

Translingo: AI Powered Multilingual Translator

V.Ashok Gajapathi Raju¹, Jami Jasmitha², Vajja Satwika³, Gangi Vineela⁴, Kanchari Uday Vinay Maharana⁵

¹Assistant Professor, Information Technology, Aditya Institute of Technology and Management, Tekkkali, Andhra Pradesh, India

^{2,3,4,5}UG Students, Information Technology, Aditya Institute of Technology and Management, Tekkkali, Andhra Pradesh, India

Abstract

In an increasingly more globalized global, seamless conversation across language limitations is crucial. The undertaking affords specific Lingo, an AI video voice translation platform leveraging automatic Speech recognition (ASR), Neural device Translation (NMT), and text-to-Speech (TTS) technology. not like traditional translation structures, explicit Lingo presents actual-time, context-aware, and adaptive translations, making it notably appropriate for industries. the integration of transformer-based models complements contextual knowledge, while its patentable layout gives substantial commercialization capacity. Translingo is constructed with a modular architecture that supports a couple of languages and permits for non-stop gaining knowledge of, ensuring stepped forward translation accuracy over time. furthermore, the platform contains noise reduction algorithms and accessory normalization strategies to decorate speech recognition accuracy.

Keywords: Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Text-to-Speech (TTS), Multilingual Translation, AI Localization, Real-Time Voice Translation, Video Accessibility, Natural Language Processing (NLP).

1. Introduction

The Project, "Translingo: AI Powered Multilingual Translator," ambitions to bridge language boundaries via imparting real time voice and video translations the use of transformer based totally device getting to know fashions. It integrates computerized Speech recognition (ASR), Neural Machine Translation (NMT), and textual content-to-Speech (TTS) to deliver seamless, context-aware translations, making multilingual verbal exchange handy.

The number one goal is to increase and put in force a system capable of translating video and audio content into more than one language with high accuracy and low latency. The challenge includes audio extraction, speech-to-text conversion, language translation, and speech synthesis, ensuring clean and herbal-sounding outputs. the use of deep mastering fashions like Wav2Vec 2.zero, T5, the machine methods audio in real-time at the same time as minimizing noise and slang disturbances for greater clarity.

The task holds significant programs in training, commercial enterprise, healthcare, and global conversation, allowing seamless interactions throughout unique languages. The method follows established pipeline, inclusive of audio processing, device translation, and real-time speech synthesis. by means of combining pre-trained transformer's and adaptive getting to know techniques, the device continuously improves its translation accuracy and enhancing accessibility and breaking down language boundaries throughout more than one Industries.

The expected outcome is a completely purposeful AI-powered translation platform, evaluated based on accuracy, latency, and value. Its success will contribute to technological improvements in speech translation, improving accessibility and verbal exchange throughout international groups.

2. Literature Survey

Dr. Shabina Modi et al., developed Dubify, a real-time multilingual video translation platform. Using machine learning and NLP, Dubify translates spoken words into text, then into multiple languages, while synthesizing audio and preserving cultural nuances, making video content globally accessible.[1]

Marcus Rohrbach et al., created a system that generates natural language descriptions of video content by combining visual recognition and language generation. The system, tested on the TACoS dataset, provided improved accuracy and human-like descriptions using CRFs and statistical machine translation. [2]

Muhammad Usman Tariq Abu explored the impact of AI-powered translation tools in multilingual classrooms. The study highlighted how neural network advancements enhance accessibility, inclusivity, and communication across diverse linguistic groups, promoting equal learning opportunities.[3]

Kolhar and Alameen developed a real-time translation system for non-native Arabic-speaking teachers in Saudi classrooms. The system, integrated with digital podiums, helps improve comprehension and engagement in English-taught subjects, with positive feedback from students.[4]

Ann Lee et al., introduced a direct speech-to-speech translation model, bypassing text generation. The model demonstrated significant improvements on the Fisher Spanish-English dataset, showcasing its potential for unwritten language translation.[5]

Ye Jia et al., developed a sequence-to-sequence model for speech-to-speech translation. The model translates speech spectrograms into target spectrograms, preserving the speaker's voice, with advantages in latency and paralinguistic feature retention.[6]

Keyu et al., developed FunAudioLLM, a framework for human-LLM interaction through voice, incorporating SenseVoice for speech recognition and CosyVoice for natural voice generation. It supports multilingual speech, emotion detection, and voice cloning for various applications. [7]

Barrault et al., introduced SeamlessM4T, a unified multilingual model supporting 100 languages across text and speech. It utilizes multimodal datasets to improve translation quality, setting new benchmarks in multilingual communication systems.[8]

Bigioi and Corcoran reviewed the state of automated multilingual video dubbing. They discussed various methodologies and challenges, particularly in data management and synchronization, and emphasized the need for improved models to enhance dubbing quality.[9]

Linden Wang implemented an automated translation system for Khan Academy videos to improve accessibility for non-English speakers. The system integrated ASR, machine translation, and TTS technologies to produce synchronized translations. User corrections enhanced translation quality, demonstrating its potential to bridge educational gaps globally.[10]

3. Existing System

Modern-day AI video voice translation systems combine more than one additives to facilitate seamless multilingual communication. Automatic Speech Recognition (ASR) era converts spoken language into text, allowing correct transcription of video content. Neural Machine Translation (NMT) models, including transformer-based totally architectures process the text and provide context-conscious, high-accuracy translations.

Textual content-to-Speech (TTS) structures, such as Tacotron and FastSpeech, synthesize human-like speech, ensuring natural and expressive voice output. Actual-time processing Advanced noise reduction and accessory normalization techniques enhance ASR performance, addressing versions in speech clarity, slang, and nearby dialects. Adaptive mastering mechanisms constantly refine translation accuracy based on person interactions and feedback.

Systems vary based on language guide, processing speed, and computational resources. Technological improvements retain to force improvements in real-time synchronization and cultural model.

4. Methodology

The proposed methodology for "Translingo: AI Powered Multilingual Translator" is an integration of our best in the class artificial intelligence (AI) technologies to provide an easy, convenient and accurate solutions for multilingual video voice translation. Methodology is comprised of multi-stage pipeline to process the input video files, convert it to spoken content and transcribe it into the desired language and then synthesize it to the natural sounding audio. The system focuses on real time processing, accuracy and user customization.

4.1 Automatic Speech Recognition (ASR)

The process begins with the speech extraction from the input video through state-of-the-art ASR models like Whisper or Kaldi. These models are pre-trained on large-scale multilingual corpora, enabling them to process various accents, speech rates, and background noise effectively. The ASR module provides high-accuracy text transcripts with timestamps, which are essential for lip-sync and subtitle alignment in the subsequent stages.

ASR Process Equation:

$$\hat{W} = \arg \max_W P(X|W) \cdot P(W) \quad (1)$$

Where:

- $P(X|W)$: Acoustic model — likelihood of audio features X given word sequence W .
- $P(W)$: Language model — probability of word sequence W .

4.2 Neural Machine Translation (NMT)

After transcription is finished, the text is input into a Neural Machine Translation system, which translates it into the target language. NMT models such as OpenNMT utilize attention mechanisms and transformer-based architecture to learn about sentence structure and maintain meaning.

The translation process predicts a sequence of target words $y=(y_1,y_2,...,y_T)$

$y = (y_1, y_2, ..., y_T)$, $y=(y_1,y_2,...,y_T)$

given a source sentence $x=(x_1,x_2,...,x_S)$ $x = (x_1, x_2, ..., x_S)$ $x=(x_1,x_2,...,x_S)$.

This is modeled as:

$$P(y|x) = \prod_{t=1}^T P(y_t|y_{<t}, x) \quad (2)$$

Where Each word y_t is predicted based on:

- Previous words $y_{<t}$
- And the entire input sentence x .

4.3 Text-to-Speech (TTS) Synthesis

The system must generate natural-sounding, expressive voiceovers in multiple languages using neural TTS models. Users should be able to customize voice parameters (accent, gender, tone, speaking style).

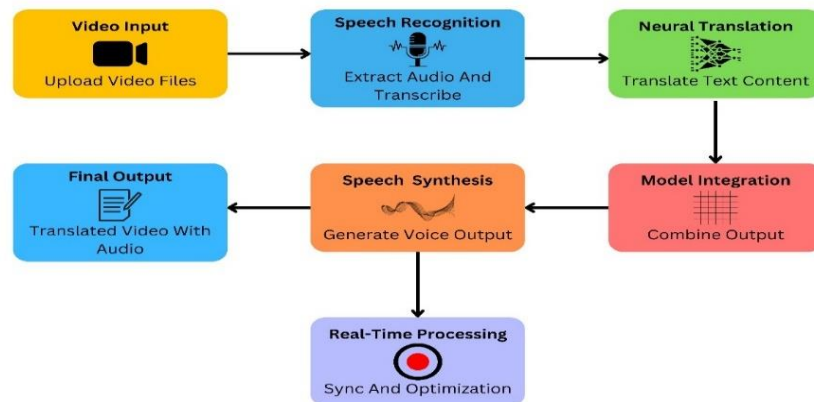
4.4 Real-Time Processing and Synchronization

The proposed system facilitates real-time operation, enabling instantaneous translations or retranslations of live or pre-recorded videos. Additionally, it implements real-time voice synthesis for both live and pre-recorded content. To ensure seamless synchronization, the timestamps generated during the Automatic Speech Recognition (ASR) stage are matched with the audio output from the Text-to-Speech (TTS) stage. This harmonization of video frames and translated audio guarantees a natural and timely voiceover that perfectly complements the original video content, creating an unobtrusive and immersive viewing experience.

4.5 Continuous Improvement and Feedback

The system also incorporates a feedback loop for continuously improving long-term reliability and performance. The ASR/NMT/TTS models that we use are retrained with feedback collected from users and metrics of their performance. Through a series of iterations, the system can learn from real-world use and come to understand new processes.

Fig 5 Methodology of video voice translator



5. Designing The System

5.1 Architecture And Components

The Translingo system is designed with a scalable and modular architecture to ensure efficiency, flexibility, and adaptability. It integrates multiple AI components to enable real-time video translation by leveraging speech recognition, neural translation, and speech synthesis.

Video and Audio Processing: Users upload a video file to the system. The system extracts the audio from the video while preserving the original quality.

Speech Recognition (ASR): Converts spoken content in the audio into text. Utilizes deep learning models for accurate transcription.

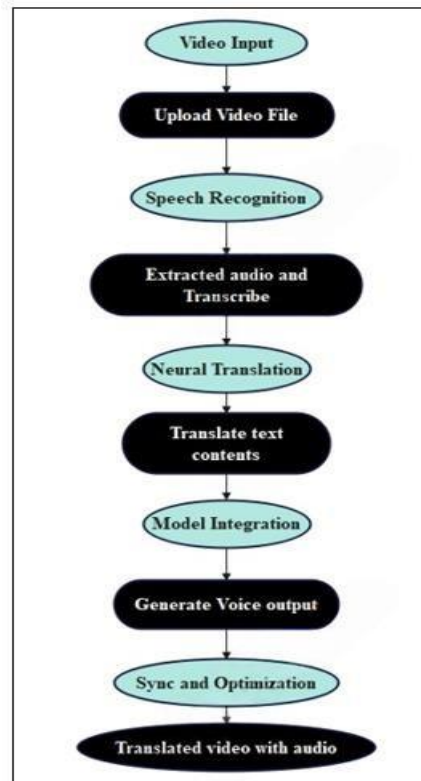
Neural Machine Translation (NMT): Translates the transcribed text into the target language. Uses transformer-based for high-quality translation.

Text-to-Speech (TTS) Synthesis: Converts the translated text into natural-sounding speech. Implements AI-powered TTS models.

Sync and Optimization: The newly generated voice is synchronized with the video while maintaining lip sync accuracy.

Translated Video with Audio: The final video is generated with translated voice output. Users can download or stream the translated video.

Fig 6.1 System Architecture of Translingo: AI video voice translation system



6. Results and Discussion

6.1 Translation Accuracy

BLEU (Bilingual Evaluation Understudy) score was used to evaluate translation module's accuracy, the score measures the to what extent the translated text comes close to human translations of the references. In Table 1, we summarize BLEU scores of different language pairs run on a set of video datasets. The results achieved as consistently high BLEU scores suggest that it is able to generate contextually accurate translations

Table 1: BLEU Scores for language pairs

Language pair	BLEU Score (%)
English → Spanish	85.6
English → Mandarin	83.2
English → French	87.1
English → Arabic	81.4
English → Hindi	82.9

6.2 Latency Analysis

Latency was measured as the time taken by the system to process a one-minute video for translation and dubbing. Table 2 provides a comparative analysis of Translingo's latency against existing translation

systems. The results demonstrate the system's efficiency, with significantly lower latency due to its optimized AI models.

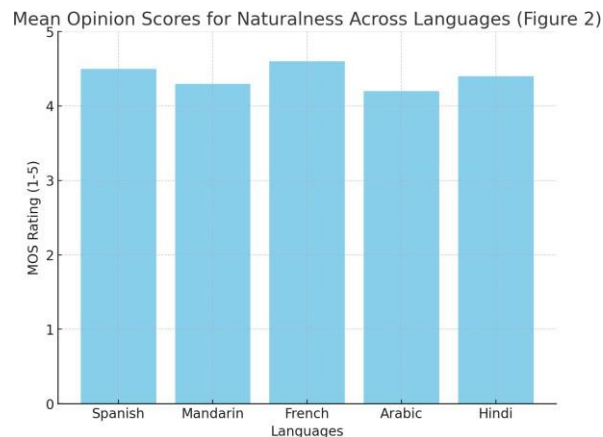
Table 2: Latency Comparison with Existing Systems

System	Latency(seconds)
Translingo	5.4
System A	9.2
System B	11.7
System C	8.6

6.3 Naturalness of Voice Synthesis

To assess the quality of voice synthesis, a Mean Opinion Score (MOS) survey was conducted with participants rating the naturalness of the synthesized voice on a scale of 1 (poor) to 5 (excellent) . The graph will show MOS ratings for languages such as Spanish, Mandarin, French, Arabic, and Hindi, with most scores falling in the range of 4 to 5.

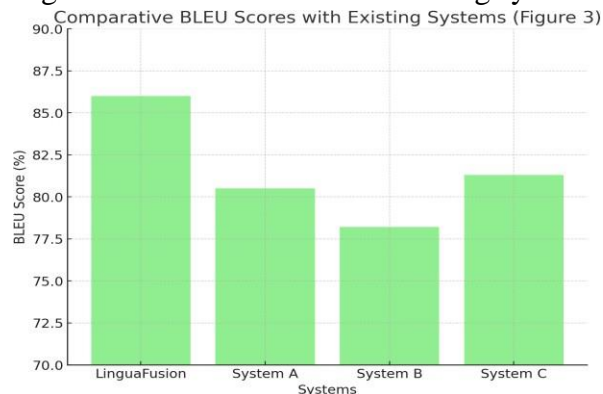
Figure 1: MOS for voice synthesis



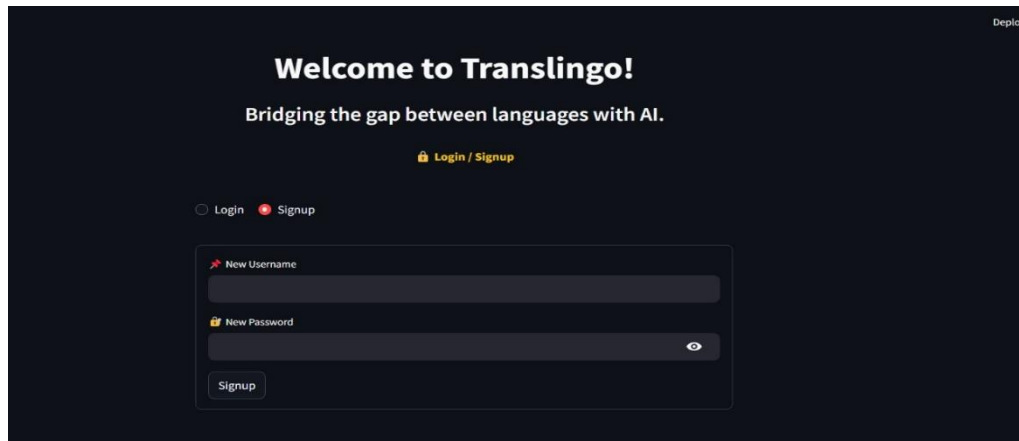
6.4 Comparative Analysis of System Performance

A comparative analysis was performed to evaluate Translingo against leading translation tools in terms of translation accuracy and synthesis naturalness.

Figure 2: BLEU Scores with existing systems



To provide a real-time demonstration of the Translingo system, we have developed a dedicated project website. This interactive platform allows users to upload videos, select source and target languages, and receive translated voice-over output directly in their browser. The website showcases the working of core components including Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) synthesis in a seamless end-to-end pipeline. The interface is user-friendly and designed to support multiple formats, making it ideal for testing and educational purposes. The web-based implementation emphasizes the practical utility and scalability of Translingo in real-world applications.



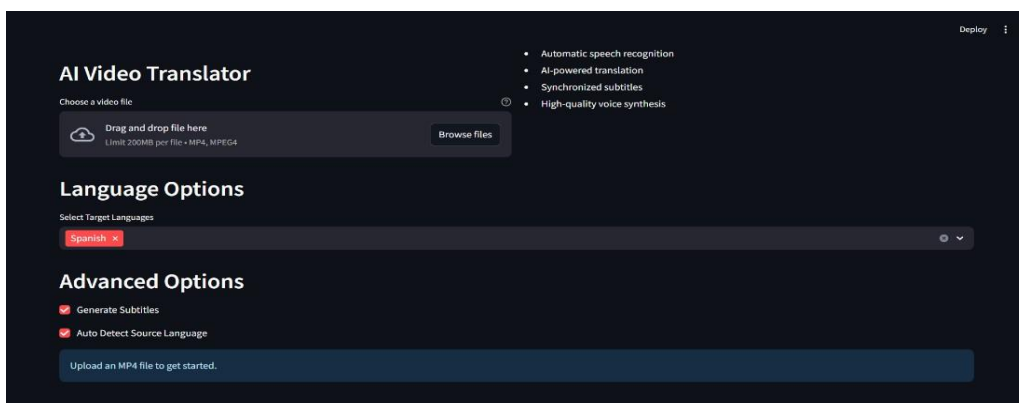
The screenshot shows the 'Welcome to Translingo!' login page. It features a dark blue background with white text. The main heading is 'Welcome to Translingo!' followed by the tagline 'Bridging the gap between languages with AI.' Below this is a 'Login / Signup' link. There are two radio buttons for 'Login' and 'Signup', with 'Signup' selected. The 'Signup' form includes fields for 'New Username' and 'New Password' (with a toggle for visibility), and a 'Signup' button. A 'Deploy' link is visible in the top right corner.

Login Page



The screenshot shows the 'Welcome to Translingo!' home page. It features a dark blue background with white text. The main heading is 'Welcome to Translingo!' followed by the tagline 'Bridging the gap between languages with AI.' Below this are 'Home' and 'AI' buttons, and a 'Logout' button. The 'How It Works' section lists five steps: 1. Upload your video file (MP4 format), 2. Select target languages for translation, 3. Configure optional settings in the sidebar (for non-English/Hindi videos, manually select the source language), 4. Click 'Start Translation', and 5. Download your translated files. The 'Features' section lists multi-language support (including English, Spanish, French, German, Chinese, Japanese, Telugu, Hindi, Tamil, Kannada). A sidebar shows language options: English, Español, Português, 한국어, Русский, Deutsch, Bahasa Indonesia, and Français. A 'Deploy' link is visible in the top right corner.

Home Page



The screenshot shows the 'AI Video Translator' configuration screen. It features a dark blue background with white text. The main heading is 'AI Video Translator'. Below this is a 'Choose a video file' section with a 'Drag and drop file here' area (limit: 200MB per file, MP4, MPEG4) and a 'Browse files' button. The 'Language Options' section has a 'Select Target Languages' dropdown menu with 'Spanish' selected. The 'Advanced Options' section has two checkboxes: 'Generate Subtitles' and 'Auto Detect Source Language', both of which are checked. A 'Deploy' link is visible in the top right corner.

AI Video Translator – Uploading video and Configuration Screen

References

1. Dr. Shabina Modi, Mr. Aniket Kamble, Mr. Pratik Gawande, and Ms. Pritee Ithape, “Dubify: Multilingual Video Translation Platform,” *IJARST*, pp. 791–796, May 2024, doi: 10.48175/IJARST-18389.
2. M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, “Translating Video Content to Natural Language Descriptions,” in *2013 IEEE International Conference on Computer Vision*, Sydney, Australia: IEEE, Dec. 2013, pp. 433–440. doi: 10.1109/ICCV.2013.61.
3. M. U. Tariq, “AI-Powered Language Translation for Multilingual Classrooms:,” in *Advances in Educational Technologies and Instructional Design*, R. Doshi, M. Dadhich, S. Poddar, and K. K. Hiran, Eds., IGI Global, 2024, pp. 29–46. doi: 10.4018/979-8-3693-2440-0.ch002.
4. M. Kolhar and A. Alameen, “Artificial Intelligence Based Language Translation Platform,” *Intelligent Automation & Soft Computing*, vol. 28, no. 1, pp. 1–9, 2021, doi: 10.32604/iasc.2021.014995.
5. A. Lee et al., “Direct speech-to-speech translation with discrete units,” Mar. 21, 2022, arXiv: arXiv:2107.05604. doi: 10.48550/arXiv.2107.05604.
6. Y. Jia et al., “Direct speech-to-speech translation with a sequence-to-sequence model,” Jun. 25, 2019, arXiv: arXiv:1904.06037. doi: 10.48550/arXiv.1904.06037.
7. K. An et al., “FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs,” Jul. 11, 2024, arXiv: arXiv:2407.04051. doi: 10.48550/arXiv.2407.04051.
8. S. Communication et al., “SeamlessM4T: Massively Multilingual & Multimodal Machine Translation,” Oct. 25, 2023, arXiv: arXiv:2308.11596. doi: 10.48550/arXiv.2308.11596.
9. D. Bigioi and P. Corcoran, “Multilingual video dubbing—a technology review and current challenges,” *Front. Signal Process.*, vol. 3, p. 1230755, Sep. 2023, doi: 10.3389/frsip.2023.1230755.
10. L. Wang, “Applying automated machine translation to educational video courses,” *Educ Inf Technol*, vol. 29, no. 9, pp. 10377–10390, Jun. 2024, doi: 10.1007/s10639-023-12219-0.
11. C. Le et al., “TransVIP: Speech to Speech Translation System with Voice and Isochrony Preservation,” 2024, arXiv. doi: 10.48550/ARXIV.2405.17809.
12. R. S. A. Pratama and A. Amrullah, “ANALYSIS OF WHISPER AUTOMATIC SPEECH RECOGNITION PERFORMANCE ON LOW RESOURCE LANGUAGE,” *pilar*, vol. 20, no. 1, pp. 1–8, Mar. 2024, doi: 10.33480/pilar.v20i1.4633.
13. Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, “Translatotron 2: High-quality direct speech-to-speech translation with voice preservation,” 2021, arXiv. doi: 10.48550/ARXIV.2107.08661.
14. A. Waibel et al., “Face-Dubbing++: Lip-Synchronous, Voice Preserving Translation of Videos,” 2022, arXiv. doi: 10.48550/ARXIV.2206.04523.
15. P. K. R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar, “Towards Automatic Face-to-Face Translation,” 2020, doi: 10.48550/ARXIV.2003.00418.
16. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep Audio-Visual Speech Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022, doi: 10.1109/TPAMI.2018.2889052.
17. S. K. Pulipaka, C. K. Kasaraneni, V. N. Sandeep Vemulapalli, and S. S. Mourya Kosaraju, “Machine Translation of English Videos to Indian Regional Languages using Open Innovation,” in *2019 IEEE*



International Symposium on Technology and Society (ISTAS), Medford, MA, USA: IEEE, Nov. 2019, pp. 1–7. doi: 10.1109/ISTAS48451.2019.8937988.

18. Z. Nabila, H. R. Ayu, and A. Surtono, “Implementation of Google Translate Application Programming Interface (API) as a Text and Audio Translator,” *CoreIT*, vol. 8, no. 1, p. 19, Jun. 2022, doi: 10.24014/coreit.v8i1.15629.
19. J. Lee et al., “Challenge on Sound Scene Synthesis: Evaluating Text-to-Audio Generation,” 2024, arXiv. doi: 10.48550/ARXIV.2410.17589.