https://doi.org/10.33472/AFJBS.6.Si2.2024.2811-2817

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Transformative Perspectives: AI-Enabled Dubbing Software for Multilingual Content Localization

[1]Nami Susan Kurian, [2]Kalaivani S, [3]Sinchana K S, [4]Subha T D, [5]Sornisha N L, [6]Sai Charan G, [7]Kundana Sree P

[1]Assistant Professor, Department of Electronics and Communication Engineering, Jyothy Institute of Technology, Bengaluru, India

[2] Assistant Professor, Department of Electronics and Communication Engineering, Rajalakshmi Institute of Technology, Chennai, India

[3]UG Scholar, Department of Electronics and Communication Engineering, Jyothy Institute of Technology, Bengaluru, India

[4]Assistant Professor, Department of Electronics and Communication Engineering, RMK Engineering College, Kavaraipettai, India

[5,6,7] UG Scholar, Department of Electronics and Communication Engineering, RMK Engineering College, Kavaraipettai, India

[1]namisusan7@gmail.com, [2]kalaivani.s@ritchennai.edu.in, [3]sinchana2907@gmail.com, [4]tdsubha2010@gmail.com,[5]nlso22091.ec@rmkec.ac.in, [6]gali21134.ec@rmkec.ac.in, [7]pula22115.ec@rmkec.ac.in

Volume 6, Issue Si2, 2024

Received: 09 March 2024

Accepted: 10 April 2024

Published: 20 May 2024

doi:10.33472/AFJBS.6.Si2.2024.2811-2817

Abstract—Dubbing serves as a critical element in rendering content accessible to global audiences, facilitating intercultural adaptation, multilingual marketing endeavors, and upholding the integrity of original intentions. Its essence lies in the substitution of the original soundtrack with translated or modified versions, thereby ensuring comprehension and resonance across diverse linguistic landscapes. The dubbing process encompasses several fundamental steps and considerations. Foremost among these is linguistic precision, which is indispensable for effectively conveying the intended message. Translators are tasked with capturing nuances, cultural references, and colloquialisms while synchronizing dialogue with lip movements to maintain authenticity. Furthermore, voice casting assumes a crucial role in achieving fidelity to the original work, necessitating the involvement of skilled individuals capable of embodying characters across cultural boundaries. From this perspective, video/audio dubbing emerges as an indispensable phase in the broader process of in-formation dissemination, substituting translations for the original track to enable global accessibility. This paper underscores the imperative of preserving authenticity by delineating the foundational phases and factors inherent in global content localization and dissemination efforts, thereby ensuring the retention of the intended message across linguistic and cultural frontiers.

Keywords—Natural Language Processing, Automatic Speech Recognition, Machine Translation, Google Web Speech API, Google Translator

I. INTRODUCTION

India is a profoundly multi-lingual nation, boasting a rich tapestry of living languages, including 18 major ones that are constitutionally recognized. In our increasingly interconnected global society, effective communication fosters cross-cultural understanding. Imagine reaching millions beyond your own mother tongue, expanding your business or sharing ideas with an entirely new audience. Translation, here, is in most demand, may it be audio/video format. Dating back, traditional translation and caption generation methodologies were time consuming and involved skilled bunch of human translators. But with the raise of technological advancements, AI, ML to NLP and Cloud service approaches have been contemplated since late 20s. Any verbal, non- verbal form of information requires a sophisticated procedural alignment comprising various techniques; namely (a) ASR -Automatic Speech Recognition, (b) NLP- Natural language Processing, (c) MT-Machine Translation etc., for extraction and revival of information, further aiding audio/video translation, facilitating seamless understanding and connection among diverse communities.

The sole objective of the paper is to enable an ordered conversion of English language video/audio into Indian regional language of choice with a supportive high quality, systematic voice modulation, lip synchronization and accurate translation of speech in real time. Ensures an intuitive interface with customizable preferences for generation of localized video content.

II. RELATED WORKS

Researchers have been exploring Machine Translation for ages witnessing the accuracy after the recent evolution, the deep learning. Some of the related works that served as foundation for the suggested technique besides inspiring for taking up this research work is discussed.

In [1] ANGLABHARTI, a method for machine translation from English to numerous Indian languages is mentioned. Utilizing a pattern-directed methodology that integrates grammatical components devoid of context, it generates a pseudo-targe that is appropriate for a collection of Indian languages. Rules are acquired through corpus analysis, and semantic tags resolve sense ambiguity. The pseudo-target is translated into the target language via the text generator, which also includes a sentence structure correction process. The approach balances between transfer and interlingual but falls short of complete disambiguation. The prototype for English to Hindi aims for 90percent machine completion, leaving 10percent for human post-editing, facilitated by a designed interface. The system employs QUINTUS-PROLOG with 50 implemented rules, capable of translating most common sentences, producing multiple valid target sentences due to source language ambiguity. In paper [2] researchers introduces a Rule-Based Machine Translation (RBMT) system for translating English sentences to Malayalam, focusing on the cricket domain. Utilizing a transfer approach and Stanford parser for preprocessing, the system achieves an 86percent correct translation rate in testing. It employs a bilingual dictionary and a verb dictionary for inflections, contributing to meaningful translations. Despite successful outcomes, challenges arise with incorrect parse trees for lengthy sentences, leading to their division. The paper suggests refining the parser and dictionaries to enhance accuracy. The domain-specific RBMT system demonstrates potential for broader application across domains with the incorporation of additional rules and domain-specific words. Combining multiple domain-specific systems could create a comprehensive RBMT system for English to Malayalam translation.

In [3] work emphasizes the significance of providing educational content in one's mother tongue for effective and adaptive learning in eLearning platforms. It highlights the challenges in translating eLearning videos, citing manual coordination and involvement of various stakeholders as major obstacles. The paper acknowledges the recent advancements in Machine Learning and Natural Language Processing, suggesting that intelligent techniques can streamline the translation process. The implementation of machine learning in video translation demonstrates significant time reduction and resource optimization. The conclusion anticipates that machine translation using neural-based techniques will minimize manual effort, increase productivity, and address overlapping issues through separate subtitles. The proposed model, integrating machine learning and intelligent automation, is seen as a means to expedite the video translation process, with potential for further enhancement through intelligent splitting and post-editing.

In [4] the integration of popular screen readers (NVDA and ORCA) with Text-to Speech (TTS) systems for six Indian languages, addressing the needs of the visually challenged through a participatory design approach. A uniform framework for syllable-based TTS synthesis across Hindi, Tamil, Marathi, Bengali, Malayalam, and Telugu is developed, catering to India's multilingual context. Usability studies, including MOS and System Usability tests, were conducted, achieving a Mean Opinion Score (MoS) of 62.27 percent. The paper highlights efforts, challenges, and solutions in integrating TTS with screen readers, aiming to replace JAWS in training programs for the visually challenged community at IIT Madras. Active discussions during computer training sessions using these integrated systems are also depicted.

In [5] work outlines the integration of popular screen readers (NVDA and ORCA) with Text-to-Speech (TTS) systems for six Indian languages, addressing the needs of the visually challenged through a participatory design approach. A uniform framework for syllable-based TTS synthesis across Hindi, Tamil, Marathi, Bengali, Malayalam, and Telugu is developed, catering to India's multilingual context. Usability studies, including MOS and System Usability tests, were conducted, achieving a Mean Opinion Score (MoS) of 62.27 percent. The paper highlights efforts, challenges, and solutions in integrating TTS with screen readers, aiming to

replace JAWS in training programs for the visually challenged community at IIT Madras. Active discussions during computer training sessions using these integrated systems are also depicted.

Developing fully-automatic, high-quality machine translation systems for general purposes (FGH-MT) poses challenges owing to the creative and interpretative aspects inherent in translation [6]. No existing system qualifies as FGH-MT for any language pair, primarily because machines lack the capability to interpret texts accurately, especially considering audience and purpose variations. The Anusaaraka approach is proposed, breaking the MT system into two modules: A domain-specific module that is dependent on statistical data and world knowledge complements the basic Anusaaraka system, which is based on linguistic information. The output generated is close to the target language, requiring some training for human comprehension.

Experience with Indian languages suggests that training needs may be minimal due to linguistic similarities and shared cultural elements. This approach potentially be expanded to construct an Anusaaraka for English to Hindi, although additional training time might be required for practical use. The goal of the work being discussed is to enhance text-to-speech (TTS) quality [7] by using a feedback neural net- work to predict pitch frequency (F0) contours in Kannada speech. The model applies data driven approaches and uses the parameters of the Fujisaki intonation model as input characteristics. The festival architecture included a 4-layer feedback neural network. Both CART and neural network models are capable of reliably predicting F0 contours, ac- cording to performance comparisons with CART and Tilt models, with CART demonstrating higher naturalness and intelligibility. However, the neural network outperforms the Tilt model.

The research findings suggest that although the feedback neural network can effectively predict F0 contours, ensuring the production of high-quality speech requires the incorporation of additional factors, such as duration models. Overall, CART model synthesis exhibits superior naturalness compared to the neural network model.

III. Primary Components in the Proposed Framework

A. Objective

- Develop the software using Python and VS Code to enable seamless conversion of English language videos into Indian regional languages.
- Implement advanced natural language processing (NLP) techniques to accurately translate and synthesize speech in real- time.
- Optimize the software for efficient voice modulation and lip synchronization to enhance the overall quality of dubbed videos.
- Create a user-friendly interface for content creators, allowing them to customize dubbing preferences and easily generate localized video content
- Evaluate and refine the AI-based dubbing software through rigorous testing and user feedback to ensure high- quality and a culturally relevant output for a diverse Indian audience

B. Tools Deployed

A. movie.py

An audio extraction tool is software application, specifically to separate or extract audio from various file types, including video files and audio-visual me- dia. These tools enable users to isolate and save the audio component in a standalone format, typically in formats like MP3, WAV, or other common audio file types. The dubbed audio is reviewed and evaluated for accuracy, lip syncing, and overall quality. Any necessary adjustments or corrections are made to ensure a high-quality dubbing output..

B. Speech recognition

Speech recognition, sometimes referred to as voice recognition or automatic speech recognition (ASR), is a technique that transcribes spoken words into written language. Accurately identifying and transcription of human speech is the main objective of speech recognition systems. Speech recognition systems translate spoken words into written text by analyzing audio data, finding pattern in spoken languages, and applying machine learning techniques and algorithms.

C. gTTS (Google Text-to-Speech)/GoogleTrans)

GoogleTrans can be employed to translate the original script or dialogue from one language to another. The translated text can be processed to ensure accuracy and naturalness in the target language. Once the translated text is available, gTTS can be utilized to generate synthetic speech for the translated content. Users can specify parameters such as voice type, speed, and intonation to match the desired characteristics for the dubbed content. The generated speech can then be synchronized with the corresponding video or audio segments to create the dubbed version of the content.

D. pyttsx3

Voice Synthesis: The translated text is converted into spoken words in the target language using voice synthesis technology. This step generates audio output which sounds like a native speaker of the target language.

E. pyaudio

Audio Mixing: The software combines the original audio with the synthesized voice in the target language, adjusting the timing and volume levels to create a seamless dubbed audio track.

F. Google Web Speech API:

The Google Web Speech API, now part of the broader Web Speech API, enables developers to integrate speech recognition capabilities into web applications. It allows users to interact with websites using voice commands and speech-to -text transcription. Developers can access this feature through JavaScript APIs, facilitating voice-driven user experiences across various platforms and devices[8].

IV. FUNCTIONALITY OF THE MODEL

A. Setting Up the Environment:

- Set up VS Code and Python
- Install the ‘speechrecognition, ‘gTTS‘, ‘pyttsx3‘, and‘ pyaudio libraries,as needed.
- Create credentials for the Google Web Speech API.

B. Understanding the Movie Module (‘movie.py‘)

The ‘movie.py‘ module contains functions to handle video processing. It utilizes libraries like OpenCV to read and manipulate video files. Functions include loading video files, extracting audio tracks, and syncing audio with video frames.

C. Implementing Speech Recognition

Use the ‘speech.recognition‘ library to transcribe spoken words from video. Utilize its functions to capture and convert speech into text format. Handle various languages and accents for improved accuracy

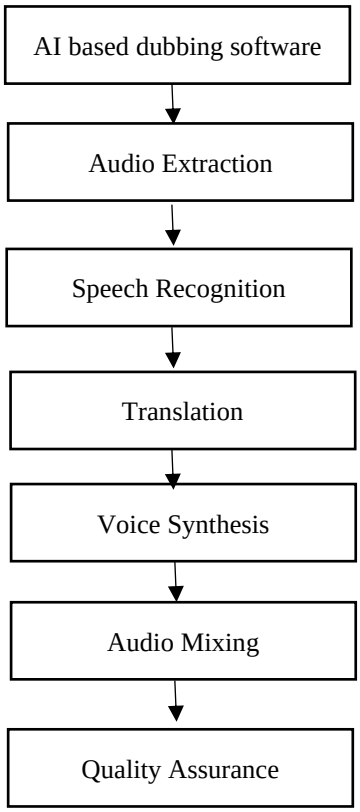

Fig. 1. Workflow diagram

D. Integrating Google Text-to-Speech (gTTS) and Google- Trans

- Utilize ‘gTTS‘ to convert the translated text into synthesized speech.
- Leverage ‘GoogleTrans‘ to perform language translation if needed.
- Ensure translated text maintains synchronization with the original video’s timings.

The work flow diagram us shown in the Fig.1. It depicts the different stages of translation.

E. Employing pyttsx3 for Offline Text-to-Speech Conversion

- Use ‘pyttsx3‘as an alternative text-to-speech engine for offline processing.
- Customize speech parameters such as voice pitch, rate, and volume.
- Ensure compatibility and smooth integration with the dubbing process.

F. Managing Audio Input and Output with PyAudio

- Utilize ‘pyaudio‘ to handle audio input and output functionalities.
- Capture the user’s voice for dubbing purposes.
- Streamline audio playback and recording processes.

G. Leveraging Google Web Speech API

- Integrate Google Web Speech API [10] for advanced speech recognition capabilities.
- Enhance accuracy and performance in recognizing speech patterns and accents.
- Leverage cloud-based processing for handling complex speech data.

Fig. 2. Step by step process of translation

The step by step work flow process in Fig.2. is as follows:

- Load the original video file into the dubbing software.
- Transcribe the original audio using speech recognition.
- Translate the transcribed text, if necessary, using Google-Trans.
- Convert the translated text into synthesized speech using gTTS or pyttsx3.
- Overlay the synthesized speech onto the video frames, ensuring synchronization.
- Save the dubbed video with the new audio track

TABLE I. IMPORTANCE OF AI BASED MODEL

Feature	**AI based dubbing model**	**Conventional model**
Customization Options	Extensive options for tone and pacing	Limited options for customization
Cost-efficiency	Optimized price model	Fixed/Expensive
User Interface	Friendly and Intuitive	Complex and former

Feature	AI based dubbing model	Conventional model
Integration Capabilities	Seamless integration of video and content	Lacks Integration capabilities
Multilingual	Supports wide range of languages	Limited language option

V. Outcomes

A. *Advantages of AI based Dubbing Software*

The software offers benefits in different sectors [9] in different ways.

- Language Accessibility
- Cost Efficiency
- Localization
- Scalability
- Consistency

B. *Multifaceted Benefits Across Diverse Domains*

a. *Entertainment Industry*

AI-based dubbing software fa- cilitates the localization of movies, TV shows, and digital content, allowing for wider global distribution and audience engagement. It enables seamless dubbing and subtitling processes, enhancing accessibility and enjoyment for viewers worldwide.

b. *Education Sector*

Dubbing software assists in creating educational videos and tutorials in multiple languages, catering to diverse student populations. It enables the translation of educational content, making it more accessible to non-native speakers and promoting inclusive learning environments.

c. *Corporate Training:*

AI-powered dubbing software helps in producing training materials and corporate presentations in various languages, enabling multinational companies to deliver consistent messaging across their global workforce. It enhances employee engagement and comprehension by providing training content in the learners' preferred language.

d. *Media and Advertising*

AI-based dubbing software enables advertisers and marketing agencies to create localized advertisements and promotional content for different regions and target audiences. It helps in maintaining brand consistency while adapting marketing messages to local languages and cultural nuances.

e. *Accessibility and Inclusion*

Dubbing software enhances accessibility for individuals with hearing impairments by pro- viding audio descriptions and dubbing options in different languages. It promotes inclusivity by ensuring that multimedia content is accessible to individuals with diverse linguistic and cultural backgrounds.

f. *Cultural Exchange and Diplomacy*

AI-powered dubbing software fosters cultural exchange and diplomacy by facilitating the translation and dubbing of diplomatic speeches, conferences, and cultural events. It promotes cross-cultural understanding and communication by breaking language barriers and enabling effective communication between nations and communities.

VI. Results and Discussion

As the program code runs, primarily the audio is separated and stored in a distinct file. This is accomplished by usage of movie.py tool which is efficient in extraction of audio from various file types. This enables users to isolate the audio component typically in formats like MP3 (or) WAV. Further, a copy of video is stored without audio. This video sample is available under Windows C - Videos. [Fig.4, Fig.5]. Target Language is asked for input which undergoes various steps.

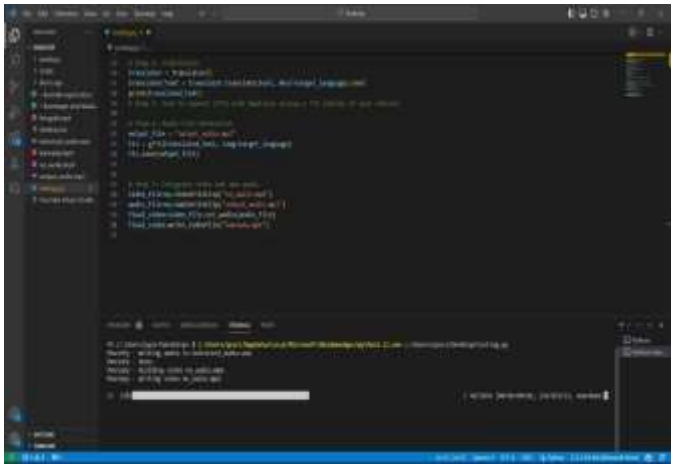

Fig.4. Processing Video without audio

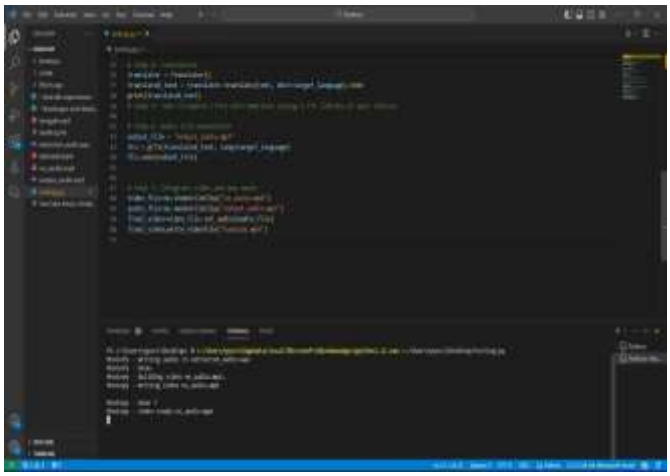

Fig.5. Video without audio output

A. *Implementation of Speech Recognition*

Automatic Speech Recognition [ASR] technique transcribes human speech into written language with utmost accuracy by analysing auto data incorporating machine learning techniques and algorithms.

B. *Integration of Google Text-to-Speech(gTTS) and Google-Trans*

GoogleTrans can be employed to translate the original script or dialogue from one language to another and the

translated text can be processed to ensure accuracy and naturalness of the target language.

Google Text-to-Speech(gTTS) can be utilized to generate synthetic speech for the translated content. Various specific parameters such as voice type, speed, and intonation to match the desired characteristics for the dubbed content can be customized by the user. Generated speech is then synchronized with the corresponding video or audio segments to create the dubbed version of the content.

C. Implementation of pyttsx3

Generated audio output similar to native speaker of the defined target language is shown in Fig.6 and Fig.7. The Synthesized Speech obtained as a result of above procedure now undergoes audio mixing with the pyaudio software to fine tune timing and volume levels. These are overlayed onto video frames ensuring synchronization. Google Web Speech API is leveraged to enhance accuracy and performance in recognizing speech patterns and accents. Additionally leverage cloud-based processing for handling complex speech data is also done in Fig.8.

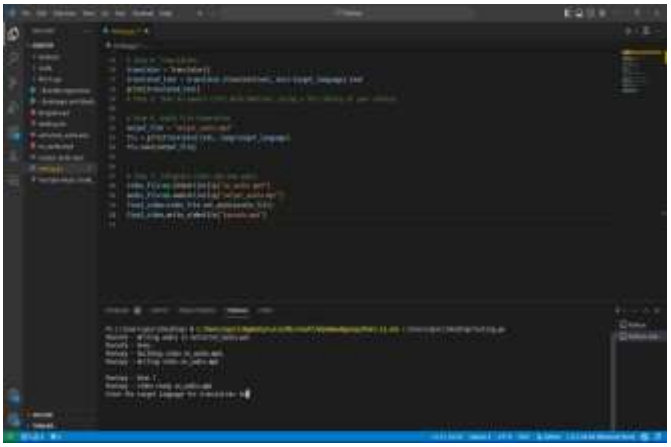

Fig.6. Target language selection and processing

Fig.7. Target language Conversion

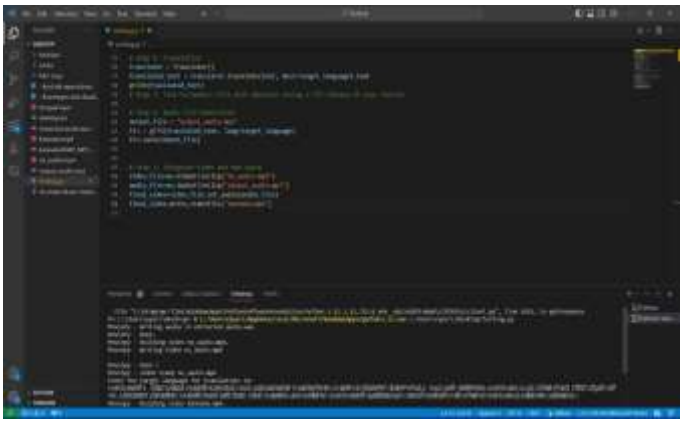

Fig.8. Fine tuning of timing and volume levels

VII. CONCLUSION

In this approach, the process of dubbing stands as an indispensable tool for fostering cross-cultural communication and accessibility in the realm of multimedia content. This paper underscores the critical significance of dubbing in facilitating intercultural adaptation, multilingual marketing, and preserving the original intentions of the content creators. Emphasizing the importance of retaining the essence of the original material, this paper highlights the fundamental phases and considerations involved in the art of dubbing, ultimately championing its role in making information accessible and resonant across diverse linguistic backgrounds.

REFERENCES

[1] Sinha, Rohit Mahesh K., et al. ”ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages,” IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, Vol. 2, IEEE, 1995.

[2] Aasha, V. C, and Amal Ganesh, ”Machine translation from English to Malayalam using transfer approach, ”2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)”, IEEE, 2015.

[3] Shenbagaraj, Shenbagaraj, and Sailesh Iyer, ”An Intelligent System for Automated translation of Videos from English to Native Language applying Artificial Intelligence Techniques for Adaptive eLearning,” International Journal of Intelligent Systems and Applications in Engineering 12.3s (2024), pp. 620-640.

[4] Karthikadevi M, and K. G. Srinivasagan, ”The development of syllable-based text to speech system for Tamil language,” 2014 International Conference on Recent Trends in Information Technology, IEEE, 2014.

[5] Kurian, Anila Susan, et al. ”Indian language screen readers and syllable- based festival text-to-speech synthesis system,” Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, 2011.

[6] Bharati, Akshar, et al. ”Anusaaraka: Machine translation in stages.” VIVEK-BOMBAY- 10,1997, pp. 22-25.

[7] Trilla, Alexandre, and Francesc Alias. ”Sentence-based sentiment analysis for expressive text-to-speech,” IEEE transactions on audio, speech, and language processing, 21.2, 2012, pp. 223-233.

[8] Tushar Chikte, Abhijeet Hatte, Khushboo Bavsar, Dr. N. N. Khalsa, ”Talk With AI,” International Research Journal of Modernization in Engineering Technology and Science, Vol. 5, 2023.

[9] Berrak Sisman, Junichi Yamagishi, Simon King, Haizhou Li, “An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning”, IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 29, 2021.

[10] William Brannon, Yogesh Virkar, Brian Thompson, “Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing,” Transactions of the Association for Computational Linguistics, Vol.11,2023, pp. 419–435.