Voice-controlled Assistant with Chat LLM Integration

Ahmed Mohamed Galal
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt

Ahmed Osama
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt

Tony Begemy
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt

Ali Mohamed Ali
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt

Omar El Medani
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt

Abstract*— This project aims to develop a voice-controlled assistant integrating advanced speech processing and a large language model (LLM), enabling users to interact with the system through natural language voice commands and receive responses generated by the LLM. The methodology involves utilizing custom speech recognition algorithms, natural language processing (NLP) techniques, and text-to-speech (TTS) generation to ensure seamless communication between the user and the assistant. Experimental results demonstrate the assistant's high accuracy and efficiency in understanding and executing user queries, providing an intuitive and engaging user experience. This project contributes significantly to the fields of speech processing and machine learning by addressing current limitations and enhancing accessibility, productivity, and user satisfaction across various domains.

I. Introduction

Despite the increasing prevalence of voice-controlled assistants, many existing systems lack the flexibility and customization necessary to meet the diverse needs of users effectively. Current assistants often struggle with accurately understanding and processing natural language commands, especially in noisy environments or when dealing with uncommon words and accents. The rapid advancement of technology has heightened the demand for effective and user-friendly human-computer interaction methods. Voice-controlled assistants offer a promising solution by facilitating hands-free and natural interactions. However, to fully realize their potential, it is essential to overcome the limitations of existing systems, particularly in terms of speech recognition accuracy, natural language understanding, and user customization.

As voice-controlled technology becomes increasingly integrated into various sectors, including smart homes, automotive systems, healthcare, and education, enhancing the capabilities of voice assistants is crucial [9]. Our project addresses this need by developing an advanced voice-controlled assistant that leverages the latest advancements in speech processing and machine learning.

II. Related Works

Our project is situated within the context of a vast body of research spanning deep learning and speech processing domains. Over the years, significant strides

have been made in both areas, leading to transformative advancements in technology and applications.

In the realm of deep learning, neural network architectures have evolved from simple feed-forward models to sophisticated structures capable of handling complex tasks. Researchers have explored various architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [1], [2], to address challenges in speech recognition, natural language processing, and dialogue systems. Early research in speech recognition focused on traditional methods such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). However, with the emergence of deep learning, particularly deep neural networks (DNNs), there has been a paradigm shift towards end-to-end speech recognition systems that can directly map acoustic signals to text. Models like Deep Speech have demonstrated the efficacy of this approach, leveraging deep learning techniques to achieve state-of-the-art performance in speech recognition tasks [3]. Concurrently, advancements in dialogue systems have led to the development of conversational agents capable of engaging in natural language interactions with users. Generative hierarchical neural network models have been employed to build end-to-end dialogue systems, enabling conversational agents to generate contextually relevant responses based on user input [4].

Attention mechanisms have emerged as a key component in deep learning architectures, allowing models to selectively focus on relevant information while processing input sequences. This attention-based approach has been instrumental in improving the performance of various natural language processing tasks, including machine translation, text summarization, and dialogue generation [5].

Furthermore, techniques such as query-by-example keyword spotting using long short-term memory networks (LSTMs) have enabled efficient retrieval of spoken queries from large datasets, enhancing the usability of speech-based applications [6]. Additionally, neural belief trackers have been deployed in dialogue systems to track the state of conversations and generate more coherent responses [7].

Our project builds upon the foundations laid by these advancements, integrating cutting-edge techniques in deep learning and speech processing to develop a voice-controlled assistant. By leveraging state-of-the-art approaches and methodologies from a diverse range of research areas, we aim to create an intuitive and seamless interaction experience for users across different domains and applications.

What is missing is the work on Noise Cancellation and Wake Word Accuracy, while wake word detection is enhanced by deep learning, attaining high accuracy in noisy settings requires constant work.

There is an opportunity to improve upon accessibility features, such as integrating features that cater to users with disabilities. Consider voice commands for visually impaired users, or text transcripts of conversations for those with hearing impairments.

III. METHODOLOGY

The project aims to design, develop, and test a voice-command controlled virtual assistant prototype. This virtual assistant will be capable of understanding and executing voice commands to perform various tasks.
In order to accomplish our objective, there are multiple things to for us to consider:
We'll develop a system capable of recognizing speech by transfer learning and fine tuning the Whisper model without relying on existing APIs. This involves processing audio input to transcribe spoken words into text accurately.
The Whisper model is an advanced speech recognition system developed by OpenAI. Designed to understand and transcribe spoken language with high accuracy, Whisper leverages deep learning techniques, specifically transformer architectures, to process audio data. This model can handle various accents, dialects, and background noises, making it versatile for different environments [8].
We'll utilize various programming tools and libraries. Using Python programming language and utilizing the many libraries it has such as TensorFlow/Keras for training the model, also use of Tkinter to build a GUI, as well as using pyttsx3 for text to speech for the response. These tools will serve as the foundation for developing our virtual assistant system.
We'll design a flexible and scalable architecture for the virtual assistant, encompassing modules for speech recognition, and task execution. This architecture will ensure the system's scalability, maintainability, and adaptability to future enhancements.
We will leverage a publicly available speech dataset for training and testing our speech recognition model. We used Fleurs by Google [9] which provides a diverse dataset of human voices in multiple languages, which can be used to train the speech recognition model. The dataset includes recordings of people reading sentences, enabling us to build a robust speech recognition system capable of transcribing various accents and speech patterns. The model will only start listening to the user after hearing the stop word, after which the model will either do a command or answer the user's question using OpenAI's GPT model.

IV. DISCUSSION

Now, let us discuss how the project findings relate to the project goals and the available knowledge in the literature. The main objective was therefore to build a voice assistant that incorporates current speech technologies with an LLM with the objective of creating a natural interface. This was achieved through the use of customer speech recognition, natural language processing, and text to speech engines.

It majorly contributes to enhancing the overall capacity of the project in speech processing. This approach shows how the Whisper model can be fine-tuned and applied to achieve a high level of accuracy in interpreting queries and intonations, even in noisy settings or with accents. This helps to solve a major drawback in the existing systems and takes the development to the level that contemporary technology offers. Furthermore, the integration of an LLM such as the GPT makes the assistant a better place in establishing an appropriate response thereby enhancing the quality of the interaction.

The project is a versatile one as it creates an opportunity for several applications. In smart homes, the assistant can provide control over tasks via voice commands, which increases comfort. In automotive systems, it delivers voice control in a hands-free mode, thereby enhancing safety. In the healthcare sector, it helps practitioners by transcribing patient details through voice while in the education sector, it helps learners in their lessons. It versatility makes it applicable to all domains; productivity and satisfaction are improved.

However, there is some limitation with the current project which include; One key area is a complete dependence on high-quality audio input in order to achieve the best results. In terms of its strengths, one should mention that Whisper model performs stably with noise, although there are some issues with noise cancellation for its stable performance. Further,

integration with an LLM may bring issues of efficient computation and may not practically be deployable in scenarios with restricted computing power.

Some of the challenges that came up during the development process include the following. A major challenge was to maintain high recognition accuracy for different accents and background noises, which has been solved by applying the Whisper model. Another problem was the interaction between the speech recognition system and the LLM, which was solved at the architectural level. Making it easier for disabled users included the use of voice commands and text transcripts, which were still in the process of development.

In conclusion, the project offer significant improvement on the properties of voice-controlled assistants. Due to problematic areas in speech recognition and incorporating improved LLMs, it helps towards further developing human-Computer Interfaces. In future developments, more work will be devoted to increasing the accessibility and toward searching for new uses for the assistant.

V. Conclusion

In conclusion, this research demonstrates the potential of speech recognition, using fine tuning and transfer learning to build upon large speech recognition models such as Whisper, to create highly functional and user friendly voice controlled assistant that can answer your questions by being connected to a large language model such as the GPT model, and being able to understand given commands such as opening applications. We have developed a system that is capable of accurately transcribing speech and effectively handling natural language.
Our methodology, which involves custom speech recognition algorithms, natural language processing, and text-to-speech generation, ensures seamless communication between the user and the speech assistant enhancing the overall user experience, and being more accessible for wider users.

By improving accessibility, productivity, and user satisfaction across different domains, our work builds the way for more intuitive and engaging human to computer interactions.
Future work will focus on exploring the integration of more accessibility features to cater to users with disabilities, ensuring that our voice-controlled assistant is inclusive and beneficial to a broader audience [9].

VI. References

[1]Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016, March). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

[2]Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[3]Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

[4]Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, I.

[5]Chen, G., Parada, C., & Sainath, T. N. (2015, April). Query-by-example keyword spotting using long short-term memory networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5236-5240). IEEE.

[6]Goyal, T., & Durrett, G. (2020). Neural syntactic preordering for controlled paraphrase generation. *arXiv preprint arXiv:2005.02013*.

[7]Mrkšić, N., Séaghdha, D. O., Wen, T. H., Thomson, B., & Young, S. (2016). Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

[8]S. Wang, C. -H. Yang, J. Wu and C. Zhang, "Can Whisper Perform Speech-Based In-Context Learning?," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 13421-13425, doi: 10.1109/ICASSP48485.2024.10446502.

[9]Google/fleurs · datasets at hugging face. Available at: https://huggingface.co/datasets/google/fleurs

[10] R. Puviarasi1 , Mritha Ramalingam2 , Elanchezhian Chinnavan, “Low Cost Self-assistive Voice Controlled Technology for Disabled People”, Jul.-Aug. 2013 pp-2133-2138