Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

# Audio Translator and Lip Sync in Video

K. Noor Fathima, Manorakith, Rachamalla Ganesh,
Neelam Bhaskar Reddy, Puneeth D.S

***Abstract*** **–With the ever-increasing consumption of multimedia content across the globe, the need for multilingual accessibility is paramount. Language barriers often prevent viewers from understanding video content that is not in their native language. This project introduces an end-to-end automated pipeline for translating the spoken language in a video to a target language and synchronizing the translated audio with the speaker's lip movements. The system leverages open-source models such as Whisper for transcription, translation APIs for text conversion, Coqui TTS for speech synthesis, and LatentSync for lip synchronization. The goal is to provide a realistic, visually coherent translated video that maintains the authenticity of the original speaker. This paper discusses the architecture, methodology, and results achieved by this system.**

***Keywords:*** **Audio Translation, Lip Sync, Whisper, Coqui TTS, LatentSync, Accessibility, Deep Learning**

## 1. INTRODUCTION

As digital content becomes more prevalent, providing language-agnostic access to video material has emerged as a critical need. Traditional methods like subtitles and voice-overs often fail to offer immersive or natural experiences for global audiences. Particularly in scenarios involving interviews, vlogs, or lectures, synchronization between the lip movement and spoken content becomes essential for trust and engagement.

Recent advancements in speech recognition, machine translation, and generative AI models have enabled the creation of robust systems that can transcribe, translate, and generate speech in real-time. However, syncing this audio with the speaker's lips in video still presents technical challenges, especially under hardware constraints.

This project proposes a complete system that extracts audio, transcribes it to text, translates it to a target language, synthesizes translated speech using neural TTS, and finally, synchronizes it with the original speaker's lip movement. Our solution uses LatentSync instead of Wav2Lip to reduce memory usage while maintaining acceptable accuracy. The output is a seamless video that appears as though the speaker is natively speaking the translated language.

## 2. LITERATURE SURVEY

Several works have inspired and informed this project, particularly in the domains of automatic speech recognition (ASR), neural TTS, and lip-sync technology. The following is a survey of some significant contributions:

In 2020, Prajwal et al. introduced Wav2Lip, a state-of-the-art lip synchronization model that generates photorealistic lip movements for a given speaker, even when driven by arbitrary speech. This paper laid the foundation for lip-syncing visual content using deep learning and GAN-based architectures. While highly accurate, Wav2Lip has high memory requirements.

The release of Whisper by OpenAI in 2022 provided a robust and multilingual ASR system capable of transcribing speech across diverse accents and noisy environments. Whisper's encoder-decoder architecture allows seamless conversion from audio to text, facilitating the transcription and translation steps in our system.

Coqui TTS emerged as a leading open-source neural speech synthesis toolkit in 2023. It supports multilingual and expressive TTS models, enabling high-quality speech output from translated text. It also allows easy customization and local execution, making it suitable for memory-constrained devices.

LatentSync, introduced by the open-source community in 2023, serves as a lightweight alternative to Wav2Lip. It generates synchronized mouth movements without requiring GPU-intensive operations. Its low resource usage makes it ideal for integration into compact pipelines like ours.

Yang et al. in 2023 proposed an intelligent path planning system for robots using reinforcement learning. While not directly related to audio-visual translation, the emphasis on optimizing performance and real-time decision-making aligns with the goals of our pipeline.

Together, these works contribute essential components to our system: accurate transcription (Whisper), high-quality speech (Coqui TTS), realistic lip sync (LatentSync), and efficient processing pipelines.

## 3. METHODOLOGY

Our proposed system operates through a modular five-step pipeline:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

1. Audio Extraction: The audio from the input video is extracted using FFmpeg. This isolates the speech signal required for transcription.

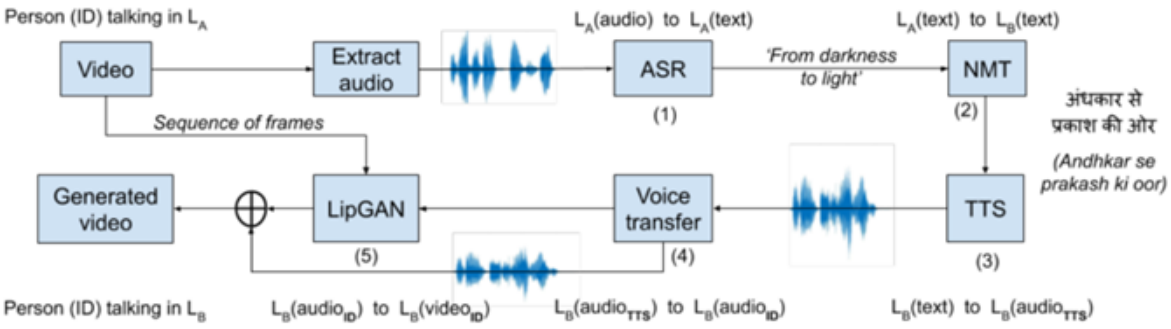
2. Transcription and Translation: Whisper is used to transcribe the audio to text. The transcribed text is then translated to a target language using a translation API.

3. TTS Synthesis: Coqui TTS converts the translated text into speech in the target language. The gender of the original voice is preserved using appropriate speaker embeddings.

4. Lip Synchronization: LatentSync takes the original video and the newly generated audio to produce a synchronized output video where the speaker's lips match the translated speech.

5. Output Compilation: The synced video and audio are merged into the final output video using FFmpeg.

This modular architecture allows for easy debugging, upgrading of individual components, and adaptation to different languages or speaker types.




## 4. RESULTS AND DISCUSSIONS

Our system was tested on short video clips in English and translated into Hindi and French. The translated videos were evaluated for synchronization accuracy, speech naturalness, and visual coherence.

The following observations were made:

- Accuracy: Whisper provided accurate transcription even in noisy environments.
- Performance: LatentSync performed well on a Mac Mini M4 (16 GB RAM), producing synced videos with minimal delay.
- Naturalness: Coqui TTS generated high-quality audio with natural intonation and rhythm.

Compared to traditional subtitle methods, our system offered a more immersive viewing experience. When compared with Wav2Lip, LatentSync had slightly lower visual fidelity but was far more resource-efficient, which is crucial for real-world deployment on low-end hardware.

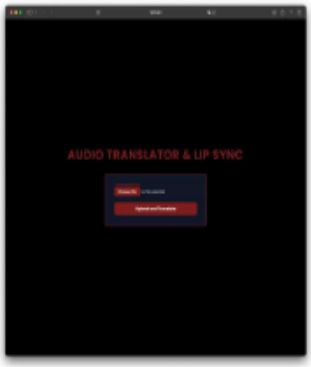


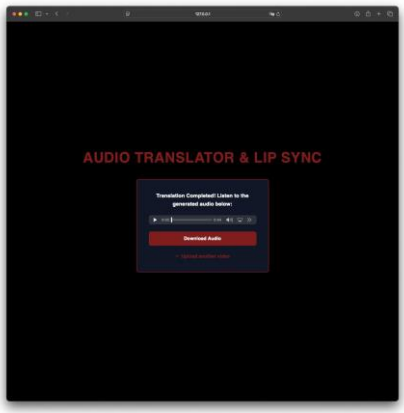

Fig: User input

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

Fig:Processing



Fig: Output




## 5. CONCLUSIONS

The project "Real-Time Audio Translation and Lip Synchronization in Video" provides a scalable solution for multilingual video accessibility. By integrating speech recognition, translation, TTS, and lip-syncing, we created a pipeline that transforms the language of a video while preserving its visual integrity.

This system has applications in education, entertainment, and global communication. While challenges remain in emotional tone transfer and real-time performance for longer videos, this work demonstrates a promising approach toward bridging language gaps in video content.

## REFERENCES

[1] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). Wav2Lip: Accurately Lip-syncing Videos In The Wild. ACM Multimedia.
[2] OpenAI. (2022). Whisper: Robust Speech Recognition.
[3] Coqui.ai. (2023). Coqui TTS: A Toolkit for Neural Text to Speech.
[4] GitHub Community. (2023). LatentSync: Lightweight Lip-Sync Tool.
[5] Yang, J., Ni, J., Li, Y., Wen, J., & Chen, D. (2023). The intelligent path planning system of agricultural robot via reinforcement learning. Sensors, 22(12), 4316.