Grenze journal name, Volume. 1, Issue.1, May 2015

Personal Voice Assistant

Sanskruti Bhujade[1] Aditya Shahu[2] ,Yashashwi Wahie[3] ,Hritik Pandey[4], Mridula Korde[5]
[1-5] , Ramdeobaba College of Engineering and Management, Nagpur, India 440013
Email:bhujadesr@rknec.edu[1] , shahuaa_3@rknec.edu[2], wahieyv@rknec.edu[3], pandeyhr_2@rknec.edu[4], kordems@rknec.edu[5]

***Abstract*—In the age of smart devices and the Internet of Things (IoT), the integration of Personal Voice Assistants (PVAs) with advanced AI models like ChatGPT has opened up exciting possibilities for enhancing the capabilities of edge devices. This research paper explores the integration of a PVA with ChatGPT on the ESP32 microcontroller platform, aiming to empower edge devices with natural language processing and conversational intelligence. The suggested PVA creates human-like replies to user queries using ChatGPT, a potent AI language model. ChatGPT has been trained on a massive dataset allowing it to comprehend and react to a variety of requests. The integration of ChatGPT with the ESP32 microcontroller, a low-power and flexible platform for IoT applications, forms the basis of this research. To ensure the seamless cohabitation of AI capabilities within resource-constrained edge devices, the implementation specifics, including model optimization, deployment techniques, and hardware considerations, are thoroughly covered.**

***Index Terms*—Text to Speech, AI, ChatGPT**

I. INTRODUCTION

In the era of technological advancement in AI, integration of AI and edge computing has ushered in a new wave of innovation. Strong language models, like OpenAI's, ChatGPT, have been made possible in recent years by developments in natural language processing (NLP). These models are helpful for a range of applications, including customer service, content creation, and personal assistants, since they can produce writing that appears human and carry on conversations.

The transformer architecture, first presented by Vaswani et al. in 2017, serves as the foundation for ChatGPT. The transformer design is particularly suited for text generation since it can process a lot of context in parallel, unlike standard recurrent neural networks (RNNs).

It makes conversational communication between users and machines possible. It was much awaited and is expected to completely change the way people use technology. It was just released to the public. Without requiring any prior programming or computer science knowledge or experience, ChatGPT allows users of all ages and backgrounds to interact organically in a number of languages by understanding context, intent, sentiment, and more. It can be utilized for a variety of purposes, including banking, healthcare, entertainment, education, and customer service.

In the proposed device, the communication is established through Google API and ESP32 controller which has gained popularity among IoT enthusiasts and amateurs because of its strong processing capabilities, integrated WiFi and Bluetooth connectivity, and low power consumption. With ChatGPT, building more complex and interactive programs that can respond to text input in a human-like manner, like voice assistants or chatbots is possible. The most effective means of communication for people to convey their ideas and emotions across linguistic boundaries is speech.

The characteristics of speech vary among languages. Nonetheless, each person has a different dialect and tempo even when speaking the same language. For some, this makes it difficult to understand the message being delivered. Long speeches can occasionally be challenging to follow for a variety of reasons, including varying pronunciation, tempo, and other factors. The interdisciplinary topic of computational linguistics known as voice recognition contributes to the development of systems that enable speech recognition and text translation. Text summarizing takes a text as a source, selects the most crucial information from it, and summarizes it adequately.

Text to speech (TTS) stands as a technology powered by artificial intelligence, capable of converting written text into synthesized speech. This technology employs advanced algorithms and speech synthesis methods to produce voices that are remarkably natural and high in quality. TTS brings the ability for machines to engage in spoken communication with users, introducing an auditory dimension to their interactions. While prominent technology giants like Amazon, Microsoft, and Google have made substantial investments in text to speech research, OpenAI has not yet ventured into this domain.

As a pioneer in generative AI, OpenAI possesses the necessary resources to potentially compete with leading text to speech providers should it decide to launch a TTS product or feature. The integration of TTS would further enhance ChatGPT's versatility, catering to learning, content creation, and various other applications. Users could benefit from having study materials read aloud, listening to drafts of their written work, or simply enjoying the experience of hearing ChatGPT's explanations. On the whole, the incorporation of a text to speech tool into ChatGPT promises to elevate the user experience, rendering interactions more captivating and accessible.

II. LITERATURE SURVEY:

A thorough analysis of the literature suggests that extensive dataset was used to train ChatGPT to generate writing that is logical and human-like. Predicting the next word in a series of text was the language modeling objective used to train the model. Preprocessing was done on the data to guarantee excellent quality and get rid of duplicates. ChatGPT is the recommended option because of its strong performance, which frequently outperformed other models (ChatGPT, Google Bard, PDF).

A neural network chatbot called ChatGPT can construct responses depending on user input and retains the specifics of previous talks. It provides examples, asks questions about hobbies, and responds to inquiries for gifts. The chatbot does not discuss its goal, race, or religion, nor does it respond to contentious questions or offer personal viewpoints. It is incapable of independent thought and is equipped with filters to stop it from writing messages about immoral or unlawful behavior. Editing ChatGPT responses is possible, and users can ask a leading question or provide more details to receive the right answer if the bot interprets the context incorrectly. It's crucial to remember that ChatGPT lacks autonomy and might not be able to respond to inquiries regarding contentious subjects.

Voice assistants, which include Cortana, Alexa, Google Assistant and Siri are software agents that run on smartphones or specially designed speaker devices. They record the user's voice, listen for a key word, and transfer the recording to a dedicated server. After processing and interpreting the instruction, the server gives the voice assistant the tasks, media, or information it needs. Voice command-enabled services are becoming more prevalent, and Internet-of-things device makers are incorporating voice control into their products as well. The goal of this technology is to transform computer-human interaction.

The most crucial component of human communication is speech. Speech is regarded as the primary medium for communication, despite the fact that there are other ways in which we can convey our ideas and emotions. The process of teaching a machine to recognize different people's voice based on specific words or phrases is called speech recognition. It is easy to distinguish differences in pronunciation in each person's speech. Speech originates as a signal, which is then processed to transfer all of its information into text format. The process of taking a signal and applying specific logic to convert it to the needed format is called feature extraction.

Even if speaking is the most straightforward form of communication, there are still certain concerns with speech recognition, such as issues with pronunciation, fluency, broken words, stuttering, etc. When processing a speech, each of these needs to be taken into account. One of the key ideas in the documentation industry is text summary. Because they take a lot of time to read and comprehend, lengthy publications are challenging. This issue is resolved by text summarization, which offers a condensed, semantic summary of the text.

A combination of text summarization and speech to text conversion is used in the proposed study. Applications that call for a concise synopsis of lengthy speeches can benefit from this hybrid approach, which is especially helpful for documentation. The suggested approach's flow diagram shows males using voice recognition

Any application that calls for summarization benefits from the combination of these two modules. The first and most important stage in using natural language processing, or NLP, is to extract the speech's valuable qualities. The act of summarizing is hampered if a word or sentence is understood to be meaningless. Semantics is necessary while summarizing the text, therefore even punctuation is important in the process.

In [1] , it was proposed that Artificial Intelligence, interprets user requests and provides answers, saving time and ensuring up-to-date information. Chatbots, with memory, can recognize mood changes, respond to emotions, and run automated workflows. Modern AI uses machine learning, natural language processing, and sentiment analysis for personalized, real-time interactions were elaborated.

AI solutions like Chat GPT revolutionize business operations with automation and faster processing times, but may soon become obsolete due to poor performance and lack of contextual understanding, especially in data-driven scenarios up to 2021[2].

In [4], the proposed model uses automatic speech recognition to convert user's speech into text format, utilizing Google's gTTS engine and Microsoft's SAPI5, enabling voice assistant for reading, searching, and document conversion.

ChatGPT, an AI-based chatbot, has the potential to revolutionize technology interactions, improving customer service and content creation. Google, Bard, and Baidu are launching AI chatbots. However, caution is needed due to its limitations and the need for further research [5].

In [5], it was explored that development of a virtual assistant capable of answering questions, controlling IoT devices, and controlling peripherals, comparing speech recognition systems with artificial neural networks.

In [6],the comparison between voice recognition technology and voice assistant software, developed by companies like Apple, Amazon, Google, and Microsoft, offers natural language commands.

In [7],Google cloud is used for speech to text conversion API . In [8], it is proposed that Google Cloud Speech an 80% recognition rate for speech impaired and 100% for normal voice speech recognition, influenced by tone, pronunciation, and speech speed.

III. PROPOSED METHODOLOGY:

The proposed flow is provided below.

A. *Voice Recording:*

This is the initial step where the audio data is captured. This could be through a microphone or from a pre-recorded audio file.

B. *File Conversion:*

The audio data captured in the first step is then converted into a digital format, typically a string. This is done using a process called speech recognition, which involves complex algorithms to interpret and transcribe the spoken words into text.

C. *Text Parsing:*

Once the audio data is converted into text, it is then parsed. Parsing involves processing the text and preparing it for summarization. This could involve removing stop words (commonly used words like 'is', 'an', 'the', etc.), punctuation, and other unnecessary elements.

D. *Tokenization:*

This is the process of breaking down the text into individual words or 'tokens'. Each word in the text is considered a separate token. This is a crucial step for understanding the frequency of each word in the text.

E. *Summarization:*

After tokenization, the text is summarized. This involves identifying the most important sentences or phrases in the text. The importance of a sentence can be determined based on the frequency of the words it contains. The sentences are then ranked based on their importance, and the top-ranked sentences form the summary.

F. *Writing to File:*

The summarized text is then written to a file. This could be a text file, a Word document, or any other format that can store text.

G. Statement Recognition:

The final step involves recognizing if the summarized text forms a valid statement. This could involve checking for grammatical correctness, coherence, and relevance.

H. End of Process:

The process ends after the statement is recognized. The output is a summarized text that is a valid statement.

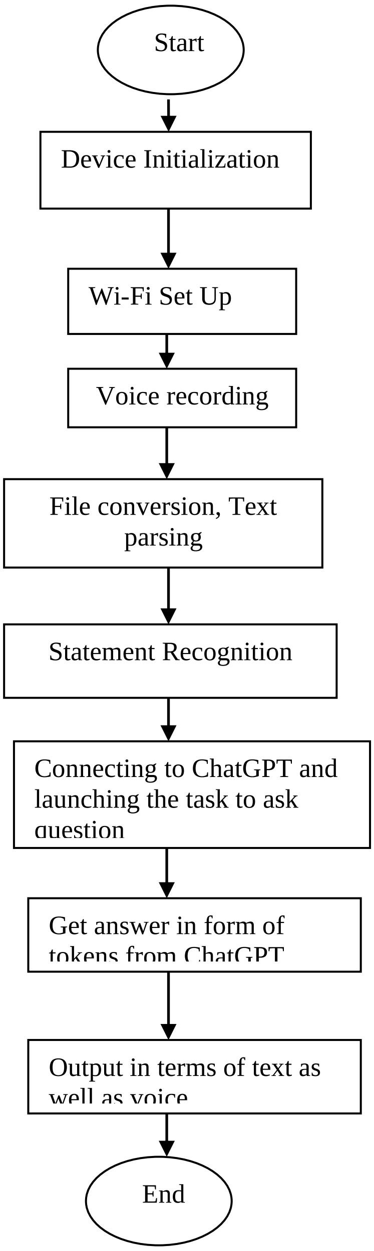

Start
Device Initialization
Wi-Fi Set Up
Voice recording
File conversion, Text parsing
Statement Recognition
Connecting to ChatGPT and launching the task to ask question
Get answer in form of tokens from ChatGPT
Output in terms of text as well as voice
End

Fig 1: Flow of proposed methodology

IV. IMPLEMENTATION

Being a frontrunner in generative AI, OpenAI possesses the necessary resources to compete with leading text-to-speech providers, should it choose to introduce a TTS product or feature. The incorporation of TTS would not only enhance ChatGPT's functionality for learning and content creation but also broaden its appeal. Users could benefit from having study materials read aloud, listening to drafts of their writing, or simply enjoying ChatGPT's explanations. In essence, the integration of a text-to-speech tool into ChatGPT would significantly elevate the user experience, rendering interactions more captivating and accessible.

A. System Testing

Testing was conducted by speaking the numbers 1 to 10, performed by three speech impaired voices and three normal voices and the average recognition rate by using following relation

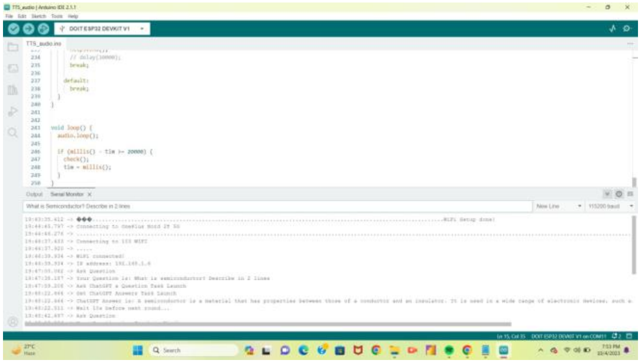
$$SR = \frac{NCorrect}{NTotal} x100\%$$

Where,
SR is the Success Rate,
NCorrect is the Number of correctly spoken tokens
NTotal is the Total number of samples of spoken tokens

Blackbox testing involves evaluating a predetermined function by examining its output to determine whether the function operates as intended or not. This assessment is conducted without knowledge of the internal workings of the system. The test is executed by running the system and observing the resulting output. Table 1 displays the results of blackbox testing specifically designed for the Automatic Speech Recognition (ASR) system.

TABLE I. RESULT OF BLACKBOX TESTING.

Description	Expected Result	Result
Run a task to ask question on semiconductor	The system can display output text along with a voice output	success
Run a task to ask question on college	The system can display output text along with a voice output	success
Run a question to give answer in Hindi	The system can display output text along with a voice output	success
Run a task to speak in Hindi	The system can display output text along with a voice output	success
Run a task to ask question on semiconductor	The system can display output text along with a voice output	success

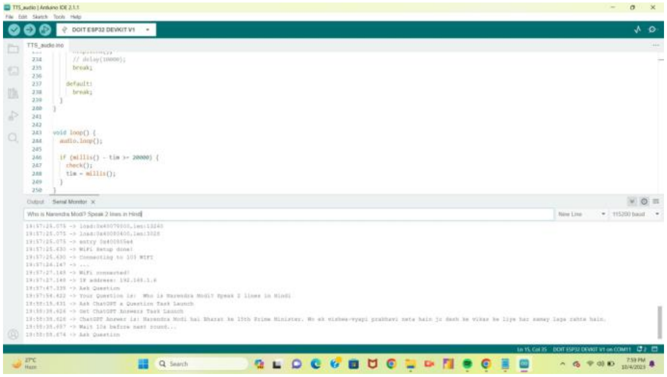


```
      // delay(10000);
      break;

    default:
      break;
  }
}


void loop() {
  audio.loop();

  if (millis() - tim >= 20000) {
    check();
    tim = millis();
  }
}
```

What is Semiconductor? Describe in 2 lines

```
19:43:35.412 -> ◆◆◆..........................................................WiFi Setup done!
19:44:45.797 -> Connecting to OnePlus Nord 2T 5G
19:44:46.276 -> ..................................................................................
19:46:37.433 -> Connecting to 103 WIFI
19:46:37.920 -> .....
19:46:39.934 -> WiFi connected!
19:46:39.934 -> IP address: 192.168.1.6
19:47:05.082 -> Ask Question
19:47:35.137 -> Your Question is: What is semiconductor? Describe in 2 lines
19:47:59.208 -> Ask ChatGPT a Question Task Launch
19:48:22.466 -> Get ChatGPT Answers Task Launch
19:48:22.466 -> ChatGPT Answer is: A semiconductor is a material that has properties between those of a conductor and an insulator. It is used in a wide range of electronic devices, such a
19:48:22.511 -> Wait 15s before next round...
19:48:42.497 -> Ask Question
```

Fig. 2. Initial setup of connection and connection established

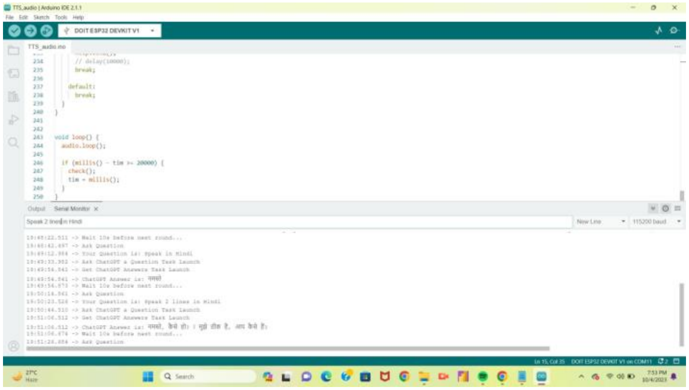


```
        // delay(10000);
        break;

      default:
        break;
    }
}


void loop() {
  audio.loop();

  if (millis() - tim >= 20000) {
    check();
    tim = millis();
  }
}
```

Who is Narendra Modi? Speak 2 lines in Hindi

```
19:57:25.075 -> load:0x40078000,len:13260
19:57:25.075 -> load:0x40080400,len:3028
19:57:25.075 -> entry 0x400805e4
19:57:25.430 -> WiFi Setup done!
19:57:25.430 -> Connecting to 103 WIFI
19:57:26.147 -> ....
19:57:27.148 -> WiFi connected!
19:57:27.148 -> IP address: 192.168.1.6
19:57:47.339 -> Ask Question
19:57:54.422 -> Your Question is:  Who is Narendra Modi? Speak 2 lines in Hindi
19:58:15.431 -> Ask ChatGPT a Question Task Launch
19:58:35.426 -> Get ChatGPT Answers Task Launch
19:58:35.426 -> ChatGPT Answer is: Narendra Modi hai Bharat ke 15th Prime Minister. Wo ek vishwa-vyapi prabhavi neta hain jo desh ke vikas ke liye har samay lage rahte hain.
19:58:35.457 -> Wait 10s before next round...
19:58:55.474 -> Ask Question
```

Fig. 3. Query given after connection establishment

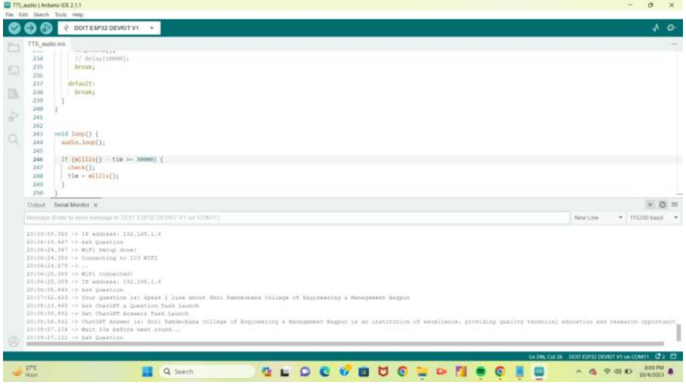


```
        // delay(10000);
        break;

      default:
        break;
    }
}


void loop() {
  audio.loop();

  if (millis() - tim >= 20000) {
    check();
    tim = millis();
  }
}
```

Speak 2 lines in Hindi

```
19:48:22.511 -> Wait 10s before next round...
19:48:42.497 -> Ask Question
19:49:12.984 -> Your Question is: Speak in Hindi
19:49:33.902 -> Ask ChatGPT a Question Task Launch
19:49:54.542 -> Get ChatGPT Answers Task Launch
19:49:54.541 -> ChatGPT Answer is: नमस्ते
19:49:54.573 -> Wait 10s before next round...
19:50:16.561 -> Ask Question
19:50:23.528 -> Your Question is: Speak 2 lines in Hindi
19:50:44.510 -> Ask ChatGPT a Question Task Launch
19:51:06.512 -> Get ChatGPT Answers Task Launch
19:51:06.512 -> ChatGPT Answer is: नमस्ते, कैसे हो? । मुझे ठीक है, आप कैसे हैं?
19:51:06.474 -> Wait 10s before next round...
19:51:28.854 -> Ask Question
```

Fig. 4. Query given in Hindi language



TTS_audio | Arduino IDE 2.1.1
File Edit Sketch Tools Help
DOIT ESP32 DEVKIT V1
TTS_audio.ino
Output
Serial Monitor
Message (Enter to send message to 'DOIT ESP32 DEVKIT V1' on 'COM11')
New Line
115200 baud
Ln 246, Col 26 DOIT ESP32 DEVKIT V1 on COM11
27°C Haze
Search
8:09 PM 10/4/2023

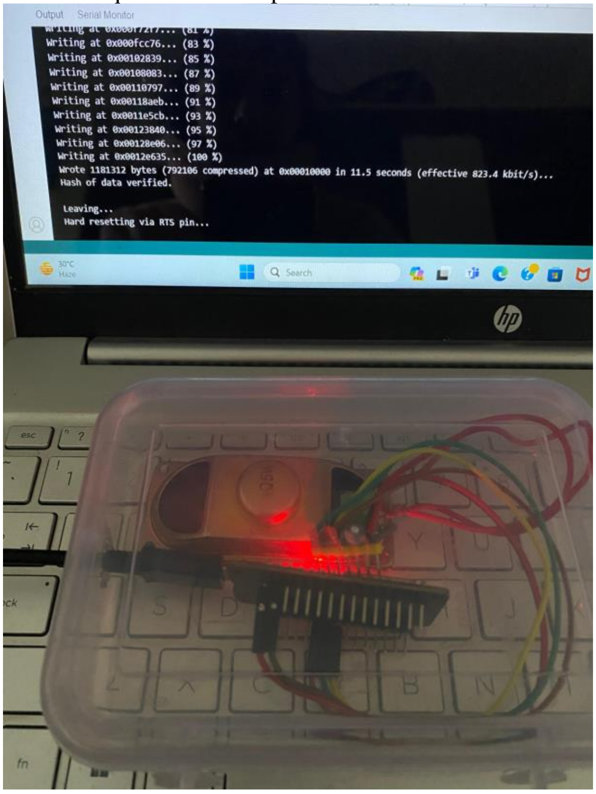
```
        // delay(10000);
        break;

      default:
        break;
    }
}


void loop() {
  audio.loop();

  if (millis() - tim >= 30000) {
    check();
    tim = millis();
  }
}
```

```
20:03:58.305 -> IP address: 192.168.1.6
20:04:18.487 -> Ask Question
20:06:24.347 -> WiFi Setup done!
20:06:24.308 -> Connecting to 103 WIFI
20:06:24.878 -> ..
20:06:25.388 -> WiFi connected!
20:06:25.309 -> IP address: 192.168.1.6
20:06:55.448 -> Ask Question
20:07:52.633 -> Your Question is: Speak 1 line about Shri Ramdeobaba College of Engineering & Management Nagpur
20:08:23.465 -> Ask ChatGPT a Question Task Launch
20:08:56.882 -> Get ChatGPT Answers Task Launch
20:08:56.842 -> ChatGPT Answer is: Shri Ramdeobaba College of Engineering & Management Nagpur is an institution of excellence, providing quality technical education and research opportunit
20:08:57.104 -> Wait 10s before next round...
20:09:27.122 -> Ask Question
```

Fig. 5. Query given after connection establishment

V. CONCLUSION

AI voice assistants have revolutionized the way we interact with technology by providing convenience, efficiency, and personalization. With continuous improvement in voice recognition and natural language understanding, it becomes more intuitive and user-friendly. Enabling the integration of ChatGPT with ESP32 microcontrollers has the potential to unlock a multitude of enhanced IoT applications, including chatbots, voice assistants, and natural language interfaces. We intend to harness the ChatGPT API within the ESP32 platform for this purpose. It's important to note that privacy and security concerns also exist with AI voice assistant, as it often require access to personal data and voice recordings



```
Writing at 0x000f72f7... (81 %)
Writing at 0x000fcc76... (83 %)
Writing at 0x00102839... (85 %)
Writing at 0x00108083... (87 %)
Writing at 0x00110797... (89 %)
Writing at 0x00118aeb... (91 %)
Writing at 0x0011e5cb... (93 %)
Writing at 0x00123840... (95 %)
Writing at 0x00128e06... (97 %)
Writing at 0x0012e635... (100 %)
Wrote 1181312 bytes (792106 compressed) at 0x00010000 in 11.5 seconds (effective 823.4 kbit/s)...
Hash of data verified.

Leaving...
Hard resetting via RTS pin...
```

Fig. 6. Experimental Setup

REFERENCES

[1] Shafeeg, A., Shazhaev, I., Mihaylov, D., Tularov, A., & Shazhaev, I. (2023). Voice assistant integrated with chat gpt. Indonesian Journal of Computer Science, 12(1)

[2] George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. Partners Universal International Innovation Journal, 1(1), 9-23

[3] Janokar, S., Ratnaparkhi, S., Rathi, M., & Rathod, A. (2023). Text-to-Speech and Speech-to-Text Converter—Voice Assistant. In Inventive Systems and Control: Proceedings of ICISC 2023 (pp. 653-664). Singapore: Springer Nature Singapore.

[4] Ram, B., & Pratima Verma, P. V. (2023). Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI. World Journal of Advanced Engineering Technology and Sciences, 8(01), 258-261.

[5] Abougarair, A. J., Aburakhis, M., & Zaroug, M. (2022). Design and implementation of smart voice assistant and recognizing academic words. International Robotics & Automation Journal, 8(1), 27-32.

[6] Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.

[7] Choi, J., Gill, H., Ou, S., Song, Y., & Lee, J. (2018, December). Design of voice to text conversion and management program based on Google Cloud Speech API. In 2018 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1452-1453). IEEE.

[8] Anggraini, N., Kurniawan, A., Wardhani, L. K., & Hakiem, N. (2018). Speech recognition application for the speech impaired using the android-based google cloud speech API. TELKOMNIKA (Telecommunication Computing Electronics and Control), 16(6), 2733-2739.