Una
Matata

Grab Voice Assistant

How does an AI Voice Assistant work?

An AI voice assistant works by listening to a user's voice, interpreting the speech into text, understanding the intent behind the message, and responding appropriately—either by speaking back, taking an action, or both. The process starts with speech recognition, where the audio input is converted to text. Then comes natural language processing (NLP), which helps the assistant understand the meaning and context of the request. Based on this, it uses natural language understanding (NLU) to interpret the user's intent. The assistant then finds or generates a response, which is converted back to speech through text-to-speech (TTS) technology.

What is the STT → LLM → TTS Pipeline?

This pipeline is the backbone of how modern AI voice assistants and real-time conversational systems work. It converts your voice into a response that talks back to you — like what ChatGPT does when you use voice mode.

STT
Speech-to-Text

- Converts your spoken voice into written text.
- Think of it as voice typing.
- Example: You say, “What’s the weather today?” → it becomes: "What's the weather today?"

LLM
Large Language Model

- “Brain” of the system.
- Reads the transcribed text, understands what you mean, and generates a smart reply.
- Example: It processes the question about weather and decides the best answer might be: "Today is sunny with a high of 75°F."

TTS
Text-to-Speech

- Converts the reply (text) from the LLM back into a realistic voice.
- You hear: “Today is sunny with a high of 75°F.”
- It can use human-like voices, different languages, even emotions!

What is a Real-Time Model?

A real-time model is an AI model designed to process and respond to inputs immediately or with minimal delay — typically within milliseconds to a few seconds.
It's all about speed + accuracy so that users feel like they're interacting with something instantaneous and natural.

Real-time model enables this seamless experience by being:

- Fast – Low latency (very short delays)
- Lightweight – Runs on edge devices or with minimal cloud delay
- Efficient – Optimized to use fewer resources (memory, power)

AI model that is used in this project is **GPT-4o Mini Realtime** which is designed to handle text, audio (speech), and image

Unlike older systems that used separate models for each input type (like separate STT, LLM, and TTS), GPT-4o handles all of it in one neural network. This reduces delay and improves efficiency.

Real-Time vs. Non-Real-Time Models

Speed

Real-time model: designed for speed and responsiveness.

Non-real-time model: focuses more on depth, power, and accuracy

Goal

Real-time models: deliver a natural, conversational experience with minimal delay.

Non-real-time model: suited for tasks like long-form writing, detailed question answering, or data analysis, where speed is less critical.

Size

Real-time model: small, optimized, and efficient enough to run on devices like smartphones, smartwatches, or edge hardware.

Non-Real-time model: larger, more resource-intensive, and typically run in powerful cloud servers or data centers.

Therefore we implement Real-time model to our project because of:

- Fast response – Enables smooth, natural conversations with minimal delay.
- Lightweight – Runs efficiently on phones, smartwatches, and other devices.
- Real-time interaction – Feels like you're talking to a person, not a machine.
- Resource-friendly – Uses less power and memory, ideal for on-device use.
- Immediate action – Perfect for quick voice commands like “Turn off the lights.”

What is the API?

API (Application Programming Interface)

It acts a messenger between different software systems. It allows one program to talk to another and ask it to do things — like give data or perform a task

e.g. when you use an app like a weather app, that app sends a request to a weather API. The API says, “Hey weather server, this user wants the current weather in KL.” The server responds, and the app shows you the result.

API used in this project is:

1. METMalaysia

MetMalaysia provides a Web Service API that allows developers to access a variety of meteorological data, including weather forecasts, warnings, and real-time observations.

The API offers various endpoints to access different types of data:

- Weather Forecasts
- Weather Warnings
- Real-Time Observations
- Climate Data

2. Mapbox

Powerful mapping platform that lets developers add customizable maps and location features to apps and websites. It works by providing APIs and tools to render interactive maps, display markers, routes, and geolocation data.

Its offers:

- real-time tracking
- delivery apps
- navigation
- data visualization

Problem Statement

The objective is to build a robust voice interaction system that enables reliable driver-assistant communication in challenging audio environments. You need to use some of your creative spirit to design a scenario in which audio assistance is specifically useful. The solution should:

1.Maintain high accuracy in noisy conditions

- Implement noise cancellation and filtering techniques to enhance voice recognition accuracy despite background interference.

2.Adapt to diverse speech patterns

- Utilize NLP models capable of understanding and processing regional accents, dialects, and colloquial expressions.

3.Provide clear, reliable functionality with partial audio clarity

- Design systems that can interpret incomplete or unclear voice inputs and still deliver accurate responses.

4.Demonstrate resilience across various environmental challenges Ensure the solution is adaptable to different environmental conditions and can maintain reliable communication

Solution

1. High Accuracy in Noisy Conditions

Solution with GPT-4o Mini Realtime:

- GPT-4o processes audio natively, meaning it can handle raw, noisy speech better than traditional systems that separate STT and LLM.
- With fine-tuned preprocessing (e.g., noise reduction filters before input), GPT-4o's real-time capabilities allow it to extract meaning even when background noise is present (e.g., traffic, music, or rain).
- Built-in robustness to audio distortion enables it to outperform traditional STT in challenging environments.

2. Adapt to Diverse Speech Patterns

Solution:

- GPT-4o has been trained on a diverse range of global accents and dialects, making it highly effective at understanding regional speech variations and colloquial expressions.
- It doesn't rely on exact phrasing but it understands intent even with informal or non-standard language.
- Can be fine-tuned or prompted to prioritize contextual understanding, which helps when interpreting local driver commands (e.g., "lah" or "can park here ah?" in Malaysian English).

3. Handle Partial or Unclear Audio Inputs

Solution:

- Since GPT-4o is multimodal and context-aware, it can infer missing words or guess meaning from partial sentences.
- For example, if a user says, "Turn on the…" and stops, GPT-4o can use context (previous commands, time of day, environment) to guess "air conditioning" or "headlights."
- This makes the system feel smart and natural, even when drivers speak unclearly or are interrupted.

4. Resilience to Environmental Challenges

Solution:

- GPT-4o Mini Realtime is optimized for low-latency, making it ideal for real-time use in dynamic environments like a moving car.
- Because it operates efficiently on edge or with minimal delay, it's more reliable in poor network conditions.
- Can be paired with onboard sensors (e.g., GPS, car diagnostics) to adjust responses based on weather, speed, or driving conditions — enabling proactive, situation-aware assistance.

thank you