INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 1654-1661

SURVEY ON SOPHISTICATED VIDEO DUBBING SOFTWARE

Dr. CH. Vijaya Kumar[1], Muthireddy Rishita[2], Gongati Bhavana[3], Gunda Gouthami[4], Kodali Venkata Vishnu Vamshi[5]

[1]Associate Professor, CSE Dept, ACE Engineering College Hyderabad, India

[2,3,4,5] Student, CSE Dept, ACE Engineering College Hyderabad, India

ABSTRACT

The primary objective is to bridge linguistic gaps and promote inclusivity by making video content accessible and comprehensible to a wider audience. Key features of the software include user-friendly interfaces, efficient video processing capabilities, and support for multiple Indian languages. The development process involves the integration of cutting-edge technologies such as neural machine translation and robust language models. The anticipated impact of this project is significant, contributing to the democratization of information and fostering cross-cultural understanding. By enabling the translation of videos into languages associated with different religions, the software aspires to promote cultural harmony and facilitate the sharing of diverse perspectives. The primary objectives include implementing advanced machine translation techniques, ensuring precise translations, supporting multiple video formats, and promoting cross-cultural communication. The software emphasizes linguistic diversity, cultural understanding, and the democratization of information through cutting-edge technologies and user-friendly interfaces.

Keywords: Linguistic diversity, tandem, Multilingual

1. INTRODUCTION

The exponential growth of digital content consumption has accentuated the demand for diverse and regionally relevant video content. However, a substantial gap exists in the availability of high-quality, localized multimedia content for Indian audiences who speak various regional languages. This gap is particularly evident in the lack of efficient and accessible tools for dubbing *videos* from the dominant English language into multiple Indian languages. In this context, addressing the need for sophisticated dubbing technologies and fostering collaboration between content creators, technology developers, and policymakers becomes imperative to bridge the linguistic gap and create a more inclusive digital content ecosystem.

The project at hand is a pioneering initiative aimed at the development of a Multilingual Video Dubbing Software tailored specifically for the rich linguistic landscape of Indian regional languages, utilizing state-of-the-art Natural Language Processing (NLP) techniques. India, with its diverse cultural tapestry, is home to a plethora of languages and dialects, each with its unique nuances and linguistic intricacies. The objective of this project is to bridge the linguistic divide by creating a sophisticated solution that enables seamless dubbing of videos into various Indian regional languages, thereby fostering inclusivity and accessibility in the digital content space.

The cornerstone of the software lies in the judicious selection and integration of NLP models. Leveraging advanced speech recognition models, such as Google's Speech-to-Text API, ensures accurate conversion of spoken words into text. Complementary to this, sophisticated text-to-speech synthesis models, like WaveNet, are harnessed for generating natural-sounding speech. These models are fine-tuned to account for the intricacies of Indian languages, with an emphasis on maintaining linguistic authenticity and cultural relevance.

The development process encompasses several key components, including language processing modules, translation services, voice synthesis systems, and intricate lip-syncing algorithms. These elements work in tandem to facilitate the seamless transformation of the original video content into a multilingual format, wherein the spoken words align harmoniously with the on-screen lip movements. The software is designed with scalability in mind, ensuring adaptability for future updates, additions of new languages, and optimization for real-time dubbing performance.

In essence, this Multilingual Video Dubbing Software for Indian Regional Languages is not just a technological venture; it is a cultural bridge, connecting people through the universal medium of audio-visual content in their native languages. As the digital landscape continues to evolve, this project stands as a testament to the commitment to inclusivity, technological innovation, and the celebration of linguistic diversity in the vibrant mosaic that is India. The advent of globalization and digital connectivity has led to an unprecedented influx of content consumption across diverse linguistic and cultural landscapes. However, this surge in digital content is not always accessible or relatable to individuals whose primary language and cultural context differ from the content's origin. In this context, the project emerges as a response to the growing need for a sophisticated software solution that seamlessly translates and buds videos from English to various Indian languages, with a specific emphasis on accommodating religious nuances.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 1654-1661

Understanding the multifaceted nature of linguistic diversity in India, the software goes beyond mere translation. It employs advanced budding algorithms to ensure a smooth and contextually appropriate transition between segments of the video content. Moreover, the software recognizes the importance of religious sensitivity in language translation, particularly when dealing with content related to diverse religions. As such, it incorporates mechanisms to handle religious terminologies with respect and accuracy, enhancing the overall quality of the translated content. One of the distinguishing features of the software is its user-centric approach. It provides users with the flexibility to customize the budding process based on their preferences, allowing for a personalized and culturally sensitive viewing experience. By placing control in the hands of the users, the software aims to cater to individual preferences and ensure that the translated content aligns with their cultural and religious backgrounds. In essence, this project is not merely a technological solution but a cultural bridge that facilitates a deeper understanding and appreciation of content across linguistic and religious boundaries. As we delve into the project's development and functionalities, it becomes evident that it addresses a significant gap in the digital content landscape, contributing to cultural inclusivity and technological innovation.

2. LITERATURE SURVEY

Neural Machine Translation for Low-resource Languages by Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, Rishemjit Kaur: - March 2023 [1]

Neural Machine Translation (NMT) has evolved significantly over the past decade, gaining widespread adoption since the early 2000s and reaching a mature phase. Despite its dominance in machine translation, its efficacy diminishes when applied to low-resource language pairs, primarily due to the scarcity of extensive parallel corpora. The spotlight is now on implementing NMT techniques tailored for low-resource languages, sparking notable research in this domain. The aim is to furnish guidelines for selecting suitable NMT techniques in diverse LRL data settings based on empirical findings.

Speech Driven Video Editing via an Audio-Conditioned Diffusion Model by Dan Bigioi, Shubhajit Basak, Michał Stypułkowski, Maciej Zieba, Hugh Jordan, Rachel McDonnell, Peter Corcoran - May 2023 [2]

A novel neural pipeline allowing one to generate pose aware 3D animated facial landmarks synchronised to a target speech signal is proposed for the task of automatic dubbing. The goal is to automatically synchronize a target actors' lips and facial motion to an unseen speech sequence, while maintaining the quality of the original performance. Given a 3D facial key point sequence extracted from any reference video, and a target audio clip, the neural pipeline learns how to generate head pose aware, identity aware landmarks and outputs accurate 3D lip motion directly at the inference stage.

Dubbing in Practice: A Large scale Study of HumanLocalization With Insights for Automatic Dubbing by William Brannon, Virkar Thompson- May 2023 [3]

The results challenge a number of assumptions commonly made in both qualitative literature on human dubbing and machine-learning literature on automatic dubbing, arguing for the importance of vocal naturalness and translation quality over commonly emphasized isometric (character length) and lip-sync constraints, and for a more qualified view of the importance of isochronic (timing) constraints.

Multilingual video dubbing—a technology review and current challenges by Dan Bigioi, Corcoran – Sep 2023 [4]

The proliferation of multi-lingual content on today's streaming services has created a need for automated multi-lingual dubbing tools. In this article, current state-of-the-art approaches are discussed with reference to recent works in automatic dubbing and the closely related field of talking head generation. A taxonomy of papers within both fields is presented, and the main challenges of both speech-driven automatic dubbing, and talking head generation are discussed and outlined, together with proposals for future research to tackle these issues.

Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization by Soumik Mukhopadhyay1, Saksham Suri1, Ravi Teja Gadde2, Abhinav Shrivastava1 – Aug 2023 [5]

The task of lip synchronization (lip-sync) seeks to match the lips of human faces with different audio. It has various applications in the film industry as well as for creating virtual avatars and for video conferencing. This is a challenging problem as one needs to simultaneously introduce detailed, realistic lip movements while preserving the identity, emotions, and image quality.

Audio Driven Talking Head video Generation By Yizhe Zhua; Chunhui Zhanga; Qiong Liub; Xi Zhoubn – June 2023 [6]

Synthesizing high-fidelity talking head videos by fitting input audio sequences is a highly anticipated technique in many applications, such as digital humans, virtual video conferences, and human-computer interaction. Popular GAN-based

methods aim to align speech audio with lip motions and head poses. However, existing methods are prone to training instability and even mode collapse, resulting in low-quality video generation. In this paper, we propose a novel audio-driven diffusion method for generating high-resolution realistic videos of talking heads with the help of the denoising diffusion model.

Using Video Dubbing to Foster EFL College Students' English News Broadcasting Ability by Meng-lian Liu – May 2023 [7]

Data included deep speaking pre-and post-test scores, initial and final dubbing videos, learning logs, and reflective essays. The results showed that after engaging in video dubbing the students improved their English-speaking and English news broadcasting ability, in terms of accuracy and fluency. Coaching and modeling were ranked as the two most useful processes which supported students as they improved their English-news broadcasting ability through repeatedly listening, echoing, and imitating.

Technology Pipeline for Large Scale Cross-Lingual Dubbing of Lecture Videos into Multiple Indian Languages by Ishika Gupta, Arun Kumar, Ashish Seth, Bhagyashree Mukherjee- Dec 2023 [8]

Cross-lingual dubbing of lecture videos requires the transcription of the original audio, correction and removal of disfluencies, domain term discovery, text-to-text translation into the target language, chunking of text using target language rhythm, text-to-speech synthesis followed by isochronous lip-syncing to the original video. This task becomes challenging when the source and target languages belong to different language families, resulting in differences in generated audio duration.

Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers by Yasheng Sun, Hang Zhou, Qianyi Wu, Zhibin Hong, Jingtuo Liu – Dec 2022 [9]

Previous studies have explored generating accurately lip-synce talking faces for arbitrary targets given audio conditions. However, most of them deform or generate the whole facial area, leading to non-realistic results. In this work, we delve into the formulation of altering only the mouth shapes of the target person. This requires masking a large percentage of the original image and seamlessly inpainting it with the aid of audio and reference frames. To this end, we propose the Audio-Visual Context-Aware Transformer (AV-CAT) framework, which produces accurate lip-sync with photo-realistic quality by predicting the masked mouth shapes.

Using video dubbing to foster college students' English-speaking ability by Cheng-Yueh Jao, Hui-Chin Yeh, Nian-Shing Chen, Wan-Rou Huang - June 2022 [10]

Data included GEPT speaking pre-and post-test scores, initial and final dubbing videos, learning logs, and reflective essays. The results showed that after engaging in video dubbing the students improved their English-speaking ability, in terms of accuracy and fluency. Coaching and modeling were ranked as the two most useful processes which supported students as they improved their English-speaking ability through repeatedly listening, echoing, and imitating. Implications and limitations of the study are discussed.

3. COMPARISON ANALYSIS

S.NO	PAPER TITLE	WORK DONE	PERFORMANCE ANALYSIS	FUTURE WORK	DRAWBACKS
1	Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, Rishemjit Kaur, "Neural Machine Translation for Low-resource Languages" - 2023	This survey shed light on the evolving landscape of NMT, particularly in the context of low-resource languages. By synthesizing current research trends, the article contributes to the identification of effective NMT	The recommendations presented aim to inspire continued research efforts, fostering advancements in addressing the unique challenges posed by low-resource language pairs in the realm of Neural Machine Translation.	Refining recommendat ions, advanced video dubbing, easy real time, further language support, user education.	Lack of standardization, Time consuming.

		techniques for diverse LRL data.			
2	Dan Bigioi, Shubhajit Basak, Michał Stypułkowski, Maciej Zieba, Hugh Jordan, Rachel McDonnell, Peter Corcoran, "Speech Driven Video Editing via an Audio-Conditioned Diffusion Model" - 2023	A novel neural pipeline allowing one to generate pose aware 3D animated facial landmarks synchronised to a target speech signal is proposed for the task of automatic dubbing	The goal is to automatically synchronize a target actors' lips and facial motion to an unseen speech sequence, while maintaining the quality of the original performance.	Model Speed and In The Wild Training: It is no secret that diffusion models are slow, both to train and to sample from. Our models are no exception, taking us approximately 6 minutes/epoch to train the single-speaker model, and 40 minutes/epoch	Lack of performance evaluation, privacy concerns, comparison with other software.
3	William Brannon, Virkar Thompson, "Dubbing in Practice: A Large Scale Study of HumanLocalization With Insights for Automatic Dubbing" - 2023	The results challenge a number of assumptions commonly made in both qualitative literature on human dubbing and machine-learning literature on automatic dubbing.	The importance of vocal naturalness and translation quality over commonly emphasized isometric (character length) and lip-sync constraints, and for a more qualified view of the importance of isochronic (timing) constraints.	Appearance Consistency: As previously discussed, our multi-speaker model's generated output appearance for unseen identities occasionally deviates from the original. To investigate this phenomenon, we intend to delve deeper into the underlying causes.	Time-consuming, Lacking user support.
4	Dan Bigioi, Corcoran, "Multilingual video dubbing—a technology review and current challenges" - 2023	This work may serve both as an introduction and reference guide for researchers new to the fields of automatic dubbing, and talking head generation, but also seeks to draw attention to the latest techniques and new approaches and methodologies for	Evaluating the precision of language translation, ensuring that the meaning, tone, and cultural nuances are accurately conveyed. This involves comparing the original and dubbed scripts and possibly getting feedback from native speakers.	Although significant progress has been made in the fields of talking head generation and automatic dubbing, these areas of research are constantly evolving, and several open challenges still need to be addressed, offering plenty of opportunities for future work.	One of the primary drawbacks of video dubbing software is the potential loss of audio and video quality.

		those who already have some familiarity with the field.			
5	Soumik Mukhopadhyay1, Saksham Suri1, Ravi Teja Gadde2, Abhinav Shrivastava1, "Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization" - 2023	Lip-sync methods can be roughly classified into the following four categories. Please note that there may be over-laps between these categories. Embedding-based head reconstruction.	Assessing the naturalness and clarity of synthesized voices. This can involve subjective evaluations by listeners and objective measures such as intelligibility scores. The goal is to ensure voices sound as human-like as possible, with appropriate emotional tones matching the original audio.	Analyzing the time it takes to dub a video, from processing the original audio and video to synthesizing the dubbed speech and integrating it into the video. Faster processing times improve efficiency, especially for large-scale or real-time dubbing projects.	Depending on the software and the quality of the original recording, the dubbed version may suffer from synchronization issues, loss of lip-sync accuracy, or distortion of audio quality.
6	Yizhe Zhua; Chunhui Zhanga; Qiong Liub; Xi Zhoubn, "Audio Driven Talking Head video Generation" - 2023	Popular GAN-based methods aim to align speech audio with lip motions and head poses. However, existing methods are prone to training instability and even mode collapse, resulting in low-quality video generation.	Assessing how well the software manages computational resources. This involves measuring CPU and GPU usage, memory consumption, and any other relevant resources during the dubbing process. Optimizations in this area can make the software more accessible to users with varying hardware capabilities.	Enhancing lip-sync technology to ensure that the dubbed audio matches the lip movements of the actors in the video perfectly. This involves advancements in computer vision and machine learning to analyze and predict lip movements more accurately.	Dubbing videos can be a time-consuming process, especially for longer videos or projects with numerous scenes.
7	Meng-lian Liu, "Using Video Dubbing to Foster EFL College Students' English News Broadcasting Ability" - 2023	The results showed that after engaging in video dubbing the students improved their English-speaking and English news broadcasting ability, in terms of accuracy and fluency.	Measuring how well the dubbed speech matches the lip movements of the original video. This can be done through computer vision techniques and human evaluation to ensure the synchronization feels natural to viewers.	Leveraging more advanced deep learning models to better understand context, tone, and nuances in speech. This could lead to more accurate and natural-sounding dubbing that better matches the original performance in terms of emotion	The process involves recording new audio tracks, editing, synchronizing, and fine-tuning to ensure the dubbed version matches the original content seamlessly.

				and intention.	
8	Ishika Gupta, Arun Kumar, Ashish Seth, Bhagyashree Mukherjee, "Technology Pipeline for Large Scale Cross-Lingual Dubbing of Lecture Videos into Multiple Indian Languages" - 2023	Cross-lingual dubbing of lecture videos requires the transcription of the original audio, correction and removal of disfluencies, domain term discovery, text-to-text translation into the target language, chunking of text using target language rhythm, text-to-speech synthesis followed by isochronous lip-syncing to the original video.	Analyzing the time it takes to dub a video, from processing the original audio and video to synthesizing the dubbed speech and integrating it into the video. Faster processing times improve efficiency, especially for large-scale or real-time dubbing projects.	Working on reducing the processing time for video dubbing to enable real-time or near-real-time dubbing capabilities. This could be particularly useful for live broadcasts or streaming content in multiple languages.	While many video dubbing software support multiple languages, the availability of languages may be limited compared to the vast array of languages spoken worldwide.
9	Yasheng Sun, Hang Zhou, Qianyi Wu, Zhibin Hong, Jingtuo Liu, "Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers" - 2022	Explored generating accurately lip-synce talking faces for arbitrary targets given audio conditions. However, most of them deform or generate the whole facial area, leading to non-realistic results.	Assessing how well the software manages computational resources. This involves measuring CPU and GPU usage, memory consumption, and any other relevant resources during the dubbing process. Optimizations in this area can make the software more accessible to users with varying hardware capabilities.	Expanding the range of languages and dialects the software can handle, making video content accessible to a broader audience worldwide. This includes improving translation quality and cultural relevance in dubbing.	Adapting content for different cultural contexts during the dubbing process can be complex.
10	Cheng-Yueh Jao, Hui-Chin Yeh, Nian-Shing Chen, Wan-Rou Huang, "Using video dubbing to foster college students' English-speaking ability" - 2022	Results showed that after engaging in video dubbing the students improved their English-speaking ability, in terms of accuracy and fluency. Coaching	Assessing the software's compatibility with different video and audio formats, as well as its ability to integrate with other production tools. This ensures that users can easily incorporate the	Addressing ethical and legal considerations related to voice replication and copyright, ensuring that the use of an actor's likeness and voice in dubbed content respects	Consistency in voice quality, tone, and style across multiple dubbed versions of the same content can be difficult to achieve, especially when

		and modeling were ranked as the two most useful processes which supported students as they improved their English-speaking ability through repeatedly listening, echoing, and imitating.	dubbing software into their existing workflows.	their rights and complies with regulations.	using different voice actors or dubbing studios.

4. FUTURE SCOPE

The envisioned future enhancements for the video translation software geared towards translating from English to various Indian languages encompass a multifaceted approach aimed at elevating its functionality and user experience. The foremost objective revolves around expanding language support, broadening the software's reach to encompass additional Indian languages, thereby enhancing accessibility for a diverse user base. Simultaneously, efforts to bolster translation accuracy stand pivotal, promising users more precise and contextually relevant translations, thereby ensuring a superior viewing experience. Real-time translation capabilities are poised to revolutionize the software, ushering in a dynamic dimension to translation processes, facilitating instantaneous conversion of content to diverse languages. Integral to this vision is the integration of user feedback mechanisms, fostering an iterative process of refinement and improvement driven by user insights. Moreover, customization options are slated for implementation, empowering users to tailor translation settings according to their preferences and requirements, thereby enhancing flexibility and user satisfaction. Furthermore, leveraging machine learning capabilities, the software aspires to adapt to user preferences over time, refining its translation algorithms and recommendations based on user interactions and feedback. This iterative approach promises continuous improvement and refinement, ensuring that the software remains at the forefront of video translation technology, delivering exceptional value and utility to its users.

5. CONCLUSION

In conclusion, the development of the video translation software marked a significant stride towards bridging linguistic gaps and fostering cultural inclusivity. The project successfully showcased the feasibility of translating videos from English to various Indian religious languages, contributing to enhanced accessibility and understanding across diverse communities. The incorporation of advanced language processing techniques, such as audio transcription, language detection, and translation, underscored the project's commitment to delivering accurate and culturally sensitive results.

While the software demonstrated notable achievements, there is always room for improvement. Future iterations could focus on refining translation algorithms, expanding language support, and addressing occasional discrepancies to ensure a more seamless user experience. The positive user feedback and increased engagement with translated content emphasize the software's potential impact in promoting global communication and mutual understanding. As technology continues to evolve, this project lays a foundation for future innovations in the realm of multilingual video translation.

6. REFERENCES

[1] Dan Bigioi, Shubhajit Basak, Michał Stypułkowski, Maciej Zieba, Hugh Jordan, Rachel McDonnell, Peter Corcoran () Speech Driven Video Editing via an Audio-Conditioned DiffusionModel (https://doi.org/10.48550/arXiv.2301.04474)

[2] Yogesh Virkar, Marcello Federico, Robert Enyedi, Roberto Barra-Chicote (April 2022) Prosodic Alignment for off-screen Automatic Dubbing (https://doi.org/10.48550/arXiv.2204.02530)

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen (December 2015) End-to-end speech recognition in English (https://doi.org/10.48550/arXiv.1512.02595)

[4] Jindrich Matousek, Zdenek Hanzlicek, Daniel Tihelka and Martin Mener (December 2010) Automatic Dubbing of TV Programmes for the Hearing Impaired (DOI: 10.1109/ICOSP.2010.5655861)

[5] Zdenek Hanzlicek, Jindrich Matousek, Daniel Tihelka (December 2008) Towards Automatic Audio Track Generation for Czech TV Broadcasting: Initial Experiments with Subtitles-to-Speech Synthesis (DOI:

10.1109/ICOSP.2008.4697710)

[6] E. Moulines and F. Charpentier (August 1990) Pitch-Synchronous waveform processing techniques for text-to-speech synthesis using Diphones (DOI: 10.21437/Eurospeech.1989-172)

[7] Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, Marcello Federico (April 2022) Duration modeling of neural TTS for Automatic Dubbing (DOI: 10.1109/ICASSP43922.2022.9747158)

[8] Srikar Kashyap Pulipaka, Chaitanya Krishna Kasaraneni, Venkata Naga Sandeep Vemulapalli, Surya Sai Mourya Kosaraju (December 2019) Machine Translation of English Videos to Indian Regional Languages using Open Innovation (DOI:10.1109/ISTAS48451.2019.8937988)

[9] Bigioi, D., Basak, S., Jordan, H., McDonnell, R., and Corcoran, P. (2023). Speech driven video editing via an audio-conditioned diffusion model. https://arxiv.org/abs/2301.04474. doi:10.1109/ACCESS.2022.3231137

[10] Bigioi, D., Jordan, H., Jain, R., McDonnell, R., and Corcoran, P. (2022). Pose-aware speech driven facial landmark animation pipeline for automated dubbing. IEEE Access 10, 133357–133369.

[11] Du, C., Chen, Q., He, T., Tan, X., Chen, X., Yu, K., et al. (2023). Dae-talker: high fidelity speech-driven talking face generation with diffusion autoencoder. https://arxiv.org/abs/2303.17550.

[12] Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., et al. (2023). SeamlessM4t—massively multilingual and multimodal machine translation. Menlo Park, California, United States: Meta.

[13] Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., and Lu, Y. (2023). High-fidelity and freely controllable talking head video generation. https://arxiv.org/abs/2304.10168.