Curious PM Project

Flow of Project : https://miro.com/app/board/uXjVLRKRgGg=/

Overview

This project focuses on improving the quality of audio within videos by extracting the audio, converting it into text (transcription), correcting grammar, and eliminating filler words using Azure OpenAI. After cleaning the transcript, it is converted back into audio and synchronised with the original video.

The most challenging part of the project was remapping the newly generated audio back onto the original video without causing any synchronisation issues, ensuring there was no delay or mismatch between the video frames and the new audio.

Problem Statement

When processing video content to improve its audio quality, there are several key challenges:

1. **Audio Extraction: Isolating the audio from the video file for processing.**
2. **Speech Recognition: Converting the extracted audio into text using a reliable speech-to-text engine.**
3. **Grammar Correction and Filler Word Removal: Correcting grammatical issues and eliminating filler words to enhance clarity.**
4. **Text-to-Speech Conversion: Re-converting the cleaned transcript into an audio format.**
5. **Synchronisation (The Most Challenging Part): Ensuring that the newly generated audio, which may have slightly altered timing due to the removal of filler words, matches the original video precisely without any audio-video delay.**

Solution Process

1. Audio Extraction

The first step involved isolating the audio from the video. This audio track was then prepared for further processing.

2. Speech-to-Text Conversion

The audio was converted into a text transcript using deepgram API. This transcript formed the basis for the next steps in the process.

3. Text Cleaning and Grammar Correction

Azure OpenAI was used to clean the transcript by correcting grammatical errors and removing common filler words like “uh,” “um,” and “hmm.” This resulted in a more professional and clear transcript, which could then be converted back into speech.

4. Text-to-Speech Conversion

The cleaned transcript was converted back into an audio format using text-to-speech technology(Deepgram API). This generated the new, corrected audio track, which was now shorter due to the removal of filler words and pauses.

Tackling the Most Challenging Part: Audio-Video Synchronisation

The Core Challenge: After correcting the transcript and removing filler words, the new audio became shorter than the original, creating a significant issue in ensuring that the new audio would align perfectly with the video. This required careful attention to maintain synchronisation between the speech and video frames, especially to avoid any lip-sync mismatches.

Solution Approach:

- **Breaking the Audio into Chunks**: To solve this, I divided the original audio into chunks based on the natural breaks at the end of each sentence. Each sentence served as a checkpoint for mapping the audio back to the video. These checkpoints allowed for better control over the timing of the new audio.
- **Processing Sentences Individually**: Each sentence from the cleaned transcript was converted into an audio segment. This ensured that even though the duration of each sentence might have changed due to the removal of filler words, the remapping would still respect the original video's structure.
- **Remapping Audio Using Checkpoints**: Using the time checkpoints of the original sentences, I remapped each newly generated audio segment onto the video. Since each chunk was individually timed and aligned with its corresponding part in the video, this approach maintained perfect synchronisation. By working on sentence-level chunks, the issue of time discrepancy between the old and new audio was effectively managed.

Outcome: This method ensured that the new audio fit seamlessly with the original video without introducing any delays or synchronisation issues. By breaking down the audio into sentence chunks, it allowed for precise alignment and perfect timing, resulting in a natural and polished video output.

Conclusion

This project demonstrates a complete solution for enhancing the audio content of videos while addressing the critical challenge of maintaining synchronisation between the modified audio and the video. By leveraging Azure OpenAI to clean up the transcript and applying advanced techniques to remap the new audio to the original video, the project improves the clarity and professionalism of spoken content in videos.

This solution is particularly useful for content creators and educators looking to enhance the quality of their video content without sacrificing synchronisation between the audio and the visual elements.