AI BASED MULTILINGUAL HUMAN VIDEO TRANSLATION

Project Reference No.: *49S_BE_6737*

College : *Sri Venkateshwara College of Engineering, Bengaluru*
Branch : *Department of Computer Science and Engineering - Data Science*
Guide(s) : *Prof. Meghana K L*
Prof. Anbulakshmi S
Student(s) : *Mr. Abhinek Kumar Pandey*
Mr. Md Raqibul Islam Khan
Mr. Anil Kumar G N
Mr. Bharath R

Keywords:

Artificial Intelligence, Automatic Speech Recognition, Neural Machine Translation, Text-to-Speech, Lip-Sync.

Introduction:

The digital era has led to an explosion of video content across platforms, but language remains a significant barrier to global accessibility. Traditional video dubbing and translation processes are manual, time-consuming, and expensive. Furthermore, dubbed videos often lack the natural feel of the original speaker due to mismatched lip movements and altered vocal prosody. To address this challenge, an AI-powered multilingual video translation system is proposed to automate the conversion of spoken content into multiple target languages with high accuracy. This system bridges the communication gap by enabling seamless localization for education, entertainment, and communication sectors. By integrating state-of-the-art Deep Learning models, the system ensures that translated videos retain the original speaker's essence through synchronized facial movements and expressive, natural-sounding audio.

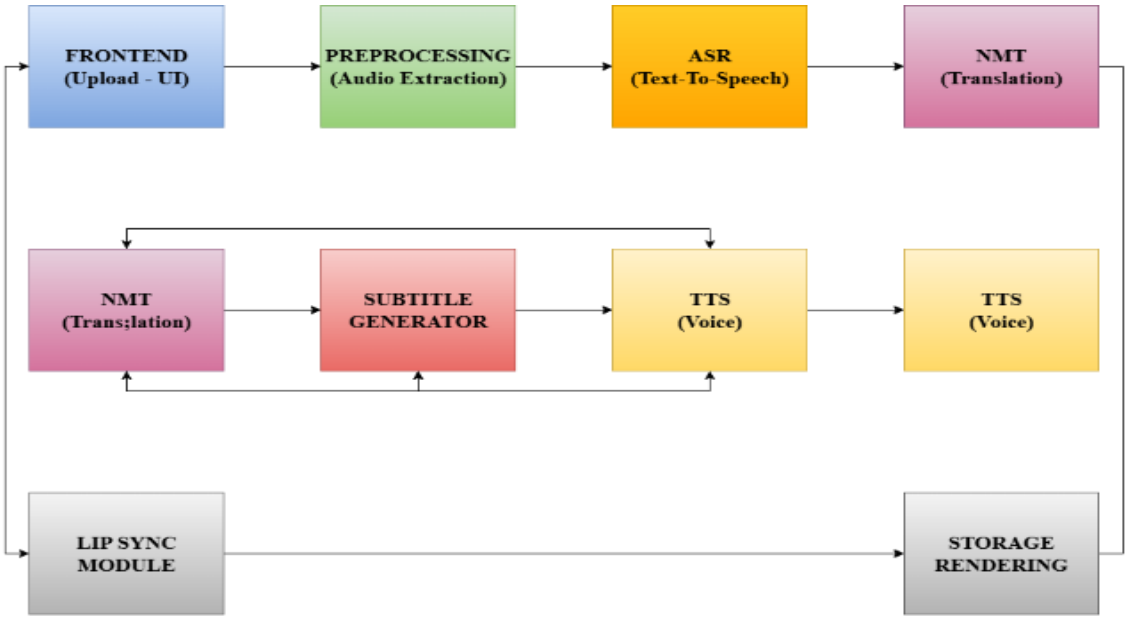

Figure 1: System Architecture of the Human Video Translation Pipeline

Objectives:

1. To develop an AI-powered multilingual video translation system capable of converting spoken content into multiple target languages with high accuracy.
2. To integrate Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Text-to-Speech (TTS), and Lip-Sync modules into a unified pipeline.
3. To ensure natural-sounding translated speech with prosody-aware synthesis and accurate subtitle alignment.
4. To achieve realistic lip synchronization between translated audio and the speaker's facial movements.
5. To improve accessibility and global reach of video content by enabling seamless localization.

Methodology:

The project follows a modular, AI-driven pipeline that processes input video and generates a fully translated, lip-synchronized output.

1. **Video Preprocessing:** Audio and frames are extracted from the input video to prepare data for the subsequent modules.
2. **Automatic Speech Recognition (ASR):** Speech is converted to text using a transformer-based ASR (Whisper) with CTC loss to generate time-stamped transcripts.
3. **Neural Machine Translation (NMT):** The source transcript is translated into the target language using transformer-based NMT with beam search and length normalization.
4. **Subtitle Alignment:** Dynamic Time Warping (DTW) is utilized to align translated text with the original timing, generating readable and time-accurate subtitles.
5. **Prosody-Aware Text-to-Speech (TTS):** The translated text is converted into natural-sounding speech by modeling pitch, duration, and energy for expressive output.
6. **Lip-Sync Refinement:** Phonemes are mapped to visemes, and GAN-based neural refinement is applied to synchronize mouth movements with the synthesized audio.
7. **Rendering & Output Generation:** Finally, the synchronized video frames, translated audio, and aligned subtitles are combined using FFmpeg to produce the final localized video.

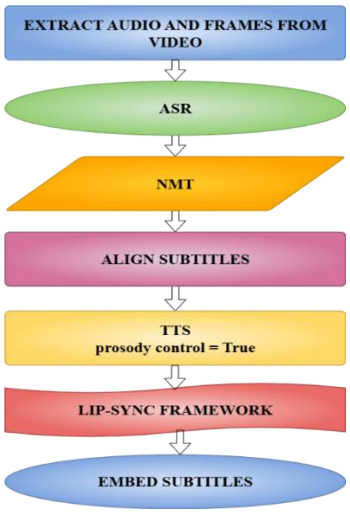

Figure 2: Flow Chart showing the Processing Stages from Upload to Rendered Output

Result and Conclusion:

In conclusion, the project culminates in a fully functional AI-powered multilingual video translation system. The integration of the ASR engine ensures highly accurate transcriptions, while the NMT module provides contextually precise translations. By utilizing DTW-based alignment and prosody-aware TTS, the system successfully generates subtitles and synthesized speech that match the emotional tone and timing of the original video. Furthermore, the lip-sync module seamlessly adjusts the speaker's facial movements to the new audio track, eliminating the visual discrepancies common in traditional dubbing. In conclusion, the system delivers natural, synchronized, and high-quality localized videos, providing a scalable and modular architecture for audiovisual translation.

References:

1. [1] A. Vaswani et al., "Attention is All You Need," NeurIPS, 2017.
2. [2] A. Graves et al., "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data," ICML, 2006.
3. [3] N. Kalchbrenner et al., "Sequence to Sequence Learning with Neural Networks," NIPS, 2014.
4. [4] R. Prabhavalkar et al., "Automatic Speech Recognition with Transformers," IEEE ASRU, 2020.
5. [5] K. Park et al., "Lip Sync Error Detection for Automated Dubbing," ACM MM, 2021.
6. [6] B. Li et al., "Multilingual End-to-End Speech Translation," ICASSP, 2021.
7. [7] TRAVID: An End-to-End Video Translation Framework, arXiv:2309.11338v1, 2023.
8. [8] VideoDubber: Machine Translation with Speech-Aware Length Control, arXiv:2211.16934v2, 2023.
9. [9] Bridging Language Barriers: AI in Real-Time Multilingual Translation, IJIRT, 2024.

Future Scope:

The future scope of this project includes:

1. Scaling the architecture to support additional languages and upgrading the core modules with future advancements in speech recognition, translation, and synthesis technologies.
2. Integrating multi-speaker diarization to accurately handle videos with multiple participants and improving emotion-preserving TTS to capture subtler vocal nuances.
3. Deploying the application on optimized cloud GPU clusters to facilitate real-time or near-real-time streaming video translation and to minimize rendering times for high-resolution formats.