Laboratorio 3 – Aplicación con ASR, LLM y Voice Cloning

Fecha de entrega de código: Lunes 27 de Octubre - 8:00 PM
Fecha de presentación: Martes 28 de Octubre - 8:00 PM (horario de clases)

1. Descripción del Problema

En este laboratorio, el objetivo es construir una aplicación interactiva de voz que combine tres componentes:

a. **Automatic speech recognition (ASR):** convertir audio en texto.
b. **Large language model (LLM):** generar una respuesta breve y coherente a partir del texto reconocido.
c. **Text to speech (TTS con Voice Cloning):** reproducir la respuesta generada con la voz de un integrante del grupo.

El usuario debe poder hablar con la aplicación y recibir una respuesta hablada.

2. Librerías sugeridas

Las herramientas a usar son libres, siempre que sean gratuitas y ejecutables en Google Colab.
Se recomienda lo siguiente:

a. **Automatic speech recognition (ASR):** Whisper.
b. **Large language model (LLM):** FLAN-T5, Qwen2.5-3B-Instruct.
c. **Text to speech (TTS con Voice Cloning):** xTTS.

Código ejemplo aquí.

3. Tareas

Cada grupo deberá desarrollar las siguientes etapas:

a. **Automatic speech recognition (ASR):** Transcribir el contenido usando el modelo de ASR elegido. Registrar el texto obtenido y el tiempo de inferencia.
b. **Large language model (LLM):** Tomar el texto transcrito y generar una **respuesta breve (1–3 oraciones)** usando un modelo de lenguaje. El LLM debe ser ejecutado localmente en Colab. Registrar el tiempo de respuesta y la longitud del texto generado.
c. **Text to speech (TTS con Voice Cloning):** Generar el audio de salida que lea la respuesta del LLM usando esa voz clonada. Comparar la calidad frente a una voz base (sin clonación).
d. **Integración:** Integrar las tres partes en un único flujo de ejecución. El sistema debe recibir un audio de pregunta y devolver un audio de respuesta. Mostrar los tiempos por etapa (ASR, LLM, TTS) y el tiempo total del pipeline.

4. Entrega (Martes 28)

Deberán subir sus soluciones al formulario de entrega. El envío debe incluir:

- **Código fuente (Google Colab Notebook):**
    a. Implementación de cada módulo (ASR, LLM, TTS).
    b. Función final que integre las tres etapas.
    c. Resultados intermedios y finales impresos (textos y tiempos).

5. Presentación en clase (Martes 28)

Durante el horario de clase, cada grupo deberá **demostrar en vivo** su aplicación.
La presentación consiste en:

1. Ejecutar un ejemplo en vivo: hablar con la aplicación y oír la respuesta.
2. Explicar brevemente:
    a. Qué modelo usaron en cada parte y por qué.
    b. Qué ajustes realizaron (parámetros, latencia, calidad).
    c. Qué observaciones obtuvieron sobre la calidad de la voz clonada.