Voice-Driven AI Counselor: Integrating Real-Time LLM Interaction for Personalized Assistance

Yassine Ibork, Kshirsagar Shruti

Wichita State University

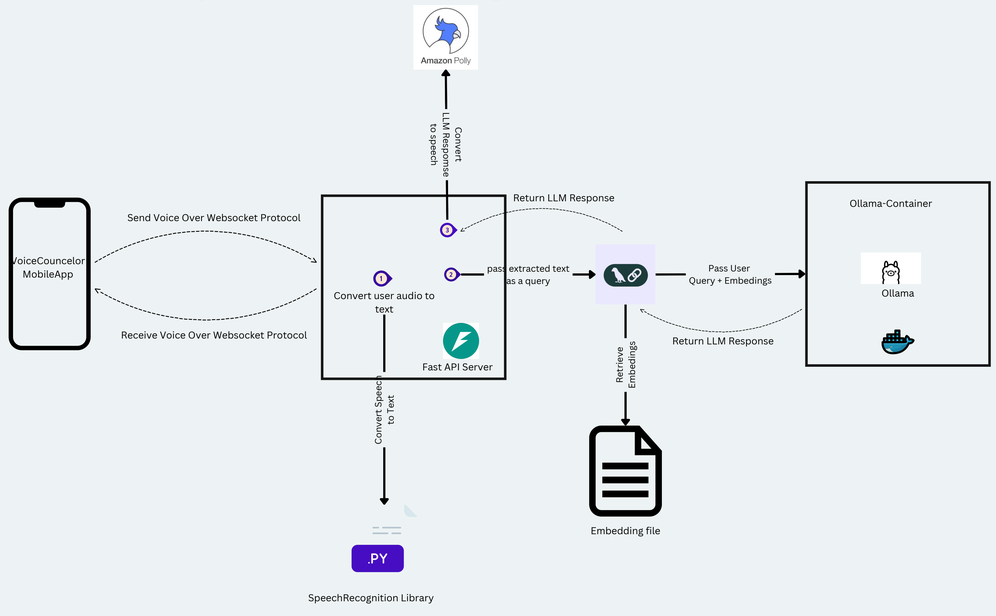
Introduction

This project presents a voice-based chat counselor that enables real-time, conversational interactions with a Large Language Model (LLM), simulating a counselor experience through audio. The system integrates React Native for cross-platform support, FastAPI for backend operations, Amazon Web Sevices (AWS) Polly for text-to-speech, and LangChain for managing LLM queries. The initial implementation is operational, proving the feasibility of a voice-driven support tool. Future work will focus on optimizing text-to-speech by evaluating locally-run pre-trained models to enhance accuracy, speed, and cost-efficiency, improving the overall user experience.

Methodology

The architecture comprises the following components:

1. **Mobile App (React Native):** Captures user audio and sends it to the backend via WebSocket for processing.
2. **FastAPI Server:** Converts audio to text using SpeechRecognition, constructs and sends queries to the LLM, and receives responses. The responses are converted to audio using AWS Polly and sent back to the app.
3. L**LM (Ollama Container):** Processes queries with the help of LangChain for structured query management, using embeddings from the Psych8k dataset for Retrieval-Augmented Generation (RAG).
4. **Audio Processing:** The SpeechRecognition library captures and converts user speech into text, enabling voice input functionality. This setup allows users to interact naturally, with real-time responses facilitated by a robust backend(progress report).



Background

Previous work on developing a chat-based counseling agent, as discussed in [3], introduced an LLM-based solution aimed at supporting users with mental health concerns. While effective, this approach relies solely on text, missing the added depth that spoken interaction can provide, as many people find it easier to express emotions and feelings through speech. To build on this, our project integrates a text to speech and speech to text systems with an LLM, enabling users to communicate verbally, thereby creating a more natural and emotionally resonant experience.

Results

The project successfully implements the Retrieval-Augmented Generation (RAG) architecture within the large language model framework, demonstrating its potential for delivering personalized mental health support. Word embeddings generated from the Psych8k dataset [2] were effectively utilized in the RAG process, enabling the LLM to provide contextually relevant and tailored responses to user queries. Additionally, the integration of AWS Polly allowed for natural-sounding text-to-speech conversion, while a speech recognition library enabled seamless speech-to-text processing. These components together create a robust, interactive voice-based system that enhances user engagement and accessibility.

Conclusion and Future Work

Our Work Combines various technologies to create an interactive voice-based chat counselor. While the core functionality is established, further enhancements are necessary, particularly in the TTS module. Future work will involve testing pre-trained TTS models locally to inform decisions on the best TTS system based on accuracy, cost, and speed, using the LJ Speech dataset for comprehensive evaluation[1].

References

[1] “The LJ Speech Dataset.” Accessed: Sep. 08, 2024. [Online]. Available: https://keithito.com/LJ-Speech-Dataset

[2] “EmoCareAI/Psych8k · Datasets at Hugging Face.” Accessed: Sep. 08, 2024. [Online]. Available: https://huggingface.co/datasets/EmoCareAI/Psych8k

[3] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, “ChatCounselor: A Large Language Models for Mental Health Support,” Sep. 27, 2023, arXiv: arXiv:2309.15461. doi: 10.48550/arXiv.2309.15461.