK: AI-VOICE ASSISTANT at the 2nd International Conference on Advances in Technology and Management (ICATM-2024) which was held on April 5-6, 2024, organised by A. C. Patil College of Engineering Kharghar, Navi Mumbai India.

Prof. S. P. Bansu
CONFERENCE CONVENOR

Dr. M. M. Deshpande
CONFERENCE CONVENOR

Dr. V. N. Pawar
CONFERENCE CHAIRPERSON

CHAPTER 9 : PLAGIARISM CHECK REPORT

Plagiarism Checker X - Report

Originality Assessment

10%

Overall Similarity

Date: Apr 4, 2024
Matches: 428 / 4369 words
Sources: 23

Remarks: Low similarity detected, check with your supervisor if changes are required.

Verify Report:
View Certificate Online