Brief Report

1. How the System Works (High-Level Overview)

This system takes an audio or video file as input, performs automatic speech recognition (ASR) using the Whisper model, and uses face recognition to identify speakers in the video. It detects languages automatically and can translate if needed, while also generating timestamps for each transcript segment. The final transcript includes speaker labels, timestamps, and the spoken content.

2. Challenges and Solutions

One challenge was integrating speaker identification with the transcript, since audio-only recognition does not directly provide speaker labels. I solved this by detecting faces in the video, clustering them, and mapping time ranges to transcript segments. Another challenge was running the system entirely in Google Colab without a local Python environment; I addressed this by ensuring all dependencies are installed directly in Colab and providing clear run instructions.

3. What Works Well and Possible Improvements

The speech recognition works well for clear audio and provides accurate timestamps. Speaker identification is effective when faces are clearly visible and well-lit. However, accuracy decreases when faces are partially hidden or far from the camera. In the future, the system could be improved by adding voice-based speaker identification for better performance when face detection fails.

4. Instructions for Running the Code

1. Open the provided .ipynb file in Google Colab.
2. Install the required packages in Colab:
3. Upload the necessary attachment files, and unzip filemayuse.zip before uploading its contents.
4. Run each notebook cell sequentially from top to bottom.

5. Please note:

For Part 1 (Automatic Speech Recognition) and Part 2 (Speaker Identification), the results are printed directly to the console (Colab cell output) and are not saved to a file.

For Part 3 (Language Detection and Translation), the output is saved in the file “sentence_transcript_formatted.txt” in the working directory.