# AUDIO TRANSLATOR AND LIP SYNC IN VIDEO

***Project Reference No.:*** *48S_BE_4318*

*College : Sri Siddhartha Institute of Technology, Tumakuru*
*Branch : Department of Computer Science and Engineering*
*Guide(s) : Mrs. K Noor Fathima*
*Student(s): Mr. Manorakith*
*Mr. Rachamalla Ganesh*
*Mr. Neelam Bhaskar Reddy*
*Mr. Puneeth D S*

**Keywords:**

Speech Translation, Lip Sync, Video Processing, Deep Learning, TTS (Text-to-Speech), Whisper, Coqui TTS, Wav2Lip, Audio-Visual Synchronization, Multilingual Content.

**Introduction:**

With the rise of globalization and online content consumption, language barriers have become a significant challenge for reaching diverse audiences. While subtitles are a common solution, they often fail to deliver an immersive experience. Our project addresses this gap by developing a system that translates spoken audio in a video to another language and synchronizes the translated speech with the speaker's lip movements. This combination of audio translation and lip synchronization enhances accessibility and realism, especially in education, entertainment, and communication. The system leverages state-of-the-art pre-trained models for speech-to-text, translation, text-to-speech (TTS), and lip-syncing, ensuring efficiency and accuracy

**Objectives:**

1. To develop a tool that translates spoken language in a video into another language
2. To generate natural-sounding speech using a TTS engine while preserving the speaker's original gender
3. To synchronize lip movements with the translated audio for realistic output

4. To enhance accessibility and engagement for multi-lingual content consumers

**Methodology:**

The project works as follows:

1. Speech-to-Text (STT): Extract audio from the input video and transcribe using Whisper.
2. Translation: Translate the transcribed text using APIs such as Google Translate.
3. Gender Detection: Analyze original voice to detect gender using ML-based models.
4. Text-to-Speech (TTS): Convert translated text into audio using Coqui TTS with appropriate gendered voice.
5. Lip Synchronization: Use Wav2Lip to synchronize video with the new translated audio.
6. Video Editing & Merging: Merge updated visuals and audio using FFmpeg and Python scripting

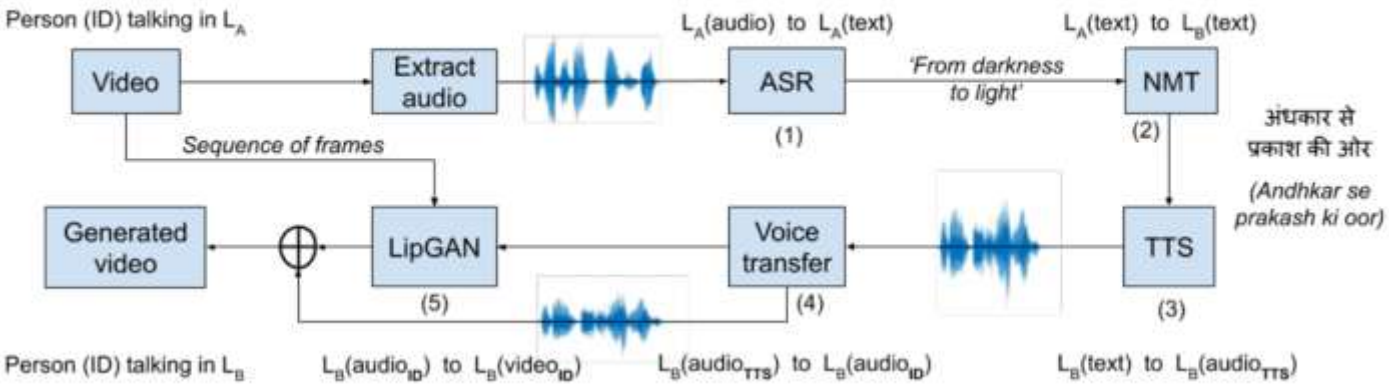


Figure 1: Basic workflow

**Result and Conclusion:**

In conclusion, The system was successfully implemented using publicly available pre-trained models. The translated audio matched the lip movements in the video with high accuracy. The gender-detection mechanism preserved the speaker's voice identity, enhancing the naturalness of the output. Visual results showed synchronized mouth

movements across various video scenarios. Audio-visual evaluations confirmed improved viewer comprehension and immersion.

**Future Scope:**

The future scope of this project includes:

1. Extending the system to support real-time video conferencing with live translation and lip sync
2. Expanding the language base to include regional languages.
3. Censoring inappropriate words using “beep”.
4. Optimizing the pipeline for real-time performance on resource-constrained devices.