Hume.ai

Hume.ai is an AI voice system designed to respond empathetically. It has applications in healthcare for monitoring patients and mental health conditions, as well as in schools for monitoring students to help them focus on classes and create emotionally intelligent chatbots. Hume.ai measures 53 expressions through emotional language and 48 expressions through facial cues, vocal bursts, and speech prosody.

- **Emotional Language:** The words we say and the emotions they convey.
- **Facial Cues:** Facial expressions.
- **Vocal Bursts:** Non-linguistic vocal utterances like sighs, ohs, ahs, umms, etc.
- **Speech Prosody:** The way we use words while speaking.

Features:

1. **Empathic Voice Interface (EVI)**
2. **Expression Measurement API**
3. **Custom Model API**

1. Empathic Voice Interface (EVI)

Hume's Empathic Voice Interface (EVI) is an emotionally intelligent AI that takes audio input and returns both audio and a transcript. Based on tone and rhythm, EVI generates empathetic responses.

- **Connection:** A WebSocket connection is established using an API key and sometimes a WSS URI. Once connected, users can start speaking through their device. Audio is sent to EVI via WebSocket, and EVI returns transcribed text, vocal expression measurements, generated text, and audio responses.
- **Responses:**
    - The text of EVI’s reply
    - EVI’s expressive audio response
    - A transcript of the user’s message with vocal expression measures
    - Messages if the user interrupts EVI
    - Notifications when EVI finishes responding
    - Error messages if issues arise

2. Expression Measurement API

Hume's Expression Measurement API is a cutting-edge tool for capturing and analyzing human expressions across different modalities.

- **Input Processing:** The API can analyze various types of media:
    - Audio files for vocal expressions
    - Video files for facial expressions
    - Images for facial expressions
    - Text for linguistic emotional content
- **Measurement Capabilities:**
    - **Facial Expression:** Captures 48 dimensions of emotional meaning, including subtle facial movements.
    - **Speech Prosody:** Analyzes the tone, rhythm, and timbre of speech across 48 dimensions.
    - **Vocal Burst:** Identifies and analyzes vocal sounds (e.g., laughs, sighs, cries) along 48 dimensions.
    - **Emotional Language:** Evaluates the emotional tone of transcribed text across 53 dimensions.
- **Processing Methods:**
    - Batch Processing: For analyzing large volumes of files asynchronously.
    - Real-time Streaming: For continuous, real-time analysis of data streams.

3. Custom Model API

The Custom Model API integrates patterns of language, vocal expression, and facial expression to predict human preferences and needs more accurately than traditional language models. This API allows users to create custom multimodal models tailored to specific applications, such as predicting well-being, satisfaction, mental health, and more.

Work Summary:

1. **Integrating EVI with Expression Measurement API:**
    - Integrated EVI with the Expression Measurement API to enable EVI to understand both facial and vocal expressions. The camera detected facial expressions while EVI detected vocal expressions simultaneously. We explored features in Hume.ai but found no built-in functionality for this integration. Attempts to connect them through prompts were unsuccessful, as prompts are only sent at the start of a conversation and do not update with each new expression. Ultimately, we concluded that EVI and the Expression Measurement API cannot work together as intended.
2. **Connecting RAG with EVI:**
    - Attempted to connect RAG to EVI. During this process, it became clear that audio is converted to bytes and sent to EVI via WebSocket, making it impossible to add RAG in between this process. After consulting with the Hume.ai team on Discord, we learned that to make EVI work according to our requirements, we would need to train our own model using the Custom Model API.