Multilingual Voice RAG Assistant for Visually Impaired – Design Document

Overview

The Voice RAG Assistant is a multilingual voice-enabled assistant built with Streamlit. It accepts audio or text queries, performs retrieval-augmented generation (RAG) using Google Gemini, and returns AI-generated answers with audio output. It supports multiple languages and integrates technologies like OpenAI Whisper, FAISS, Sentence Transformers, and Google Text-to-Speech.

Pipeline Design

Input Layer

- **Voice Input:** Users upload .wav or .mp3 audio files containing their query.
- **Text Input:** Users can type their queries directly.

Speech-to-Text (STT) Layer

- **Transcription Engine:**
    - **Tool:** OpenAI Whisper
    - Converts audio files to text with detected language.
    - Ensures accurate transcription of multilingual audio.

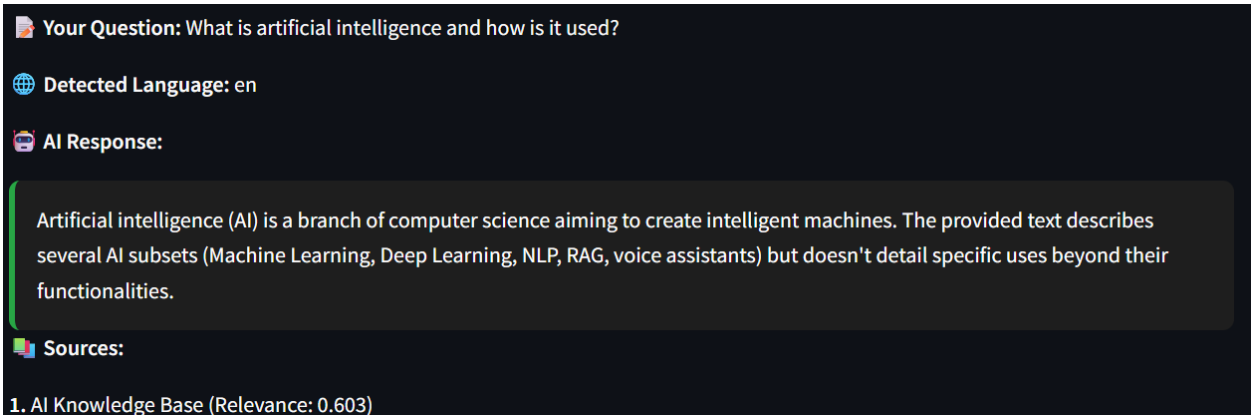
Embedding Layer

- **Embedding Generation:**
    - **Tool:** Sentence Transformers (all-MiniLM-L6-v2 model)
    - Converts both the **query text** and the **knowledge base chunks** into high-dimensional embeddings.
    - Embeddings are normalized using **FAISS L2 normalization** for cosine similarity.

Vector Search (Retriever)

- **Index Creation:**
    - **Tool:** FAISS (Facebook AI Similarity Search)
    - **Index Type:** IndexFlatIP for inner product search (cosine similarity).
    - Stores pre-encoded knowledge base document chunks.

- **Retrieval Logic:**
    - Given a query embedding, retrieves top-**k** most similar document chunks (default k=3).
    - Computes **semantic similarity** using **FAISS**.
    - Each retrieved chunk includes its **source** and **similarity score**
- **Generation Layer (RAG)**
    - **LLM Integration:**
        - **Tool:** Google Gemini API
        - Configures Gemini with the provided **API key**.
    - **Prompt Construction:**
        - Combines retrieved document chunks into a contextual **prompt**:



        - Sends the prompt to Gemini's **gemini-1.5-flash** model.
        - Receives a **concise, context-aware response**.

Text-to-Speech (TTS) Layer

- **Audio Generation:**
    - **Tool:** [Google Text-to-Speech (gTTS) or pyttsx3 for offline.
    - Converts generated text response into audio (.wav).
    - Plays the response using an **HTML5 audio player**.

User Interface

- **Frontend:**
    - **Tool:** Streamlit
    - Provides:
        - Tabs for **Voice Query**, **Text Query**, and **About**.
        - File uploader for audio files.
        - Display of transcription and generated answers.
        - Playback of audio responses.
        - Display of retrieval **sources** with relevance scores.

🛠️ Technology Stack

Component	Tool/Library
Frontend	Streamlit
Speech Recognition	OpenAI Whisper
Embedding Generation	Sentence Transformers (`MiniLM`)
Vector Search	FAISS
Language Model	Google Gemini API (`gemini-1.5-flash`)
Text-to-Speech	Google TTS / pyttsx3
Audio Playback	HTML5 Embedded Audio Player

Detailed Retrieval & Generation Logic

Retrieval (FAISS)

1. **Document Preprocessing:**
    - Text documents are **split into overlapping chunks** (~300 words with 50-word overlap) using split_text().
    - Each chunk is embedded via **Sentence Transformers**.
2. **FAISS Index Construction:**
    - Chunks' embeddings are normalized and added to a **FAISS IndexFlatIP** index. **Cosine similarity** is computed for semantic search.
3. **Query Execution:**
    - User query is embedded and normalized.
    - **Top-k matching document chunks** are retrieved along with similarity scores.

Generation (RAG)

1. **Context Construction:**
    - Combine retrieved chunks to create a **context** for the query.
2. **Prompt Engineering:**
    - Build a structured prompt for **Gemini** with clear instructions and the context.
3. **Gemini Generation:**
    - **LLM (gemini-1.5-flash)** generates the final response text.
4. **Post-processing:**
    - Response is returned as text and converted into speech using **TTS**.

Multilingual Support

- **Whisper** detects query language automatically. **Gemini** supports multi-language prompt understanding and response.
- **gTTS/pyttsx3** generates audio in the detected language.