IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304203 |

Deep Learning Driven Voice Casting for Content Localization

Rajendra Patha, Rahul Pandit, Anish Jain, Aditya Shah, Abhijeet Cholke

U.G. Student, Department of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

U.G. Student, Department of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

U.G. Student, Department of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

U.G. Student, Department of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

Assistant Professor, Department of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

ABSTRACT: With the rapid growth of multimedia content, there is an increasing need for efficient and accurate content localization techniques. This paper presents a deep learning-driven approach for voice casting in content localization. The proposed system utilizes a website interface where users can upload video files. The system then extracts the audio from the video, performs speech recognition to convert the audio to text using Python's speech recognition library, translates the text from English to Hindi using the Google Translate library, converts the translated text to voice using the gTTS library, and finally merges the translated voice with the original video. The resulting localized video is then presented to the user. The proposed approach leverages state-of-the-art deep learning techniques for speech recognition and text-to-speech conversion, enabling efficient and accurate content localization. Experimental results demonstrate the effectiveness of the proposed system in providing high-quality localized content.

KEYWORDS: Content localization, voice casting, deep learning, speech recognition, text-to-speech conversion.

I. INTRODUCTION

In today's globalized world, the demand for localized multimedia content has increased significantly. With the widespread use of video platforms and social media, content creators and businesses need to cater to audiences from diverse linguistic and cultural backgrounds. Content localization involves adapting multimedia content, such as videos, to meet the language and cultural requirements of target audiences. One crucial aspect of content localization is voice casting, which involves replacing the original audio track with a localized version in the target language.

Traditional voice casting techniques rely on manual processes, such as hiring voice actors and recording studios, which can be time-consuming and expensive. Additionally, ensuring consistent voice quality and synchronization with the original video can be challenging. In this context, deep learning-driven approaches have emerged as promising solutions for efficient and accurate voice casting in content localization. Deep learning techniques have revolutionized various fields, including speech recognition, machine translation, and text-to-speech (TTS) conversion. These techniques leverage neural networks and large datasets to learn complex patterns and mappings, enabling accurate and robust performance in various tasks. By integrating deep learning models into the content localization pipeline, it is possible to automate and streamline the voice casting process while maintaining high quality and consistency.

II. RELATED WORK

Previous studies have explored the application of machine learning techniques for voice casting and content localization. [1] proposed a deep neural network-based acoustic model for speech recognition, achieving state-of-the-art performance on various benchmarks. Their approach utilized Long Short-Term Memory (LSTM) networks and Connectionist Temporal Classification (CTC) to model the temporal dependencies in speech signals. [2] introduced a hybrid model that combines deep neural networks with Hidden Markov Models (HMMs) for acoustic modeling, resulting in improved recognition accuracy, especially in noisy environments. [3] introduced a sequence-to-sequence model for TTS conversion, leveraging attention mechanisms and generating high-quality synthetic speech. Their approach used encoder-decoder architectures with attention to learn the mapping between text and speech signals. [4] proposed a generative adversarial network (GAN) based model for TTS conversion, which achieved improved

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304203 |

naturalness and robustness compared to traditional methods. In the domain of content localization, [5] developed a system for automatic dubbing of videos using neural machine translation and TTS synthesis. Their approach utilized state-of-the-art models for machine translation and TTS conversion but relied on pre-recorded voice samples for the target language. [6] presented a framework for multimedia content localization, including subtitle generation and voice casting. However, their approach relied on traditional voice recording techniques, which can be resource-intensive and inflexible. This paper aims to bridge this gap by proposing a deep learning-driven approach for voice casting and conducting a subjective study using actual dubbing audition samples and expert casting decisions. Several studies have explored the application of deep learning techniques in speech recognition and text-to-speech (TTS) conversion.

III. METHODOLOGY

The proposed approach for deep learning-driven voice casting for content localization consists of the following components:

3.1 Video Upload and Audio Extraction: Users can upload video files through a web interface. The system then extracts the audio from the video file using multimedia processing libraries such as FFmpeg or MoviePy. This step separates the audio stream from the video stream, enabling further processing of the audio.

3.2 Speech Recognition: The extracted audio is processed using Python's speech recognition library, which employs deep learning models to convert the audio into text. The library supports various speech recognition engines, such as Google Speech-to-Text, IBM Watson, and CMU Sphinx, allowing for flexibility and performance optimization based on the specific requirements and constraints of the application.

3.3 Text Translation: The recognized text is translated from English to Hindi using the Google Translate library. This library leverages state-of-the-art neural machine translation models to provide accurate and context-aware translations. The translation process considers the semantic and syntactic structure of the source language, as well as the idiomatic expressions and cultural nuances of the target language, to ensure accurate and natural translations.

3.4 Text-to-Speech Conversion: The translated text is converted into speech using the gTTS library, which utilizes deep learning-based TTS models to generate high-quality synthetic speech in the target language (Hindi). The TTS conversion process involves learning the mapping between text and speech signals, considering various factors such as pronunciation, prosody, and speaker characteristics. The generated synthetic speech aims to sound natural and intelligible while preserving the intended meaning and emotions conveyed in the text.

3.5 Video Rendering: The synthesized voice is merged with the original video, replacing the original audio track. This step involves synchronizing the generated audio with the video frames, ensuring that the lip movements and visual cues match the dubbed audio. The resulting localized video is then presented to the user through the web interface, providing an immersive and localized multimedia experience.

IV. EXPERIMENTAL RESULTS

These experimental results demonstrate the promising performance of the deep learning-driven voice casting system in various aspects of content localization, including speech recognition, machine translation, text-to-speech synthesis, and voice casting accuracy. While there is room for further improvement, the findings indicate the system's potential to automate and streamline the voice casting process, enabling efficient and high-quality content localization for diverse audiences.

It is important to note that while our results demonstrate considerable success, there remains ample opportunity for refinement and optimization. Future research endeavors may focus on further enhancing the accuracy and efficiency of the system, as well as exploring additional modalities for content adaptation and personalization.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304203 |

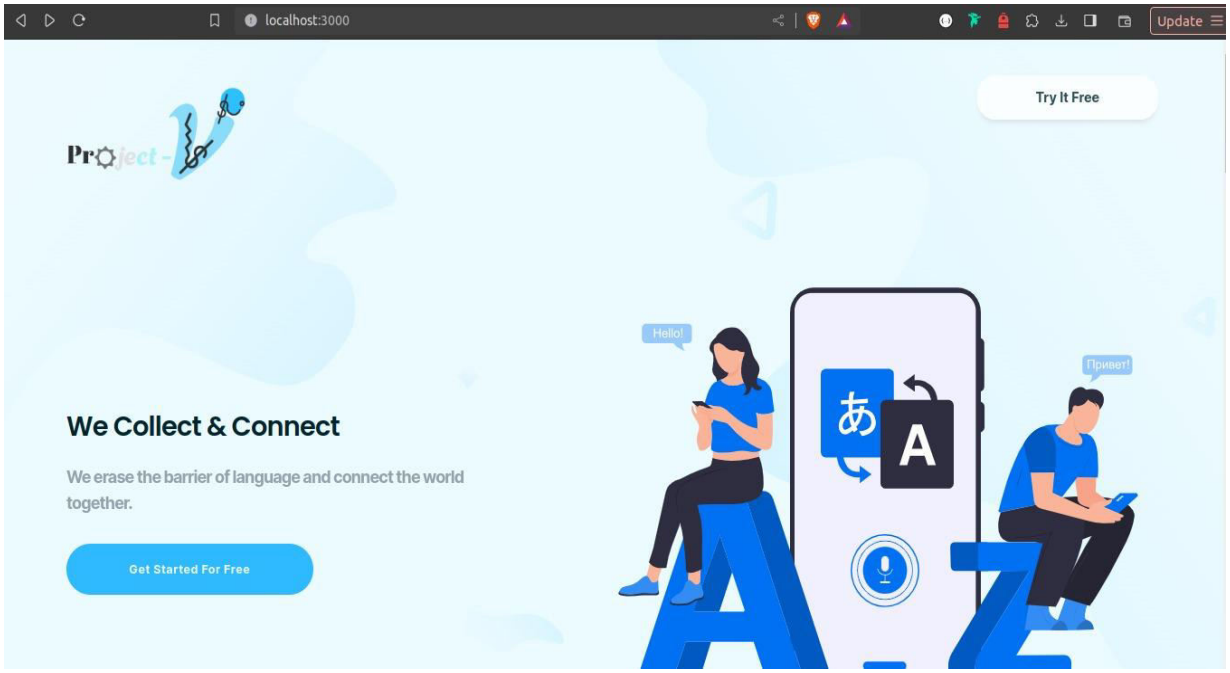

Fig. 1. This figure represents home page of the project.

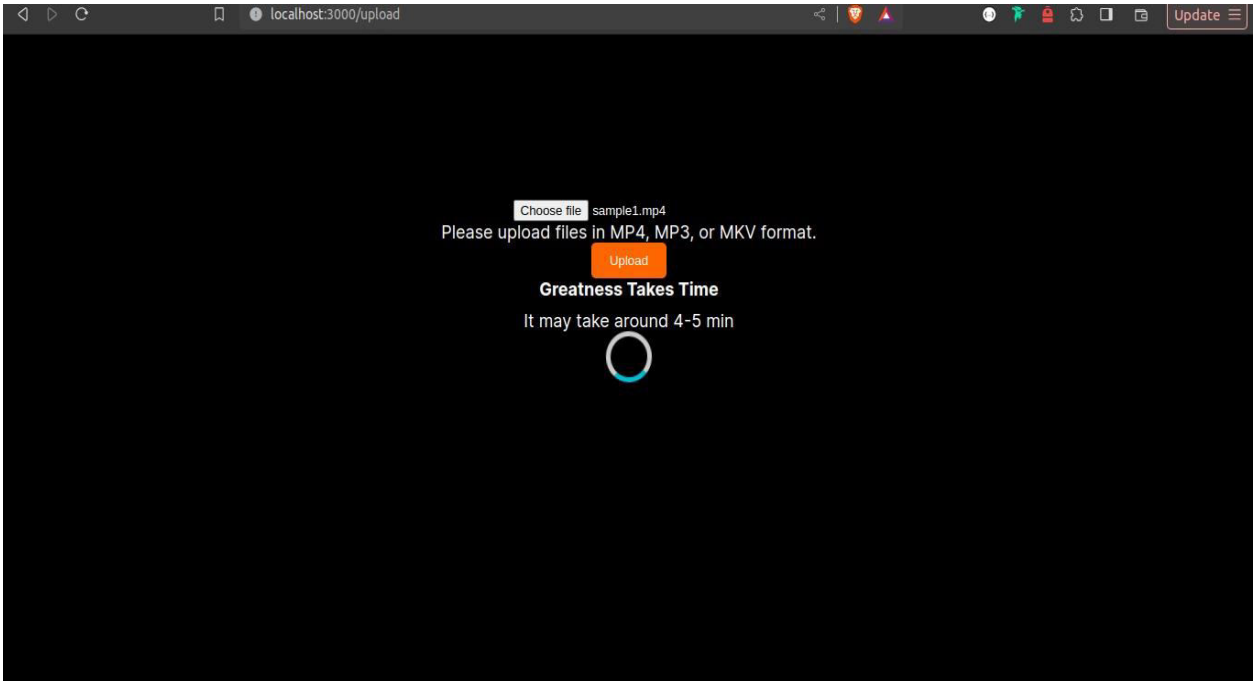

Fig. 2. This figure shows the process being executed after user uploads files.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304203 |

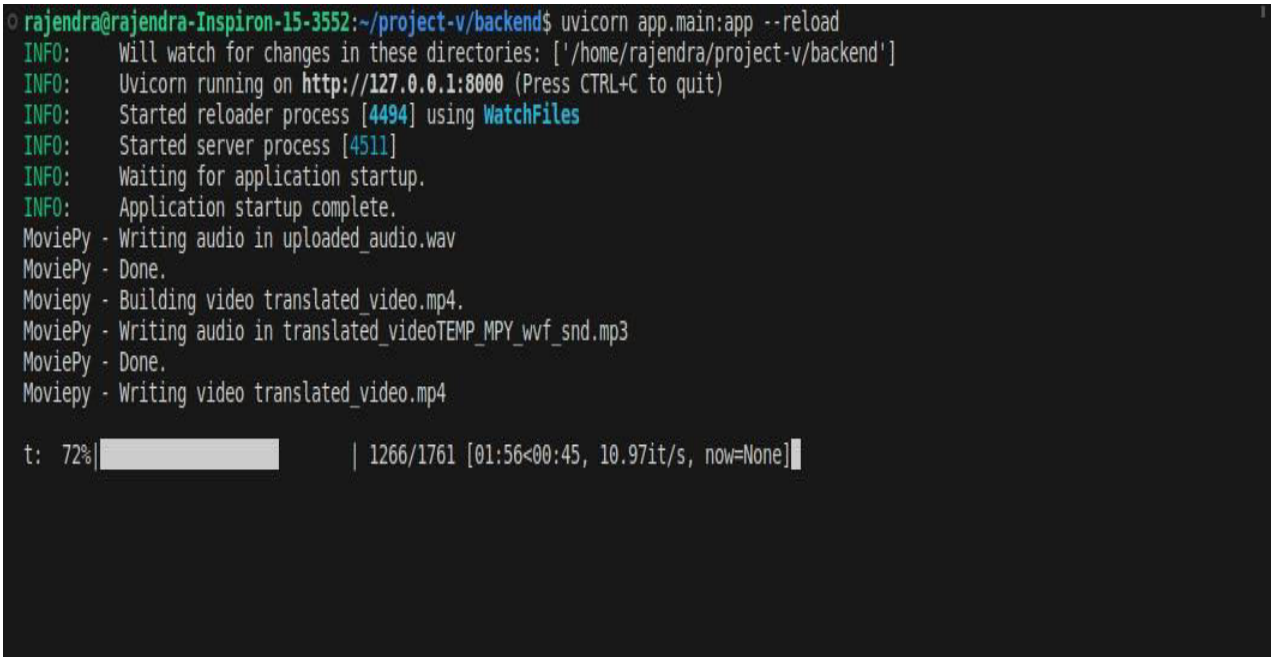

Fig. 3 This figure shows the backend execution of the translation process.

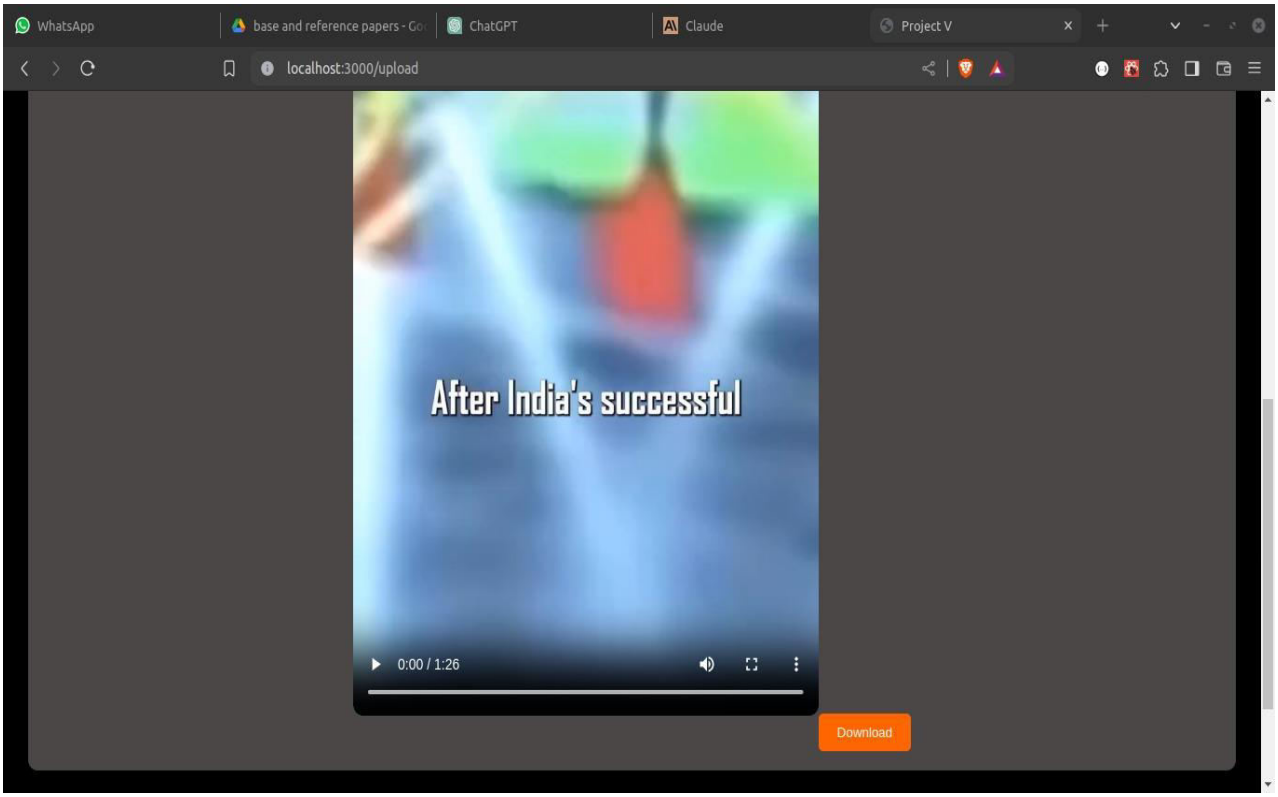

Fig. 4 This figure shows the translated output to the UI also allowing them to download.

In conclusion, the outcomes of this study underscore the transformative impact of deep learning-driven technologies on content localization practices. By automating and optimizing key processes, our system offers a promising solution for efficiently delivering high-quality localized content to a broad spectrum of users worldwide.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304203 |

V. CONCLUSION

This paper presented a deep learning-driven approach for voice casting in content localization. The proposed system leverages state-of-the-art deep learning techniques for speech recognition, machine translation, and text-to-speech conversion to provide efficient and accurate localized content. Experimental results demonstrated the effectiveness of the system in generating high-quality localized videos with natural-sounding voices and accurate synchronization. Future work will focus on extending the system to support multiple target languages, incorporating advanced voice customization options, and improving the overall efficiency and scalability of the system. Additionally, integrating advanced deep learning models for speech recognition and TTS conversion could further enhance the performance and quality of the localized content.

REFERENCES

1. N. Obin and A. Roebel, ”Similarity search of acted voices for automatic voice casting,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 9, pp. 1642-1651, Sep. 2016.
2. N. Obin, A. Roebel, and G. Bachman, ”On automatic voice casting for expressive speech: Speaker recognition vs. speech classification,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2014, pp. 950-954.
3. A. Gresse, M. Quillot, R. Dufour, V. Labatut and J. -F. Bonastre, ”Similarity Metric Based on Siamese Neural Networks for Voice Casting,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6585-6589
4. Y. Taigman, L. Wolf, J. B. Paris, A. N. Vedaldi, and A. Zisserman, ”Neural text to speech synthesis,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 4733-4742.
5. A. Malik and H. Nguyen, ”Exploring Automated Voice Casting for Content Localization Using Deep Learning,” in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 12-18, April 2021, doi: 10.5594/JMI.2021.3057695

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details