CONDITIONED REPLACEMENT OF SPECIFIC WORDS IN HINDI LANGUAGE VIDEOS WITH CORRESPONDING LIP-SYNC.

IE 643 COURSE PROJECT

Team : Mind Guardians

Harshil singh Juneja (22B0666)

Shivamkumar (22B0745)

OUTLINE

PROBLEM STATEMENT

Given a video and two words, the task is to replace a specific word spoken in the video with a new word, ensuring accurate lip-sync. The solution involves identifying the timestamp of the word to be replaced, generating the audio for the new word, and synchronizing the new audio with modified lip movements in the video, resulting in a seamless replacement of the original word

WORKFLOW

•**Input:** Receive a video and two words (old word to replace, new word to insert).
•**Timestamp Extraction:** Use Google API to extract the timestamp of the word to be replaced.
•**Audio Generation:** Generate audio for the new word using GTTS (Google Text-to-Speech).
•**Audio Replacement:** Replace the old word’s audio segment with the new word’s audio using the identified timestamp.
•**Lip-Sync Modification:** Modify the lip movements in the video using Wav2Lip model for accurate synchronization.
•**Voice Cloning :**Use Real-Time Voice Cloning (CorentinJ) for matching the original speaker's voice tone.
•**Output:** Produce a final video with the new word seamlessly replacing the old word with correct lip-sync.

WORK DONE BEFORE PREP-PRESENTATION

•**Initial Approach:**
•Extracted the audio from the video and performed Automatic Speech Recognition (ASR) to get the transcript.
•Used Whisper to locate timestamps for the identified words in the transcript.
•Faced challenges with timestamp accuracy as Whisper provided timestamps for phrases, not individual words.
•**Voice Cloning Research:** Explored Real-Time Voice Cloning (CorentinJ) to match the voice of the speaker in the original video for generating new word audio.
•**Limitations Identified:** Inaccurate word-level timestamps from Whisper led to synchronization issues. Decided to switch to Google API for precise timestamp extraction for individual words.

WORK DONE AFTER PREP-PRESENTATION

- **Timestamp Extraction Improvement:** Replaced Whisper with Google API for extracting accurate word-level timestamps, solving synchronization issues.
- **Audio Generation:** Generated audio for the new word using GTTS (Google Text-to-Speech), ensuring it aligns with the identified timestamp.
- **Lip-Sync Enhancement:** Integrated Wav2Lip model (Rudrabha's GitHub repo) for realistic lip-sync adjustments in the video based on the new word's audio.

CONTINUE...

•**Voice Cloning :** Real-Time Voice Cloning (CorentinJ) to match the original speaker's voice tone, to enhance the naturalness of the new audio.
•**Evaluation:** Used Frame to frame optical consistency score to know the flow of frames in video
•**Gradio:** Integrated frontend in the code using Gradio to take user input and display the result.

PROPOSED-APPROACH

- **Approach 1: Timestamp Extraction Using Whisper**
- **Description:**
  - Extracted audio from the input video and used an Automatic Speech Recognition (ASR) model for transcript generation.
  - Applied Whisper to locate timestamps for phrases containing the target word.
- **Challenges:** Only phrase-level timestamps were provided, requiring equal distribution of time across words, leading to inaccurate word-level synchronization.
- **Approach 2: Google API for Accurate Timestamps**
- **Description:**
  - Switched to using Google Speech-to-Text API for precise word-level timestamp extraction.
  - Ensured higher accuracy and synchronization by directly obtaining the exact timestamp of the word to be replaced.
- **Outcome:** Achieved better word localization and minimized sync errors.

CONTINUE...

- **Approach 3: Lip-Sync Enhancement Using Wav2Lip**
- **Description:**
  - Leveraged the Wav2Lip model to modify the lip movements in the video based on the newly generated audio.
  - Ensured that the lips of the speaker in the video match the pronunciation of the replaced word.
- **Outcome:**
  - Realistic and accurate lip-sync, even for unseen identities.
- **Approach 4: Voice Cloning Using Real-Time Voice Cloning (CorentinJ)**
- **Description:**
  - Used Real-Time Voice Cloning to generate audio for the new word that mimics the voice tone and style of the original speaker.
  - Enhanced the naturalness of the audio, making the replacement seamless..

HARDWARE USED

- GPU: NVIDIA Tesla P100 (16 GB VRAM) / NVIDIA Tesla T4 (16 GB VRAM)
- CPU: Intel Xeon Processor (Up to 2.3 GHz)
- RAM: 13 GB

CONCLUSION

•**Project Summary:**
•Developed an innovative pipeline for replacing specific words in a video with another word while maintaining accurate lip-sync.
•Integrated state-of-the-art models for timestamp extraction (Google API), lip-sync (Wav2Lip).
•Explored and refined multiple approaches to achieve seamless word replacement.
•**Key Achievements:**
•Successfully overcame challenges in timestamp accuracy and lip-sync quality.
•Achieved word replacement with correct time stamp.
•Demonstrated the effectiveness of combining advanced AI techniques for audio and video synchronization tasks.
•**Drawback :** The lip sync is not visually indistinguishable. This is due to less effective audio converstion.

NOVELITY ASSESSMENT

- **Objective:**
- Develop a customized GAN-based lip-sync model specifically trained for Hindi language videos.
- **Proposed Steps:**
- **Dataset Collection:**
    - Gather a diverse dataset of Hindi language talking face videos covering various accents and speakers.
- **Model Development:**
    - Design a GAN-based architecture tailored for Hindi lip-sync, leveraging the specific phonetic characteristics of the language.