Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

AI Video Translation Using Lip Synchronization

Mrs. Ketakee Ghawade[1], Darshana Menghare[2], Tejaswini Samrit[3], Om Padgilwar[4], Param Chauragade[5]

[1] Assistant Professor, Department of Computer Technology, KITS Ramtek, India
[2,3,4,5] Department of Computer Technology, KITS Ramtek, India

Abstract
As digital content continues to grow around the world, language differences still make it hard for people to communicate effectively in video calls. This project introduces an AI-based video translation system that automatically matches the translated audio with the speaker's facial movements. This allows videos to be translated into several languages while keeping the speaker's expressions natural. The system first turns spoken words into text using speech recognition technology. Then, it uses neural machine translation to convert the text into the desired language. After that, text-to-speech technology creates the translated audio. A separate lip-syncing model ensures that the audio matches the speaker's mouth movements, making the results look and sound realistic. This solution helps make videos more accessible, improves how people interact with content, and supports communication across different languages in areas like education, entertainment, and business. Tests show that the system gives accurate translations and smooth video outputs, cutting down the need for manual voice-over work. This method shows how AI can improve communication in videos and help people from different language backgrounds understand and enjoy content better.

Keywords: Artificial Intelligence, Video Translation, Lip-Syncing, Speech Recognition, Neural Machine Translation

1. Introduction

In today's digital world, video content has become a major way of communicating in education, entertainment, business, and social media. Every day, millions of videos are posted on different platforms, reaching people from all around the world, including those who speak different languages and come from different cultures. However, language differences often make it hard for people to fully understand and access this content. Older methods like subtitles and manual voice-over translations can be slow, expensive, and sometimes don't feel natural to watch. Recent developments in AI have made it possible to use automatic speech recognition, machine translation, and speech synthesis with high accuracy. These tools allow spoken words in videos to be translated into other languages. But just translating and dubbing the voice doesn't always match the speaker's mouth movements, which can make the video feel less real and less engaging.

To fix this, new AI techniques have been created to match translated audio with the speaker's face movements. By using speech-to-text, neural machine translation, text-to-speech, and lip-syncing models together, it's now possible to create videos that look and sound natural, with the voice and mouth

movements in sync. This project introduces an AI-powered video translation system that automatically syncs translated audio with the speaker's lips.

The goal is to remove language barriers and help people from different language backgrounds connect better. The system makes videos more accessible, interesting, and clear for people who speak multiple languages. It can be used in online learning, content creation, advertising, and sharing ideas across cultures**.**

2. MAIN SECTION – I

2.1 LITERATURE REVIEW

Video translation and dubbing used to be done by hand, which took a lot of time, money, and human work. Old methods mostly used subtitles or voice-over, but these often didn't match the speaker's lip movements. This mismatch made the videos feel less natural and reduced how much people engaged with the content.

New research in artificial intelligence has brought automated tools for speech recognition, language translation, and speech synthesis. Automatic Speech Recognition (ASR) systems are now very good at turning spoken words into text. Neural Machine Translation (NMT) has made translations better by using deep learning models that understand language context more deeply. Some studies have also looked into AI-based lip-syncing techniques.

Deep learning models like generative adversarial networks (GANs) and facial landmark detection help match lip movements with the generated speech. These techniques help make videos look more realistic and synchronized.

But many current systems only focus on either translation or dubbing and don't fully include lip-syncing. Because of this, there's a need for a system that combines translation and lip-syncing to create natural multilingual videos. The proposed system fills this gap by bringing together ASR, NMT, Text-to-Speech (TTS), and lip-syncing into one process.

2.2 METHODOLOGY

The proposed AI-driven video translation system with lip-syncing follows a clear process that uses several AI methods. The steps are as follows:

1.Video Input
The system takes in a video that includes someone speaking.
The video is broken down into its audio and visual parts.

2.Audio Extraction
The audio part of the video is taken out.
This audio is then used by the speech recognition part of the system.

3.Speech-to-Text Conversion
The system uses automatic speech recognition to turn the spoken words into text.
This makes sure the original message is captured accurately.

4.Language Translation

The text is translated into the desired language using neural machine translation.This translation is smart and considers the context for better meaning.

5.Text-to-Speech Generation
The translated text is then turned back into speech using text-to-speech technology.The speech sounds natural and clear.

6.Lip-Syncing
An AI model is used to match the generated speech with the speaker's mouth movements in the video.This is done by analyzing facial features and using deep learning techniques to make the lips move in time with the speech.

7.Output Video Generation
Finally, the synchronized audio is added back to the original video to create a translated video that looks and sounds natural with realistic lip movements.

2.3 TOOLS AND TECHNOLOGIES USED

Creating an AI-powered video translation system with lip-syncing involves using a mix of software tools, programming frameworks, and AI models. Here's a breakdown of the main tools and technologies used:

1.Programming Language

Python is the main language used because it's easy to work with and has great support for AI and machine learning tools.

2.Speech Recognition Tools

Tools like OpenAI Whisper and Google Speech-to-Text are used to turn spoken words into text with high accuracy.

3.Machine Translation

Models such as Google Translate API or transformer-based models are used to translate the text into various languages.

4.Text-to-Speech (TTS)

TTS tools like gTTS and Tacotron are used to create natural-sounding speech from the translated text.

5.Lip-Syncing Models

Deep learning models like Wav2Lip are used to align the generated speech with the lip movements in the video.

6.Video Processing

Libraries like OpenCV and MoviePy are used for editing videos, extracting frames, and combining audio with video.

7.Development Environment

Google Colab or Jupyter Notebook is used for developing and testing the models because they offer GPU support and are easy to use**.**

2.4 SYSTEM ARCHITECTURE

The system for AI-powered video translation with lip-syncing is made up of several connected parts that work one after another to create the final translated video. The design of the system makes sure that data moves smoothly from one step to the next. The process starts with the input video module, where the user gives a video that has spoken words. Then, the video goes to the audio extraction module, which takes out the audio part from the video. Once the audio is extracted, it goes to the Automatic Speech Recognition (ASR) module. This part turns the spoken words into text. The text is then sent to the Neural Machine Translation (NMT) module, which translates the text into the language the user wants. After the translation is done, the text moves to the Text-to-Speech (TTS) module.

This module creates speech in the target language. The new speech is then given to the lip-syncing module, which uses deep learning to match the audio with the speaker's mouth movements. Finally, the lip-synced audio is combined with the original video in the output generation module, resulting in a translated video that looks natural and realistic**.**

A . Workflow Architecture

Figure 1 shows the overall process of the system. When a user uploads a video through the interface, the system takes out the audio for further processing. The audio is then sent to the ASR module, which converts the spoken words into text. This text goes to the translation module, where it is translated into the desired language using a neural machine translation model. The translated text is then turned into speech by the TTS module. This speech is then combined with the original video. There is also an error-handling system that deals with problems that might occur during recognition or translation. The final video is ready for preview and download. This workflow creates a continuous and automatic process for translating videos into multiple languages.

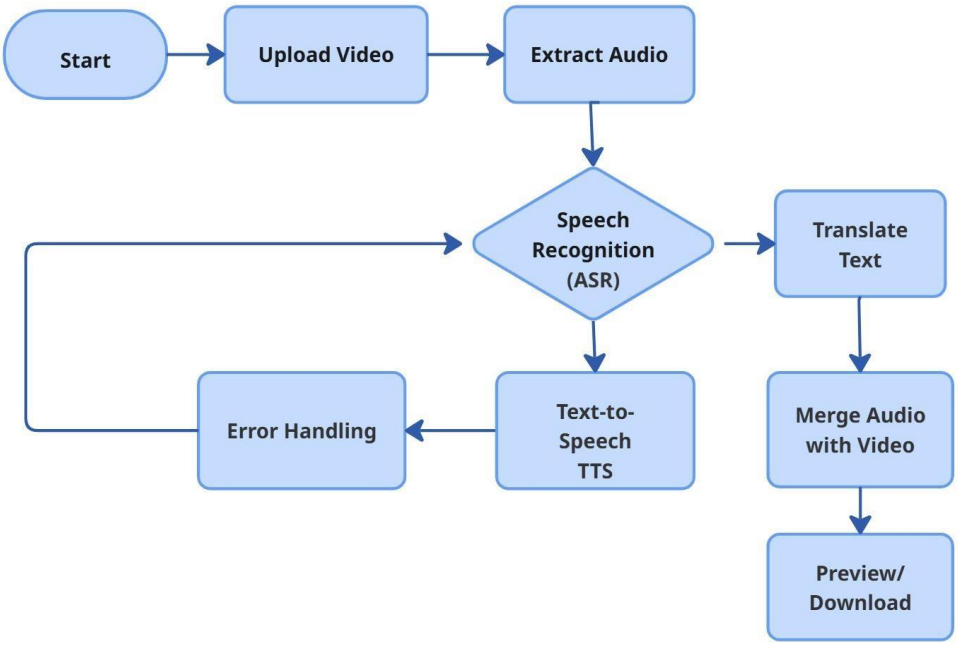

Fig. 1. System Architecture of AI-Powered Video Translation with Lip-Syncing

B. Layered Architecture

Figure 2 shows the system's layered architecture. The system is organized into several layers to make it modular and improve how data flows through it. The User Interface Layer lets users upload videos, choose languages, and preview or download the final results.This part of the system is all about making the user experience easy and accessible. The Application Layer handles communication between different parts of the system using tools like Flask and NodeJS.It takes care of processing requests and keeping everything in sync. The AI Processing Layer does the main work of the system. It handles tasks like speech recognition, translation, converting text to speech, syncing lips with audio, detecting faces, replacing audio, and rendering the final video. This layer gives the system its smart features needed for automatic video translation.

The External Services Layer uses third-party platforms and APIs to run AI models. These services help with the heavy processing needed for the models to work properly. The Data Storage Layer keeps video and audio files temporarily while they are being processed. This layered setup helps the system scale better, be easier to maintain, and perform more efficiently, making it ready for use in real-world scenarios.

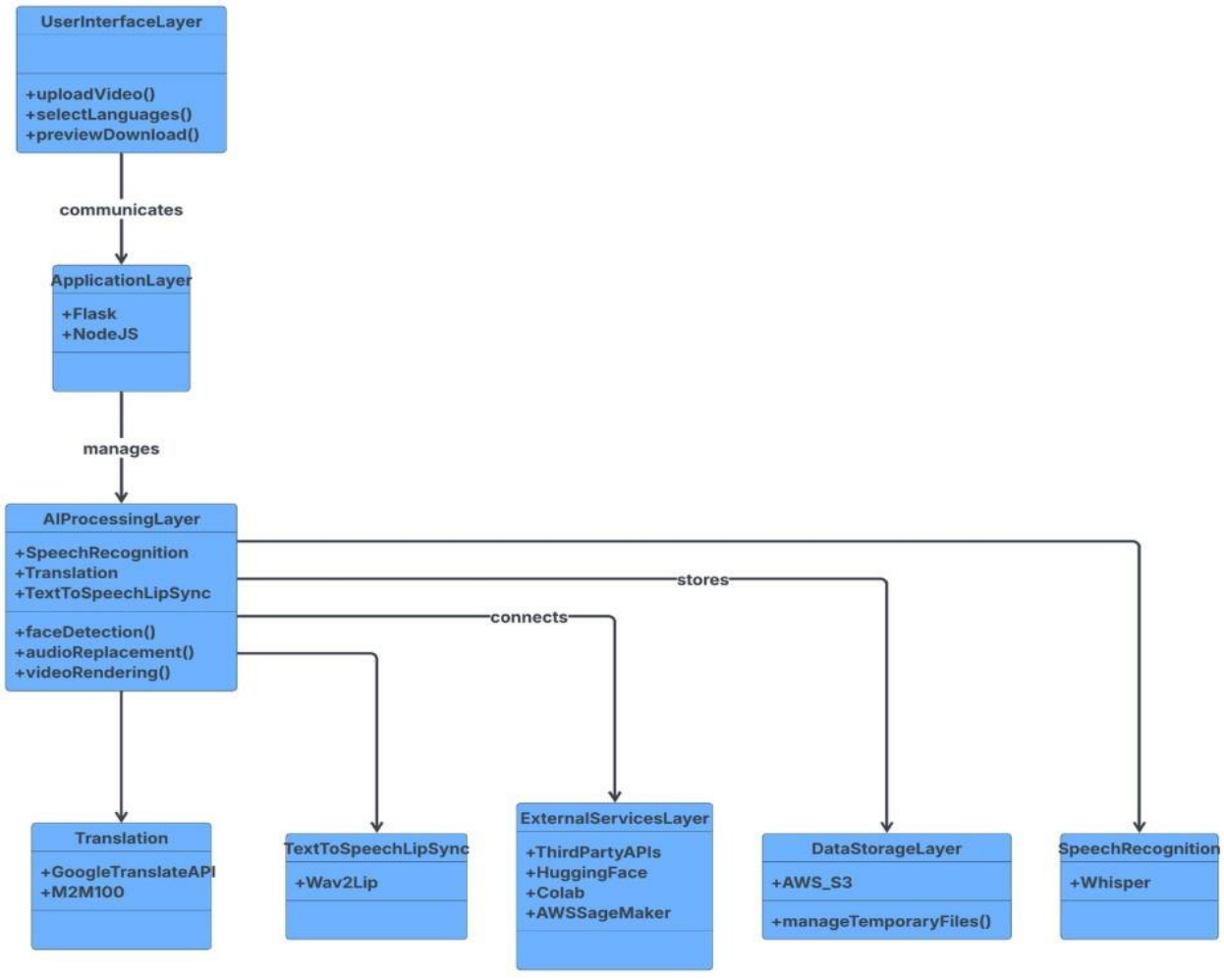

Fig. 2. Layered System Architecture of the Proposed System

3. MAIN SECTION – II

3.1 RESULTS AND DISCUSSION

The AI-powered video translation with lip-syncing system was tested using several sample videos that had spoken language. The system worked well in pulling out the audio from the videos and turning speech into text with decent accuracy when the audio was clear. The translation part of the system gave correct translations for languages that are commonly used. The text-to-speech part created speech that sounded natural, which made the final translated videos better in quality.

The lip-syncing part matched the generated speech with the speaker's lip movements very well. The final videos looked real, with very little delay between the audio and what was shown on screen. This made viewers more engaged compared to old dubbing methods. It took about 2 to 3 minutes to process a one-minute video, depending on how fast the system was and how good the internet connection was. It was found that background noise and unclear speech slightly reduced the accuracy of speech recognition.

Overall, the system worked reliably for translating and syncing videos in multiple languages.Using AI-based modules made the process faster and less work-intensive than traditional video dubbing methods.

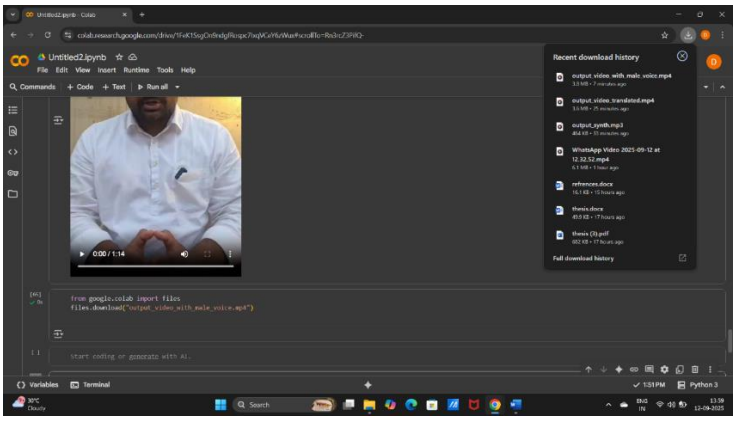

Fig. 3. Output Video Generated After Translation and Lip-Syncing

4. CONCLUSION

The AI-powered video translation with lip-syncing system offers a great way to improve communication across different languages. It combines speech recognition, machine translation, text-to-speech technology, and lip-sync techniques to create videos that have accurate translations and matching facial movements. The tests show that this system can deliver clear translations and natural-sounding voices, while keeping the lips in sync with the spoken words. This helps cut down on the need for manual dubbing and creating subtitles, which saves time and effort.

This system is especially helpful in areas like online learning, making media content, and connecting people around the world. While it might not work perfectly in noisy environments or when speech is unclear, the results overall show it's a dependable and useful tool. Looking ahead, possible improvements could include adding more languages, faster real-time processing, and better lip-syncing using more

advanced AI models. The system shows how AI can help overcome language barriers and make digital content available to more people.

Acknowledgement

The authors would like to sincerely thank the Department of Computer Technology at KITS Ramtek for their support and resources that helped in completing this work. They also extend their thanks to their project guide, Mrs. Ketakee Ghawade, for her guidance, encouragement, and technical help throughout the project's development. The authors are also thankful to all the faculty members and colleagues who offered their suggestions and assistance during the project's implementation and testing stages. Their support played a key role in the successful completion of this research.

References

1. A. Radford et al., “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI, 2022.
2. Y. Chen et al., “Wav2Lip: Accurately Lip-syncing Videos In The Wild,” Proceedings of ACM Multimedia, 2020.
3. D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” ICLR, 2015.
4. T. Wang et al., “Tacotron: Towards End-to-End Speech Synthesis,” Google Research, 2017.
5. Google, “Google Translate API Documentation,” Available: https://cloud.google.com/translate
6. A. Vaswani et al., “Attention Is All You Need,” NeurIPS, 2017.
7. Hugging Face, “Transformers Library Documentation,” Available: https://huggingface.co/docs
8. AWS, “Amazon S3 Developer Guide,” Available: https://docs.aws.amazon.com/s3