WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

LingoFusion: An Automatic Video Dubbing Suite using Artificial Intelligence

Lalita Panika[1], Aastha Gracy[2], Sanket Mathur[3], S. Hariharan Reddy[4], Tanishqa Sahu[5]

[1]*Assistant Professor, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*
[2,3,4,5] *UG Students, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*
DOI: https://doi.org/10.55248/gengpi.5.0124.0113

A B S T R A C T

In an increasingly interconnected world, the language barrier often stands as an obstacle to the dissemination of valuable content. LingoFusion is a groundbreaking automatic video dubbing application designed to bridge this divide by seamlessly dubbing English YouTube videos into various Indian native languages, such as Hindi, Marathi, Gujarati, Telugu, and Bengali. This innovative solution harnesses a powerful stack of technologies, including the NPM package for accessing YouTube audio, Deepgram API for transcript generation, and Flutter packages for translation and text-to-speech capabilities. It facilitates a seamless cross-cultural exchange of digital content by automatically dubbing videos, making them more accessible and engaging for a wider audience. By utilizing state-of-the-art speech recognition, machine translation, and voice synthesis tools, this app simplifies the process of language adaptation, promoting inclusivity, and enhancing the reach of multimedia content.

Keywords: Computer vision, Natural Language Processing , Auto Dubbing , Artificial Intelligence , Neural Vocoding

1. Introduction

Automatic video dubbing refers to the process of automatically replacing the original audio content of a video with a newly generated audio track in a different language. This technology is designed to overcome language barriers and make digital content more accessible to diverse audiences. The term "dubbing" traditionally refers to the practice of recording and replacing voices in a film or video, typically to provide a translation or adaptation for audiences who speak a different language.

LingoFusion is more than just a tool for linguistic adaptation; it's a cultural bridge that enriches the digital experience for users and content creators alike. The essence of the app lies in its ability to democratize access to content by making it linguistically inclusive. By tapping into advanced technologies and leveraging a strategic combination of tools, LingoFusion ensures that language differences no longer hinder the global sharing of information and entertainment. At the core of LingoFusion's functionality is a sophisticated technological stack. The app employs the npm package to seamlessly retrieve audio from YouTube videos, ensuring a smooth integration with the platform where a wealth of content resides..

1.1 Problem Description

In our interconnected world, where digital content knows no geographical boundaries, the language barrier remains a persistent challenge. English may be a global lingua franca, but a significant portion of the world's population prefers content in their native languages. This preference poses a formidable obstacle to the seamless exchange of information, knowledge, and entertainment across linguistic divides. A substantial amount of online content, particularly on platforms like YouTube, is predominantly available in English. This linguistic bias creates disparities in content accessibility, leaving a vast audience with limited options to engage with the wealth of information and entertainment available on digital platforms. As a consequence, individuals who are not proficient in English may find themselves excluded from valuable educational resources, entertainment content, and informative videos.

Historically, the solution to this language barrier has involved manual dubbing, where human actors re-record the dialogue in the target language. However, this process is labor-intensive, time-consuming, and expensive. It requires skilled professionals, studio time, and meticulous attention to ensure accurate lip synchronization and cultural adaptation. As a result, the vast majority of content creators, especially those with limited resources, are often unable to invest in extensive dubbing efforts. Moreover, the linguistic landscape of countries like India is incredibly diverse, with numerous languages and dialects spoken across different regions. For content creators aiming to cater to this diverse audience, the challenge is compounded, requiring adaptations into multiple languages to truly resonate with viewers

1.2 Purpose of the Project

In this context, LingoFusion emerges as a solution poised to break down these language barriers. By automating the video dubbing process, LingoFusion empowers content creators to effortlessly adapt their videos into a variety of languages, including Hindi, Marathi, Gujarati, Telugu, and Bengali. This innovative application leverages advanced technologies, such as speech recognition, machine translation, and text-to-speech synthesis, to provide a cost-effective and efficient alternative to manual dubbing. By addressing the challenges posed by linguistic diversity and manual dubbing limitations, LingoFusion aims to democratize access to digital content. It seeks to create a more inclusive digital landscape, where individuals can engage with videos in their native languages, fostering cross-cultural understanding and enabling content creators to connect with a broader, more diverse audience. As LingoFusion addresses these language barriers head-on, it stands at the forefront of a transformative shift in the way we consume and share digital content globally. The purpose of the LingoFusion project is multi-faceted and driven by the recognition of several critical challenges in the current digital landscape. The overarching goal is to enhance the accessibility and inclusivity of digital content, particularly videos on platforms like YouTube, by overcoming language barriers.

2. Related work

Saad A. Baza et. al. [1] proposed an end-to-end architecture that automatically translates videos and produces synchronized dubbed voices using deep learning models, in a specified target language. Our architecture takes a modular approach, allowing the user to tweak each component or replace it with a better one. We present our results from said architecture, and describe possible future motivations to scale this to accommodate multiple languages and multiple use cases.

Chenxu Hu et. al. [2] proposed Neural Dubber, the first neural network model to solve a novel automatic video dubbing (AVD) task: synthesizing human speech synchronized with the given video from the text. Neural Dubber is a multi-modal text-to-speech (TTS) model that utilizes the lip movement in the video to control the prosody of the generated speech. Furthermore, an image-based speaker embedding (ISE) module is developed for the multi-speaker setting, which enables Neural Dubber to generate speech with a reasonable timbre according to the speaker's face. Experiments on the chemistry lecture single-speaker dataset and LRS2 multi-speaker dataset show that Neural Dubber can generate speech audios on par with state-of-the-art TTS models in terms of speech quality. Most importantly, both qualitative and quantitative evaluations show that Neural Dubber can control the prosody of synthesized speech by the video, and generate high-fidelity speech temporally synchronized with the video.

Mattia Di Gangi et. al. [3] proposed AppTek Dubbing, a product that will be available in Q3 2022 to automatically dub a video into a target language. We plan multiple releases of the product with incremental features, as well as the possibility to allow human intervention for increased quality.

Here are some of the most popular and well-rated NFT marketplaces where the user can buy and sell these digital assets.

Marcello Federico et. al. [4] proposed enhancements to a speech-to speech translation pipeline in order to perform automatic dubbing. Our architecture features neural machine translation generating output of preferred length, prosodic alignment of the translation with the original speech segments, neural text-to-speech with fine tuning of the duration of each utterance, and, finally, audio rendering to enriches text-to-speech output with background noise and reverberation extracted from the original audio.

3. Approach

LingoFusion is a pioneering automatic video dubbing application that empowers users to effortlessly transform English YouTube videos into Indian native languages such as Hindi, Marathi, Gujarati, Telugu, and Bengali. This innovative solution harnesses a powerful stack of technologies, including the NPM package for accessing YouTube audio, Deepgram for transcript generation, and Flutter packages for translation and text-to-speech capabilities.

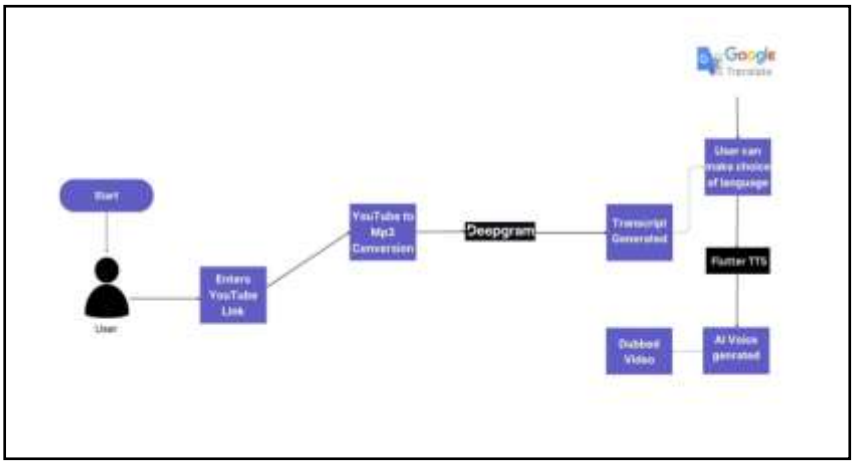

Fig 1.0 Application Block Diagram

Step 1 - The first step in our application is to collect YouTube link from the user . The user will paste the youtube link in the required text field and it will generate an Mp3 file of it. After pasting the YouTube link the request goes to backend server for generating mp3 using node.js packages.

Step 2 - For Transcript generation we are using Deepgram API which is a public API. Deepgram offers a potent blend of capabilities beyond basic STT, transforming audio into actionable insights.

Step 3 - The next step is to translate the text into native Indian languages such as Hindi, Marathi, Gujarati, Telugu, Bengali. For the translation we are using the flutter translator package which basically uses Google translate.

Step 4 - Before giving the translated text for speech generation it is important to clean the text as the transcript contains the timestamps which are not a part of speech but required during mapping process . So to clean the text we are using the concept of Regular Expression to bypass that part.

Step 5 - The next step is to get the Speech generation i.e converting the cleaned and translate text to AI voice. For this we are using the Flutter TTS which is basically the package made by Flutter community which in backend used Google Translate API and we don't require a paid account of Google Cloud for it.

4. Output

The project has successfully developed and deployed a user-friendly software platform that automates the translation and dubbing of videos from English into various regional languages. Translated and dubbed videos with voiceovers that accurately capture the cultural nuances and sensitivities of the selected regional languages, ensuring a culturally relevant and authentic multimedia experience. Streamlined processes for video localization, significantly reducing the time and resources required for content creators, educators, and businesses to reach a wider and more diverse audience.

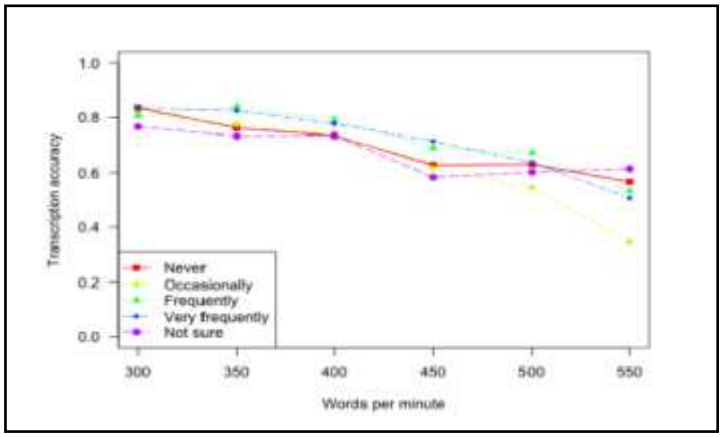

Fig 1.1 Transcription Accuracy Graph

In our exploration of language translation, Google Translate emerges as a pivotal player, revolutionizing online translation services. This tool, developed by Google, is distinguished by its widespread accessibility and dynamic utilization of statistical machine translation (SMT) and neural machine translation (NMT) approaches.

Introduction	**Trained Hours**
Hindi	163:25:47
Marathi	168:13:50
Gujarati	146:23:04
Telugu	50:51:36
Bengali	138:18:47

Table 1.0 Google Translate Indian language database

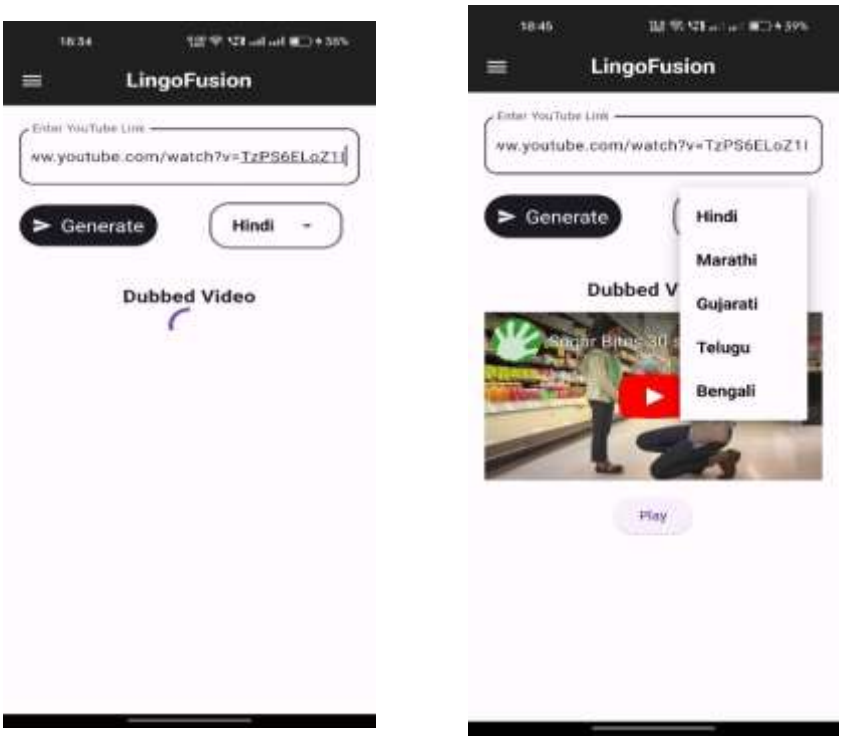

Fig 1.2 Final App Output

5. Discussion and Future Work

In the ever-expanding digital universe, where information knows no linguistic bounds, LingoFusion emerges as a groundbreaking solution poised to reshape the landscape of content accessibility. With a laser focus on transcending language barriers, LingoFusion is not merely an automatic video dubbing app; it is a transformative force with a profound mission. At its core, LingoFusion embodies the essence of inclusivity. In a world where a plethora of content is primarily available in English, the application stands as a beacon, illuminating a path toward a more linguistically diverse and inclusive digital space. It is more than just a tool for translation; it is a bridge that connects individuals from different linguistic backgrounds, fostering cross-cultural communication and understanding.

The future scope of LingoFusion extends beyond its current capabilities, presenting a dynamic roadmap for innovation and expansion. As technology evolves and user needs shift, the application is poised to adapt and introduce new features that further enhance its impact and relevance. Here, we explore the potential future directions and advancements that could shape the next phase of LingoFusion's journey. Enhanced Language Support: LingoFusion's future lies in expanding its language repertoire. While it currently focuses on translating English content into Indian languages, there is immense potential to broaden its scope to include a more extensive array of global languages. By incorporating additional language models and leveraging advancements in machine translation, LingoFusion could become a truly global platform, catering to diverse linguistic communities around the world. Customization and Personalization: Future iterations of LingoFusion may explore ways to enhance user customization. This could involve providing users with options to personalize the dubbing experience, such as selecting preferred accents, adjusting speech speed, or even choosing specific regional dialects. The ability to tailor the dubbing output to individual preferences would contribute to a more personalized and engaging user experience. Integration with AI and Natural Language Processing (NLP): Integrating AI and NLP technologies can elevate LingoFusion's language adaptation capabilities. By leveraging these advanced technologies, the app could refine its understanding of context, idiomatic expressions, and cultural nuances, resulting in more accurate and contextually relevant translations. This would contribute to a more natural and immersive viewing experience for users.

References

[1]. Saad A. Bazaz et. al. (2022) " Automated Dubbing and Facial Synchronization using Deep Learning". https://ieeexplore.ieee.org/abstract/document/9773697

[2]. Chenxu Hu et. al. (2021) "Neural Dubber: Dubbing for Videos According to Scripts" in 35th Conference on Neural Information Processing Systems (NeurIPS 2021). https://arxiv.org/abs/2110.08243

[3]. Mattia Di Gangi et. al. (2022) "Automatic Video Dubbing at AppTek" in Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.65/

[4]. Marcello Federico et. al. (2020) "From Speech-to-Speech Translation to Automatic Dubbing" in Proceedings of the 17th International Conference on Spoken Language Translation https://aclanthology.org/2020.iwslt-1.31/

[5] Christoph Bregler et. al. (2019) “Video Rewrite: Driving Visual Speech with Audio” at ACM SIGGRAPH. https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/human/bregler-sig97.pdf.

[6]. R. Tolosana et. al. (2020) “Deepfakes and beyond: A survey of face manipulation and fake detection”. https://www.researchgate.net/publication/338355353_DeepFakes_and_Beyond_A_Survey_of_Face_Manipulation_and_Fake_Detection.

[7]. Hideyuki Tachibana et. al. “Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention”. https://arxiv.org/abs/1710.08969

[8]. Tae-Hyun Oh et. al. (2019) “ Speech2face: Learning the face behind a voice,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/1905.09773

[9]. Konstantinos Vougioukas et. al. (2019) “Realistic Speech-Driven Facial Animation with GANs” in International journal of computer vision. https://arxiv.org/abs/1906.06337

[10]. Ohad fried et. al. (2019) “Text-based editing of talking-head video,” ACM Trans. Graph”. https://arxiv.org/abs/1906.01524