ECE 445

SENIOR DESIGN LABORATORY

PROJECT PROPOSAL

AMADEUS

Augmented Modular AI Dialogue and Exchange User System

Team No.33

Qiran Pang (qpang2@illinois.edu)

Chengyuan Peng (cpeng14@illinois.edu)

Ryan Fu (ryfu2@illinois.edu)

TA: Jason Zhang

Professor: Cunjiang Yu

October 1, 2024

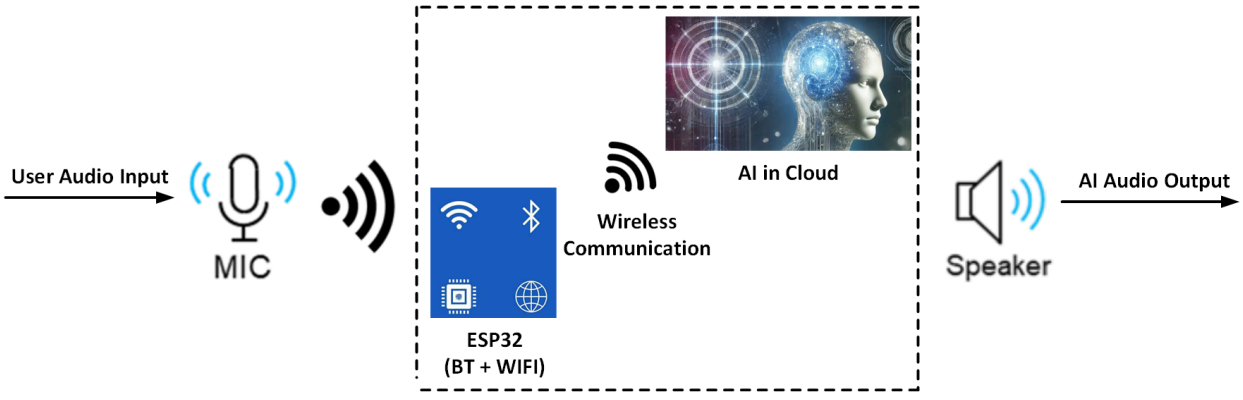
Introduction:

Problem:

People have envisioned engaging in natural, conversational interactions with robots for many years to fulfill emotional and lifestyle needs. However, most current interactive AI systems are either too bulky or rely heavily on smartphones, detracting from the organic nature of such interactions. A more tangible, interactive medium—such as a child talking to a familiar toy or a headset with built-in AI—would offer a more immersive experience, which corresponds to an increasing market need[1]. Embedding a trained AI model in each toy or device would be cost-prohibitive since it would require powerful and expensive embedded computers. To address this, we propose leveraging cloud-based AI models, such as ChatGPT or similar character-driven AI, in our embedded system, which can process data remotely and send responses back to the device in real time.

Solution:

We aim to develop an AI-based audio interactive interface, housed on a custom-designed PCB. This system will capture audio from the user, transmit it via Wi-Fi to a cloud-based AI model for processing, and play the AI's response back to the user. The ESP32 microcontroller, equipped with Wi-Fi and audio input/output capabilities, will serve as the core of our system.



High-Level Requirements:

- **Response time:** The AI model should receive audio input from the user within **5 seconds**, process it, and send a response back to the PCB within **5 seconds** (response time may vary depending on the chosen AI model and internet speed).

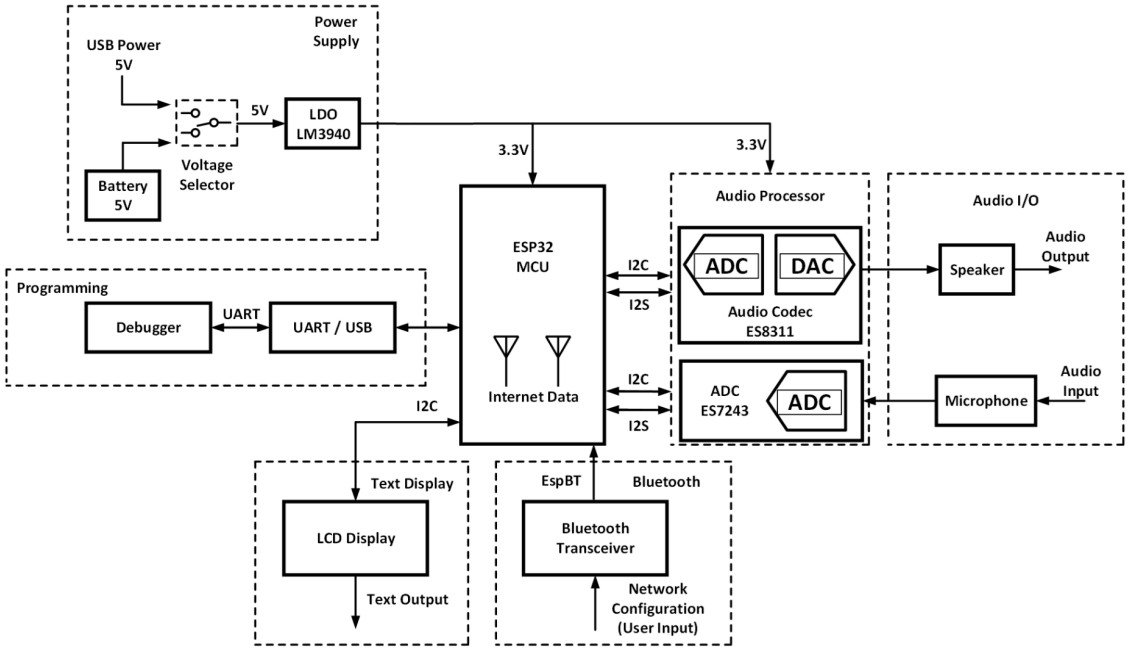
- **Voice Clarity**: The AI-generated audio must be **clear and audible** to the user, with a Signal-to-Noise Ratio (SNR) larger than 20 dB.
- **Multi-language support:** The system will support voice input in **three different languages**: **Chinese, English, and Japanese**.

Additional features:

- **Indoor and outdoor modes**: In outdoor mode, audio input will be processed only when a button is pressed, and the system will apply noise reduction to improve voice clarity.
- **Headphone/Bluetooth integration**: This feature will allow users to interact with the device using wireless headphones or earbuds.
- **Text Display**: The PCB will include a small display to show transcribed audio, offering a visual representation of conversations for users.

Design:

Block Design:



Subsystem Overview & requirements:

##Subsystem 1: AI Web Client

Overview:

Our language model will be hosted on a cloud-based server. The local MCU will transmit audio to the server via a WiFi module. We are collaborating with a local start-up that will provide some AI models[2] for us. However, we also have the option to train our own AI model to create additional characters using their interface or connect with other available AI models online such as ChatGPT.

Interaction:

The AI web client mainly interacts with the WiFi module of the ESP32 board to receive and send audio and text signals.

Requirements:

- The AI should respond with an average latency of **no more than 10 seconds** for a 10-second audio input, ensuring the user receives the reply promptly.
- AI models should support language inputs in **English, Chinese, and Japanese**.

##Subsystem 2: ESP32 with Wifi Capability

Overview:

The ESP32 with Wi-Fi capability serves as the core processing unit for the entire system. It receives audio input from the microphone via the ADC, transmits the audio to the cloud-based AI model via Wi-Fi, receives the processed audio response from the AI model, and then sends the output to the audio codec for playback. This subsystem is also responsible for interfacing with the Bluetooth module for Wi-Fi configuration and managing communications with the text display for visual output.

Interaction:

The ESP32 is the core of the system so it needs to interact with all other subsystems: audio I/O system by receiving and sending audio signal through I2C and I2S, AI web client through WiFi, LCB

display to display text through I2C, Bluetooth module to receive WiFi configuration, debug module to be debugged through UART, and power supply system to receive 3.3 V power.

Requirements:

- The ESP32 must establish a Wi-Fi connection **within 10 seconds of receiving valid credentials** and maintain a stable connection with >99% uptime during operation.
- The ESP32 must be able to transmit and receive audio data **at a minimum bitrate of 64 kbps** to ensure acceptable audio quality.

##Subsystem 3: Power System

Overview:

The system can be powered through either a USB connection or a 5V battery. The 5V supply directly powers the I/O devices and the programming module. To provide 3.3V power for the microcontroller and audio processing module, an LDO voltage regulator is used to step down the voltage.

Interaction:

The power supply system interacts with other subsystems by providing Vcc power to other components, such as the ESP32 core and audio I/O processor.

Requirements:

- The power system must be able to use a **5V power supply** source and **power all the modules** in our system with their respective power requirements (5V or 3.3V).

##Subsystem4: Bluetooth Communication

Overview:

A Bluetooth transceiver module will be connected to the ESP32 processor to receive user input for configuring the internet connection. The user will transmit the internet passcode to the Bluetooth transceiver, which will then relay this information to the microcontroller to establish the connection.

Interaction:

The Bluetooth module interacts only with the ESP32 microcontroller to configure the Wifi for network connection.

Requirements:

- The Bluetooth module must successfully receive credentials and send it to the cpu **within 5 seconds** after the user sends it out through Bluetooth.

##Subsystem5: Audio I/O & Processing

Overview:

The Audio I/O & Processing subsystem is responsible for capturing the user's voice through the microphone, converting the analog audio signals into digital form using an ADC, processing these signals, and then sending the digital audio to the ESP32 for transmission. Once the AI response is received, it is converted from digital to analog form using a DAC or Audio Codec, and the output is played through a speaker.

Interaction:

The Audio I/O & Processing subsystem interacts on one side with the user by receiving audio from the microphone and playing audio through the speaker and interacts with the ESP32 microcontroller through I2C and I2S to communicate and send signals. In addition, it also receives power from the power supply module.

Requirements:

- The microphone must have a sensitivity of **at least -42 dBV** and a **frequency response range of 20 Hz to 20 kHz** to capture a full range of human speech clearly.
- The audio processing circuit must maintain an **SNR of at least 20 dB** to ensure clear audio input and output.

##Subsystem6: Text Display

Overview:

If we have more time after finishing the baseline, an additional feature of our project will be a text display LCD. After the audio input / output are converted into texts, the LCD screen attached to the microprocessor will display the text output. It will ideally display both the input from user and output from AI on the LCD screen so users can make sure their audio is identified correctly while reading the response from AI when they did not hear the audio clearly.

Interaction:

Through I2C, the LCD text display interacts with the ESP32 microcontrollers by receiving a text to be displayed on the screen for the user to see,

Requirements:

- The LCD display resolution must be at least **128x64 pixels** so user can clearly see the text
- The LCD should display the text output within **2 seconds** of receiving data, ensuring synchronization with audio playback.

##Subsystem7: Debug Module

Overview:

The debug module will consist of a debugging serial port and a programmer. The serial port will be temporarily integrated into the PCB for debugging the output from the ESP32 processor through UART. Additionally, a programmer will be connected to the MCU for programming purposes through USB.

Interaction:

The debug module is used by the user to debug and program the ESP32 microcontroller through UART and USB.

Requirements:

- The debug module must support UART communication at **115200 bps**, with error detection and handling to ensure reliable data transmission.

Tolerance Analysis:

A common challenge that embedded system designers frequently encounter is insufficient storage space of the processor, especially if the system is related to acoustics. The audio files will typically be large enough to occupy a large portion of the flash memory.

Suppose a 30s audio data is sampled at a rate of 44.1kHz, memory overflow can easily happen:

Along the 30s duration, the total number of samples will be:

$$30 \times 44.1k = 1323$$

Also, suppose each sample is of type int (4 bytes), the total number of bytes occupied by this sample will be:

$$1323 \times 4 = 5292 = 5.16 \text{ MB}$$

As there will be both audio inputs and outputs that are processed by the MCU, the total size of the two audio samples will be $5.16 \times 2 = 10.32$MB. 10MB is an incredibly large size to be processed – even modern laptop cache can hardly meet this requirement. Not to mention ESP32 processors are much less powerful than a complete computer system.

Given that the best ESP processors only have an internal memory of 4MB with half of the storage already occupied by built-in libraries, we will have very limited space to store the audio inputs and outputs. As such, we must have a smarter implementation to reduce the sizes of the audio samples. We have come up with the two following solutions:

1. Instead of storing the entire file in the flash, we could use the flash as a buffer. In particular, we plan to use the flash as a buffer, only to store 5ms of the audio each time and send it to the cloud. In that case, the buffer only needs to use about 30b for the audio storage.

2. A 32-bit audio sample is way more than enough to produce an audible audio output, hence we should not waste our memory on unnecessary data. It is likely that we can compress all int32 audio samples into int8 forms, thus decreasing the data size from 10.32MB to 2.58MB. 2.5MB is sufficient to be stored into a MCU without any external memory components.

The potential risk of this project is that ESP32's flash size is 2 - 4 mb depending on the model of the chip, which is relatively small for storing audio files locally. For example, it takes 600kb to store a 10s wav file, which infers that if we receive a 1-minute audio file, its size will exceed the storage of the flash. Consequently, instead of storing the entire file in the flash, we use the flash as a buffer. In particular, we

plan to use the flash as a buffer, only to store 5ms of the audio each time and send it to the cloud. In that case, the buffer only needs to use about 30b for the audio storage.

Ethics and safety:

#User Privacy and Data Security[3]
The AMADEUS project involves the collection and processing of user audio data, which raises privacy and data security concerns. In accordance with the IEEE Code of Ethics, Section I.5, it is our responsibility to ensure that the privacy of the user is protected and that sensitive information is not misused. User data must be securely transmitted and stored, utilizing encryption both during transit (Wi-Fi) and at rest on cloud servers. Compliance with global data privacy regulations must be maintained. To avoid breaches, we will anonymize user data when possible and implement secure protocols for all communications between the ESP32 board and cloud server.

#Bias in AI Models[3]
As the project involves the use of AI, it is important to address the risk of bias in the AI models. According to the ACM Code of Ethics, Section 1.4, we must ensure fairness in algorithmic processes. Any AI system should be free of bias regarding race, gender, or other demographic factors. To mitigate this, we will work with the developers of the AI model to ensure that training data is diverse and representative. Additionally, continuous monitoring and auditing of AI model performance will be established to prevent unfair treatment of users.

ce will be implemented, allowing users to review data policies and make informed decisions.

#Hardware Safety[3]
The AMADEUS system involves the use of an ESP32 microcontroller and audio-related hardware components. According to UL 60950-1 and IEC 62368-1 safety standards for audio-visual and IT equipment, the PCB must be designed to avoid electrical hazards such as short circuits or overheating. Additionally, any exposed parts of the system must be properly insulated to protect users from accidental electrical shocks.

#Power System Safety[3]

The power supply system uses both USB and battery-powered configurations. It is crucial to ensure that these power sources are properly regulated to avoid potential fire hazards or battery explosions. The IEEE Code of Ethics, Section I.1, requires us to prioritize public safety and welfare. Therefore, we will conduct rigorous testing of the power supply circuit, ensure compliance with FCC Part 15 regulations regarding electromagnetic interference, and adopt safety protocols for battery usage, such as overvoltage and temperature protection circuits.

References

[1] Marr, B. (2024, July 2). Generative AI is coming to your home appliances. *Forbes*. https://www.forbes.com/sites/bernardmarr/2024/03/29/generative-ai-is-coming-to-your-home-appliances/(visited on 10/1/2024)

[2] FalcoTK. (n.d.). GitHub - FalcoTK/character-ai: Unofficial API for character.ai, Support chat v2, support voice module (BETA TES). GitHub. https://github.com/FalcoTK/character-ai

[3] IEEE. “”IEEE Code of Ethics”.” (2016), [Online]. Available:IEEE - IEEE Code of Ethics(visited on 9/17/2024).