**Multimodal Voice Agent**

**Implementation Overview**

I extended my Caroline AI agent from previous assignments with voice interaction capabilities by adding speech-to-text (STT) and text-to-speech (TTS). The system follows a simple pipeline: the user speaks a question, Whisper transcribes it to text, my agent processes it, and gTTS speaks the response back.

**Technologies Used**

**Speech-to-Text: OpenAI Whisper**

* I chose Whisper because it's accurate, free, and works offline
* Used the "base" model which balances speed and accuracy
* Takes about 1-2 seconds to transcribe a 5-second voice clip

**Text-to-Speech: Google TTS (gTTS)**

* Selected gTTS for its natural-sounding voices and simple API
* Generates clear, professional audio responses
* Takes less than 1 second to create speech from text

**Audio Processing:**

* sounddevice for recording from microphone
* soundfile for saving audio files
* Records 5-second clips at 16kHz (standard for speech)

**Integration with Caroline AI**

The voice system wraps my existing Caroline AI agent, which contains knowledge about my cardiovascular research at Harvard, healthcare AI experience at Tencent, and technical skills. When someone asks about these topics via voice, the system recognizes keywords and responds with relevant information from my background.

**Example Run: Demo Walkthrough**

**User Input:** I spoke into the microphone saying "What is your research at Harvard?"

**Step 1 - Recording (5 seconds)** The system captured my voice and saved it as a WAV file.

**Step 2 - Speech-to-Text (~1.2 seconds)** Whisper transcribed my speech perfectly: "What is your research at Harvard"

* All words were correct despite natural speaking pace
* Automatic capitalization was applied

**Step 3 - Agent Processing (<0.1 seconds)** My agent detected keywords "harvard" and retrieved the response: "I focus on cardiovascular risk modeling using cooperative learning and survival analysis. My thesis involves developing polygenic hazard scores using multi-view data integration across genomic, cardiac MRI, and clinical data..."

**Step 4 - Text-to-Speech (~0.8 seconds)** gTTS converted the response into natural-sounding speech (18 seconds of audio).

**Step 5 - Playback (18 seconds)** The system played Caroline’s voice response through the speakers.

**Total Time:** User waits ~2 seconds for processing, then hears an 18-second response. The entire interaction feels natural and conversational.

**Insights**

**1. Voice Makes AI More Accessible**. The same information delivered through voice feels more engaging than reading text. Speaking to the agent and hearing responses makes it feel like a real conversation.

**2. Latency Matters.** The 2-second wait time feels natural because it’s similar to how humans pause before answering questions. Anything under 3 seconds works well for conversational AI.

**3. Audio Quality is Critical.** Clear input audio leads to accurate transcriptions. Background noise or mumbling causes errors. The TTS output sounds professional and matches the technical expertise of Caroline’s responses.

**4. Open-Source Tools Are Production-Ready.** Whisper and gTTS perform as well as paid APIs for this use case. Using free, local tools makes the project more accessible and teaches the full pipeline instead of just calling APIs.

**5. Voice Reveals Personality.** Hearing Caroline’'s expertise spoken aloud adds authenticity. The same content feels more credible when delivered through natural speech versus text.

**Conclusion**

This project successfully demonstrates how voice interaction transforms a text-based AI agent into a more natural, engaging system. By combining Whisper for speech recognition and gTTS for speech synthesis, I created a complete voice interface that maintains Caroline AI’s personality and expertise while enabling hands-free, conversational interaction.

The implementation shows that building multimodal agents is practical with modern open-source tools, requiring no API costs or complex infrastructure. The ~2-second response time provides a real-time conversational feel, and the voice modality makes the agent more accessible and engaging than text alone.

**Future improvements** could include streaming processing for lower latency, voice activity detection to automatically stop recording, and support for multiple languages.