Integrating Speech to Text, Text to Speech, and Data Access in ChatGPT

Vuppala Pavan Sai(A04275614)

Lekhan Guru Sai Dachepalli(A04276317)

Manmohan Reddy Karingula(A04283261)

Department of Computer Science

COSC-6340-W01, Human Computer Interaction

Texas A&M University, Corpus Christi

Abstract–In recent times we have seen the inception of LLM (Large language models), these models use deep learning techniques particularly GPT (Generative Pretrained Text). The large models like GPT have many parameters which enables them to capture all the complex patterns in the data that is given, the data that is being given to the model is in english language. then the model works through the data to generate the output, hence these large models perform very impressively in a question-answering activity. The model is first pre-trainded with vast amounts of data often from various internet sources that tends to be very informative and large in size, and this pre-training helps the model learn patterns. The trained model is used by the end user to generate the answers for the questions given by the end user. Our research and proposal aims at eliminating the barrier for users who are unable to type in the question for LLM models particularly the Chat GPT to interact, by integrating Speech to Text and Text to Speech which aids people with various disabilities to interact with.

1. INTRODUCTION

In recent times we have seen the inception of LLM (Large language models), these models use deep learning techniques particularly GPT (Generative Pretrained Transformer). The large models like GPT have many parameters which enables them to capture all the complex patterns in the data that is given, the data that is being given to the model is in english language. then the model works through the data to generate the output, hence these large models perform very impressively in a question-answering activity. The model is first pre-trained with vast amounts of data often from various internet sources that tends to be very informative and large in size, and this pre-training helps the model learn patterns. The trained model is used by the end user to generate the answers for the questions given by the end user. The chatbot systems like GPT-3, BERT, TransformerXL, XLNet, T5, DialoGPT, RoBERTa, ALBERT Turing NLG etc. which uses Artificial Intelligence (AI) to engage in conversations with the end users. Our research and proposal aim at eliminating the barrier for users who are unable to type in the question for LLM models in particularly the chatGPT to interact, by integrating Speech to Text and

Text to Speech which aids people with various disabilities to interact with.

ChatGPT is an Artificial language model which is developed by OpenAI and was introduced to the market in 2020 and has been trained on the information until September 2021 as it was the limited training period of the model. Its architecture is based on GPT-3.5 which is the 3rd Generative pre-trained model which has the capability of answering questions, generating texts and interactive behavior with the user accordingly and generating responses. These text-based responses are generated by utilizing the NLP algorithms which are trained on the vast amount of data sets and learning patterns and language representation which can predict and generate the outcome of the given context. In ChatGPT it can generate the answer in the form of text manner..

In present industry communication is the key to progress and this communication can happen by any means whether it can be through verbal communication through Cell phones over calls and it is also a crucial means of communication device where an user can contact others using emails, text messages, etc. Over past few years many innovations have been taking place in SMS technology where user can generate text using voice messages in form of text using speech recognition technology, keeping in mind of disabled persons to assist them the usage of few applications like Speech to Text (STT), Text to Speech (TTS) and speech translation.

Speech-to-text is a technology that converts spoken language into written text. The system converts users speech into written words while allowing for natural speaking. Applications for this technology include voice-activated gadgets, voice assistants like Siri and Alexa, and numerous others. To convert spoken words into text, speech-to-text technology first separates the words' sounds. The corresponding text is then generated after these sounds are compared to a database of recognized words.

Text-to-speech is the reverse process of speech-to-text. It involves converting written text into spoken language. Text-to-speech systems take text input and generate synthetic speech output that sounds human-like. Text-to-speech technology is used in applications like voice assistants which uses Nuance Communication to convert Text-to-Speech, audiobooks, and accessibility features in electronic devices.

2. MOTIVATION

In our research proposal we are trying to integrate Text-To-Speech and Speech-To-Text with the Large Language Model ChatGPT for several reasons but we are here mainly focusing on two reasons the first one is it can provide an advantage of listening to the output by doing this people who are visually impaired can understand the output more clearly and precisely and the second one which are trying to propose is the functionality where it can help the ChatGPT to dictate instead of typing which will be very helpful to the people who face difficulty in typing which makes easier.

As the current version of the ChatGPT has the knowledge cutoff till 2021. Which means the AI model has the ability to answer the questions until 2021 where the later data after 2021 is not trained to the model so in our research proposal an update that integrates the ChatGPT to access and understand data in the PDF or document formats so that it can be trained based on the present data available which helps the model to stay updated about the information and updates.

There are many benefits of integrating the model with data access via PDF or documents. One is users can receive more accurate information interacting with ChatGPT, more valuable and reliable. The other reason is the ChatGPT making the model stay up to date with the information. This can be achieved by Providing the data either in documentation or pdf format so the data is inputted and whenever it is asked about the new data the data is analyzed and the responses are out.

3. LITERATURE REVIEW

There are many models and approaches to find accurate results that have been researched and developed over many years for the Speech-to-Text (STT) and Text-to-Speech (TTS). In STT the traditional methods like Hidden Markov models(HMMs) and Gaussian Mixture Models (GMMs) were commonly used in the early days they require a extensive manual engineering as they had limitations and had to be handled manually for the pronunciations and accents as the model is relied on acoustic features, linguistic models, and language-specific rules[1].

With the rise of the convolutional neural networks (CNNs), recurrent neural networks (RNNs) and deep learning revolutionized STT. Later the feature engineering was unnecessary as End-to-end models like Listen, Attend, and Spell (LAS), Connectionist Temporal Classification (CTC) emerged. In our research we are using Whisper for STT for a reason as Whiper's is an open-source automatic speech recognition system which is an open source developed by OpenAI's. It employs the Whisper ASR architecture, which is based on a hybrid approach that combines CTC and Transformer-based models which ensures and adapts to the two different languages and accents. This makes a strong competition in the ASR landscape.

In early traditional Text-to-Speech models they employee rule-based methods, which are prerecorded and linguistic speech segments that are combined to from an sentence. Because of the limitations of rule-based approaches, these systems were frequently limited in their naturalness and expressiveness. The introduction of statistical parametric methods transformed TTS. These models, such as Hidden Markov Models (HMMs) and Unit Selection, enabled more natural-sounding speech synthesis by concatenating speech units. However, they frequently required significant data and expert knowledge to achieve satisfactory results. As the technology has been advancing there is also an improvement in the Text-to-Speech which have been driven by the neural networks based techniques like WaveNet and Tacotron. WaveNet is a generative model which was developed by DeepMind that directs raw audio waveforms which means it can generate speech in both natural and realistic way, for more complex sounds like environment etc.Tacotron on other hand is another neural network based technique, it is an sequence -to -sequence based model that means it takes input as an model and generate output as a speech. Inorder to do this there are several steps first text gets converted to spectrogram which is nothing but the visual representation of the sound waves, now the neural networks is used to generate corresponding speech waveform, it is mainly used in commercial products including amazon Alexa and Google's assistant.

With the rapid growth in computer hardware and software speech recognition technology is becoming a major technology in various devices, a diverse group of disciplines like information theory, acoustics, pattern recognition, phonetics, linguistics, and neuroscience are involved in speech recognition technology.

During 50's the research work for speech recognition began at Bell Labs Speech recognition system-Audrey where they first identified the ten English digits and here after the speech recognition made great progress, further in 1980 speech recognition used HMM model for the first time and was successfully used. The VQ/IIMM approach is used by FULEE Kai and others to create the Continuous Speech Recognition System-SPHINX, It is a high-performance, continuous speech recognition system with a broad vocabulary that is the first of its kind in the world. Finally, people are able to overcome the three main challenges of a big vocabulary, continuous speaking, and non-specific. Additionally, it identified the dominant statistical techniques and models used for language processing and speech recognition[4].

Speech recognition technology has already advanced from the lab to the real world; there are more established market products. Many wealthy nations, including the United States, Japan, and South Korea, as well as well-known corporations like IBM, Apple, Microsoft, AT&T, and others, have made significant investments in the study and development of useful voice recognition systems[4].

The speech recognition system, which includes feature extraction, pattern matching, and the reference model library, is fundamentally a pattern recognition system. composition.

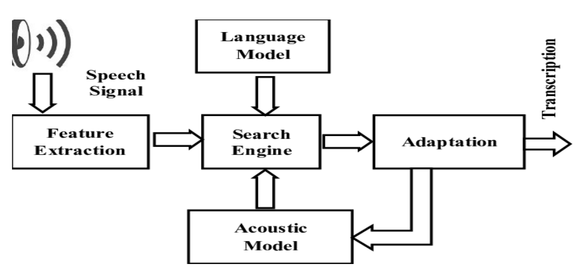

Figure 1: Speech Recognition Architecture

The process of turning unprocessed audio signals into a manageable representation that the model can use is known as feature extraction in speech recognition models. A concise representation of the voice waveform is sought after throughout the feature extraction stage. This format should offer a good match with the distributional assumptions made by the acoustic models and limit the loss of information that distinguishes between words.

A search engine is a piece of software that looks through a database of websites to locate those that correspond to a user's query. Search engines are employed in speech recognition models to locate pertinent information about the subject being spoken.

A statistical model that assigns probabilities to word sequences is known as a language model. Language models are employed in voice recognition models to foretell which words will most likely follow one another in a specific situation.

The link between an audio signal and the phonemes or other linguistic components that make up speech is statistically represented by an acoustic model. Acoustic models are used in speech recognition models to identify spoken words by comparing them to the phonemes or other linguistic units contained in the model. The process of changing a current acoustic or linguistic model to better match fresh data is referred to as adaptation in the context of voice recognition. The model can be retrained using fresh data, or strategies like speaker adaptation or domain adaptation might be used. Some useful voice recognition systems have been put into commercial operation in an effort to accelerate the world's research and development of speech recognition applications.

AT&T created the standard speech recognition system-VRCP system in 1992.The system, which uses the five words "collect," "person," "third number," "the operator," and "calling card," as well as a general small-vocabulary speech recognition system, is used in AT&T Communications' online services to enable automatic operator assistance in place of the operator's completion of five different call types[4].

The first extensive commercial speech recognition application system, the stock quote system, was introduced by Charles Schwab in September 1996. The speech recognition system was also the first in the finance industry. The solution works well to lower call center costs while increasing customer happiness and service quality. Schwab soon started his explanation on the stock trading system.

Sprint PCS, a significant U.S. telecom provider, has the largest digital wireless network at the same time and is renowned for its excellent and cutting-edge customer care.Since 2000, voice-driven solutions have been available to customers. Customer support, voice dialing, number verification, address changes, and other services are offered via the system. Additionally, China Telecom has introduced the CELL-VVAS (VOICEVALUE-ADDED SYSTEM), a voice recognition integration of value-added services system that employs a distributed superb recognition engine and has produced a reliable and effective application. In order to offer users a choice of user-friendly, individualized services, the system includes flawlessly integrated telecoms switching network application.

Bell Labs is a leader in this field and is developing telephone voice recognition technology, which is another branch of speech recognition technology. This technology will aid people with telephone inquiries, automatic wiring, and some of the specialized tasks like tourist information and other tasks. The bank will be able to offer customers 24-hour Phone Banking Service once it has implemented the voice query system with speech recognition technology. The user can speak out the stock name or code to validate their requirements in the securities business using a telephone speech recognition audio system, and the system will automatically read the most recent stock price, substantially assisting the user.

By allowing the computer to automatically respond to consumers' demands and then playback the phone number of the inquiry, the artificial voice technology used in the 114 directory assistance allows you to save human resources.

Text to speech conversion is done using speech synthesis technology. It's a piece of technology that lets computers speak with a voice that sounds like ours. The technology has been around for a while and has been put to use in many different applications, including virtual assistants, automated phone systems, and navigational systems. With the advancement of deep learning and artificial intelligence, speech synthesis technology has advanced significantly in the recent years. The quality of synthesized speech has substantially increased recently thanks to neural network-based speech synthesis.

Digital signal processing (DSP) and natural language processing (NLP) are the two fundamental components of speech synthesis. NLP is the process of transforming text into a format that the speech synthesis system can understand. Text analysis, morphological analysis, contextual analysis, syntactic-prosodic analysis, and semantic analysis are just a few of the jobs that fall under this category. The output of the NLP system is transformed using DSP into an audio signal that can be played back through speakers or other audio equipment[2].

Digital Signal Processing (DSP) is the use of mathematical algorithms to transform signals from analog to digital form and vice versa. It involves the manipulation of signals using various techniques such as filtering, modulation, and compression. DSP is used in a wide range of applications such as audio and video processing, speech recognition, image processing, and control systems. DSP has revolutionized the way we process and analyze signals by providing faster and more accurate results than traditional analog methods. It has become an essential tool in many fields of engineering and science[2].

A branch of computer science and artificial intelligence called "Natural Language Processing" (NLP) is concerned with the use of natural language in communication between machines and people. It entails the creation of models and algorithms that give computers the ability to comprehend, decipher, and produce human language. NLP is utilized in many different applications, including speech recognition, sentiment analysis, chatbots, and machine translation. NLP has transformed how we communicate with machines by providing more natural and intuitive interfaces. It has become an essential tool in many fields of engineering and science.

A technique for producing voice signals is called acoustic speech synthesis. Speech synthesis aims to produce speech in such a way and of such a quality that it closely resembles human speech (sometimes even the voice of a specific individual); this includes not only the voice itself and its quality but also the style of speaking, etc. Text-to-speech (TTS) technology is used to generate speech automatically by a machine. Its job is to translate any text into the appropriate speech. TTS can therefore be thought of as a collection of unique modules and algorithms that perform automatic translation of written text into speech. The modules cover voice creation as well as word processing (such as analysis and normalization), text to pronunciation conversion (such as phonetic transcription and prosody generation), and text conversion. Two methods dominate speech creation at the moment:

Signal Based Approach: Signal-based technique, with unit selection being a well-known example of this approach, in which the final speech is formed by the concatenation of appropriately chosen speech segments (speech units) such context-dependent phones, diphones, or halfphones[7];

Model Based Approach: model-based method, where speech is produced using (speech production) models; the most well-known example of this is statistical parametric synthesis, or hidden Markov model based synthesis, in which speech is produced from statistical models[7].

Users with the following populations may gain from speech recognition systems for learning:

1. Disabilities in learning, such as dyslexia and dysgraphia
2. damage caused by repeated strain, such as carpal tunnel syndrome
3. insufficient or poor motor skills
4. vision problems
5. Physical limitations
6. Language Barriers in English

With the inclusion of speech recognition and synthesis students may take advantage of the LLM's ability to take audio input and give back audio output making it more convenient for disabled people in using the LLM.

In our research we have found many articles about how Chatgpt uses its OpenAI servers to gather all the information after finding all the complex patterns, but we came across an article on how to build a LLM that only reads a local file and answers the question in context to the file. The article called "How to create a private ChatGPT that interacts with your local documents" demonstrates how to build a LLM that uses the given documents and answers questions regarding the file.

4. EXPERIMENTAL SETUP

For the experiments, we have used google colab as IDE to run all the experiments and we have used libraries such as openAI, and langchain to retrieve the model and use it. We have used variety of models by openAI, for speech recognition, completion and embeddings. We used TikToken, OpenAI, and Langchain to provide capabilities for working with large language models like GPT-3. TikToken efficiently tokenizes text for us. We used OpenAI as the API client. Langchain simplifies chaining models for us. We leverage Faiss and OpenAI embeddings to enable creating vector indexes of text for efficient retrieval and contextual ranking. This provides the knowledge source to ChatGPT. We utilize Whisper, gTTS, and Gradio to provide the speech recognition, synthesis, and interface components. Whisper transcribes audio to text for us. gTTS converts text to speech for us. Gradio creates the web UI for us.

The code for our research can be found in the repository[10], the repository consists of

5. METHODOLOGY

i) Giving data access to ChatGPT:

In our experiment, we would like to introduce data access to ChatGPT, where we allow ChatGPT to access through data files given by the user. This experiment enhances the capabilities of ChatGPT by providing additional textual context that it can reference when answering questions. The methodology involves setting up a pipeline that ingests text files, embeds them into vectors, indexes the vectors for fast retrieval, and makes those retrievals available to ChatGPT.

The first step is preparing the data. We gather relevant text files containing contextual information into a directory. These files are in unstructured text format across various topics that may be useful for ChatGPT. To process the text, we leverage the langchain library's tools for embedding and retrieval. First, the DirectoryLoader loads the text files into memory as Document objects. Then, the RecursiveCharacterTextSplitter splits the Documents into smaller Text chunks of 1000 characters each, with no overlap between chunks.

Next, the OpenAIEmbeddings class embeds the textual chunks into high-dimensional vectors by interfacing with OpenAI's text embedding API. The FAISS vector store indexes these text embeddings for fast nearest neighbor lookup. At query time, we can quickly find the most contextually relevant text chunks. With the index built, we construct a RetrievalQA pipeline in langchain. This chains together the components needed to retrieve texts and query the LLM. As the ChatGPT uses gpt-3.5-turbo model for completions, we load the OpenAI GPT-3.5 Turbo model via api key which can be found by openAI account and configure the pipeline to search over the FAISS index when running queries. To use this context-enhanced ChatGPT, we simply provide a query question and the RetrievalQA chain automatically retrieves related texts from the index and passes them along with the question to ChatGPT. This provides supplementary contextual information that ChatGPT can leverage when formulating its response, enhancing its knowledge. The result is an augmented conversational agent that accesses additional textual context beyond just its pre-training.

ii) Giving audio input and output to ChatGPT:

Enabling audio-in and audio-out capabilities for conversational AI systems like ChatGPT requires integrating speech recognition and speech synthesis technologies. This allows users to interact in a natural, voice-driven manner rather than via text.Our methodology involves building a pipeline that transcribes incoming audio to text that can be input to ChatGPT. Then, we synthesize ChatGPT's text response back into human-like speech.

For audio input processing, we use the Whisper open-source speech recognition framework. Whisper provides pre-trained models for speech-to-text transcription including support for multiple languages. We load the Whisper "base" model which works well for short-form audio like conversational snippets.

When our system receives audio data, we first load it into memory using Whisper's audio loading utility. This converts the raw audio into a format usable by Whisper models. Next, we pad or trim the audio as needed to fit Whisper's required input length.We extract log-Mel spectrogram features from the preprocessed audio clip, which represent the frequency content in a compact form. This spectrogram is passed to the Whisper model to detect the spoken language, then decode the speech into text. The decoding step leverages beam search to generate the most likely text transcription.

The resulting text is then sent to ChatGPT to generate a response. We use the OpenAI API to query the ChatGPT model, providing the Whisper-transcribed text as the prompt. This allows ChatGPT to produce a relevant response as if the text was typed.

To synthesize the text response back into natural sounding speech, we utilize the gTTS Python library. gTTS leverages Google's text-to-speech pre-trained models supporting multiple languages. We call gTTS to convert ChatGPT's text into an audio file.The synthesized audio, along with the original input text and ChatGPT's response, are returned by our pipeline. This end-to-end feature enables the user to have voice-driven conversations.

For the development of front-end, we have used the Gradio library to create an easy-to-use web interface. Gradio allows quickly building UIs with Python using pre-packaged widgets. We use a microphone input to capture voice queries and an audio player output to play back ChatGPT's responses. When deployed, a user can simply speak into their microphone and hear ChatGPT's reply instantly. The integration of Whisper and gTTS models into the pipeline makes the conversational interaction feel natural and human-like.

Our methodology provides a template for augmenting text-based conversational agents with voice capabilities. The modular design connecting speech modules with large language models can extend these techniques to other domains beyond ChatGPT.

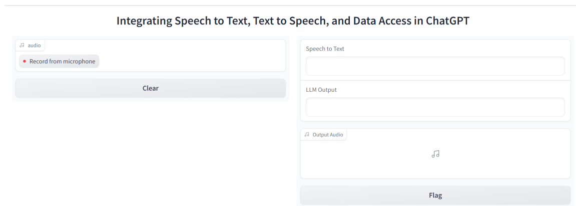

Figure 2: UI of the extended version of ChatGPT.

6. EXPERIMENTS & RESULTS

For the experiment, we need to compare with the existing system, here we have taken a research paper as example and we have loaded it to the model for the reference. The paper is divided into chunks and if there is need of any relevant information to the question, then the model(ChatGPT) refers to the paper and pulls out all the necessary details and adds as the context to the existing question, to take it as the reference and answer it.

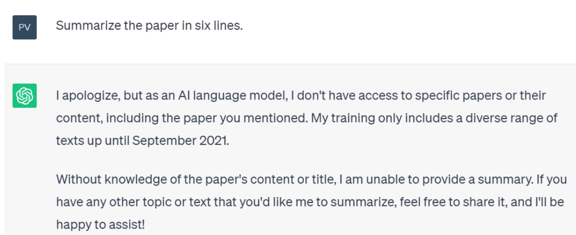

Figure 3 : ChatGPT without enhancements

Given are the two figures(Figure 3 and Figure 4), one is the ChatGPT, without data access feature, when asked the question"Summarize the paper in six lines." to two of them, the results of the existing system is "I apologize, but as an AI language model, I don't have access to specific papers or their content, including the paper you mentioned. My training only includes a diverse range of texts up until September 2021."

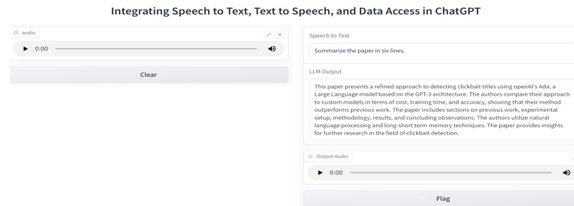

Figure 4: Enhanced ChatGPT results.

but the proposed system's output is : "The given paper presents a refined approach to detecting clickbait using openAI's Ada, which is a Large Language model based on the GPT-3 architecture. The authors claim that their approach easily outperforms custom models in terms of cost, training time, and accuracy. The paper discusses the previous work on classification of titles as clickbait or non-clickbait, followed by the experimental setup and methodology used in their research. The results of their study are presented in a separate section. Finally, the paper concludes with some observations and directions for further research."

As you can the clear difference of the response of the two systems. For the existing system, the answer given was, that it doesn't have access to the paper provided, where the proposed system has given access to the paper and the model is able to refer and answer to the question efficiently. And also, here the input is given as audio and it is converted to text and later it is feeded to the model and the output generated which is known as completion is again converted to speech and the audio file is given as output where it can be played many time for the user to understand it easily.

7. ADVANTAGES

The combined ability of LLM to use speech recognition and speech synthesis paired with access to local documents will be very useful to a wide variety of applications and users have a very good impact in terms of interaction and functionality that will not only aid disabled people but also enhance the productivity of users in daily life.

The following are the features added

1. Speech recognition ability for ChatGPT.
2. Speech delivering ability for ChatGPT
3. Data access to local files with in ChatGPT.

8. CONCLUSION

From the above experiments and results we can clearly see that the interaction between a human and a LLM (Large Language Model) can be taken a step ahead with the speech recognition, speech synthesis , and the data access will also provide the user a more sophisticated experience as it now has the ability to read and look into users local documents this is also secure when the user doesn't want to share the document online and want to get the insights of the document on the go.

9. REFERENCES

[1] J. Solomon, "Speech synthesis techniques," 1981 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, New York, NY, USA, 1981, pp. 98-99,

[2] Y. Tabet and M. Boughazi, "Speech synthesis techniques. A survey," International Workshop on Systems, Signal Processing and their Applications, WOSSPA, Tipaza, Algeria, 2011, pp. 67-70,

[3] Z. Yin, "An Overview of Speech Synthesis Technology," 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 2018, pp. 522-526,

[4] J. Meng, J. Zhang and H. Zhao, "Overview of the Speech Recognition Technology," 2012 Fourth International Conference on Computational and Information Sciences, Chongqing, China, 2012, pp. 199-202,

[5] Yu Tiecheng. The current development of speech recognition [J]. Communication World, 2005.

[6] Ren Tianping. Application of speech recognition technology [J]. Henan Science and Technology, 2005.

[7] Acoustic Speech Synthesis | Acoustic speech synthesis :: Department of Cybernetics, UWB (zcu.cz)

[8] D.H. Klatt "Review of text-to-speech conversion for English," Journal of the Acoustical Society of America, vol. 82(3), 1987.

[9] How to create a private ChatGPT that interacts with your local documents| https://bdtechtalks.com/2023/06/01/create-privategpt-local-llm/