

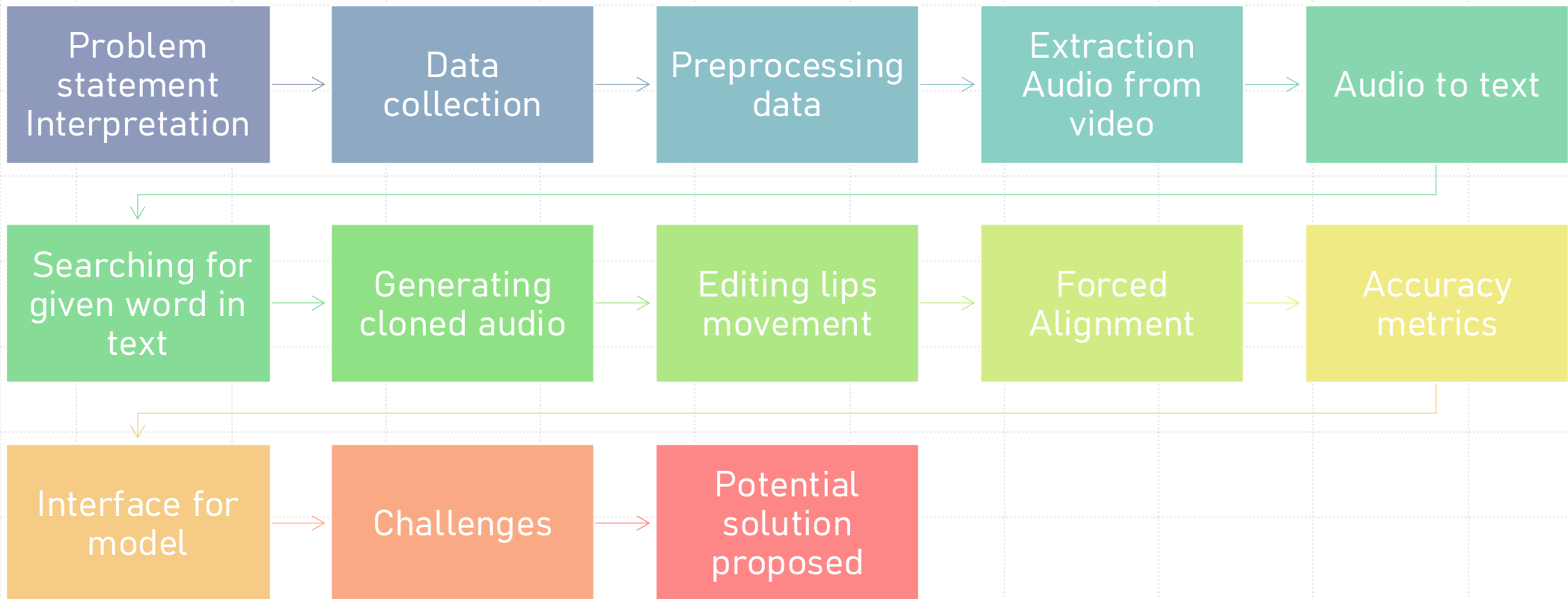
Mind Guardians

Lead: Harshil Singh Juneja
(22B0666)

Member : ShivamKumar (22B0745)



Overview



Problem statement Interpretation

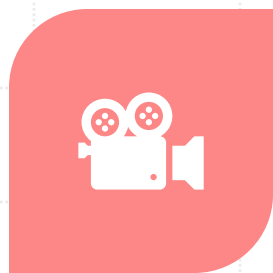
- Develop a system that can accurately replace specific words in Hindi videos with user-provided words, while ensuring the replaced word's lip-sync matches the video seamlessly.
- Give an interface for user to upload video and select the word for replacement as well as to pass the word for the replaced word.
- Give an output matrix to compare the precision of lip-sync and frame alignment



Dataset Collection



COMMON VOICE
DATASET FOR AUDIO
TO TEXT




HINDI CORPUS FOR VIDEO
FILES



INDIC-TTS



GRID DATASET FOR
LIPSYNC DATA



Preprocessing data

Preprocessing video data

- Normalizing
- Converting voice to 16Hz
- Aspect ratio

Text to speech data

- Text Normalization {Convert Numbers, date, word to its full form }
- Hindi Specific Tokenizer
- Remove Punctuation

Audio to text

- Trimiing silence
- Noice Reduction
- Feature extractions like Mel spectrogram, Mel Frequency Cepstral coefficient

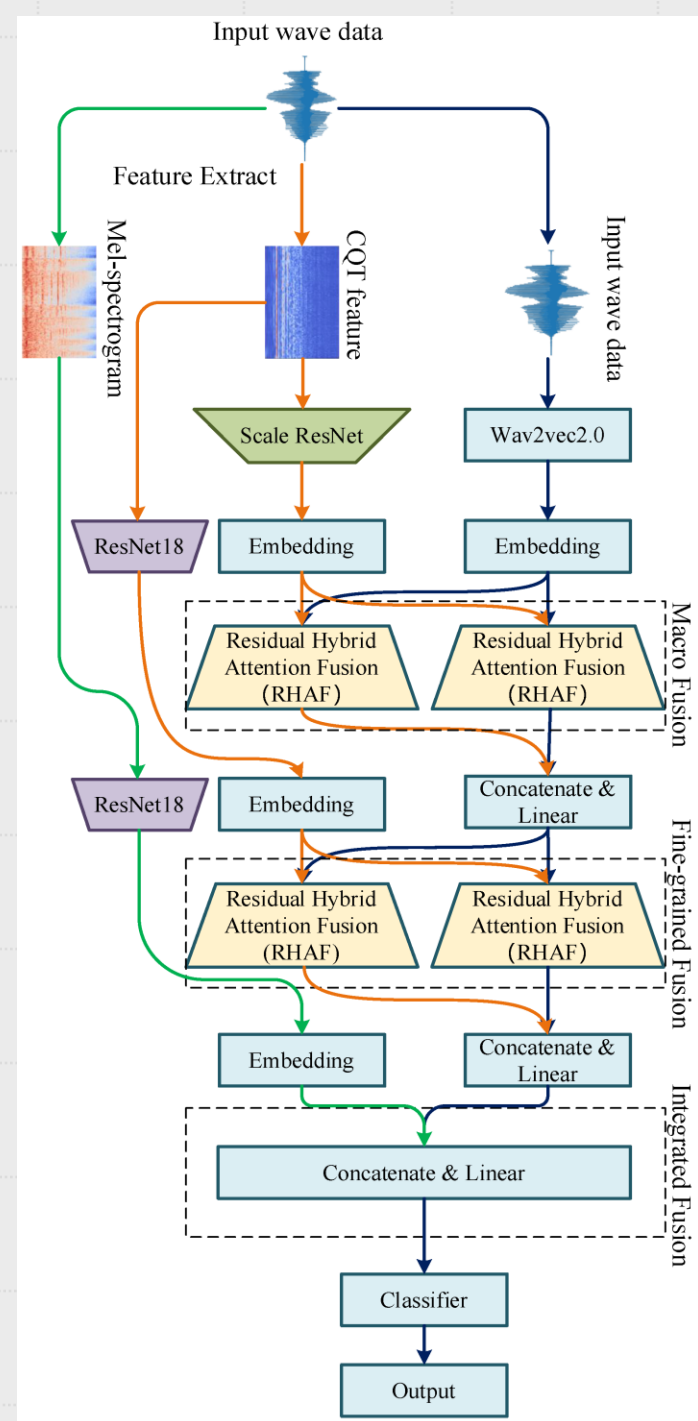
Extracting audio from video


- Extracting audio allows us to apply speech recognition models to transcribe spoken content
- MoviePy is a Python library that deals with multimedia processing, specifically for editing and manipulating video and audio files and can extract audio from videos for us



Audio to text

- Converting audio to text allows you to identify specific words in the spoken content that need to be replaced or edited in the video
- Wav2Vec is a state-of-the-art Automatic Speech Recognition model developed by Facebook AI that converts spoken audio into text by learning speech representations from raw audio data
- Wav2Vec 2.0 is highly accurate and works well with Hindi audio, ensuring that the transcription of your spoken content is precise





Searching for the given word in text



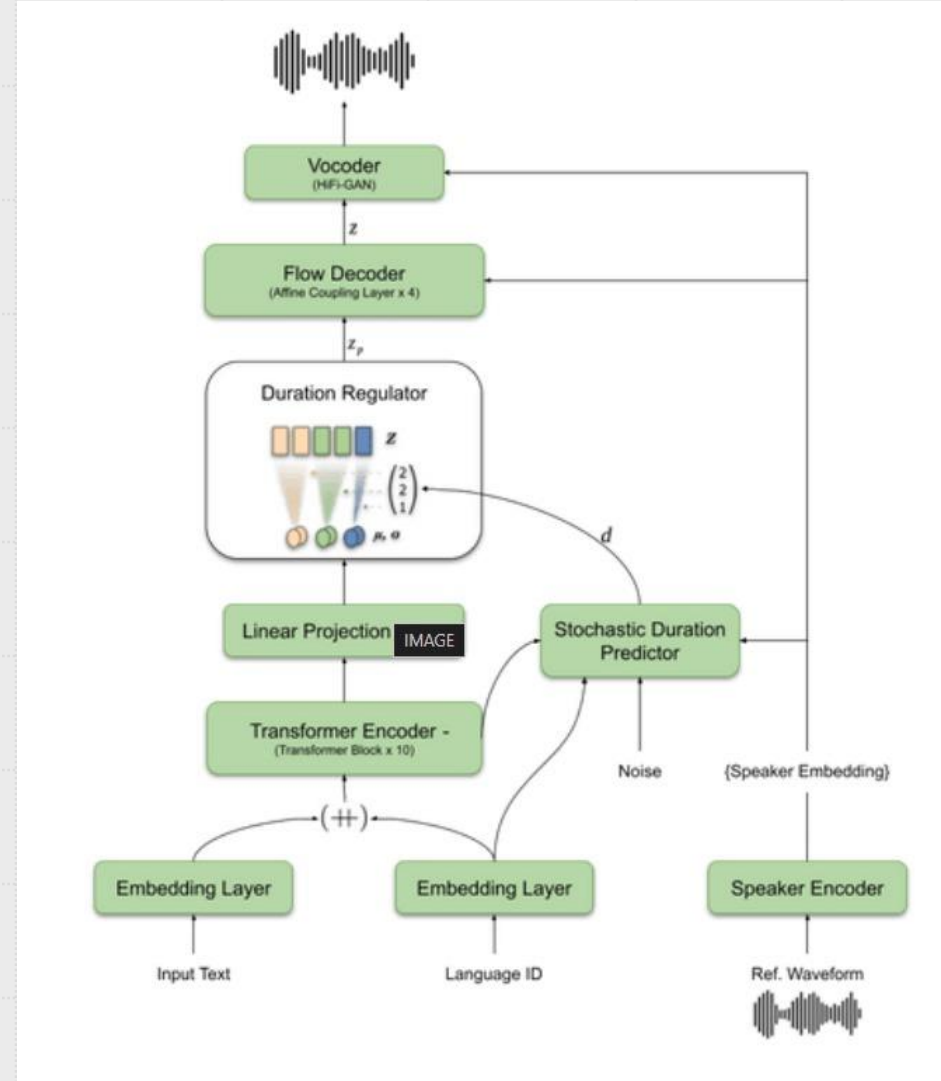
Using Python's `replace()` method for basic word substitution or tokenize the sentence with regex to replace exact words without affecting partial matches.



Will be considering punctuation handling and hindi tokenization for precise word replacement in Hindi sentences.

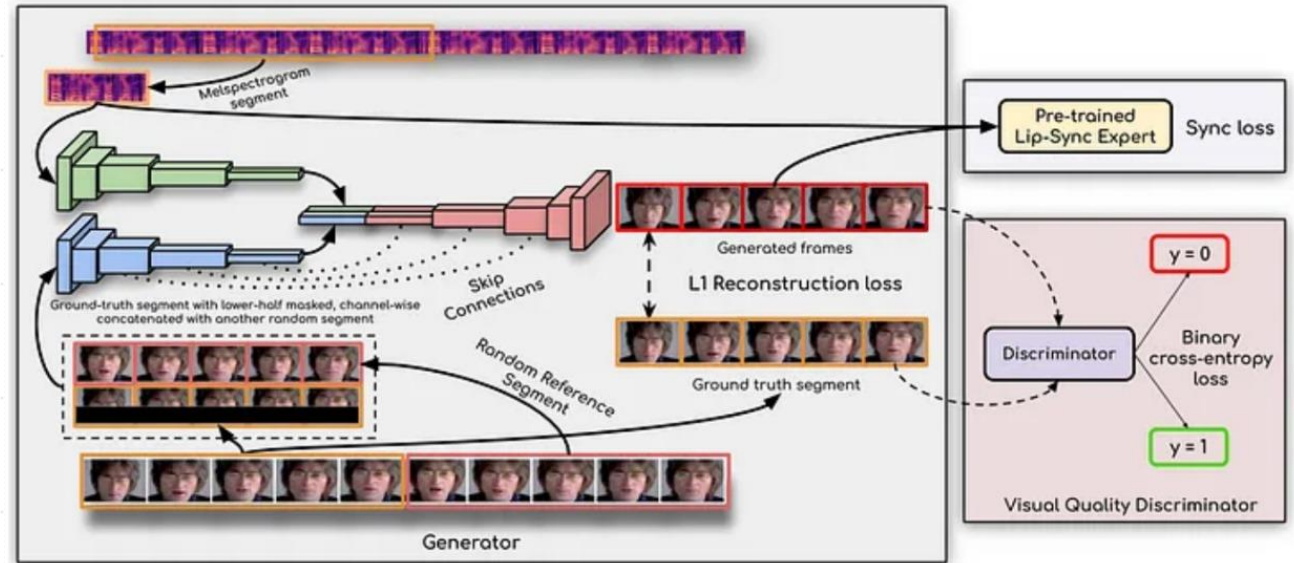
Generating cloned audio

- It ensures that the newly generated speech matches the tone, pitch, and style of the original speaker
- Coqui TTS is an open-source text-to-speech (TTS) framework designed for creating high-quality synthetic voices
- Delivers natural-sounding speech that maintains the nuances of the original speaker, improving overall audio quality



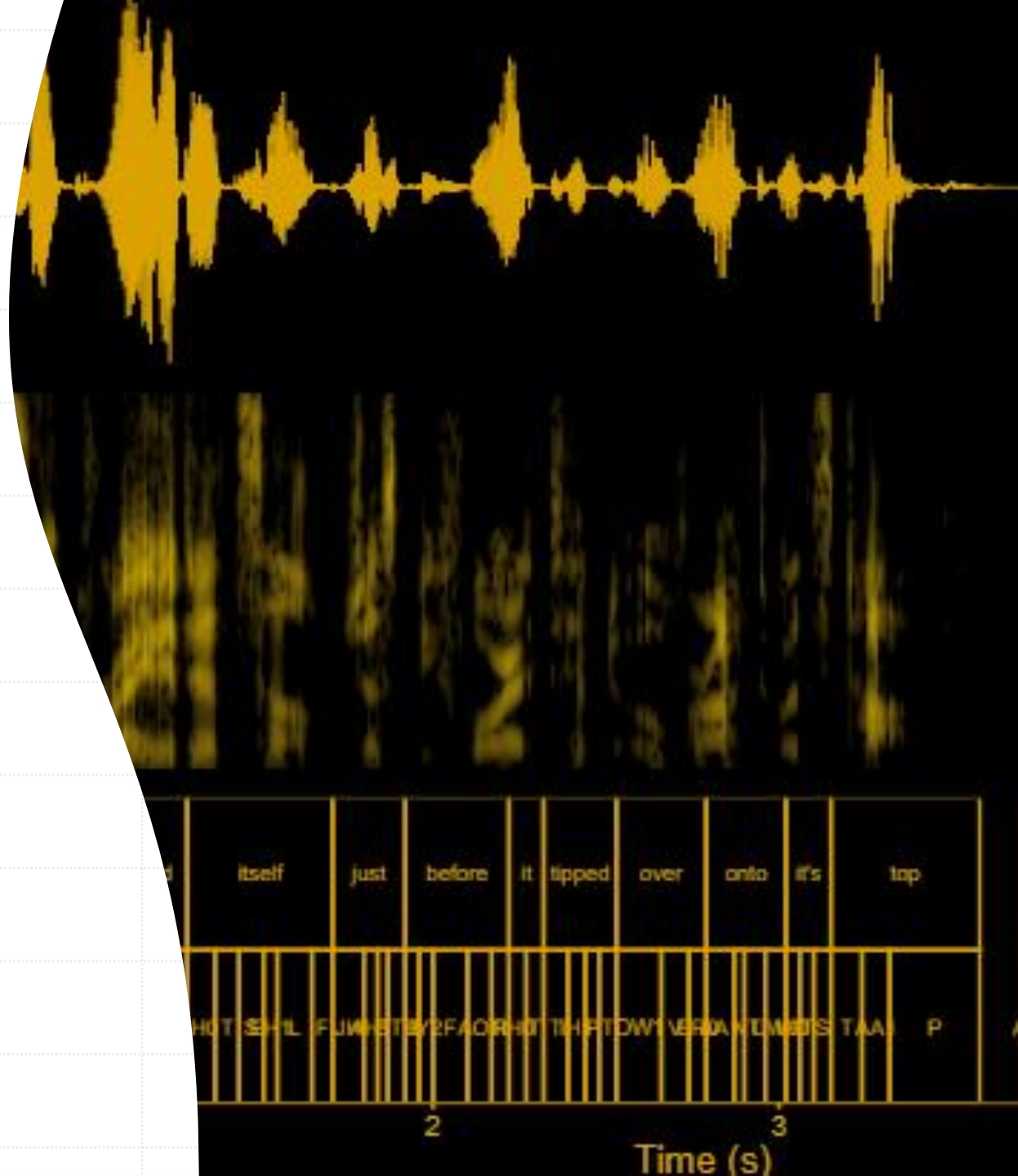
Editing lips movement

- Wav2Lip is a deep learning-based state of art model designed specifically for lip-syncing
- Wav2Lip provides highly accurate lip movements that align perfectly with the audio input, ensuring a natural and seamless visual experience, especially when replacing audio in video content



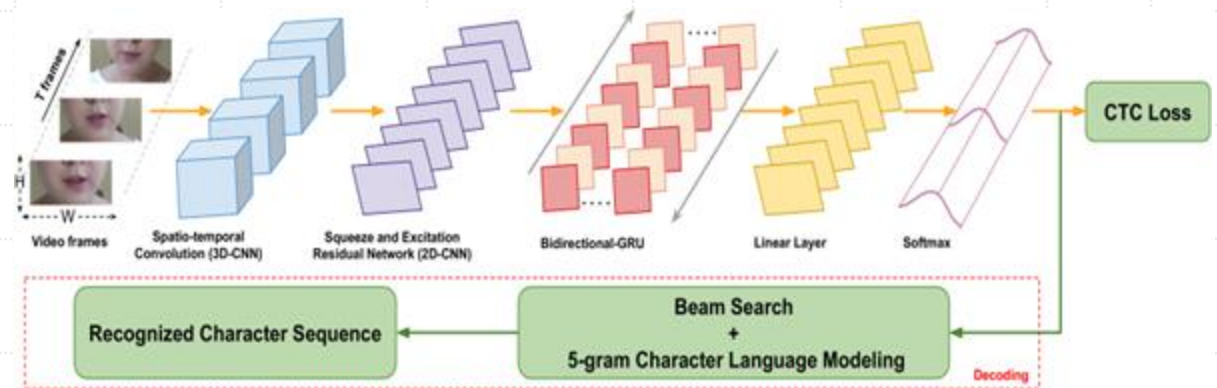
Forced Alignment tools

- Forced alignment tools are used to automatically align audio with the corresponding text transcript
- For this we are using Montreal Forced Aligner (MFA) which is an open-source tool used for aligning audio with text transcripts at the word or phoneme level. It automatically provides precise timestamps for when words and phonemes are spoken in audio files



Accuracy metrics

- **Phoneme-to-Viseme Alignment:**
Ensure the smallest unit of sound produced by the audio matches the corresponding smallest unit of visual speech on the lip-sync video
- **Frame difference metrics:**
Ensures that there are no sudden changes in the visual appearance of the face (especially around the mouth) between consecutive frames, which would make the lip-sync appear unnatural.



Interface for model

Input section

1. Allows user to upload a Video
2. Select a word to replace
3. Provide a replacement word

Output Section

1. Display the output video along with input video
2. Display the matrix of comparison





Challenges

- In a feasible time to maintain the original voice of the speaker while plugging the voice of the replaced word
- Ensuring precise word detection and replacement in the video without disturbing surrounding words or audio.
- Maintaining accurate synchronization between the modified audio and the speaker's lip movements, a critical component for visual consistency.
- Handling various accents, pronunciations, and speech speeds in Hindi, which might affect the word replacement process.





Potential solution proposed

- **Maintaining Original Speaker's Voice:**
Using **voice cloning** or **speech synthesis** techniques, such as **voice conversion models**, to generate the replaced word in the speaker's voice.
- **Precise Word Detection and Replacement:**
Apply **forced alignment** techniques to precisely time-align the original audio transcription with the video. Using **phoneme-level ASR**, the exact start and end of the word can be detected and replaced without affecting surrounding words, ensuring smooth transitions in both audio and video.
- **Lip-Sync Accuracy:**
Implement **phoneme-to-viseme mapping** and using models like **SyncNet** for accurate lip synchronization. This will ensure the replaced word matches the speaker's lip movements frame-by-frame, preserving visual consistency even during dynamic facial movements.
- **Handling Accents and Pronunciations:**
Utilize a **multilingual or accent-robust ASR model** such as **wav2vec 2.0** fine-tuned on diverse Hindi datasets.