**Maya AI– Technical Overview**

**Introduction**

Maya is an AI-powered conversational companion designed to simulate human-like, real-time voice interactions.  
It integrates speech recognition, AI-driven natural language understanding, and text-to-speech synthesis to deliver empathetic, context-aware, and emotionally intelligent conversations.

1. **Speech-to-Text (STT)**

**Current Choice → Google Speech Recognition**

**Why Chosen**

* Fast, lightweight, and easy integration using the speech\_recognition Python library.
* High accuracy for short, conversational phrases.
* Zero-cost → ideal for MVPs and early prototypes.

**Performance**

* Latency → ~0.5 – 1 sec per transcription
* Accuracy → High for English conversational speech
* Dependency → Requires internet connection

**Alternatives 🔄**

| **Service** | **Pros** | **Cons** | **Best Use** |
| --- | --- | --- | --- |
| **OpenAI Whisper** | Works offline, multilingual support | Slower, heavy model (~1GB) | Privacy-focused, offline apps |
| **AssemblyAI** | Streaming STT, high accuracy | Paid API | Real-time assistants |
| **Deepgram** | Ultra-low latency, very accurate | Paid API | Live call bots |
| **Azure Cognitive STT** | Enterprise-grade accuracy, streaming-ready | Paid | Enterprise AI systems |

**2. Text-to-Speech (TTS)**

**Current Choice → Microsoft Edge Neural TTS (edge\_TTS)**

**Why Chosen**

* Uses neural voices for natural, human-like speech.
* Lightweight Python integration via edge-tts.
* Offers 100+ voices and 45+ languages.
* Free tier available → scalable with Azure for production.

**Performance 🎧**

* Latency → ~300 – 700 ms per sentence
* Voice Quality → Highly natural, expressive, and smooth
* Dependency → Requires internet connection

**Alternatives**

| **Service** | **Pros** | **Cons** | **Best Use** |
| --- | --- | --- | --- |
| **ElevenLabs** | Most human-like voices, emotional tones | Paid | High-end assistants |
| **Amazon Polly** | Scalable, multiple neural voices | Paid | Cloud-based apps |
| **Google Cloud TTS** | High quality, multilingual support | Paid | Google ecosystem |
| **Coqui TTS** | Open-source, works offline | Requires tuning | Privacy-first apps |

**3. AI Engine**

* Groq GPT Model → Delivers fast, context-driven responses.
* Integrated with LangChain, ConversationBufferMemory → Remembers the last 10 interactions for context-aware replies.
* Uses a concise, empathetic system prompt to make Maya feel human-like.

**Why This Setup Works Now**

| **Aspect** | **Choice** | **Why It Works** |
| --- | --- | --- |
| **STT** | Google Speech API | Accurate, lightweight, and free |
| **TTS** | Edge Neural TTS | Natural voices, fast, scalable |
| **Memory** | LangChain BufferMemory | Context-aware, human-like chats |
| **AI Engine** | Groq - openai/gpt-oss-120b | High-speed inference, ideal for real-time apps |

**Conclusion** → This setup is perfect for MVPs and small-scale production deployments.

**Future Enhancements 🚀**

1. **Streaming STT** → Use **Whisper(production)** for **real-time transcription**.
2. **Expressive TTS** → Integrate **Edge Neural TTS** for **emotion-rich, human-like voices**.
3. **Offline Privacy Mode** → Combine **OpenAI Whisper** + **Coqui TTS** for **fully local deployments**.
4. **Streaming AI Responses** → Allow Maya to **talk while thinking** for **seamless conversations**.